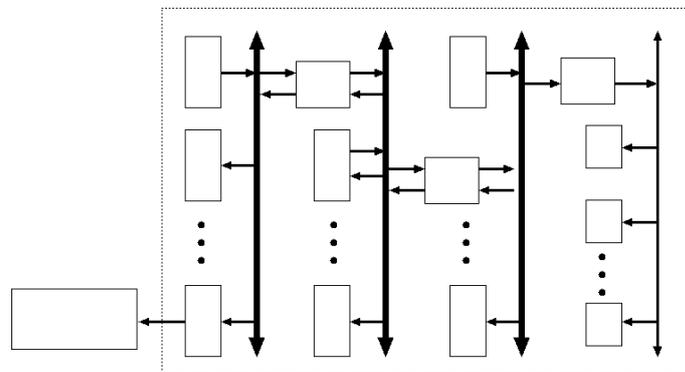


# System on Chip Design and Modelling

## D&M Patterns Portfolio



University of Cambridge  
Computer Laboratory

Dr. David J Greaves

(C) 2008-18 All Rights Reserved DJG.

Printed January 2018

- (1) **Basic SoC Components**
- (2) **Energy use in Digital Hardware.**
- (3) **Masked versus Reconfigurable Technology & Computing**
- (4) **RTL, Interfaces, Pipelining and Hazards**
- (5) **Formal Methods and Assertion-Based Design**
- (6) **Architectural Exploration using ESL Modelling**
- (7) **High-level Design Capture and Synthesis**
- (8) **SystemC: Hardware Modelling Library**
- (9) **SoC Engineering and Associated Tools**
- (10) **Custom Accelerator Structures**

### **0.0.1 System On Chip Design And Modelling Patterns - Portfolio Notes Printout**

This document consists of lecture notes and other materials accumulated over many years and all bundled together in one PDF file. Some parts may now be only of historical interest. In any given year or course, only about half of this material was lectured.

A current-day system on a chip (SoC) consists of several different microprocessor subsystems together with memories and I/O interfaces. This course covers SoC design and modelling techniques with emphasis on architectural exploration, assertion-driven design and the concurrent development of hardware and embedded software. This is the “front end” of the design automation tool chain. (Back end material, such as design of individual gates, layout, routing and fabrication of silicon chips is not covered.)

### **0.0.2 Recommended Reading**

Subscribe for webcasts from ‘Design And Reuse’: [www.design-reuse.com](http://www.design-reuse.com) Ed Lipianski’s Book

*Embedded Systems Hardware for Software Engineers* Ed Lipiansky. McGraw-Hill Professional (1 Feb. 2012)

Mishra K. (2014) *Advanced Chip Design - Practical Examples in Verilog*. Pub Amazon Martson Gate.

Multicore field-programmable SoC: Xilinx Zync Product Brief

Atmel, ARM-based Embedded MPU AT91SAM Datasheet

OSCI. *SystemC tutorials and whitepapers* . Download from OSCI [www.accelera.org](http://www.accelera.org) or copy from course web site.

Brian Bailey, Grant Martin. *ESL Models and Their Application: Electronic System Level Design*. Springer.

*The Simple Art of SoC Design - Closing the gap between ESL and RTL* Michael Keating (2011), Springer ISBN 978-1-4419-8586-6

Ghenassia, F. (2006). *Transaction-level modeling with SystemC: TLM concepts and applications for embedded systems* . Springer.

Eisner, C. & Fisman, D. (2006). *A practical introduction to PSL* . Springer (Series on Integrated Circuits and Systems).

Foster, H.D. & Krolnik, A.C. (2008). *Creating assertion-based IP*. Springer (Series on Integrated Circuits and Systems).

Grotker, T., Liao, S., Martin, G. & Swan, S. (2002). *System design with SystemC*. Springer. E-BOOK (PDF)

Wolf, W. (2002). *Modern VLSI design (System-on-chip design)*. Pearson Education. Old now. Link broken. LINK.

### 0.0.3 Example: A Cellphone.

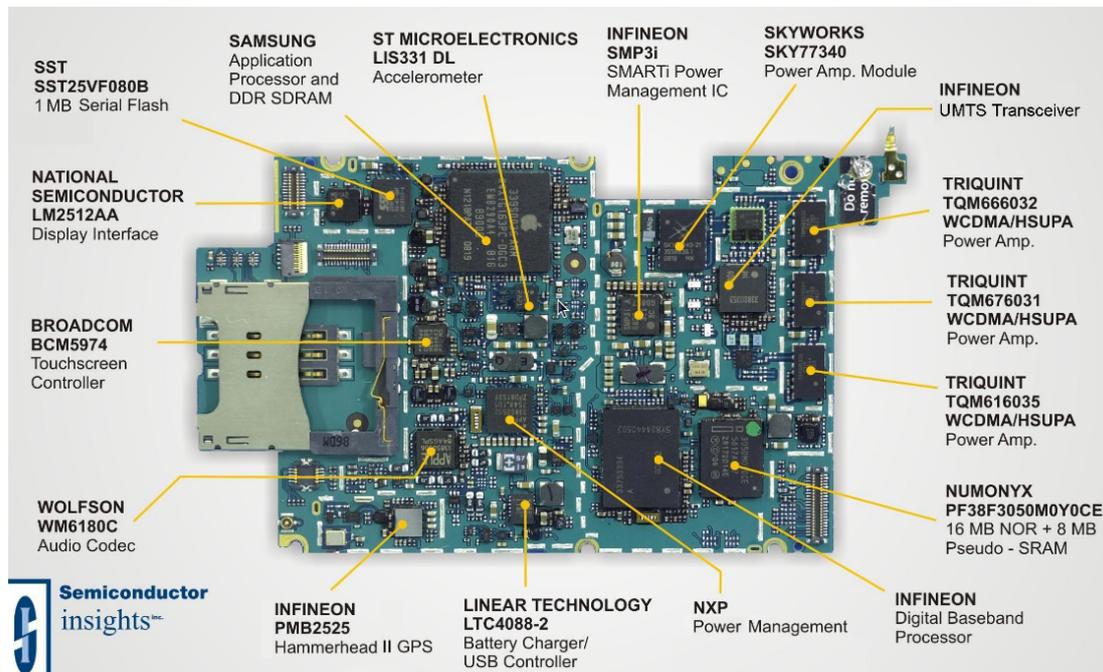


Figure 1: One of the two main PCBs of an Apple iPhone. Main SoC is top, left-centre.

A modern mobile phone contains eight or more radio transceivers, counting the various cellphone standards, GPS, WiFi, near-field and Bluetooth. For the Apple iPhones, all use off-SoC mixers and some use on-SoC ADC/DAC. Another iPhone teardown link

Further examples: [iFixit Teardowns iPhone Dissected](#)

[Samsung Galaxy](#)[Numonyx Flash Datasheet](#)[Cellphone physical components - bill of materials:](#)

- Main SoC - Application Processor, Caches and DRAM
- Display (touch sensitive) + Keypad + Misc buttons
- Audio ringers and speakers, microphone(s) (noise cancelling),
- Infra-red IRDA port
- Multi-media codecs (A/V capture and replay in several formats)
- Radio Interfaces: GSM (three or four bands), BlueTooth, 802.11, GPS, Nearfield, .... plus antennas.
- Power Management: Battery Control, Processor Speed, on/off/flight modes.
- Front and Rear Cameras, Flash/Torch and ambient light sensor,
- Memory card slot,
- Physical connectors: USB, Power, Headset,

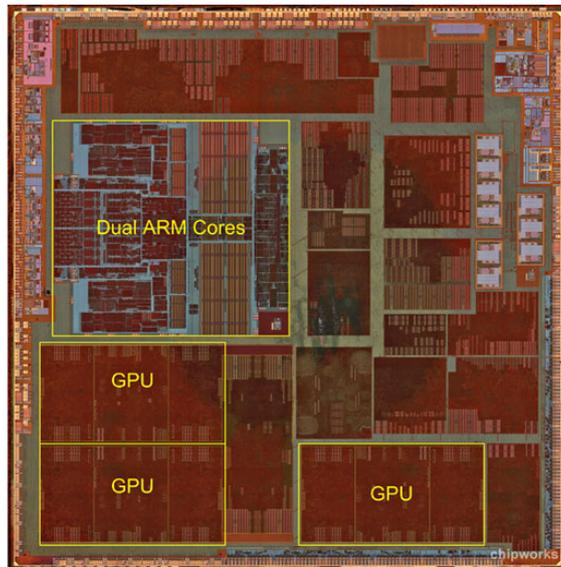


Figure 2: An Apple SoC - Two ARM and 3 GPU cores. Made by arch rival Samsung.

- Case, Battery and PCBs
- Java VM and Operating System.

#### 0.0.4 Introduction: What is a SoC 1/2 ?

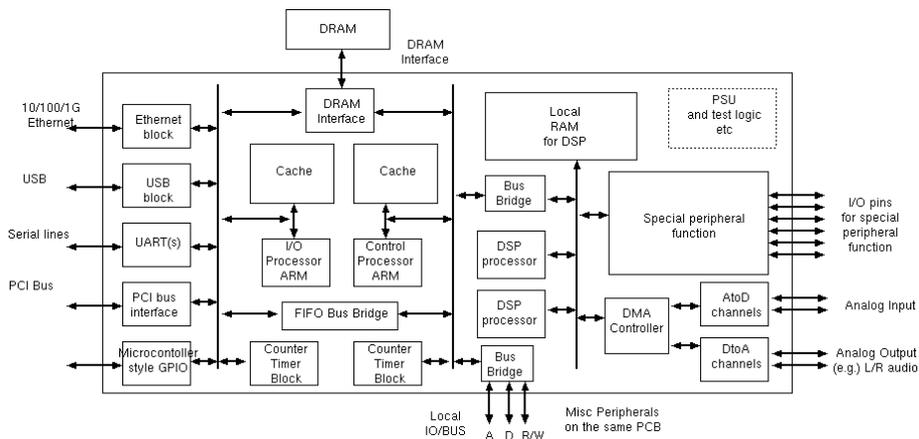


Figure 3: Block diagram of a multi-core 'platform' chip, used in a number of networking products.

A System On A Chip: typically uses 70 to 140 mm<sup>2</sup> of silicon.

Multicore field-programmable SoC Xilinx Product Brief: PDFAtmel ARM-Based Platform Chip: PDF **Advert ...**

*Embedded Systems Hardware for Software Engineers describes the electrical and electronic circuits that are used in embedded systems, their functions, and how they can be interfaced to other devices.*

*Basic computer architecture topics, memory, address decoding techniques, ROM, RAM, DRAM, DDR, cache memory, and memory hierarchy are discussed. The book covers key architectural features of widely used microcontrollers and microprocessors, including Microchip's PIC32, ATMEL's AVR32, and Freescale's MC68000. Interfacing to an embedded system is then described. Data acquisition system level design considerations and a design example are presented with real-world parameters and characteristics. Serial interfaces such as RS-232, RS-485, PC, and USB are addressed and printed circuit boards and high-speed signal propagation over transmission lines are*

covered with a minimum of math. A brief survey of logic families of integrated circuits and programmable logic devices is also contained in this in-depth resource.

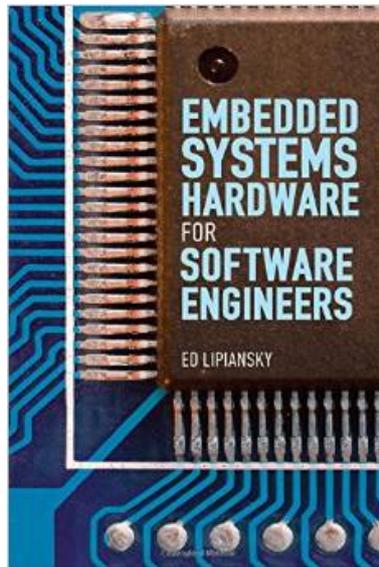


Figure 4: Embedded Systems Hardware for Software Engineers.

Ed Lipianski's Book

### 0.0.5 Introduction: What is a SoC 2/2 ?

A SoC is a complete system on a chip. A 'system' includes a microprocessor, memory and peripherals. The processor may be a custom or standard microprocessor, or it could be a specialised media processor for sound, modem or video applications. There may be multiple processors and also other generators of bus cycles, such as DMA controllers. DMA controllers can be arbitrarily complex and are only really distinguished from processors by their complete or partial lack of instruction fetching.

Processors are interconnected using a variety of mechanisms, including shared memories and message-passing hardware entities such as general on-chip networks and specialised channels and mailboxes.

SoCs are found in every consumer product, from modems, mobile phones, DVD players, televisions and iPods.

## 0.1 Hardware Design Flow for an ASIC

ASIC: Application-Specific Integrated Circuit. The ASIC hardware design flow is divided by the **Structural RTL** level into:

- **Front End:** specify, explore, design, capture, synthesise  $\rightsquigarrow$  **Structural RTL**
- **Back End:** **Structural RTL**  $\rightsquigarrow$  place, route, mask making, fabrication.

There is a companion software design flow that must mesh perfectly with the hardware if the final product is to work first time. This course will put as much emphasis on field-programmable parts (FPGAs) as on ASICs, since FPGA has now grown in importance.

Figure 6 shows a typical design and manufacturing flow that leads from design capture to ASIC fabrication. This might take six months. The FPGA equivalent can be done in half a day!

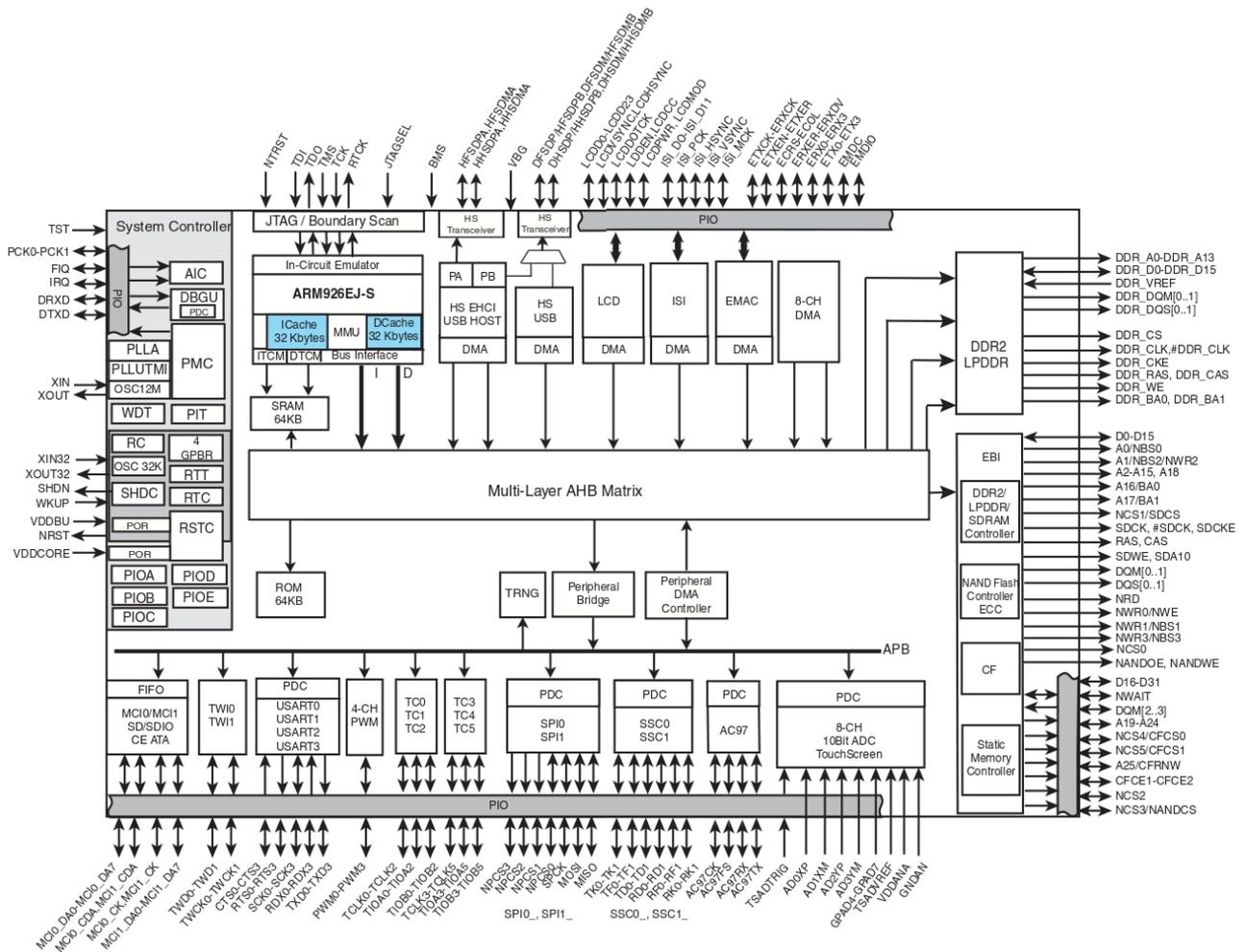


Figure 5: Platform Chip Example: Atmel SAM Series 9645.

### 0.1.1 Front End

The design must be specified in terms of high-level requirements, such as function, throughput and power consumption.

Design capture: it is transferred from the marketing person’s mind, back of envelope or wordprocessor document into machine-readable form.

Architectural exploration will try different combinations of processors, memories and bus structures to find an implementation with good power and load balancing. A loosely-timed high-level model is sufficient to compute the performance of an architecture.

Detailed design will select IP (interlectual property) providers for all of the functional blocks, or else they will exist from previous in-house designs and can be used without license fees, or else freshly written.

Logic synthesis will convert from behavioural RTL to structural RTL. Synthesis from formal high-level forms, including C,C++, SysML statecharts, formal specifications of interfaces and behaviour is beginning to be used.

Instruction set simulators (ISS) for embedded processors are needed: purchased from third parties such as ARM and MIPS, or as a by-product of custom processor design.

The interface specifications (register maps and other APIs) between components need to be stored: the IP-XACT format may be used.

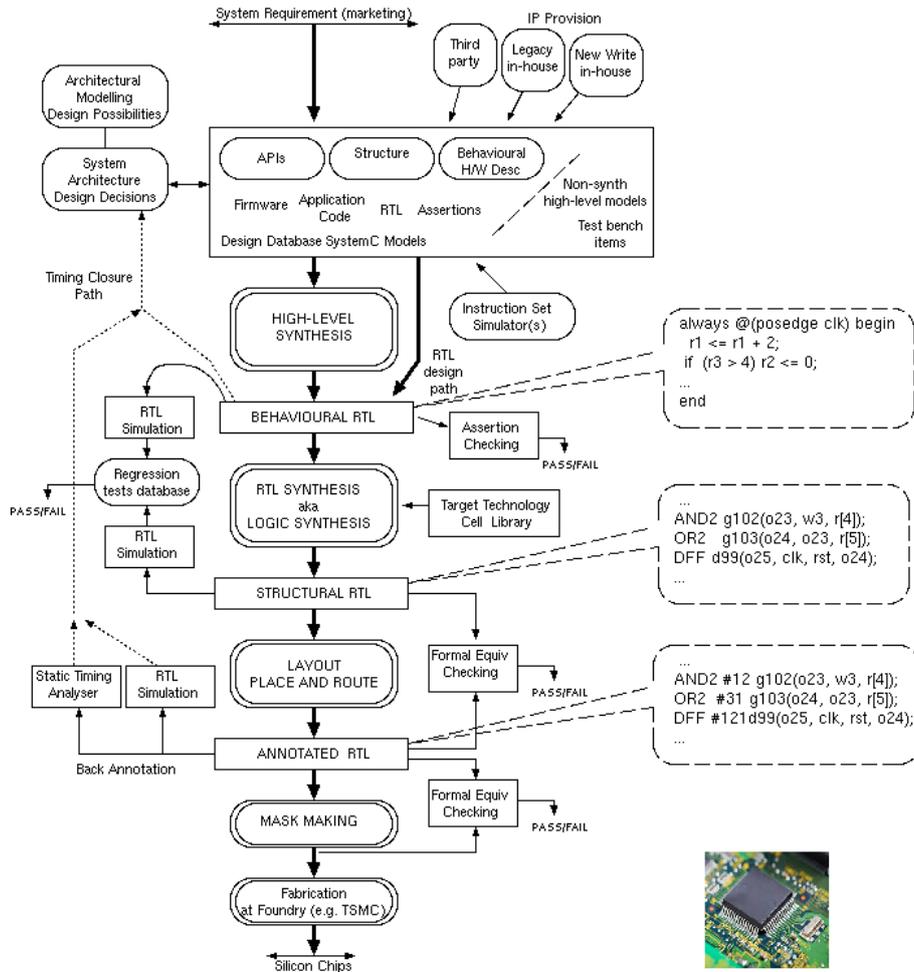


Figure 6: Design and Manufacturing Flow for SoC.

High-level models that are never intended to be synthesisable and test bench components will also be coded, typically using SystemC.

### 0.1.2 Back End

After RTL synthesis using a target technology library, we have a structural netlist that has no gate delays. Place and route gives 2-D co-ordinates to each component, adds external I/O pads and puts wiring between the components. RTL annotated with actual implementation gate delays gives a precise power and performance model. If performance is not up to par, design changes are needed.

Fabrication of masks is commonly the most expensive single step (e.g. one million pounds), so must be correct first time.

Fabrication is performed in-house by certain large companies (e.g. Intel, Samsung) but most companies use foundries (UMC, TSMC).

At all stages (front and back end), a library of standard tests will be run every night and any changes that cause a previously-passing test to fail (regressions) will be automatically reported to the project manager.

### 0.1.3 Levels of Modelling Abstraction

Our modelling system must support all stages of the design process, from design entry to fabrication. But we cannot model a complete SoC in detail and expect to simulate the booting of the O/S in reasonable time. We typically use a model called an **ESL Virtual Platform**. And where we are interested in the details of a specific module, we need to mix components using different levels of modelling abstraction within a single virtual platform.

Levels commonly used are:

- **Functional Modelling:** The ‘output’ from a simulation run is accurate.
- **Memory Accurate Modelling:** The contents and layout of memory is accurate.
- **Untimed TLM:** No time stamps recorded on transactions.
- **Loosely-timed TLM:** The number of transactions is accurate, but order may be wrong.
- **Approximately-timed TLM:** The number and order of transactions is accurate.
- **Cycle-Accurate Level Modelling:** The number of clock cycles consumed is accurate.
- **Event-Level Modelling:** The ordering of net changes within a clock cycle is accurate.

Other terms in use are:

- **Programmer View Accurate:** The contents of visible memory and registers is as per the real hardware, but timing may be inaccurate and other registers or combinational nets that are not designated as part of the ‘programmers view’ may not be modelled accurately.
- **Behavioural Modelling:** Using a threads package, or other library (e.g. SystemC), hand-crafted programs are written to model the behaviour of each component or subsystem. Major hardware items such as busses, caches or DRAM controllers may be neglected in such a model.

The Programmer’s View is often abbreviated as ‘PV’ and if timing is added it is called ‘PV+T’.

The Programmer’s View contains only architecturally-significant registers such as those that the software programmer can manipulate with instructions. Other registers in a particular hardware implementation, such as pipeline stages and holding registers to overcome structural hazards, are not part of the PV.

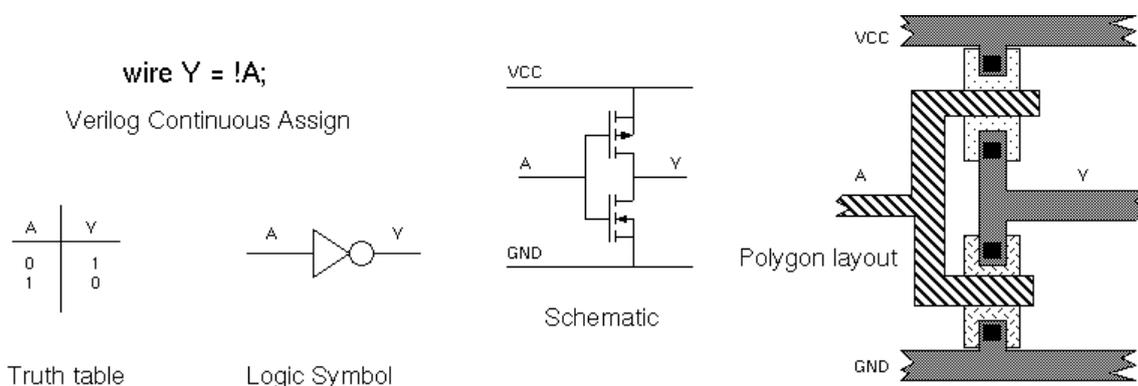


Figure 7: An inverter viewed at various levels of abstraction.

# KG 1 — Basic SoC Components

This section is a tour of actual hardware components (IP blocks) found on chips, presented with schematics and illustrative RTL fragments, and connected using a simple bus. Later we will look at other busses and networks on chip.

In the old-fashioned approach, we notice that the hand-crafted RTL used for the hardware implementation has no computerised connection with the firmware, device drivers or non-synthesisable models used for architectural exploration. Today, XML representations of IP-block meta-info resolve this (IP-XACT and OVM/UVM will be mentioned in the last lecture if time permits).

## 1.1 Simple Microprocessor: Bus Connection and Internals

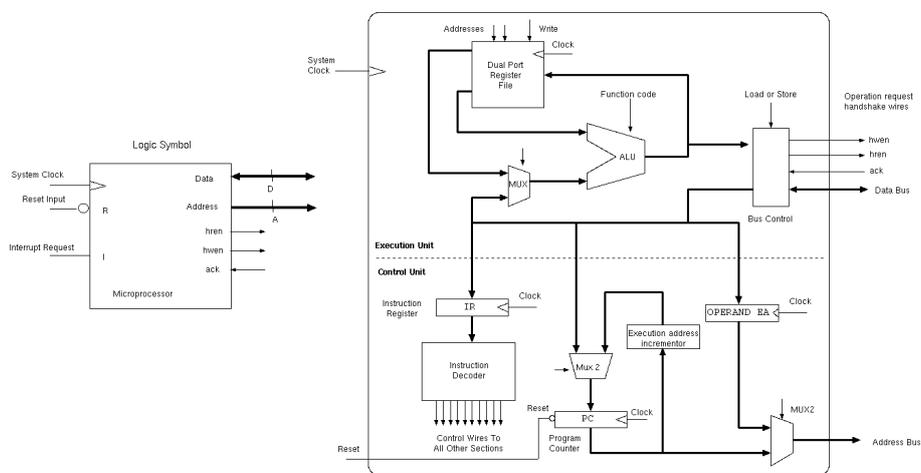


Figure 1.1: Schematic symbol and internal structure for a microprocessor (CPU).

This device is a bus master or *initiator* of bus transactions. It makes a load/read by asserting host read enable: **hren**. It writes to address space (a store) by asserting host write enable **hwen**. In this course we are concerned with the external connections only.

A central processor unit (CPU) is an execution unit and a control unit. A microprocessor (MPU) is a processor (CPU) on a chip. Early microprocessors such as the original Intel 8080 device had a 16 bit address bus and an 8 bit data bus so can address 64 Kbytes of memory. We say it had an A16/D8 memory architecture. Modern MPUs commonly have on-chip caches and an MMU for virtual memory.

It executes a handshake with external devices using the hren/hwen signals as requests and the ack signal as an acknowledge. In the following slides every device can respond immediately and so no ack signal is shown. In practice, contention, cache misses and operations on slow busses will cause wait states for the processor. Simple processors stall entirely during this period, whereas advanced cores carry on with other work and can receive responses out of order.

The interrupt input makes it save the current PC and load an agreed value that is the entry point for an interrupt service routine.

The high-order address bits are decoded to create chip enable signals for each of the connected peripherals, such as the RAM, ROM and UART.

As we shall see, perhaps the first SoCs, as such, were perhaps the microcontrollers. The Intel 8051 used in the mouse shipped with the first IBM PC is a good example. For the first time, RAM, ROM, Processor and I/O

devices are all on one piece of silicon. We all now have many of these such devices : one in every card in our wallet or purse. Today's SoC are the same, just much more complex.

### 1.1.1 A canonical D8/A16 Micro-Computer

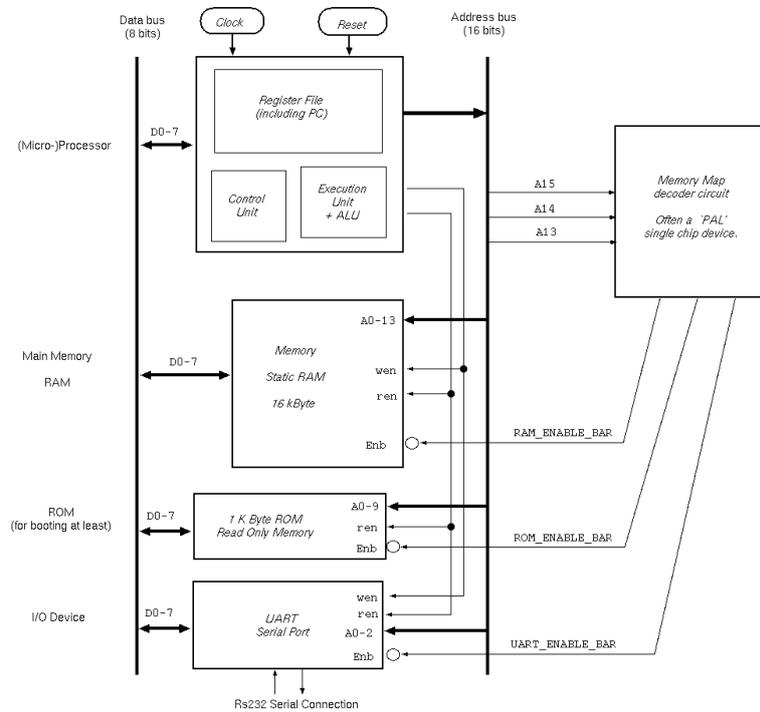
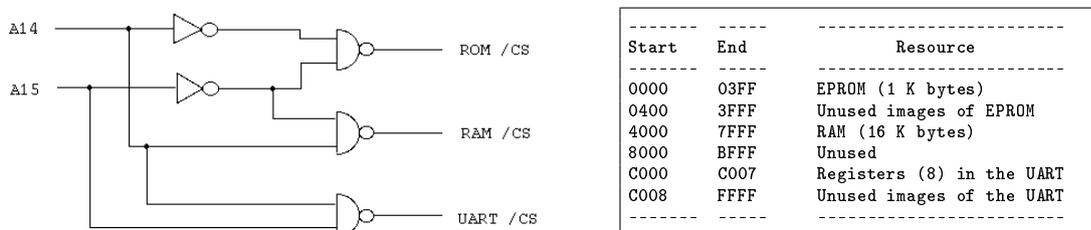


Figure 1.2: Early microcomputer structure, using a bidirectional/tri-state data bus.

Figure 1.36 shows the inter-chip wiring of a basic microcomputer (i.e. a computer based on a microprocessor).



The following RTL describes the required glue logic for the **memory map**:

```

module address_decode(abus, rom_cs, ram_cs, uart_cs);
input [15:14] abus;
output rom_cs, ram_cs, uart_cs;
assign rom_cs = !(abus == 2'b00); // 0x0000
assign ram_cs = !(abus == 2'b01); // 0x4000
assign uart_cs = !(abus == 2'b11); // 0xC000
endmodule
    
```

For a thorough example, cat /proc/iomem on your linux machine.

The 64K memory map of the processor has been allocated to the three addressable resources as shown in the table. The memory map must be allocated without overlapping the resources. The ROM needs to be at address zero if this is the place the processor starts executing from when it is reset. The memory map must be known

at the time the code for the ROM is compiled. This requires agreement between the hardware and software engineers concerned.

In the early days, the memory map was written on a blackboard where both teams could see it. For a modern SoC, there could be hundreds of items in the memory map. An XML representation called IP-XACT is being adopted by the industry and the glue logic may be generated automatically.

### 1.1.2 ROM - Read Only Memory

The switches cannot be switched! ROM is either mask programmed at manufacture or field-programmable.

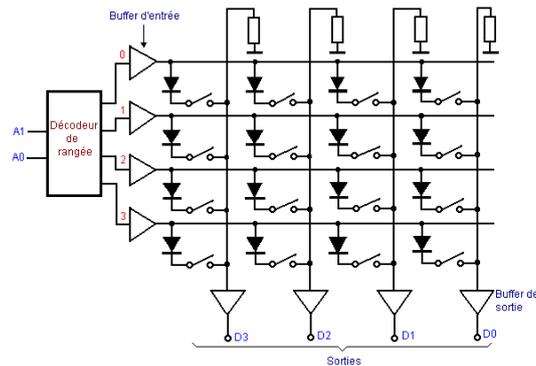


Figure 1.3: ROM Structure - the 'switches' have various implementation technologies.

This is a tiny ROM: four words of four bits.

An addressed row causes the column wires to become one or zero according to whether the diodes are installed (or connected) at the crosspoints. FLASH is a common type of ROM used in USB-sticks and SD cards. The 'switches' in FLASH are transistors with floating gates that are charged and discharged using electron tunnelling when ten or more volts are applied, but which retain their static charge for many years under normal conditions. By 'floating' we mean totally insulated from the rest of the electronic circuit (an envelope of silicon dioxide surrounds each floating gate).

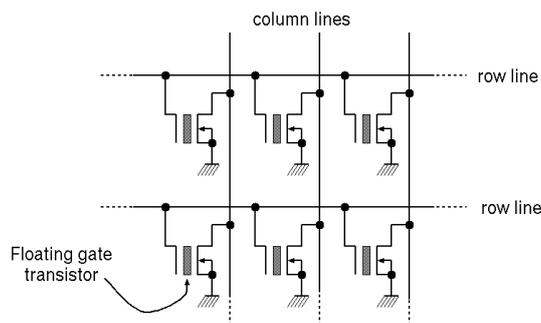


Figure 1.4: FLASH ROM structure (not examinable).

*Flash circuit-level details not examinable for part II CST*

### 1.1.3 A Basic Micro-Controller

A microcontroller has all of the system parts on one piece of silicon. First introduced in 1979-85 (e.g. Intel 80C51). Such a microcontroller has an internal D8/A16 architecture and is used in things like a door lock, mouse or smartcard. Many smartcards today contain a microcontroller that runs a cut-down, integer only, Java VM.



Figure 1.5: Intel 80C51 Microcontroller Chip from 1980.

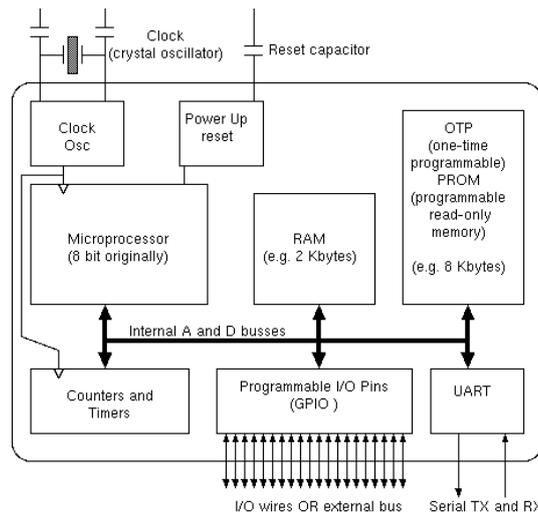


Figure 1.6: A typical single-chip microcomputer (microcontroller).

### 1.1.4 Switch/LED Interfacing

Figure 1.8 shows an example wiring structure for hardwired functionality with switches and LEDs. Figure 1.9 shows an example of memory address decode and simple LED and switch interfacing for programmed I/O (PIO) using a microprocessor. When the processor generates a read of the appropriate address, the tri-state buffer places the data from the switches on the data bus. When the processor writes to the appropriate address, the broadside latch captures the data for display on the LEDs until the next write.

### 1.1.5 UART Device

The RS-232 serial port was widely used in the 20th century for character I/O devices (teletype, printer, dumb terminal). A pair of simplex channels (output and input) make it full duplex. Additional wires are sometimes used for hardware flow control, or a software Xon/Xoff protocol can be used. Baud rate and number of bits per word must be pre-agreed.



Figure 1.7: Contact plate for a smartcard - Reader supplies VCC power, clock and reset. I/O is via the one-bit, bidirectional data pin .

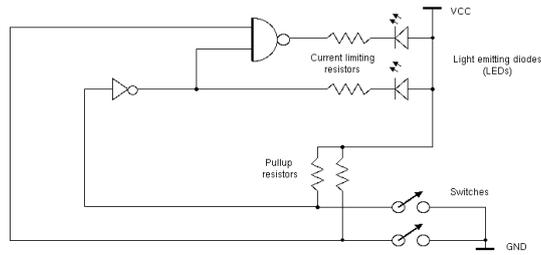


Figure 1.8: Connecting LEDs and switches to digital logic.

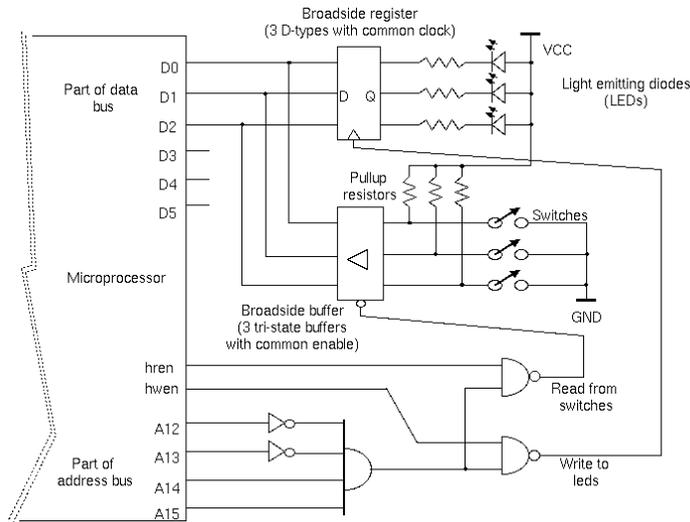


Figure 1.9: Connecting LEDs and switches for CPU programmed I/O (PIO)

The request signal is called strobe. The other signals on the connector are not important.

We'll cover this under the protocol and interface part of the course.

### 1.1.6 Programmed I/O

Programmed Input and Output (PIO). Input and output operations are made by a program running on the processor. The program makes read or write operations to address the device as though it was memory. Disadvantage: Inefficient - too much polling for general use. Interrupt driven I/O is more efficient.

Here is C preprocessor code to define the I/O locations in use by a simple UART device (universal asynchronous receiver/transmitter).



Figure 1.10: Typical Serial and Parallel Ports of 20th Century

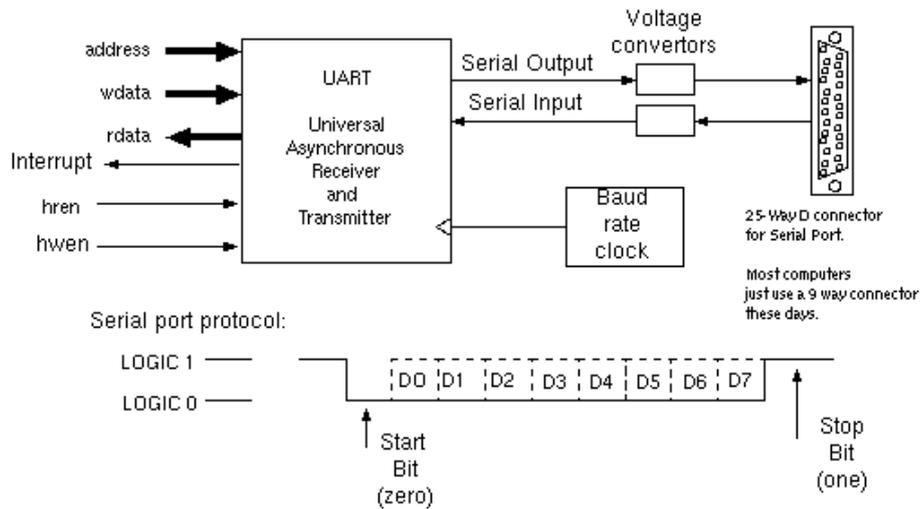


Figure 1.11: Typical Configuration of a Serial Port with UART



Figure 1.12: Serial Port Connector (9 pin instead of original 25 pin).

```
//Macro definitions for C preprocessor
//Enable a C program to access a hardware
//UART using PIO or interrupts.

#define IO_BASE 0xFFFC1000 // or whatever

#define U_SEND 0x10
#define U_RECEIVE 0x14
#define U_CONTROL 0x18
#define U_STATUS 0x1C

#define UART_SEND() \
    (*(volatile char *) (IO_BASE+U_SEND))
#define UART_RECEIVE() \
    (*(volatile char *) (IO_BASE+U_RECEIVE))
#define UART_CONTROL() \
    (*(volatile char *) (IO_BASE+U_CONTROL))
#define UART_STATUS() \
    (*(volatile char *) (IO_BASE+U_STATUS))

#define UART_STATUS_RX_EMPTY (0x80)
#define UART_STATUS_TX_EMPTY (0x40)

#define UART_CONTROL_RX_INT_ENABLE (0x20)
#define UART_CONTROL_TX_INT_ENABLE (0x10)
```

The receiver spins until the empty flag in the status register goes away. Reading the data register makes the status register go empty again. The actual hardware device might have a receive FIFO, so instead of going empty, the next character from the FIFO would become available straightaway:

```
char uart_polled_read()
{
    while (UART_STATUS() &
           UART_STATUS_RX_EMPTY) continue;
    return UART_RECEIVE();
}
```

The output function is exactly the same in principle, except it spins while the device is still busy with any data written previously:

```
uart_polled_write(char d)
{
    while (!(UART_STATUS() &
            UART_STATUS_TX_EMPTY)) continue;
    UART_SEND() = d;
}
```



Figure 1.13: Teletype/teleprinter: These devices were almost entirely mechanical with the electric circuit being little more than one switch activated by cams for transmit and one solenoid for receive.

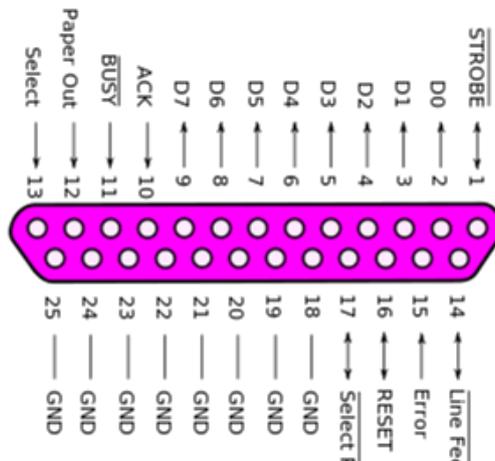


Figure 1.14: (Centronix) Parallel Port Pin Connections

Interrupt driven UART device driver:

```

char rx_buffer[256];
volatile int rx_inptr, rx_outptr;

void uart_reset()
{ rx_inptr = 0; tx_inptr = 0;
  rx_output = 0; tx_output = 0;
  UART_CONTROL() |= UART_CONTROL_RX_INT_ENABLE;
}
// Here we call wait() instead of 'continue'
// in case the scheduler has something else to run.
char uart_read() // called by application
{ while (rx_inptr==rx_outptr) wait(); // Spin
  char r = buffer[rx_outptr];
  rx_outptr = (rx_outptr + 1)&255;
  return r;
}

char uart_rx_isr() // interrupt service routine
{ while (1)
  {
    if (UART_STATUS()&UART_STATUS_RX_EMPTY) return;
    rx_buffer[rx_inptr] = UART_RECEIVE();
    rx_inptr = (rx_inptr + 1)&255;
  }
}

uart_write(char c) // called by application
{ while (tx_inptr==tx_outptr) wait(); // Block if full
  buffer[tx_inptr] = c;
  tx_inptr = (tx_inptr + 1)&255;
  UART_CONTROL() |= UART_CONTROL_TX_INT_ENABLE;
}

char uart_tx_isr() // interrupt service routine
{ while (1)
  {
    if (!(UART_STATUS()&UART_STATUS_TX_EMPTY)) return;
    UART_SEND() = tx_buffer[tx_outptr];
    tx_outptr = (tx_outptr + 1)&255;
  }
  UART_CONTROL() &= 255-UART_CONTROL_TX_INT_ENABLE;
}
    
```

This second code fragment illustrates the complete set of five software routines needed to manage a pair of circular buffers for input and output to the UART using interrupts. If the UART has a single interrupt output for both send and receive events, then two of the five routines are combined with a software dispatch between their bodies. Not shown is that the ISR must be prefixed and postfixed with code that saves and restores the processor state (this is normally written in assembler).

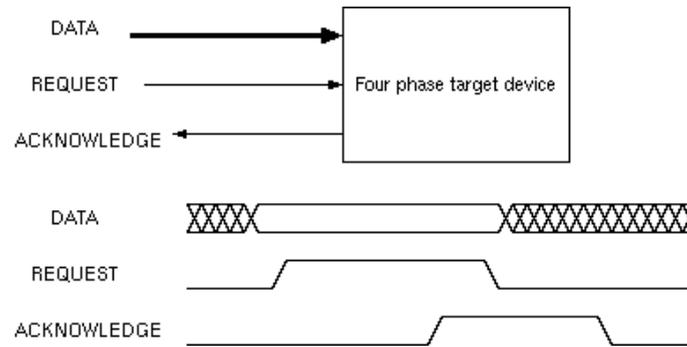


Figure 1.15: Timing diagram for an asynchronous, four-phase handshake.

## 1.2 A SoC consists of interconnected IP Blocks

In this section, we tour a number of IP (intellectual property) blocks. All will be **targets**, most will also generate **interrupts** and some will also be **initiators**. Interrupts will be conveyed by custom point-to-point nets. Data will use a 'bus'. A bus conveys a **transaction** from an initiating IP block on a target block and a response back again. A transaction contains a command that typically selects one of the these forms:

- **Single Word Write:** as generated by a CPU store instruction,
- **Single Word Read:** ditto for CPU load,
- **Burst Reads or Writes:** as cache lines are loaded or evicted or for block moves (e.g. Ethernet datagram DMA),
- **Laned Writes:** byte stores mean that only certain parts of a word are dirty,
- **I/O Read or Write:** programmed input/output requires cache bypass,
- **Load-locked:** (aka load-linked) takes a lock on a memory location,
- **Store-conditional:** (aka store-exclusive) performs a write if lock not lost in the meantime,
- **Memory Fence:** (aka barrier) preserves RaW and WaW ordering,
- **Evict or Snoop:** for inter-cache consistency updates (e.g. ARM ACE protocol)
- **(Misc Others:** Presence probe, debug access, read-ahead warm up ...)

In complex SoCs, burst transactions are often conveyed as packets over a network-on-chip (NoC). Such a packet may be broken up into *flits*

### 1.2.1 Illustrative, very simple SoC Bus.

In another section of these notes, we shall look at real-world busses like AXI, but we'll start simple here. For capacitance reasons, and owing to the small area use of transistors compared with the area used by busses, we normally do not use bi-directional (tri-state) busses within our SoC: instead our 'bus' consists of dedicated nets and multiplexor trees. In this section we use the following RTL net names:

- **addr[31:0]:** Input. Selection of internal address - not all 32 bits will be used,
- **hwen:** Input. Asserted during a write from host to target,
- **hren:** Input. Asserted during a read from target to host,

- **wdata[31:0]**: Input. Data to a target when writing/storing,
- **rdata[31:0]**: Output. Data read from target is reading/loading,
- **interrupt**: Output. Asserted by target when wanting attention.

On an **initiator** the net directions will be reversed. For simplicity, in this section, we assume a synchronous bus with no acknowledgement signal, meaning that every addressed target must respond in one clock cycle with no exceptions. Hence a cycle acknowledge handshake signal is not needed. Also we assume only complete words are ever stored, so no byte lane qualifiers for bytes and halfwords are shown.

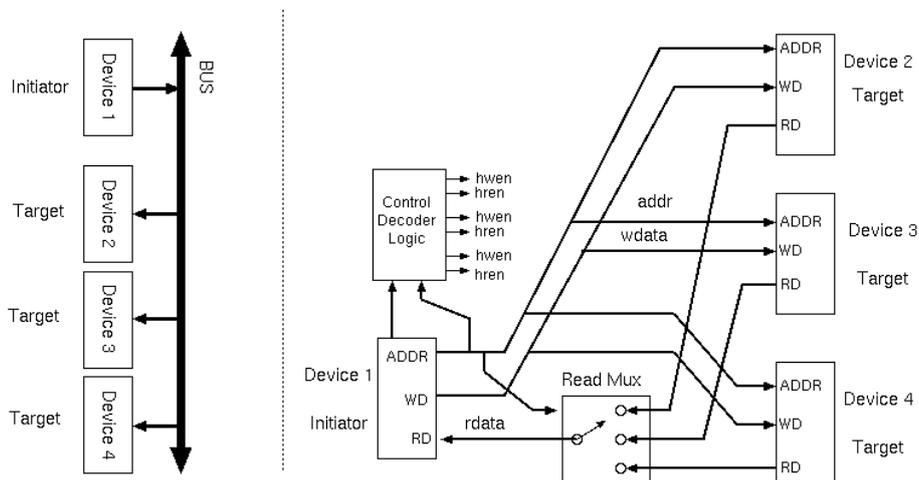


Figure 1.16: A basic SoC bus structure where one initiator addresses three targets (macroview and detailed wiring).

Figure 1.16 shows such a bus with one initiator and three targets. No tri-states are used: on a modern SoC address and write data outputs use wire joints or buffers; read data uses multiplexors. There is only one initiator, so no bus arbitration is needed.

Max throughput is unity (i.e. one word per clock tick). Typical SoC bus capacity:  $32 \text{ bits} \times 200 \text{ MHz} = 6.4 \text{ Gb/s}$ .

The most basic bus has one initiator and several targets. The initiator does not need to arbitrate for the bus since it has no competitors. Bus operations are just reads or writes of single 32-bit words. In reality, most on-chip busses support burst transactions, whereby multiple consecutive reads or writes can be performed as a single transaction with subsequent addresses being implied as offsets from the first address.

Interrupt signals are not shown in these figures. In a SoC they do not need to be part of the shared bus standard as such: they can just be dedicated wires running from device to device.

Un-buffered wiring can potentially serve for the write and address busses, whereas multiplexors are needed for read data. Buffering is needed in all directions for busses that go a long way over the chip.

### 1.2.2 Interrupt Wiring: General Structure for Uniprocessor

Nearly all devices have a master interrupt enable control flag that can be set and cleared by under programmed I/O by the controlling processor. Its output is just ANDed with the local interrupt source. We saw its use in the UART device driver, where transmit interrupts are turned off when there is nothing to send.

The programmed I/O uses the write enable (**hwen**) signal to guard the transfer of data from the main data bus into the control register. A **hren** signal is used for reading back stored value (shown on later slides).

The principal of programming is (see UART device driver):

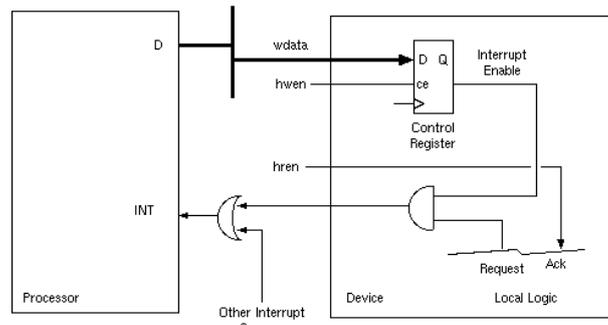
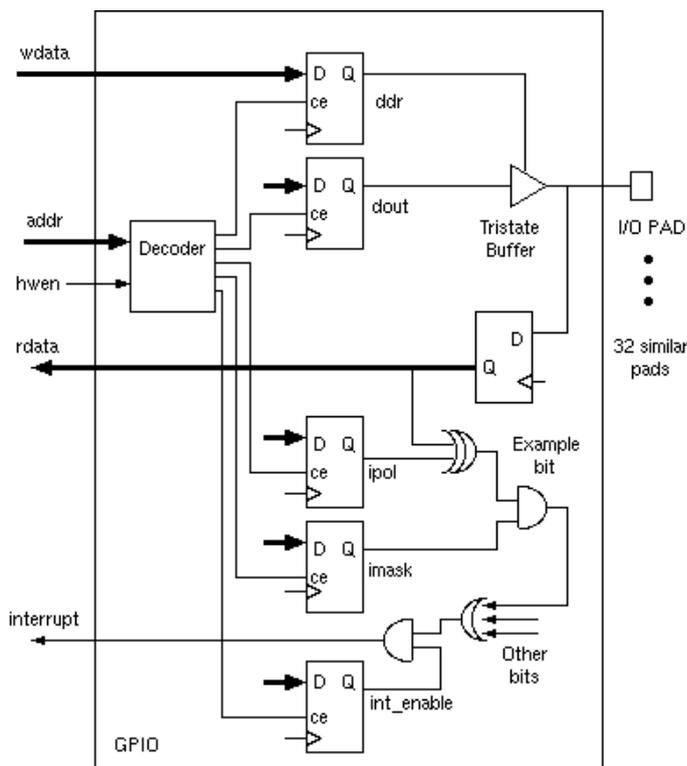


Figure 1.17: Interrupt generation: general structure within a device and at system level.

- Receiving device: Keep interrupt enabled: device interrupts when data ready.
- Transmit device: Enable interrupt when S/W output queue non-empty: device interrupts when H/W output queue has space.

With only a single interrupt wire to the processor, all interrupt sources share it and the processor must poll around on each interrupt to find the device that needs attention. Enhancement: a **vectored interrupt** makes the processor branch to a device-specific location. Interrupts can also be associated with priorities, so that interrupts of a higher level than currently being run preempt.

### 1.2.3 GPIO - General Purpose Input/Output Pins



RTL implementation of 32 GPIO pins:

```
// Programming model
reg [31:0] ddr; // Data direction reg
reg [31:0] dout; // output register
reg [31:0] imask; // interrupt mask
reg [31:0] ipol; // interrupt polarities
reg [31:0] pins_r; // register'd pin data

reg int_enable; // Master int enable (for all bits)

always @(posedge clk) begin
    pins_r <= pins;
    if (hwen && addr==0) ddr <= wdata;
    if (hwen && addr==4) dout <= wdata;
    if (hwen && addr==8) imask <= wdata;
    if (hwen && addr==12) ipol <= wdata;
    if (hwen && addr==16) int_enable <= wdata[0];
end

// Tri-state buffers.
bufif b0 (pins[0], dout[0], ddr[0]);
.. // thirty others here
bufif b31 (pins[31], dout[31], ddr[31]);

// Generally the programmer can read all the
// programming model registers but here not.
assign rdata = pins_r;

// Interrupt masking
wire int_pending = (!((pins_r ^ ipol)&imask));
assign interrupt = int_pending && int_enable;
```

Microcontrollers have a large number of GPIO pins (see earlier slide).

*Exercise:* Show how to wire up a push button and sketch out the code for a device driver that returns how many times it has so far been pressed. Sketch polled and interrupt driven code.

Some state registers inside an I/O block are part of the **programmer's model** in that they can be directly ad-

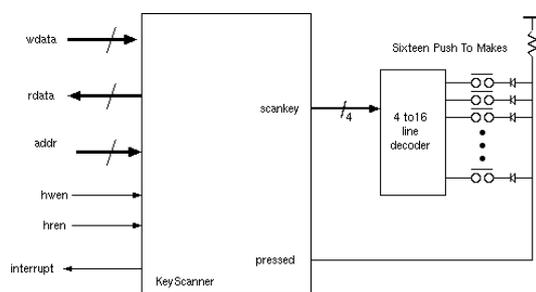
dressed with software (read and/or written), whereas other bits of state are for internal implementation purposes.

The general structure of GPIO pins has not changed since the 6821 controller chip designed in about 1972 that provided 20 such pins. A number of pins are provided that can either be input or output. A data direction register sets the direction on a per-pin basis. If an output, data comes from a data register. Interrupt polarity and masks are available on a per-pin basis for received events. A master interrupt enable mask is also provided.

The slide illustrates the schematic and the Verilog RTL for such a device. All of the registers are accessed by the host using programmed I/O.

### 1.2.4 Scan Multiplexing

Resistive switches shown (most keyboards and touch screens now use capacitive rather than resistive).



```

output [3:0] scankey;
input pressed;
reg int_enable, pending;
reg [3:0] scankey, pkey;

always @(posedge clk) begin
    if (!pressed) pkey <= scankey;
    else scankey <= scankey + 1;

    if (hwen) int_enable <= wdata[0]
    pressed1 <= pressed;
    if (!pressed1 && pressed) pending <= 1;
    if (hren) pending <= 0;
end
assign interrupt = pending && int_enable;
assign rdata = { 28'b0, pkey };

```

This simple keyboard scanner scans each key until it finds one pressed. It then loads the scan code into the pkey register where the host finds it when it does a programmed I/O read.

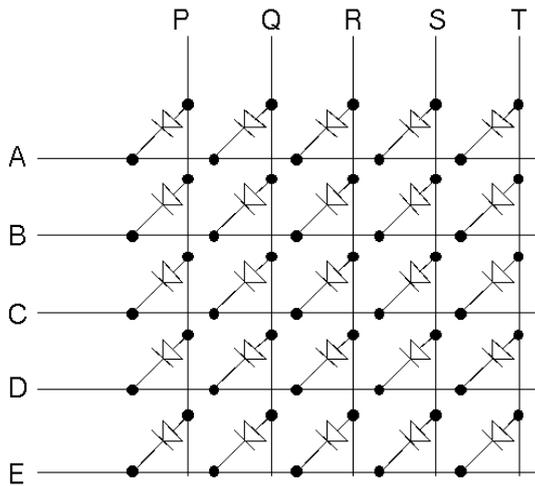
The host will know to do a read when it gets an interrupt. The interrupt occurs when a key is pressed and is cleared when the host does a read hren.

The details of this simple sketch are a little unrealistic. In practice, one would not scan at the speed of the processor clock. One would scan more slowly to stop the wires in the keyboard generating radio-frequency interference (RFI). Also, one should use extra register on asynchronous input pressed (see crossing clock domains) to avoid metastability. Finally, one would put the keys in a close to square grid, with as many 'pressed' column outputs form the array as row wires feeding the array.

And today, typically, one might use a dedicated microcontroller to scan the keyboard rather than design a hardware circuit.

Note, a standard PC keyboard generates an output byte on press and release and implements a short FIFO internally.

*Not lectured or examinable in 2015/16.*

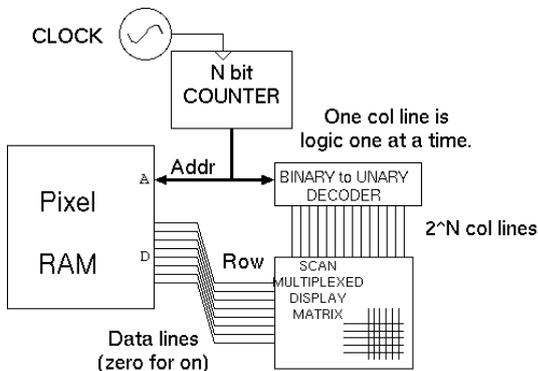


When a large number of leds or switches need to be connected, as in a display or keyboard, the number of connections can normally be reduced by connecting them in a matrix. Figure ?? shows how 25 LEDs can be connected to just ten signals.

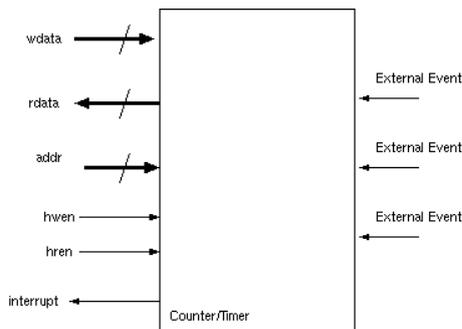
Because of persistence of vision (for displays) and slowness of fingers (for keyboards and touchpads) we do not need to drive/sample every element continuously.

The matrix gives a **scan-multiplexed** display or keyboard. In the display, one vertical column line can be driven to logic one at a time and a zero placed on the horizontal lines that should be illuminated in that column. A circuit to repeatedly read out the contents of a RAM and display it is shown in figure ?. Clearly, the desired LEDs are not all on at once, but by scanning the display faster than the human eye can detect flashes (about 50 Hz) and by using sufficiently large currents, so that the elements are brighter than would otherwise be required, this is overcome. The current is set by the value of a series current limiting resistor that is not shown.

For a scan-multiplexed keyboard, the switches take the place of the LEDs. Push-to-make, normally open switches must be used and the user should not press two at once. The scanning circuit must take one row line low in turn. Pull-up resistors keep the column lines at logic one unless a switch to a low row line is pressed.



### 1.2.5 Counter/Timer Block



```
// RTL for one channel of a typical timer

// Programmers' Model State
reg int_enable, int_pending;

reg [31:0] prescalar;
reg [31:0] reload;

// Programmer-invisible internal state
reg ovf'
reg [31:0] counter, prescale;

// Host write operations
always @(posedge clk) begin
    if (hwen && addr==0) int_enable <= wdata[0];
    if (hwen && addr==4) prescalar <= wdata;
    if (hwen && addr==8) reload <= wdata;
    // Write to addr==12 to clear interrupt
end
wire host_op = hwen && addr == 12;

// Host read operations
assign rdata =
    (addr==0) ? {int_pending, int_enable}:
    (addr==4) ? prescalar:
    (addr==8) ? reload: 0;

// A timer counts system clock cycles.
// A counter would count transitions from external input.
always @(posedge clk) begin
    ovf <= (prescale == prescalar);
    prescale <= (ovf) ? 0: prescale+1;
    if (ovf) counter <= counter -1;
    if (counter == 0) begin
        int_pending <= 1;
        counter <= reload;
    end
    if (host_op) int_pending <= 0;
end

// Interrupt generation
assign interrupt = int_pending && int_enable;
```

The counter/timer channel is essentially a counter that counts internal clock pulses or external events and which interrupts the processor on a certain count value. An automatic re-load register accommodates poor interrupt latency, so that the processor does not need to re-load the counter before the next event.

The counter/timer block will have perhaps 8 such channels. One is commonly dedicated as the system **watch-dog** that performs a hard reboot if not serviced within some time out by the CPU (e.g. 500 ms).

Timer (illustrated in the RTL) : counts pre-scaled system clock, but a counter has external inputs as shown on the schematic (e.g. car rev counter). Four to eight, versatile, configurable counter/timers generally provided in one block. All registers also configured as bus slave read/write resources for programmed I/O. In this example, the interrupt is cleared by host programmed I/O (during host\_op).

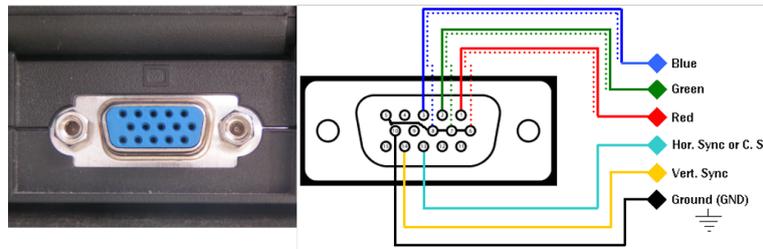
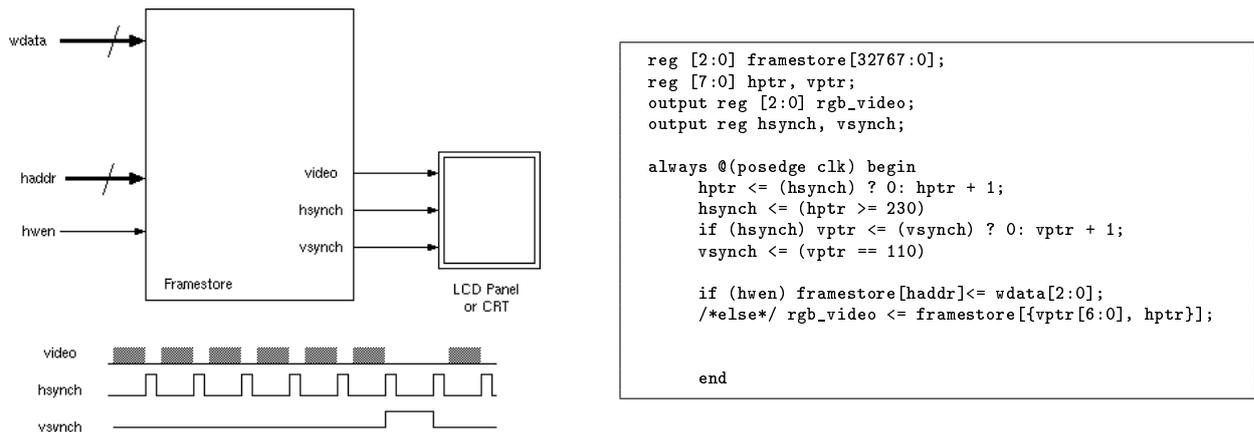


Figure 1.18: Super video graphics adaptor (SVGA) analog framestore output port.

### 1.2.6 Video Controller: Framestore



The framestore reads out the contents of its frame buffer again and again. The device driver needs to know the mapping of RAM addresses to screen pixels and has zeroed the locations read out during horizontal and vertical synchronisation.

No DAC is needed to render the basic saturated 8 colours, black, white, magenta, cyan, red etc.. To show other colours a DAC (digital to analog convertor) drives the R G and B signals. Modern DVI ports put the DACs at the display end of the monitor cable which is then digital.

The memory is implemented in a Verilog array and this has two address ports. Another approach is to have a single address port and for the RAM to be simply 'stolen' from the output device when the host makes a write to it. This will cause noticeable display artefacts if writes are at all frequent.

This framestore has fixed resolution and frame rate, but real ones have programmable values read from registers instead of the fixed numbers 230 and 110 (see the linux Modeline tool for example numbers). It is an output only device that never goes busy, so it generates no interrupts.

The framestore in this example has its own local RAM. This reduces RAM bandwidth costs on the main RAM but uses more silicon area. A delicate trade off! A typical compromise, also used on audio and other DSP I/O, is to have a small staging RAM or FIFO in the actual device but to keep as much as possible in the main memory.

Video adaptors in PC computers have their own local RAM or DRAM and also a local processor that performs polygon shading and so on (GPU).

### 1.2.7 Arbiter

*(Internal details not lectured).*

When multiple clients wish to share a resource, an arbiter is required. An arbiter decides which requester should be serviced. Arbiter circuits may be synchronous or asynchronous. Typical shared resources are busses,

memories and multipliers.

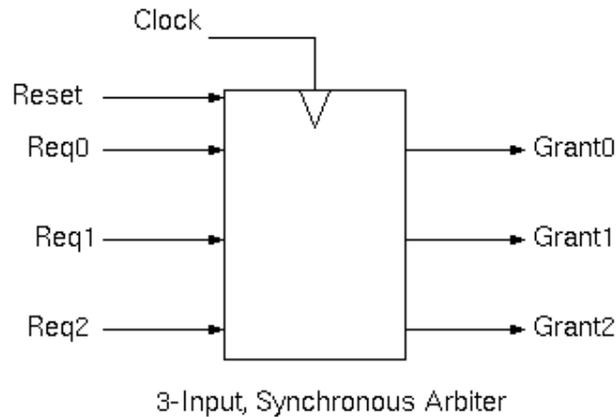


Figure 1.19: Typical Arbiter Schematic (three port/synchronous example)

There are two main arbitration disciplines:

- Static Priority - based on input port number (stateless).
- Round Robin - based on last user (held in internal state).

Another major policy variation is *preemptive* or not: can a granted resource be deassigned while the request is still asserted.

Complex disciplines involve dynamic priorities based on use history that avoid starvation or might implement 'best matchings' between a number of requesters and a number of resources.

```
//RTL implementation of synchronous, static priority arbiter with preemption.
module arbiter(input clk,
               input reset,
               input [2:0] reqs,
               output reg [2:0] grants);

  always @(posedge clk) if (reset) grants <= 0;
  else begin
    grants[0] <= reqs[0]; // Highest static priority
    grants[1] <= reqs[1] && !(reqs[0]);
    grants[2] <= reqs[2] && !(reqs[0] || reqs[1]);
  end
end
```

*Exercise:* Give the RTL code for a non-preemptive version of the 3-input arbiter.

*Exercise:* Give the RTL code for a round-robin, non-preemptive version of the 3-input arbiter.

### 1.2.8 Basic bus: Multiple Initiators.

The basic bus may have multiple initiators, so additional multiplexors select the currently active initiator. This needs arbitration between initiators: static priority, round robin, etc.. With multiple initiators, the bus may be busy when a new initiator wants to use it, so there are various arbitration policies that might be used. Preemptive and non-preemptive with static priority, round robin, and others mentioned above.

The maximum bus throughput of unity is now shared among initiators.

Since cycles now take a variable time to complete we need acknowledge signals for each request and each operation (not shown). How long to hold bus before re-arbitration? Commonly re-arbitrate after every burst.

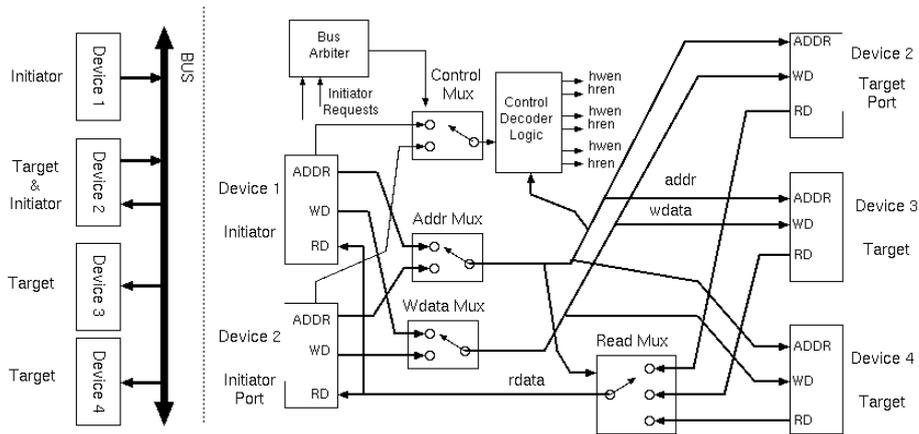
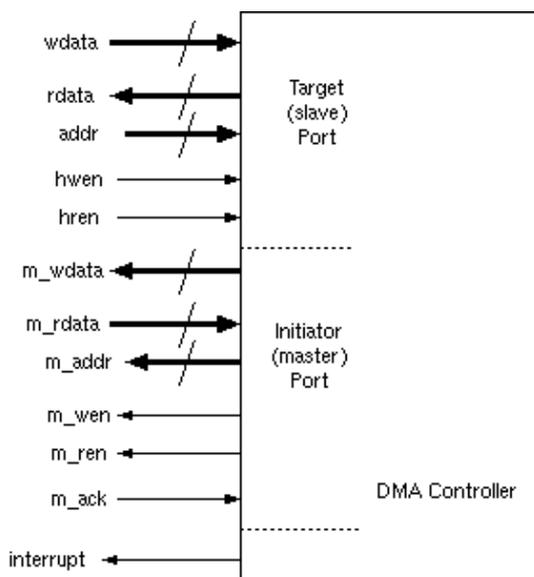


Figure 1.20: SoC bus structure where one of the targets is also an initiator (e.g. a DMA controller).

Practical busses support bursts of up to, say, 256 words, transferred to/from consecutive addresses. Our simple bus for this section does not support bursts. The latency in a non-preemptive system depends on how long the bus is held for. Maximum bus holding times affect response times for urgent and real-time requirements.

### 1.2.9 DMA Controller



This controller just block copies: may need to keep src and/or dest constant for device access. DMA controllers may be built into devices: SoC bus master ports needed.

```
// State for Programmers' Model
reg [31:0] count, src, dest;
reg int_enable, active;

// Other local state
reg [31:0] datareg;
reg intt, rwbar;

always @(posedge clk) begin // Target
  if (hwen && addr==0) begin
    { int_enable, active } <= wdata[1:0];
    intt <= 0; rwbar <= 1;
  end
  if (hwen && addr==4) count <= wdata;
  if (hwen && addr==8) src <= wdata;
  if (hwen && addr==12) dest <= wdata;
end
assign rdata = ...// Target readbacks

always @(posedge clk) begin // Initiator
  if (active && rwbar && m_ack) begin
    datareg <= m_rdata;
    rwbar <= 0;
    src <= src + 4;
  end
  if (active && !rwbar && m_ack) begin
    rwbar <= 1;
    dest <= dest + 4;
    count <= count - 1;
  end
  if (count==1 && active && !rwbar) begin
    active <= 0;
    intt <= 1;
  end
end
assign m_wdata = datareg;
assign m_ren = active && rwbar;
assign m_wen = active && !rwbar;
assign m_addr = (rwbar) ? src:dest;
assign interrupt = intt && int_enable;
```

The DMA controller is the first device we have seen that is a bus initiator as well as a bus target. It has two complete sets of bus connections. Note the direction reversal of all nets on the initiator port.

This controller just makes block copies from source to destination with the length being set in a third register. Finally, a status/control register controls interrupts and kicks of the procedure.

The RTL code for the controller is relatively straightforward, with much of it being dedicated to providing the target side programmed I/O access to each register.

The active RTL code that embodies the function of the DMA controller is contained in the two blocks qualified with the active net in their conjunct.

Typically, DMA controllers are multi-channel, being able to handle four or so concurrent or pending transfers. Many devices have their own DMA controllers built in, rather than relying on dedicated external controllers. However, this is not possible for devices connected the other side of bus bridges that do not allow mastering (initiating) in the reverse directions. An example of this is an IDE disk drive in a PC.

Rather than using a DMA controller one can just use another processor. If the processor runs out of (i.e. fetches its instructions from) a small, local instruction RAM or cache it will not impact on main memory bus bandwidth with code reads and it might not be much larger in terms of silicon area.

An enhancement might be to keep either of the src or destination registers constant for streaming device access. For instance, to play audio out of a sound card, the destination address could be set to the programmed I/O address of the output register for audio samples and set not to increment.

For streaming media with hard real-time characteristics, such as audio, video and modem devices, a small staging FIFO is likely to be needed in the device itself because the initiator port may experience latency when it is serviced. The DMA controller then initiates the next burst of its transfer when the local FIFO reaches a trigger depth.

### 1.2.10 Network and Streaming Media Devices

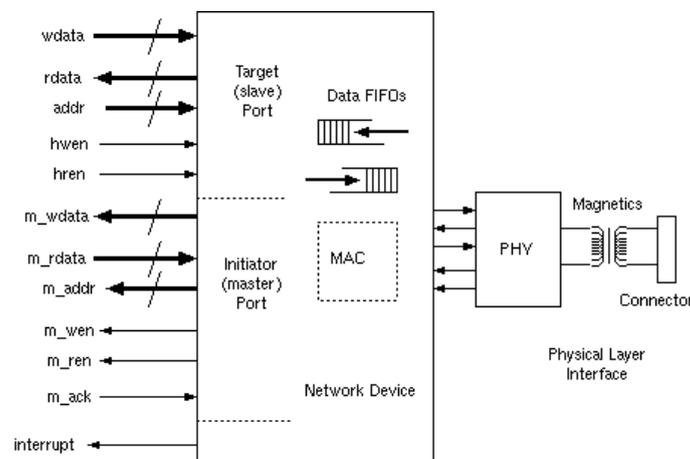


Figure 1.21: Connections to a DMA-capable network device.

Network devices, such as Ethernet, USB, Firewire, 802.11 are similar to streaming media devices, such as audio, and modem devices, and commonly have embedded DMA controllers. Only low throughput devices like the UART are likely not to use DMA.

DMA offloads work from the main processor, but, equally importantly, using DMA requires less staging RAM or data FIFO in device. In the majority of cases, RAM is the dominant cost in terms of SoC area.

Another advantage of a shared RAM pool is **statistical multiplexing gain**. It is well known in queueing theory that having a monolithic server performs better than having a number of smaller servers, with same total capacity, that each are dedicated to one client. If the clients all share one server and jobs arrive more or less at random, the system can be more efficient in terms of service delay and overall buffer space needed. The same

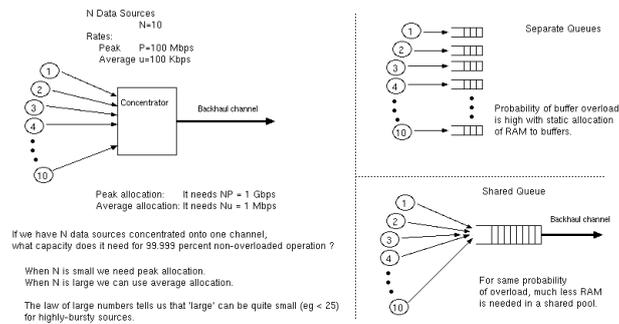


Figure 1.22: Extra diagram illustrating statistical multiplexing gain [Non-examinable].

effect applies to buffer allocation: having a central pool requires less overall RAM, to meet a statistical peak demand, than having the RAM split around the various devices.

The DMA controller in a network or streaming media device will often have the ability to follow elaborate data structures set up by the host CPU, linking and de-linking buffer pointers from a central pool.

### 1.2.11 Bus Bridge

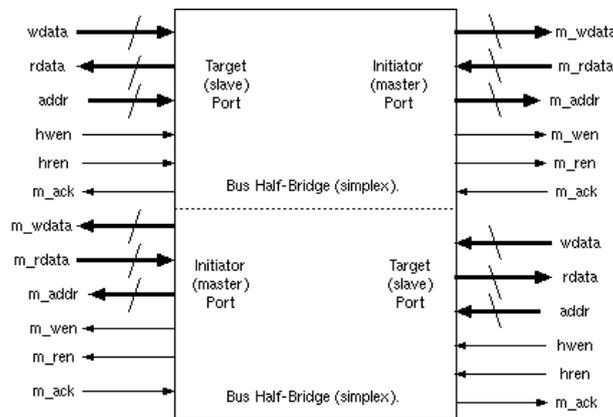


Figure 1.23: Bi-directional bus bridge, composed from a pair of back-to-back simplex bridges.

The essential behaviour of the bus bridge is that bus operations slaved on one side are mastered on the other. The bridge need not be symmetric: speeds and data widths may be different on each side.

A bus bridge connects together two busses that are potentially able to operate independently when traffic is not crossing. However, in some circumstances, especially when bridging down to a slower bus, there may be no initiator on the other side, so that side never actually operates independently and a unidirectional bridge is all that is needed.

The bridge need not support a unified or **flat address space**: addresses seen on one side may be totally reorganised when viewed on the other side or un-addressable. However, for debugging and test purposes, it is generally helpful to maintain a flat address space and to implement paths that are not likely to be used in normal operation.

A bus bridge might implement write posting using an internal FIFO. However it will generally block when reading. This is a motivation for having a network-on-chip.

As noted, the 'busses' on each side use multiplexors and not tri-states on a SoC. These multiplexors are different from bus bridges since they do not provide **spatial reuse** of bandwidth. Spatial reuse occurs when different busses are simultaneously active with different transactions. Multiple busses is a poor-man's network-on-chip.

With a bus bridge, system bandwidth ranges from 1.0 to 2.0 bus bandwidth: inverse proportion to bridge crossing cycles.

### 1.2.12 Inter-core Interrupter (Doorbell/Mailbox)

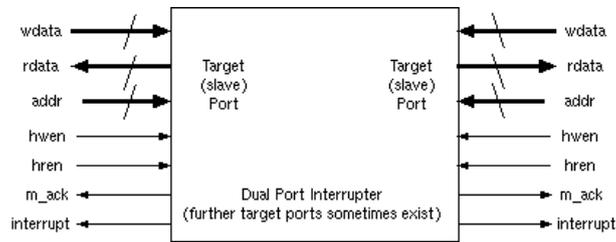


Figure 1.24: Dual-port interrupter (doorbell) or mailbox.

The inter-core interrupter (Doorbell/Mailbox) is a commonly-required component for basic synchronisation between separate cores. Used, for instance, where one CPU has placed a message in a shared memory region for another to read. Sometimes the interrupter is part of a central interrupt distributor, such as the 'GIC' from ARM, that enables any device interrupt to be routed to any core with any priority. Such a device offers multiple target interfaces, one per client bus. It generates interrupts to one core at the request of another.

Operational sequence: one core writes a register that asserts an interrupt wire to another core. The interrupted core, in its service routine, reads or writes a register in the interrupter to clear the interrupt.

Mailbox variant allows small data items to be written to a queue in the interrupter. These are read out by the (or any) core that is (or wants to) handle the interrupt. Link: Doorbell Driver Fragments.

### 1.2.13 Remote Debug (JTAG) Access Port

(Not lectured for part II).

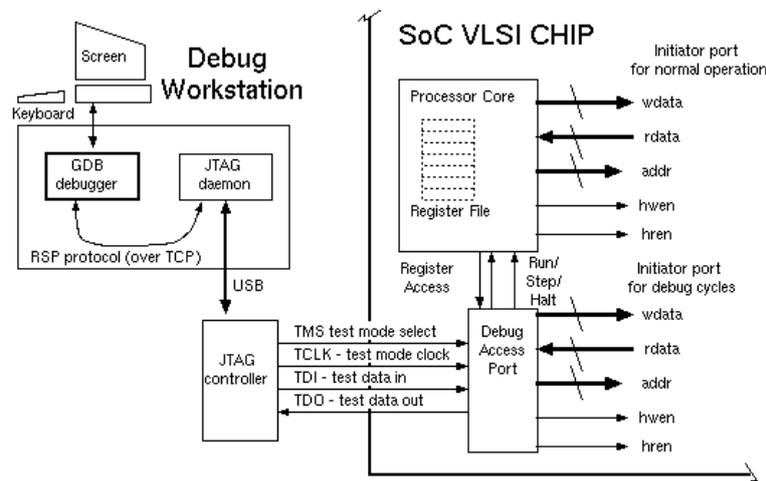


Figure 1.25: Remote Access Port connected to H/W SoC (can also connect to SystemC model).

There are various forms of debug access port, they can be connected to bus or connected to a CPU core or both. External access is often via the JTAG port which is fairly slow, owing to bit-serial data format, so sometimes parallel bus connections are provided. The basic facilities commonly provided are

- Perform a bus read or write cycles,

- Halt/continue/single-step the processor core,
- Read/modify processor core registers,
- Provide ‘watchpoints’ which halt on certain address bus values.

In a typical setup the debugger (such as GNU gdb) runs on a remote workstation via a TCP connection carrying the RSP protocol to the debug target. For real silicon, the target is a JTAG controller (e.g. connected to the workstation via USB) whereas on a SystemC model it is an SC\_MODULE that is listening for RSP on a unix socket.

### 1.2.14 Clock Source - Crystal Oscillator Example

(Not examinable for part II CST.)

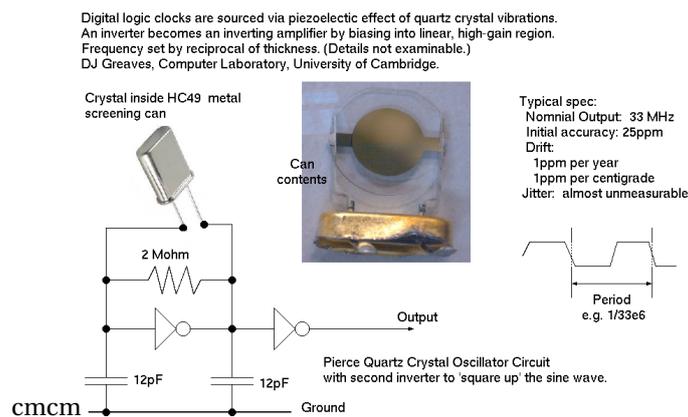


Figure 1.26: Clock multiplication using a PLL and distribution using a fractal H-tree.

Most electronic products use a soundwave inside a quartz crystal as a clock source. Ultra cheap products, like a musical greetings card, will use an R/C oscillator, but this will have typically only a 10 percent initial accuracy - (one semitone is 6 percent).

For above 20 MHz or so, the crystal cannot be cut thinly enough, so overtones or PLL multipliers are used.

### 1.2.15 Clock Frequency Multiplier PLL and Clock Tree

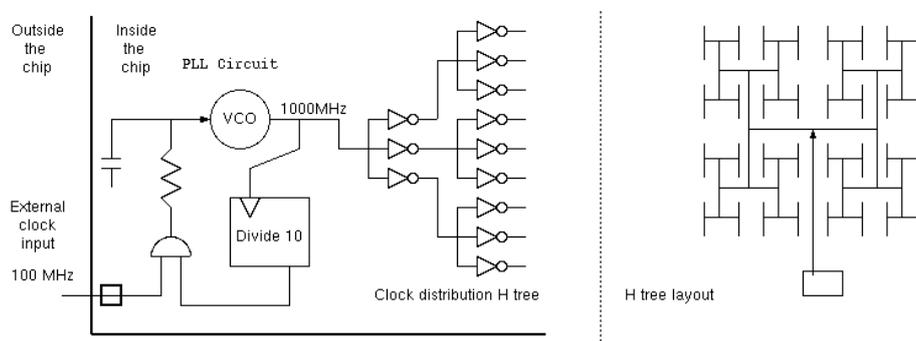


Figure 1.27: Clock multiplication using a PLL and distribution using a fractal H-tree.

- Clock sourced from a lower-frequency external (quartz) reference.
- Multiplied up internally with a phase-locked loop.

- Dynamic frequency scaling (future topic) implemented with a programmable division ratio.
- Skew in delivery is minimised using a balanced clock distribution tree.
- Physical layout: fractal of H's, ensuring equal wire lengths.
- Inverters are used to minimise pulse shrinkage (duty-cycle distortion).
- Clock may be gated at the leaves before sending to idle flip-flops to save power.

The clock tree delivers a clock to all flops in a domain with sufficiently low skew to avoid shoot-thru. This is achieved by balancing wire lengths between the drivers. The clock frequency is a multiple of the external reference which is commonly sourced from the piezo-effect of sound waves in a thin slice of quartz crystal. Later on, under power management, we will talk about having a programmable clock frequency, so it's worth noting that the multiplication factor of 10 illustrated in the slide can be variable and programmed in some systems (e.g. laptops).

### 1.2.16 Clock Domain Crossing Bridge

A clock-domain-crossing bridge is needed between clock domains. The basic techniques are the same whether implemented as part of an asynchronous FIFO, a SoC bus bridge or inside an IP block (e.g. network receive front end to network core logic). The same techniques apply when receiving asynchronous signals into a clock domain.

The following figure illustrates the key design aspects for crossing in one direction, but generally these details will be wrapped up into a carefully-design library block like the domain-crossing FIFO shown elsewhere.

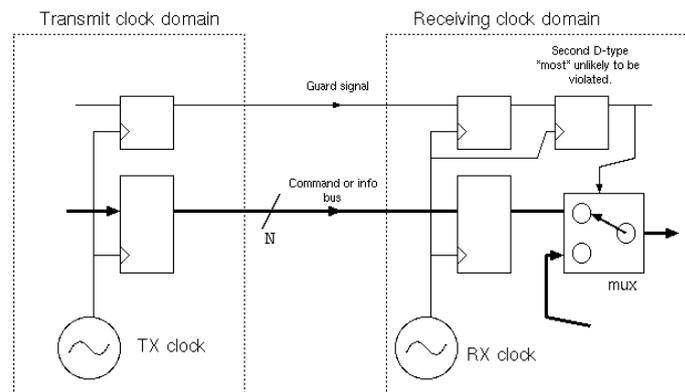


Figure 1.28: Generic setup when sending parallel data between clock domains.

#### Design principle:

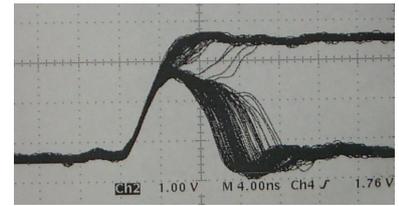
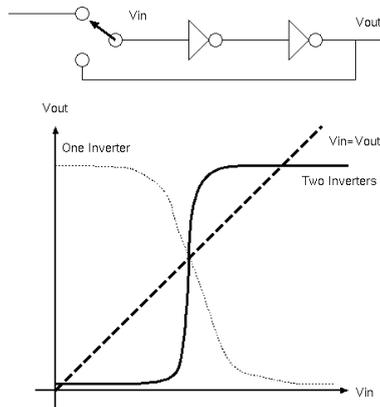
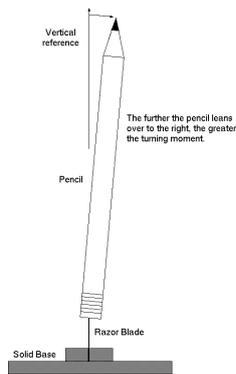
- Have a one-bit signal that is a guard or qualifier signal for all the others going in that direction.
- Make sure all the other signals are settled in advance of guard.
- Pass the guard signal through two registers before using it (metastability avoidance).
- Use a wide bus (crossing operations less frequent).

#### Receiver side RTL:

```
input clk; // receiving domain clock
input [31..0] data;
input req;
output reg ack;

reg [31:0] captured_data;
reg r1, r2;
always @(posedge clk) begin
    r1 <= req;
    r2 <= r1;
    ack <= r2;
    if (r2 && !ack) captured_data <= data;
```

#### Metastability Theory:



A pencil balancing on a razor blade can be metastable, but normally flops to one side or the other. A bistable is two inverters connected in a ring. This has two stable states, but there is also a metastable state. If a D-type is clocked while its input is changing, it might be set close to its metastable state and then drift to one level or the other. Sometimes, it will take a fair fraction of a clock period to settle. The oscillogram shows metastable waveforms at the output of a D-type when set/hold times are sometimes violated.

Two quartz crystal oscillators, each of 10 MHz frequency will actually be different by tens of Hz and drift with temperature. Atomic clocks are better: accuracy is one part in ten to the twelve or better, but infeasible to incorporate in everyday equipment and still not good enough to avoid rapid metastable failure.

A simplex clock domain crossing bridge carries information in only one direction. Duplex carries in both directions. Because the saturated symbol rates are not equal on each side, we need a protocol with insertable/deletable padding states or symbols that have no semantic meaning. Or, in higher-level terms, the protocol must have elidable idle states between transactions.

Clock domain crossing is needed when connecting to I/O devices that operate at independent speeds: for example, an Ethernet receiver sub-circuit works at the exact rate of the remote transmitter that is sending to it. Today's microprocessors also have separated clock domains for their cores viz their DRAM interfaces.

The data signals can also suffer from metastability, but the multiplexer ensures that these metastable values never propagate into the main logic of the receiving domain.

100 percent utilisation is impossible when crossing clock domains. The four-phase handshake limits utilisation to 50 percent (or 25 if registered at both sides) Other protocols can get arbitrarily close to saturating one side or the other provided we know the maximum tolerance in the nominal clock rates. Since clock frequencies are different, 100 percent of one side is either less than 100 percent of the other or else overloaded.

### 1.2.17 SoC Example - Raspberry Pi

The Raspberry Pi is a SoC with on-chip DRAM. A relatively new development in SoC implementation.

- Original Raspberry Pi Model A : 256 MByte
- Raspberry Pi Model B : 512 MByte
- Raspberry Pi 2 : 1GByte

Die stacking (putting two pieces of silicon in one package) enables short connections between main memory and logic. Growing technology maturity allows DRAM to be put on the same silicon die with the SoC logic. Memory density is still going up, even if CPU core speed has stalled.

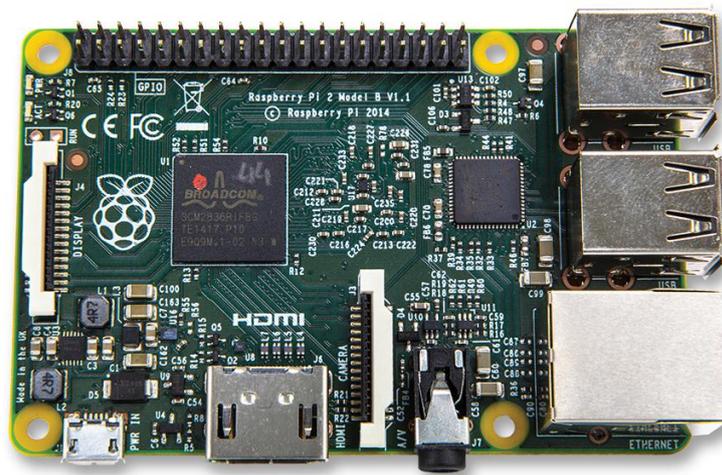


Figure 1.29: Raspberry Pi Photograph.

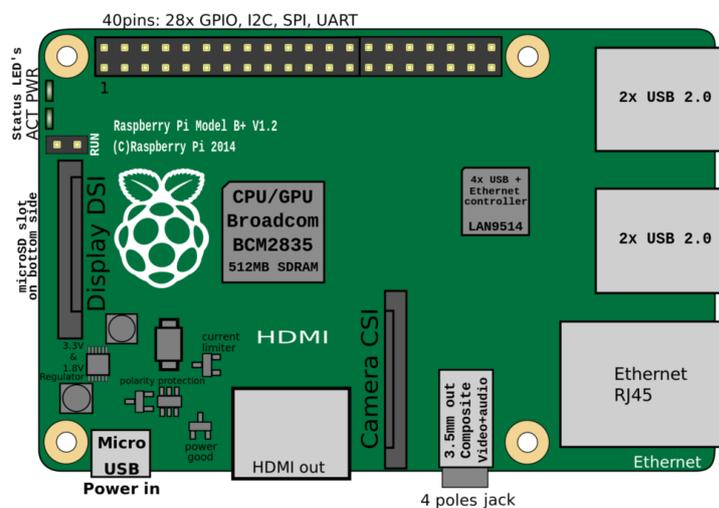


Figure 1.30: Raspberry Pi (Model B) PCB Layout.

### 1.2.18 Cache Design

Implementing 4-way, set-associative cache is relatively straightforward. One does not need an associative RAM macrocell: just synthesise four sets of XOR gates from RTL using the '=' operator!

```
reg [31:0] data0 [0:32767], data1 [0:32767], data2 [0:32767], data3 [0:32767];
reg [14:0] tag0 [0:32767], tag1 [0:32767], tag2 [0:32767], tag3 [0:32767];

always @(posedge clk) begin
    miss = 0;
    if (tag0[addr[16:2]]==addr[31:17]) dout <= data0[addr[16:2]];
    else if (tag1[addr[16:2]]==addr[31:17]) dout <= data1[addr[16:2]];
    else if (tag2[addr[16:2]]==addr[31:17]) dout <= data2[addr[16:2]];
    else if (tag3[addr[16:2]]==addr[31:17]) dout <= data3[addr[16:2]];
    else miss = 1;
end
```

Of course we also need a write and evict mechanism... (not shown). Rather than implement least-recently-used (LRU) one tends to do 'random' replacement which can be as simple as using keeping a two bit counter to say which 'way' to evict next. Typically an IP company like ARM will provide a high-quality, carefully-tuned implementation, ready to go.

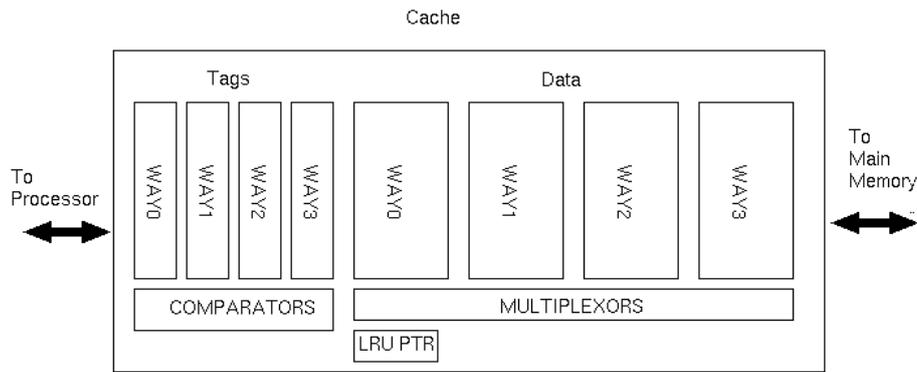


Figure 1.31: Memory blocks and tag comparator needed for a 4-way, set-associative cache.

### 1.2.19 Cache Modelling

Depending on our needs, we may want to measure the hit ratio in the I or D caches, or the effect on performance from the misses, or neither, or all such metrics. [Virtutec Simics.] or XPARCH Prazor. So a cache can be modelled at various levels of abstraction:

- Not at all - after all it does not affect functionality,
- Using an estimated hit ratio and randomly adding delay to main memory transactions accordingly,
- Fully modelling the tags and their lookup (while making backdoor access to the main memory for the data),
- Modelling the cache data RAMs as well (needed for detailed power modelling).

An instruction cache (I-cache), when modelled, may or may not be accessed by an emulator or instruction set simulator (ISS). For instance, the ISS may use backdoor access to the program in main memory, or it might use JIT (just-in-time) techniques where commonly executed inner loops of emulated code are converted to native machine code for the modelling workstation.

There are highly-configurable SystemC cache models in our memories folder: size, ways, write through/write back, consistent/non-consistent, directory based.

### 1.2.20 SoC Example: Helium 210

A platform chip is the modern equivalent of a microcontroller: it is a flexible chip that be programmed up to serve in a number of embedded applications. The set of components remains the same as for the microcontroller, but each has far more complexity: e.g. 32 bit processor instead of 8. In addition, rather than putting a microcontroller on a PCB as the heart of a system, the whole system is placed on the same piece of silicon as the platform components. This gives us a system on a chip (SoC).

The example illustrated in figure 1.33 has two ARM processors and two DSP processors. Each ARM has a local cache and both store their programs and data in the same off-chip DRAM.

The left-hand-side ARM is used as an I/O processor and so is connected to a variety of standard peripherals. In any typical application, many of the peripherals will be unused and so held in a power down mode.

The right-hand-side ARM is used as the system controller. It can access all of the chip's resources over various bus bridges. It can access off-chip devices, such as an LCD display or keyboard via a general purpose A/D local bus.

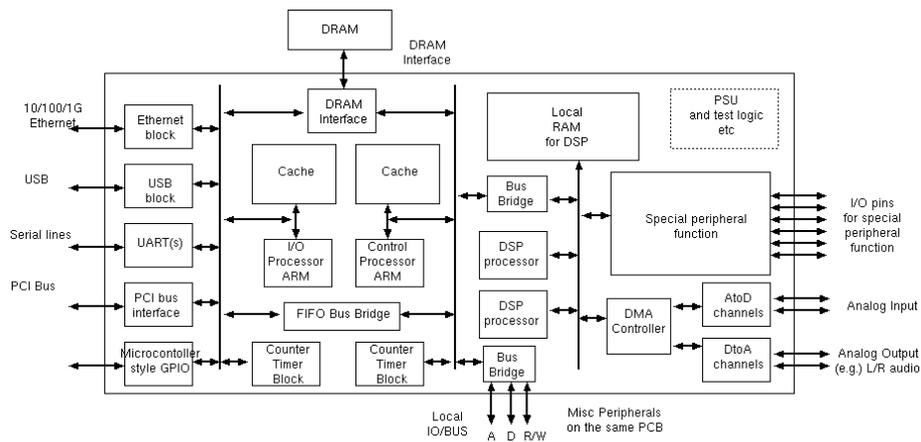


Figure 1.32: Platform Chip Example: Virata Helium 210



Figure 1.33: Helium chip as part of a home gateway ADSL modem (partially masked by 802.11 module).

The bus bridges map part of one processor's memory map into that of another so that cycles can be executed in the other's space, albeit with some delay and loss of performance. A FIFO bus bridge contains its own transaction queue of read or write operations awaiting completion.

The twin DSP devices run completely out of on-chip SRAM. Such SRAM may dominate the die area of the chip. If both are fetching instructions from the same part of the same RAM, then they had better be executing the same program in lock-step or else have some own local cache to avoid huge loss of performance in bus contention.

The rest of the system is normally swept up onto the same piece of silicon and this is denoted with the 'special function peripheral.' This would be the one part of the design that varies from product to product. The same core set of components would be used for all sorts of different products, from iPods, digital cameras or ADSL modems.

### 1.2.21 SoC Example: Atmel SAM9645

The SAM9645 integrates a 400 MHz ARM core and a large number of DMA-capable peripheral controllers using a central bus 'matrix': PDF DataSheet

A platform chip is an SoC that is used in a number of products although chunks of it might be turned off in any one application: for example, the USB port might not be made available on a portable media player despite being on the core chip.

At the architectural design stage, to save the cost of a full crossbar matrix interconnect, devices can be allocated to busses with knowledge of the expected access and traffic patterns. Commonly there is one main bus master

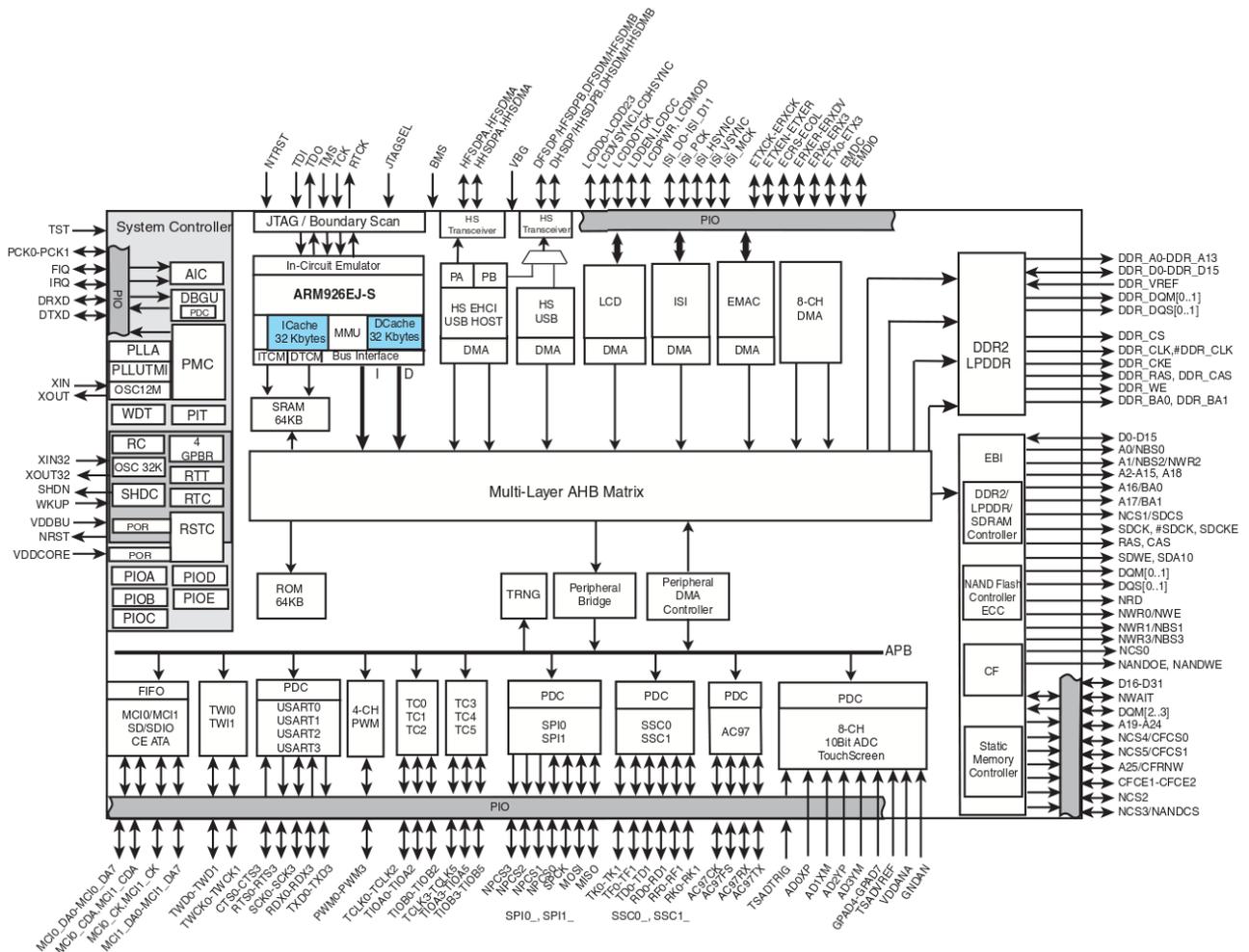


Figure 1.34: Platform Chip Example: Atmel SAM Series 9645.

per bus. The bus master is the device that generates the address for the next data movement (read or write operation).

Busses are connected to bridges, but crossing a bridge has latency and also uses up bandwidth on both busses. So we should allocate devices to busses so that inter-bus traffic is minimised based on a priori knowledge of likely access patterns.

Lower-speed busses may go off chip.

DRAM is always an important component that is generally off chip as a dedicated part. Today, some on-chip DRAM is being used in SoCs.

### 1.3 Architecture: Bus and Device Structure

Transmitting data consumes energy and causes delay. Basic physical parameters:

- Speed of light on silicon and on a PCB is 200 metres per microsecond.
- A clock frequency of 2 GHz has a wavelength of  $2E8/2E9 = 10$  cm.
- Within a synchronous digital clock domain we require connections to be less than (say) 1/10th of a wavelength.

Speed of light calculations

$$\lambda = 685 \text{ nm} = 685 \times 10^{-9} \text{ m}$$

$$v = ?$$

$$c = \lambda \cdot \nu \rightarrow \text{frequency } \frac{1}{s} \text{ or } \nu$$

Speed of light  $300 \times 10^8$

wavelength (in m)

Figure 1.35: Speed of light is a constant (and in silicon it is lower).

- Conductor series resistance further slows signal propagation and is dominant source of delay.
- So need to register a signal in several D-types if it passes from one corner of an 8mm chip to the other!
- Can have several thousand wires per millimetre per layer: fat busses (128 bits or wider) are easily possible.
- Significant DRAM is several centimetres away from the SoC and also has significant internal delay.

Hence we need to use protocols that are tolerant to being registered (passed through D-type pipeline stages). The four-phase handshake has one datum in flight and degrades with reciprocal of delay. We need something a bit like TCP that keeps multiple datums in flight. (Die stacking and recent DRAM-on-SoC approaches reduce wire length to a few mm for up to 500 MB of DRAM.)

But first let's revisit the simple hwen/rwen system used in the 'socparts' section.

### 1.3.1 A canonical D8/A16 Micro-Computer from 1980's

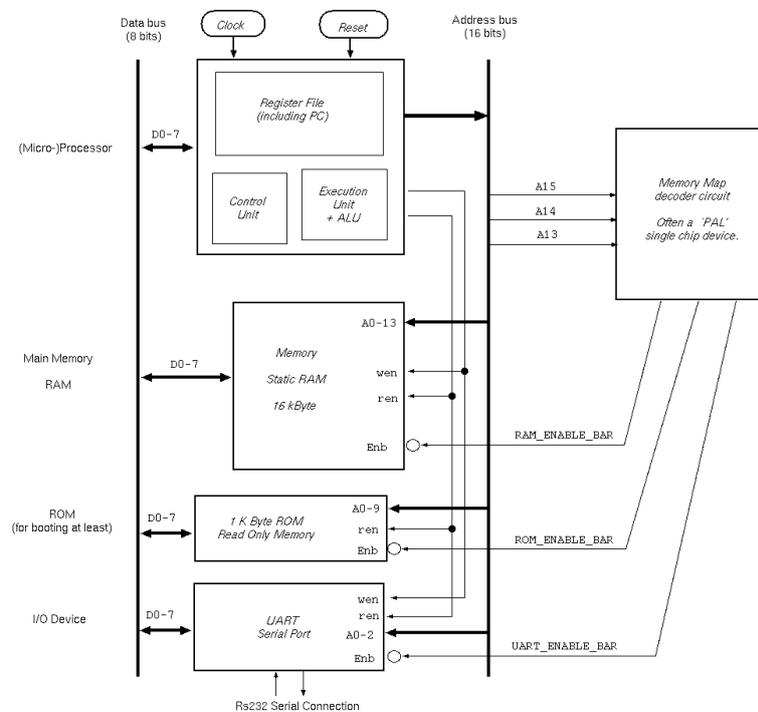


Figure 1.36: Early microcomputer structure, using a bidirectional/tri-state data bus.

Figure 1.36 shows the inter-chip wiring of a basic microcomputer (i.e. a computer based on a microprocessor). Busses of this nature were about 30cm long and had a cycle time of 250 ns or so.

### 1.3.2 Basic Bus: One initiator (II).

The bus in our early microcomputer was a true bus in the sense that data could get on and off at multiple places. SoCs do not use tri-states but we still use the term ‘bus’ to describe the point-to-point connections used today between IP blocks.

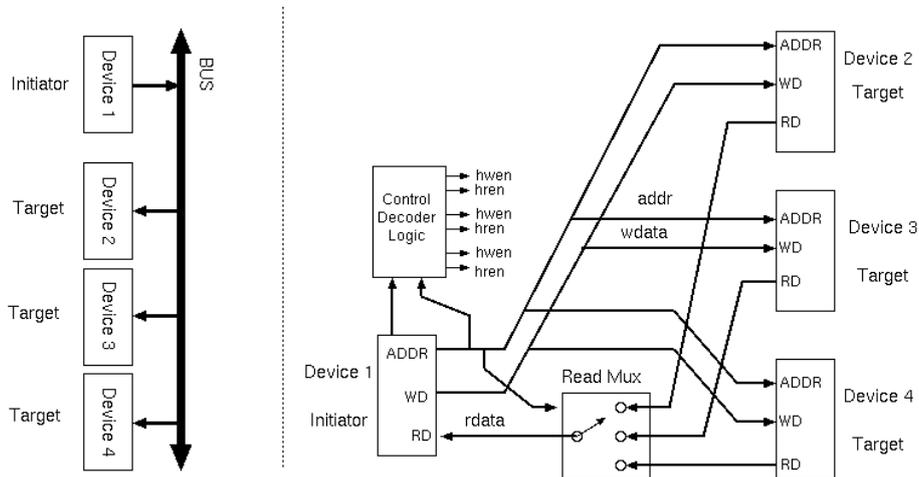


Figure 1.37: Example where one initiator addresses three targets.

Figure 1.37 shows such a bus with one initiator and three targets.

No tri-states are used: on a modern SoC address and write data outputs use wire joints or buffers, read data uses multiplexors.

Max throughput is unity (i.e. one word per clock tick). Typical SoC bus capacity:  $32 \text{ bits} \times 200 \text{ MHz} = 6.4 \text{ Gb/s}$ , but owing to protocol degrades with distance. This figure can be thought of as unity (i.e. one word per clock tick) in comparisons with other configurations we shall consider.

The most basic bus has one initiator and several targets. The initiator does not need to arbitrate for the bus since it has no competitors.

Bus operations are reads or writes. In reality, on-chip busses support burst transactions, whereby multiple consecutive reads or writes can be performed as a single transaction with subsequent addresses being implied as offsets from the first address.

Interrupt signals are not shown in these figures. In a SoC they do not need to be part of the physical bus as such: they can just be dedicated wires running from device to device. (For ESL higher-level models and IP-XACT representation, interrupts need management in terms of allocation and naming in the same way as the data resources.)

Un-buffered wiring can potentially serve for the write and address busses, whereas multiplexors are needed for read data. Buffering is needed in all directions for busses that go a long way over the chip.

### 1.3.3 Basic bus: Multiple Initiators (II).

Basic bus, but now with two initiating devices. Needs arbitration between initiators: static priority, round robin, etc.. With multiple initiators, the bus may be busy when a new initiator wants to use it, so there are

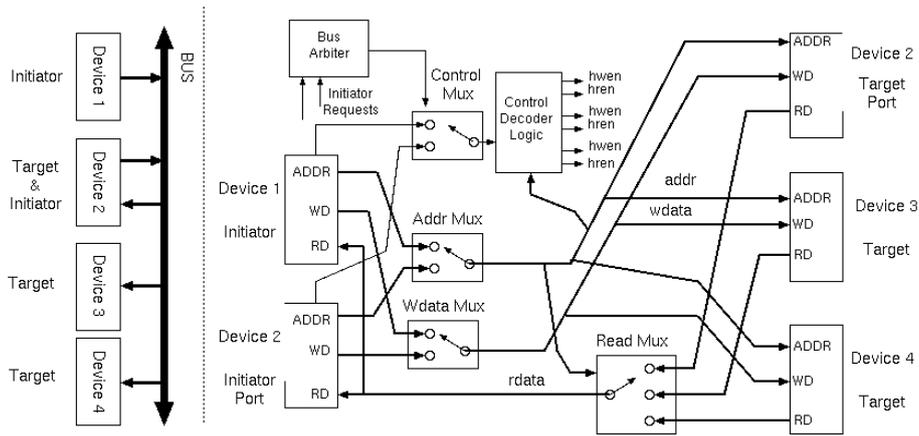


Figure 1.38: Example where one of the targets is also an initiator (e.g. a DMA controller).

various arbitration policies that might be used. Preemptive and non-preemptive with static priority, round robin and so on. The maximum bus throughput of unity is now shared among initiators.

Since cycles now take a variable time to complete, owing to contention, we certainly need acknowledge signals for each request and each operation (not shown).

How long to hold bus before re-arbitration? Commonly re-arbitrate after every burst. The latency in a non-preemptive system depends on how long the bus is held for. Maximum bus holding times affect response times for urgent and real-time requirements.

### 1.3.4 Bridged Bus Structures.

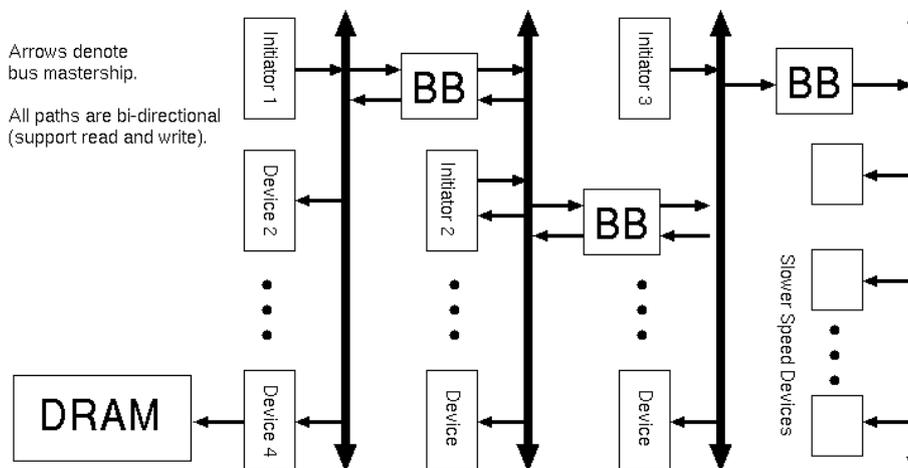


Figure 1.39: A system design using three main busses.

To make use of the additional capacity from bridged structures we need at least one main initiator for each bus. However, a low speed bus might not have its own initiators: it is just a slave to one of the other busses.

Bus bridges provide full or partial connectivity and some may write post. Global address space, non-uniform access time (NUMA). Some busses might be slower, narrower or in different clock domains from others.

The maximum throughput is the sum of that of all the busses that have their own initiators, but the achieved throughput will be lower if the bridges are used a lot: a bridged cycle consumes bandwidth on both sides.

How and where to connect DRAM is always a key design issue. The DRAM may be connected via a cache. The

cache may be dual ported on to two busses, or more.

Bus bridges, other glue components and top-levels of structural wiring are typically instantiated by an automated tool inside the SoC designer's favourite design environment. Schematic entry for the SoC top level is common. From the open source world, this may be an IP-XACT plugin for Eclipse. All major EDA vendors also supply their own design environment tools.

### 1.3.5 Classes of On-Chip Protocol

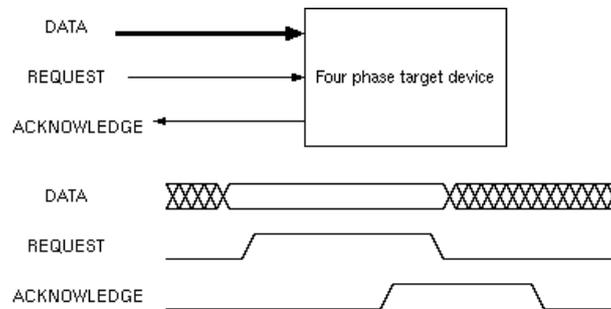


Figure 1.40: Timing diagram for an asynchronous, four-phase handshake.

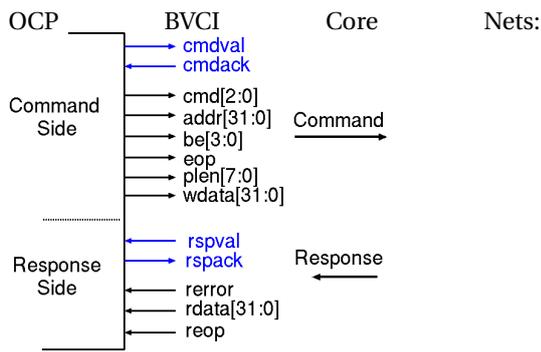
1. Reciprocally-degrading: such as handshake protocols studied earlier: throughput is inversely proportional to target latency in terms of clock cycles,
2. Delay-tolerant: such as AXI-lite and OCP's BVCI (below): new commands may be issued while awaiting responses from earlier,
3. Reorder-tolerant: such as full AXI: responses can be returned in a different order from command issue: helpful for DRAM access and needed for advanced NoC architectures.
4. Virtual-circuit: (beyond scope of this course): rather than putting a destination address and port number in each message, traffic is routed at each hop via a short circuit identifier (or routing tag) where mappings have been set up in advance in the routing nodes.
5. Separated send and acknowledge circuits: A decoupling between send and reply or send and acknowledge, perhaps using a priority mechanism or perhaps using physical separation of the two directions of flow, exists, to ensure responses can always be returned, hence avoiding a form of deadlock.
6. Credit flow controlled: (beyond scope of this course): each source has a credit counter per destination or per destination/port number pair, controlling how many packets it can send without receiver buffer over-run.

Labels or tags need to be added to each transaction to match up commands with responses.

The EACD+ARCH part Ib classes use the 'Avalon' bus on the Altera devices: Avalon Interface Specifications

For those interested in more detail: Comparing AMBA AHB to AXI Bus using System Modelling

Last year you used the Altera Avalon bus in part IB ECAD+Arch workshops. Many real-world IP blocks today are wired up using OCP's BVCI and ARM's AHB. Although the port on the IP block is fixed, in terms of its protocol, it can be connected to any system of bus bridges and on chip networks. Download full OCP documents from OCIP.org. See also bus-protocols-limit-design-reuse-of-ip



- All IP blocks can sport this interface.
- Separate request and response ports.
- Data is valid on overlap of req and ack.
- Temporal decoupling of directions:
- Allows pipeline delays for crossing switch fabrics or crossing clock domains.
- Sideband signals: interrupts, errors and resets: vary on per-block basis.
- Two complete instances of the port are needed if block is both an initiator and target.
- Arrows indicate signal directions on initiator. All are reversed on target.

A prominent feature is totally separate request and response ports. This makes it highly tolerant of delays over the network and amenable to crossing clock domains. Older-style handshake protocols where targets had to respond within a prescribed number of clock cycles cannot be used in these situations. However BVCi requests and responses must not get out of order since there is no id token.

For each half of the port there are request and acknowledge signals, with data being transferred on any positive edge of the clock where both are asserted.

If a block is both an initiator and a target, such as our DMA controller example, then there are two complete instances of the port.

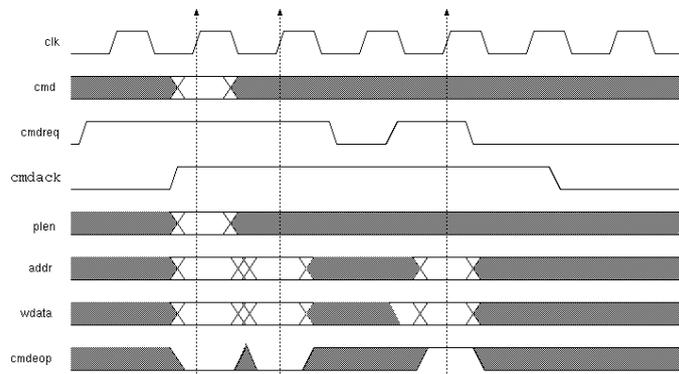
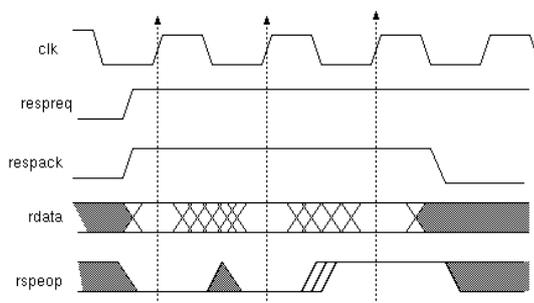


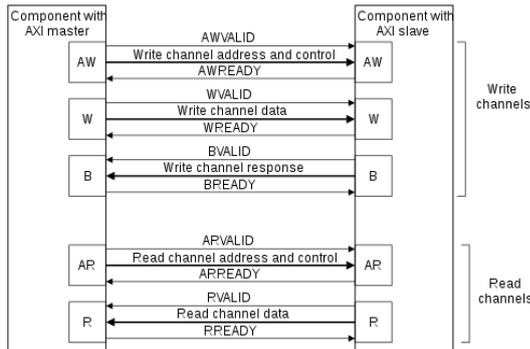
Figure 1.41: BVCi Protocol, Command Timing Diagram

Operations are qualified with conjunction of req and ack. Response and acknowledge cycles maintain respective ordering. Bursts are common. Successive addressing may be implied.



BVCi Response Portion Protocol Timing Diagram

### 1.3.6 ARM AXI Bus: The Current Favourite



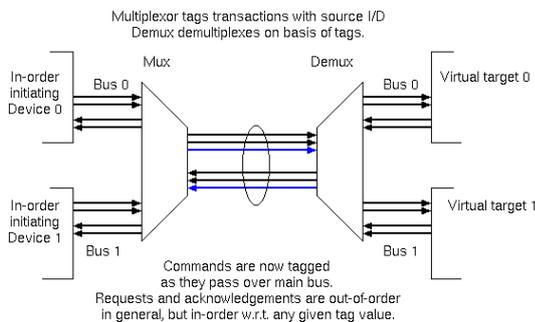
AXI-lite Bus Structure - Five temporally floating sub ports. One

AXI port if formed of five separate interfaces that are called channels: two for read, three for write. Each of the five has its own contra-directional READY/VALID pair with all other nets running in the VALID direction and qualified by the conjunction of ready and valid on a clock edge. In simple applications, the address and data channels for write will run close to lockstep, making a more natural total of four logical interfaces. Sequential consistency: the complete decoupling of the read and write aspects immediately raises the prospect of RaW and WaR hazards. A RaW hazard is where a read does not connect with the most-recently written data in the correct/nominal temporal order. AXI can be used with and without (AXI-lite) ordering tokens/tags.

AXI is widely used even for non-ARM products ARM AXI

### 1.3.7 Supporting out-of-order operation using tags.

Some initiators, particularly out-of-order CPU cores, issue multiple outstanding reads and can do useful work as soon as any of them are serviced. Some targets, particularly DRAM, can provide better performance by servicing requests out-of-order. If we multiplex a pair of in-order busses onto a common bus, yet tag all of the traffic from each bus on the common bus according to its in-order initiator, we have a tagged, out-of-order bus.



Out-of-order bus with tags.

The semantics are that for any given tag the requests and replies must be kept in order. The devices on the left may be separate initiator blocks, like processors and DMA controllers, or they may be different load/store ports on a common IP block, or, in theory, any mix. For the targets on the right, there is no difference between demultiplexing to separate in-order targets and a single target that understands tags.

The tags above are used to correlate results to replies over an out-of-order bus. To preserve sequential consistency between, say, separate load/store ports on a CPU, which would have their own IDs a fence mechanism of some sort is also needed. Fences preserve RaW and WaW orderings: no transaction is allowed to overtake others in a way that would make it jump over a fence in the time domain. (In the OCP/BVCI bus, tag numbers are/were used in a different way from AXI: a fence is implied when an initiator increased a tag number.) On the AXI bus there are no fences. Instead, an initiator must just wait for all outstanding responses to come back before issuing a transaction on any of its load/store ports that is after a fence.

For AXI load-linked, store-conditional and other atomic operations, the command fields in the issuing channels contain additional nets and code points that indicate whether the transaction is exclusive in this way.

### 1.3.8 Network on Chip: Simple Ring.

A two-level hierarchy of bridged rings is sometimes a sweet spot for SoC design. For example, IBM Cell Broadband Engine uses dual rings. At moderate size, using a fat ring (wide bus links) is better than a thin X-bar design for same throughput in terms of power consumption and area use.

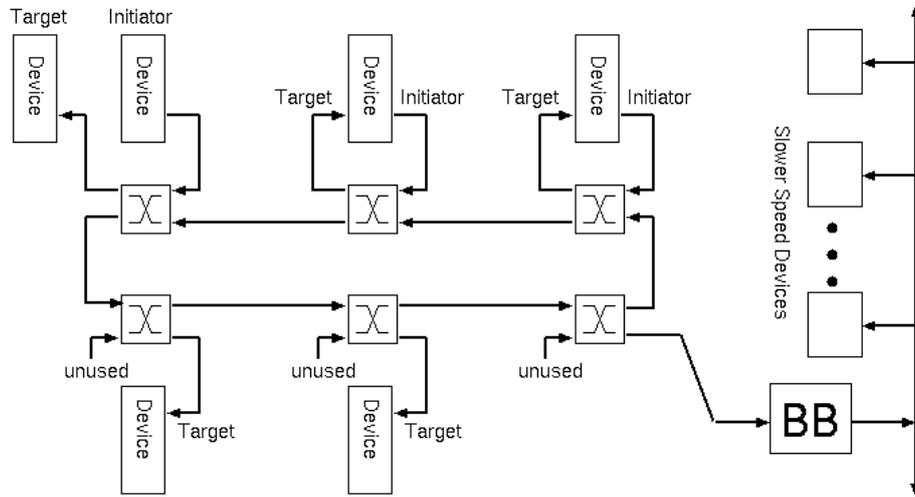


Figure 1.42: A ring network: a low-complexity network on chip structure.

A two-by-two switch element enables formation of rings (and other NoC structures). The switch element is registered: hence ring network can span the chip. A higher-radix element allows more devices to be connected at a 'station'. Performance: Single ring: throughput=2. Dual counter-rotating rings: throughput=4.

With ring (and certainly with all more complex NoCs) IP block protocol/interface needs to support decoupled requests and response packets.

Ring has local arbitration in each element, but global policies are required to avoid deadlock and starvation.

Ring gives priority to traffic already on the ring and uses LAN-like buffering at source, hence no requirement for queuing in element.

Ring does not carry interrupts or other sideband signals.

Switched networks require switching elements. With a 2x2 element it is easy to build a ring network. The switching element may contain buffering or it may rely on back-pressure to make sources reduce their load.

Single ring: throughput=2. Counter-rotating ring (one ring in each direction): throughput=4 since a packet only travels 1/4 of the way round the ring on average.

Using a network, the delay may be multiple clock cycles and so a **write posting** approach is reasonable. If an initiator is to have multiple outstanding read requests pending it must put a token in each request that is returned in the response packet for identification purposes.

Although there can be effective local arbitration in each element, a network on a chip can suffer from deadlock. Some implementations use separate request and response networks, so that a response is never held up by new requests, but this just pushes deadlock to the next higher logical level when some requests might not be serviceable without the server issuing a subsidiary request to a third node. Global policies and careful design are required to avoid deadlock and starvation.

### 1.3.9 Network on chip: Switch Fabrics.

A simple ring is not very effective for above small tens of nodes. Instead, richer meshes of elements are used and the elements can have a higher radix, such as 4x4.

There are a number of well-known switch wiring schemes, with names such as Benes, Clos, Shuffle, Delta, Torus, Mesh, Express-Mesh, Butterfly. These vary in terms of the complexity and contention ratios. Note even a full-crossbar (any input to any output in unit time), which is very costly, still suffers from output port contention, so rarely justified on performance grounds, but uniform access delays make it easy to provide sequential consistency (see my Comparative Architecture notes).

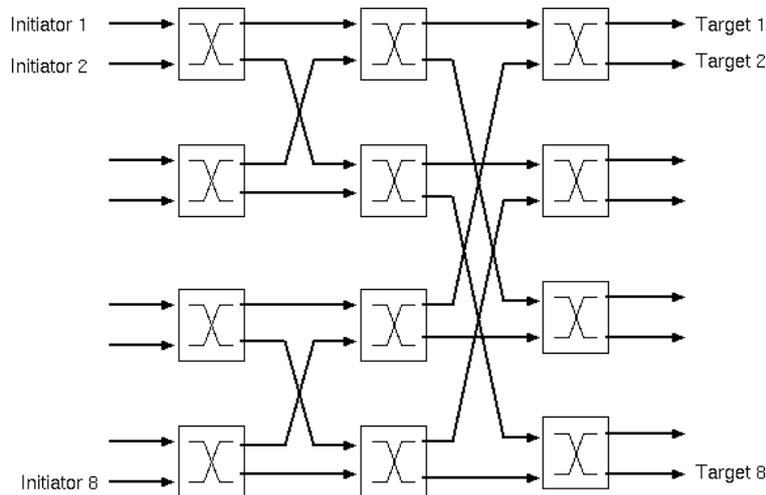


Figure 1.43: A more-complex switching fabric: more wiring, more bandwidth and less fabric contention than ring (but still has output port contention).

Illustrated is using two-by-two switch element connects eight devices in three stages. Using a higher-radix (e.g. 4) is common. The throughput is potentially equal to the number of ports, but the fabric may partially block and there may be uneven traffic flows leading to receiver contention. These effects reduce throughput. Typically will not need quite as many initiators as targets, so a symmetric switch system will be over provisioned.

Can be overly complex on the small scale, but scale ups well. See Network On Chip Synthesis Tool: Mullins NetGen Network Generator. RDM NoC Notes

### 1.3.10 Network on Chip: Higher Dimensions.

(Not examinable for part II).

Can we consider higher-dimensional interconnect? The hypercube has lowest diameter for number of customers. But it has excessive wiring. Chips are two-dimensional so perhaps it's good to use a 2-D network? But this may be overly conservative. Maybe use 2.5-D? have a small number of 'multi-hop' links?

On benign (load-balanced) traffic, the flattened butterfly approaches the cost/performance of a butterfly network and has roughly half the cost of a comparable performance Clos network.

Further details (non examinable): "The advantage over the Clos is achieved by eliminating redundant hops when they are not needed for load balancing." See '*Flattened butterfly: a cost-efficient topology for high-radix networks*' by John Kim, William J. Dally, Dennis Abts.

Typical NoC designs use credit-based flow control where counters remotely track the amount of receive buffer space available at the destination. A so-called 'cut-through' routing node starts forwarding the start of a packet before the end of it is received. This is the alternative to store-and-forward.

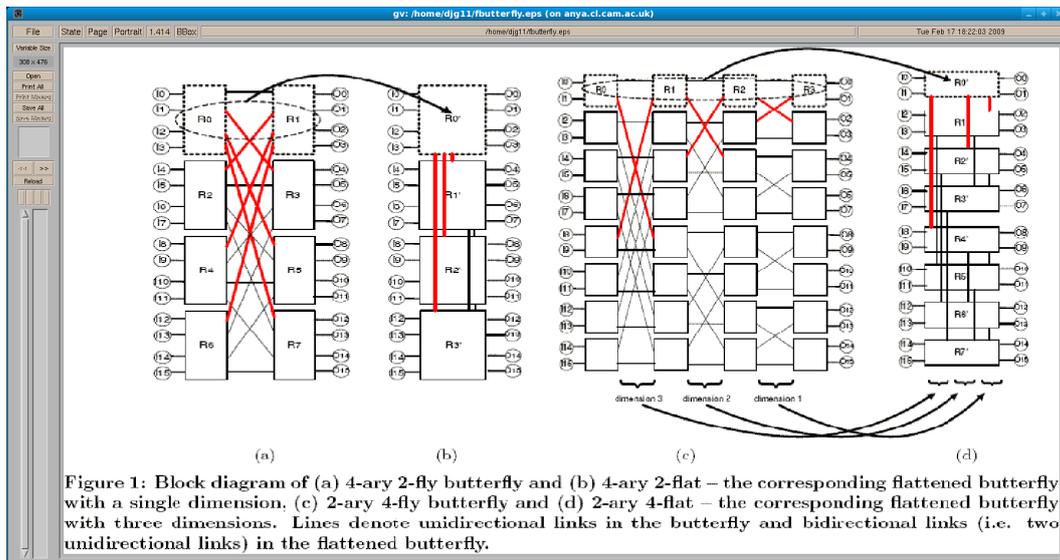


Figure 1.44: The 'Flattened Butterfly' network topology.

### 1.3.11 NoC Modelling

Do we want to model every contention point and queuing detail ?

Use a high-level model: Treat the NoC just as a square array corresponding to the floor plan of the chip and in each entry we hold a running average local utilisation.

- Add delay penalty to traversing transaction based on  $1/(1-p)$ , Computer Systems Modelling
- Log local energy consumption proportional to delay,
- Target routing protocol can be used unmodified or skipped.

Or model even more abstractly: uses

Problems with higher than cycle-accurate modelling:

- Transactions may be out of order if using large quantum L/T model (see later).
- Deadlock may be missed.

### 1.3.12 On-chip Busses Summary.

Multiplexing using tri-states is common at the PCB level but active multiplexors result in less wire being charged and so are used on-chip to save energy.

It is handy if all of the IP blocks to be integrated conform to a common bus bus port standard.

Automatic synthesis of glue logic and memory maps is possible (see HLS section of these notes, if lectured).

Formal specifications of bus ports are widely used, assisting in tool automation and ABD.

The AMBA AHB bus from ARM Cambridge was widely used: but is quite complex (e.g. when resuming from a split burst transaction) and was intolerant of additional pipeline stages being inserted between initiator and responder (i.e. had no temporal decoupling).

The OCP BVCi supports temporal decoupling, but requests and responses must not overtake: hence it can cross clock domains and tolerate pipeline stages. But it cannot tolerate re-ordered responses (e.g. from a DRAM).

Full AXI includes tokens/tags on each operation so support out-of-order request/response association. This suits many advanced initiators and targets, but on-chip networks must preserve transaction ordering within an tag value to avoid WaW hazards.

Other busses: The Wishbone bus and IBM CoreConnect bus: used by various public domain IP blocks and various designs (e.g. RTL OpenRISC). The OpenRISC used in some practical materials on the course web site uses Wishbone. [Wikipedia Wishbone Core Connect](#)

## KG 2 — Energy use in Digital Hardware.

Battery life is very important for convenience. The electricity bill is sometimes the biggest cost in a data centre [citation needed!]. Saving energy in computing is always a good idea. In this section we will examine energy and performance and energy saving techniques.

Energy in an electronic device gets used in several different ways: for a Mobile Phone we might see the following budget:

- **Screen Backlight:** 1 to 2 watts
- **RF Transmissions:** via the various transmit antennae: up to 4 watts
- **Sound and Vibrations:** through the speaker and little shaker motor 300 mW
- **Heat:** 'wasted' in the electronics: up to 5 watts

A mobile phone might have a 10 Wh capacity (36 kilo Joules).

### 2.1 Basic Physics

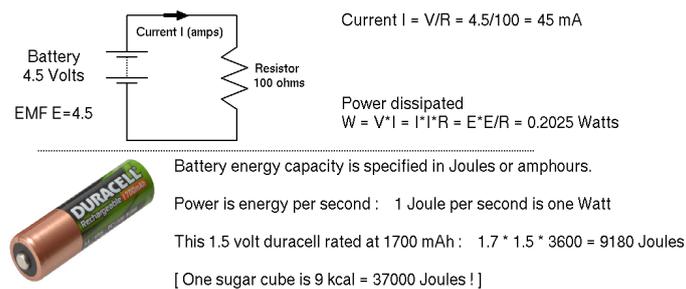


Figure 2.1: Ohms Law, Power Law and Battery Capacity.

We need to know that power (watts) is voltage (volts) times current (amps). Power (watts) is also energy (joules) per unit time (second). And a joule of energy is one coulomb of charge dropping one volt in potential.

$$P = V \times I = E \times f$$

We also need to understand a little about capacitance. A capacitor stores energy.

The wiring capacitance for a pair of conductors, surrounded by silicon dioxide (permittivity 4), is  $\pi\epsilon_0\epsilon_r$  in Farads per metre, which is roughly 100 pF/m. This applies fairly well for VLSI nets where the units of 0.1 pF/mm or 0.1 fF/um are more useful.

#### 2.1.1 Energy of Computation (1)

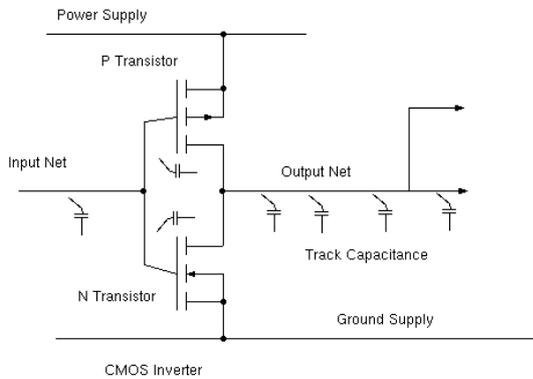
It is critical to understand where the energy goes in hardware in physical terms. All our models build up from this basis. Dynamic energy arising from stray capacitance turns out to be our main energy user today.

Gate current  $I = \text{Static Current (leakage)} + \text{Dynamic Current}$ .

Dynamic current = Short-circuit current + Dynamic charge current.

Early CMOS (VCC 5 volts): negligible static current, but today at VCC of 0.9 to 1.3 volts it can be 30 percent of consumption for some (high-leakage) designs.

The short-circuit power can generally be neglected. It is the energy wasted when both the P and N transistors are briefly conducting at once as the output. It is only significant when an input is left floating or slowly changing.



Dynamic charge current computation:

- All energy in a net is wasted each time output goes from one to zero.
- The energy in a capacitor was  $E = CV^2/2$ .
- The same amount of energy is wasted charging it up.
- Dominant capacitance is proportional to net length.
- Gate input and output capacitance also contribute to C.

The **toggle rate** for a net is the average frequency of a change in value, which is twice the activity ratio multiplied by the clock frequency.

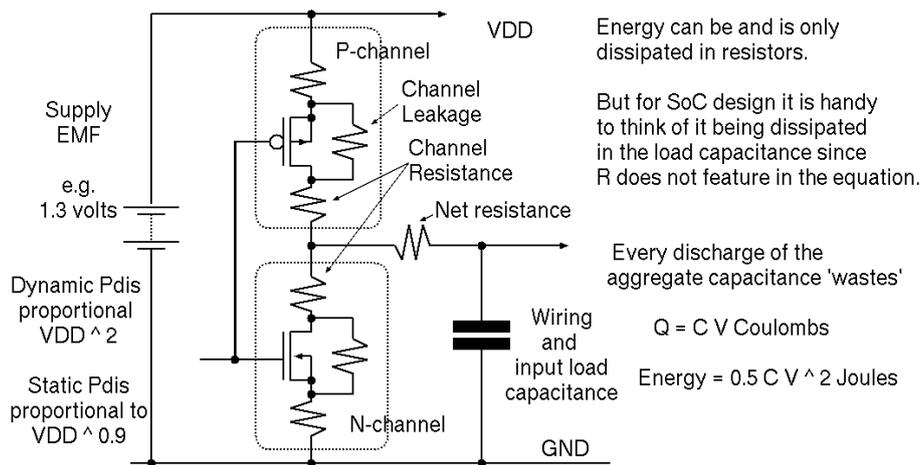


Figure 2.2: Dynamic energy dissipation mechanism.

Capacitors do not consume energy - they only store it temporarily. Only resistors dissipate energy in logic circuits, but their resistance does not feature in the energy use formula. The energy in the wiring capacitance is 'wasted' on each complete transition cycle. If the clock frequency is  $f$  and a net has activity ratio  $\alpha$  (the fraction of clock cycles it transitions from one to zero) then the energy used is

$$E = f * \alpha * C * V^2$$

As we increase voltage, dynamic power grows quadratically. Static power grows a little better than (i.e. less than) linear since transistors may turn off more fully.

Note that although we divide by two as part of the standard formula for the energy in a capacitor, this quantity of energy is wasted both in the charging network on the zero-to-one transition and in the discharging network on the one-to-zero transition. So we can drop the half.

Note: static power consumption is static current multiplied by supply voltage ( $P=IV$ ). Page 30 or so of this

cell library has real-word examples: 90nm Cell Library See also the power formula on the 7400A data sheet: 74LVC00A.pdf Further details: Power Management in CPU Design.

### 2.1.2 Landauer Limit and Reversible Computation

There are theoretical limits on the energy an irreversible computation requires. Current technology is a long way off these in two respects:

- we use too much energy representing and communicating bits and
- we use Vonn Neumann based computation which does not scale well.

Let's consider electrical computers:

- If we build a computer using a network of standard components (such as transistors), where the interconnection pattern expresses our design intent, then the components must be at different spatial locations. The computer will have some physical volume.
- If we interconnect our components using electrical wires these will have capacitance, resistance and inductance that stop them behaving like ideal conductors. The smaller the volume, the less wire we need and the better the wires (and hence computer) will work.
- If we use transistors that need a swing of about 0.7 volts on their gates to switch them reliably between off and on, then our wires need to move through at least that much potential difference.

The capacitance of the wires turns out to be our main enemy. Given a prescribed minimum voltage swing, the energy used by switching a wire between logic levels can only be made smaller by reducing its area and hence capacitance. Hence smaller computers are better.

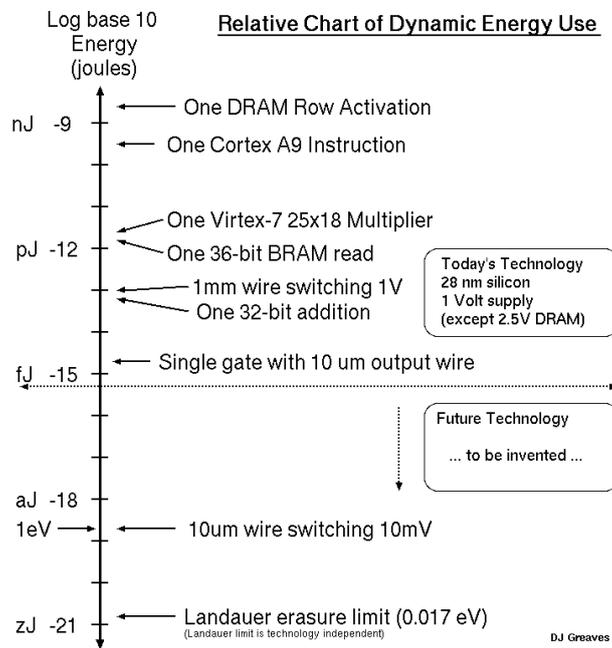


Figure 2.3: Relative chart of dynamic energy use.

The traditional approach of wasting the energy of each transition can be countered by returning the charge to the power supply (eg. Asynchrobatic logic for low-power VLSI design by David John Willingham, Univ Westminster, 2010) and reversible logic (eg. Toffoli logic) can get below the Landauer limit. Some standard algorithms

such as encryption and lossless compression are reversible at the large - the trick is mooted to be to code in a way that does not delete intermediate results during the computation. Such techniques may be in wide use within a decade or perhaps two.

Figure 2.3 shows ballpark figures for dynamic energy use in today's (2016/2018) mainstream 28 nm silicon technology. We see that contemporary computers are about six orders of magnitude above the Landauer limit in terms of energy efficiency, so there is a lot of improvement still possible before we have to consider reversibility. (Log to base 18 months of one million is roughly 30 years, so 2050 is the reversible computing deadline.)

### 2.1.3 Typical Numerical Values

Let's just run some numbers for a fictional chip made solidly of logic:

22 um geometry: Capacitance per micron 0.3 fF

Energy of discharge for a net of 1mm at VDD of 1 volt is 0.15 pJ

Simple gate area: 0.3 square um.

If we assume a sub-system of 1000 gates, its area is  $1000 \times 0.3 = 300$  sq um

By Rent or otherwise, we have an average wiring length of  $0.3 \times \text{sqrt}(300) = 5.2$  um.

Clocking at 200 MHz with an activity ratio of 0.1, dynamic power will be  $1000 \times 200e6 \times 0.1 \times 5.2 \times 0.3e-15 = 31$  uW.

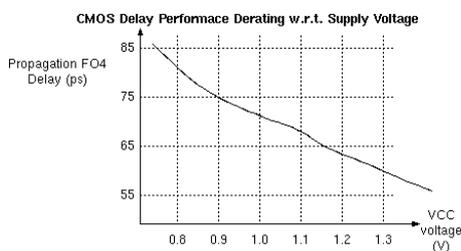
We will assume the inputs transition quickly enough for us to neglect short-circuit currents. Now we must add on the static power and the energy to drive the outputs from the sub-system:

Static power depends on leakage, but might be one fifth the dynamic power at 1 volt, although depends greatly on activity ratios and choice and semiconductor doping levels.

If we assume ten output nets with an activity of 0.1 and length 250 micron, their energy will be  $10 \times 200e6 \times 0.1 \times 250 \times 0.3e-15 = 150$  uW.

If we assume a 1 sq centimetre chip is made up of  $1e8/300 = 3e5$  such units, the chip power would be  $3e5(150e-6 + 31e-6 + 6e-6) = 56$  watts. Clearly, we either need fluid cooling (water or ethanol heat pipe) or else the Dark Silicon approach. The activity ratio of 0.1 (toggle rate of 20 percent) reflects very-busy logic, such as an AES encoder. Most subsystems have lower activity ratios when in use. And the average activity ratio depends on how frequently used the block is, of course.

### 2.1.4 Gate Delay as a Function of Supply Voltage



Transistors have a gate threshold voltage around which they switch from off to on. This limits our lowest possible supply voltage. Above this, logic delay in CMOS is roughly inversely proportional to supply voltage. Accordingly, to operate faster, we need a higher supply voltage for a given load capacitance. CMOS Delay Versus Supply Voltage

$$\text{Gate Delay} \propto \frac{C * V}{(V - V_t)^2}$$

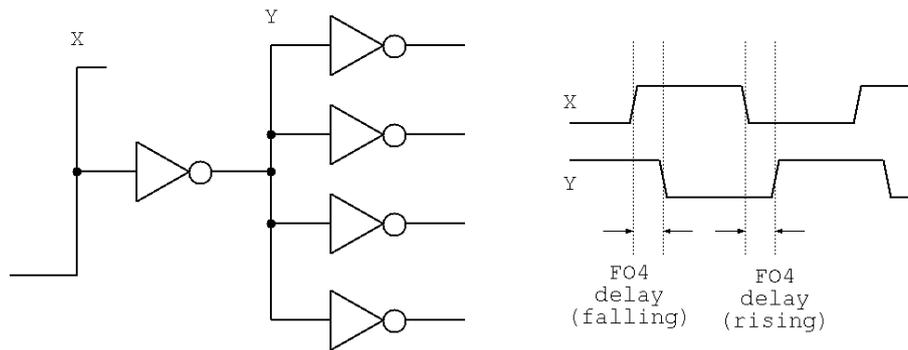


Figure 2.4: Fanout-4 delay measurement.

The **FO4 delay** is the delay through an inverter that is feeding four other nearby inverters (fan out of four). This is often quoted as a reference characteristic of a logic technology. The combinational delay of a particular design can also be expressed in a technology-independent way by quoting it in units of FO4 delay.

Here is a complete demo of simulating a CMOS inverter made from two MOSFETs using hspice.

```
// spice-cmos-inverter-djg-demo.hsp
// Updated 2017 by DJ Greaves
// Based on demo by David Harris harrisd@leland.stanford.edu

////////////////////////////////////
// Set supply voltage

////////////////////////////////////
// Declare global supply nets and connect them to a constant-voltage supply
.global Vdd Gnd
Vsupply Vdd Gnd DC 'VddVoltage'

////////////////////////////////////
// Set up the transistor geometry by defining lambda

.opt scale=0.35u * Define lambda // This is half the minimum channel length.

// Set up some typical MOSFET parameters.
//http://www.seas.upenn.edu/~jan/spice/spice.models.html#mosis1.2um

.MODEL CMOSN NMOS LEVEL=3 PHI=0.600000 TOX=2.1200E-08 XJ=0.200000U
+TPG=1 VTO=0.7860 DELTA=6.9670E-01 LD=1.6470E-07 KP=9.6379E-05
+UO=591.7 THETA=8.1220E-02 RSH=8.5450E+01 GAMMA=0.5863
+NSUB=2.7470E+16 NFS=1.98E+12 VMAX=1.7330E+05 ETA=4.3680E-02
+KAPPA=1.3960E-01 CGDO=4.0241E-10 CGSO=4.0241E-10
+CGBO=3.6144E-10 CJ=3.8541E-04 MJ=1.1854 CJSW=1.3940E-10
+MJSW=0.125195 PB=0.800000

.MODEL CMOSP PMOS LEVEL=3 PHI=0.600000 TOX=2.1200E-08 XJ=0.200000U
+TPG=-1 VTO=-0.9056 DELTA=1.5200E+00 LD=2.2000E-08 KP=2.9352E-05
+UO=180.2 THETA=1.2480E-01 RSH=1.0470E+02 GAMMA=0.4863
+NSUB=1.8900E+16 NFS=3.46E+12 VMAX=3.7320E+05 ETA=1.6410E-01
+KAPPA=9.6940E+00 CGDO=5.3752E-11 CGSO=5.3752E-11
+CGBO=3.3650E-10 CJ=4.8447E-04 MJ=0.5027 CJSW=1.6457E-10
+MJSW=0.217168 PB=0.850000
```

```

////////////////////////////////////
// Define the inverter, made of two mosfets as usual, using a subcircuit.

.subckt myinv In Out N=8 P=16 // Assumes 5 lambda of diffusion on the source/drain
m1 Out In Gnd Gnd CMOSN l=2 w=N
+ as='5*N' ad='5*N'
+ ps='N+10' pd='N+10'
m2 Out In Vdd Vdd CMOSP l=2 w=P
+ as='5*P' ad='5*P'
+ ps='P+10' pd='P+10'
.ends myinv

////////////////////////////////////
// Top-level simulation netlist
// One instance of my inverter and a load capacitor
x1 In Out myinv // Inverter
C1 Out Gnd 0.1pF // Load capacitor

////////////////////////////////////
// Stimulus: Create a waveform generator to drive In
// Use a "Piecewise linear source" PWL that takes a list of time/voltage pairs.

Vstim In Gnd PWL(0 0 1ns 0 1.05ns 'VddVoltage' 3ns VddVoltage 3.2ns 0)

////////////////////////////////////
// Invoke transient simulation (that itself will first find a steady state)

.tran .01ns 6ns // Set the time step and total duration
.plot TRAN v(In) v(Out)
.end
// To get an X-windows plot, run the following interactive sequence of commands:
// $ ngspice spice-cmos-inverter-djg-demo.hsp
// ngspice 11 -> run
// ngspice 12 -> plot v(In) v(Out)

// end of file

```

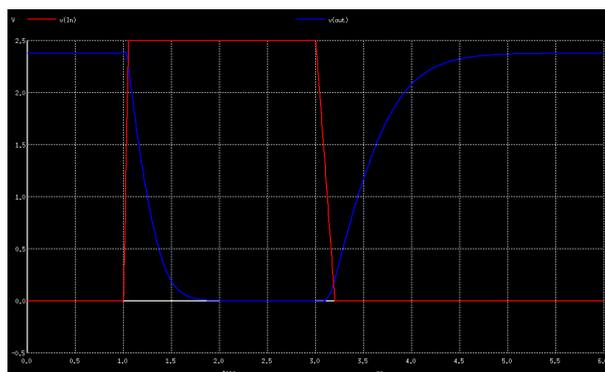


Figure 2.5: Plot when running from a VCC supply of 2.5 volts. Red is stimulus and blue is output.

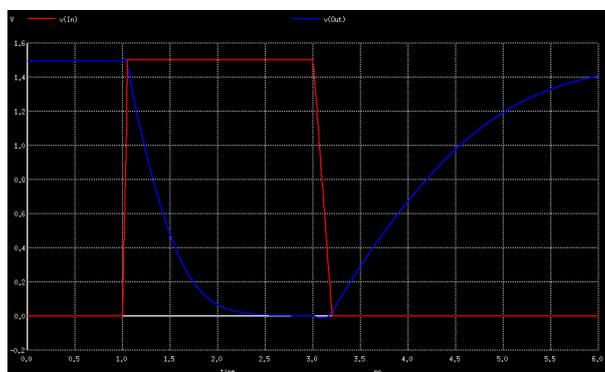


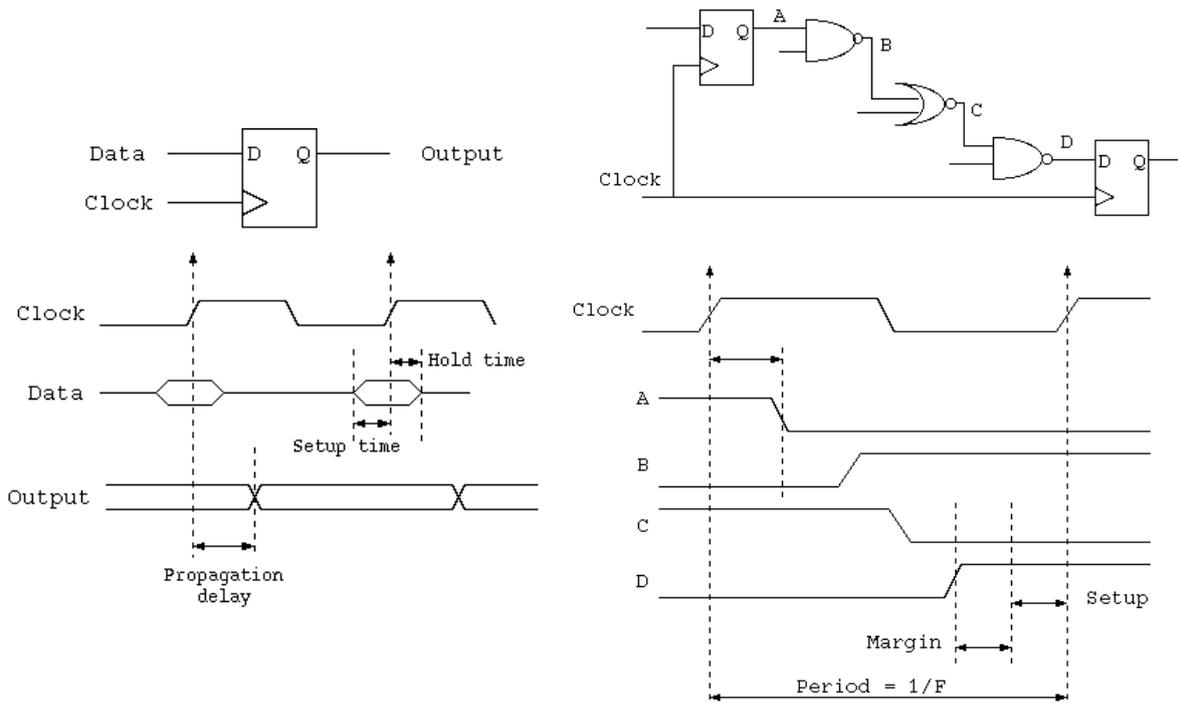
Figure 2.6: Plot when running at 1.5 volts.

The output load capacitor experiences a fairly typical exponential charge and discharge shape. The shape is

not a true  $1 - \exp(-t/CR)$  curve owing to some non-linearity in the MOSFETs. But it is pretty close. If the FETs had equal on resistance at the two supply voltages, although the swing of the output in the two plots would be different, the delays before they cross the half-supply level would be identical. The difference arises owing to the on resistance being less when the gate voltage is less (i.e. when it is closer to the transistor threshold voltage).

**2.1.5 Detailed Delay Model.**

Now we have looked at the energy model, the other primary metric for a design can be considered: its delay performance that limits the clock rate.



The maximum clock frequency of a synchronous clock domain is set by its critical path. The longest path of combinational logic must have settled before the setup time of any flip-flop starts.

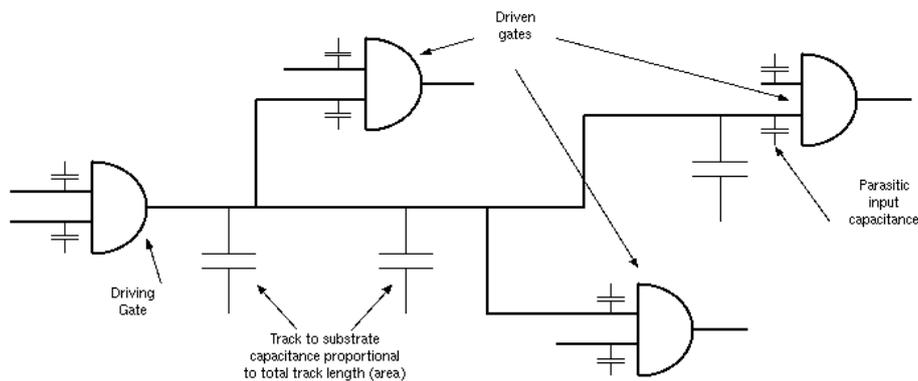


Figure 2.7: Logic net with tracking and input load capacitances illustrated.

Both the power consumption and effective delay of a gate driving a net depend mainly on the length of the net driven.

The delay is modelled with this formula:

$$\text{device delay} = (\text{intrinsic delay}) + (\text{output load} \times \text{derating factor}).$$

The track-dependent output loading is a library constant times the track area. The load-dependent part is the sum of the input loads of all of the devices being fed. For short, non-clock nets (less than 0.1 wavelength), we just include propagation delay in the gate derating and assume the signal arrives at all points simultaneously.

Precise track lengths are only known after place and routing (Figure 6). Pre-layout and pre-synthesis we can predict net lengths from Rent's Rule and RTL-level heuristics.

Figure 2.7 shows a typical net, driven by a single source. To change the voltage on the net, the source must overcome the stray capacitance and input loads. The fanout of a gate is the number of devices that its output feeds. The term *fanout* is also sometimes used for the maximum number of inputs to other gates a given gate is allowed to feed, and forms part of the design rules for the technology.

The speed of the output stage of a gate, in terms of its propagation delay, decreases with output load. Normally, the dominant aspect of output load is capacitance, and this is the sum of:

- the capacitance proportional to the area of the output conductor,
- the sum of the input capacitances of the devices fed.

To estimate the delay from the input to a gate, through the internal electronics of a gate, through its output structure and down the conductor to the input of the next gate, we must add three things:

- the internal delay of the gate, termed the intrinsic delay
- the reduction in speed of the output stage, owing to the fanout/loading, termed the derating delay,
- the propagation delay down the conductor.

The propagation delay down a conductor obeys standard transmission line formula and depends on the distributed capacitance, inductance and resistance of the conductor material and adjacent insulators. For circuit board traces, resistance can be neglected and the delay is just the *speed of light* in the circuit board material: about 7 inches per nanosecond, or 200 metres per microsecond. On the other hand, for shorter nets on chip, less than one tenth a wavelength long, we commonly assume the signal arrives at all destinations at once and model the propagation delay as an additional inertial component of the driving gate and include this via the gate derating.

On a chip, the speed of light can be neglected because chips are physically small, but the resistance of the aluminium conductors is sufficiently large to have an effect for critical applications, such as master clock signals.

Older technologies used polysilicon interconnections which had significant resistance and so the different gates connected to a polysilicon net would experience different arrival times of a signal at their inputs.

A conductor made of aluminium on a chip is small enough for the speed of light delay to be ignored. Its resistance is generally small enough for it to be considered at the same voltage at all points. (critical clock nets are sometimes studied in more detail).

On a circuit board, a copper conductor is often too long for the speed of propagation of signals to be neglected, but its resistance is always so low that the influence of resistance may be neglected (except for power supply traces — but we are not interested in the speed of signals on power conductors).

In both cases, it is generally sufficient to model the propagation delay of the net as an increase in the output delay of its driver(s) using the formula below. In addition, output drivers slow down in themselves as the output net length and fanout is increased owing to capacitance and this effect normally dominates net propagation delay. Hence the standard model above.

The upshot of this is that on a chip, we can mostly neglect the propagation aspects of delay, whereas on a circuit board, we need to model certain critical conductors as components in themselves. These have a simple delay model, whose value can be set by post-routing back annotation.

## 2.2 Memory Power and Performance

As well as logic gates, our main users of power will be SRAM, DRAM and I/O pads.

### 2.2.1 RAM - On-chip Static Random-Access Memory (Static RAM).

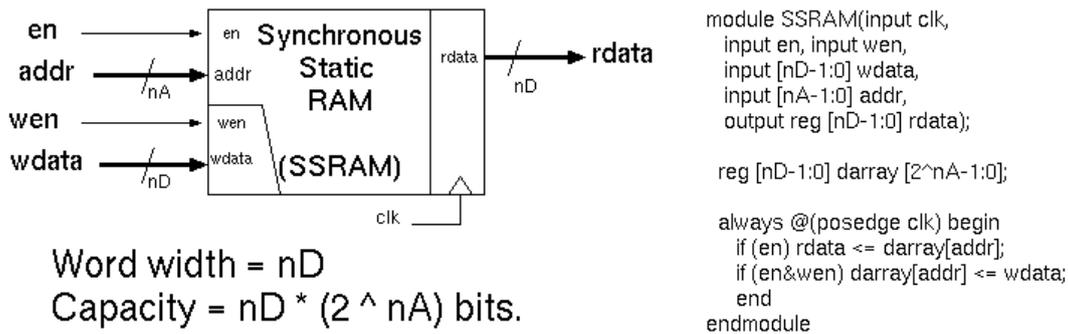


Figure 2.8: Synchronous static RAM with single port: logic symbol and internal RTL model.

RAMs vary in their size and number of ports. Single-ported SRAM is the most important and most simple resource. It connects to a bus as an addressable target. It is also used inside caches for tags and data. Today's SoC designs have more than fifty percent of their silicon area devoted to SRAM for various purposes. Commonly, synchronous RAMs are used, requiring typically one clock cycle of delay between address input and corresponding data output. The same address can be written with fresh data during the same clock cycle, if desired.

The illustrated RAM has a one clock cycle read latency. When a write occurs, the old value at the location is still read out, which is commonly a useful feature.

The 'en' input signal is not strictly needed since the RAM could deliver read data on all cycles. However, this wastes power, so without an enable input we should ensure the address inputs are relatively stable when a RAM result is not needed.

Owing to RAM fabrication overheads, RAMs below a few hundred bits should typically be implemented as register files made of flip-flops. But larger RAMs have better density and power consumption than arrays of flip-flops.

RAMs for SoCs were originally supplied by specialist companies such as Virage and Artizan (although these are now part of larger EDA companies). A 'RAM compiler' tool is run for each RAM in the SoC. It reads in the user's size, shape, access time and port definitions and creates a suite of models, including the physical data to be sent to the foundry.

High-density RAM (e.g. for L2 caches) may clock at half the main system clock rate and/or might need error correction logic to meet the system-wide reliability goal.

RAM consumes static and dynamic energy. The ESL section of these notes gives high-level modelling figures that include about 10 pJ per read or write operation and a leakage of 82 nW per bit.

**Additional notes:**

On-chip SRAM needs a test mechanism. Various approaches:

- Can test with software running on embedded processor.
- Can have a special test mode, where address and data lines become directly controllable (JTAG or otherwise).
- Can use a built-in hardware self test (BIST) wrapper that implements 0/F/5/A and walking ones typical tests.

Larger memories and specialised memories are normally off-chip for various reasons:

- Large area: would not be cost-effective on-chip,
- Specialised: proprietary or dense VLSI technology cannot best be made on a SoC die where the process is optimised for general logic,
- Specialised: e.g. non-volatile process (such as FLASH)
- Commodity parts: economies of scale (ZBT SRAM, DRAM, FLASH)

But in the last five years DRAM and FLASH have found their way onto the main SoC as maturing technology shifts the economic sweet spot.

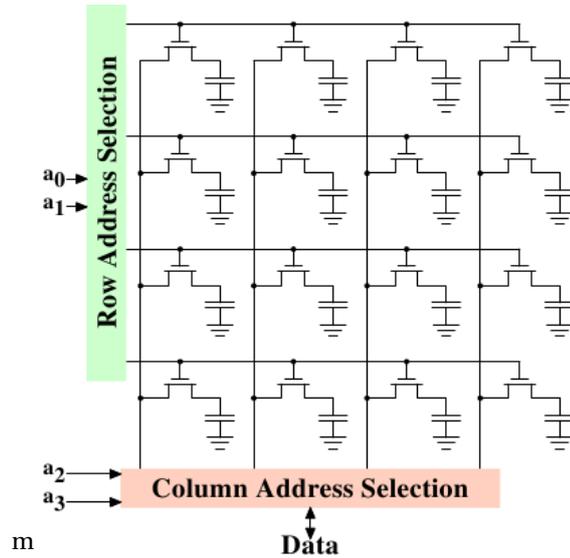
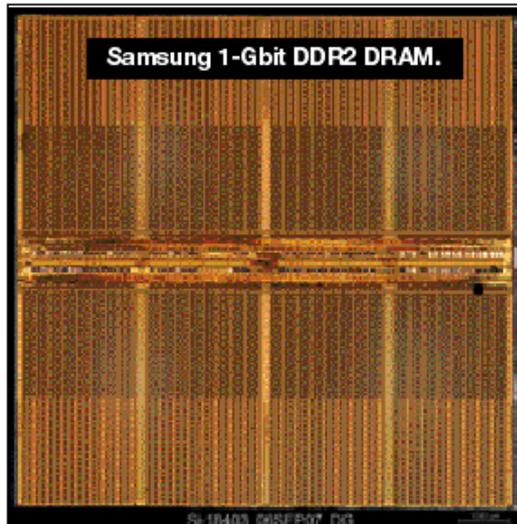
### 2.2.2 Dynamic RAM : DRAM



Figure 2.9: DRAM single-in-line memory module (SIMM).

DRAMs for use in PCs are mounted on SIMMS or DIMMS. But for embedded applications, they are typically just soldered to the main PCB. Normally one DRAM chip (or pair of chips to make D=32) is shared over many sub-systems in, say, a mobile phone. SoC DRAM compatibility might be a generation behind workstation DRAM: e.g. using DDR3 instead of DDR4. Also, the most recent SoCs embed some DRAM on the main die or flip-chip/die-stack the DRAM directly on top of the SoC die in the same package (multi-chip module — MCM).

Modern DRAM chip with  
8 internal memory banks.



These are the pin connections of a typical DIMM from 2010:

Clk+/-	Clock (200MHz)	wq[7:0]	Write lane qualifiers
Ras-	Row address strobe	ds[7:0]	Data strobes
Cas-	Column address strobe	dm[7:0]	Data masks
We-	Write enable	cs-	Chip select
dq[63:0]	Data in/out	addr[15:0]	Address input
reset	Power on reset	bs[2:0]	Bank select
		spd[3:0]	Serial presence detect

DRAM performance is often quoted in MT/s which is mega-transfers per second. Our DIMM example has a 200 MHz clock and hence 400 MT/s. This is low performance by today's standards: 64 bits times 400 MHz gives 25.6 Gb/s peak (4 GB/sec). The capacity is a 1 Gbyte DIMM made of 8 chips.

The latest (Jan 2018) DDR4 memories operate at 4332 MT/sec. Each transfer carries a word the width of the DRAM data bus (e.g. 16 bits) and transfers are performed on both edges of a clock. This clock would be at 2.166 GHz. But that is the burst transfer rate. To manage a burst, DRAM timings of 19-21-21 are used, which is the number of clock cycles to send the row address, the column address and for writeback, respectively.

In the worst case, if the DRAM is currently 'open' on the wrong row, 61 clock cycles will be needed to change to the new location. Roughly the same number of clock cycles again will be used in pipeline stages through the various memory hierarchy levels of the controlling device.

This DRAM has four data I/O pins and four internal planes, so no bank select bits. (Modern, larger capacity DRAMs have multiple such structures on their die and hence additional bank select inputs select which one is addressed.)

Dynamic RAM keeps data in capacitors. The data will stay there reliably for up to four milliseconds and hence every location must be read out and written back (refreshed) within this period. The data does not need to leave the chip for refresh, just transferred to the edge of its array and then written back again. Hence a whole

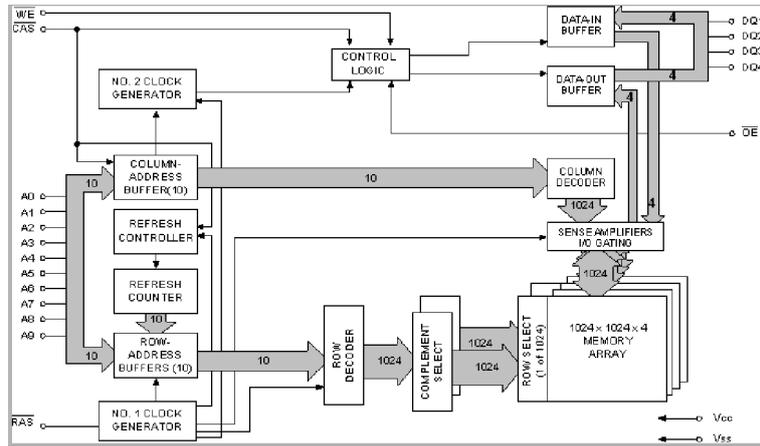


Figure 2.10: Single-bank, 4-bit wide, DRAM Chip Internal Block Diagram.

- ▶ Die Size Efficiency
    - DRAM is as much as 6x smaller in comparison to SRAM on a per bit basis.
    - Cell size for SRAM roughly 6x the size of a DRAM cell
      - 6 transistor cell versus 1 transistor
    - For DRAM approximately 55% - 70% of the die size is array
    - Periphery circuitry is 45% to 50% larger for SRAM
- Non multiplexed addressing

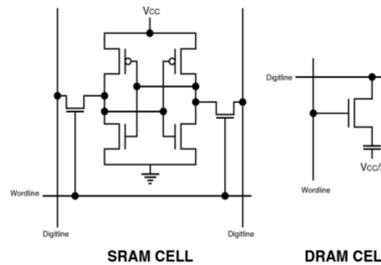


Figure 2.11: SRAM cell complexity versus DRAM cell

row of each array is refreshed as a single operation.

DRAM is not normally put on the main SoC chip(s) owing to its specialist manufacturing steps and large area needs. Instead a standard part is put down and wired up. (DRAM is traded as a commodity like corn and gold.)

A row address is first sent to a bank in the DRAM. This is called opening the row or a row activation. One a row is open, one has random access to the columns of that row using different column addresses. The DRAM cells internally have destructive read out because the capacitors get discharged into the row wires when accessed. Therefore, whenever finished with a row, the bank containing it goes busy while it writes back the data and gets ready for the next operation (charging row wires to mid-way voltage etc.).

DRAM is slow to access and certainly not ‘random access’ compared with on-chip RAM. A modern PC might take 100 to 300 clock cycles to access a random part of DRAM, but the ratio may not be as severe in embedded systems with lower system clocks. Nonetheless, we typically put a cache on the SoC as part of the memory controller. The controller may embody error detection or correction logic using additional bit lanes in the DRAM.

The cache will access the DRAM in localised bursts, saving or filling a cache line, and hence we arrange for cache lines to lie within DRAM rows.

The controller may keep multiple banks open at once to exploit tempo-spatial access locality.

DRAM controller is typically coupled with a cache or at least a write buffer.

DRAM: high latency and write-back overheads imply we must select a bank closing policy. The best controllers will lookahead in a pool of pending requests to assist decisions on when to do write back (aka close or deactivate). It is normal to prioritise reads over writes, but overtaking must be avoided or else reads can be served

from the write queue. But a new request falling in a previously open line can arrive just after we closed it! It is best if clients can tolerate responses out-of-order (hence use bus/NoC structure that supports this).

Controller must

- set up DRAM control register programming,
- set clock frequency and calibrate delay lines,
- implement specific RAS-to-CAS latencies and many other timing details,
- and ensure refresh happens.

Controller often contains a tiny CPU to interrogate serial device data. DRAM refresh overhead has minimal impact on bus throughput. For example, if 512 refresh cycles are needed in 4 ms and the cycle rate is 200E6 the overhead is 0.1 percent.

Another design consideration is how the system address bus is mapped to the various row, bank and column physical bits. This will alter the effect of memory layout on performance. Hardware is normally programmable in terms of physical address interleave. An example is, starting with most significant in physical address space: bank, row, col, burst offset, byte lane. Byte lane and burst offset are always at the bottom and col is kept lower than row, but having bank lower than col gives interleaving of accesses to open pages which is sensible when the system workload has a lot of localised activity to one large area, whereas having bank higher makes sense when the system has various concurrent active hot spots, such as typical with heap, stack, code and static segments. etc..

Here we are using the definition of a bank as a region where only one row is active within one bank. Multiple banks may be within the same channel (which always occurs for banks within one DRAM chip) or arranged over separate channels were a channel is defined as a data bus (or DRAM controller backside port). Multiple channels gives higher data throughput owing to spatial diversity.

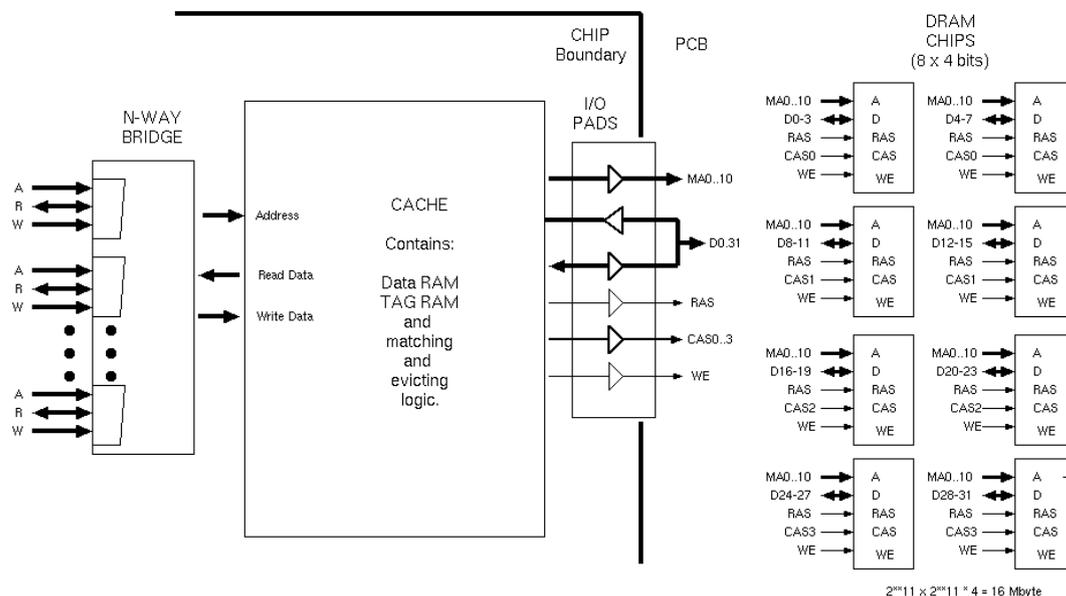


Figure 2.12: Typical structure of a small DRAM subsystem.

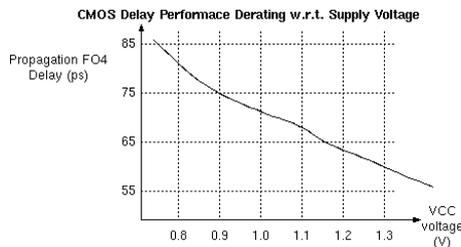
Figure 2.12 shows a 32-bit DRAM subsystem. Four CAS wires are used so that writes to individual byte lanes are possible. For large DRAM arrays, need also to use multiple RAS lines to save power by not sending RAS to un-needed destinations.

We will discuss DRAM energy use in the ESL section of the course. A main energy use is static power in the high-speed PCB driving and receiving pads. Each row activation and column selection takes dynamic energy.

## 2.3 The Voltage and Frequency Relationship

Looking at the derating graph for the standard cell libraries, we see that in the operating region, the frequency/voltage curve is roughly linear. CMOS delay is inversely proportional to supply voltage. A technique that exploits such curves is DVFS – dynamic voltage with frequency scaling.

Logic with higher-speed capabilities is smaller which means it consumes greater leakage current which is being wasted while we are halted. Also leakage energy is proportional to supply voltage (or perhaps sublinear with exponent 0.9ish : as we raise voltage, transistors are indeed turned off more, but  $P=IV$  is still increasing).



If we vary the voltage to a region dynamically, while keeping  $f$  constant, a higher supply voltage uses more power (square law) but would allow a higher  $f$ .

Let's only raise VCC when we ramp up  $f$ : classical DVFS.

For a fixed task size, energy use is proportional to  $V$  squared, so DVFS is the ideal method (i.e. for predictable, real-time tasks in low-leakage technology):

1. Adjust clock  $f$  for just-in-time completion (e.g. in time to decode the next frame of a video),
2. then adjust VCC to minimal value for reliably meeting the set-up time.

In general servers (ie. not for a static/finite workload), ramping voltage up linearly with clock frequency ( $f$ ) results in dynamic power consumption with a cubic dependence on  $f$ . But work may be bursty, so DVFS is applied (e.g. by a laptop governor).

DVFS obtains peak performance under heavy loads, yet avoid cubic overhead when idle. We adjust VCC so that, at all times, the logic just works. However, we need to keep close track of whether we are meeting real-time and timing closure deadlines.

In a server farm processing blade we may be thermally limited, so DVFS will be throttled back by rack-level governors or Intel's RAPL.

### Additional notes:

Combinational logic cannot be clock gated (e.g. PAL and PLA). For large combinational blocks: can dip power supply to reduce static current when block is completely idle (detect with XORs).

So a typical SoC uses not only many dynamic clock gated islands, but also some sub-containers with automatic frequency and voltage variation. Power isolation originally used on a longer and larger scale (complete continents) but now a lot of power islands are being used.

It is possible to locally and quickly adjust supply voltage with a series transistor - but wasteful compared with an off-chip switched-mode regulator.

An off-chip power supply can be efficiently adjusted, but limited to only a few voltage islands and tens of milliseconds inertia.

### 2.3.1 DVFS Example

Example: core area  $64 \text{ mm}^2$ ; average net length  $0.1 \text{ mm}$ ;  $400\text{K gates/mm}^2$ ,  $a = 0.25$ .

Net capacitance =  $0.1 \text{ mm} \times 1 \text{ fF/mm} \times 400\text{K} \times 64 \text{ mm}^2 = 2.5 \text{ nF}$

Supply Voltage (V)	Clock Freq (MHz)	Static Power (mW)	Dynamic Power (mW)	Total Power (mW)
0.8	100	40	24	64
1.35	100	67	68	135
1.35	200	67	136	204
1.8	100	90	121	211
1.8	200	90	243	333
1.8	400	90	486	576

The table shows example power consumption for a circuit when clocked at different frequencies and voltages. The important thing to ensure is that the supply voltage must be sufficient for the clock frequency in use: too low a voltage means that signals do not arrive at D-type inputs in time to meet set up times.

Power consumption versus frequency is worse than linear: it goes with a power law.

In the past, chips were often **core-bound** or **pad-bound**. Pad-bound meant that the chip had too many I/O signals for its core logic area: the number of I/O's puts a lower bound on the perimeter of the chip. Today's VLSI technology allows I/O pads in the middle of the chip and designs are commonly **power-bound**.

### 2.3.2 Silicon Power and Technology

Total consumption = Gate Power + Wiring Power.

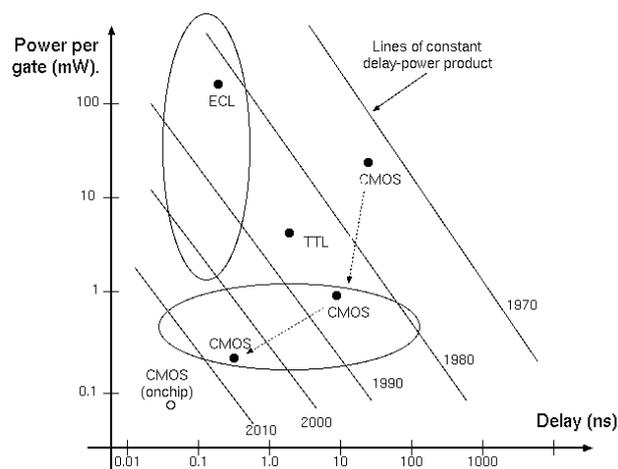


Figure 2.13: Historic energy used when a discrete NAND2 gate switches: delay-power product.

Gate switching speed is dominated by electron mobility (drift velocity) in transistor gates. Gate intrinsic energy used to be the main power consumer.

Historically, for a given technology, the product of delay and power consumption of a gate was largely constant, leading to design trade offs. Units of the product are the Joule: the energy for a logic transition in a gate.

We can improve by shifting to faster materials, such as GaAs, or just by making the gates smaller and the wires shorter. How small can we go: what is the *silicon end point*? Unanswered still!

But for modern designs, wiring power owing to capacitance is the main obstacle to performance (unless we return to water cooling again). (Remember, in reality the 'wiring power' is mostly dissipated in the gates, but the nets also have resistance.)

At any one time, there is a choice of implementation technologies. Here is the speed-power product for three



Figure 2.14: Water-cooled PC (built by Russell Hardle)

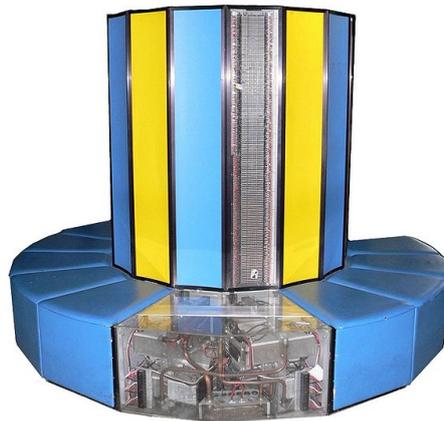


Figure 2.15: Cray XMP - wiring distance minimised by circular design.

versions of the 7400-format quad NAND gate, fabricated from different contemporary technologies in 1985. (This is a board-level part and on-chip much less driving power is needed).

Year	Technology	Device	Propagation delay (ns)	Power (mW)	Product (pJ)
1975	CMOS	CD4011BE	120 ns	(10 mW)	(1200 pJ)
1985	CMOS	74HC00	7 ns	1 mW	7 pJ
1985	TTL	74F00	3.4 ns	5 mW	17 pJ
1985	ECL	SP92701	0.8 ns	200 mW	160 pJ
2007	CMOS	74LVC00A	2.1 ns	120 uW	0.25 pJ
2012	CMOS	74LVC00A	... the same	: dominated by packaging	

CMOS has been dominant, and in 2007 is the only surviving technology: 74LVC00A.pdf

Within each of the the RTL and ECL technology, further exchange of power for speed is possible by adjusting

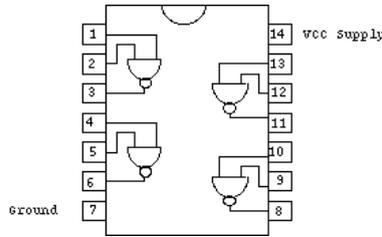


Figure 2.16: The 7400 standard part has been in manufacture using this pinout for about 40 years, so allows comparison, but is seldom used today.

resistor values.

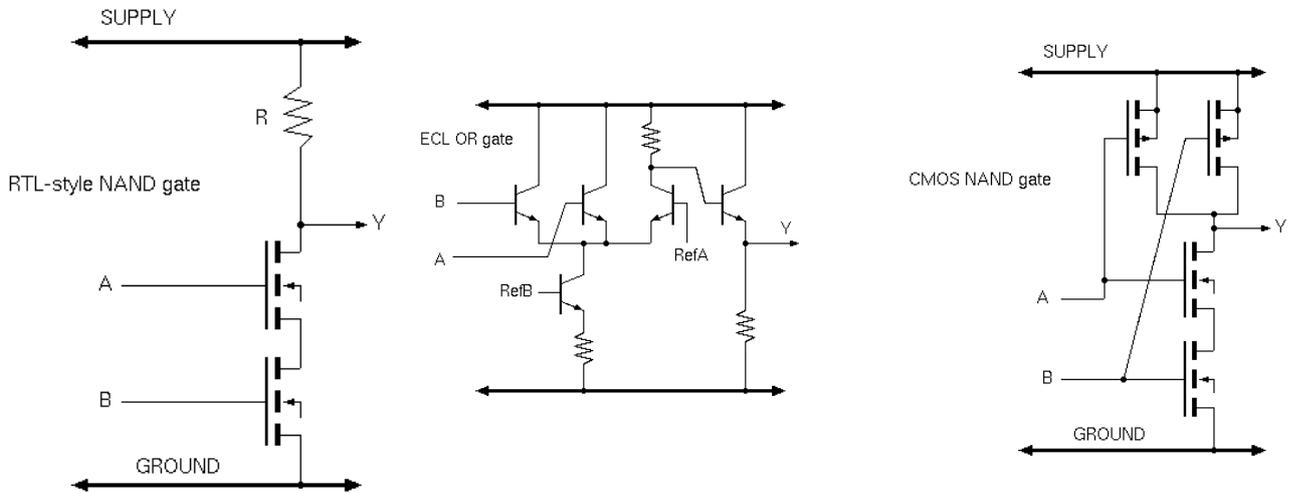
The 5 volt CMOS gate has the property that it consumes virtually no power when not changing its output. Today's lower voltage CMOS does not turn the transistors off as much, leading to significant static leakage currents.

The ECL gate is an older technology, hence a slightly higher speed-power product, but was still useful at that time since still the fastest.

Gates of medium complexity or larger (rather than SSI gates as these are) tend to be an order better in speed or power, since they do not have output stages designed for driving long nets.

Alternatives to silicon, such as GaAs have been proposed for general purpose logic. GaAs has four times higher electron mobility and so transistors of a given size switch on and off that much faster. However, increases in the speed of silicon, simply by making things smaller, have turned out to be a more effective way forward. So far!

Sleep Transistors



**Additional notes:**

Resistor-transistor logic (RTL) dissipates power all the time while output is logic zero. This *static* power is inversely proportional to R. R needs to be smaller for constant loading capacitance to get RC, the switching time constant, smaller for higher-frequency operation.

Note, RTL pre-dates field effect transistors (FETs): bipolar transistors and thermionic valves were actually used. Also a pair of diodes might be used as the AND part of a NAND gate with a single transistor or valve used for the inverter stage.

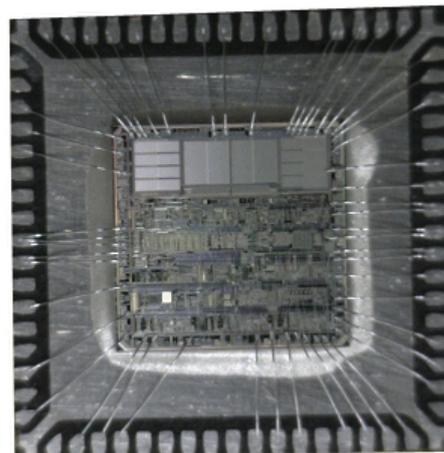
Emitter-coupled logic also uses an output load resistor that has the same power/speed trade off problem. It goes faster than equivalent power RTL and TTL since it does not saturate the transistors and hence they recover faster.

The breakthrough in CMOS was that either the upper or lower transistor path was always broken while the output was static, hence zero static power consumption and energy use was entirely proportional to the load capacitance being charged on a zero-to-one output transition. Modern CMOS runs on lower supply voltages, meaning that transistors do not turn off entirely and hence static leakage exists, bringing back the problems mentioned above.

**2.3.3 90 Nanometer Gate Length.**

The mainstream VLSI technology in the period 2004-2008 was 90 nm. This had low leakage and very high wafer yields. Now the industry is using 22 nanometer and smaller. Parameters from a 90 nanometer standard cell library:

Parameter	Value	Unit
Drawn Gate Length	0.08	$\mu\text{m}$
Metal Layers	6 to 9	layers
Max Gate Density	400K	gates/ $\text{mm}^2$
Finest Track Width	0.25	$\mu\text{m}$
Finest Track Spacing	0.25	$\mu\text{m}$
Tracking Capacitance	1	fF/mm
Core Supply Voltage	0.9 to 1.4	V
FO4 Delay	51	ps
Leakage current		nA/gate



Typical processor core: 200k gates + 4 RAMs: one square millimeter. Typical SoC chip area is 50-100  $\text{mm}^2$   $\rightsquigarrow$  20-40 million gates (semi-custom/standard cell). Actual gate and transistor counts are higher owing to full-custom blocks (RAMs mainly).

- 2007: Dual-core Intel Itanium2: 1.6 billion transistors (90 nm).
- 2010: 8-core Intel Nehalem: 2.3 billion transistors (45 nm).
- 2010: Altera Stratix IV FPGA: 2.5 billion transistors (40 nm).
- 2015: Intel CPU: circa 10 billion transistors (19 nm).
- 2018: ITRS predicts 7 nm technology for this year (2018)!

Moore's Law Transistor Count Dennard Scaling

Dimension Increase in Metal-Oxide-Semiconductor Memories and Transistors

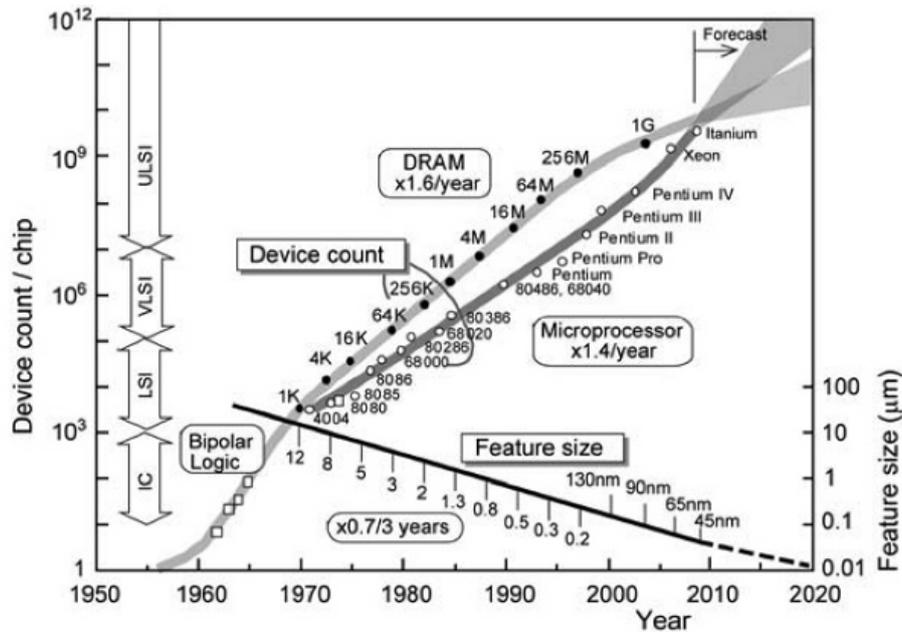


Figure 2.17: Technology Scaling Prediction.

**Dennard's Rule** stated that as transistors get smaller, their power density stays constant, so that the power use stays in proportion with area: both voltage and current scale (downward) with length. This meant that no new heat extraction technology was needed as VLSI capabilities improved. But once a supply voltage of 1 volt was reached, silicon CMOS cannot be run at much lower voltages without leakage (static power) greatly increasing.

#### Additional notes:

Typical modern datasheet rubric:

"The Xilinx Kintex UltraScale™ FPGAs are available in -3, -2, -1, and -1L speed grades, with -3 having the highest performance. The -1L devices can operate at either of two VCCINT voltages, 0.95V and 0.90V and are screened for lower maximum static power. When operated at VCCINT = 0.95V, the speed specification of a -1L device is the same as the -1 speed grade. When operated at VCCINT = 0.90V, the -1L performance and static and dynamic power is reduced."

The slide shows typical parameters from a 90 nanometer standard cell library. This figure refers to the width of the gate in the field effect transistors. The smaller this width, the faster the transistor can operate, but also it will consume more power as static leakage current. The 90 nm figure was the mainstream VLSI technology in the period 2004-2008, but then 40-45 nanometer technology was widely used with smaller 22 nm now mainstream.

Typical processor core: 200k gates + 4 RAMs: one square millimeter.

A typical SoC chip area is 50-100 mm<sup>2</sup> with 20-40 million gates. Actual gate and transistor count would be higher owing to custom blocks (RAMs mainly), that achieve a better density than standard cells.

Moore's Law has been tracked for the last three plus decades, but have we now reached the *Silicon End Point*? That is, can we no longer make things smaller (at the same cost)? Modern workstation processors have certainly demonstrated a departure from the previous trend of ever rising clock frequencies: instead they have several cores.

The **Power Wall** is currently the limiting factor for practical VLSI. As Horowitz points out, the fixed threshold voltage of transistors means that supply voltages cannot be reduced further as we go to smaller and smaller geometries, hence the previous technology trajectory will change direction: Scaling, Power, and the Future of CMOS. The limiting factor for commercial products has become the cost of thermal management. We can put more-and-more transistors on our chip but we cannot use them all at once - hence **Dark Silicon**.

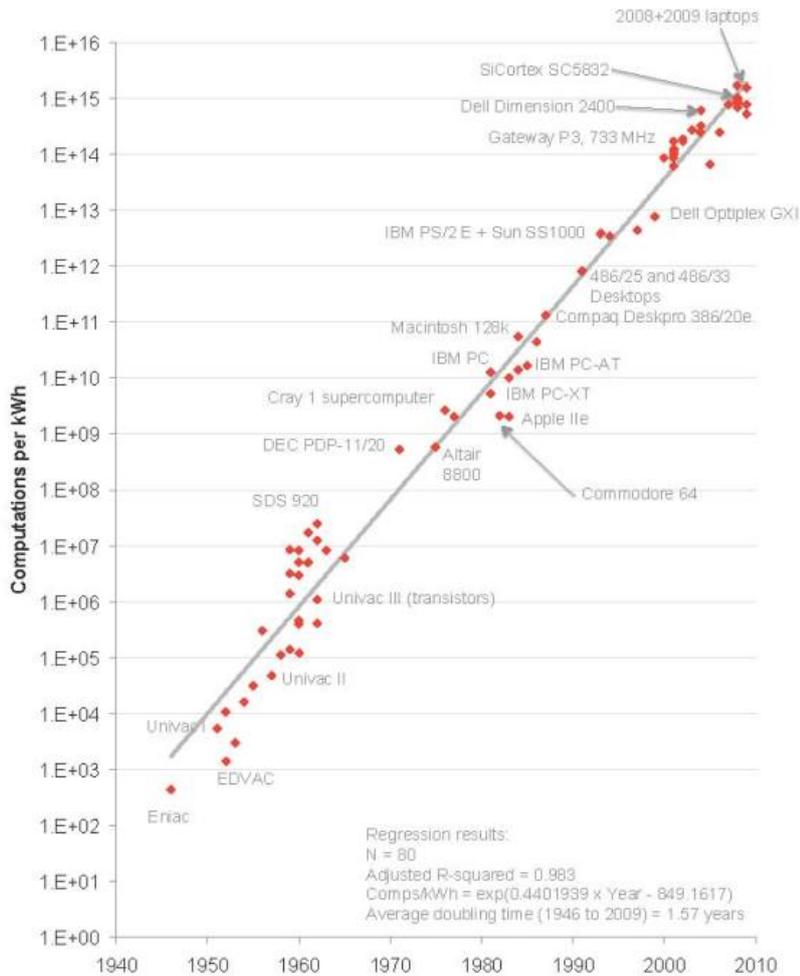


Figure 2.18: Computations per kWh. Divide by  $3.6E6$  for per Joule.

### 2.3.4 Deep Submicron and Dark Silicon

Basic physical parameters for different technologies. Metal and polysilicon resistance are growing (bad) as we shrink.

## 2.4 Further Power Saving Techniques

Turning off and slowing down are our friends when it comes to saving power.

We can save power by controlling power supplies and clock frequencies: Figure 2.24.

Our first power saving technique uses DVFS. We clock quickly and halt or clock slowly and finish just in time. We aim to clock at the lowest sufficient voltage and clock frequency and with  $\alpha$  as high as possible and minimal halt cycles.

Frequency scaling means adjusting the clock frequency to a subsystem. The voltage will have to be scaled in tandem to ensure timing is met without wasting energy. Frequency adjustment can be instant if divider

Technology	0.25	0.18	0.15	0.13	0.09
Vdd (Volts)	2.5	1.8	1.5	1.2	1
No. of Metal layers	5	6	7	9	10
Dielectric constant	4.1	3.7	3.6	2.9	2.9
Resistivity - M1 (ohm-cm)	75	80	120	90	100
Resistivity - N+/Poly (Ohm/Ct)	7	8	13	11	20
Resistivity - via (Ohm/Ct)	4	7	9	1	1.5
Ca (aF/um)	15	11	11	15	14
Gate Delay (ps)	40	32	20	17	12

Figure 2.19: Technology Scaling (transistor on and off states getting closer!).

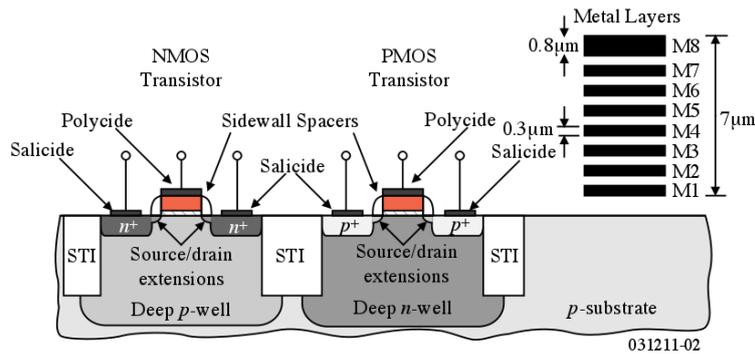


Figure 2.20: Schematic of transistor construction on silicon wafer.

output is directly used. (But when PLLs with analog filters are used, there is inertia, e.g. 1 millisecond). Voltage adjustment also has inertia and there is design and implementation complexity supporting more than a few voltage regions.

	<b>Clock Gating</b>	<b>Supply Gating</b>	<b>DVFS</b>
Control:	automatic	various	software
Granularity:	register / FSM	larger blocks	macroscopic.
Clock Tree:	mostly free runs	turned off	slows down.
Response time:	instant	2 to 3 cycles	instant (or ms if PLL adjusted)
Proportionally vary voltage:	not possible	n/a	yes.

### 2.4.1 Save Power 2: Dynamic Clock Gating

Clock trees consume quite a lot of the power in an ASIC and considerable savings can be made by turning off the clocks to small regions. A region of logic is idle if all of the flip-flops are being loaded with their current contents, either through synchronous clock enables or just through the nature of the design. EDA DESIGNLINE

Instead of using synchronous clock enables, current design practice is to use a clock gating insertion tool that gates the clock instead. One clock control logic gate serves a number of neighbouring flip-flops: state machine or broadside register.

Problem with AND gate: if CEN changes when clock is high: causes a glitch. Problem with OR gate: if CEN changes when clock is low: causes a glitch. Hence, care must be taken not to generate glitches on the clock as it is gated. Transparent latches in the clock enable signal prevent these glitches.

Care needed to match clock skew when crossing to/from non-gated domain: avoid *shoot-through* by building out the non-gated parts as well. Shoot-through occurs when a D-type is supposed to register its current D input value, but this has already changed to its new value before the clock signal arrives.

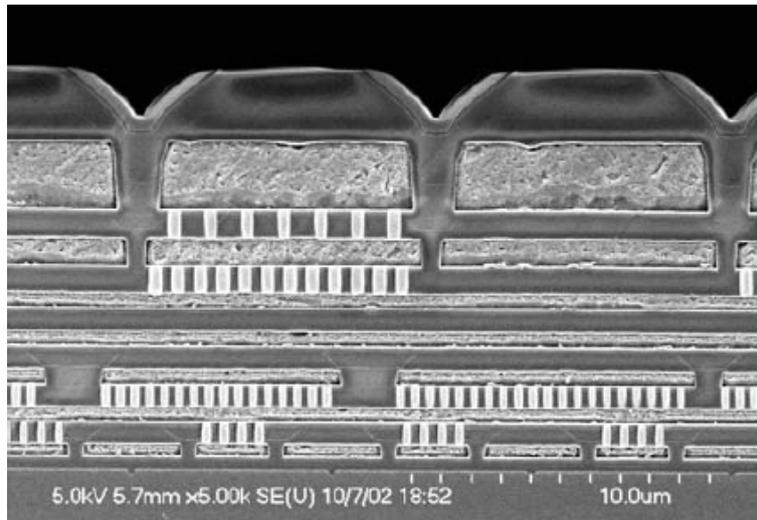


Figure 2.21: Cross section showing stacked wiring metal layers.

The following table shows the VDD power consumption of the E16G301 as a function of frequency and voltage with all 16 cores executing a heavy duty workload. The measurements were taken at room temperature without a heat sink and with 0 m/s airflow. The maximum operating frequency at each voltage level is specified next to the data point in the figure

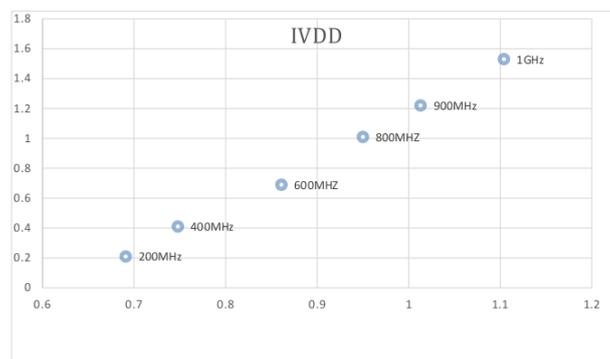


Figure 2.22: Epiphany 16 Core ‘Supercomputer’ Chip DVFS Points (a plot, not a table).

How to generate clock enable conditions ? One could have software control for complete blocks (additional control register flags, as per power gating). But today’s designs automatically detect on a finer-grain basis. Synthesiser tools can automatically insert clock required conditions and insert the additional logic. Automatic tools compute ‘clock needed’ conditions. A clock is ‘needed’ if any register will change on a clock edge.

A lot of clock needed computation can get expensive, resulting in no net saving, but it can be effective if computed once at head of a pipeline.

If not a straightforward pipeline, need to be sure there are no ‘oscillating’ stages that retrigger themselves or an ‘earlier’ stage (add further runtime checks or else statically know their maximum settling time and use a counter). The maximum settling time, if it exists, is computed in terms of clock cycles using static analysis. Beyond the settling time, all registers will be being re-loaded with their current data on each clock cycle.

Beyond just turning off the clock or power to certain regions, we can consider further power saving techniques: dynamic frequency and voltage scaling.



Figure 2.23: Thermal management heat pipes in a modern laptop.

	Clock	Power
On./Off	Clock gating	Power supply gating
Variable	Dynamic frequency scaling (DFS)	Dynamic voltage scaling (DVS)

Figure 2.24: Terminology and Overview of Power Saving Techniques.

### 2.4.2 Save Power 3: Dynamic Supply Gating

Increased tendency towards multi-product platform chips means large functional blocks on silicon may be off for complete product lifetime. The 'dark silicon' future scenario implies all chips must be mostly powered off. Battery powered devices will also use macro-scale block power down (e.g. the audio or video input and output subsystems).

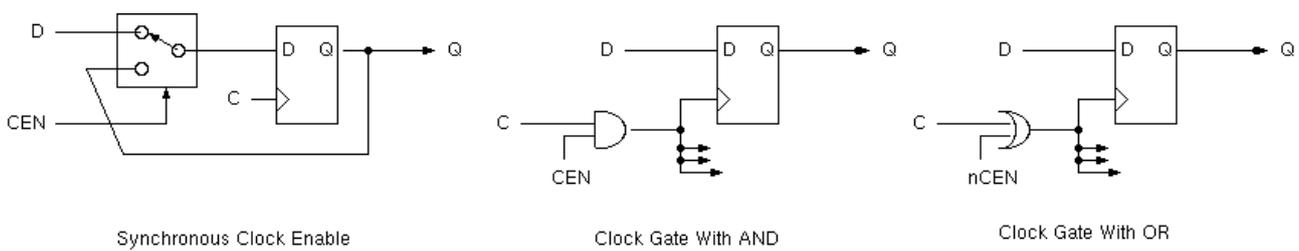


Figure 2.25: Clock enable using multiplexor, AND and OR gate.

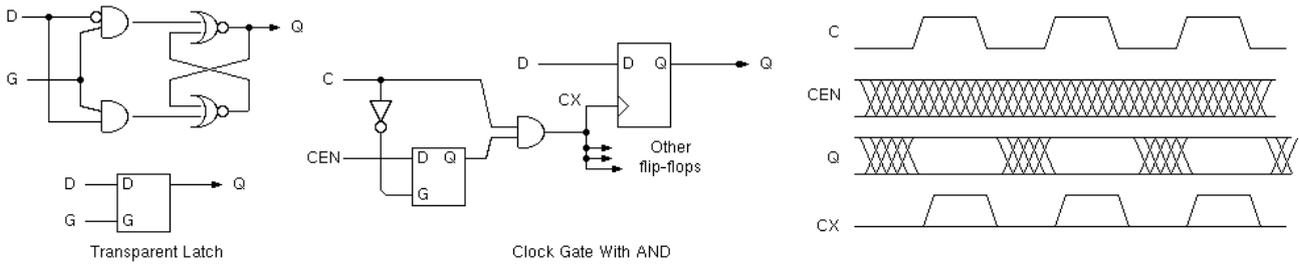


Figure 2.26: Illustrating a transparent latch and its use to suppress clock gating glitches.

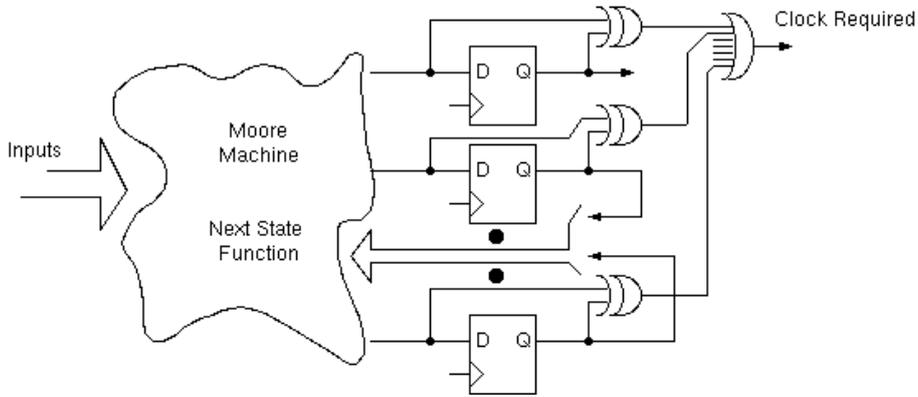
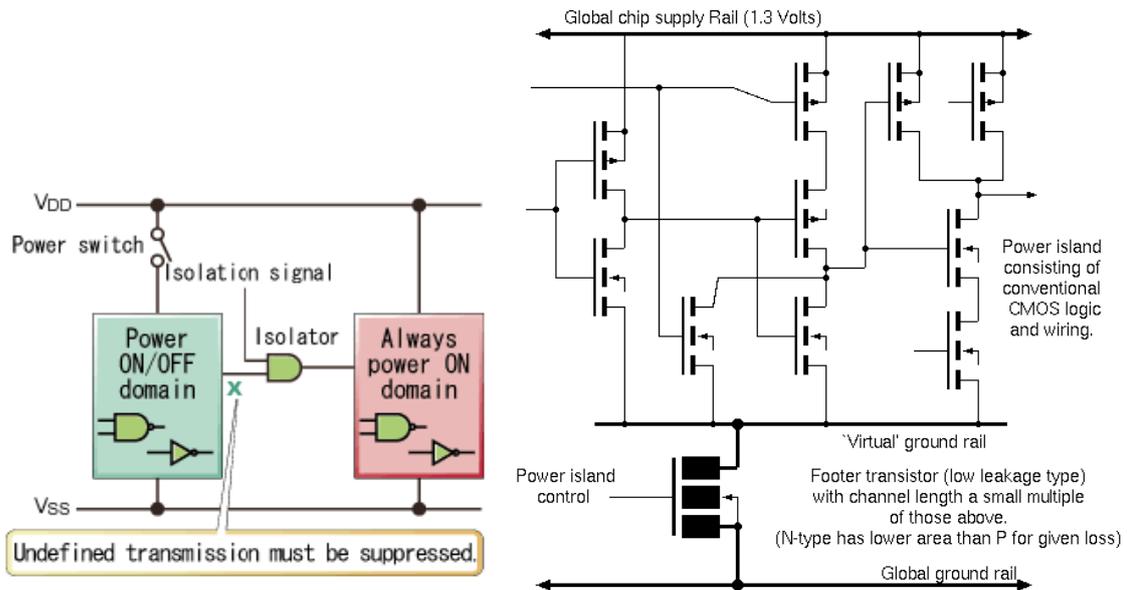


Figure 2.27: Using XOR gates to determine whether a clock edge would have any effect.



Dynamic power gating techniques typically require some sequencing: several clock cycles to power up/down a region and enable/disable isolation gates.

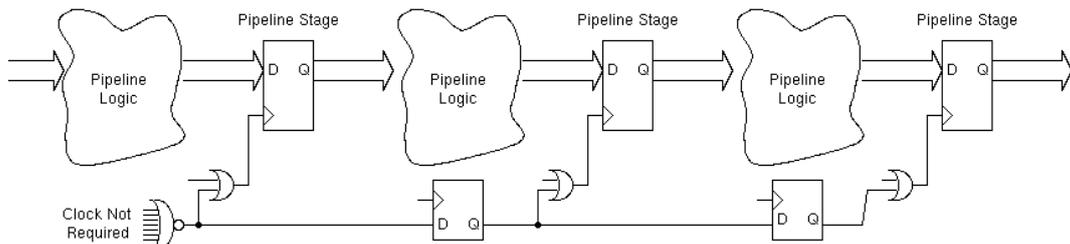


Figure 2.28: Clock needed computations forwarded down a pipeline.

Fujitsu Article: Design of low power consumption LSIs

Previously we looked at dynamic clock gating, but we can also turn off power supply to regions of a chip with fine or coarse grain, creating so-called power islands. We use power gating cells in series with supply rails. These are large, slow, low-leakage transistors. (Best to disconnect the ground supply since an N-channel transistor can be used which has smaller area for same resistance.)

Signal isolation and retention cells (t-latches) on nets that cross in and out of the region are needed. There is no register and RAM data retention in a block while the power is off. This technique is suitable at coarse grain for complete sub-systems of a chip that are not in use on a particular product or for quite a long time, such as a bluetooth transceiver or audio input ADC. It can also be used on a fine grain with automated control similar to clock gating.

However, power gating requires some sequencing to activate the enables to the isolation cells in the correct order and hence several clock cycles or more are needed to power up/down a region. Additionally, gradual turn on over tens of milli-seconds avoids creating noise on the global power rails. Originally, power off/on was controlled by software or top-level input pads to the SoC. Today, dedicated microsequencer hardware might control a hundred power islands within a single subsystem.

A common practice is to power off a whole chip except for a one or two RAMs and register files. This was particularly common before FLASH memory was invented, when a small battery is/was used to retain contents using a lower supply (CMOS RAM data holding voltage). Today, most laptops, tablets and PCs have a second, tiny battery that maintains a small amount of running logic when the main power is off or battery removed. This runs the real-time clock (RTC).

Another technique that saves power is to half-turn-on a power gating transistor and thereby run an island at a lower voltage. This is not as efficient as adjusting standard switched-mode power supplies, since the half-turned on transistor will waste energy itself.

### 2.4.3 DVFS in Low-Leakage Technology

Considering DFVS again, there are (were?) two potential strategies (Aesop: Tortoise v Hare):

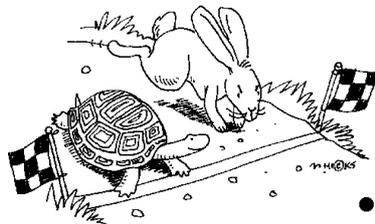


Figure 2.29: Aesop's Fable - Whose approach is better?

1. Compute quickly with halt(s), or
2. Compute more slowly and finish just in time.

To compute quickly and halt we need a higher frequency clock but consume the same number of active cycles. So the work-rate product,  $af$ , unchanged, so no power difference? No. Running the same number of work cycles at a lower frequency requires a lower voltage and hence we save energy according to  $V^2$ .

But: current geometries only have a narrow operating voltage range Too low -> too much leakage. Too high -> tunnelling and wear and heat. So today, we operate at around one volt always and the trade off is just between high and low leakage technology - a static fab-time decision. There are further (unexaminable) prospects on the table, such as dynamic body bias ...

Consider adjusting the clock frequency (while keeping VCC constant for now). What does this achieve? For a fixed task, it will take longer to complete. If the processor is to halt at the end of the task, it will spend less time halted. If the main clock tree (Figure 1.27) keeps going while halted, yet most of the chip uses local clock gating, then we do save some power in that fewer useless clock cycles are executed by the main clock tree. And we save power in logic that is not clock gated.

This sort of frequency scaling can be software controlled: update PLL division ratio. Figure 1.27 illustrates the PLL. The PLL has inertia: e.g. 1 millisecond, but this is similar to the rate at which an operating system services interrupts, and hence the clock frequency to a system can be ramped up as load arrives. Alternately, a digital divider ratio change is instant.

Let's compare with dynamic clock gating: the table shows the main differences, but the most important difference is still to come: we can reduce the supply voltage if we have reduced the clock frequency.

#### 2.4.4 Static and Dynamic Power Tradeoff

**But:** for sub 45nm, voltage scaling is less viable and transistors with much higher static leakage current are commonly used: so can now be better to operate within the voltage/frequency band that works and then power off until next deadline.

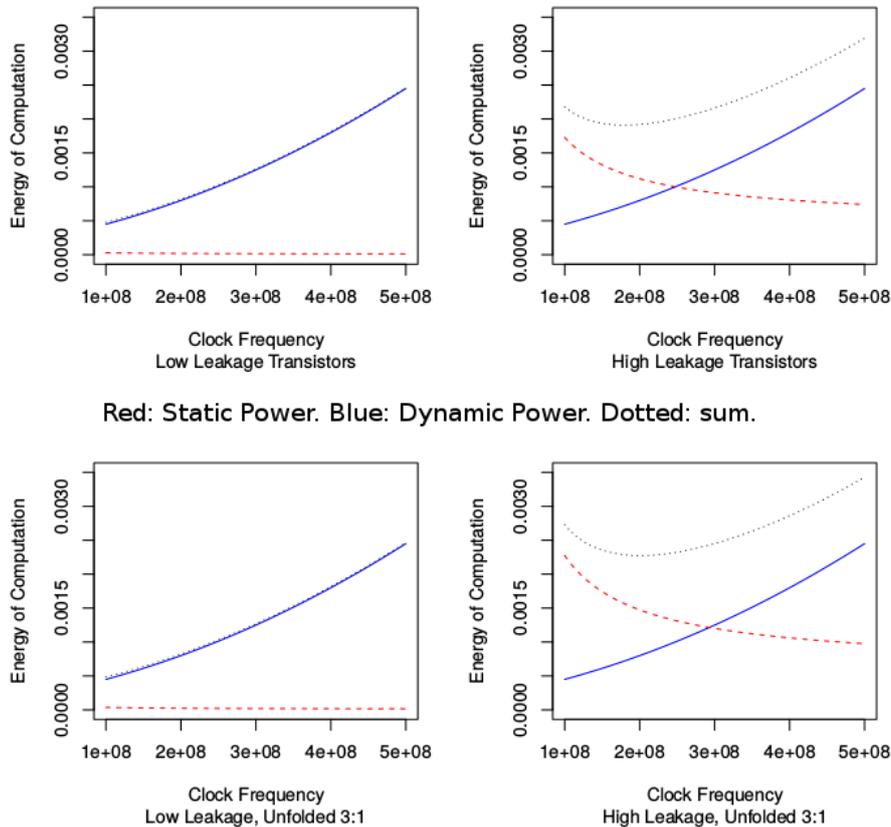


Figure 2.30: Sweetspot shift in DVFS approach for higher leakage on a real-time task.

```

# Trade off of Hare and Tortoise for increasingly leaky technology.
# For a hard-realtime computation we know the number of clock cycles needed but should we do them quickly
# and halt (Archilles) or slowly and finish just in time (Tortoise). For a higher leakage technology,
# switching off as much as possible tends to become preferable to running at low supply voltage.

# Unfold=1 is baseline design. Unfold=3 uses three times more silicon.
static_dynamic_tradeoff <- function(clock_freq, leakage, unfold, xx)
{
  op_count <- 1e5;
  execution_time = op_count / clock_freq / (unfold ~ 0.75); // Model: Pollack-like unfold benefit.
  vdd <- 1 + 0.5 * (clock_freq/100e6); // Model: Higher supply needed for higher clk.
  static_power <- leakage * vdd ~ 0.9 * unfold * 0.4; // Model: Leakage slightly sublinear.
  static_energy <- static_power * execution_time; // Integrate static power
  dynamic_energy <- op_count * vdd ~ 2.0 / 0.5 * 1e-9; // Use CV^2/2 for dynamic
}

```

For the 90nm technology, there was low static leakage and considerable scope for DVFS. With the smaller geometries performance can be traded off for greater leakage. Transistor dopant levels (and hence leakage) can be adjusted in regions or globally. We will want a low leakage, large slow transistor for power gating but may choose higher leakage transistors for logic since these will either be faster or can be run off a lower V<sub>dd</sub>, hence reducing dynamic power.

The simple R plot illustrates the shift in operating frequency sweet spot (minimal total power) with higher leakage transistors. We considered leakage of 0.05 and 0.3 (arbitrary units). With low leakage it is best to compute slowly and finish just in time. With high leakage it is best to compute more quickly and then turn off for longer. A more-detailed analysis would reflect that designers today may have a choice of doping levels and hence transistor leakage at a given voltage and the performance also depends on the doping...

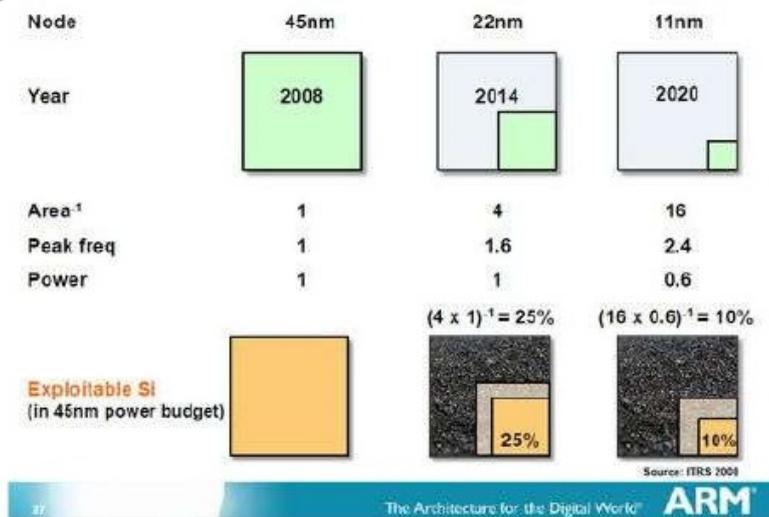
### 2.4.5 Future Trends

Transistors are still being made smaller and smaller.

We have hit the Power Wall resulting in a **Dark Silicon** approach. We can no longer turn on all of the chip and get the heat out cheaply: perhaps one tenth maximum for today's 22 nanometer chips. Even less in the future. Water cooling remains an option to mitigate the Power Wall.

Insights Article

- ⊙ Systems increasingly limited by power consumption, not number of transistors
- ⊙ → **“Dark Silicon”** : Most of the chip will be OFF to meet thermal limits



Slow, bulky power transistors will turn thousands of power islands on and off under automated or manual control.

**Conservation cores:** use of high-level synthesis (HLS) of standard software kernels into application-specific hardware coprocessors and putting them on the chip in case they are needed? Afterall, they have negligible cost if not turned on. Venkatesh

### 2.4.6 Architectural Design: Partition and Exploration

A collection of algorithms and functional requirements must be implemented using one or more pieces of silicon. Each major piece of silicon contains one or more custom or standard microprocessors. Some silicon is custom for a high-volume product, some is shared over several product lines and some is third party or standard parts. The result of the architectural design process is a mapping of the design into physical components. Certain electronic requirements, such as high voltage, microwave radio frequencies and optimum memory bit density are still met with optimised silicon (or GaAs) processes, but today, almost everything is either a 'standard part' or else can be mapped onto a single System-on-Chip. Beyond fundamental silicon capabilities, the design partition process must take into account non-technical aspects, such as stability of requirements, design lifetime, ease of reuse, and other market forces such as whether a third party source is required by the target customers.

When designing a sub-system we must choose what to have as hardware, what to have as software and whether custom or standard processors are needed. When designing the complete SoC we must think about sharing of sub-system load over the chosen processors. Estimates of the instruction fetch and data bandwidth for each processor are needed for deciding how many memories to instantiate and which processors will operate out of which memories. The envisioned system data flow between sub-systems is another important consideration in terms of how busses should be interconnected or whether a NoC is justified. When a SoC is intended for a single target application there is greater certainty about the likely data flow compared with a general purpose chip. Although transistor count does not present a significant design constraint in modern VLSI, hard-wired datapaths are more efficient than switched structures, wiring length and hence energy is minimised when less area is used. A solution providing non-blocking, full crossbar interconnection will generally be over-engineered for all applications.

Energy efficiency is today also often a critical consideration: whether for battery-operation or in a server farm, low-power design principles are applicable and power control mechanisms will affect the design at all levels.

Design Partition: Deciding on the number of processors, number of custom processors, and number of custom hardware blocks. The **system architect** must make these decisions. SystemC helps them rapidly explore various possibilities.

Co-design and co-synthesis: two basic methods (can do different parts of the chip differently):

- **Co-design:** Manual partition between custom hardware and software for various processors,
- **Co-synthesis:** Automatic partitioning: simple 'device drivers' and inter-core message formats are created automatically:

Co-synthesis is still not in mainstream use (2018). Example algorithm: MPEG compression:

- A-to-D capture to framestore,
- Colour space conversion (RGB->YUV),
- DCT transform and variable Q quantisation,
- Motion detection,
- Huffman encoding.

Can any of this be best done on a general purpose (say ARM) core ?

MPEG Encoding 1MPEG algorithm 2

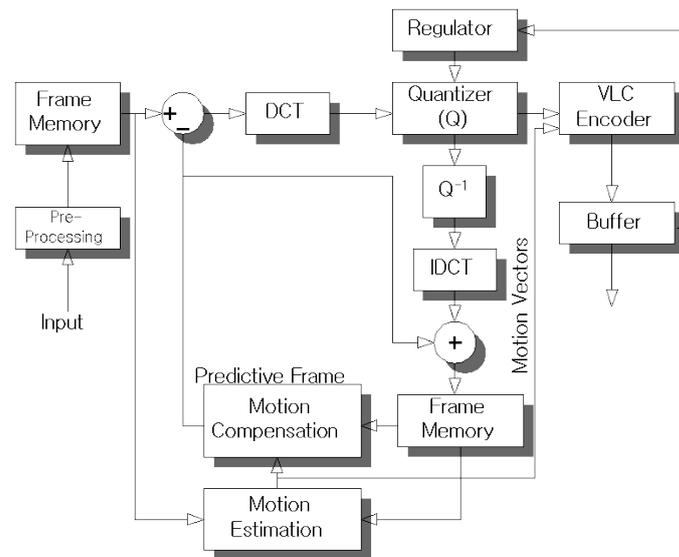


Figure 2.31: Flow chart for MPEG video compression.

### 2.4.7 H/W Design Partition

A number of separate pieces of silicon are combined to form the product. Reasons for H/W design partition:

- Modular Engineering At Large (Revision Control/Lifetime/Sourcing/Reuse),
- Size and Capacity (chips 6-11 mm in size),
- Technology mismatch (Si/GaAs/HV/Analog/Digital/RAM/DRAM/Flash)
- Supply chain: In-house versus Standard Part.
- Isolation of sensitive RF signals,
- Cost: a new chip spin of old IP is still very expensive.

## 2.5 H/W versus S/W Design Partition Principles

Many functions can be realised in software or hardware. Decide what to do in hardware:

- physical I/O (line drivers/transducers/media interfaces),
- highly compute-intensive, **fixed** functions,

what to do on custom processors or with custom instructions/coprocessors on an extensible processor:

- bit-oriented operations,
- highly compute-intensive SIMD,

- other algorithms with custom data paths,
- algorithms that might be altered post tape out.

and what to do in S/W on standard cores:

- highly-complex, non-repetitive functions,
- low-throughput computations of any sort,
- functions that might be altered post tape out,
- generally, as much as possible.

Custom processor synthesis commercial offering: See (link updated) Tensilica Customizable Processor and DSP IP

Example: if we are designing a digital camera, how many processors should it have and can the steadicam and motion estimation processing be done in software? Would a hardware implementation use less silicon and less battery power?

- The functions of a system can be expressed in a programming language or similar form and this can be compiled fully to hardware or left partly as software
- Choosing what to do in hardware and what to do in software is a key decision. Hardware gives speed (throughput) but software supports complexity and flexibility.
- Partitioning of logic over chips or processors is motivated by interconnection bandwidth, raw processing speed, technology and module reuse.

### 2.5.1 An old partitioning example: An external RS-232/POTS Modem.



Figure 2.32: A POTS modem.

Figure 2.33 shows the block diagram of a typical modem circa 1985. The illustrated device is an external modem, meaning that it sits in a box beside the computer and has an RS-232 serial connection to the computer. It also requires its own power supply.

The device contains a few analog components which behave broadly like a standard telephone, but most of it is digital. A relay is used to connect the device to the line and its contacts mirror the 'off-hook' switch which is part of every telephone. It connects a transformer across the line. The relay and transformer provide isolation of the computer ground signal from the line voltages. Similarly the ringing detector often uses an opto-coupler to provide isolation. *Clearly, these analog aspects of the design are particular to a modem and are designed by a telephone expert.*

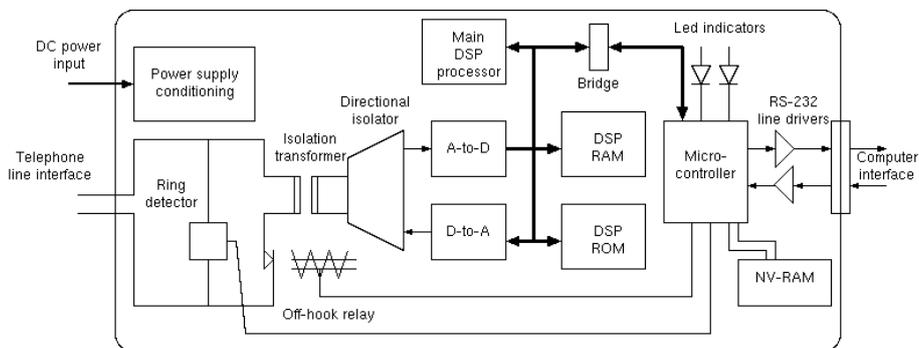


Figure 2.33: Example of a design partition — internal structure of the original modem.



Figure 2.34: PCB of a similar modem - but offchip RAM and ROM for the 80C31 microcontroller.

Modems from the 1960's implemented everything in analog circuitry since microprocessors and DSP were not available. In 1985, two microprocessors were often used.

Note that the non-volatile RAM required (and still does) a special manufacturing processing step and so is not included as a resource on board the microcontroller. Similarly, the RS-232 drivers need to handle voltages of +/- 12 volts and so these cannot be included on chip without increasing the cost of the rest of the chip by using a fabrication process which can handle these voltages. The NV-RAM is used to store the owner's settings, such as whether to answer an incoming call and what baud rate to attempt a first connection, etc..

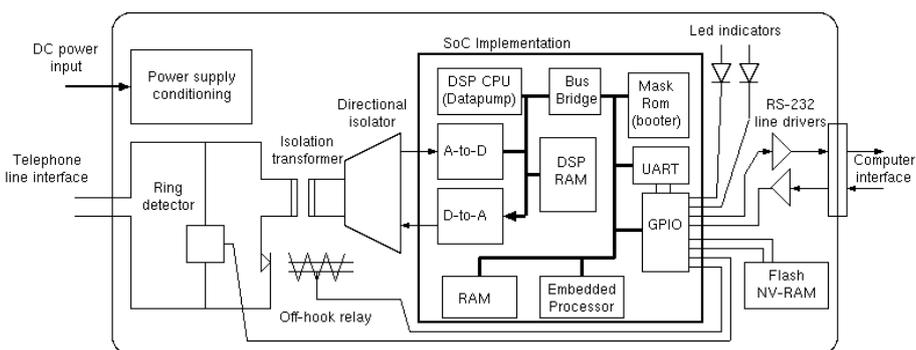


Figure 2.35: Typical structure of the modem product today (using a SoC approach).

A modern implementation would integrate all of the RAM, ROM, ADC and DAC and processors on a single SoC. The RS-232 remains off chip owing to 24 volt and negative supply voltages whereas the SoC itself may be run at 3.3 volts. The NV store is a large capacity Flash ROM device with low-bandwidth serial connection. At system boot, the main code for both processors is copied from the Flash to the two on-chip RAMs by the small, mask-programmed booter. Keeping the firmware in Flash allows the modem to be upgraded to correct bugs or encompass new communications standards.

GPIO is used for all of the digital I/O, with the UART transmit and receive paths being set up as special modes of two of the GPIO connections.

In the next two photos we track the integration of internal modem NICs (network interface cards). These replace the RS-232 serial port with a UART that is connected to the system bus.



Figure 2.36: Modem NIC (network interface card) Pre-SoC integration.



Figure 2.37: PCB of a similar modem - Post SoC technology.

### 2.5.2 Typical Radio/ Wireless Link Structure.

Radio communication above the VHF frequency range (above 150 MHz) uses high-frequency waveforms that cannot be directly processed by A-to-D or D-to-A technology. Hetrodyming is analogue multiplication with a sine wave carrier to perform frequency conversion. This exploits the  $\sin(A)\sin(B) = -\cos(A+B)/2$  part of the standard trig identity for converting upwards and the other half for converting downwards.

The high frequency circuitry is almost always implemented on a separate chip from the digital signal processing (DSP) for the baseband logic. The radio transmitter is typically 50 percent efficient and will use a about 100 mW for most indoor purposes. A cell phone transmitter has a maximum power of 4W which will be used when a long distance from the mast. (Discuss: Having a mast in a school playground means the children are beaming far less radio signal from their own phones into their own heads.) The backlight on a mobile phone LCD may use 300mW (100 LEDs at 30 mW each).

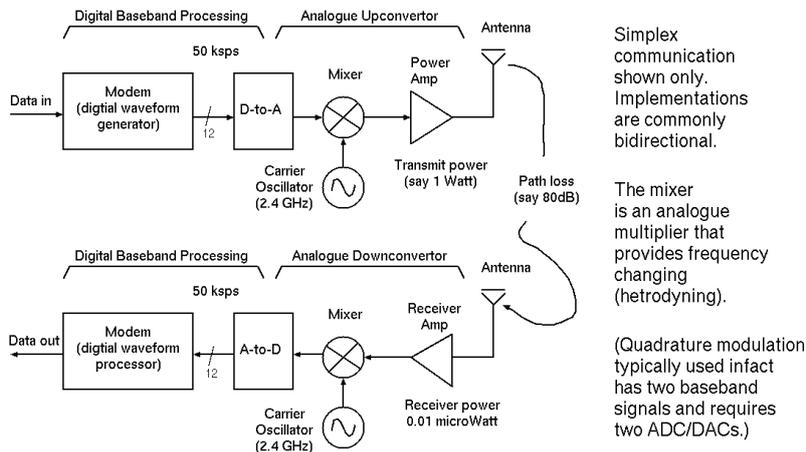


Figure 2.38: Typical structure of modern simplex radio link.

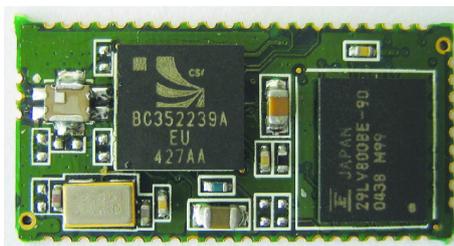


Figure 2.39: Broadcom (Cambridge Silicon Radio) Bluetooth Module circa 2000.

**2.5.3 Partitioning example: A Bluetooth Module.**

An initial implementation of the Bluetooth radio was made of three pieces of silicon bonded onto a small fibre-glass substrate...

An initial implementation of the Bluetooth radio was made of three pieces of silicon bonded onto a small fibre-glass substrate with overall area of 4 square centimetres. The module was partitioned into three pieces of silicon partly because the overall area required would give a low yield, but mainly because the three sections used widely different types of circuit structure.

The analog integrated circuit contained amplifiers, oscillators, filters and mixers that operate in the 2.4 GHz band. This was too fast for CMOS transistors and so bipolar transistors with thin bases were used. The module amplifies the radio signals and converts them using the mixers down to an intermediate frequency of a few MHz that can be processed by the ADC and DAC components on the digital circuit.

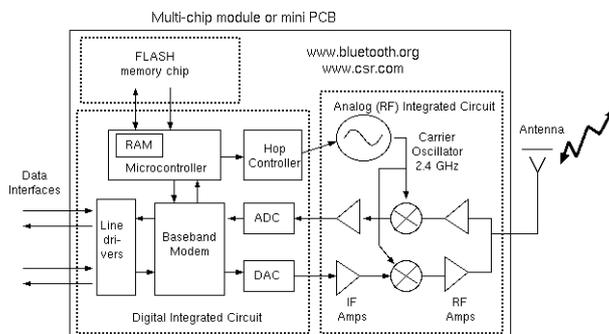


Figure 2.40: Example of a design partition — Block diagram of Bluetooth radio module (circa 2000).

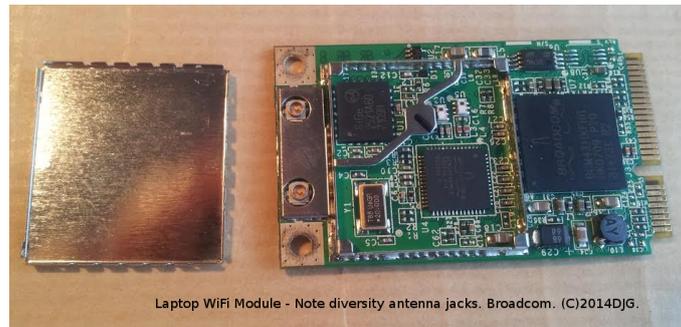


Figure 2.41: WiFi laptop module. Shielding lid, shown left, has been lifted off.

The digital circuit had a small amount of low-frequency analog circuitry in its ADC and DACs and perhaps in its line drivers if these are analog (e.g. HiFi). However, it was mostly digital, with random logic implementations of the modem functions and a microcontroller with local RAM. The local RAM holds a system stack, local variables and temporary buffers for data being sent or received.

The FLASH chip is a standard part, non-volatile memory array that can hold firmware for the microcontroller, parameters for the modem and encryption keys and other end application functions. The flash memory is a standard 29LV800BE (Fujitsu) - 8m (1m X 8/512 K X 16) Bit

Today, the complete Bluetooth module can be implemented on one piece of silicon, but this still presents a major technical challenge owing to the diverse requirements of each of the sub-components.

#### 2.5.4 Structured ASIC

Structured ASIC (masked gate array on smaller die) closes the gap between ASIC and FPGA. The design is prototyped on FPGA and early customer shipments likewise. But FPGA vendor offers a turnkey cost reduction path. For example, two implementations of the same design (Xilinx EasyPath in 2005):

Device	NRE	Unit cost
Spartan-3 FPGA:	0	12 USD
EasyPath E3S1500	75 KUSD	1 USD

Crossover at 6250 units.

Twelve year's later, no public pricing information is available, but the crossover point may be 100x greater. Xilinx EasyPath 2017 Perhaps the cost-reduced part is just a faulty FPGA yield where the faults are known to be irrelevant for the customer's application?

## 2.6 Super FPGAs: Example Xilinx Zynq

We use the terms hard and soft to differentiate between functions that are determined by the fabrication masks and are loaded into the programmable fabric. Although it was common to put so-called **soft** CPU cores in the programmable logic, today's devices have hardened CPUs and many other **hard** IP blocks. Connecting DRAM to FPGAs has become a common requirement in the last decade and hardened DRAM controllers are now common.

The high cost of ASIC masks now makes FPGA suitable for most medium volume production runs (e.g. sub 10,000 units) which includes most recording studio equipment and passenger-in-the-road detection for high-end cars. The dark silicon trend means we can put all IP blocks on one chip provided we leave them mostly turned off.

The Zynq from Xilinx has two ARM cores, all the standard SoC IP blocks and an area of FPGA programmable logic, all on one die. The same DRAM bank is accessible to both the hardened ARM cores and the programmable logic.

Xilinx Zynq-7000 Product Brief (PDF)

Flexible I/O routing means physical pads can be IP block bond outs, GPIOs or FPGA I/O blocks.

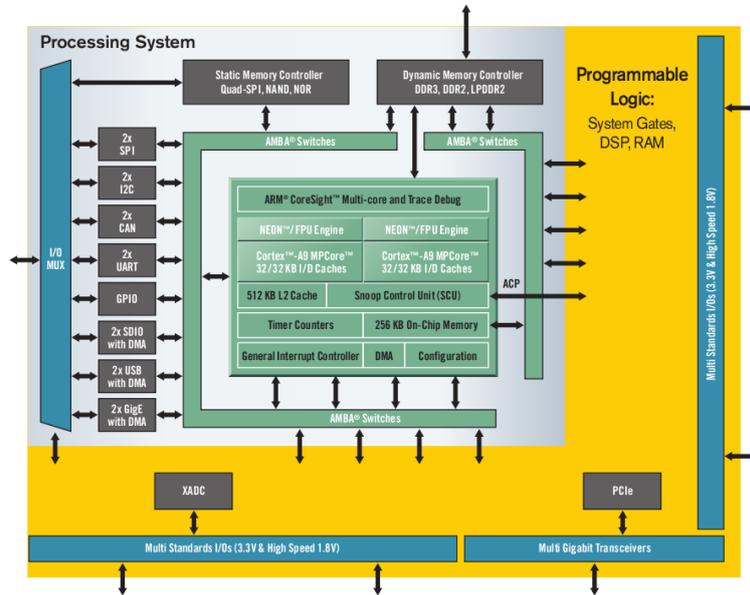


Figure 2.42: Xilinx Zynq 7000 Overview.

Vital statistics for the first Zynq offerings:

Zynq-7000 Product Table (Hardware View)				
Device Name	Z-7010	Z-7020	Z-7030	Z-7045
Part Number	XC7Z010	XC7Z020	XC7Z030	XC7Z045
Processing System (Dual ARM® Cortex™-A9 MPCore™ with NEON™ & Double Precision FPU Cache, Memory Controllers, DMA, Security and Peripherals)	Same processing system for all devices			
Xilinx 7 Series Programmable Logic Equivalent	Artix™-7 FPGA	Artix™-7 FPGA	Kintex™-7 FPGA	Kintex™-7 FPGA
Programmable Logic Cells (Approximate ASIC Gates <sup>(2)</sup> )	28K Logic Cells (~430K)	85K Logic Cells (~1.3M)	125K Logic Cells (~1.9M)	350K Logic Cells (~5.2M)
Logic Cells	28,160	85,120	125,760	349,760
Look-Up Tables LUTs	17,600	53,200	78,600	218,600
Flip Flops	35,200	106,400	157,200	437,200
Extensible Block RAM (# 36 Kb Blocks)	240 KB (60)	560 KB (140)	1,060 KB (265)	2,180KB (545)
Programmable DSP Slices (18x25 MACCs)	80	220	400	900
Peak DSP Performance (Symmetric FIR)	58 GMACS	158 GMACS	480 GMACS	1080 GMACS
PCI Express™ (Root Complex or Endpoint)	—	—	Gen2 x4	Gen2 x8
Agile Mixed Signal (AMS)/XADC	2x 12 bit, 1 MSPS ADCs with up to 17 Differential Inputs			
Security <sup>(1)</sup>	AES and SHA 256b for secure configuration			

Figure 2.43: Xilinx Zynq 7000 FPGA Resources.

Figures for some more-recent devices:

### 2.6.1 Legacy H/W S/W Design Partition

In the past (nineteen-eighties), it was best to use a standard processors as a separate chip. Today, it is no problem to put down one or more 'standard' processors on a SoC. It is also quite easy to design your own, so MIPS, Tensilica, ARM and other CPU core providers have to compete against in-house design teams. For instance, we use the the totally free OR 1000 in the practical materials of this course.

Device Name	VU31P	VU33P	VU35P	VU37P	VU11P	VU13P
System Logic Cells (K)	962	962	1,907	2,852	2,835	3,780
CLB Flip-Flops (K)	879	879	1,743	2,607	2,592	3,456
CLB LUTs (K)	440	440	872	1,304	1,296	1,728
Max. Distributed RAM (Mb)	12.5	12.5	24.6	36.7	36.2	48.3
Total Block RAM (Mb)	23.6	23.6	47.3	70.9	70.9	94.5
UltraRAM (Mb)	90.0	90.0	180.0	270.0	270.0	360.0
HBM DRAM (GB)	4	8	8	8	–	–
HBM AXI Interfaces	32	32	32	32	–	–
Clock Mgmt Tiles (CMTs)	4	4	8	12	12	16
DSP Slices	2,880	2,880	5,952	9,024	9,216	12,288
Peak INT8 DSP (TOP/s)	8.9	8.9	18.6	28.1	28.7	38.3
PCIe® Gen3 x16 / Gen4 x8	4	4	5	6	3	4
CCIX Ports <sup>(1)</sup>	4	4	4	4	–	–
150G Interlaken	0	0	2	4	6	8
100G Ethernet w/ RS-FEC	2	2	5	8	9	12
Max. Single-Ended HP I/Os	208	208	416	624	624	832
GTY 32.75Gb/s Transceivers	32	32	64	96	96	128

Figure 2.44: Parts available from Xilinx in 2018

**20X** more bandwidth than a DDR4 DIMM

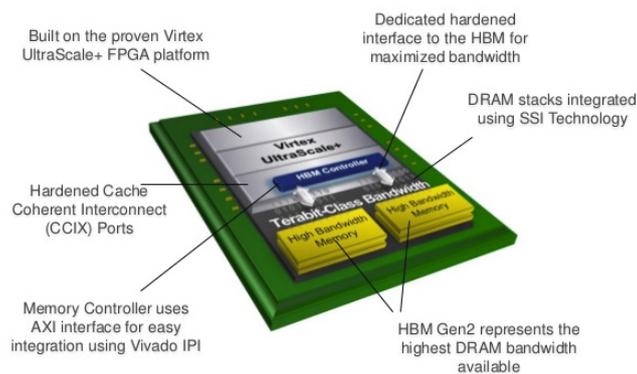


Figure 2.45: FPGA connected closely to several DRAM chips in Multi-chip Module (MCM).

### 2.6.2 An old example example: The Cambridge Fast Ring two chip set.

Two devices were developed for the CFR local-area network (1983), illustrating the almost classical design partition required in high-speed networking. They were never given grander names than the *ECL* chip and the *CMOS* chip. The block diagram for an adaptor card is shown in the Figure 2.47.

The ECL chip clocked at 100 MHz and contained the minimal amount of logic that needed to clock at the full network clock rate. Its functions were:

- implement serial transmission modulator and demodulator,
- convert from 1 bit wide to 8 bits wide and the other way around,
- perform reception byte alignment (when instructed by logic in the CMOS chip).

Other features:

- ECL logic can support analogue line receivers at low additional cost so can receive the incoming signal directly on to the chip.
- ECL logic has high output power if required (1 volt into 25 ohms) and so can drive outgoing twisted pair lines directly.

The CMOS chip clocks at one-eighth the rate and handles the complex logic functions:

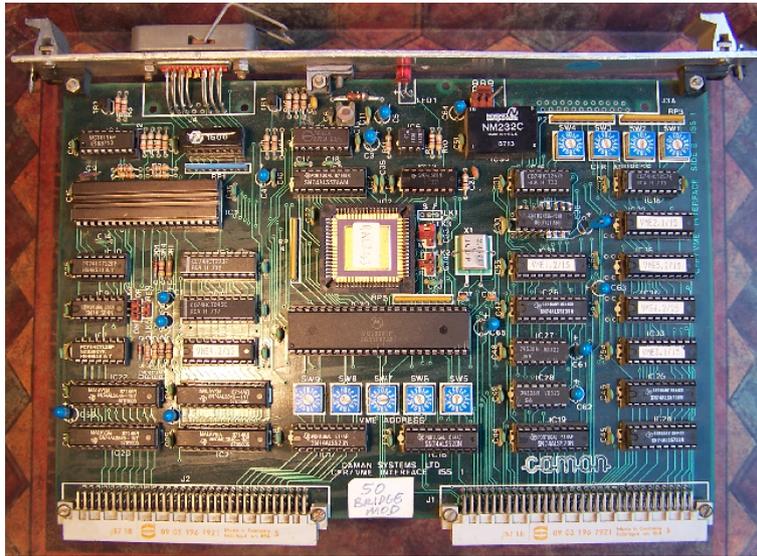


Figure 2.46: The two-chip CFR set using PALs as glue logic for the VME bus.

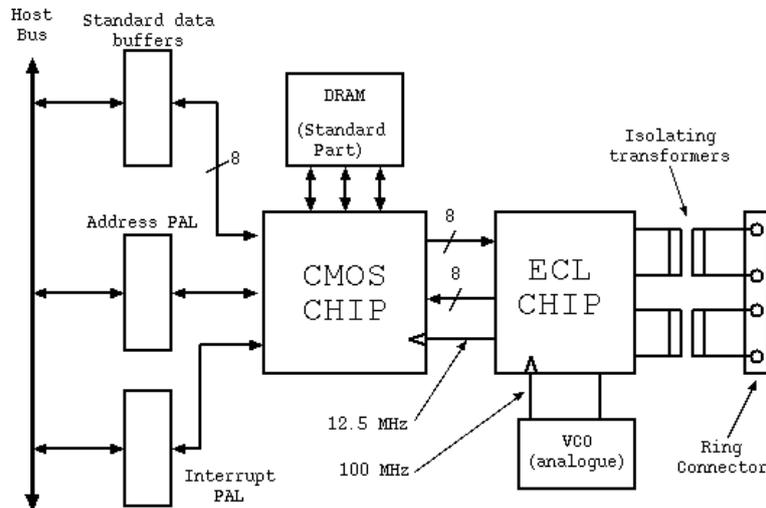


Figure 2.47: Example of a design partition — the adaptor card for the Cambridge Fast Ring.

- CRC generation
- full/empty bit protocol
- minipacket storage in on-chip RAM
- host processor interface
- ring monitoring and maintenance functions.

The ECL chip had at least 50 times the power consumption of the CMOS chip. The CMOS chip had more than 50 times the gates of the ECL chip. Rolling forward to 2010, we might make a similar design partition with a high-performance bipolar subsystem clocking at 4 GHz connected to a CMOS 'baseband' component running where some small parts operating at 500 MHz and the remainder at 250 MHz.

Standard parts were used to augment the CFR set: the DRAM chip incorporates a dense memory array which could not have been achieved for anywhere near the same cost onboard the CMOS chip and the VCO (Voltage Controlled Oscillator) device used for clock recovery was left off the ECL chip since it was a difficult-to-design analogue component where the risk of having it on the chip was not desired.

PALs are used to 'glue' the network interface itself to a particular host system bus. Only the glue logic needs to be redesigned when a new machine is to be fitted with the chipset. PALs have a short design turn-around time since they are field-programmable.

For a larger production run, the PALs would be integrated onto a custom variant of the CMOS chip.

## KG 3 — Masked versus Reconfigurable Technology & Computing

There are different types of silicon chip and different types of manufacturer. Manufacturers can be broadly classified as:

1. So-called 'IDM' or 'vertical market' chip makers such as IBM and Intel that design, manufacture and sell their chips (Integrated Device Manufacturers).
2. Fabless manufacturers such as NVIDIA and Xilinx that design and sell chips but outsource manufacturing to foundry companies.
3. The foundry companies (such as TSMC and UMC) that manufacture chips designed and sold by their customers.

The world's major foundries are SMC and TSMC: Taiwan Semiconductor Manufacturing Company Limited but some verticals also provide fab services, perhaps even to competitors in the application space.

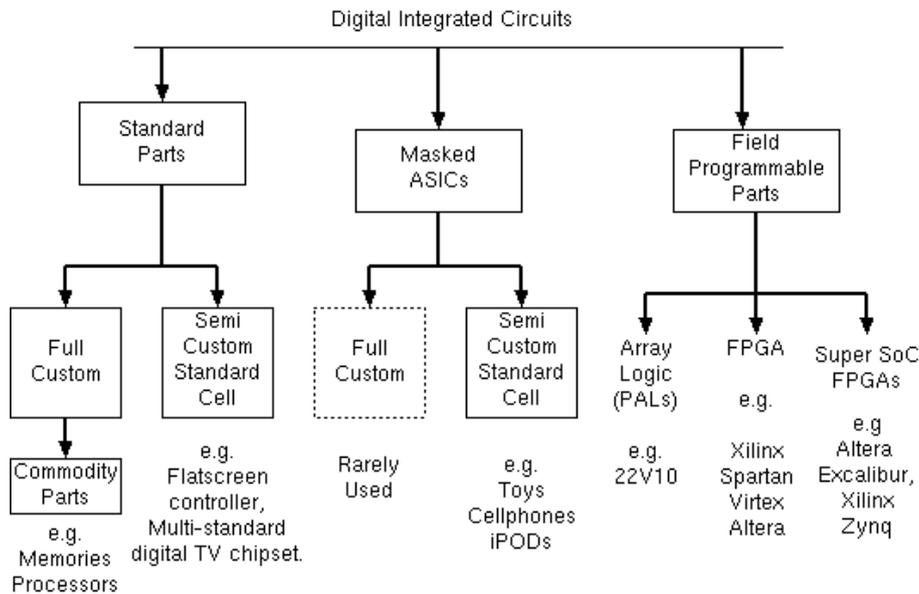


Figure 3.1: A rough taxonomy of digital integrated circuits.

Example Standard Cell Project: 8 Bit Adder 0.5 Micron Cell Library

Figure 3.1 presents a historical taxonomy of chip design approaches. The top-level division is between standard parts, ASICs and field-programmable parts. Where a standard part is not suitable, the choice between full-custom and semi-custom and field-programmable approaches has to be made, depending on performance, production volume and cost requirements.

### Additional notes:

There are deviations from this taxonomy: Complex PLDs cross between PALs and FPGA with low pin-to-pin delay. Structured ASICs were mask-programmed FPGAs popular around 2005. Today (2012-16), super FPGAs such as Zynq are obliterating semi-custom masked ASICs for all but very-high-volume products. When Will FPGAs Kill ASICs?

Chips can be classified by function: Analog, Power, RF, Processors, Memories, Commodity: logic, discrettes, FPGA and CPLD, SoC/ASIC, Other high volume (disk drive, LCD, ... ).

### 3.0.3 Chip Types and Classifications

An SSD drive, such as the one in figure 3.2, originally used only standard parts: a microprocessor/microcontroller with off-chip DRAM and twenty FLASH chips. But the huge volume of production means that a custom microcontroller was soon preferred, but the FLASH chips themselves remain standard parts.



Figure 3.2: FLASH is now replacing spinning media in all but archival applications.

### 3.0.4 Standard Parts

A **standard part** is essentially any chip that a chip manufacturer is prepared to sell to someone else along with a datasheet and EDA (electronic design automation) models. The design may actually previously have been an ASIC for a specific customer that is now on general release. Many standard parts are general-purpose logic, memory and microprocessor devices. These are frequently full-custom designs designed in-house by the chip manufacturer to make the most of in-house fabrication line, perhaps using optimisations not made available to others who use the line as a foundry. Other standard parts include graphics controllers, digital TV chipsets, GPS receivers and miscellaneous useful chips needed in high volume.

### 3.0.5 Masked ASICs.

A masked ASIC (application-specific integrated circuit) is a device manufactured for a customer involving a set of masks where at least some of the masks are used only for that device. These devices include full-custom and semi-custom ASICs and masked ROMs.

A full-custom chip (or part of a chip) has had detailed, manual design effort expended on its circuits and the position of each transistor and section of interconnect. This allows an optimum of speed and density and power consumption.

**Full-custom design is used for devices which will be produced in very large quantities:** e.g. millions of parts where the design cost is justified. Full-custom design is also used when required for performance reasons. Microprocessors, memories and digital signal processing devices are primary users of full-custom design.

In semi-custom design, each cell has a fixed design and is repeated each time it is used, both within a chip and across many devices which have used the library. This simplifies design, but drive power of the cell is not optimised for each instance.

**Semi-custom is achieved using a library of logic cells and is used for general-purpose VLSI design.**

### 3.0.6 ASIC - Application-Specific Integrated Circuit

The cost of developing an ASIC has to be compared with the cost of using an existing part or an FPGA. The existing part may not perform the required function exactly, requiring either a design specification change, or some additional *glue logic* to adapt the part to the application.

More than one ASIC may be needed under any of the following conditions:

- application-specific functions are physically distant,
- application-specific functions require different technologies,
- application-specific functions are just too big for one ASIC,
- it is desired to split the cost and risk or reuse part of the system later on.

Factors to consider on a per-chip basis:

- power consumption limitation (powers above 5 Watts need special attention),
- die size limitation (above 11 mm on a side might escalate cost per mm<sup>2</sup>),
- speed of operation — clock frequencies above 1 GHz raise issues,
- special considerations :
  - special static or dynamic RAM needs
  - analogue parts - what is compromised if these are integrated onto the ASIC ?
  - high power/voltage output capabilities for load control: e.g. motors.
- availability of a developed module for future reuse.

### 3.0.7 Semi-custom (cell-based) Design Approach

Standard cell designs use a set of well-proven logic cells on the chip, much in the way that previous generations of standard logic have been used as board-level products, such as Texas Instruments' System 74.

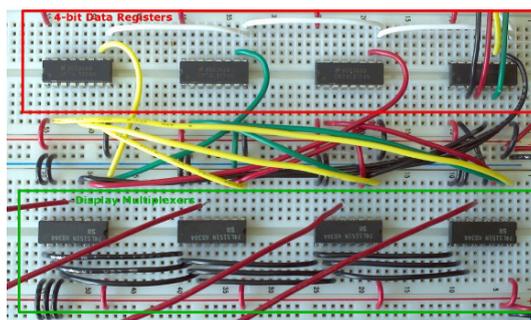


Figure 3.3: Discrete Logic Gates: Semi-custom design puts millions of them all on one die.

A library of standard logic functions is provided. Cells are placed on the chip and wired up by the user, in the same way that chips are placed on the PCB.

- Standard Cell - free placement and free routing of nets,
- Gate Array - fixed placement, masked or electrical programmable wiring.

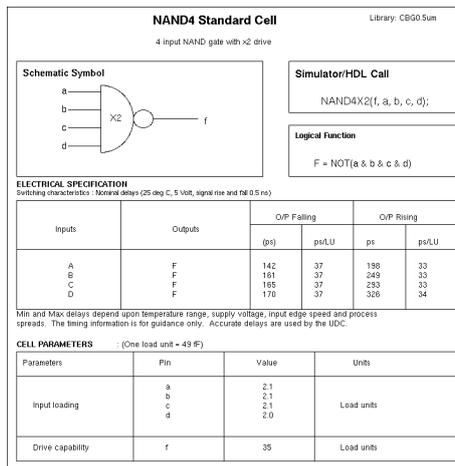


Figure 3.4: Typical cell data sheet from a standard cell library.

Figure 3.4 shows a cell from the data book for a standard cell library. This device has twice the ‘normal’ drive power, which indicates one of the compromises implicit in standard cell over full-custom, which is that the size (driving power) of transistors used in a cell is not tuned on a per-instance basis.

Mask-programmed gate array has been mostly replaced with the field-programmed FPGA except for analog/mixed-signal niches, such the example from TRIAD

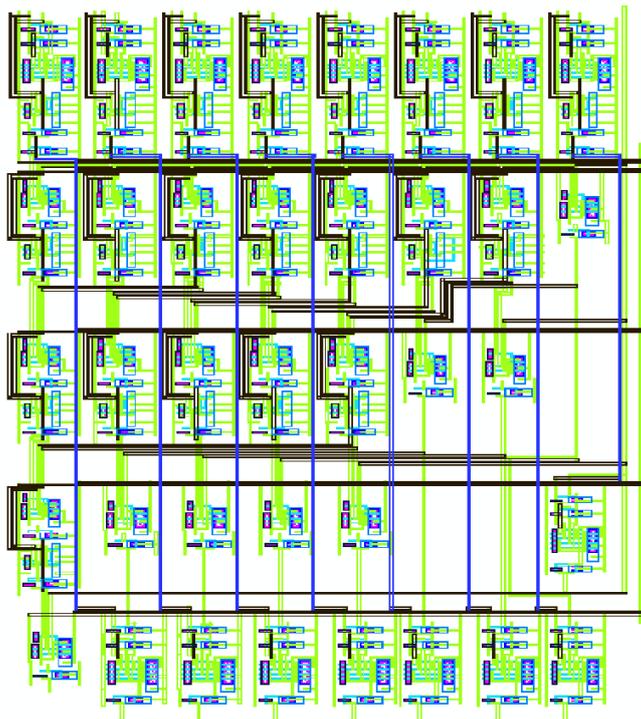


Figure 3.5: Standard cell layout for a Kogge-Stone adder. Taken from a student project (PDF on course web site).

In standard cell designs, cells from the library can freely be placed anywhere on the silicon and the number of I/O pads and the size of the die can be freely chosen. Clearly this requires that all of the masks used for a chip are unique to that design and cannot be used again. Mask making is one of the largest costs in chip design. (When) Will FPGAs Kill ASICs?

### 3.0.8 Cell Library Tour

In the lecture we will have a look at (some of) the following documents:

Standard Cell Student Project: Kogge Stone Adder

Cell libraries in the public domain: 0.5 Micron Cell Library Another 90nm Cell Library Some others: VLSI TECH Things to note: there's a good variety of basic gates, including quite a few multi-level gates, such as AND-OR gate. There's also I/O pads, flip-flops and special function cells. Many gates are available with various output powers. For each gate there are comprehensive figures that enable one to predict its delay and energy use, taking into account its track loading, how many other gates it is feeding and the current supply voltage.

### 3.0.9 ASIC Costs: RE and NRE.

The cost of a chip splits into two parts: non-recurring engineering (NRE) and per-device cost.

Item	Cost (KUSD)	Total (KUSD)
NRE: 6 months : 10 H/W Engineers	250 pa	1250
NRE: 12 months : 20 S/W Engineers	200 pa	4000
NRE: 1 Mask set (45nm)	3000	3000
RE: An 8 inch wafer	5	5n
<b>TOTAL</b>	5	8125 + 5n

For small quantities: share cost of masks with other designs e.g. the MOSIS programme offers multiproject wafer (MPW).

### 3.0.10 Chip cost versus area

The per-device cost is influenced by the yield — the fraction of working dice. The fraction of wafers where at least some of the die work is the 'wafer yield'. Historically yields have been low, but was typically close to 100 percent for mature 90 nm fabrication processes, but has again be a problem with smaller geometries in recent years.

The fraction of die which work on a wafer (often simply the 'yield') depends on wafer impurity density and die size. Die yield goes down with chip area. The fraction of devices which pass wafer probe (i.e. before the wafer is diced) and fail post packaging tests is very low. However, full testing of analog sections or other lengthy operations are typically skipped at the wafer probe stage.

Assume processed wafer sale price might be 5000 dollars: A six inch diameter wafer has area  $(3.14r^2) = 18000 \text{ mm}^2$ . A chip has area  $A$ , which can be anything between 2 to  $200 \text{ mm}^2$  (including scoring lines). Dies per wafer is  $18000/A$ .

Probability of working = wafer yield  $\times$  die yield (assume wafer yield is 1.0 or else included in the wafer cost).

Assume 99.5 percent of square millimetres are defect free. Die yield is then

$$P(\text{All } A \text{ squares work}) = 0.995^A$$

cost of working dice is

$$\frac{5000}{\frac{18000}{A} 0.995^A} \text{ dollars each.}$$

Cost of a working die given a six inch wafer with a processing cost of 5000 dollars and a probability of a square millimetre being defect free of 99.55 percent.

Area	Wafer dies	Working dies	Cost per working die
2	9000	8910	0.56
3	6000	5910	0.85
4	4500	4411	1.13
6	3000	2911	1.72
9	2000	1912	2.62
13	1385	1297	3.85
19	947	861	5.81
28	643	559	8.95
42	429	347	14.40
63	286	208	24.00
94	191	120	41.83
141	128	63	79.41
211	85	30	168.78
316	57	12	427.85
474	38	4	1416.89

For a chip with regular structure, such as a memory or an FPGA, additional hidden capacity can be deployed by burning fusible straps (aka links) during wafer probe test. This increases yield despite the larger area shipped in defect-free dies. AMD marketed a range of 3-core CPUs where the 4th, present on the die, had been strapped off.

### 3.0.11 Gate Arrays and Field-Programmable Logic.

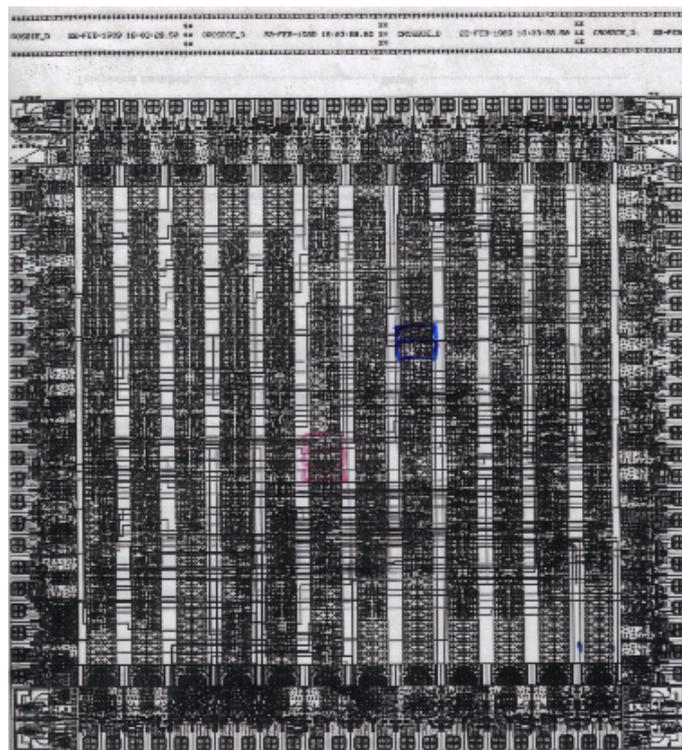


Figure 3.6: Mask for a Mask-Programmed Gate Array: (Greaves 1995, ECL for Ring Network)

Figure 3.6 reveals the regular layout of a masked gate array showing bond pads around the edge and wasted silicon area (white patches). A gate array comes in standard die sizes containing a fixed layout of configurable cells. Historically, there were two main forms of gate array:

- Mask Programmable,

- Field Programmable (FPGA).

In gate array designs, the silicon vendor offers a range of chip sizes. Each size of chip has a fixed layout and the location of each transistor, resistor and IO pad is common to every design that uses that size. Gate arrays are configured for a particular design by wiring up the transistors, gates and other components in the desired way. Many cells will be unused. For mask-programmed devices, the wiring up was done with the top two or three layers of metal wiring. Therefore only two or three custom masks were needed to be made to make a new design. In FPGAs the programming is purely electronic (RAM cells control pass transistors).

The disadvantage of gate arrays is their intrinsic low density of active silicon. This arises from rounding up to the next available die size and the area overhead to support programming. The programming area overhead is especially severe for the FPGA.

### 3.0.12 FPGA - Field Programmable Gate Array

About 25 to 40 percent of chip sale revenue now comes from field-programmable logic devices. These are chips that can be programmed electronically on the user's site to provide the desired function. PALs and CPLDs are forms of programmable logic that are fast and small. But the most important form today is the FPGA.

Recall the Xilinx/Altera FPGA parts used in the Part IB E+A classes. Field-programmable devices may be volatile (need programming every time after power up), reprogrammable or one-time programmable. This depends on how the programming information is stored inside the devices, which can be in RAM cells or in any of the ways used for ROM, such as electrostatic charge storage (e.g. FLASH).

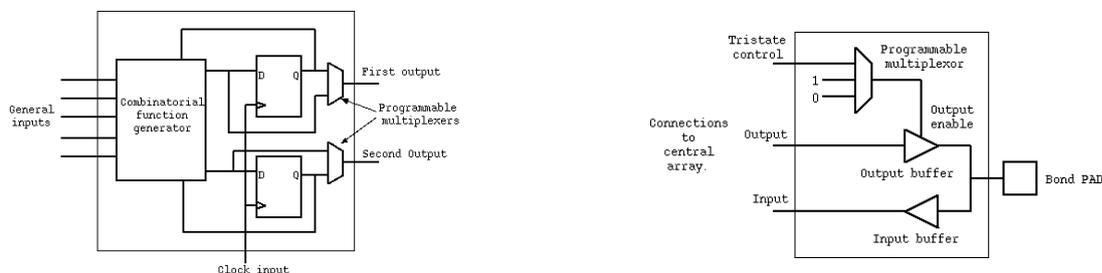
Except for niche applications (such as GaAs instead of Si), FPGAs are now always used instead of masked gate arrays and are starting to kill ASICs (see link above).

Example: The part Ib practical classes use FPGAs from Altera: ECAD and Architecture Practical Classes

Summary data for a Virtex 5 Xilinx FPGA:

Part number	XC5VLX110T-2FFG1136C
Vendor	Xilinx Inc
Category	Integrated Circuits (ICs)
Number of Gates	110000
Number of I / O	640
Number of Logic Blocks/Elements	8640
Package / Case	1136-FCBGA
Operating Temperature	0C 85C
Voltage - Supply	1.14 V 3.45 V

Circa 2009, 65 nm technology, 6-input LUT, 64 bit D/P RAMs. Today Xilinx has the Virtex 7 series that includes the Zync SoC (of which more later) wikipedia: Virtex FPGA



An FPGA (field-programmable gate array) consists of an array of configurable logic blocks (CLBs), as shown in Figure 3.7. Not shown is that the device also contains a good deal of hidden logic used just for programming it. Some pins are also dedicated to programming. Such FPGA devices have been popular since about 1990.

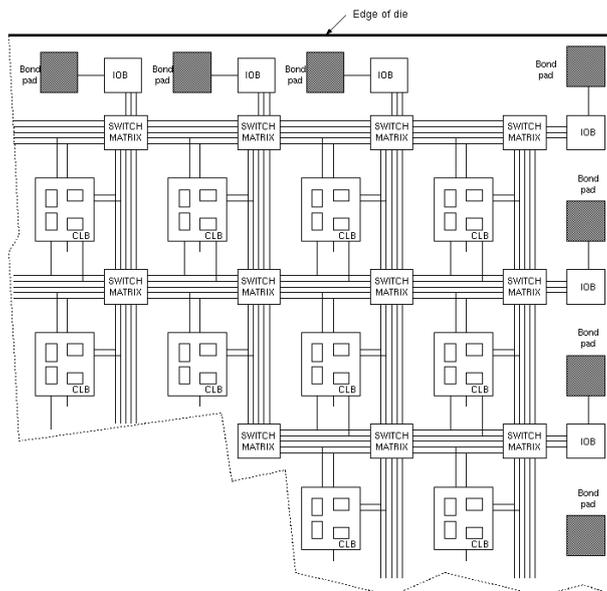


Figure 3.7: Field-programmable gate array structure, showing I/O blocks around the edge, interconnection matrix blocks and configurable logic blocks. In recent parts, the regular structure is broken up by custom blocks, including RAMs and multiplier (aka DSP) blocks.

Each CLB (configurable logic block) or slice typically contains two or four flip-flops, and has a few (five shown) general purpose inputs, some special purpose inputs (only a clock is shown) and two outputs. The illustrated CLB is of the look-up table type, where the logic inputs index a small section of pre-configured RAM memory that implements the desired logic function. For five inputs and one output, a 32 by 1 SRAM is needed. Some FPGA families now give the designer write access to this SRAM, thereby greatly increasing the amount of storage available to the designer. However, it is still an expensive way to buy memory.

FPGAs also soon started to contain RAM blocks (called block RAM or BRAM) and multiplier blocks called DSP (digital signal processing) blocks. The BRAM and DSP blocks are automatically deployed by the design tools by matching specific patterns in the user's RTL when coded appropriately. Today's FPGAs also contain many other 'hard-IP' blocks, such as PCIe, Ethernet and USB controllers that need to be manually instantiated as structural components in the RTL.

FPGAs tend to be slow, achieving perhaps one third of the clock frequency of a masked ASIC, owing to larger die area and because the signals pass through the programmable wiring junctions.

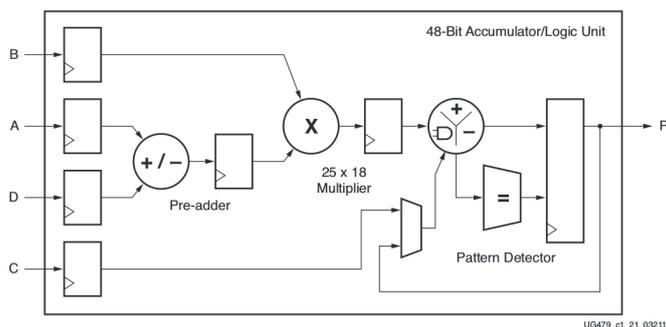


Figure 3.8: So-called DSP block in Xilinx Virtex 7 ((C) Xilinx Inc).

The Xilinx DSP block mostly contains a multiplier that delivers a 48 bit result and an adder for accumulating results where the output from one block has a high-performance programmable connection to a neighbour. The multiplier operands are two's complement, 25 and 18 bit operands. *Exercise: How many DSP blocks are needed for a 32x32 multiplier? What is its latency? What differences does it make if only 32 bits of the result are*

needed?

### 3.0.13 Circuit Switching

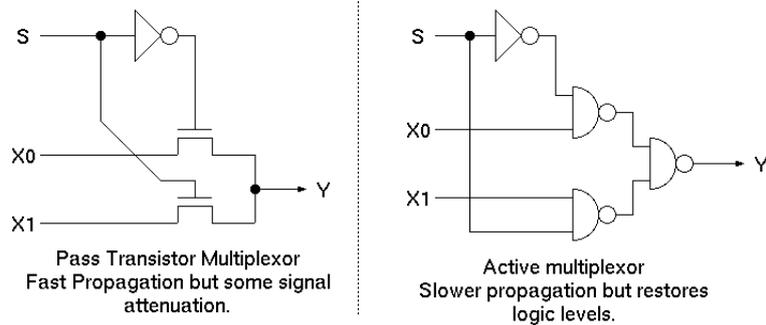


Figure 3.9: Pass transistor multiplexor compared with active path multiplexor.

Much of the area of an FPGA is taken up with programmable wiring. The **pass transistor** is a cheap (in area terms) and efficient (in delay terms) form of programmable wiring, but it does not amplify the signal.

FPGAs dominate the recent history of reconfigurable computing but are fine-grain owing to heritage in hardware circuits. There is an argument for having wider busses as the lowest programmable feature, which amortises the programming overhead to some extent, and yields the CGRA - coarse grain reconfigurable array.

## KG 4 — RTL, Interfaces, Pipelining and Hazards

An RTL description is at the so-called **Register Transfer Level**

A hardware design consists of a number of modules interconnected by wires known as ‘nets’ (short for networks). The interconnections between modules are typically structured as mating interfaces. An interface nominally consists of a number of terminals but these ultimately may have no physical manifestation owing to inter-module logic optimisation. In a modern design flow, the protocol at an interface is ideally specified once in a master file that is imported for the synthesis of each module that sports it. But a lot of low-level manual entry of RTL design is still (2017) used.

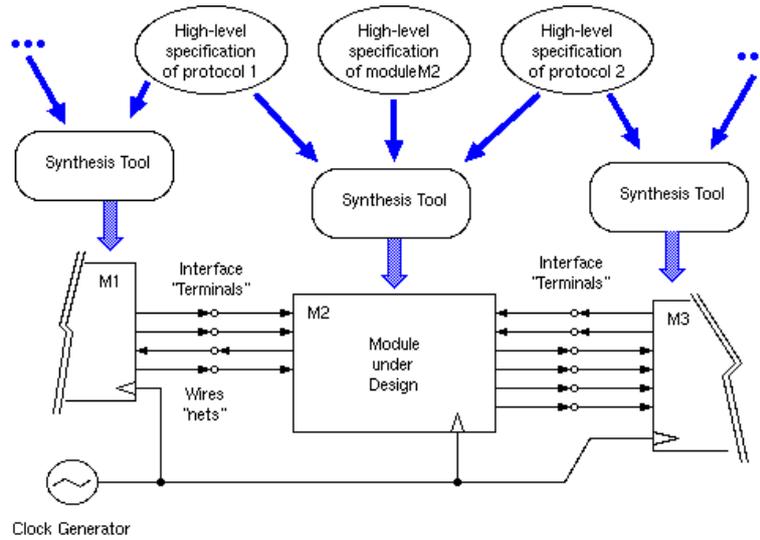


Figure 4.1: Generic (net-level) Module Interconnection Using Protocols and Interfaces.

A clock domain is a set of modules and a clock generator. Within a synchronous clock domain all flip-flops have their clocks commoned.

### 4.1 Protocol and Interface

At the electrical/net level, a **port** consists of an **interface** and a **protocol**. The interface is the set of pins or wires that connect the components. The protocol defines the rules for changing the logic levels and the meaning of the associated data. For example, an asynchronous interface might be defined in RTL as:

```
Transmit view of interface:    Receive view of interface:    // This is a four-phase asynchronous interface
output [7:0] data;            input [7:0] data;             // where the idle state has strobe and ack
output strobe;                input strobe;                // deasserted (low) and data is valid while
input ack;                    output ack;                  // the strobe signal is asserted (high).
```

Ports commonly implement **flow-control** by handshaking. Data is only transferred when both the sender and receiver are happy to proceed.

A port generally has an **idle** state which it returns to between each transaction. Sometimes the start of one transaction is immediately after the end of the previous, so the transition through the idle state is only nominal. Sometimes the beginning of one transaction is temporarily overlaid with the end of a previous, so the transition through idle state has no specific duration.

**Additional notes:**

There are four conceivable clock strategies for an interface:

Left Side	Right Side	Name
1. Clocked	Clocked	Synchronous (such as Xilinx LocalLink)
2. Clocked	Different clock	Clock Domain Crossing (see later)
3. Clocked	Asynchronous	Hybrid.
3. Asynchronous	Clocked	Hybrid (swapped).
4. Asynchronous	Asynchronous	Asynchronous (such a four-phase parallel port)

Any interface that does not have any or the same clock on both sides is asynchronous. (For example, the clock-domain crossing bridge discussed elsewhere in these notes.)

**4.1.1 Transactional Handshaking**

The mainstream RTL languages, Verilog and VHDL, do not provide synthesis of handshake circuits (but this is one of the main innovations in some more recent HLDs such as Bluespec). We'll use the word **transactional** for protocol+interface combinations that support flow-control. If synthesis tools are allowed to adjust the delay through components, all interfaces between components must be transactional and the tools must understand the protocol semantic.

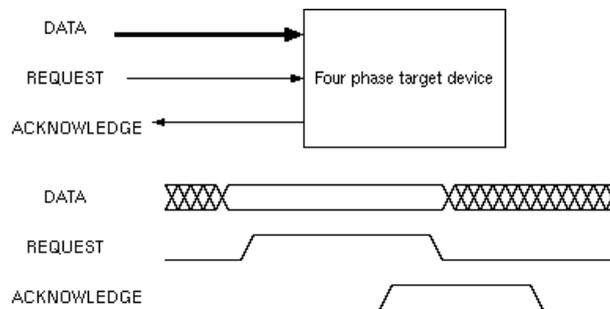


Figure 4.2: Timing diagram for an asynchronous, four-phase handshake.

Here are two imperative (behavioural) methods (non-RTL) that embody the protocol for Figure 4.2:

```
//Output transactor:
putbyte(char d)
{
    wait_until(!ack); // spin till last complete.
    data = d;
    settle(); // delay longer than longest data delay
    req = 1;
    wait_until(ack);
    req = 0;
}
```

```
//Input transactor:
char getbyte()
{
    wait_until(req);
    char r = data;
    ack = 1;
    wait_until(!req);
    ack = 0;
    return r;
}
```

Code like this is used to perform programmed I/O (PIO) on GPIO pins (see later). It can also be used as an ESL transactor (see later). It's also sufficient to act as a formal specification of the protocol.

**4.1.2 Transactional Handshaking in RTL (Synchronous Example)**

The four-phase handshake just described is suitable for asynchronous interfaces. It does not refer to a clock. A very common paradigm for synchronous flow control of a uni-directional bus is to have a handshake net in each direction with bus data being qualified as valid on **any positive clock edge where both handshake nets are asserted**. The nets are typically called 'valid' and 'ready' with valid being in the direction from initiator to target (or from the data source for simplex protocols). This paradigm forms the essence of the LocalLink

protocol from Xilinx and AXI-streaming protocols defined by ARM.

Like the four-phase handshake, LocalLink has contra-flowing The interface nets for an eight-bit transmitting interface are:

```

input clk;
output [7:0] xxx_data; // The data word - here just a byte but any size is possible.
output xxx_sof_n;      // Start of frame
output xxx_eof_n;      // End of frame
output xxx_src_rdy_n;  // This is the 'valid' signal
input  xxx_dst_rdy_n;  // reverse direction ready signal.

```

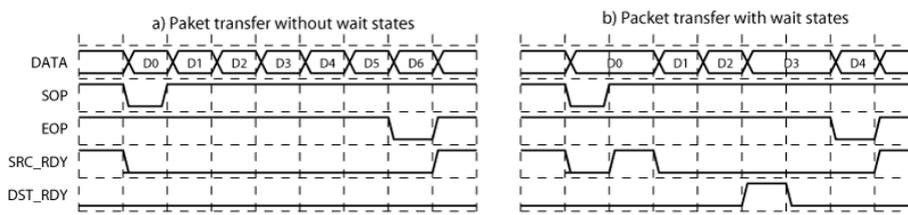


Figure 4.3: Timing diagram for the synchronous LocalLink protocol.

On top of the word-level handshake signals, LocalLink defines start-of-frame and end-of-frame signals. These provide a packet delineation layer on top of the word-level protocol. Note: all control signals are active low (denoted with the underscore n RTL suffix) in LocalLink.

A feature of this protocol is that both sides can freely assert and deassert their handshake net at will. A common implementation error is to wait for the other side to assert its handshake net first. This is a bad idea, since if both sides wait we have deadlock.

**Additional notes:**

Here is a data source in Verilog RTL for LocalLink that generates a stream of packets containing arbitrary data with arbitrary gaps.

```

module LocalLinkSrc(
    input reset,
    input clk,
    output [7:0] src_data,
    output src_sof_n,
    output src_eof_n,
    output src_src_rdy_n,
    input src_dst_rdy_n);

// The source generates 'random' data using a pseudo random sequence generator (prbs).
// The source also makes gaps in its data using bit[9] of the generator.
reg [14:0] prbs;
reg started;
assign src_data = (!src_src_rdy_n) ? prbs[7:0] : 0;
assign src_src_rdy_n = !(prbs[9]);

// The end of packet is arbitrarily generated when bits 14:12 have a particular value.
assign src_eof_n = !(src_src_rdy_n && prbs[14:12]==2);

// A start of frame must be flagged during the first new word after the previous frame has ended.
assign src_sof_n = !(src_src_rdy_n && !started);

always @(posedge clk) begin
    started <= (reset) ? 0 : (!src_eof_n) ? 0 : (!src_sof_n) ? 1 : started;
    prbs <= (reset) ? 100 : (src_dst_rdy_n) ? prbs : (prbs << 1) | (prbs[14] != prbs[13]);
end
endmodule

```

And here is a corresponding data sink:

```

module LocalLinkSink(
    input reset,
    input clk,
    input [7:0] sink_data,
    input sink_sof_n,
    input sink_eof_n,
    output sink_src_rdy_n,
    input sink_dst_rdy_n);

// The sink also maintains a prbs to make it go busy or not on an arbitrary basis.
reg [14:0] prbs;
assign sink_dst_rdy_n = prbs[0];

always @(posedge clk) begin
    if (!sink_dst_rdy_n && !sink_src_rdy_n) $display(
        "%m LocalLinkSink sof_n=%d eof_n=%d data=0x%h", sink_sof_n, sink_eof_n, sink_data);
    // Put a blank line between packets on the console.
    if (!sink_dst_rdy_n && !sink_src_rdy_n && !sink_eof_n) $display("\n\n");
    prbs <= (reset) ? 200 : (prbs << 1) | (prbs[14] != prbs[13]);
end

endmodule // LocalLinkSink

```

**Additional notes:**

And here is a testbench that wires them together:

```

module SIMSYS();

    reg reset;
    reg clk;
    wire [7:0] data;
    wire sof_n;
    wire eof_n;
    wire ack_n;
    wire req_n;

    // Instance of the src
    LocalLinkSrc src (.reset(reset),
                    .clk(clk),
                    .src_data(data),
                    .src_sof_n(sof_n),
                    .src_eof_n(eof_n),
                    .src_src_rdy_n(req_n),
                    .src_dst_rdy_n(ack_n));

    // Instance of the sink
    LocalLinkSink sink (.reset(reset),
                      .clk(clk),
                      .sink_data(data),
                      .sink_sof_n(sof_n),
                      .sink_eof_n(eof_n),
                      .sink_src_rdy_n(req_n),
                      .sink_dst_rdy_n(ack_n)
                      );

    initial begin clk =0; forever #50 clk = !clk; end
    initial begin reset = 1; #130 reset=0; end

endmodule // SIMSYS

```

## 4.2 RTL: Register Transfer Language

Everybody attending this course is expected to have previously studied RTL coding or at least taught themselves the basics before the course starts.

The Computer Laboratory has an online Verilog course you can follow: Cambridge SystemVerilog Tutor. Please note that this now covers ‘System Verilog’ whereas most of my examples are in plain old Verilog. There are a few, unimportant, syntax differences.

RTL is compiled to logic gate instances in a target library using a process called **Logic Synthesis**. RTL is also simulatable pre and post synthesis.

### 4.2.1 RTL Summary View of Variant Forms.

For the sake of this course, Verilog and VHDL are completely equivalent as register transfer languages (RTLs). Both support simulation and synthesis with nearly-identical paradigms. Of course, each has its proponent’s.

Synthesisable Verilog constructs fall into these classes:

- **1. Structural RTL** enables an hierarchic component tree to be instantiated and supports wiring (a netlist) between components.
- **2. Lists of pure (unordered) register transfers** where the r.h.s. expressions describe potentially complex logic using a rich set of integer operators, including all those found in software languages such as C++ and Java. There is one list per synchronous clock domain. A list without a clock domain is for combinational logic (continuous assignments).



```

wire dout[39:0];
reg[3:0] values[0:4] = {5, 6, 7, 8, 15};

generate
  genvar i;
  for (i=0; i < 5; i++) begin
    MUT mut[i] (
      .out(dout[i*8+7:i*8]),
      .value_in(values[i]),
      .clk(clk),
    );
  end
endgenerate
    
```

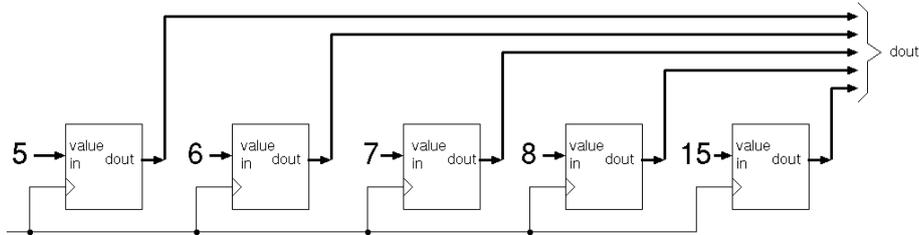


Figure 4.5: Example Generate Statement in RTL.

Heirarchic Netlist	Equivalent Flattened Netlist
<pre> module MOD1(output b, input a);   wire c;    INV inv1(c, a);   MODX modx1(b, c); endmodule  module MOD2(output q, input s, input r);   wire c;    INV inv2(c, s);   MODY mody1(q, c, r); endmodule  module MODTOP(output rr, input aa, input bb);   wire l, m;    MOD1 m(l, aa);   MOD1 n(m, bb);   MOD2 o(rr, l, m); endmodule                 </pre>	<pre> module MODTOP (output rr, input aa, input bb);   wire l, m;   wire m_c, n_c, o_c;    INV m_inv1(m_c, aa);   INV n_inv1(n_c, bb);   INV o_inv2(o_c, l);   MODX m_modx1(m_c, l);   MODX n_modx1(n_c, m);   MODY o_mody1(rr, o_c, m);  endmodule                 </pre>
	<p style="text-align: center;"> </p> <p style="text-align: center;">             For many designs the flattened netlist is often bigger than the hierarchic netlist owing to multiple instances of the same component. Here it was smaller.         </p>

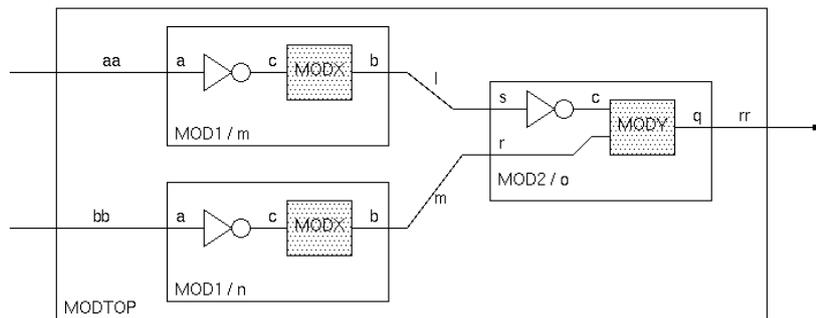
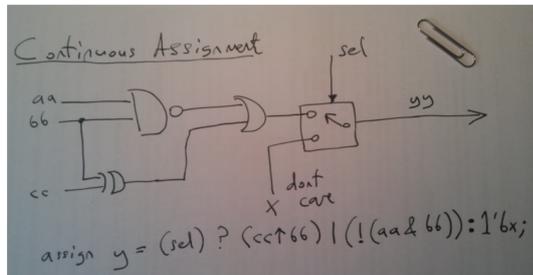


Figure 4.6: Example RTL fragment, before and after flattening.

Figure 4.6 shows structural RTL before and after flattening as well as a circuit diagram showing the component boundaries.

**2a/3: Continuous Assignment:** an item from a pure RT list without a clock domain.



```
// Continuous assignments define combinational logic circuit:
assign a = (g) ? 33 : b * c;
assign b = d + e;
```

- Order of continuous assignments is un-important,
- Tools often insist that continuous logic is loop-free, otherwise: intentional or un-intentional level-sensitive latches are formed (e.g. RS latch),
- Right-hand side's may range over rich operators (e.g. mux ?: and multiply \*),
- Bit inserts to vectors are allowed on left-hand sides (but not combinational array writes).

```
assign d[31:1] = e[30:0];
assign d[0] = 0;
```

### 2b/3: Pure RTL: unordered synchronous register transfers.

Two coding styles (it does not matter whether these transfers are each in their own always statement or share over whole clock domain):

```
always @(posedge clk) a <= b ? c + d;
always @(posedge clk) b <= c - d;
always @(posedge clk) c <= 22-c;
```

```
// or
always @(posedge clk) begin
    a <= b ? c + d;
    b <= c - d;
    c <= 22-c;
end
```

In System Verilog we would use `always_ff` in the above cases.

Typical example (illustrating pure RT forms):

```
module CTR16(
    input mainclk,
    input din,
    output o);

    reg [3:0] count, oldcount;

    always @(posedge mainclk) begin
        count <= count + 1;
        if (din) oldcount <= count; // Is 'if' pure ?
    end

    // Note ^ is exclusive-or operator
    assign o = count[3] ^ count[1];

endmodule
```

Registers are assigned in clock domains (one shown called 'mainclk'). Each register is assigned in exactly one clock domain. RTL synthesis does not generate special hardware for clock domain crossing (described later).

In a stricter form of this pure RTL, we cannot use 'if', so when we want a register to sometime retain its current value we must assign this explicitly, leading to forms like this:

```
oldcount <= (din) ? count : oldcount;
```

**3/3: Behavioural RTL:** a thread encounters order-sensitive statements.

In ‘behavioural’ expression, a thread, as found in imperative languages such as C and Java, assigns to variables, makes reference to variables already updated and can re-assign new values.

For example, the following behavioural code (inside an always block)

```
if (k) foo = y;
bar = !foo;
```

can be compiled down to the following, unordered ‘pure RTL’:

```
foo <= (k) ? y : foo;
bar <= !(k) ? y : foo;
```

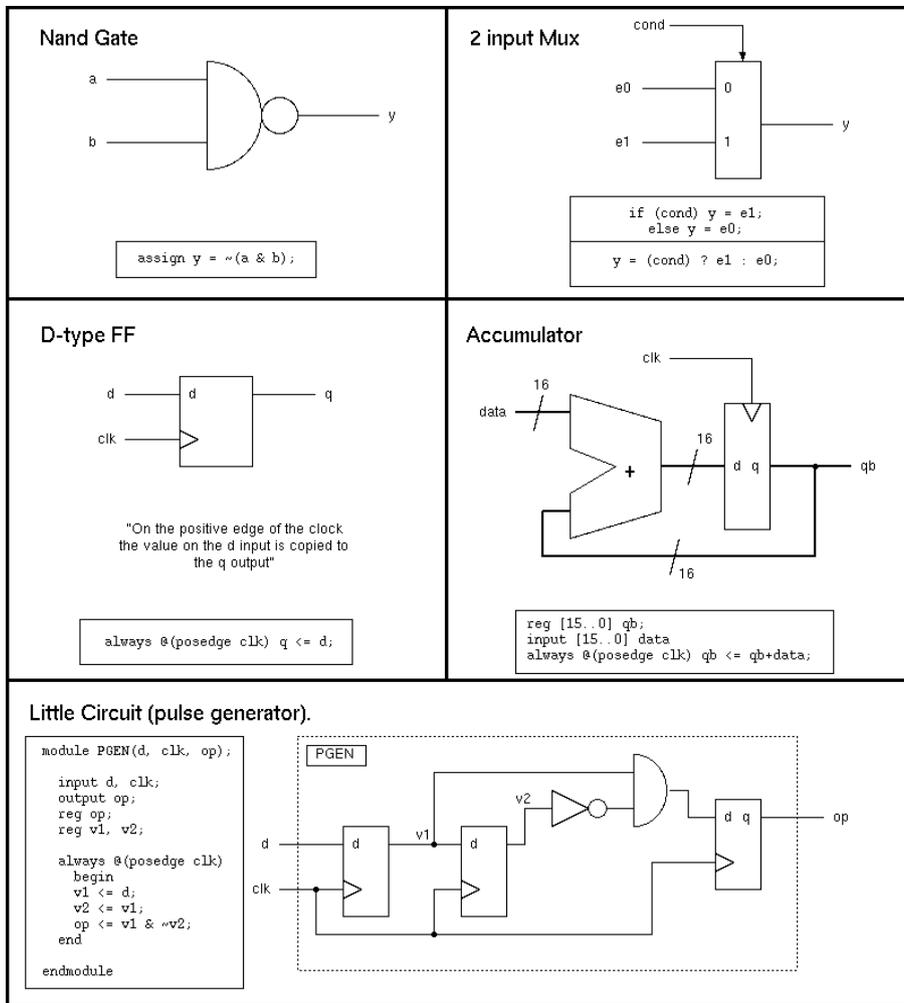


Figure 4.7: Elementary Synthesisable Verilog Constructs

Figure 4.7 shows synthesisable Verilog fragments as well as the circuits typically generated. The ‘little circuit’ uses old-style syntax for input and output designations but this is still valid today.

The RTL languages (Verilog and VHDL) are used both for simulation and synthesis. Any RTL can be simulated but only a subset is standardised as ‘synthesisable’ (although synthesis tools can generally handle a slightly larger synthesisable subset).

Simulation uses a top-level test bench module with no inputs.

Synthesis runs are made using points lower in the hierarchy as roots. We should certainly leave out the test-bench wrapper when synthesising and we typically want to synthesise each major component separately.

## 4.2.2 Synthesisable RTL

### Additional notes:

Abstract syntax for a synthesisable RTL (Verilog/VHDL) without provision for delays:

Expressions:

```
datatype ex_t = // Expressions:
  Num of int // integer constants
| Net of string // net names
| Not of ex_t // !x - logical not
| Neg of ex_t // ~x - one's complement
| Query of ex_t * ex_t * ex_t // g?t:f - conditional expression
| Diadic of diop_t * ex_t * ex_t // a+b - diadic operators + - * / << >>
| Subscript of ex_t * ex_t // a[b] - array subscription, bit selection.
```

Imperative commands (might also include a ‘case’ statement) but no loops.

```
datatype cmd_t = // Commands:
  Assign of ex_t * ex_t // a = e; a[x]=e; - assignments
| If1 of ex_t * cmd_t // if (e) c; - one-handed IF
| If2 of ex_t * cmd_t * cmd_t // if (e) c; else c - two-handed IF
| Block of cmd_t list // begin c; c; .. end - block
```

Our top level will be an unordered list of the following sentences:

```
datatype s_t = // Top-level forms:
  Sequential of edge_t * ex_t * cmd_t // always @(posedge e) c;
| Combinational of ex_t * ex_t // assign e1 = e2;
```

The abstract syntax tree for synthesisable RTL supports a rich set of expression operators but just the assignment and branching commands (no loops). (Loops in synthesisable VHDL and Verilog are restricted to the structural generation statements, mentioned above, that are fully unwound by the compiler front end and so have no data-dependent exit conditions).

An example of RTL synthesis:

Results in structural RTL netlist:

Example input:

```
module TC(clk, cen);
  input clk, cen;
  reg [1:0] count;
  always @(posedge clk) if (cen) count<=count+1;
endmodule
```

```
module TC(clk, cen);
  wire u10022, u10021, u10020, u10019;
  wire [1:0] count;
  input cen; input clk;
  CVINV i10021(u10021, count[0]);
  CVMUX2 i10022(u10022, cen, u10021, count[0]);
  CVDFDFF u10023(count[0], u10022, clk, 1'b1, 1'b0, 1'b0);
  CVXOR2 i10019(u10019, count[0], count[1]);
  CVMUX2 i10020(u10020, cen, u10019, count[1]);
  CVDFDFF u10024(count[1], u10020, clk, 1'b1, 1'b0, 1'b0);
endmodule
```

Here the behavioural input was converted, by a **Logic Synthesis Compiler**, also known as an **RTL Compiler**, to an implementation technology that included inverters, multiplexors, D-type flip-flops and XOR gates. For each gate, the output is the first-listed terminal.

**Verilog RTL Synthesis Algorithm: 3-Step Recipe:**

1. First we remove all of the blocking assignment statements to obtain a 'pure' RTL form. For each register we need exactly one assignment (that becomes one hardware circuit for its input) regardless of however many times it is assigned, so we need to build a multiplexor expression that ranges over all its sources and is controlled by the conditions that make the assignment occur.

For example:

```
if (a) b = c;
d = b + e;
if (q) d = 22;
```

is converted to

```
b <= (a) ? c : b;
d <= q ? 22 : ((a) ? c : b) + e;
```

2. For each register that is more than one bit wide we generate separate assignments for each bit. This is colloquially known as 'bit blasting'. This stage removes arithmetic operators and leaves only boolean operators. For example, if  $v$  is three bits wide and  $a$  is two bits wide:

```
v <= (a) ? 0 : (v>>1)
```

converted to

```
v[0] <= (a[0] | a[1]) ? 0 : v[1];
v[1] <= (a[0] | a[1]) ? 0 : v[2];
v[2] <= 0;
```

3. Build a gate-level netlist using components from the selected library of gates. (Similar to a software compiler when it matches operations needed against instruction set.) Sub-expressions are generally reused, rather than rebuilding complete trees. Clearly, logic minimization (Karnaugh maps and Espresso) and multi-level logic techniques (e.g. ripple carry versus fast carry) as well as testability requirements affect the chosen circuit structure. Gate Building, ML fragment

When generating gates a target technology cell library needs to be read in by the logic synthesiser. Likewise, when generating FPGA logic, the details of the CLBs and limitations of the programmable wiring need to be known by the logic synthesiser.

#### Further detail on selected constructs:

##### Additional notes:

1. How can we make a simple adder ?

The following ML fragment will make a ripple carry adder from lsb-first lists of nets:

```
fun add c (nil, nil) = [c]
| add c (a::at, b::bt) =
  let val s = gen_xor(a, b)
      val c1 = gen_and(a, b)
      val c2 = gen_and(s, c)
      in (gen_xor(s, c))::(add (gen_or(c2, c1)) (at, bt))
  end
```

2. Can general division be bit-blasted ? Yes, and for some constants it is quite simple.

For instance, division by a constant value of 8 needs no gates - you just need wiring! For dynamic shifts make a **barrel shifter** using a succession of broadside multiplexors, each operated by a different bit of the shifting expression. See link Barrel Shifter, ML fragment.

3. Can we do better for constant divisors? To divide by a constant 10 you can use that 8/10 is 0.11001100 recurring, so if  $n$  and  $q$  are 32 bit unsigned registers, the following computes  $n/10$ :

```
q = (n >> 1) + (n >> 2);
q += (q >> 4);
q += (q >> 8);
q += (q >> 16);
return q>>3;
```

### 4.2.3 Arrays and RAM Inference in RTL

RTL languages support bits, bit vectors (words) and arrays of bit vectors (RAMs). Arrays in the RTL can be synthesised to structural instances of RAM memories or else to register files made of flip flops. Certain patterns of array use are defined to trigger **RAM inference**, where a RAM is instantiated in the net list. RAM inference is

supported in FPGA logic synthesis tools. ASIC synthesis tools require the use of the alternative, which is for the RTL to contain explicit structural instances.

```
reg [31:0] myram [32767:0]; // 32K words of 32 bits each.
// To execute the following in one clock cycle needs two RAM ports
always @(posedge clk) myram[a] <= myram[b] + 2;
```

Even when RAM inference is available, it is sometimes easiest to write a leaf module that behaves like the RAM and then structurally instantiate that in the main RTL. The RAM inference will then just act inside the leaf module and edits to the main RTL cannot violate the pattern that triggers the inference procedure.

The pattern needed to trigger RAM inference is nothing more than an RTL model of the real RAM. This example is for a dual-ported (one read, one write) SRAM. SRAM is synchronous RAM with a read latency. Here the latency is one cycle.

```
module R1W1RAM(din, waddr, clk, wen, raddr, dout); // This is both a behavioural model
input clk, wen; // of the SRAM and a pattern that
input [14:0] waddr, raddr; // should trigger RAM inference in
input [31:0] din; // FPGA tools.
output [31:0] dout;

reg [31:0] myram [32767:0]; // 32K words of 32 bits each.
always @(posedge clk) begin
dout <= myram[raddr]; // Data out is registered once without otherwise being
if (wen) myram[waddr] <= din; // looked at. Write data in is synchronous with the write
end // address.
endmodule
```

The behavioural model will be replaced with a RAM macrocell in the silicon implementation. Each port of a RAM has an address input. The two essential rules for RAM inference are that

1. there is one expression that is clearly recognisable as the address for each port, and
2. the data read out is registered by the required number of pipeline broadside registers to match the latency of the target technology without any use (peeking) of the data in that pipeline.

Similar rules facilitate automated deployment of other structural resources (or FUs as we shall call them later). One example is the clock-enable flip flop (as per clock gating) and another is multiplier inference. The FPGA's DSP unit, which is essentially a pipelined multiplier, will be deployed where the tools can make sufficient structural matches

#### 4.2.4 Behavioural - 'Non-Synthesisable' RTL

Not all RTL is officially synthesisable, as defined by language standards. However, commercial tools tend to support larger subsets than officially standardised.

RTL with event control in the body of a thread defines a state machine. This is compilable by some tools. This state machine requires a program counter (PC) register at runtime (implied):

```
input clk, din;
output reg [3:0] q; // Four bits of state are define here.

always begin
q <= 1;
@(posedge clk) q <= 2;
if (din) @(posedge clk) q <= 3;
q <= 4;
end
```

How much additional state in the form of PC bits are needed? Is conditional event control synthesisable? Does the output 'q' ever take on the value 4?

As a second non-synthesisable example, consider the dual-edge-triggered flip-flop in Figure 4.8.

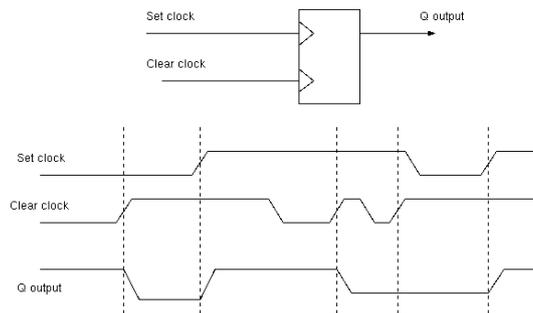


Figure 4.8: Schematic symbol and timing diagram for an edge-triggered RS flop.

```
reg q;
input set, clear;

always @(posedge set) q = 1;
always @(posedge clear) q = 0;
```

Here a variable is updated by more than one thread. This component is used mainly in specialist phase-locked loops. It can be modelled in Verilog, but is **not** supported for Verilog synthesis. A real implementation typically uses 8 or 12 NAND gates in a relatively complex arrangement. We do not expect general-purpose logic synthesis tools to create such circuits: they were hand-crafted by experts of previous decades.

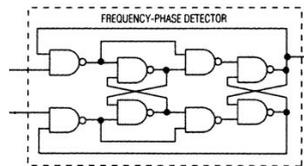


Figure 4.9: Hand-crafted circuit for the edge-triggered RS flop used in practice.

Another common source of non-synthesisable RTL code is testbenches. Testbenches commonly uses delays:

```
// Typical RTL testbench contents:

// Set the time in seconds for each clock unit.
'timescale 1 ns

reg clk, reset;
initial begin clk=0; forever #5 clk = !clk; end // Clock source 100 MHz
initial begin reset = 1; # 125 reset = 0; end // Power-on reset generator
```

**Take-away summary:** The industry has essentially zeroed-in on a very narrow synthesisable RTL subset. Behavioural input forms are essentially 'syntactic sugar' that are mapped down to pure RTL before logic minimisation and gate mapping.

#### 4.2.5 Further Logic Synthesis Issues

There are many combinational circuits that have the same functionality. Synthesis tools can accept additional guiding metrics from the user, that affect

- Power consumption (with automatic clock gating synthesis),

- Area use,
- Performance,
- Testability (with automatic test modes generation).

Our basic 3-step recipe did not have an optimisation function weighted by such metrics.

Gate libraries have high and low drive strength forms of most gates (see later). The synthesis tool will chose the appropriate gate depending on the fanout and (estimated) net length during routing. Some leaf cells are broadside and do not require bit-blasting.

The tool will use Quine/McCluskey, Espresso or similar for logic minimisation. Liberal use of the 'x' don't care designation in the source RTL allows the synthesis tool freedom to perform this logic minimisation.

```
reg[31:0] y;
...
if (e1) y <= e2;
else if (e3) y <= e4;
else y <= 32'bx;           // Note, assignment of 'x' permits automated logic minimisation.
```

Can share sub-expressions or re-compute expressions locally. Reuse of sub-expressions is important for locally-derived results, but with today's VLSI, sending a 32 bit addition result more than one millimeter on the chip may use more power then recomputing it locally! Logic synthesis is an underconstrained optimisation problem. Both choosing what cubes to use in a Boolean expression and finding which subexpressions are useful when generating several output functions are exponentially complex. Iteration and hill climbing must be used. Also, we can sometimes re-encode state so that output function is simple to decode (covered in ABD notes not lectured this year but critical for HLS where the controlling sequencer hugely benefits). (The most famous logic synthesiser is Design Compiler from Synopsys which has been used for the majority of today's chips.)

## 4.3 Simulation

Simulation of real-world systems generally requires quantisation in time and spatial domains.

There are two main forms of simulation modelling:

- FDS: finite-difference time-domain simulation, and
- EDS: event-driven simulation.

Finite-difference simulation is used for analogue and fluid-flow systems. An example is the SPICE simulator used in the Power section of these notes to model an inverter. It is rarely used in SoC design (just for low-level electrical propagation and crosstalk modelling). Variable element size (and variable temporal step size) can be used to make finite-element simulations approximate even-driven behaviour.

```
Finite-element difference equations (without midpoint rule correction):
tnow += deltaT;
for (n in ...) i[n] = (v[n-1]-v[n])/R;
for (n in ...) v[n] += (i[n]-i[n+1])*deltaT/C;
```

Basic finite-difference simulation uses fixed spatial grid (element size is deltaL) and fixed time step (deltaT seconds). Each grid point holds a vector of instantaneous local properties, such as voltage, temperature, stress, pressure, magnetic flux. Physical quantities are divided over the grid. Three examples:

1. Sound wave in wire:  $C = \text{deltaL} * \text{mass-per-unit-length}$ ,  $R = \text{deltaL} * \text{elasticity-per-unit-length}$

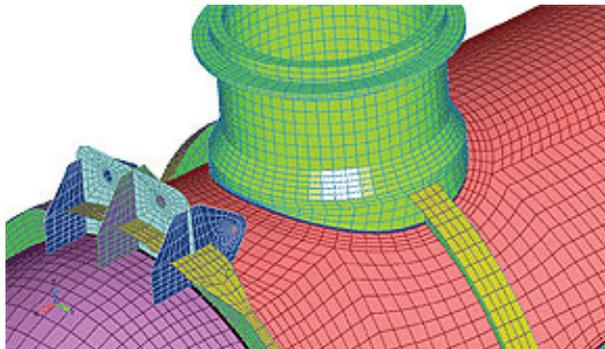


Figure 4.10: Finite Element Grid

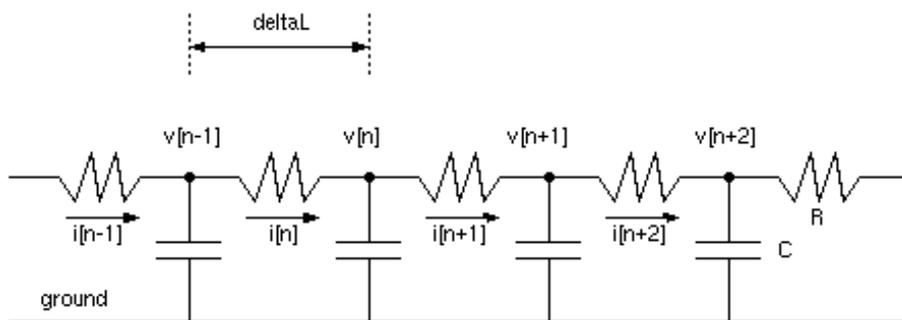


Figure 4.11: Baseline finite-difference model for bidirectional propagation in one dimension.

2. Heat wave in wire:  $C = \text{deltaL} * \text{heat-capacity-per-unit-length}$ ,  $R = \text{deltaL} * \text{thermal-conductance-per-unit-length}$
3. Electrical wave in wire:  $C = \text{deltaL} * \text{capacitance-per-unit-length}$ ,  $R = \text{deltaL} * \text{resistance-per-unit-length}$

Larger modelling errors with larger  $\text{deltaT}$  and  $\text{deltaL}$ , but faster simulation. Keep them less than 1/10th wavelength for good accuracy.

Generally use a 2D or 3D grid for fluid modelling: 1D ok for electronics. Typically want to model both resistance and inductance for electrical system. When modelling inductance instead of resistance, then need a '+=' in the  $i[n]$  equation. When non-linear components are present (e.g. diodes and FETs), SPICE simulator adjusts  $\text{deltaT}$  dynamically depending on point in the curve.

### 4.3.1 Digital Logic Modelling

In the four-value logic system each net (wire or signal), at a particular time, has one of the following logic values:

- 0 logic zero
- 1 logic one
- Z high impedance — not driven at the moment
- X uncertain — the simulator does not know

In this model, nets jump from one value to another in an instant. Real nets have a transit time.

The symbol 'X' has a different meaning according to tool applied: it means 'uncertain' during simulation and 'dont-care' during logic synthesis. The dont-care in logic synthesis enables logic minimisation (as done visually with Karnaugh maps).

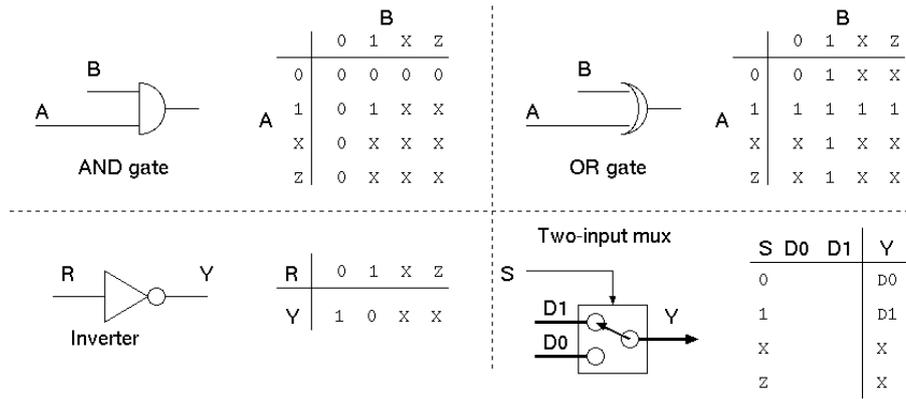


Figure 4.12: Illustrating the four-value logic level encoding for common gates.

Verilog and VHDL generally use more complex logic than the four-value logic system, with various strengths of logic zero and one so that weak effects such as pull-up resistors and weedy pass transistors can be modelled. This enables a net-resolution function to be applied when a net is driven by more than one source. For instance, an equal drive-strength one and zero will resolve to an X but the stronger will win when strengths are not matched.

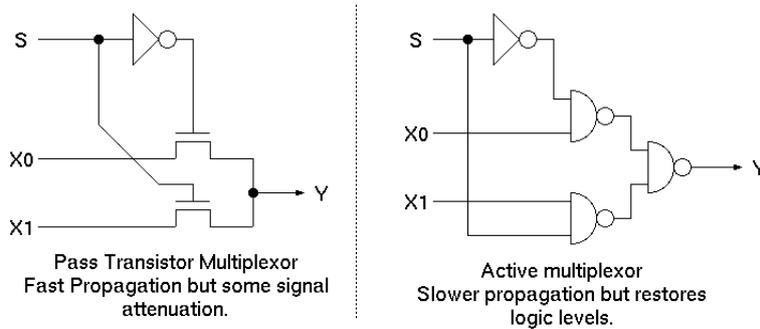


Figure 4.13: Pass transistor multiplexor compared with an active path multiplexor that needs greater silicon area, especially when the control inverter is amortised over a word.

The **pass transistor** is a cheap (in area terms) and efficient (in delay terms) form of programmable wiring, but it does not amplify the signal.

### 4.3.2 Event Driven Simulation

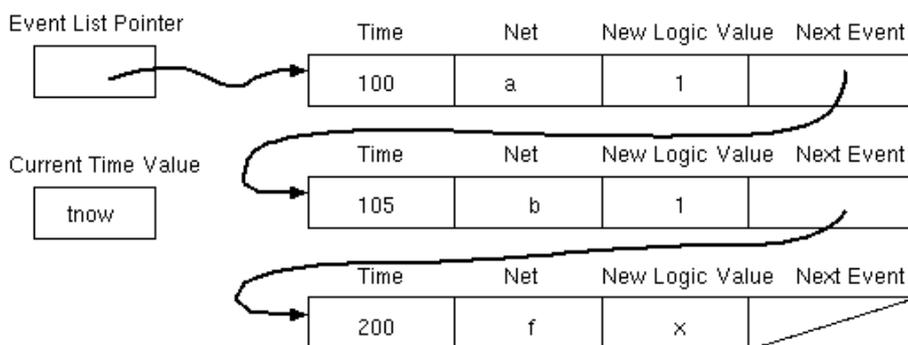


Figure 4.14: Event queue, linked list, sorted in ascending temporal order.

The following ML fragment demonstrates the main datastructure for an EDS kernel. EDS ML fragments

```

// A net has a string name and a width.
// A net may be high z, dont know or contain an integer from 0 up to 2**width - 1.
// A net has a list of driving and reading models.

type value_t = V_n of int | V_z | V_x;

type net_t = {
  net_name:      string;          // Unique name for this net.
  width:         int;             // Width in bits if a bus.
  current_value: value_t ref;     // Current value as read by others
  net_inertia:   int;            // Delay before changing (commonly zero).
  sensitives:   model_t list ref; // Models that must be notified if changed.
};

// An event has a time, a net to change, the new value for that net and an
// optional link to the next on the event queue:
type event_t = EVENT of int * net_t * value_t * event_t option ref

```

This reference implementation of an event-driven simulation (EDS) kernel maintains an ordered queue of events commonly called the **event list**. The current simulation time, **tnow**, is defined as the time of the event at the head of this queue. An event is a change in value of a net at some time in the future. Operation takes the next event from the head of the queue and dispatches it. Dispatch means changing the net to that value and chaining to the next event. All component models that are sensitive to changes on that net then run, potentially generating new events that are inserted into the event queue.

Code fragments (*details not examinable*):

Create initial, empty event list:

```
val eventlist = ref [];
```

Constructor for a new event: insert at correct point in the sorted event list:

```

fun create_and_insert_event(time, net, value) =
  let fun ins e = case !e of
        (A as EMPTY) => e := EVENT(time, net, value, ref A)
      | (A as EVENT(t, n, v, e')) => if (t > time)
        then e := EVENT(time, net, value, ref A)
        else ins e'
      in ins eventlist
  end

```

Main simulation: keep dispatching until event list empty:

```

fun dispatch_one_event() =
  if (!eventlist = EMPTY) then print("simulation finished - no more events\n")
  else let val EVENT(time, net, value, e') = !eventlist in
        ( eventlist := !e';
          tnow := time;
          app execute_model (net_setvalue(net, value))
        ) end

```

### 4.3.3 Inertial and Transport Delay

Consider a simple 'NOR' gate model with 250 picosecond delay. It has two inputs, and the behavioural code inside the model will be something like (in SystemC-like syntax, possibly covered elsewhere)

```

SC_MODULE(NOR2)
{
  sc_in < bool > i1, i2; sc_out < bool > y;
  void behaviour()
  { y.write(!(i1.read() || i2.read()), SC_TIME(250, SC_PS));
  }
  SC_CTOR(NOR2) { SC_METHOD(behaviour); sensitive << i1 << i2;
  }
}

```

The above model is run when either of its inputs change and it causes a new event to be placed in the event queue 250 picoseconds later. This will result in a pure **transport delay**, because multiple changes on the input within 250 picoseconds will potentially result in multiple changes on the output that time later. This is unrealistic: a NOR gate made of transistors will not respond to rapid changes on its input, and only decisively change

its output when the inputs have been stable for 250 picoseconds. In other words, it exhibits inertia. To model **inertial delay**, the event queue insert function must scan for any existing scheduled changes before the one about to be inserted and delete them. This involves little overhead since we are scanning down the event queue anyway.

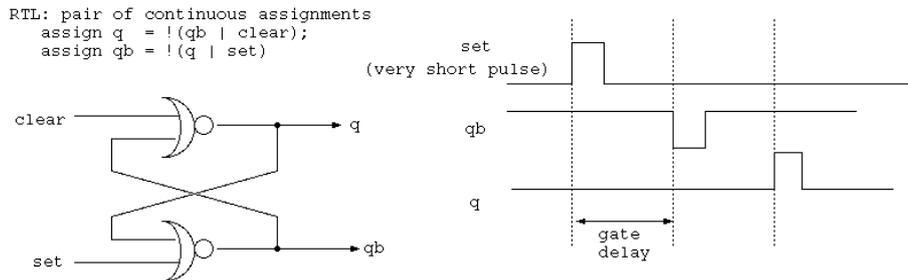


Figure 4.15: RS-latch: behaviour of runt pulse when modelling with transport delay.

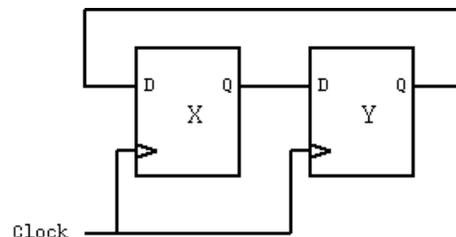
Consider the behaviour of the above RS-latch when a very short (runt) pulse or glitch tries to set it. What will it do with transport models?: the runt pulse will circulate indefinitely. What will it do with inertial models?: ignore the glitch.

#### 4.3.4 Modelling Zero-Delay Components - The Delta Cycle

At early stages of design exploration, we may not know anything about the target technology. We do not wish to insert arbitrary delay figures in our source code, yet some sort of delay is needed to make synchronous hardware work correctly. The solution is the **delta cycle**.

For correct behaviour of synchronous edge-triggered hardware, the propagation delay of D-types must be greater than their hold time. Question : How can we ensure this in a technology-neutral model that does not have any specific numerical delays ?

```
// Example: swap data between a pair of registers
reg [7:0] X, Y;
always @(posedge clock) begin
  X <= Y;
  Y <= X;
end
// E.g. if X=3 and Y=42 then Y becomes 3 and X becomes 42.
```



Answer: Hardware simulators commonly support the compute/commit or '**signal**' paradigm for non-blocking updates. The signal has **current** and **next** values.

*Delta Cycles are not examinable CST 15/16 onwards.*

All three of VHDL, Verilog RTL and SystemC support the **compute/commit** paradigm (also known as evaluate/update) using **delta cycles**. Delta cycle: a complete compute/commit cycle that does not advance global time. Zero-delay models generate new events at the current time, now. To avoid **shoot-through**, these need to be delayed until all current evaluation is complete. Models write such events to the **next** field of the signal and such writes add the signal to the pending list.

EDS kernel with pending queue (simplified):

Raw EDS without deltacycles

```
while (eventlist <> EMPTY)
{ e = hd eventlist;
  eventlist = tl eventlist;
  tnow = e.time;
  e.net.current = e.value;
  for (m in e.net.models) do m.exec()
}
```

```
while (eventlist <> EMPTY)
{ e = hd eventlist;
  if (e.time > tnow) and (pending<>EMPTY)
  { // Commit pendings and commence new delta cycle
    for (net in pending) do net.current=net.next;
    pending = EMPTY;
    for (n in nets just updated) for (m in n.models) m.exec();
  } else
  { eventlist = tl eventlist;
    tnow = e.time;
    e.net.current = e.value;
    for (m in e.net.models) do m.exec();
  }
}
```

Details fully lectured by MJCG (slide 110 onwards): Temporal Logic Course

One implementation is to have an auxiliary list containing nets, called the pending queue. The `net.write(value, when)` method checks whether the new time is the same as the current time and if so, instead of inserting an event for the net in the event list, the net is placed on the pending queue and the new value stored in a 'next value' field in the net. The kernel is then modified as shown above, to empty the pending queue when the next event would advance simulation time.

Note that committing a pending queue update can wake up a model sensitive to the updated net and cause further events to be generated. All pending updates must be committed atomically and new zero-delay events are added to a fresh pending queue.

Hence, when zero-delay models are active and the output of one feeds another (e.g. a zero delay gate in the clock path), the value of system time, `tnow`, may not advance for several consecutive delta cycles. Clock generators or other components for which we can put in delay figures operate normally, causing real advances in simulation time.

A net that is to have its updated deferred in VHDL (and SystemC) is called a **signal**, whereas immediate updates when variables are written to. In Verilog, all nets can be assigned in either way and instead two different assignment operators are provided (called blocking and non-blocking, denoted `=` and `<=` respectively).

(As we shall see, a SystemC 'sc\_signal' is implemented with a current and a next value and it is necessary to use the 'net.read()' method to read the value of a SystemC signal because C++ does not allow override of the read operator.)

### 4.3.5 Mixed Analog/Digital Simulation (Verilog-AMS)

**Hybrid System** simulations: typically a digital (embedded) controller interacts with analogue plant.

Example: Vehicle hybrid power system and automatic braking system.

We need to be able to combine digital (event-driven) simulation with analogue models, such as fluid-flow nodal analysis and FDTD modelling.

Cyberphysical trends: Cyberphysical Functional Mock Up Interface Relevant tools: Verilog-AMS, VHDL-AMS, Modelica, Simulink and FMI. Wikipedia: Verilog-AMS

Verilog Analogue and Mixed Signal (AMS) extends RTL to support:

- Signals of both analogue and digital types can be declared in the same module.
- Initial, always and **analogue procedural blocks** can appear in the same module.
- Digital signal values can be set (write operations) from any context outside of an analogue procedural block

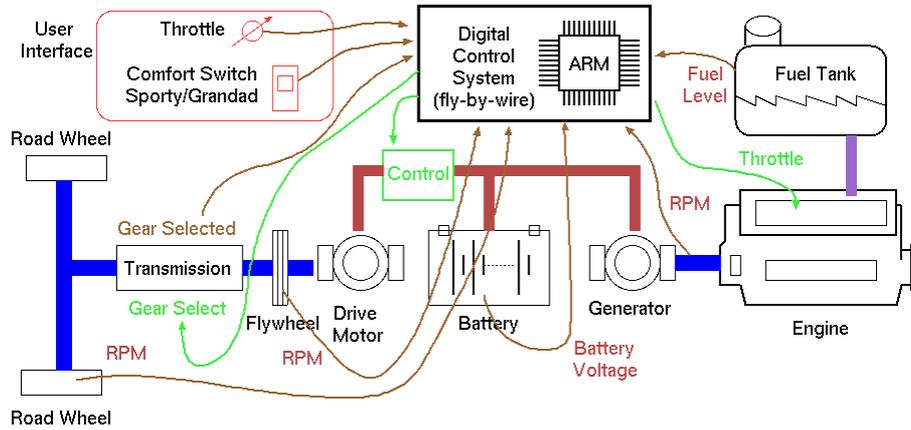


Figure 4.16: Hybrid Automobile Transmission System.

- Analogue potentials and flows can only receive contributions (write operations) from inside an analogue procedural block
- An **analog initial begin ... end** statement sets up initial voltages on capacitors or levels in a fuel tank.
- A new sensitivity enables triggering actions **always @(cross(Vt1 - 2.5)) begin ... end**.

Examples:

```
// Three 1.5 cells in series make a 4.5 volt battery.
module Battery4V5(input voltage anode, output voltage cathode);
  voltage t1, t2;
  analog begin
    V(anode) <+ 1.5 + V(t2);
    V(t2) <+ 1.5 + V(t1);
    V(t2) <+ 1.5 + V(cathode);
  end
endmodule
```

```
module resistor (inout electrical a, inout electrical b);
  parameter real R = 4700;
  analog V(a,b) <+ R * I(a,b);
endmodule
```

Verilog simulation cycle extended to support solving nodal flow equations. When we potentially de-queue a time-advancing event from EDS queue we first roll forward the FDTD simulations which themselves may contain 'cross' and similar sensitivity that insert new events on the EDS queue.

### 4.3.6 Mixed Analog/Digital Simulation: An interesting problem attributable to Zeno?

Problem: Remember Zeno's paradox, with Achilles and the tortoise? *And so you can never catch up, and the Tortoise concluded sympathetically.* AMS simulations often suffer from unintentionally revisiting that story. A Zeno hybrid system model is a hybrid system with an execution that takes an infinite number of discrete transitions during a finite time interval.

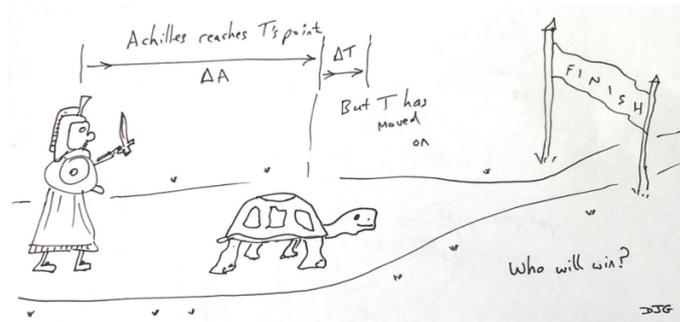


Figure 4.17: Archilles can never catch the tortoise? Sum of GP never converges?

```
// AMS simulation of of a ball bouncing -> infinite bounce frequency!
module ballbounce();
  real height, velocity;

  analog initial begin height = 10.0; velocity = 0.0; end

  analog begin // We want auto-timestep selection for this FDTD
    height <+ -velocity; // Falling downwards
    velocity <+ 9.8; // Acceleration due to gravity.
  end

  // We want discrete event triggered execution here
  always @(cross height) begin
    velocity = -0.9 * velocity; // Inelastic bounce
    height = 0.000001; // Hmm some fudge here!
  end
endmodule
// NB: Precise syntax above may not be accepted by all tools.
```

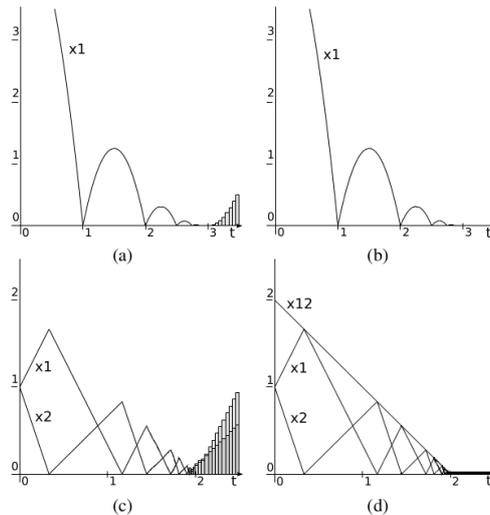


Fig. 2: Enclosures for systems with Zeno behavior. Figures (a) and (c) show the results for the bouncing ball and water tank systems of Lygeros [33, Figures 3.1 and 3.7]. However, we make these models more challenging to simulate by replacing some inequalities with equalities (such as the bouncing conditions). Figures (b) and (d) show the results for the examples when additional (redundant) constraints are added to achieve more precise enclosures.

Figure 4.18: Bounding boxes in one approach to useful simulation of Zeno-generating systems (Konecny).

A simple heuristic on the minimum timestep competes directly with adaptive timestep tuning needed to accurately model critical inflection points. Zeno suppression research is ongoing.

Note, the ball bounce example does not involve any nodal equations but the problem is fairly common with real-world examples that do tend to have wire and pipes splitting and joining.

### 4.3.7 Higher-level Simulation

Simulating RTL is slow. Every net (wire) in the design is modelled as a shared variable. When one component writes a value, the overheads of waking up the receiving component(s) may be severe. The event-driven simulator kernel contains an indirect jump instruction which itself is very slow on modern computer architectures since it will not get predicted correctly.

Much faster simulation is achieved by disregarding the clock and making so-called TLM calls between the components. Subroutine calls made between objects convey all the required information. Synchronisation is achieved via the call and its return. This is discussed in the ESL section of this course.

## 4.4 Hazards

Definitions (some authors vary slightly):

- **WaW hazard** - write-after-write: one write must occur after another otherwise the wrong answer persists,
- **RaW or WaR hazard** - write and read of a location are accidentally permuted,
- **Other Data hazard** - when an operand simply has not arrived in time for use,
- **Control hazard** - when it is not yet clear whether the results of operation should be committed (computation might still start speculatively),
- **Name Alias hazard** - we do not know if two array subscripts are equal,
- **Structural hazard** - insufficient physical resources to do everything at once.

(Where the address to a register file has not yet arrived we have a data hazard on the address itself, but this could be regarded as a control hazard for the register file operation itself (read or write).)

We have a structural hazard when an operation cannot proceed because a resource is already in use. Resources that might present structural hazards are:

- Memories and register files with insufficient ports,
- Memories with variable latency, especially DRAM,
- Insufficient number of ALUs for all the arithmetic to be scheduled in current clock tick,
- Anything **non-fully pipelined** i.e. something that goes busy, such as long multiplication (e.g. Booth Multiplier or division or a floating point unit).

A **fully-pipelined** component can start a new operation on every clock cycle. It will have fixed latency (pipeline delay). These are commonly encountered and are easiest to form schedules around. A non-fully pipelined components generally have handshake wires that start it and inform the client logic when it is busy. This is needed for computations better performed with variable latency. Another form that is non-fully pipelined has a **reinitiation interval** greater than one: for example, it might accept new data every third clock cycle, but still be fixed-latency.

Synchronous RAMs, and most complex ALUs excluding divide, are generally fully pipelined and fixed-latency.

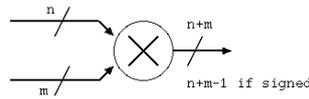


Figure 4.19: Multiplier schematic symbol.

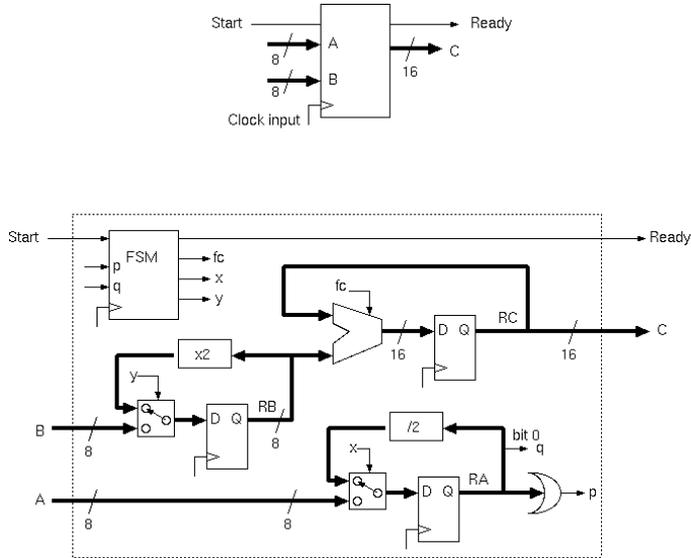
An example of a component that cannot accept new input data every clock cycle (i.e. something that is non-fully-pipelined) is a sequential long multiplier, that works as follows:

Behavioural algorithm:

```

while (1)
{
    wait (Start);
    RA=A; RB=B; RC=0;
    while(RA>0)
    {
        if odd(RA) RC=RC+RB;
        RA = RA >> 1;
        RB = RB << 1;
    }
    Ready = 1;
    wait(!Start);
    Ready = 0;
}
    
```

(Either HLS or hand coding can give the illustrated datapath and sequencer structure:)



This implements conventional long multiplication. It is certainly not fully-pipelined, it goes busy for many cycles, depending on the log of the A input. The illustration shows a common design pattern consisting of a **datapath** and a **sequencer**. Booth's algorithm (see additional material) is faster, still using one adder but needing half the clock ticks.

### 4.4.1 Multiplier Answer

```

module LONGMULT8b8(clk, reset, C, Ready, A, B, Start);

    input clk, reset, Start;
    output Ready;
    input [7:0] A, B;
    output [15:0] C;
    reg [15:0] RC, RB, RA;
    reg Ready;

    // Behavioural code:
    // while (1)
    // {
    //     wait (Start);
    //     RA=A;RB=B;RC=0;
    //     while(RA>0)
    //     { if odd(RA) RC=RC+RB;
    //       RA = RA >> 1;  RB = RB << 1;
    //     }
    //     Ready = 1;
    //     wait(!Start);
    //     Ready = 0;
    // }

    reg xx, yy, qq, pp; // Control and predicate nets
    reg [1:0] fc;
    reg [3:0] state;
    always @(posedge clk) begin
        xx = 0; // default settings.
        yy = 0;
        fc = 0;

        // Predicates
        pp = (RA!=16'h0); // Work while pp holds
        qq = RA[0]; // Odd if qq holds
        // Sequencer
        if (reset) begin
            state <= 0;
            Ready <= 0;
        end
        else case (state)
            0: if (Start) begin
                xx = 1;
                yy = 1;
                fc = 2;
                state <= 1;
            end

            1: begin
                fc = qq;
                if (!pp) state <= 2;
            end

            2: begin
                Ready <= 1;
                if (!Start) state <= 3;
            end

            3: begin
                Ready <= 0;
                state <= 0;
            end

        endcase // case (state)

        // Datapath
        RB <= (yy) ? B: RB<<1;
        RA <= (xx) ? A: RA>>1;
        RC <= (fc==2) ? 0: (fc==1) ? RC+RB: RC;
    end

    assign C = RC;
endmodule

```

Suitable test wrapper:

```

module SIMSYS();
  reg clk, reset, Start;
  wire Ready;
  wire [7:0] A, B;
  wire [15:0] C;

  // Reset Generator
  initial begin reset = 1; # 55 reset = 0; end

  // Clock Generator
  initial begin clk = 0; forever #10 clk = !clk; end

  // Stimulus
  assign A = 6;
  assign B = 7;

  // Handshake control
  always @(posedge clk) Start <= !Ready;

  // Console output logging:
  always @(posedge clk) $display("Ready=%h C=%d");

  // Device under test.
  LONGMULT8b8 the_mult(clk, reset, C, Ready, A, B, Start);
endmodule // SIMSYS

```

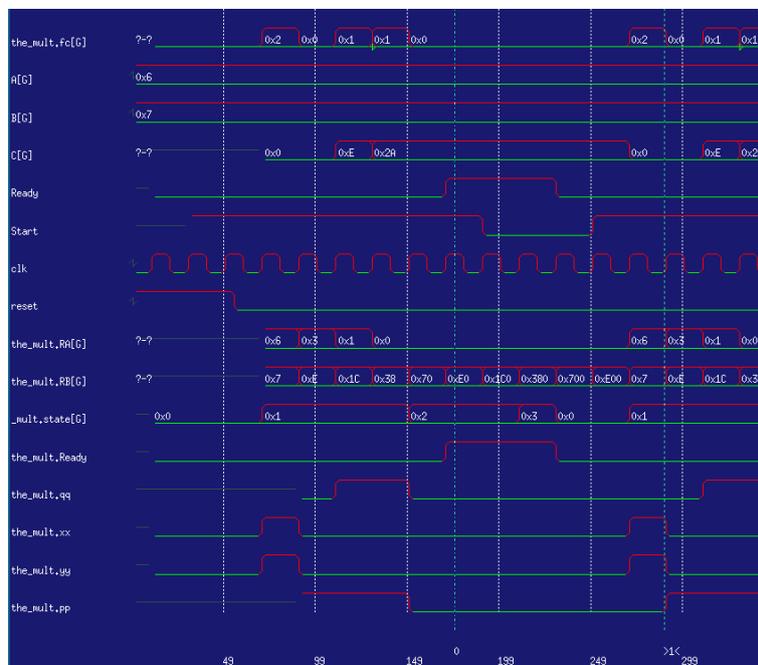


Figure 4.20: Long Multiplier Timing Waveforms

#### 4.4.2 Hazards From Array Memories

A structural hazard in an RTL design can make it non synthesisable. Consider the following expressions that make liberal use of array subscription and the multiplier operator:

Structural hazard sources are numbered:

```
always @(posedge clk) begin
    q0 <= Boz[e3]           // 3
    q1 <= Foo[e0] + Foo[e1]; // 1
    q2 <= Bar[Bar[e2]];    // 2
    q3 <= a*b + c*d;       // 4
    q4 <= Boz[e4]         // 3
end
```

1. The RAMs or register files Foo Bar and Boz might not have two read ports.
2. Even with two ports, can Bar perform the double subscription in one clock cycle?
3. Read operations on Boz might be a long way apart in the code, so hazard is hard to spot.
4. The cost of providing two 'flash' multipliers for use in one clock cycle while they lie idle much of the rest of the time is likely not warranted.

A multiplier that operates combinationaly in less than one clock cycle is called a 'flash' multiplier and it uses quadratic silicon area.

RAMs have a small number of ports but when RTL arrays are held in RAM it is easy to write RTL expressions that require many operations on the contents of a RAM in one operation, even from within one thread. For instance we might need three operations on a RAM to implement

```
A[x] <= A[y + A[z]]
```

Because RTL is a very-low-level language, RTL typically requires the user to do manual scheduling of port use. (However, some current FPGA tools do a certain amount of scheduling for the user.)

Multipliers and floating point units also typically present hazards.

To overcome hazards automatically, stalls and holding registers must be inserted. The programmer's original model of the design must be stalled as ports are re-used in the time domain, using extra clock cycles to copy data to and from the holding registers. This is not a feature of standard RTL so it must either be done by hand or automatically (see HLS section of this course).

Expanding blocking assignments can lead to **name alias** hazards:

Suppose we know nothing about *xx* and *yy*, then consider:

```
begin
    ...
    if (g) Foo[xx] = e1;
    r2 = Foo[yy];
```

To avoid **name alias** problems, this must be compiled to non-blocking pure RTL as:

```
begin
    ...
    Foo[xx] <= (g) ? e1: Foo[xx];
    r2 <= (xx==yy) ? ((g) ? e1: Foo[xx]): Foo[yy];
```

Quite commonly we do know something about the subscript expressions. If they are compile-time constants, we can decidedly check the equality at compile time. Suppose that at ... or elsewhere beforehand we had the line '*yy = xx+1;*' or equivalent knowledge? Then with sufficient rules we can realise at compile time they will never alias. However, no set of rules will be complete (decidability). And commonly they are a linear function of a loop variable of an enclosing loop (an induction expression) and, after strength reduction, the *xx+k* pattern is readily manifest.

#### 4.4.3 Overcoming Structural Hazards using Holding Registers

One way to overcome a structural hazard is to deploy more resources. These will suffer correspondingly less contention. For instance, we might have 3 multipliers instead of 1. This is the **spatial** solution. For RAMs and register files we need to add more ports to them or mirror them (i.e. ensure the same data is written to each copy).

In the **temporal solution**, a **holding register** is commonly inserted to overcome a structural hazard (by hand or by a high-level synthesis tool). Sometimes, the value that is needed is always available elsewhere in the design (and needs forwarding) or sometimes an extra sequencer step is needed.

If we know nothing about  $e_0$  and  $e_1$ : then load holding register in additional cycle:

```
always @(posedge clk) begin
  ...
  ans = Foo[e0] + Foo[e1];
  ...
end
```

```
always @(posedge clk) begin
  pc = !pc;
  ...
  if (!pc) holding <= Foo[e0];
  if (pc) ans <= holding + Foo[e1];
  ...
end
```

If we can analyse the pattern of  $e_0$  and  $e_1$ :

```
always @(posedge clk) begin
  ...
  ee = ee + 1;
  ...
  ans = Foo[ee] + Foo[ee-1];
  ...
end
```

then, apart from first cycle, use holding register to forward value from previous iteration (**loop forwarding**):

```
always @(posedge clk) begin
  ...
  ee <= ee + 1;
  holding <= Foo[ee];
  ans <= holding + Foo[ee];
  ...
end
```

We can implement the program counter and holding registers as source-to-source transformations, that eliminate hazards, as just illustrated. One algorithm is to first to emit behavioural RTL and then to alternate the conversion to pure form and hazard avoidance rewriting processes until closure.

For example, the first example can be converted to old-style behavioural RTL that has an implicit program counter (state machine) as follows:

```
always @(posedge clk) begin
  holding <= Foo[e0];
  @(posedge clk) ;
  ans <= holding + Foo[e1];
end
```

The transformations illustrated above are NOT performed by mainstream RTL compilers today: instead they are incorporated in HLS tools such as Kiwi. KiwiC Structural Hazard ExampleSharing structural resources may require additional multiplexers and wiring: so not always worth it. A good design not only balances structural resource use between clock cycles, but also critical path timing delays.

These example fragments handled one hazard and used two clock cycles. They were localised transformations. When there are a large number of clock cycles, memories and ALUs involved, a global search and optimise procedure is needed to find a good balance of load on structural components. Although these examples mainly use memories, other significant structural resources, such as fixed and floating point ALUs also present hazards.

## 4.5 Folding, Retiming & Recoding

Generally we have to chose between high performance or low power. (We can see this also in the selection of drive strengths for standard cell gates.) The **time/space fold** and **unfold** operations trade execution time for silicon area. A given function can be computed with fewer clocks by 'unfolding' in the the time domain, typically by loop unwinding (and predication).

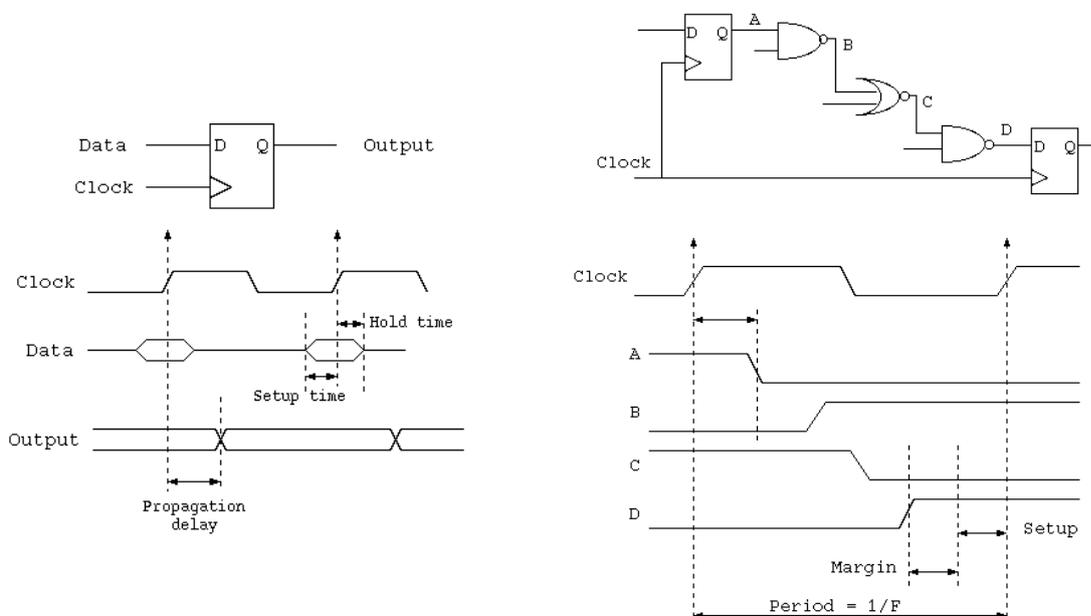
LOOPED (time) option: <pre> for (i=0; i &lt; 3 and i &lt; limit; i++)   sum += data[i] * coef[i+j]; </pre>	UNWOUND (space) option: <pre> if (0 &lt; limit) sum += data[0] * coef[j]; if (1 &lt; limit) sum += data[1] * coef[1+j]; if (2 &lt; limit) sum += data[2] * coef[2+j]; </pre>
---	---

The '+' operator is an **associative reduction** operator. When the only interactions between loop iterations are outputs via such an operator, the loop iterations can be executed in parallel. On the other hand, if one iteration stores to a variable that is read by the next iteration or affects the loop exit condition then unwinding possibilities are reduced.

We can **retime** a design with and without changing its state encoding. We will see that adding a pipeline stage can increase the amount of state without **recoding** existing state. *Note: these folding operations, discussed here in their RTL form, are precursors to considering their automated deployment by HLS tools.*

### 4.5.1 Critical Path Timing Delay

Meeting **timing closure** is the process of manipulating a design to meet its target clock rate (as set by the Marketing Department for instance).



The maximum clock frequency of a synchronous clock domain is set by its critical path. The longest path of combinational logic must have settled before the setup time of any flip-flop starts.

Pipelining is a commonly-used technique to boost system performance. Introducing a pipeline stage increases latency but also the maximum clock frequency. Fortunately, many applications are tolerant to the processing delay of a logic subsystem. Consider a decoder for a fibre optic signal: the fibre might be many kilometers long and a few additional clock cycles in the decoder increase the processing delay by an amount equivalent to a few coding symbol wavelengths: e.g. 20 cm per pipeline stage for a 1 Gbaud modulation.

Pipelining introduces new state but does not require existing state flip-flops to change meaning.

**Flip-flop migration** does alter state encoding. Migration may be manually turned on or off during logic synthesis by typical RTL compiler tools. It exchanges delay in one path for delay in another - aim to achieve balance. A sequence of such transformations can lead to a shorter critical path overall.

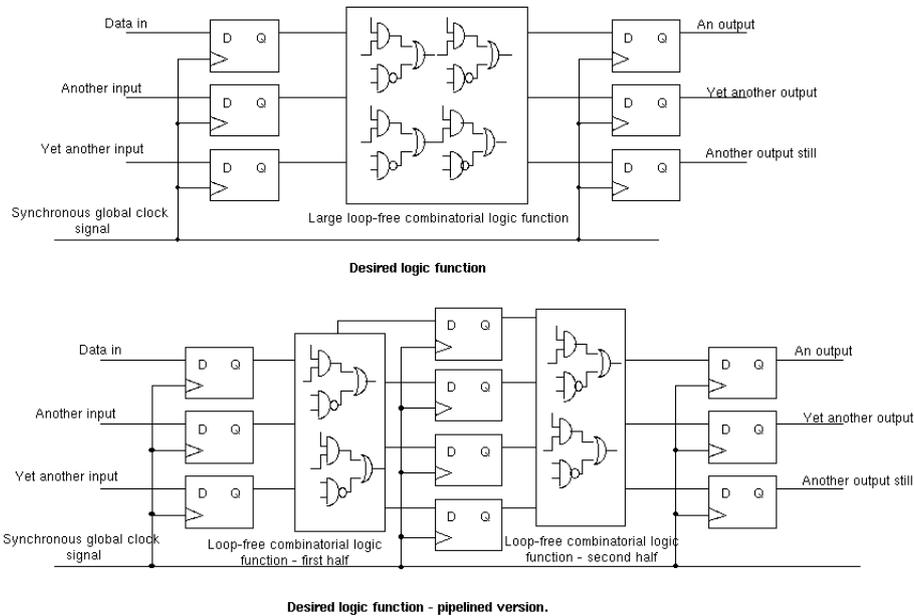


Figure 4.21: A circuit before and after insertion of an additional pipeline stage.

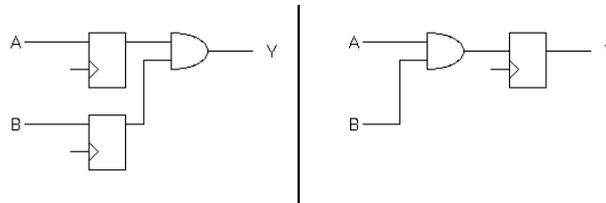


Figure 4.22: Flip-flop migration: two circuits of identical behaviour, but different state encoding.

In the following example, the first migration is a local transformation that has no global consequences:

Before:	Migration 1:	Migration 2 (non causal):
<code>a &lt;= b + c;</code>	<code>b1 &lt;= b; c1 &lt;= c;</code>	<code>q1 &lt;= (dd) ? (b+c): 0;</code>
<code>q &lt;= (d) ? a:0;</code>	<code>q &lt;= (d) ? b1+c1:0;</code>	<code>q &lt;= q1;</code>

The second migration, that attempts to perform the multiplexing one cycle earlier will require an earlier version of `d`, here termed `dd` that might not be available (e.g. if it were an external input we need knowledge of the future). An earlier version of a given input can sometimes be obtain by delaying all of the inputs (think of delaying all the inputs to a bookmakers shop), but this cannot be done for certain applications where system response time (in-to-out delay) is critical.

Problems arising:

- Circuits containing loops (proper synchronous loops) cannot be pushed very far (for example, the control hazard in a RISC pipeline).
- External interfaces that do not use transactional handshakes (i.e. those without flow control) cannot tolerate automatic re-timing since the knowledge about when data is valid is not explicit.
- Many structures, including RAMs and ALUs, have a pipeline delay (or several), so the hazard on their input port needs resolving in a different clock cycle from hazards involving their result values.

but retiming can overcome structural hazards (e.g. the ‘write back’ cycle in RISC pipeline).

Other rewrites commonly used: automatically recode for one-hot or gray encoding, or invert for reset as preset. Large FSMs are generally recoded by FPGA tools by default so that the output function is easy to generate. This is critical for good performance with complex HLS sequencers.

### 4.5.2 Static Timing Analyser Tool

A static analysis tool does not run a program or simulate a design - instead it 'stares' at the source code.

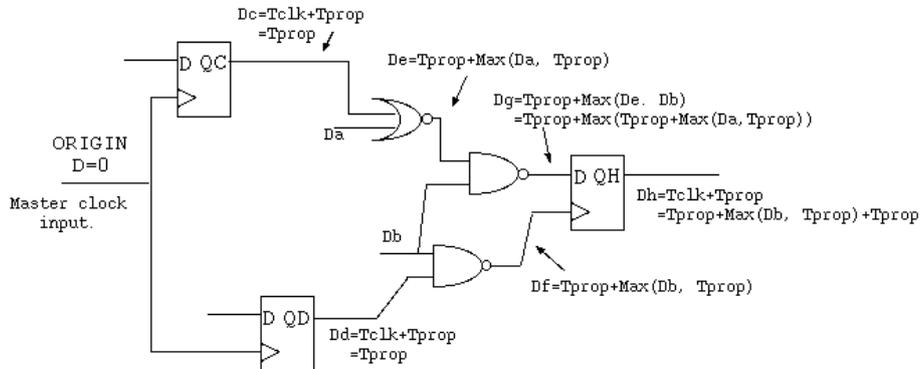


Figure 4.23: An example circuit with static timing annotations

A static timing analyser computes the longest event path through logic gates and clock-to-Q paths of edge-triggered flops. The longest path is generally the critical path that sets the maximum clock frequency. However, sometimes this is a false result, since this path might never be used during device operation.

Starting with some reference point, taken as  $D=0$ , such as the master clock input to a clock domain, we compute the relative delay on the output of each gate and flop. For a combinational gate, the output delay is the gate's propagation time plus the maximum of its input delays. For an edge-triggered flop, such as a D-type or a JK, there is no event path to the output from the D or JK inputs, so it is just the clock delay plus the flop's clock-to-Q delay. There are event paths from asynchronous flop inputs however, such as preset, reset or transparent latch inputs.

Propagation delays may not be the same for all inputs to a given output and for all directions of transition. For instance, on deassert of asynchronous preset to a flop there is no event path. Therefore, a tool may typically keep separate track of high-to-low and low-to-high delays.

### 4.5.3 Back Annotation and Timing Closure

Once the system has been placed and routed, the length and type of each conductor is known. These facts allow fairly accurate delay models of the conductors to be generated (Section 2.1.5).

The accurate delay information is fed into the main simulator and the functionality of the chip or system is checked again. This is known as **back annotation**. It is possible that the new delays will prevent the system operating at the target clock frequency.

The marketing department have commonly pre-sold the product with an advertised clock frequency. Making the actual product work at this frequency is known as meeting **timing closure**.

With low-level RTL, the normal means to achieve timing closure is to migrate logic either side of an existing register or else to add a new register - but not all protocols are suitable for registering (Section 1.3.5).

With transactional interfaces, a one-place FIFO can help with timing closure.

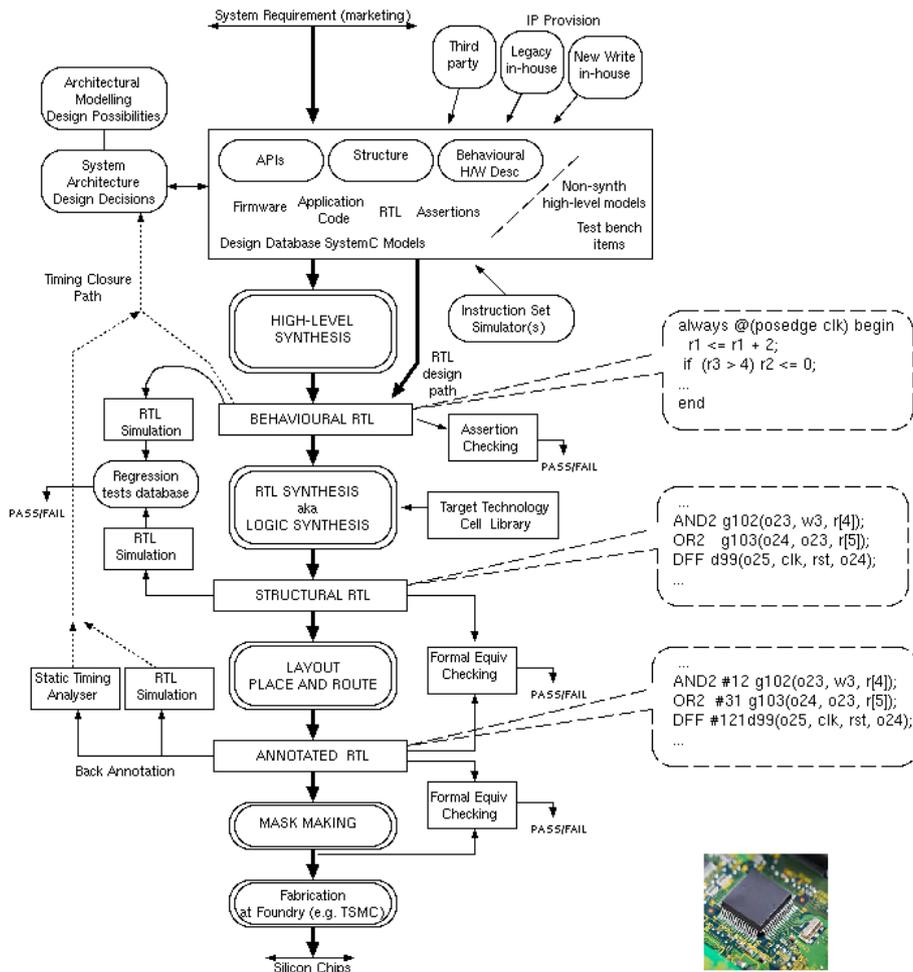


Figure 4.24: Design and Manufacturing Flow for SoC.

### 4.5.4 FIFOs

A FIFO is a first-in, first-out queue. A FIFO typically has some internal storage limit (its capacity) but the interface is independent of that.

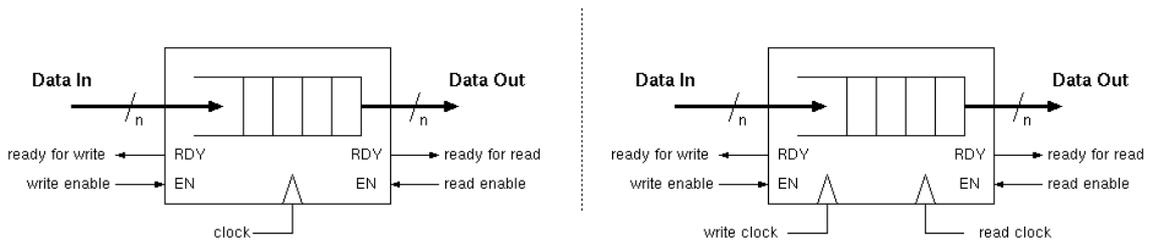


Figure 4.25: FIFO schematic symbols: Synchronous and Clock Domain Crossing.

FIFOs are either synchronous or clock-domain crossing. Synchronous FIFOs are either *bubble-free* (simultaneous read and write operations are always possible) or *internally-pipelined* (effect of a read or write is only visible at the opposite port a clock cycle later, also known as *fully-registered*).

Each of the two ports has a pair of handshake nets. Some FIFOs provide further status output signals such as half or nearly full.

A FIFO can help achieve timing closure, but again the protocol must be transactional and a *bubble-free* FIFO

cannot always be used owing to its combinational path(s), hence latency is introduced.

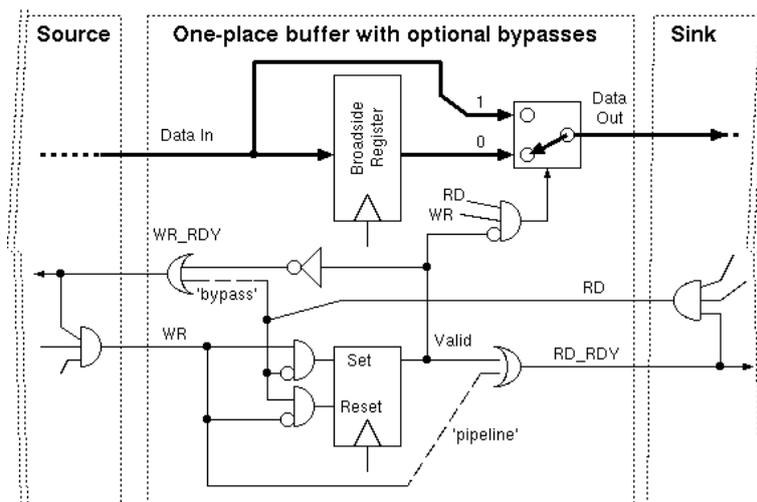


Figure 4.26: A generic, one-place, synchronous FIFO, illustrating combinational path hazards.

The dashed paths show a pair of optional bypasses, called *bypass* and *pipeline*, that may sometimes be installed. Using either of them can provide speedup in terms of latency - but they may not help with timing closure because they can extend the critical path. In other words, their use can remove clock cycles but can also restrict clock frequency. The output multiplexor is only needed if the dashed bypass wire is installed.

(BSV combination path details non-examinable.)

(Bluespec users' note: the default BSV FIFO is fully-registered and so introduces bubbles in both directions: it provides two variants with combinational paths which avoid bubbles in each direction respectively, called *bypass* and *pipelined*, but note that neither variant is fully-pipelined/fully-registered using the above definitions.) The AND gates in the source and sink are typical in most application use styles: they only assert the enable if the ready signal is present, but if both the internal speedups are present we have, overall, a combinational loop. Many EDA tools do not support combinational loops. In this FIFO case, we have a parasitic RS latch. Careful analysis shows that the latch is never set and that there is an equivalent circuit that would provide the same behaviour without the loop. Automatic reworking of that nature is not embodied in typical logic synthesiser tools and so manual editing or a rewriting pass over the circuit from a special tool is needed.

### 4.5.5 Conventional RTL Compared with Software

Synthesisable RTL looks a lot like software at first glance, but we soon see many differences.

RTL is statically-allocated and defines a finite-state machine. Threads do not leave their starting context and all communication is through shared variables that denote wires. There are no thread synchronisation primitives, except to wait on a clock edge. Each variable must be updated by at most one thread.

Software on the other hand uses far fewer threads: just where needed. The threads may pass from one module to another and thread blocking is used for flow control of the data. RTL requires the programmer to think in a massively-parallel way and leaves no freedom for the execution platform to reschedule the design.

RTL is not as expressive for algorithms or data structures as most software programming languages.

The concurrency model is that everything executes in lock-step. The programmer keeps all this concurrency in his/her mind. Users must generate their own, bespoke handshaking and flow control between components.

Verilog and VHDL do not express when a register is **live** with data - hence automatic refactoring and certain correctness proofs are impossible.

For programmers wanting conventional software paradigms, High-Level Synthesis (HLS) should be applied to the high-level program to produce RTL.

## 4.6 Alternatives to RTL

Higher-level entry forms are ideally needed, perhaps scheduling within a thread at compile-time and between threads at run time ?

**High-level Synthesis (HLS)** essentially converts software to hardware. Classically it takes one thread and a fixed body of code and it

- unwinds inner loops by some factor,
- generates a custom **datapath** containing registers, RAMs and ALUs
- and a custom **sequencer** that implements an efficient, static **schedule**

that achieves the same behaviour as the original program. It will generally deploy state re-encoding and re-pipelining to meet timing closure and power budgets.

Greaves wrote an HLS program in around 1990 that was commercially licensed: CTOV Bubble SorterExample. These ideas are only now being seriously considered by industry, with all major tools providers now offering a C-to-gates compiler. LegUp from Toronto is a modern equivalent. The Kiwi Compiler uses similar approaches for acceleration of big data algorithms expressed in concurrent CSharp on FPGA.

For the future, the following two look promising: Chisel: Constructing Hardware in a Scala-Embedded Language and HardCaml - Register Transfer Level Hardware Design in OCaml

### 4.6.1 High-Level Logic Synthesis from Programming Languages Example

Kiwi (Greaves/Singh) Scientific Accelerator: C# programs are implemented on FPGA for high-performance with low energy.

```
public static class test4
{
    static int limit = 10;
    [Kiwi.OutputBitPort("dout")] static bool dout;
    [Kiwi.InputBitPort("din")] static bool din;

    static public void arraypart()
    {
        int odd = 0, even = 0; // These are V_0 and V_1 in the ast.cil file.
        int[] arr = new int [] {0, 1, 222221, 5, 7, 8, 1121, 2021, 2048}; // arr=V_
        arr[2] = 2;
        // V_3 is the current element of the array
        // V_4 is a copy of V_2
        int pr = 0;
        foreach (int i in arr) // i=V_6
        {
            if (i%2 == 0)
                even++;
            else
                odd++;
            Kiwi.Pause();
            Console.WriteLine("{1} i={0}: ", i, pr++);
            Console.WriteLine("so far {0} Odd Numbers, and {1} Even Numbers.", odd, even);
        }
        Console.WriteLine("Found {0} Odd Numbers, and {1} Even Numbers.", odd, even);
        Kiwi.Pause();
    }
}
```

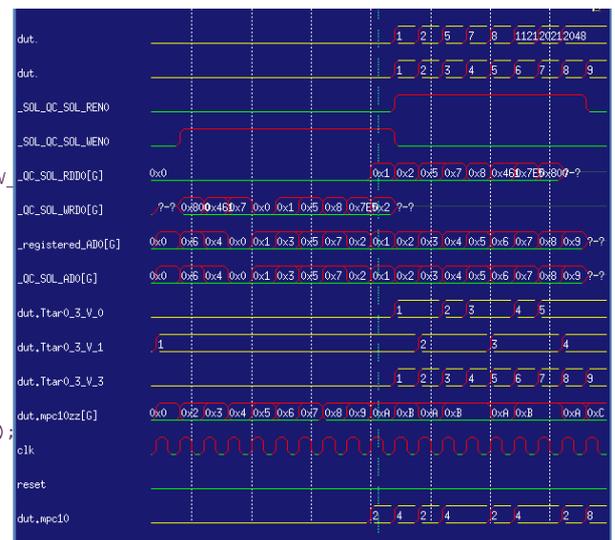


Figure 4.27: Input and Hardware Waveforms from a tiny Kiwi HLS example.

Net names generated by the Kiwi tool include program counter 'mpc10' and address, read and write busses for a RAM called QC\_SOL with suffixes AD0, RD0 and WD0 respectively. The local variables have lost their original names but can be seen to be named V\_n.

This example was compiled in a mode where one clock cycle should be associated with each instance of `Kiwi.Pause` in the `C#` source code and with even tiny arrays being implemented in RAMs. The waveforms reveal (after careful study) that multiple clock cycles were needed to initiate the RAM despite there being no `Kiwi.Pause` commands at that point in the user's program.

## KG 5 — Formal Methods and Assertion-Based Design



### 5.1 Formal Techniques

A simulation testbench consists of an instance of a device-under-test (DUT) and stimulus generators that feed input signals and data to the DUT. A checker may also be present that yields a pass/fail observable output. A testbench for a formal proof of correctness will likewise contain an instance of the DUT and a checker, but instead of stimulus generation and simulation, a proof is made that the property to be checked will hold under all possible input sequences. Such input sequences may be potentially infinite.

For a formal proof to be made, the DUT must be expressed in a language that has a precise formal semantics that is understood by the proof system. Likewise, the property to be checked, known as the *verification condition* must be expressed in either the same language or another language whose formal semantics are understood by the proof system. All input languages will be converted to a unified internal representation format in which the proofs are actually made. Although there is no precise definition of the term ‘formal semantics’, it can be taken to denote a description of precisely how the system will or should behave in a machine-readable form. Normally the formal language is declarative and, like standard mathematics, respects substitution and associativity rules.

Using formal techniques in hardware design has become mainstream. Formal semantics of the synthesisable subsets of both Verilog and VHDL exist and languages such as PSL (property specification language) are widely used to define the allowable transactions on standard busses.

Individual formal proof engines are either automatic or manual. An automatic tool requires no input from the engineer to prove the result or generate a counter-example. Manual tools require the engineer to dictate the steps of the proof, but the proof engine ensures that each step is valid. In practice, the manual tools embody a lot of automation and manual intervention is often needed for the so-called automatic tools.

#### 5.1.1 Assertions

Assertion-based design (ABD) is an approach that encourages writing assertions as early as possible, preferably *before* coding/implementation starts.

1. Writing assertions at design capture time *before* detailed coding starts.
2. Writing further assertions as coding/development progresses.
3. Using the same assertions at product test time.
4. Embedding the assertions as run-time monitors in the product for reporting or automatic shutdown/failsafe.

```
assert(x<4);           // Example imperative safety assertion
x := x + 1000;
assert(x<1004);       // Some programmers think this is the only form of assertion.
```

Programmers are used to dynamically validating **imperative** assertions embedded in procedural programming:

- Imperative (aka immediate) safety checks (like `assert.h` in C++ and `expect` in SystemVerilog)
- Also dynamic coverage checks (log that flow of control has passed a point or a property held).

On the other hand **declarative** programming (aka logic programming) involves writing assertions that hold for all time. For instance, on an indicator panel *never is light A on at the same time as light B*.

Declarative assertions are (conjunctions of):

- Declarative safety properties, that always hold, such as ‘Never are both the inner and outer door of the airlock open at once unless we are on the ground’. Safety properties normally use the keywords **never** or **always**.
- Liveness and deadlock properties, such as ‘If the emergency button is pressed, eventually at least one of the doors will be unlocked.’ Liveness properties normally use keywords such as **eventually** or **it will always be possible to**.

All four forms can be proved by formal techniques such as pen and paper, theorem provers and model checkers. **Dynamic validation** is simulation (or execution) while checking properties. This can sometimes find safety violations and, with careful construction, detect deadlock, but it cannot prove the liveness or that safety won’t be violated.

Assertions can be imported from previous designs or other parts of the same design for global consistency. Formal proof shows up corner case problems not encountered in simulation. A formally-verified result may be *required* by the customer.

### 5.1.2 Validation using Simulation

The alternative to formal verification is validation using extensive simulation and overnight testing of the day’s work using **regression testing**.

Tests can be **unit** tests or larger subsystems or complete system (H/W + S/W).

Can either write a RTL or ESL yes/no automaton as part of the test bench, or one can spool the outputs to file and **diff** against **golden** with PERL script.

Downfall of simulation: it’s non-exhaustive and time consuming.

ABD benefits in theory (and challenges in practice):

- We capture what the system is supposed to do,
- Completeness in theory (but how to define/determine this?),
- Scalability (but tools are limited in practice),
- Rare corner situations in the exponential state space (unusual conjunctions of events) are covered.

But simulations

- are needed for performance analysis and general design confidence,
- can generate some production **test vectors**.

- can be partly formal: using bus monitors for dynamic validation and Specman/VERA constrained pattern generators for stimulus.

Simulation is effective at finding many early bugs in a design. It can sometimes find safety violations and sometimes find (or accidentally encounter) deadlock but it cannot prove liveness.

Once the early, low-hanging bugs are fixed, formal proof can be more effective at finding the remainder. These tend to lurk in unusual corner cases, where particular alignment or conjunction of conditions is not handled correctly.

If a bug has a one in ten million chance of being found by simulation, then it will likely be missed, since fewer than that number clock cycles might typically be simulated in any run. However, given a clock frequency of just 10 MHz, the bug might show up in the real hardware in one second!

Simulation is generally easier to understand. Simulation gives performance results. Simulation can give a golden output that can be compared against a stored result to give a pass/fail result. A large collection of golden outputs is normally built up and the current version of the design is compared against them every night to spot **regressions**.

Simulation **test coverage** is expressed as a percentage. Given any set of simulations, only a certain subset of the states will be entered. Only a certain subset of the possible state-to-state transitions will be executed. Only a certain number of the disjuncts to the guard to an IF statement may hold. Only a certain number of paths through the block-structured behavioural RTL may be taken. Medical, defense and aerospace generally require much higher percentage coverage than commercial products.

There are many ways of defining coverage: for instance do we have to know the reachable state space before defining the state space coverage, or can we use all possible states as the denominator in the fraction? In general software, a common coverage metric is the percentage of lines of code that are executed.

Scaling of formal checking is a practical problem: today's tools certainly cannot check a complete SoC in one pass. An incremental approach based around individual sub-systems is needed.

### 5.1.3 Formally Synthesised Bus Monitor

A bus normally conforms to a well-established protocol and investment in a formal specification of the protocol is normally worthwhile. A bus monitor is a typical example of dynamic validation: it is a checker that flags protocol violations:

- safety violations are indicated straightaway,
- for a liveness property the monitor can indicate whether it has been tested at least once and also whether there is a pending antecedant that is yet to be satisfied.

For implementation in silicon, or if we are using an old simulator (e.g. a Verilog interpreter) that does not provide PSL or other temporal logic, the assertions can be compiled to an RTL checker automaton.

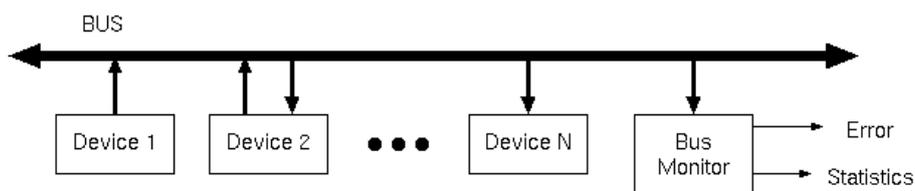


Figure 5.1: Dynamic validation: Monitoring bus operation with a hardware monitor.

A bus monitor connects to the net-level bus in RTL or silicon. (TLM formal monitoring is also being developed.)

The monitor can keep statistics as well as detect protocol violations.

Example of checker synthesis from a formal spec: [www.cl.cam.ac.uk/research/srg/han/hprls/orangepath/transactors](http://www.cl.cam.ac.uk/research/srg/han/hprls/orangepath/transactors) and Bus Monitors

#### 5.1.4 Is a formal specification complete ?

##### Additional notes:

Is a formal specification complete ?

- Does it fully-define an actual implementation (this is overly restrictive) ?
- Does it exactly prescribe all allowable, observable behaviours ?

By 'formal' we mean a machine-readable description of what is correct or incorrect behaviour. A **complete** specification might describe all allowable behaviours and prohibit all remaining behaviours, but most formal definitions today are not complete in this sense. For instance, a definition that consists of a list of safety assertions and a few liveness assertions might still allow all sorts of behaviours that the designer knows are wrong. He can go on adding more assertions, but when does he stop ?

One might define a 'complete specification' as one that describes all observable behaviours. Such a specification does not restrict or prescribe the internal implementation in **black box** terms since this is not observable.

When evaluating an assertion-based test program for an IP block, we can think of various, ad hoc, coverage metrics: e.g. What percentage of rule disjuncts (terms that are ORed) held as dominators (a term that makes the disjunction hold) on their own (without any other term in that disjunction holding) ? Or, e.g. What percentage of reachable state space was spanned? But there are no widely accepted such metrics in the industry, but the system Verilog 'cover' statement keeps statistics. System Verilog Assertions

#### 5.1.5 Assertion forms: State/Path, Concrete/Symbolic.

Many assertions are over **concrete state**. For instance '*Never is light A off when light B is on*'. Other assertions need to refer to **symbolic values**. For instance '*The value in register X is always less than the value in register Y*'.

**State properties** describe the current state only. For instance '*Light A is off and light B is on*'. **Path properties** relate successive state properties to each other. For instance '*light A always goes off before light B comes on*'.

We shall see PSL requires the symbolic values be embedded in the bottommost 'modelling layer' and that its temporal layer cannot deal with symbolic values. For instance, we cannot write ' $\{A(x); B(y)\} \mid \Rightarrow \{C(x, y)\}$ '.

(Note: the internal representation used by a checker tool for a concrete property can commonly use a symbolic encoding, such as a BDD, to handle an exponentially-large state space using reasonable memory, but that is orthogonal to the nature of the conditions being proved.)

Typically, for a bus or network, an important requirement is **data conservation**: what went in came out. Data conversation is a symbolic path property.

The DUT is the device under test. **Black Box testing** is where tests are phrased only in terms of the externally visible behaviour of DUT. **White Box testing** enables assertions to range over internal variables of the DUT.

### 5.1.6 Property Specification Language (PSL)

PSL is a linear-time temporal algebra designed for RTL engineering. The PSL language is now part of the System Verilog assertion language.

[www.project-veripage.com/psl\\_tutorial\\_2.php](http://www.project-veripage.com/psl_tutorial_2.php)

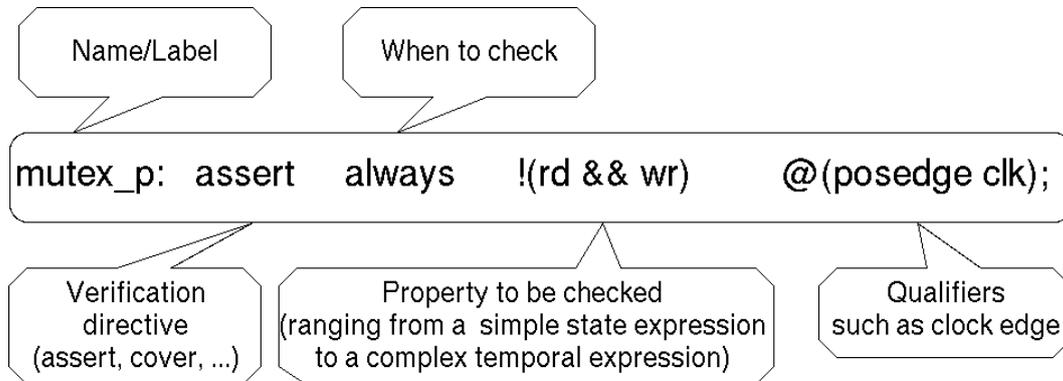


Figure 5.2: General structure of a PSL assertion

As in most temporal logics, there are three main directives:

1. `always` and `never`,
2. `next`,
3. `eventually!`

The `always` directive is the most frequently used and it specifies that the following property expression should be checked every clock. The `never` directive is a shorthand for a negated `always`.

The `next` directive relates successive state properties, as qualified by the clocking event and qualifying guard.

The `eventually!` directive is for liveness properties that relate to the future. The `eventually!` directive is suffixed with a bang sign to indicate it is strong property that cannot be (fully) checked with simulation.

For hands-on experience, see a previous ACS exercise: Dynamic validation using Monitors/Checkers and PSL

The general structure of a PSL assertion has the following parts:

- A name or label that can be used for diagnostic output.
- A verification directive, such as **`assert`**.
- When to check, such as **`always`** or **`eventually!`**.
- The property to be checked: a state expression or a temporal logic expression.
- A qualifying guard, such as a clock edge or enable signal at which time we expect the assertion to hold.

The `'assert'` keyword can be replaced with `'cover'` and instead of errors being reported statistics on hold and fail are accumulated.

### 5.1.7 ABD - PSL Four-Level Syntax Structure

The abstract syntax of PSL uses four levels:

- Since the language is embedded in the concrete syntax of several other languages, such as Verilog, SystemVerilog and VHDL, its syntactic details vary. In particular, creating state predicates involves expressions that range over the nets and variables of the host language. The precise means for this is defined by the **MODELLING LAYER** that allows one to create state properties using RTL.

Non-boolean, symbolic sub-expressions can be used in the modelling layer to generate boolean state predicates.

```
assign tempok = temperature < 99;
```

- All high-level languages and RTLs have their own syntax for boolean operators and this can be used within the modelling layer. However boolean combinations can also be formed using the PSL **BOOLEAN LAYER**.

```
not (rd && wr); // rd, wr are nets in the RTL (modelling layer).
```

- The PSL **TEMPORAL LAYER** allows one to define named sub-expressions and properties that use the temporal operators. For example:

```
// Sequence definition
sequence s1 is {pkt_sop; (!pkt_xfer_en_n [*1 to 100]); pkt_eop};

sequence s2 is {pkt_sop; (!pkt_xfer_en_n [*1 to 100]); pkt_aborted};

// Property definition
property p1 is reset_cycle_ended |=> {s1; s2};

// Property p1 uses previously defined sequences s1 and s2.
```

- The PSL **VERIFICATION LAYER** implements the declarative language itself. It includes the main keywords, such as **assert**.

PSL has a rich regular expression syntax for pattern matching. These are called *SERES* or sequences. SERES stands for Sugar Extended Regular Expression, where Sugar was an older name for PSL.

Sequence elements are state properties from Modelling and Boolean layers. Core operators are (of course): disjunction, concatenation and arbitrary repetition. As a temporal logic: interpret concatenation as a time sequencing.

- **A;B** Semicolon denotes sequence concatenation
- **A[\*]** Postfix asterisk for arbitrary repetition
- **A|B** Vertical bar (stile) for alternation.

Make easier to use with additional operators defined in terms of primitives:

- **A[+]** One or more occurrences: **A;A[\*]**
- **A[\*n]** Repeat n times
- **A[=n]** Repeat n times non-consecutively
- **A:B** Fusion concatenation (last of A occurs during first of B)

Further repetition operators denote repeat count ranges. Repeat counts must be compile-time constant (for today's standard/tools).

### 5.1.8 ABD - PSL Properties and Macros

PSL defines some simple path to state macros

- `rose(X)` means  $!X; X$
- `fell(X)` means  $X;!X$

Others are easy to define:

- `stable(X)` can be defined as  $X; X \mid \mid !X; !X$
- `changed(X)` can be defined as  $X;!X \mid \mid !X; X$
- `onehot(X)` can be defined as **X is a power of 2**
- `onehot0(X)` can be defined as **onehot(X)  $\mid \mid$  (X==0)**

### 5.1.9 ABD - Naive Path to State Conversion

Compiling regular expressions to RTL is completely straightforward (part of a typical proof that for every RE there is an FSM).

By converting a path expression to a state expression we can generate an RTL checker for use in dynamic validation. It can also be used for converting all path expressions to state expressions if the core of a proof tool can only handle state expressions, such as a raw BDD package or SAT solver.

**Additional notes:**

The following ML fragment handles the main operators: concatenation, fusion concatenation, alternation, arbitrary repetition and n-times repetition.

```

fun gen_pattern_matcher g (seres_state e) = gen_and2(g, gen_boolean e)

| gen_pattern_matcher g (seres_diop(diop_seres_alternation, l, r)) =
  let val l' = gen_pattern_matcher g l
      val r' = gen_pattern_matcher g r
  in gen_or2(l', r') end

| gen_pattern_matcher g (seres_diop(diop_seres_catenation, l, r)) =
  let val l' = gen_dff(gen_pattern_matcher g l)
      val r' = gen_pattern_matcher l' r
  in r' end

| gen_pattern_matcher g (seres_diop(diop_seres_fusion, l, r)) =
  let val l' = gen_pattern_matcher g l
      val r' = gen_pattern_matcher l' r
  in r' end

| gen_pattern_matcher g (seres_monop(mono_arb_repetition, l)) =
  let val nn = newnet()
      val l' = gen_pattern_matcher nn l
      val r = gen_or2(l', g)
      val _ = gen_buffer(nn, r)
  in r end

| gen_pattern_matcher g (seres_diop(diop_n_times_repetition, l,
  seres_state(x_num n))) =
  let fun f (g, k) = if k=0 then g else
      gen_pattern_matcher (f(g, k-1)) l
  in f (g, n) end

```

This generates a simple one-hot automaton and there are far more efficient procedures used in practice and given in the literature.

A harder operator to compile is the length-matching conjunction (introduced shortly), since care is needed when each side contains arbitrary repetition and can declare success or failure at a number of possible times.

### 5.1.10 ABD - SERES Pattern Matching Example

Suppose four events are supposed to always happen in sequence:

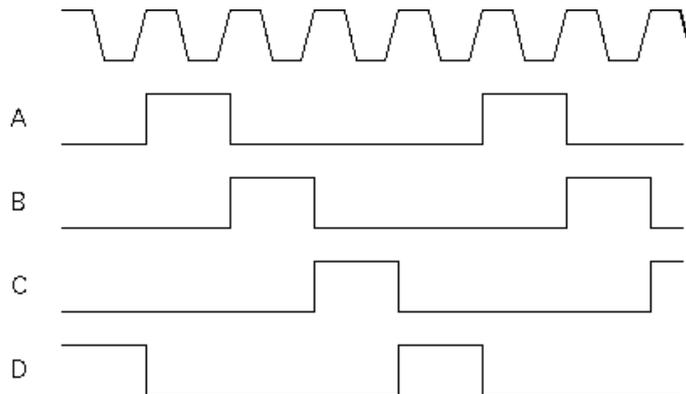
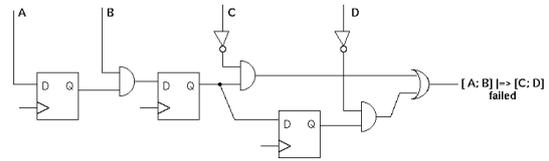


Figure 5.3: Desired behaviour of the four nets.

First attempt, we write **always true[\*]; A; B; C; D** Basic pattern matcher applied to **A;B;C;D** generates:





```

DFF(g0, A, clk);
AND2(g1, g0, B);
DFF(g2, g1, clk);
INV(g3, C);
AND2(g4, g3, g2); // Holds if C missing
DFF(g5, g2, clk);
INV(g6, D);
AND2(g7, g5, g6); // Holds if D missing
OR2(g8, g7, g4);
> val it = x_net "g8" : hexp_t // Holds on error

```

Even this is not very specific: C and D might occur at other times. It is a good idea to write protocol rules as suffix implications that range over SERES. Use a separate temporal implication for each sequential step.

What about asserting a requirement of **data conservation**? At an interface we commonly want to assert that data is not lost or duplicated. Is PSL any help? Not really, one needs a language that can range over symbolic data and tagged streams of data.

### 5.1.13 ABD - Boolean Equivalence Checker

Boolean equivalence: do the two functions produce the same output?

- For all input combinations?
- For a subset of input combinations (some input patterns are don't cares).

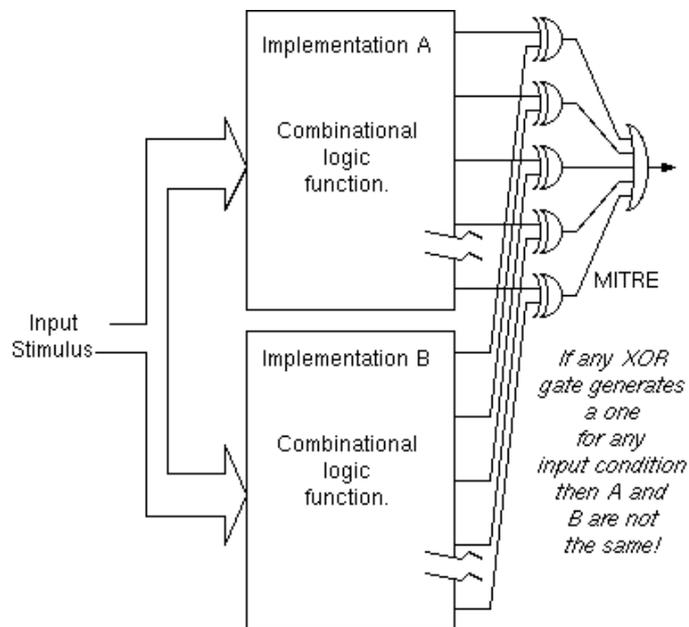


Figure 5.4: A mitre compares the outputs from a pair of supposedly-equivalent combinational components.

Often we have two implementations to check for equivalence, for instance, when RTL is turned into a gate-level netlist by synthesis we have:

- RTL version: pre-synthesis, and

- Gate-level version: post-synthesis.

Sources of difference between the designs might be manual implementation of one of them, manual edits to synthesiser outputs and EDA tool faults. For instance, after place and route operations, it is common to extract the netlist out from the masks and check that for correctness, so this is another source of the same netlist.

The **boolean equivalence problem** is do two functions produce the same output. However, are we interested for all input combinations? No, normally we are only interested in a subset of input combinations (because of don't care conditions).

The method, shown in Figure 5.4, is to create a **mitre** of the two designs using a disjunction of XOR gate outputs. Then, feed negation of mitre to a SAT solver to see if it can find any input condition that produces a one on the output.

SAT solving is a matter of trying all input combinations, so has exponential cost in theory and is NP complete. However, modern solvers such as **zChaff** essentially exploit the intrinsic structure of the problem so that they normally are quite quick at finding the answer.

Result: if there are no input combinations that make the mitre indicate a functionality difference, then the designs are equivalent.

Commercial example: Synopsys Formality

#### 5.1.14 Automated versus Manual Proof Tools

There are two main styles of mechanised proof tool:

- **Automatic Proof Tools** require no manual input. The main example is/was the model checker. Their grander descendant, the SAT-modulo checkers (SMT tool) expands the application space for automatic proof.
- **Theorem Provers** hold a database of axioms, logic rules and already derived theorems. Manual guidance is used to request which rule to apply to the current theorem set to produce further theorems. Eventually the required result can hopefully be found. A lot of semi-automatic application of rules is commonly supported (called tactics.) Previous proofs can be saved as tactics for replay in the current proof.

Contemporary, industrial SoC design only uses fully automated provers, whereas research in specification and verification often uses manually-guided provers where the computer may make suggestions about proof steps but its main role is to check the result has been derived without a false step.

Well-known model checkers are SMV and Spin. Yices and Z3 are SMT solvers.

Well-known theorem provers are HOL, Isabelle, ACL2 and Coq.

Cadence JasperGold is commonly used in industry and contains many tools in this space.

#### 5.1.15 A Simple Model Checker

For a small finite state machine we can use a simple model checker for a state safety property:

Algorithm: '**Find reachable state space**' (add successors of current set until closure):

1.  $S := \{q_0\}$  // initial state
2.  $S := S \cup \{q' \mid \exists \sigma \in \Sigma, q \in S. NSF(q, \sigma) = q'\}$
3. If safety property does not hold in any  $q \in S$  then flag error.
4. If S increased in step 2 then goto step 2.

S can be held explicitly in bit map form or symbolically as a BDD.

Variation 1: ignore safety property while finding reachable state space then finally check for all found states.

Variation 2: property to check might be a path property, so either

- Compile it to a checking automaton (becomes a state property of expanded NSF), or
- Expand it as we go (using modal mu calculus).

The PSL strong assertions need to be checked with a formal proof tool. Model checking is normally used because it is fully automated.

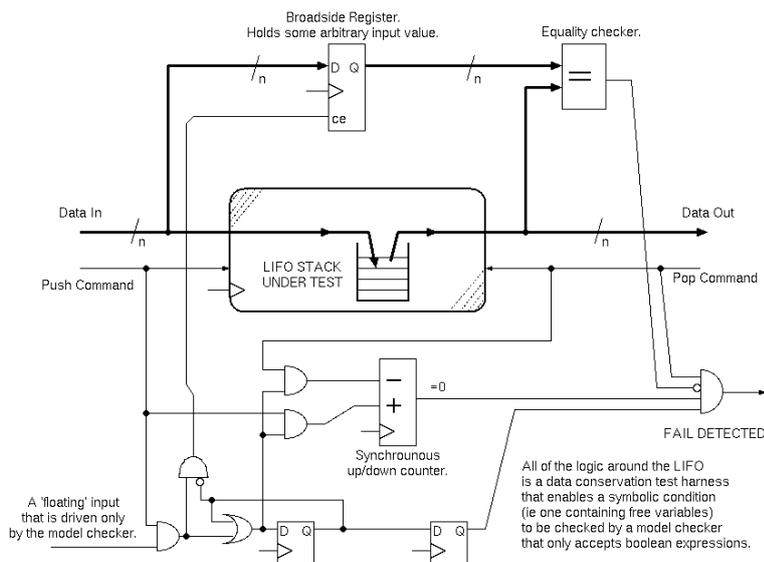
A model checker explores every possible execution route of a finite-state system by exploring the behaviour over all possible input patterns.

There are two major classes of model checker: explicit state and symbolic. Explicit state checkers actually visit every possible state and store the history in a very concise bit array. If the bit array becomes too big they use probabilistic and hashing techniques. The main example is Spin. Symbolic model checkers manipulate expressions that describe the reachable state space and these were famously implemented as BDDs in the SMV checker. There are also other techniques, such as bounded model checking, but the internal details of model checkers is beyond the scope of this course.

The most basic model checker only checks state properties. To check a path property it can be compiled into an automaton and included as part of the system itself.

To check liveness formally is beyond the scope of this course, but one algorithm is to repeatedly trim cul-de-sacs from the state transition graph so that only a core where all states are reachable from all others remains.

### 5.1.16 ABD - Model Checking a FIFO Queue and LIFO Stack



What is the equivalent circuit for a FIFO ?

If we simulate the circuit above we have to provide stimulus for the input data and push and pop command inputs. The harness built around the LIFO has a 'floating' input that would also need a stimulus waveform in a simulation. It checks that the word pushed first after its first +ve transition comes out again at the correct time. A simple state-based model checker tool applied to the above system will check correct operation for all words in and out even though only one holding register is present in the harness. (Thanks to Daryl Stewart of ARM for teaching this technique.) The condition for checking is that never does the fail output hold.

### 5.1.17 ABD - Sequential Logic Equivalence

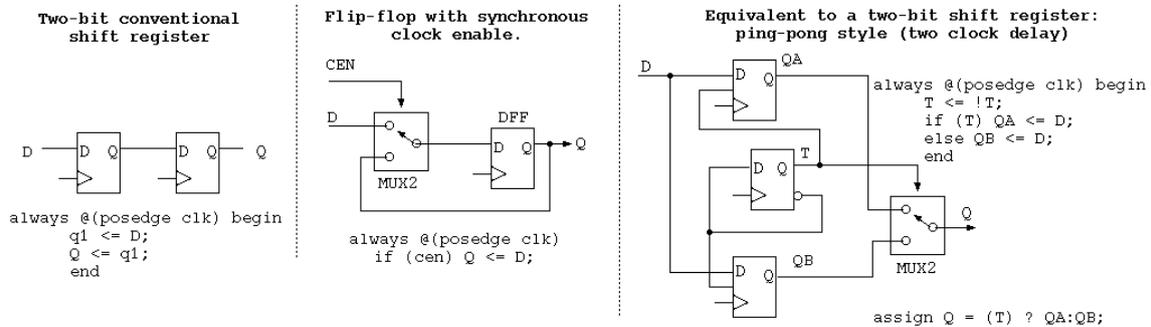


Figure 5.5: Two circuits that use different amounts of internal state to achieve the same functionality.

The figure shows implementations of a two-bit shift register. They differ in amount of internal state. They have equivalent observable behaviour (ignoring glitches). Note, to implement larger delays, the design based on multiplexors might use more logic and less power than the design based on shifting, since fewer nets toggle on each clock edge.

Another common question that needs checking is *sequential equivalence*. Do a pair of designs follow the same state trajectory ?

- Considering the values of all state variables ?
- Considering a re-encoding of the state variables ?
- For an observable subset of the state (e.g. at an interface) ?
- When interfacing with a given reactive automaton ?

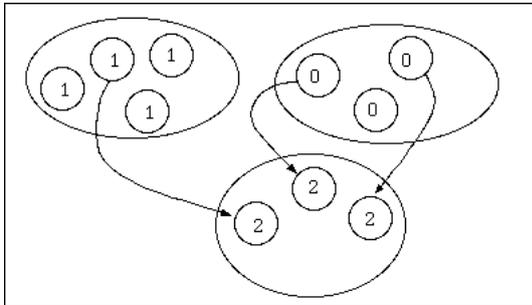
Other freedoms that could be allowed within the notion of equivalence:

- Temporally floating ports - latency independence. With floating ports we do not consider the relative timing of events between ports, only the relative timing of events within each port.
- Synchronous or asynchronous (turn-taking) composition. If a pair of circuits are combined, do they share a common clock or take it in turns to move?
- Strong or weak bi-simulation (stuttering equivalence). A stuttering equivalence between a pair of designs may exist if we disregard the precise number of clock cycles each took to achieve the result (such as different implementations of a microprocessor).

Practical problem: Designs may only be equivalent in the used portion of the state space. Hence we may need a number of side conditions that specify the required operating conditions.

### 5.1.18 ABD - Sequential Logic Simplification

A finite-state machine may have more states than it needs to perform its observable function because some states are totally equivalent to others in terms of output function and subsequent behaviour. Note that one-hot coding does not increase the reachable state space and so is **not** an example of that sort of redundancy.



A Moore machine can be simplified by the following procedure:

- 1. Partition all of the state space into blocks of states where the observable outputs are the same for all members of a block.
- 2. Repeat until nothing changes (i.e. until it closes) For each input setting:
  - 2a. Chose two blocks, B1 and B2.
  - 2b. Split B1 into two blocks consisting of those states with and without a transition from B2.
  - 2c. Discard any empty blocks.
- 3. The final blocks are the new states.

*Bisimulation algorithm not examinable in this course.*

Alternative algorithm: start with one partition per state and repeatedly conglomerate. The best algorithms use a mixture of the two approaches to meet in the middle. Wikipedia: Formal Equivalence Checking

Research example: CADP package: developed by the VASY team at INRIA. Commercial products: Conformal by Cadence, Formality by Synopsys, SLEC by Calypto.

One future use of this sort of procedure might be to generate an instruction set simulator for a processor from its full RTL implementation. This sort of de-pipelining would give a non-cycle accurate, higher-level model that runs much faster in simulation.

There are some good on-line resources. Such as Dulos System Verilog Assertions

### 5.1.19 Automated Stimulus Generation (Directed-Random Verification)

Commerical products: Verisity's Specman Elite [www.open-vera.com](http://www.open-vera.com)

Simulations and test programs require **stimulus**. This is a sequence of input signals, including clock and reset, that exercise the design.

Given that formal specifications for many of the input port protocols might exist, one can consider automatic generation of the stimulus, from random sources, within the envelope defined by the formal specification. Several commercial products do this, including Verisity's Specman Elite, Synopsys Vera.

Here is an example of some code in Specman's own language, called 'e', that defines a frame format used in networking. Testing will be inside envelope defined by **keep** statement.

```

struct LLCHeader { v: int( bits:2); 0: int(bits 14); }

struct frame {
  llc: LLCHeader;
  destAddr: uint (bits:48);
  srcAddr: uint (bits:48);
  size: int (bits:32);
  payload: list of byte;
  keep payload.size() in [0..size];
}

```

Sequences of bits that conform to the frame structure are created and presented at an input port of the design under test. An heirarchy of specifications and constraints is supported. One can compose and extend one specification to reduce its possible behaviours:

```

// Subclass the frame to make it more specialised:
extend frame { keep size == 0; };

```

### 5.1.20 OVM/UVM

The **Open Verification Methodology** (OVM) is an industry-wide methodology with a supporting building-block library for the systematic creation of documentation and verification harnesses for IP blocks. To conform, an IP block will be accompanied with a prescribed set of unit tests.

Doulos: From OVM to UVM

run OVM simulations from a web browser

Verification Methodology Cookbooks

*(The OVM/UVM topics are non-examinable for part II CST.)*

### 5.1.21 ABD - Conclusion

ABD today is often focussed on safety and liveness properties of systems and formal specifications of the protocols at the ports of a system. However, there are many other useful properties we might want to ensure or reason about, such as those involving counting and/or data conservation. These are less-well embodied in contemporary tools.

PSL deals with concrete values rather than symbolic values. Many interesting properties relate to symbolic data (e.g. specifying the correct behaviour of a FIFO buffer). Using PSL, all symbolic tokens must be wrapped up in the modelling layer which is not the core language.

Formal methods are taking over from simulation, with the percentage of bugs being found by formal methods growing. However, there is a lack of formal design entry. Low-level languages such as Verilog do not seamlessly mix with automatic synthesis from formal specification and so double-entry of designs is common.

## KG 6 — Architectural Exploration using ESL Modelling

A SoC combines hardware and software and communicates with the outside world via various interfaces. An ESL Model of a SoC is able to simulate the complete system behaviour. This includes running all of the software that the real SoC will run in a largely unmodified form. The acronym ESL stands for ‘Electronic System Level’. An alternative name for an ESL model is ‘Virtual Platform’.

The performance of an ESL model must be good enough to execute large programs in a reasonable time. This typically means achieving at least one percent of real system performance. An ESL model is normally accurate in terms of memory layout and content, but many other hardware details are commonly neglected unless they are of special relevance to the test being run. This is the principle means of gaining high performance.

Another way to apply an ESL model to complex software is ‘checkpoint and replay’. This is useful if a lot of software must run before the point of interest is approached. A checkpoint is chosen, such as the point after boot and operating system start. At the checkpoint, all of the state from the model is saved to a checkpoint file. Information could be captured from the real system in principle, but the ESL model may not be identical to the real system and minor discrepancies may arise, or instrumenting the real system may be tricky (especially if it does not exist yet). Since the checkpoint serves as the basis for a number of experiments, time invested in generating it is amortised.

To conduct an experiment, the ESL model is loaded with the checkpoint data and modelling rolls forward from that point. The model may be switched to a greater level of detail than used for preparing the checkpoint, either globally or just for some sub-systems. For instance a high-level model of an I/O block may be switched to an RTL model.

We can model our hardware system at various levels of detail following the taxonomy:

- **Functional Modelling:** The ‘output’ from a simulation run is accurate.
- **Memory Accurate Modelling:** The contents and layout of memory is accurate.
- **Untimed TLM:** No time stamps recorded on transactions.
- **Loosely-timed TLM:** The number of transactions is accurate, but order may be wrong.
- **Approximately-timed TLM:** The number and order of transactions is accurate.
- **Cycle-Accurate Level Modelling:** The number of clock cycles consumed is accurate.
- **Event-Level Modelling:** The ordering of net changes within a clock cycle is accurate.

An ESL methodology aims:

Aim 1: To model with good performance a complete SoC using full software/firmware.

Aim 2: To allow seamless and successive replacement of high-level parts of the model with low-level models/implementations when available and when interested in their detail.

So, an ESL methodology must provide:

- Tangible, lightweight **rapidly-generated prototype** of full SoC architecture.
- **Rapid Architectural Evaluation:** determine bus bandwidth and memory use for a candidate architecture. Easy to adjust major design parameters.
- **Algorithmic Accuracy:** Get real output from an early system, hosting the real application/firmware, possibly in real-time.

- **Timing information:** Get timing numbers for performance (accurate or loose timing).
- **Power information:** Get power consumption estimates to evaluate chip temperature and system battery life.
- **Firmware development:** Integrate high-level behavioural models of major components with their device drivers to run test software and applications before tape-out.

A commonly used method is SystemC Transactional-Level Modelling (TLM) using high-level C++ models running over the SystemC event-driven kernel. Enhancements beyond that are:

- Synthesise high-level models to form parts of the fabricated system (e.g. using HLS)(but today manual re-coding is mainly used).
- Embed assertions in the high-level models and use these same assertions through to tape out (Section 5 (not lectured this year)).

### 6.0.22 SystemC: Hardware Modelling Library Overview

SystemC is a free library for C++ for hardware SoC modelling. Download from [www.accelera.org](http://www.accelera.org) SystemC was developed over the last fifteen years with three major releases. Also of significance is the TLM coding style 1.0 and sub-library, TLM 2.0.

It includes (at least):

- A module description system where a module is a C++ class,
- An eventing and threading kernel,
- Compute/commit signals as well as other forms of channel,
- A library of fixed-precision integers,
- Plotting and logging facilities for generating output,
- A transactional modelling (TLM) sub-library.

Greaves developed the TLM\_POWER3 add-on library for power modelling.

Originally aimed as an RTL replacement, for low-level hardware modelling. Now being used for high-level (esp. transactional) modelling for **architectural exploration**. Also sometimes used as an implementation language with its own synthesis tools. SystemC Synthesis

Problem: hardware engineers are not C++ experts but they can be faced with confusing C++ error messages.

Benefit: General-purpose behavioural C code, including application code and device drivers, can all be modelled in a common language.

SystemC can be used for detailed net-level modelling, but today its main uses are:

- Architectural exploration: Making a fast and quick, high-level model of a SoC to explore performance variation against various dimensions, such as bus width and cache memory size.
- Transactional-level (TLM) models of systems, where handshaking protocols between components using hardware nets are replaced with subroutine calls between higher-level models of those components.
- Synthesis: RTL can be synthesised from from SystemC source code using High-Level Synthesis. SystemC Synthesis

**Additional notes:**

On the course web site, there is information on two sets of practical experiments:

- **Simple TLM 1 style:** To help investigate the key aspects of the transactional level modelling (TLM) methodology without using extensive libraries of any sort we use our own processor, the almost trivial nominalproc, and we cook our own transactional modelling library.

This practical takes an instruction set simulator of a nominal processor and then sub-class it in two different ways: one to make a conventional net-level model and the other to make an ESL version. The nominal processor is wired up in various different example configurations, some using mixed-abstraction modelling.

- **TLM 2 style:** Using the industry standard TLM 2.0 library and the Open Cores OR1K processor. This is ultimately easier to use, but has a steeper learning curve.

In this course we shall focus on the loosely-timed, blocking TLM modelling style of ESL model.

## 6.1 ESL Flow Model: Avoiding ISS/RTL overheads using native calls.

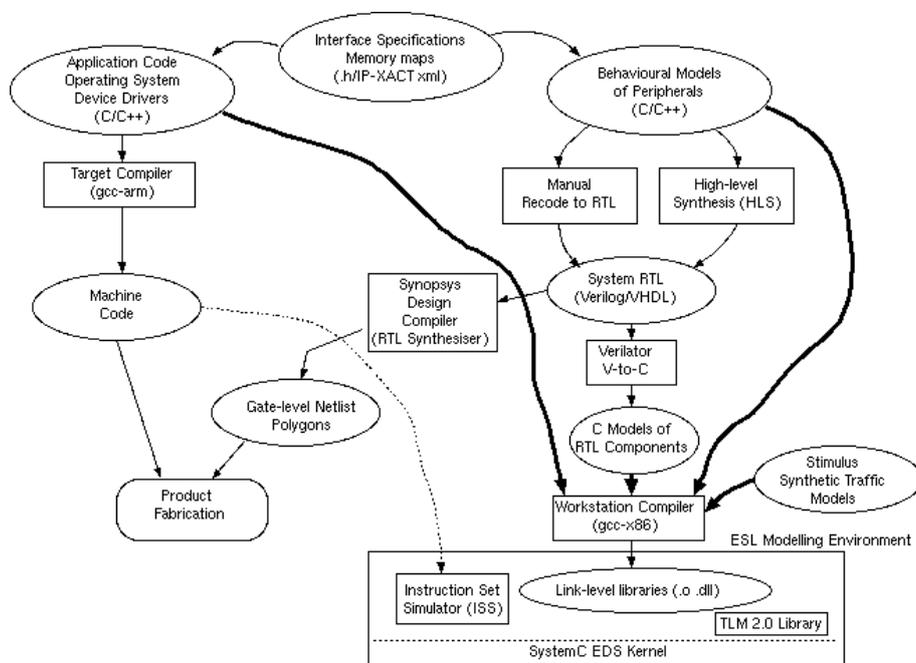


Figure 6.1: ESL Flow: Avoiding the ISS by cross-compiling the firmware and direct linking with behavioural models.

ESL flows are commonly based on C/C++. This language is used for behavioural models of the peripherals and for the embedded applications, operating system and device drivers. For fabrication, the embedded software is compiled with the target compiler (e.g. gcc-arm) and RTL is converted to gates and polygons using Synopsys Design Compiler. For ESL simulation, as much as possible, we take the original C/C++ and link it all together, whether it is hardware or software, and run it over the SystemC event-driven simulation (EDS) kernel.

Variations: sometimes we can import RTL components using a tool such as Verilator or VTOC. Sometimes we use an ISS to interpret (or JIT) the target processor machine code.

### 6.1.1 Using C Preprocessor to Adapt Firmware

We may need to recompile the hardware/software interface when compiling for ESL model as compared to the device driver installed in an OS or ROM firmware. For a 'mid-level model', differences might be minor and so implemented in C preprocessor. Device driver access to a DMA controller might be changed as follows:

```
#define DMACONT_BASE      (0xFFFFCD00) // Or other memory map value.
#define DMACONT_SRC_REG   0
#define DMACONT_DEST_REG  4
#define DMACONT_LENGTH_REG 8           // These are the offsets of the addressable registers
#define DMACONT_STATUS_REG 12

#ifdef ACTUAL_FIRMWARE

    // For real system and lower-level models:
    // Store via processor bus to DMACONT device register
    #define DMACONT_WRITE(A, D)    (*(DMACONT_BASE+A*4) = (D))
    #define DMACONT_READ(A)       (*(DMACONT_BASE+A*4))
#else

    // For high-level TLM modelling:
    // Make a direct subroutine call from the firmware to the DMACONT model.
    #define DMACONT_WRITE(A, D)    dmaunit.slave_write(A, D)
    #define DMACONT_READ(A)       dmaunit.slave_read(A)
#endif

// The device driver will make all hardware accesses to the unit using these macros.
// When compiled native, the calls will directly invoke the behavioural model, bypassing the bus model.
```

DMA Controller RTL Version (from 2016 SoC Parts Slide Pack)

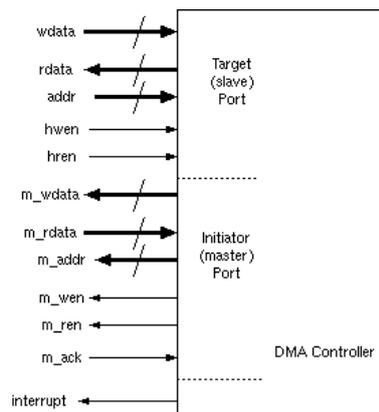


Figure 6.2: Block Diagram of one-channel DMA unit.

Behavioural model example (the one-channel DMA controller):

```

// Behavioural model of
// slave side: operand register r/w.
uint32 src, dest, length;
bool busy, int_enable;

u32_t status() { return (busy << 31)
    | (int_enable << 30); }

u32_t slave_read(u32_t a)
{
    return (a==0)? src: (a==4) ? dest:
        (a==8) ? (length) : status();
}
void slave_write(u32_t1 a, u32_t d)
{
    if (a==0) src=d;
    else if (a==4) dest=d;
    else if (a==8) length = d;
    else if (a==12)
    { busy = d >> 31;
      int_enable = d >> 30; }
}

```

```

// Bev model of bus mastering portion.
while(1)
{
    waituntil(busy);
    while (length-- > 0)
        mem.write(dest++, mem.read(src++));
    busy = 0;
}

```

We would like to make interrupt output with an RTL-like continuous assignment:

```
interrupt = int_enable&!busy;
```

But this will need a thread to run it, so this code must be placed in its own C macro that is inlined at all points where the supporting expressions might change.

## 6.2 Transactional-Level Modelling (TLM)

Recall our list of three inter-module communication styles, we will now consider the third style:

1. **Pin-level modelling:** an event is a change of a single-bit or multi-bit net,
2. **Abstract data modelling:** an event is delivery of a large data packet, such as a complete cache line,
3. **Transactional-level modelling:** avoid events as much as possible: use intermodule software calling.

(Actually, the second style was not lectured this year, but it's where an EDS event conveys a large struct instead of the new value for a single net.)

In general use, a *transaction* has atomicity, with commit or rollback. But in ESL the term means less than that. In ESL we might just mean that a thread from one component executes a method on another. However, the call and return of the thread normally achieve flow control and implement the atomic transfer of some datum, so the term remains retains some dignity.

We can have blocking and non-blocking TLM coding styles:

- **Blocking:** Hardware flow control signals implied by thread's call and return.
- **Non-blocking:** Success status returned immediately and caller must poll/retry as necessary.

In SystemC: blocking requires an SC\_THREAD, whereas non-blocking can use an SC\_METHOD. (*CST: Non-examinable 15/16 onwards.*)

Which is better: a matter of style? Non-blocking enables finer-grained concurrency and closer to cycle-accurate timing results. TLM 2.0 sockets will actually map between different styles at caller and callee.

Also, there are two standard methods for timing annotation in TLM modelling, **Approximately-timed** and **Loosely-timed** and in these notes we shall emphasize the latter.

Figure 6.3 is an example protocol implemented at net-level and TLM level:

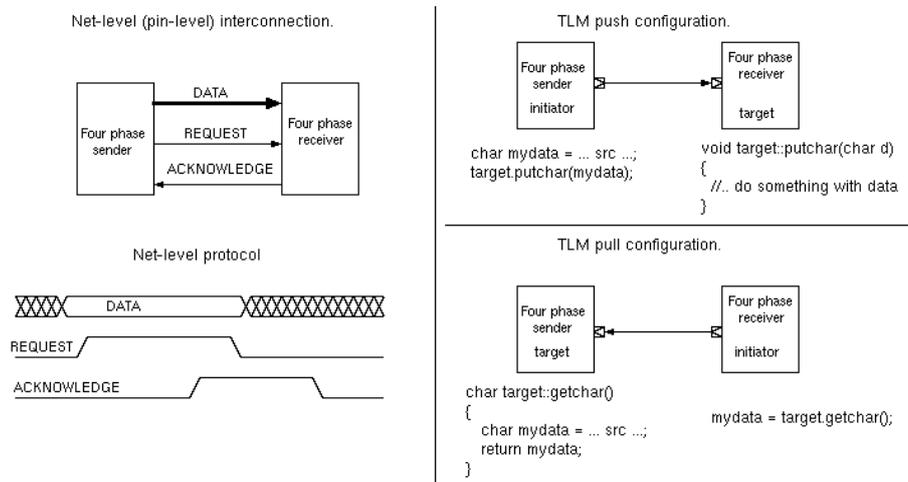


Figure 6.3: Three views of four-phase handshake between sender and receiver: net-level connection and TLM push and TLM pull configurations (untimed).

Note that the roles of initiator and target do not necessarily relate to the sources and sinks of the data. In fact, an initiator can commonly make both a read and a write transaction on a given target and so the direction of data transfer is dynamic. Perhaps try the practical materials: ultra-simple SystemC implementation ‘Toy ESL’

### 6.2.1 General ESL Interactions with Shortcuts Illustrated

Consider the Ethernet CRC Example

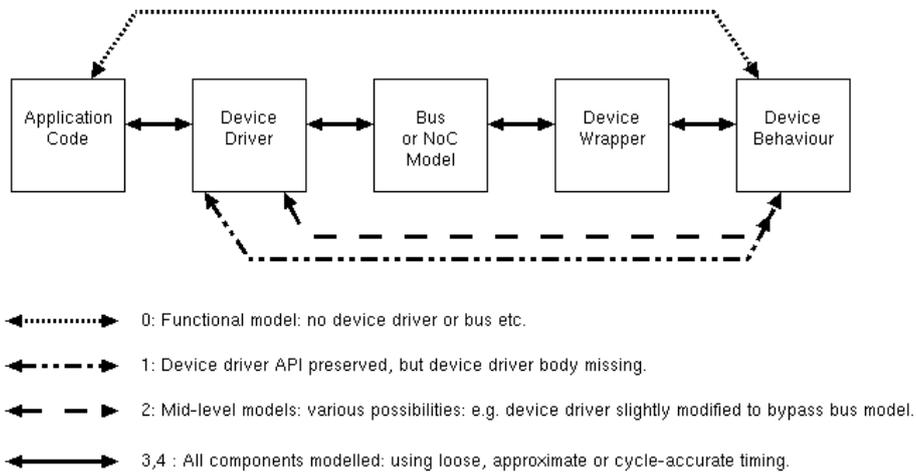


Figure 6.4: Some possible shortcuts through full system model to omit details.

Another view of the higher modelling abstractions:

1. Highest-level (vanished) model: Retains algorithmic accuracy: implemented using SystemC or another threads package: device driver code and device model mostly missing, but the **API to the device driver is preserved**, for instance, a single TLM transaction might send a complete packet when in reality multiple bus cycles are needed to transfer such a packet;

2. Mid-level model: Implemented using SystemC: the device driver is only slightly modified (using preprocessor directives or otherwise) but the interconnection between the device and its driver may be different from reality, meaning bus utilisation figures are unobtainable or incorrect;
3. Bus-transaction accurate mode: each bus operation (read/write or burst read/write and interrupt) is modelled, so bus loading can be established, but timing may be loose and transaction order may be wrong, again, minor changes in the device driver and native compilation may be used;
4. Lower-level models: Implemented in RTL or cycle-accurate SystemC: target device driver firmware and other code is used unmodified.

Point 3 encompasses mainstream TLM models, like Prazor Virtual PlatformManual

### 6.2.2 Mixing modelling styles: 4/P net-level to TLM transactors.

An aim of ESL modelling was to be able to easily replace parts of the high-level model with greater detail where necessary. So-called **transactors** are commonly needed at the boundaries.

Here is an example blocking transactor. It forms a gateway from a transactional initiator to a pin-level target.

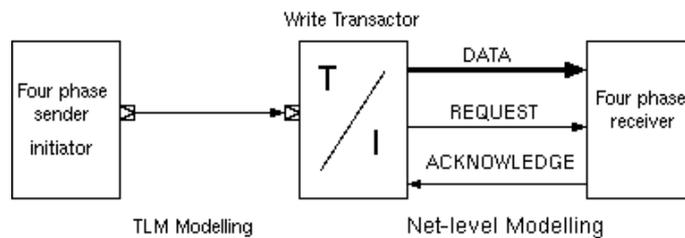


Figure 6.5: Mixing modelling styles using a transactor.

<pre>// Write transactor 4/P handshake b_putbyte(char d) {     while(ack) do wait(10, SC_NS);     data = d;     settle();     req = 1;     while(!ack) do wait(10, SC_NS);     req = 0; }</pre>	<pre>// Read transactor 4/P handshake char b_getbyte() {     while(!req) do wait(10, SC_NS);     char r = data;     ack = 1;     while(req) do wait(10, SC_NS);     ack = 0;     return r; }</pre>
---	--

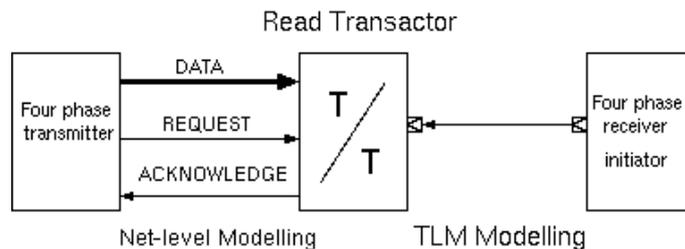


Figure 6.6: Mixing modelling styles using a transactor 2.

### 6.2.3 Transactor Configurations

Four possible transactors are envisionsable for a single direction of the 4/P handshake and in general.

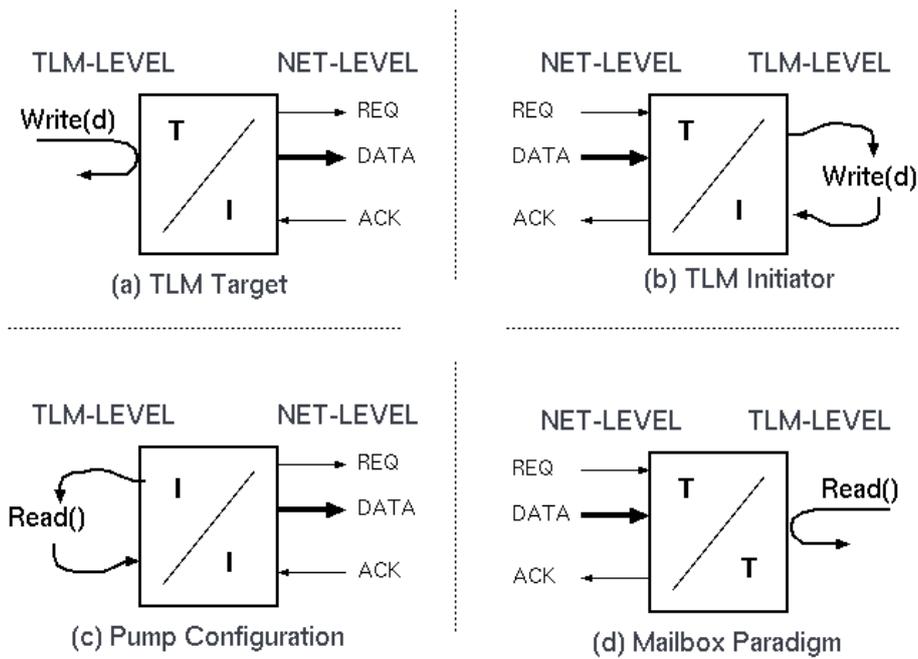


Figure 6.7: Possible configurations for simple transactors.

**Additional notes:**

An (ESL) Electronic System Level *transactor* converts from a hardware to a software style of component representation. A hardware style uses shared variables to represent each net, whereas a software style uses callable methods and up-calls. Transactors are frequently required for busses and I/O ports. Fortunately, formal specifications of such busses and ports are becoming commonly available, so synthesising a transactor from the specification is a natural thing to do.

There are four forms of transactor for a given bus protocol. Either side may be an initiator or a target, giving four possibilities.

A transactor tends to have two ports, one being a net-level interface and the other with a thread-oriented interface defined by a number of method signatures. The thread-oriented interface may be a target that accepts calls from an external client/initiator or it may itself be an initiator that makes calls to a remote client. The calls may typically be blocking to implement flow control.

The initiator of a net-level interface is the one that asserts the command signals that take the interface out of its starting or idle state. The initiator for an ESL/TLM interface is the side that makes a subroutine or method call and the target is the side that provides the entry point to be called.

Consider a transactor with a 'Read()' target port and net-level parallel input. This is an alternative generalisation of the (a) configuration but for when data is moving in the opposite direction. Considering the general case of a bi-directional net-level port with separate TLM entry points for 'Read()' and 'Write(d)' helps clarify.

**6.2.4 Example of non-blocking coding style:**

Example: Non-blocking (untimed) transactor for the four-phase handshake.

```

bool nb_putbyte_start(char d)
{
    if (ack) return false;
    data = d;
    settle(); // A H/W delay for skew issues,
             // or a memory fence in S/W for
             // sequential consistency.
    req = 1;
    return true;
}

bool nb_putbyte_end(char d)
{
    if (!ack) return false;
    req = 0;
    return true;
}

```

```

bool nb_getbyte_start(char &r)
{
    if (!req) return false;
    r = data;
    ack = 1;
    return true;
}

bool nb_getbyte_end()
{
    if (req) return false;
    ack = 0;
    return true;
}

```

Both routines should be repeatedly called by the client until returning true.

### 6.2.5 ESL TLM in SystemC: First Standard TLM 1.0.

```

class my_component: public sc_module, ethernet_if, usb_if
{
    // SC_METHODs and SC_THREADS for normal internal behaviour
    ...

    // methods to implement ethernet_if
    ...

    // methods to implement usb_if
    ...

    // Constructor and sensitivity
    ...
}

```

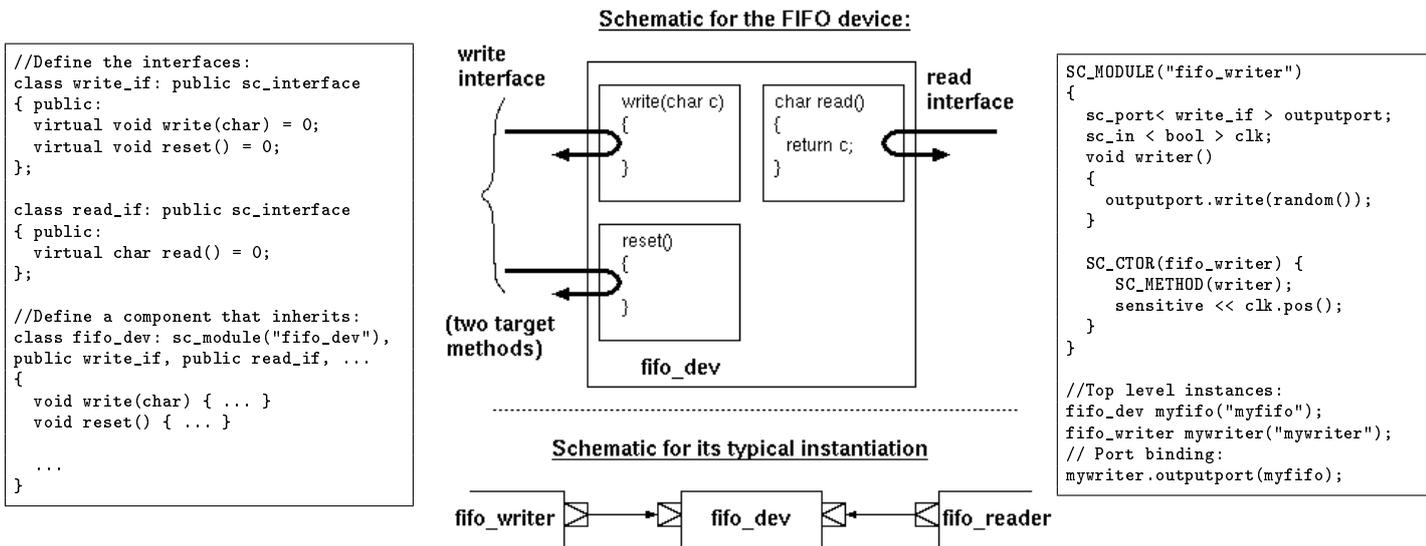
The OSCI TLM 1.0 standard used conventional C++ concepts of multiple inheritance. As shown in the ‘Toy ESL’ materials and the example here, an SC\_MODULE that implements an interface just inherits it.

SystemC 2.0 implemented an extension called `sc_export` that allows a parent module to inherit the interface of one of its children. This was a vital step needed in the common situation where the exporting module is not the top-level module of the component being wired-up.

However, TLM 1.0 had no standardised or recommended structure for payloads and no standardised timing annotation mechanisms. There was also the problem of how to have multiple TLM ports on a component with same interface: e.g. a packet router. This is not often needed for software, and hence omitted from high-level languages like C++, but it is common for hardware designs. (A work-around is to add a dummy formal parameter that is given a different concrete type in each instance of an interface ...)

However, referring back to the DMA unit behavioural model (see examples sheet), we can see that that memory operations are likely to get well out of synchronisation with the real system since this copying loop just goes as fast as it can without worrying about the speed of the real hardware. It is just governed by the number of cycles the read and write calls block for, which could be none. The whole block copy might occur in zero simulation time! This sort of modelling is useful for exposing certain types of bugs in a design, but it does not give useful performance results. We shall shortly see how to limit the sequential inconsistencies using a quantum keeper.

A suitable coding style for sending calls ‘*along the nets*’ (prior to the TLM 2.0 standard):



Here a thread passes between modules, but modules are plumbed in Hardware/EDS netlist structural style. Although sometimes called a 'standard', it is really an ad-hoc coding style.

See the slide for full details, but the important thing to note is that the entry points in the interface class are implemented inside the FIFO device and are bound, at a higher level, to the calls made by the writer device. This kind of plumbing of upcalls to endpoints formed an essential basis for future transactional modelling styles.

### 6.2.6 ESL TLM in SystemC: TLM 2.0

Although there was a limited capability in SystemC 1.0 to pass threads along channels, and hence do subroutine calls along what look like wire, this was made much easier SystemC 2.0. TLM2.0 (July 2008) tidies away the TLM1.0 interface inheritance using **convenience sockets** and defines the **generic payload**. It also defined memory/garbage ownership and transport primitives with timing and fast backdoor access to RAM models. And it provided a raft of useful features, such as automatic conversion between blocking and non-blocking styles.

```
// Filling in the fields of a TLM2.0 generic payload:
trans.set_command(tlm::TLM_WRITE_COMMAND);
trans.set_address(addr);
trans.set_data_ptr(reinterpret_cast<unsigned char*>(&data));
trans.set_data_length(4);
trans.set_streaming_width(4);
trans.set_byte_enable_ptr(0);
trans.set_response_status( tlm::TLM_INCOMPLETE_RESPONSE );

// Sending the payload through a TLM socket:
socket->b_transport(trans, delay);
```

Other standard payloads (e.g. 802.3 frame or audio sample) might be expected to have become widely used by now.

Rather than having application-specific method names, we standardise on a generic bus operation and demultiplex within various IP blocks based on register address.

The generic payload can be extended on a custom basis and intermediate bus bridges and routers can be polymorphic about this: not needing to know about all the extensions but able to update timestamps to model routing delays.

Let's consider a small SRAM example: first define the socket in the .h file:

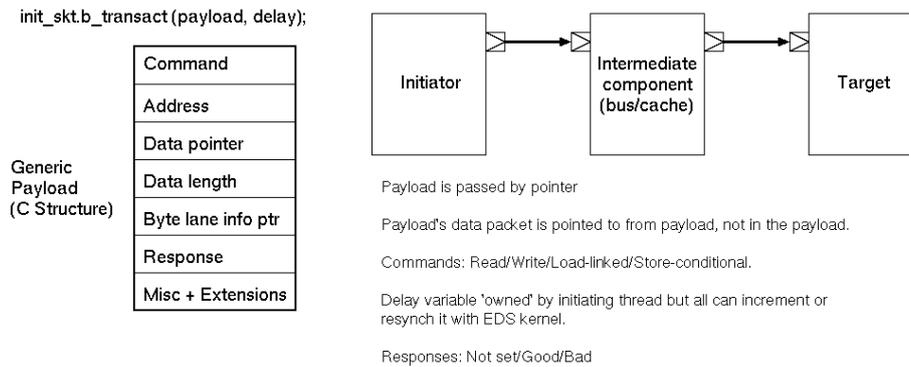


Figure 6.8: The general TLM 2.0 Setup

```

SC_MODULE(cbgram)
{
    tlm_utils::simple_target_socket<cbgram> port0;
    ...
}
  
```

Here is the constructor:

```

cbgram::cbgram(sc_module_name name, uint32_t mem_size, bool tracing_on, bool dmi_on): sc_module(name), port0("port0"),
    latency(10, SC_NS), mem_size(mem_size), tracing_on(tracing_on), dmi_on(dmi_on)
{
    mem = new uint8_t [mem_size]; // allocate memory
    // Register callback for incoming b_transport interface method call
    port0.register_b_transport(this, &cbgram::b_transact);
}
  
```

And here is the guts of b\_transact:

```

void cbgram::b_transact(tlm::tlm_generic_payload &trans, sc_time &delay)
{
    tlm::tlm_command cmd = trans.get_command();
    uint32_t adr = (uint32_t)trans.get_address();
    uint8_t * ptr = trans.get_data_ptr();
    uint32_t len = trans.get_data_length();
    uint8_t * lanes = trans.get_byte_enable_ptr();
    uint32_t wid = trans.get_streaming_width();

    if (cmd == tlm::TLM_READ_COMMAND)
    {
        ptr[0] = mem[adr];
    }
    else ...

    trans.set_response_status( tlm::TLM_OK_RESPONSE);
}
  
```

Wire up the ports in the level above:

```

busmux0.init_socket.bind(memory0.port0);
  
```

The full code, and many other examples, can be found in the Prazor virtual platform (see for example `vhls/src/memories/sram`).

Sockets of the 'multi' style can be multiply bound and so provide a mux or demux function. Sockets of 'passthrough' style enable a generic payload reference to be passed on.

Figure 6.9: Typical setup showing four socket types.

**Additional notes:**

TLM 2.0 Socket Types:

**simple\_initiator\_socket.h** version of an initiator socket that has a default implementation of all interfaces and allows to register an implementation for any of the interfaces to the socket, either unique interfaces or tagged interfaces (carrying an additional id)

**simple\_target\_socket.h** version of a target socket that has a default implementation of all interfaces and allows to register an implementation for any of the interfaces to the socket, either unique interfaces or tagged interfaces (carrying an additional id) This socket allows to register only 1 of the transport interfaces (blocking or non-blocking) and implements a conversion in case the socket is used on the other interface

**passthrough\_target\_socket.h** version of a target socket that has a default implementation of all interfaces and allows to register an implementation for any of the interfaces to the socket.

**multi\_passthrough\_initiator\_socket.h** an implementation of a socket that allows to bind multiple targets to the same initiator socket. Implements a mechanism to allow to identify in the backward path through which index of the socket the call passed through

**multi\_passthrough\_target\_socket.h** an implementation of a socket that allows to bind multiple initiators to the same target socket. Implements a mechanism to allow to identify in the forward path through which index of the socket the call passed through

The user manuals for TLM 2.0 are linked here [ACS P35 Documents Folder](#)

### 6.2.7 CSharp Implementation

Here is an approximate skeleton of the structure of the sockets in C#.

```

// Skeleton outline of the TLM2 socket structure for blocking transport only.
// This is a skeleton in that the other socket and channel types are missing and b_transact has no payload.
class TLM2_Sockets_Skeleton
{
    public interface b_transimple_target_socket_callback_t
    {
        void b_transact();
    }

    public class simple_initiator_socket
    {
        simple_target_socket who;

        public void bind(simple_target_socket whom) { who = whom; }

        public void b_transact() { who.b_transact(); }
    }

    public class simple_target_socket
    {
        b_transimple_target_socket_callback_t who;

        public void b_transact() { who.b_transact(); }

        public void register_cb(b_transimple_target_socket_callback_t s) { who = s; }
    }
}

// An initiator
class CPU
{
    public TLM2_Sockets_Skeleton.simple_initiator_socket m0 = new TLM2_Sockets_Skeleton.simple_initiator_socket();
    public TLM2_Sockets_Skeleton.simple_initiator_socket m1 = new TLM2_Sockets_Skeleton.simple_initiator_socket();

    public void step() // Cycle callable bus functional model
    {
        ...
        m0.b_transact(...);
        ...
    }
}

// Class defining the RAM target. The IO would be very similar.
class RAM : TLM2_Sockets_Skeleton.b_transimple_target_socket_callback_t
{
    // A target has a target socket
    public TLM2_Sockets_Skeleton.simple_target_socket s0 = new TLM2_Sockets_Skeleton.simple_target_socket();

    public RAM() // Constructor
    { s0.register_cb(this); }

    public void b_transact()
    { Console.WriteLine("Got To RAM");
      ...
    }
}

class Bench
{
    static CPU the_cpu = new CPU();
    static RAM the_ram = new RAM();
    static RAM the_io = new IO();
    public static void wBench() // Constructor-like
    {
        the_cpu.m0.bind(the_ram.s0);
        the_cpu.m1.bind(the_io.s0);
    }

    public static void Main()
    {
        wBench();
        Console.WriteLine("Starting");
        the_cpu.step();
    }
}

```

The simple initiator and target sockets illustrated in this skeleton offer no significant benefits. The benefit arises in real implementations where other features are present and adherence to the coding standard arises from the type checker.

### 6.2.8 Timed Transactions: Adding delays to TLM calls.

A TLM call does not interact with the SystemC kernel or advance time. To study system performance, however, we must model the time taken by the real transaction over the bus or network-on chip (NoC).

We continue to use SystemC EDS kernel with its `tnow` variable defined by the head of the event queue. This is our main virtual time reference, but we aim not to use the kernel very much, only entering it when inter-module communication is needed.

Note: In SystemC, we can always print the kernel `tnow` with:

```
cout << "Time now is : " << simcontext()->time_stamp() << "\n";
```

This reduces context swap overhead (a computed branch that does not get predicted) and we can run a large number of ISS instructions or other operations before context switching, aiming to make good use of the caches on the modelling workstation.

The naive way to add **approximate timing** annotations is to block the SystemC kernel in a transaction until the required time has elapsed:

```
sc_time clock_period = sc_time(5, SC_NS); // 200 MHz clock

int b_mem_read(A)
{
    int r = 0;
    if (A < 0 or A >= SIZE) error(...);
    else r = MEM[A];
    wait(clock_period * 3); // <-- Directly model memory access time: three cycles say.
    return r;
}
```

The preferred **loosely-timed** coding style is more efficient: we pass a time accumulator variable called 'delay' around for various models to augment where time would pass (clearly this causes far fewer entries to the SystemC kernel):

```
// Preferred coding style
// The delay variable records how far ahead of kernel time this thread has advanced.
void b_putbyte(char d, sc_time &delay)
{
    ...
    delay += sc_time(140, SC_NS); // It should be increment at each point where time would pass...
}
```

The leading ampersand on delay is the C++ denotation for pass by reference. But, at any point, any thread can **resynch** itself with the kernel by performing

```
// Resynch idiomatic form:
wait(delay);
delay = 0;
// Note: delay has units sc_time so the SystemC overload of {\ttfamily wait} is called, not the O/S posix wait.
```

Important point: **With frequent resynchs, simulation performance is reduced but true transaction ordering is modelled more closely.**

### 6.2.9 TLM - Measuring Utilisation and Modelling Contention

When more than one client wants to use a resource at once we have **contention**.

Real queues are used in hardware, either in FIFO memories or by flow control applying backpressure on the source to stall it until the contended resource is available. An arbiter allocates a resource to one client at a time.

Contention like this can be modelled using real or virtual queues:

1. In a low-level model, the real queues are modelled in detail.
2. A TLM model may queue the transactions, thereby blocking the client's thread until the transaction can be served.
3. Alternatively, the transactions can be run straightaway and the estimated delay of a **virtual queue** can be added to the client's delay account.

To support style 2, SystemC provides a TLM payload queue: `tlm_utils::peq_with_get`

### 6.2.10 Replacing Queues With Delay Estimates

With a virtual queue, although the TLM call passes through the bus/NoC model without suffering delay or experiencing the contention or queuing of the real system, we can add on an appropriate estimated amount.

Delay estimates can be based on dynamic measurements of utilisation at the contention point, in terms of transactions per millisecond and a suitable formula, such as  $1/(1-p)$  that models the queuing delay in terms of the utilisation.

```
// A simple bus demultiplexor: forwards transaction to one of two destinations:
busmux::write(u32_t A, u32_t D, sc_time &delay)
{
    // Do actual work
    if (A >= LIM) port1.write(A-LIM, D, delay) else port0.write(A, D, delay);

    // Measure utilisation (time for the last 100 transactions)
    if (++opcount == 100)
    {
        sc_time delta = sc_time_stamp() - last_measure_time;
        local_processing_delay = delay_formula(delta, opcount); // e.g. 1 + 1/(1-p) nanoseconds
        logging.log(100, delta); // record utilisation

        last_measure_time = sc_time_stamp();
        opcount = 0;
    }

    // Add estimated (virtual) queuing penalty
    delay += local_processing_delay;
}
```

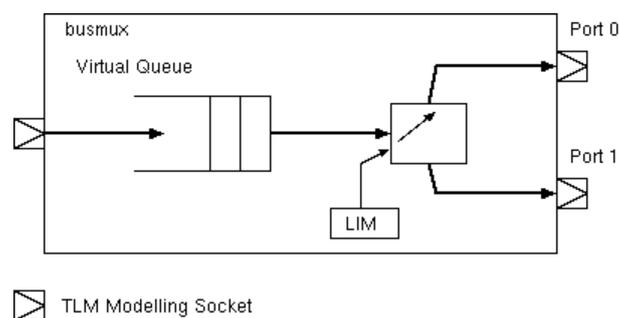


Figure 6.10: Busmux logical schematic diagram.

In the above, a delay formula function knows how many bus cycles per unit time can be handled and hence can compute and record the utilisation and queuing delays.

The value ' $p$ ' is the utilisation in the range 0 to 1. From queuing theory, with random arrivals, the queuing delay goes to infinity following a  $1/(1-p)$  response as  $p$  approaches unity. For uniform arrival and service times, the queuing delay goes sharply to infinity at unity.

### 6.2.11 Instruction Set Simulator (ISS)

An Instruction Set Simulator (ISS) is a program that interprets or otherwise models the behaviour of machine code. Typical implementation as a C++ object for ESL:

```
class mips64iss
{ // Programmer's view state:
  u64_t regfile[32]; // General purpose registers (R0 is constant zero)
  u64_t pc; // Program counter (low two bits always zero)
  u5_t mode; // Mode (user, supervisor, etc...)
  ...
  void step(); // Run one instruction
  ...
}
```

The ISS can be **cycle-accurate** or just **programmer-view accurate**, where the hidden registers that overcome structural hazards or implement pipeline stages are not modelled.

This fragment of a main step function evaluates one instruction, but this does not necessarily correspond to one clock cycle in hardware (e.g. fetch and execute would be of different instructions owing to pipelining or multiple issue):

```
void mips64iss::step()
{
  u32_t ins = ins_fetch(pc);
  pc += 4;
  u8_t opcode = ins >> 26; // Major opcode
  u8_t scode = ins & 0x3F; // Minor opcode
  u5_t rs = (ins >> 21) & 31; // Registers
  u5_t rd = (ins >> 11) & 31;
  u5_t rt = (ins >> 16) & 31;

  if (!opcode) switch (scode) // decode minor opcode
  {
    case 052: /* SLT - set on less than */
      regfile_up(rd, ((int64_t)regfile[rs]) < ((int64_t)regfile[rt]));
      break;

    case 053: /* SLTU - set on less than unsigned */
      regfile_up(rd, ((u64_t)regfile[rs]) < ((u64_t)regfile[rt]));
      break;

    ...
  }

  void mips64iss::regfile_up(u5_t d, u64_t w32)
  { if (d != 0) // Register zero stays at zero
    { TRC(trace("[ r%i := %11X ]", d, w32));
      regfile[d] = w32;
    }
  }
}
```

Various forms of ISS are possible, modelling more or less detail:

Type of ISS	I-cache traffic Modelled	D-cache traffic Modelled	Relative Speed
1. Interpreted RTL	Y	Y	0.000001
2. Compiled RTL	Y	Y	0.00001
3. V-to-C C++	Y	Y	0.001
4. Hand-crafted cycle accurate C++	Y	Y	0.1
5. Hand-crafted high-level C++	Y	Y	1.0
6. Trace buffer/JIT C++	N	Y	20.0
7. Native cross-compile	N	N	50.0

A cycle-accurate model of the processor core is normally available in RTL. Using this under an EDS interpreted simulator will result in a system that typically runs one millionth of real time speed (1). Using compiled RTL, as is now normal practice, gives a factor of 10 better, but remains hopeless for serious software testing (2).

Using programs such as Tenison VTOC and Verilator, a fast, cycle-accurate C++ model of the core can be generated, giving intermediate performance (3). A hand-crafted model is generally much better, requiring perhaps 100 workstation instructions to be executed for each modelled instruction (4). The workstation clock frequency is generally about 10 times faster than the modelled embedded system.

If we dispense with cycle accuracy, a hand-crafted behavioural model (5) gives good performance and is generally throttled by the overhead of modelling instruction and data operations on the model of the system bus.

A JIT (just-in-time) cross-compilation of the target machine code to native workstation machine code gives excellent performance (say 20.0 times faster than real time) but instruction fetch traffic is no longer fully modelled (6). Techniques that unroll loops and concatenate basic blocks, such as used for trace caches in processor architecture, are applicable.

Finally (line 7), compiling the embedded software using the workstation native compiler exposes the unfettered raw performance of the workstation for CPU-intensive code.

### 6.2.12 Typical ISS setup with Loose Timing (Temporal Decoupling)

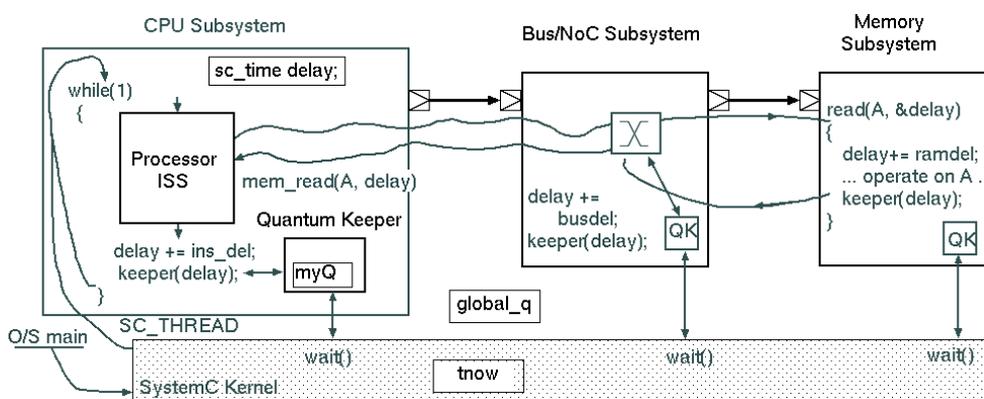


Figure 6.11: Typical setup of thread using loosely-timed modelling with a quantum keeper.

In this reference example, for each CPU core, a single thread is used that passes between components and back to the originator and only rarely enters the SystemC Kernel.

As explained above, each thread has a variable called **delay** of how far it has run ahead of kernel simulation time, and it only yields when it needs an actual result from another thread or because its delay exceeds a locally-chosen value. This is loose timing. Each component increments the delay field in the TLM calls it processes, according to how long it would have delayed the client thread under approximate timing. Every thread must encounter a quantum keeper at least once in its outermost loop.

The quantum keeper code is just a conditional resynch:

```
void keeper(ref delay) { if (delay > global_q) { wait(delay); delay = 0; } }
```

By calling `sc_wait(delay)` the simulation time will advance to where the caller has got to while running other pending processes. The `myQuantum` could be a system default value or a special value for each thread or component. (We here write `sc_wait` to emphasise it is the SystemC kernel primitive `systemc::wait`, not the unix call.)

Or where a thread needs to spin, waiting for a result from some other thread:

```

while (!condition_of_interest)
{
    wait(delay);
    delay = 0;
}

```

Generally, we can choose the quantum according to our current modelling interest:

- **Large time quantum:** fast simulation,
- **Small time quantum:** transaction order interleaving is more accurate.

Transactions may execute in a different sequence from reality: **sequential consistency** compromised ? Bugs exposed?

## 6.3 Power Estimation: RTL versus ESL

RTL simulations can give accurate power figures, especially if a full place-and-route is performed. ESL flows aim to provide a rapid power indications during the early architectural exploration phases.

### 6.3.1 RTL Operating Frequency and Power Estimation

RTL synthesis is relatively quick but produces a detailed output which is slow to process for a large chip - hence pre-synthesis energy and delay models are desirable. Place-and-route will give accurate wiring lengths but is a highly time consuming investment in a given design point. A simulation of a placed-and-routed design can give very accurate energy and critical path figures, but is likewise useless for 'what if' style design exploration.

A table of possible approaches:

-	- Without Simulation -	- Using Simulation -
<b>Without Place and Route</b>	Fast - Design exploration. Area and delay heuristics needed.	Can generate indicative activity ratios to be used instead of simulation in further runs.
<b>With Place and Route</b>	Static timing analyser will give an accurate clock frequency.	Gold standard: only bettered by measuring a real chip.

### 6.3.2 Gold standard: Power Estimation using Simulation Post Layout

Spreadsheet style power modelling from VCD and SAIF logs.

- **VCD:** Verilog Change Dump file - as generated by net-level SystemC or RTL simulations.
- **SAIF:** Switching Activity Interchange Format - the industry standard approach (aka Spatial Archive Interchange Format). Quick Tutorial

Both record the number of changes on each net of circuit from a net-level simulation. Once we know the capacitance of a net (from layout) we can accurately compute the power consumed. But, need to design down to the net-level and do a slow low-level simulation to collect adequate data.

Total Energy = Sum over all nets (net activity ratio \* net length)

Clearly, if we know the average net length and average activity ratio we get the same precise answer **regardless of design details**, hence good prospects exist for power estimation from high-level simulations.

### 6.3.3 RTL Power Estimation by Static Analysis (ie Without Simulation)

**Post** RTL synthesis we have a netlist and can use approximate models (based on Rent's rule) for wire lengths provided sufficient hierarchy exists (perhaps five or more levels). We can either use the natural hierarchy of the RTL input design or we can apply a clustering/clique finding algorithms to determine a rough placement floorplan without doing a full place and route.

**Pre** RTL synthesis we can readily collect the following certainties (and hence the static power (ignoring drive strength selection and power gating))

- Number of flip-flops
- Number and bit widths of arithmetic operators
- Size of RAMs

Random logic complexity can be modelled in gate-equivalent units. These might count a ripple-carry adder stage as 4 gates, a multiplexor as 3 gates per bit and a D-type flip-flop as 6 gates.

```

module CTR16(
  input mainclk,
  input din, input cen,
  output o);

  reg [3:0] count, oldcount; // D-types

  always @(posedge mainclk) begin
    if (cen) count <= count + 1; // ALU
    if (din) oldcount <= count; // Wiring
  end

  assign o = count[3] ^ count[1]; // Combinational
endmodule

```

But the following dynamic quantities require heuristic estimates:

- Dynamic clock gating ratios
- Flip-flop activity (number of enabled cycles/number of flipping bits)
- Number of reads and writes to RAMs
- Glitch energy in combinational logic.

DRAM power generally comes from a different budget (off chip) and can only really be estimated by dynamic modelling on a real or virtual platform. But note that for small embedded devices, the DRAM static power in its PCB track drivers can dominate DRAM dynamic power.

*Non-examinable:* There exists a technique to estimate the logic activity using balance equations.

We here use toggle rates, instead of activity ratios.

The balance equations range over a pair of values for each net, consisting of

- **average duty cycle:** the fraction of time at a logic one
- **average toggle rate:** the fraction of clock cycles where a net changes value. This is twice the activity ratio needed for the dynamic power computation at the end.

Consider an XOR gate with inputs toggling randomly. Assuming uncorrelated inputs, the output will also be random and its toggle rate can be predicted to be 50 percent. (c.f. entropy computations in Information Theory). But if we replace with an AND or OR gate, the output duty cycle will be 1 in 4 on average and its toggle rate will be given by the binomial theorem and so on.

Overall, a synchronous digital logic sub-system can be modelled with a set of balance equations (simultaneous equations) in terms of the average duty cycle and expected toggle rate on each net. D-types make no difference. Inverters subtract the duty cycle from unity. The other gates have equations as developed above.

Is this a useful measure? We need the stats for the input nets to run the model. We can look at the partial derivatives with respect to the input stats and if they are all very small, our result will hold over all inputs.

This is not widely used. Instead industry captures the average toggle rate of the nets in a subsystem during simulation runs (SAIF output) of each of the various operating phase/modes.

Some additional dynamic energy is consumed as 'short-circuit current' which is current that passes during switching when both the P and N transistors are on at once, but this is small and we mainly ignore it in these notes. Useful article: POWER MANAGEMENT IN CPU DESIGN

Short-circuit current is proportional to the toggle ratio. The toggle ratio,  $t_r$  is the percentage of clock cycles that see a transition in either direction. The net toggle rate = Operating frequency of the chip  $f \times t_r$ ;

- 1 W/cm<sup>2</sup> can be dissipated from a plastic package.
- 2-4 W/cm<sup>2</sup> requires a heat sink.
- more than 8 W/cm<sup>2</sup> requires forced air, heatpipe or water cooling.

Workstation and laptop microprocessors dissipate tens of Watts: hence cooling fans and heat pipes. In the past we were often *core-bound* or *pad-bound*. Today's SoC designs are commonly *power-bound*.

### 6.3.4 Typical macroscopic performance equations: SRAM example.

It is important to model SRAM accurately. A 45nm SRAM can be modelled at a macro level in terms of its Area, Delay and Power Consumption:

Four rules of thumb (scaling formulae) for single-ported SRAM CACTI at HP labs. Cacti RAM Models

Technology parameters:

- Read width 64 bits. Technology Size (nm):45 Vdd:1.0
- Number of banks: 1 Read/Write Ports per bank:1
- Read Ports per bank: 0 Write Ports per bank:0

Interpolated equations:

- Area = 13359.26+4.93/8\*bits squm: gradient = 0.6 squm/bit.
- Read energy = 5 + 1.2E-4 / 8 \* bits pJ.

- Leakage (static power) = 82nW per bit.
- Random access latency =  $0.21 + 3.8E-4(\sqrt{\text{bits}})$  nanoseconds \* 1.0/supply voltage.

Another rule of thumb: area is about 600 square lambda for an SRAM bit cell, where lambda is the feature size (45E-9).

### 6.3.5 Typical macroscopic performance equations: DRAM example.

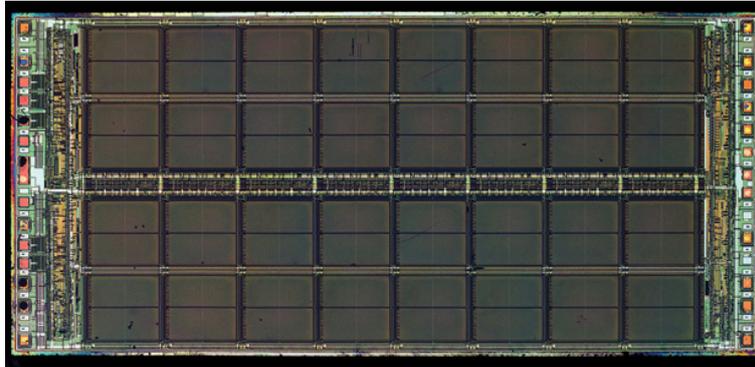


Figure 6.12: Micron Technology's MT4C1024 DRAM chip (1024x1024x1 circa 1994).

A DRAM channel is some number of DRAM die (eg. 16) connected to a set of I/O pads on a controller. The channel data width could typically be 16, 32 or 64 bits. The capacity might be 16 GByte.

- **Controller static power:** The pads forming the so-called 'phy' will typically consume half a watt or so, even when idle.
- **DRAM static power:** each die takes about 100 mW when idle but may enter a sleep mode if left unused for a millisecond or so, reducing this to 10 mW or so.
- Each **row activation** takes dynamic energy (see table).
- Each **column burst transfer** takes on-chip energy and PCB trace energy.
- Each **row closure** (writeback/de-activate) takes dynamic energy (see table).
- Refresh operations consume a small amount of dynamic energy (see exercise sheet for numbers).

Partial Row Activation for Low-Power DRAM System - Lee, Kim and Hong  
Calculating Memory System Power for DDR3

The Prazor virtual platform integrates the University of Maryland DRAM simulator.

The phy is the set of pads that operate with high performance to drive the PCB traces. Such power can be minimised if the traces are kept short with using multi-chip modules or die stacking. Micron have released a multi-channel DRAM module: the Micron HMC. This can have a number of host nodes sharing a number of die-stacked DRAM chips.

### 6.3.6 Rent's Rule Estimate of Wire Length

If we know the physical area of each leaf cell we can estimate the area of each component in a hierarchical design (sum of parts plus percentage swell).

**Table 2: DRAM die area and row activation energy breakdown of a 2Gb x8 DDR3-1600 DRAM chip.**

Area (mm <sup>2</sup> )			
DRAM cell	4.677	Sense amplifier	1.909
Row predecoder	0.067	Local wordline driver	1.617
Total area (including other components)			11.884
Energy per MAT (pJ)			
Local bitline	15.583	Local wordline	0.046
Local sense amplifier	1.257	Row decoder	0.035
Total row activation energy per MAT			16.921
Energy per bank (pJ)			
Row activation bus	17.944	Row predecoder	0.072
Total row activation energy per bank			288.752

Figure 6.13: DRAM activation energies.

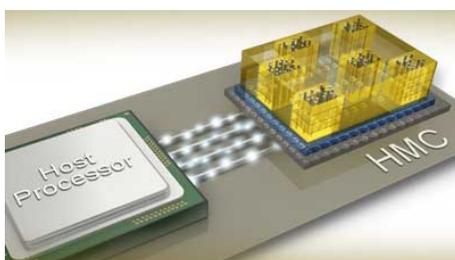


Figure 6.14: Die-stacked DRAM subsystem: The Micron Hybrid Memory Cube.

Rent's rule pertains to the organisation of computing logic, specifically the relationship between the number of external signal connections (terminals) to a logic block with the number of logic gates in the logic block, and has been applied to circuits ranging from small digital circuits to mainframe computers [Wikipedia].

$$\text{Terminals} = k (\text{Gates})^{\rho}$$

Rent's rule uses a simple power-law relationship for the number of terminals for a sub-system as function of the number of logic gates in the sub-system. Figure 6.15 shows three possible design styles that vary in rent coefficient. A circuit composed of components with no local wiring between them is the other extreme possibility, with a Rent exponent  $\rho$  of 1.0. But the rule-of-thumb is that for most 'general' subsystems a Rent exponent varying between about 0.5 and 0.7 is seen. Circuits like the shift register can be outliers, having no increase in external connectivity regardless of length: these have a Rent exponent of 0. The same situation arises with an accelerator on-a-stick, where the complexity of the accelerator (e.g. degree of unfold) will alter the rent exponent owing to the fixed-configuration bus port.

Hefeida and Chowdhury give some plots that explore typical designs. Not surprisingly we see that the average net length looks roughly like a flattened square-root function in the number of gates. The paper gives detailed formulae that have exponents in the range 0.25 to 0.5. Improved Model for Wire-Length Estimation in Stochastic Wiring Distribution

#### Lowest-Common Parent Approach Assuming Good Layout:

Knowing the average wire length and the average activity ratio is not sufficient to get the average power owing to non-linear effects (all of the activity might be on the longer nets for instance). Hence it is better to have a more detailed model when forming the product.

Generalisations of Rent's rule can model wire length distribution (with good placement). For a single level of design hierarchy, the random placement of blocks in a square of area defined by their aggregate area gives one wire length distribution (basic maths). A careful placement is used in practice, and this reduces wire length

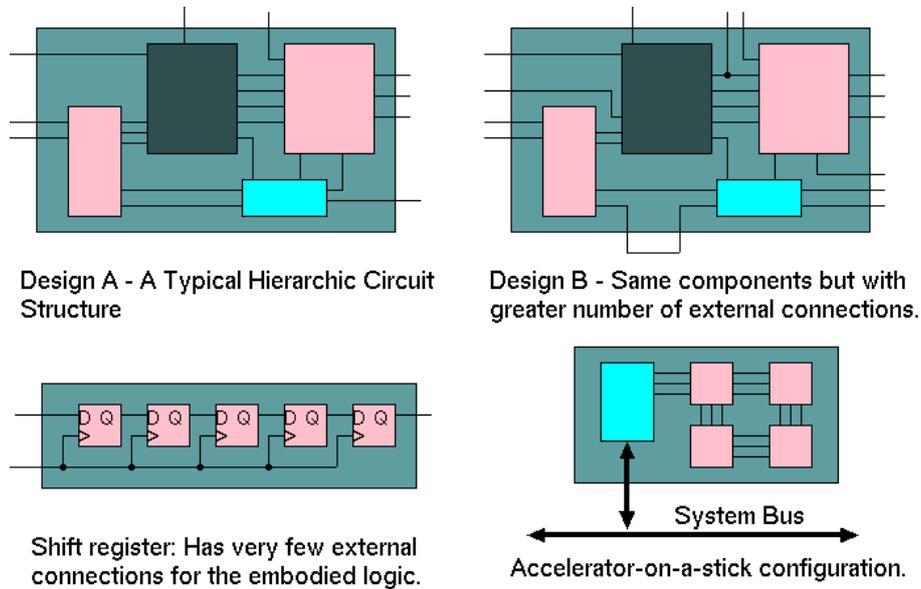


Figure 6.15: Two similar designs with different Rent exponents and two non-Rentian design points.

by a Rent-like factor (eg. by a factor of 2). With a hierarchic design, where we have the area use of each leaf cell, even without placement, we can follow a net's trajectory up and down the hierarchy and apply Rent's Rule. Hence we can estimate a signal's length by sampling a power law distribution whose 'maximum' is the square root of the area of the lowest-common-parent component in the hierarchy.

### 6.3.7 Macroscopic Phase/Mode Power Estimation Formula

An IP block will tend to have an average power consumption in each of its phases or modes. Power modes include sleep, idle, off, on etc.. Clock frequency and supply voltage are also subject to step changes and expand the discrete phase/mode operating space. Given that blocks switch between energy states a simple energy estimation technique is based percentage of time in each state. This was how the TLM POWER2 library for SystemC worked. TLM POWER3 uses this approach for static power but logs energy quanta for each transaction.

The Parallella board has two USB line drivers (circled in red). They have three operating modes with different energy use in each.

### 6.3.8 Spreadsheet-based Energy Accounting

Knowing the average number of operations per second on a unit is generally all that is needed to work out its average energy, once the joules per operation are known.

The Xilinx Xpower 28 nm Zynq spreadsheet models about 415 pJ per ARM clock cycle and 23 pJ per DSP multiply. BRAM reads of RAMB18x2 (36 Kbit) units take 8.5 pJ and writes about 10 percent less. The formula earlier for 45nm,  $5.0 + 1.2e-4 / 8.0 * \text{mbits}$  gives 5.5 pJ but twice as much for a write.

But the totals for each component drastically depend on the activity ratios and initial guesses are typically set close to worst case which is conservative, but typically wildly out. Hence SAIF-based or other dynamic trace information must be fed in for an accurate result.

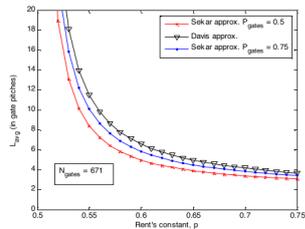


Fig. 2: Effect of Rent's exponent on the average interconnect length for a large number of gates

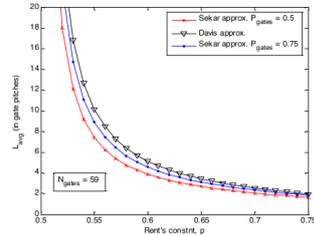


Fig. 3: Effect of Rent's exponent on the average interconnect length for a relatively small number of gates

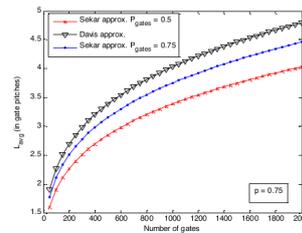


Fig. 4: Effect of gate count on the average interconnect wire length

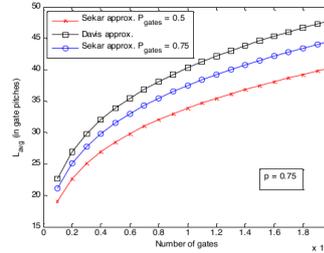


Fig. 5: Effect of gate count on the average interconnect length over an extended range

Figure 6.16: Rent Wiring Length Graphs (Hefieida-2015).

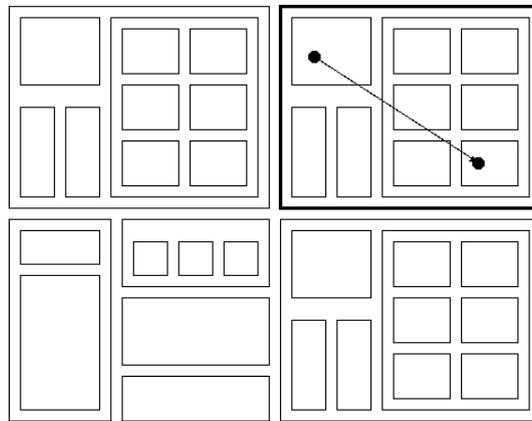


Figure 6.17: Illustrating Lowest Common Parent of the endpoint logic blocks. (This will always be roughly the same size for any sensible layout of a given design, so having a detailed layout like the one shown is not required).

### 6.3.9 Transactional Energy Modelling

SystemC TLM POWER2 library release:

- Deals with power 'modes' and 'phases' of subsystems,
- Was difficult to integrate with loosely-timed modelling,
- Could not record energy consuming events, such as individual bus transactions,
- Power consumption for a component read from a table that must always be manually created.

Cambridge TLM POWER 3 library for SystemC:

- Supports power and energy equally well, with power calculations being accurate at the end of each LT quantum.

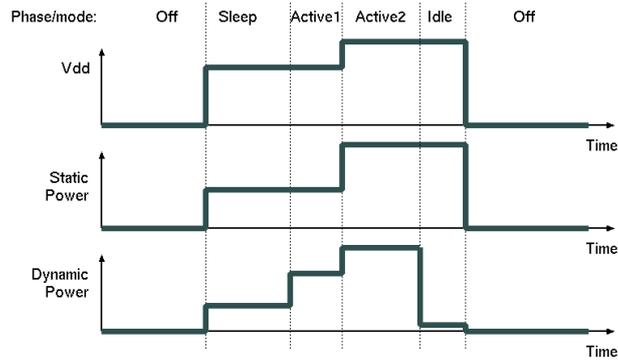


Figure 6.18: Typical Phase and Mode Time-Domain Plot.



Phase/Mode Power Model	USB Phy						
	Rail 1		Rail 2		Rail 3		Total Power
Mode	V	A	V	A	V	A	W
Idle	3.30E+000	1.80E-005	1.80E+000	7.00E-007	3.30E+000	3.00E-005	1.60E-004
L/S Mode	3.30E+000	6.30E-003	1.80E+000	1.10E-002	3.30E+000	5.00E-003	5.71E-002
H/S mode	3.30E+000	2.90E-002	1.80E+000	2.20E-002	3.30E+000	5.90E-003	1.55E-001

Figure 6.19: Phase/Mode figures for a USB Phy Line Driver.

- Requires each component to inherit the Prazor base class (current implementation),
- Enables physical size of components to be logged (e.g. as a basis for nominal place and route),
- Supports X-Y co-ordinates to be allocated to each component (a concrete layout),
- Wiring distances can be estimated using Rent's rule OR measured from X-Y coordinates if placed,
- Power/energy consumption for a component can depend on constructor args (e.g. memory size, bus width).
- Counts bit flips in generic payload and multiplies by length of the bus.
- A 'recompute\_pvt' method of each module is called when supply voltage (or temperature etc) is changed.

(Details not examinable.)

Setting the static power and area for an SRAM based on earlier 45nm formulae:

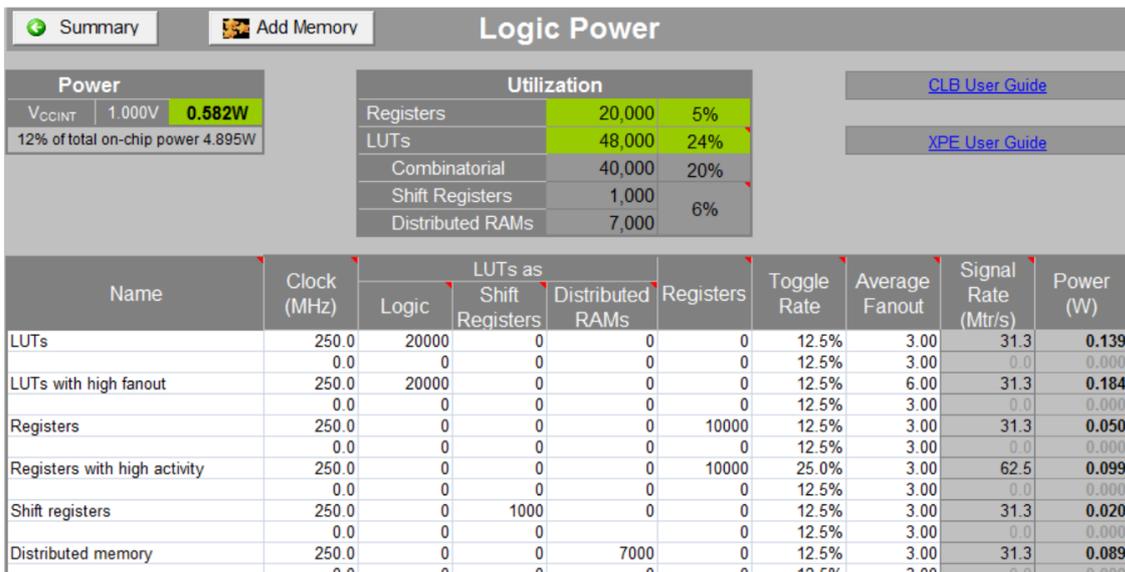


Figure 21: Effect of LUT Configuration, Toggle Rates and Average Fanout on Power Estimation (7 Series)

Figure 6.20: Example Xilinx Xpower Spreadsheet

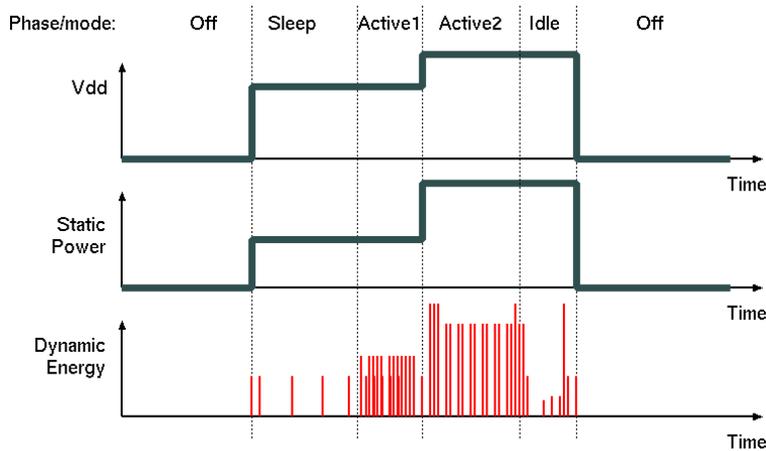


Figure 6.21: Typical Transactional Energy Modelling Plot.

```

void sram64_cbg::recompute_pvt_parameters() // Called in constructor and when Vcc is changed.
{
    m_latency = sc_time(0.21 + 3.8e-4 * sqrt(float(m_bits)), SC_NS);

    pw_power leakage = pw_power(82.0 * m_bits, PW_nW);
    set_static_power(leakage);

    set_fixed_area(pw_area(13359.0 + 4.93/8 * m_bits, PW_sqm));

    m_read_energy_op = pw_energy(5.0 + 1.2e-4 / 8.0 * m_bits, pw_energy_unit::PW_pJ);
    m_write_energy_op = 2.0 * m_read_energy_op; // rule of thumb!

    // NB: Might want different energy when high-order address bits change.

    pw_voltage vcc = get_vcc();
    m_latency = m_latency / vcc.to_volts();
    cout << name () << ": " << kind() << ": final latency = " << m_latency << "\n";
}
    
```

How to log the wiring and transaction power for an SRAM:

```

void sram64_cbg::b_access(PW_TLM_PAYLOADTYPE &trans, sc_time &delay)
{
    tlm::tlm_command cmd = trans.get_command();

    // Log wiring power consumed by transaction arriving here.
    // Also set which nets modelled by the TLM will be active after this operation:
    // For a write none (except a response ack) and for read the payload data.
    trans.pw_log_hop(this, (cmd==tlm::TLM_READ_COMMAND ? PW_TGP_DATA: PW_TGP_NOFIELDS) |
        PW_TGP_ACCT_CKP, &read_bus_tracker);

    if (cmd == tlm::TLM_READ_COMMAND)
    {
        // Log internal transaction energy for read
        pw_module_base::record_energy_use(m_read_energy_op);
        ...
    }
    else if (cmd == tlm::TLM_WRITE_COMMAND)
    {
        // Log internal transaction energy for write
        pw_module_base::record_energy_use(m_write_energy_op);
        ...
    }
}

```

### 6.3.10 TLM POWER 3 Report File Example

MODULE NAME	STATIC0 ENERGY		DYNAMIC1 ENERGY		WIRING2 ENERGY	
Standalone modules:						
...top.coreunit_0.core_0	9.997983e-05J	0.77%	3.25128e-05J	0.25%	1.35116151e-07J	0.00%
Memory 0 (DRAM)	0.00866173075J	66.65%	0.00419979737J	32.32%	1.32334593e-07J	0.00%
the_top.uart0	0J	0.00%	8.84e-07J	0.01%	2.746e-12J	0.00%
Customer Accounts:						
anonymous	0.00866173075J	66.65%	3.25128e-05J	0.25%	2.6745349e-07J	0.00%
busaccess_0	0J	0.00%	0.00420136352J	32.33%	0J	0.00%
TOP LEVEL++	0.00876171058J	67.42%	0.00423387632J	32.58%	2.6745349e-07J	0.00%
Each line is for a separately-traced subsystem. These lines may be neither disjoint or complete. The TOP LEVEL figure is simply another line in the table that relates to the highest module found. Total energy used: 12900 uJ (12995854356318 fJ)						
MODULE NAME	STATIC0 POWER		DYNAMIC1 POWER		WIRING2 POWER	
Standalone modules:						
...top.coreunit_0.core_0	0.01W	75.38%	0.00325193592W	24.51%	1.35143409e-05W	0.10%
Memory 0 (DRAM)	0.866347818W	67.35%	0.420064464W	32.65%	1.3236129e-05W	0.00%
the_top.uart0	0W	0.00%	8.84178339e-05W	100.00%	2.74655e-10W	0.00%
Customer Accounts:						
anonymous	0.866347818W	99.62%	0.00325193592W	0.37%	2.67507446e-05W	0.00%
busaccess_0	0W	0.00%	0.420221111W	100.00%	0W	0.00%
TOP LEVEL++	0.876347818W	67.42%	0.423473047W	32.58%	2.67507446e-05W	0.00%
Each line is for a separately-traced subsystem. These lines may be neither disjoint or complete. The TOP LEVEL figure is simply another line in the table that relates to the highest module found. Average power used: 1290 mW (1299847614895725 fW)						

Figure 6.22: Text Trace File generated by TLM POWER3

### 6.3.11 sPEEDO Energy Interface

The sPEEDO project was a local strawman API proposal for application code to read off its own energy use. Fine-grained Energy/Power Instrumentation for Software-level Efficiency Optimization

If you invoke the Prazor `./configure --with-speedo` and with `TLMPOWER3` included, you can read off the energy modelled by

```

// Write a simple unit value to do a basic checkpoint, copying live CTX0 to CTX1.
SPEEDO_CHKPT(0);
energy_t e = READ_SPEEDO_CTX1(SPEEDO_CTX_REG_GLOBAL_ENERGY);

```

The energyshim.c file can either read backdoor registers on the virtual platform or else read from the real probe on the real board.

### 6.3.12 Higher-Level Simulation - Virtual Platforms

There are a number of full-system simulators or 'virtual platforms' in academia and industry. Examples include QEMU, Sniper, ZSim, GEM5 and Prazor Virtual Platform For further reading in part III: ACS P35

## KG 7 — High-level Design Capture and Synthesis

In this section of the course we look at high-level synthesis and possibly some other high-level design entry methods.

### 7.0.13 Accellera IP-XACT

IP-XACT is an XML Schema for IP Block Documentation standardised as IEEE 1685. Wikipedia

It was developed by an industrial working party, the SPIRIT Consortium, as a standard for automated configuration and integration of IP blocks. IP-XACT is an IEEE standard for describing IP blocks and for automated configuration and integration of assemblies of IP blocks. It describes interfaces and attributes of a block (e.g. terminal and function names, register layouts and non-functional attributes). It includes separate RTL and ESL/TLM descriptions (future work to integrate these). It aims to provide all the front-end infrastructure for rapid SoC assembly from diverse IP supplies, support for assertions and and perhaps even some glue logic synthesis.

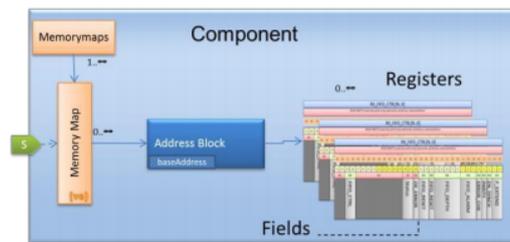


Figure 7.1: IP-XACT captures memory map and register field definitions.

All IP-XACT documents use titular attributes `spirit:vendor`, `spirit:library`, `spirit:name`, `spirit:version`. A document typically represents one of:

- bus specification, giving its signals and protocol etc;
- leaf IP block data sheet;
- or a heirarchic component wiring diagram that describes a sub-system by connecting up or abstracting other components made up of `spirit:componentInstance` and `spirit:interconnection` elements.

For each port of a component there will be a `spirit:busInterface` element in the document. This may have a `spirit:signalMap` that gives the mapping of the formal net names in the interface to the names used in a corresponding formal specification of the port. A simple wiring tool will use the signal map to know which net on one interface to connect to which net on another instance of the same formal port on another component.

There may be various versions of a component in the document, each as a `spirit:view` element, relating to different versions of a design: typical levels are gate-level, RTL and TLM. Each view typically contains a list of filenames as a `spirit:fileSet` that implement the design at that level of abstraction in appropriate language, like Verilog, C++ or PSL.

Non-functional data present includes the programmer's view with a list of `spirit:register` declarations inside a `spirit:memoryMap` or `spirit:addressBlock`.

Similar tools: Our Part Ib students currently use the Qsys System Integrator tool from Altera. ARM has its Socrates tool and Xilinx has IP Designer in Vivado.

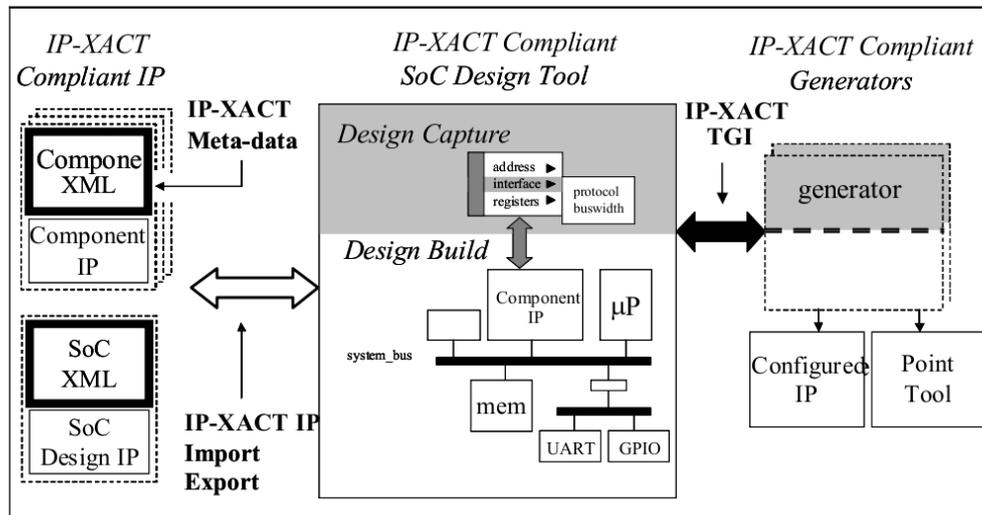


Figure 7.2: Reference Model for design capture and synthesis using IP-XACT blocks.

IP blocks and bus standards are stored in libraries and indexed using datasheets for each block in xml according to the IP-XACT schema. A schematic design capture editor supports creation and editing of a high-level block diagram for the SoC. The SoC design is then output as IP-XACT conformant XML. Various synthesis plugins, termed 'generators' produce the inter-block structural wiring RTL that is otherwise highly tedious and error prone to manually create. They may also instantiate bus bridges and port multiplexors and other glue logic.

Automatic generation of memory maps and device-driver header files is also normally supported. Header files in RTL and C are kept in synch. Testbenches following OVM/UVM coding standards might also be rendered. Other outputs, such as power and frequency estimates or user manual documentation are typically generated too. Greaves+Nam created a glue logic synthesiser Synthesis of glue logic, transactors, multiplexors and serialisers from protocol specifications.

Perhaps explore the free plugin(s) for Eclipse if you are keen.

#### 7.0.14 Start Here

##### Start Slide

The first half or majority of these slides covers 'classical HLS'. The final slides or second half discusses some alternative schemes, to be covered if time permits.

##### High-Level Synthesis (HLS)

Generally speaking, High-Level Synthesis (HLS) compiles software into hardware.

Although a research topic for decades, HLS is now seeing industrial traction. An HLS system revolves around an **HLS compiler** for a high-level language (typically C++). This

- Binds HLL arrays to RAMs and base addresses to items stored in a common RAM.
- Decides what mix of structural components (FUs) such as RAMs and ALUs to instantiate.
- Allocates work to clock cycles (aka scheduling).
- Generates an RTL output for the logic synthesis.
- May provide pre-built libraries for common I/O and mathematics.

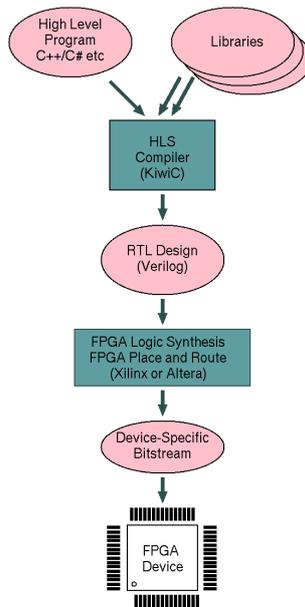


Figure 7.3: Basic Steps of an HLS Flow.

The output from an High-Level Synthesis (HLS) compiler is generally RTL which is then fed to an RTL compiler, aka **Logic Synthesiser**, that performs logic synthesis. As we have seen, the logic synthesizer

- instantiates multiplexors that transfer data according to predicates.
- performs logic minimisation or area/power optimisation.
- expands operations on broadside (aka vector) registers into simple boolean operations (aka bit blasting).
- replaces simple boolean operators with gates from an ASIC cell library or look-up tables in an FPGA.

Traditional RTL design entry (Verilog/VHDL) needs:

- Human comprehension of the state encoding,
- Human comprehension of the cycle-by-cycle concurrency, and
- Human accuracy to every low-level detail, such as which registers are **live**.

Performing a Time-for-Space re-folding (i.e. doing the same job with more/less silicon over less/more time) requires a **complete** redesign when entered manually in RTL!

Optimising schedules in terms of memory port and ALU uses? RTL requires us use Pen and paper? Can we do better than manual RTL coding? Yes, we use **High-Level Synthesis**.

Dark silicon facilitates 'Conservation Cores'. A paper at ASPLOS'10 about putting common kernels in silicon and 'Reducing the Energy of Mature Computations' by power gating. PDF

If one considers an embedded processor connected to a ROM, it may be viewed as one large FSM. Since for any given piece of software, the ROM is unlikely to be full and there are likely to be resources in the processor that are not used by that software: the application of a good quality logic minimiser to the system, while it is in the design database, could trim it greatly. In most real designs, this will not be helpful: for instance, the advantages of full-custom applied to the processor core will be lost. In fact, the minimisation function may be too complex for most algorithms to tackle on today's computers.

On the other hand, algorithms to create a good static scheduling of a fixed number of hardware resources work quite well. A processing algorithm typically consists of multiple processing stages (e.g. called pre-emphasis, equalisation, coefficient adaptation, FFT, deconvolution, reconstruction and so on). Each of these steps normally has to be done within tight real-time bounds and so parallelism through multiple instances of ALU and register hardware is needed. The Cathedral DSP compiler was an early tool for helping design such circuits. Such tools can perform time/space folding/unfolding of the algorithm to generate the static schedule that maps operations and variables in a high-level description to actual resources in the hardware. Cache misses, contention for resources and operations that have data-dependent runtime will cause time-domain deviations from a static schedule, so a potentially a dynamic schedule could then make better use of resources **but the overhead of dynamic scheduling can outweigh the cost of the resources saved if the data dependant variations are rare.**

Custom hardware is generally much more energy efficient than general-purpose processors. Reasons include (also see DJG 9-points in §10).

- All the resources deployed are in use with no wasted area,
- Dedicated data paths are not waylaid with unused multiplexors,
- Paths, registers and ALUs can have appropriate widths rather than being rounded up to general word sizes,
- No fetch/execute overhead,
- Even on an FPGA with its exaggerated dimensions, pass transistor multiplexors use less energy than active multiplexors,
- Operands are fetched in an optimised order, computed once-and-for-all rather than at each step as in today's complex out-of-order CPUs.

### 7.0.15 Higher-level: Generative, Behavioural or Declarative?

There are several primary, high-level design expression styles we can consider (in practice use a blend of them ?):

Purely **generative approaches**, like the Lava and Chisel hardware construction languages (HCLs), just 'print out' a circuit diagram or computation graph. There is also the **generate** statement in Verilog and HLD RTLs. Such approaches do not handle data-dependent IF statements: instead muxes must be explicitly printed (although Chisel has some syntactic support for mux generation with **when** like statements). Lava Compiler (Singh)

Generative approaches for super-computer programming, such as DryadLINQ from Microsoft, also elegantly support rendering large static computation trees (CSP or Kahn Networks) that can be split over processing nodes. They are agnostic as to what mix of nodes is used: FPGA, GPU or CPU. pn tool for Process Networks.

But the most interesting systems support complex data-dependent control flow. These are either:

- **Behavioural:** Using imperative software-like code, where threads have stacks and pass between modules, and so on..., or
- **Declarative/Functional/Logical:** Constraining assertions about the allowable behaviour are given, but any ordering constraints are implicit (e.g. SQL queries) rather than being based on a program counter concept. (A declarative program can be defined as an un-ordered list of definitions, rules or assertions that simultaneously hold at all times.)

Historically, the fundamental problem to be addressed was **Programmers like imperative programs but the PC concept with its associated control flow limits available parallelism.** An associated problem is that **even a fairly pure functional program has limited inferable parallelism in practice**

Using a parallel set of guarded atomic actions (as in Bluespec) is pure RTL: which is declarative (since no threads).

All higher-level styles are amenable to substantial automatic design space exploration by the compiler tool. The tool performs **datapath** and **schedule** generation, including re-encoding and re-pipelining to meet timing closure and power budgets.

So-called ‘classical HLS’ converts a behavioural thread to a static schedule (fixed at compile time). But the parallelism inferred is always limited. Sometimes no scheduler is needed. This can mean a fully-pipelined implementation can be generated. Alternatively, a systolic array or process network can be rendered, that again has no sequencer, but which accepts data with an initiation interval greater than unity (fixed for systolic array and variable for a CSP/Kahn-like network).

Transistors are abundant and having a lot of hardware is not itself a problem (We discussed the *Power Wall* in §2.3.3). Three forms of parallel speed-up are well-known for classical imperative parallel programming:

- **Task-Level Parallelism:** partition the input data over nodes and run the same program on each node without inter-node communication (aka *embarrassingly parallel*).
- **Programmer-defined, Thread-Level Parallelism:** The programmer uses constructs such as pthreads or CSharp Parallel.for loop to explicitly denote local regions of concurrent activity that typically communicate using shared variables.
- **Instruction-Level Parallelism:** The imperative program (or local region of) is converted to dataflow form, where all ALU operations can potentially be run in parallel, but operands remain pre-requisite to results and load/store operations on a given mutable object must respect program order.

A major (yet sadly less-popular) alternative to thread-level parallelism is programmer-defined channel-based communication, that bans mutable shared variables (examples: Erlang/Occam/Handel-C/Kahn Networks).

## 7.0.16 Instruction-Level Parallelism

Q. Does a program have a certain level of implicit parallelism? A. With respect to a specific compiler optimisation level and a specific input data set it does indeed. Here is a concrete example:

```
Prihozhy - Code for counting digits 1..5 in integer n.
void main()
{ unsigned long n=21414;
  int m[5], k=0;
  for (int i=0;i<5;i++) m[i]=0;
  while(n) { m[n%10]=1; n/= 10; }
  for (int j=0;j<5;j++) if (m[j]) k++;
}
```

In their 2003 paper, Prihozhy, Mattavelli and Mlynec, determine the available parallelism in various C programs. For small programs, such as the digit counting example above, with given input data, the critical path length and the total number of instructions (or clock cycles) can be drawn out. The critical path is highlighted with bold/wider arrows. (Someone volunteer to make them orange?) Their ratio gives the **available instruction-level parallelism**, such as  $174/30=5.8$ . Data Dependencies Critical Path Evaluation (Note also David Wall’s ‘Limits of instruction-level parallelism’. In Proc. ASPLOS-4, 1991. and J Mak’s ‘Limits of parallelism using dynamic dependency graphs’ 2009.)

Basic algorithm: Like the static timing analyser for hardware circuits (§4.5.2), for each computation node we add its delay to that of the latest-arriving input. But with different input data, the control flow varies and the available parallelism varies. Also the code can be re-factored to increase the parallelism.

The parallelism of the digit counter example can be improved from 30 to 25, as shown in the paper, e.g. by using variants of the **while-to-do transformation**.



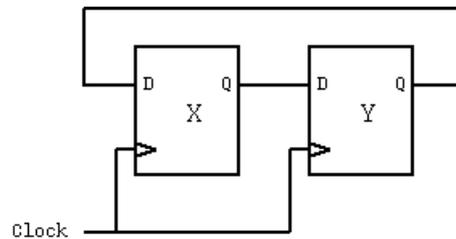


Figure 7.5: A circuit to swap two registers.

```

always @(posedge clk) begin
    x <= y;
    y <= x;
end

```

will produce the circuit of Figure 7.5. (If X and Y were busses, the circuit would be repeated for each wire of the bus.) The semantics of the above code are that the right-hand sides are all evaluated and then assigned to the left-hand sides. The order of the statements is unimportant.

However, as mentioned in §4.2.1, the same circuit may be generated using a specification where assignment is made using the = operator. If we assume there is no other reference to the intermediate register t elsewhere, and so a flip-flop named t is not required in the output logic. On the other hand, if t is used, then its input will be the same as the flip-flop for y, so an optimisation step will use the output of y instead of having a flip-flop for t.

```

always @(posedge clk) begin
    t = x;
    x = y;
    y = t;
end

```

With this style of specification the order of the statements is significant and typically such assignment statements are incorporated in various nested if-then-else and case commands. This allows hardware designs to be expressed using the conventional imperative programming style that is familiar to software programmers. The intention of this style is to give an easy to write and understand description of the desired function, but this can result in logic output from the synthesiser which is mostly incomprehensible if inspected by hand.

The word ‘behavioural’, when applied to a style of RTL or software coding, tends to simply mean that a sequential thread is used to express the sequential execution of the statements.

Despite the apparent power available using this form of expression, there are severe limitations in the officially synthesisable subset of Verilog and VHDL that might also be manifest in basic C-to-gates tool. Limitations are, for instance, each variable must be written by only one thread and that a thread is unable to leave the current file or module to execute subroutines/methods in other parts of the design.

The term ‘*behavioural model*’ is used to denote a short program written to substitute for a complex subsection of a structural hardware design. The program would produce the same useful result, but execute much more quickly because the values of all the internal nets and pipeline stages (that provide no benefit until converted to actual parallel hardware form) were not modelled. Verilog and VHDL enable limited forms of behavioural models to serve as the source code for the subsection, with synthesis used to form the netlist. Therefore limited behavioural models can sometimes become the implementation.

Many RTL synthesisers support an implied program counter (state machine inference).

```

reg [2:0] yout;
always
begin
  @(posedge clk) yout = 1;
  @(posedge clk) yout = 4;
  @(posedge clk) yout = 3;
end

```

In this example, not only is there a thread with current point of execution, but the implied ‘program counter’ advances only partially around the body of the `always` loop on each clock edge. Clearly the compiler or synthesiser has to make up flip-flops not explicitly mentioned by the designer, to hold the current ‘program counter’ value.

None of the event control statements is conditional in the example, but the method of compilation is readily extended to support this: it amounts to the program counter taking conditional branches. For example, the middle event control could be prefixed with ‘if (din)’.

```

if (din) @(posedge clk) yout = 4;

```

Take a non-reentrant function:

- generate a custom **datapath** containing registers, RAMs and ALUs
- and a custom **sequencer** that implements an efficient, **static schedule**

that achieves the same behaviour.

```

int multiply(int A, int B) // A simple long multiplier with variable latency.
{ RA=A; // Not RTL: The while loop trip count is data-dependent.
  RB=B; //
  RC=0; //
  while(RA>0) // Let's make a naive HLS of this program...
  {
    if odd(RA) RC = RC + RB;
    RA = RA >> 1;
    RB = RB << 1;
  }
  return RC;
}

```

This simple example has no multi-cycle primitives and a 1-to-1 mapping of ALUs to the source code text, so no scheduling was needed from the HLS tool. Each register has a multiplexor that ranges over all the places it is loaded from. We followed a **syntax-directed** approach (also known as a constructive approach) with no search in the solution space for minimum clock cycles or minimum area or maximum clock frequency. The resulting block could serve as a primitive to be instantiated by an HLS tool. This example is not fully-pipelined and so typically would not be used for that purpose.

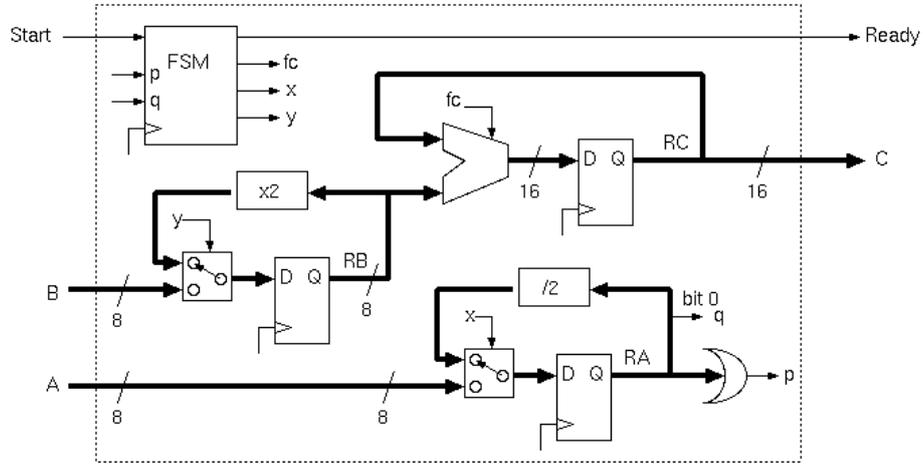
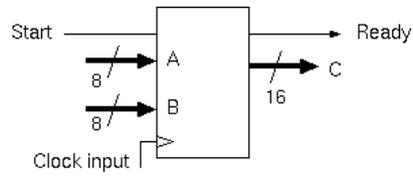


Figure 7.6: Long multiplier viewed as datapath and sequencer.

## 7.0.18 Multiplier Answer (2)

```

module LONGMULT8b8(clk, reset, C, Ready, A, B, Start);

    input clk, reset, Start;
    output Ready;
    input [7:0] A, B;
    output [15:0] C;
    reg [15:0] RC, RB, RA;
    reg Ready;

    // Behavioural code:
    // while (1)
    // {
    //     wait (Start);
    //     RA=A;RB=B;RC=0;
    //     while(RA>0)
    //     { if odd(RA) RC=RC+RB;
    //       RA = RA >> 1;  RB = RB << 1;
    //     }
    //     Ready = 1;
    //     wait(!Start);
    //     Ready = 0;
    // }

    reg xx, yy, qq, pp; // Control and predicate nets
    reg [1:0] fc;
    reg [3:0] state;
    always @(posedge clk) begin
        xx = 0; // default settings.
        yy = 0;
        fc = 0;

        // Predicates
        pp = (RA!=16'h0); // Work while pp holds
        qq = RA[0]; // Odd if qq holds
        // Sequencer
        if (reset) begin
            state <= 0;
            Ready <= 0;
        end
        else case (state)
            0: if (Start) begin
                xx = 1;
                yy = 1;
                fc = 2;
                state <= 1;
            end

            1: begin
                fc = qq;
                if (!pp) state <= 2;
            end

            2: begin
                Ready <= 1;
                if (!Start) state <= 3;
            end

            3: begin
                Ready <= 0;
                state <= 0;
            end

        endcase // case (state)

        // Datapath
        RB <= (yy) ? B: RB<<1;
        RA <= (xx) ? A: RA>>1;
        RC <= (fc==2) ? 0: (fc==1) ? RC+RB: RC;
    end

    assign C = RC;
endmodule

```

Suitable test wrapper:

```

module SIMSYS();
  reg clk, reset, Start;
  wire Ready;
  wire [7:0] A, B;
  wire [15:0] C;

  // Reset Generator
  initial begin reset = 1; # 55 reset = 0; end

  // Clock Generator
  initial begin clk = 0; forever #10 clk = !clk; end

  // Stimulus
  assign A = 6;
  assign B = 7;

  // Handshake control
  always @(posedge clk) Start <= !Ready;

  // Console output logging:
  always @(posedge clk) $display("Ready=%h C=%d");

  // Device under test.
  LONGMULT8b8 the_mult(clk, reset, C, Ready, A, B, Start);
endmodule // SIMSYS

```

### 7.0.19 Classical HLS Compiler: Operational Phases

The classical HLS tool operates much like a software compiler, but needs more time/space guidance. A single thread from an imperative program is converted to a sequencing FSM and a custom datapath.

1. **Lexing and Parsing** as for any HLL
2. **Type and reference checking**: can an int be added to a string? Is an invoked primitive supported?
3. **Trimming**: Unreachable code is deleted, register widths are reduced where it is manifest that the value stored is bounded, constants are propagated between code blocks and identity reductions are applied to operators, such as multiplying by unity.
4. **Binding**: Every storage and processing element, such as a variable or an add operation or memory read, is allocated a physical resource.
5. **Polyhedral Mapping**: A memory layout or ordering optimisation for nested loops.
6. **Schedulling**: Each physical resource will be used many times over in the time domain. A static schedule is generated. This is typically a scoreboard of what expressions are available when. (Worked example in lecturers.)
7. **Sequencer Generation**: A controlling FSM that embodies the schedule and drives multiplexor and ALU function codes is generated.
8. **Quantity Surveying**: The number of hardware resources and clock cycles used can now be readily computed.
9. **Optimisation**: The binding and schedulling phase may be revisited to better match user-provided target metrics.
10. **RTL output**: The resulting design is printed to a Verilog or VHDL file.

Some operations are intrinsically or better implemented as variable-latency. Examples are division and reading from cached DRAM. This means the static schedule cannot be completely rigid and must be based on expected execution times.

Important binding decisions arise for memories:

- Which user arrays are to have their own RAMs or which to share or which to put in DRAM?
- Should a user array be spread over RAMs for higher bandwidth?
- How is data to be packed into words in the RAMs?
- For ROMs, extra copies can be freely deployed.
- Mirroring data in RAM for more read bandwidth will require additional work when writing to keep in step.
- How should data be organised over DRAM rows? Should data even be stored in DRAM more than once with different row alignments?

### 7.0.20 Adopting a Suitable Coding Style for HLS

A coding style that works well for a contemporary Von Neumann computer may not be ideal for HLS. For now at least, we cannot simply deploy existing software and expect good results.

Here are four sections of code that all perform the same function. Each is written in CSharp. The zeroth routine uses a loop with a conditional branch on each execution. It has data-dependent control flow.

```
public static uint tally00(uint ind)
{
    uint tally = 0;
    for (int v = 0; v < 32; v++)
    {
        if (((ind >> v) & 1) != 0) tally++;
    }
    return tally;
}
```

Implementation number one replaces the control flow with arithmetic.

```
public static uint tally01(uint ind)
{
    uint tally = 0;
    for (int v = 0; v < 32; v++)
    {
        tally += ((ind >> v) & 1);
    }
    return tally;
}
```

This version uses a nifty programming trick

```
public static uint tally02(uint ind)
{ // Borrowed from the following, which explains why this works:
  // http://graphics.stanford.edu/~seander/bithacks.html#CountBitsSetParallel
  uint output = ind - ((ind >> 1) & 0x55555555);
  output = ((output >> 2) & 0x33333333) + (output & 0x33333333);
  output = ((output + (output >> 4) & 0xF0F0F0F) * 0x1010101);
  return output >> 24;
}
```

This one uses a 'reasonable-sized' lookup table.

```

// A 256-entry lookup table will comfortably fit in any L1 dcache.
// But Kiwi requires a mirror mark up to make it produce 4 of these.
static readonly byte [] tally8 = new byte [] {
    0, 1, 1, 2, 1, 2, 2, 3, 1, 2, 2, 3, 2, 3, 3, 4,
    1, 2, 2, 3, 2, 3, 3, 4, 2, 3, 3, 4, 3, 4, 4, 5,
    ...
    3, 4, 4, 5, 4, 5, 5, 6, 4, 5, 5, 6, 5, 6, 6, 7,
    4, 5, 5, 6, 5, 6, 6, 7, 5, 6, 6, 7, 6, 7, 7, 8,
};

public static uint tally03(uint ind)
{
    uint a0 = (ind >> 0)&255;
    uint a1 = (ind >> 8)&255;
    uint a2 = (ind >> 16)&255;
    uint a3 = (ind >> 24)&255;
    return (uint)(tally8[a0] + tally8[a1] + tally8[a2] + tally8[a3]);
}

```

The look-up table needs to be replicated to give four copies. The logic synthesiser might ultimately replace such a ROM with combinational gates if this will have lower area.

Which of these works best on FPGA? The experiments on the following link may be discussed in lectures: [Kiwi Bit Tally \(Ones Counting\) Comparison](#)

A good HLS tool should not be sensitive to coding style and should use strength reduction and sub-expression sharing and all standard compiler optimisations. (But we'll soon discuss that layout of data in memories is where we should exercise care and where automated techniques can help less).

(**Strength Reduction** is compiler-speak for replacing one operator with a less costly operator where possible, such as replacing multiply by -1 with subtract from 0).

### 7.0.21 HLS Synthesisable Subset.

Can we convert arbitrary or legacy programs to hardware? Not very well in general. Can we write new HLL programs that compile to good hardware? Yes. But we must stick to the supported subset for synthesis.

Typical HLS restrictions:

- **Program must be finite-state and single-threaded,**
- all recursion bounded,
- all dynamic storage allocation outside of infinite loops (or deallocated again in same loop),
- use only boolean logic and integer arithmetic,
- limited string handling,
- very-limited standard library support,
- be explicit over which loops have run-time bounds.

An early example DJG C-To-V compiler from 1995. [Bubble Sorter Example](#)

Today many commercial HLS tools are widely available: SystemCrafter, Calypto Catapult, SimVision, CoCentric, C-Level Design, Forte Cynthesizer (now acquired by Cadence), C-to-Verilog.com and xPilot now called Vivado HLS, ... other tools are/were HardwareC, SpecC, Impulse-C, NEC Cyber Workbench, Synopsis SynphonyC.

And research HLS tools, such as Kiwi and LegUp, support floating point, pointers and some dynamic storage allocation using DRAM banks as necessary,

The advantages of using a general-purpose language to describe both hardware and software are becoming apparent: designs can be ported easily and tested in software environments before implementation in hardware. There is also the potential benefit that software engineers can be used to generate ASICs: they are normally cheaper to employ than ASIC engineers! The practical benefit of such approaches is not fully proven, but there is great potential.

The software programming paradigm, where a serial thread of execution runs around between various modules is undoubtedly easier to design with than the forced parallelism of expressions found in RTL-style coding. Ideally, a new thread should only be introduced when there is a need for concurrent behaviour in the expression of the design.

A product from COMPILOGIC is typical and claimed the following:

- Compile C to RTL Verilog for synthesis to FPGA and ASIC hardware.
- Compile C to Test-Bench for Verilog simulation.
- Compiler options to control design's size and performance.
- Global analysis optimizes C-program intentions in hardware.
- Automatic and controlled parallelism and pipelining.
- Generates readable Verilog for integration and modification.
- Options to assist tracing/debugging HDL generated.
- Includes command line and GUI programmer's workbench.

but like many domain names allocated to companies in this area in the last 15 years, this one too has expired.

However, we cannot compile general C/C++ programs to hardware: they tend to use too many language features. Java and CSharp are better, owing to stronger typing and banning of arithmetic on object handles (all subscription operations apply to first-class arrays).

### 7.0.22 Unrolling: Trading time for space.

A given function can generally be done in half as many clock cycles using twice as much silicon, although name aliases and control hazards (dependence on run-time input data) can limit this. As well as the C/C++ input code we require additional directives over speed, area and perhaps power. The area directives may specify the number of RAMs or how to map arrays into shared DRAM. Trading (or folding) such time for space is basically a matter of unwinding loops or introducing new loops.

Hazards can limit the amount of unrolling possible, including limited numbers of ports on RAMs and user-set budgets on the number of certain components (FUs) instantiated, such as adders or multipliers.

Unrolling can be:

- **automatic**, where the tool aims to meet a given throughput and clock frequency,
- **manual**, based on pragmas or other mark up inserted in the source code.

### 7.0.23 Pipelined Scheduling - One Basic Block

After some amount of loop unrolling, we have an expanded control-flow graph that has larger basic blocks than in the original HLL program. In classical HLS, each basic block of the expanded graph is given a time-domain static schedule.

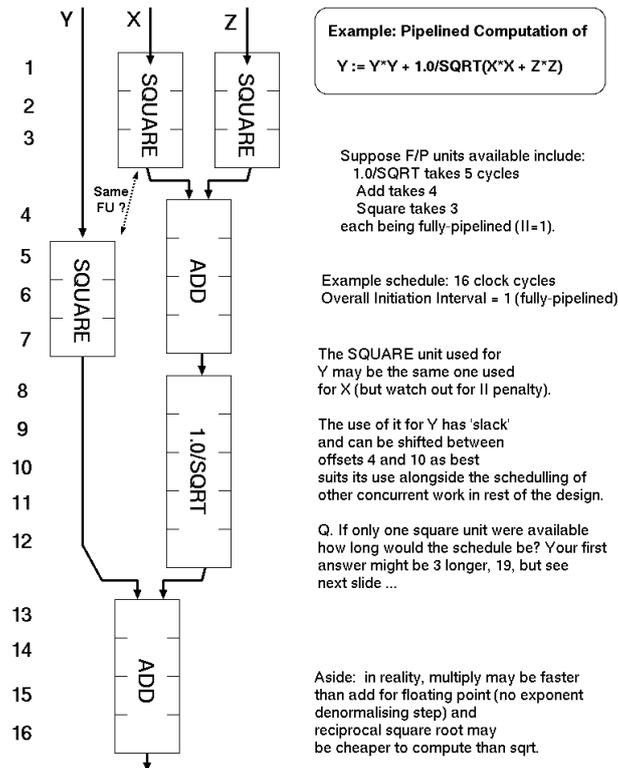


Figure 7.7: Example static schedule for a basic block containing a single assignment.

The diagram we have just considered was a little naive: it was constructed assuming that the FUs were non-pipelined (i.e. their II was the same as their latency). Non-pipelined versions of some kinds of FU are more compact (less silicon area) than pipelined equivalents, so they do crop up. But we were told our floating-point FUs had II=1.

Our new design uses only one SQUARE unit at the expense of latency being one higher. II is often more critical than latency for HLS, and our II is only 3.

A good heuristic for scheduling is to start the operations which have the largest processing delay as early as possible. This is called a *list schedule*. But integer linear programming packages are often used to find an optimum schedule and resource use trade off.

Our example generated only one output. A basic block schedule typically contains multiple assignments, with sub-expressions and functional units being reused in the time domain throughout the schedule and shared between output assignments.

To avoid RaW hazards within one basic block, all reads to a variable or memory location must be scheduled before all writes to the same.

The name alias problem means we must be conservative in this analysis when considering whether array subscripts are equal. This is 'undecidable' in general theory, but often doable in practice. Indeed many subscript expressions will be simple functions of loop 'induction variables' whose pattern we need to understand for high performance.

### 7.0.24 Pipelined Scheduling - Between Basic Blocks

Multiple basic blocks, even from one thread, will be executing at once, owing to pipelining. Frequently, an inner loop consists of one basic block repeated, and so it is competing with itself for structural resources and data hazards. This is **loop pipelining**. Where an outer loop is requested to be pipelined, all loops inside must

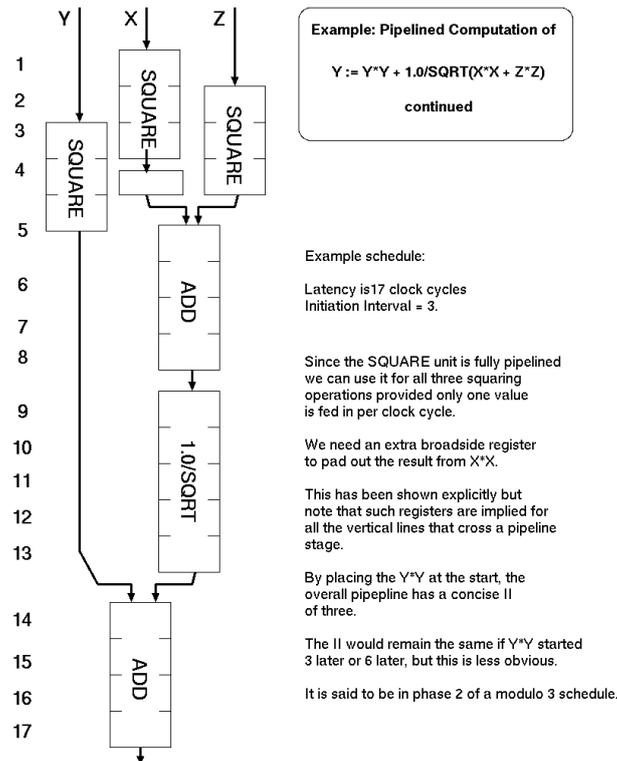


Figure 7.8: The same example but now with just one SQUARE unit and exploiting the fact it is fully-pipelined.

be pipelined.

Each time offset in a block's schedule needs to be checked for structural hazards against the resource use of all other blocks that are potentially running at the same time.

Every block has its own static schedule of determined length. The early part of the schedule generally contains control-flow predicate computation to determine which block will run next. This can be null if the block is the initialisation code for a subsequent loop (i.e. the basic block ends on a branch destination rather than on a conditional branch). The later part of the block contains data reads, data writes and ALU computations. Data operations can also occur in the control phase but when resources are tight (typically memory read bandwidth) the control work should be given higher scheduling priority and hence remain at the top of the schedule.

Per thread there will be at most one control section running at any one time. But there can be a number of data sections from successors and from predecessors still running.

The interblock schedule problem has exponential growth properties with base equal to the average control-flow fan out, but only a finite part needs to be considered governed by the maximal block length. As well as avoiding structural hazards, the schedule must contain no RaW or WaW hazards. So a block must read a datum at a point in its schedule after any earlier block that might be running has written it. Or if at the same time, forwarding logic must be synthesised.

It may be necessary to add a 'postpad' to relax the schedule. This is a delay beyond what is needed by the control flow predicate computation before following the control flow arc. This introduces extra space in the global schedule allowing more time and hence generally requiring fewer FUs.

In the highlighted time slot in figure 7.9, the D3 operations of the first block are concurrent with the control and data C1+D1 operations of a later copy of itself when it has looped back or with the D1 phase of Block 2 if it exited from its tight loop.

Observing **sequential consistency** imposes a further constraint on scheduling order: for certain blocks, the order of operations must be (partially) respected. For instance, in a shared memory, where a packet is being

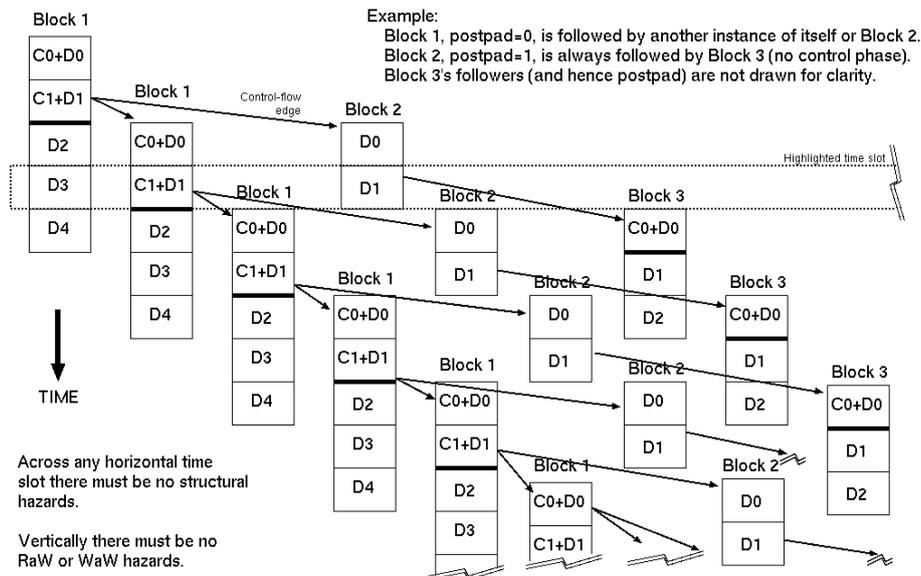


Figure 7.9: Fragment of an example inter-block initiation and hazard graph.

stored and then signalled ready with a write to a flag or pointer in the same RAM, the signalling operation must be kept last. (This is not a WaW hazard since the writes are to different addresses in the RAM.) Observing these limits typically results in an expansion of the overall schedule.

## 7.1 HLS Functional Units (FUs)

The output from HLS is RTL. The RTL will use a mixture of operators supported by the back end logic synthesiser, such as integer addition, and structural components selected from an HLS functional unit (FU) block library, such as floating-point multiply.

### 7.1.1 Functional Unit (FU) Block Properties

Apart from the specification of the function itself, such as multiply, a block that performs a function in some number of clock cycles can be characterised using the following metrics:

- int **Precision**
- bool **Referentially-Transparent (Stateless)**: always same result for same arguments.
- bool **EIS (An end in itself)**: Has unseen side effects such as turning on an LED
- bool **FL or VL**: Fixed or Variable latency
- int **Block latency**: cycles to wait from arguments in to result out (or average if VL)
- int **Initiation Interval**: minimum number of cycles between starts (arguments in time) (or average if VL)
- real **Energy**: Joules per operation - normally a few nanojoules for a ...
- real **Gate count or area**: Area is typically given in square microns or, for FPGA, number of LUTs.

A unit whose initiation interval is one is said to be **'fully pipelined'**.

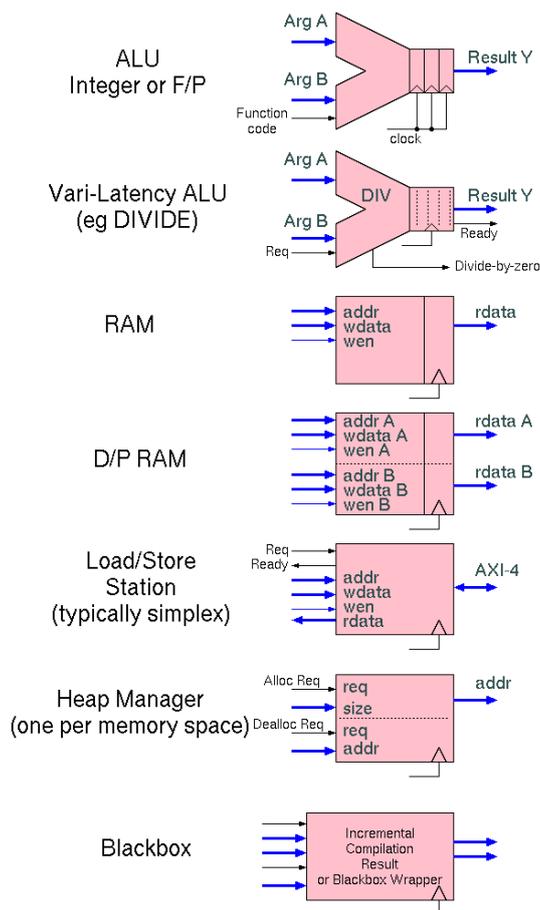


Figure 7.10: Typical examples of FUs deployed by an HLS compiler.

In today's ASIC and FPGA technology, combinational add and subtract of up to 32-bit words is typical. But RAM read, multiply and divide are usually allocated at least one pipeline cycle, with larger multiplies and all divides being two or more. For 64-bit word widths, floating point or RAMs larger than L1 size (e.g. 32 KByte), two or more cycle latency is common, but with an initiation interval of one ( $ii=1$ ).

### 7.1.2 Functional Unit (FU) Chaining

Naively instantiating standard FUs can be wasteful of performance, precision and silicon area. Generally, if the output of one FU is to be fed directly to another then some optimisation can be made and many sensible optimisations involve changes of state encoding or algorithm that are beyond the back-end logic synthesiser.

A common example is an associative reduction operator such as floating-point addition in a scalar product. In that example, we do not wish to denormalise and round-and-renormalise the operand and result at each addition. This

- adds processing latency in clock cycles or gate delay on critical path,
- requires modulo scheduling (Lam) for loops shorter than the reduction operator's latency,
- uses considerable silicon area.

For example, in 'When FPGAs are better at floating-point than microprocessors' (Dinechin et al 2007), it is shown that a fixed-point adder of width greater than the normal mantissa precision can reduce/eliminate underflow errors and operate with less energy and fewer clock cycles.

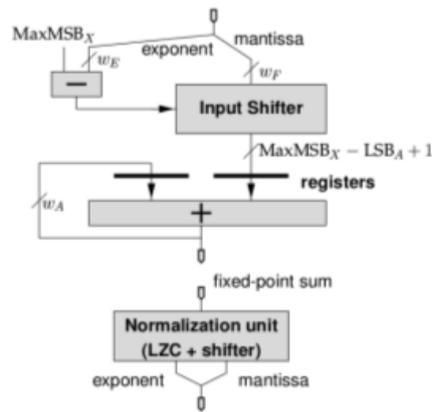


Figure 5: The proposed accumulator. Only the registers on the accumulator itself are shown, the rest of the design is combinatorial and can be pipelined arbitrarily.

Figure 7.11: Fixed-Point Accumulator Proposal.

Their approach is to denormalise the mantissa on input from each iteration and renormalise once at the end when the result is needed. Even a ‘running-average’ example is generally used in a decimated form (i.e. only every 10th or so result is looked at).

## 7.2 Discovering Parallelism: Classical HLS Paradigms

The well-known map-reduce paradigm allows the map part to be done in parallel. An associative reduction operator gives the same answer under diverse bracketings. Common associative operators are addition, multiplication, maximum and bitwise-OR. Our first example uses xor. Here we look at some examples where the bodies of an iteration may or may not be run in parallel.

```
public static int associative_reduction_example(int starting)
{
    int vr = 0;
    for (int i=0;i<15;i++) // or also i+=4
    {
        int vx = (i+starting)*(i+3)*(i+5);
        vr ^= ((vx&128)>0 ? 1:0);
    }
    return vr;
}
```

Where the loop variable evolves linearly, variables and expressions in the body that depend linearly on the loop variable are termed **linear induction variables/expressions** and can be ignored for loop classification since their values will always be independently available in each loop body with minimal overhead.

A **loop-carried dependency** means that parallelisation of loop bodies is not possible. Often the loop body can be split into a part that is and is-not dependent on the previous iteration, with the is-not parts run in parallel at least.

```
public static int loop_carried_example(int seed)
{
    int vp = seed;
    for (int i=0;i<15;i++)
    {
        vp = (vp + 11) * 31241/121;
    }
    return vp;
}
```

A value read from an array in one iteration can be **loop forwarded** from one iteration to another using a holding register to save having to read it again.

```
static int [] foos = new int [10];
static int ipos = 0;
public static int loop_forwarding_example(int newdata)
{
    foos[ipos++] = newdata;
    ipos %= foos.Length;
    int sum = 0;
    for (int i=0;i<foos.Length-1;i++)
    {
        sum += foos[i]^foos[i+1];
    }
    return sum;
}
```

**Data-dependent** exit conditions also limit parallelisation, although a degree of speculation can be harmless. How early in a loop body the exit condition can be determined is an important consideration and compilers will pull this to the start of the block schedule (as illustrated in figure 7.7). When speculating we continue looping but provide a mechanism to discard unwanted side effects.

```
public static int data_dependent_controlflow_example(int seed)
{
    int vr = 0;
    int i;
    for (i=0;i<20;i++)
    {
        vr += i*seed;
        if (vr > 1111) break;
    }
    return i;
}
```

The above examples have been demonstrated using Kiwi HLS on the following link [Kiwi Common HLS Paradigms Demonstrated](#). [Wikipedia:Loop Dependence](#)

Compared with software compilers, like gcc, HLS tools are currently relatively immature. Here is an example from Vivado HLS of a false positive on loop-carried dependency. The following refuses to compile. [LINK](#)

```
//directive PIPELINE set
uint8_t process_image(uint8_t pxVal)
{
    static unsigned int x = 0;
    static uint8_t lines[16][WIDTH];

    //directive UNROLL set
    for(int i=0; i < 15; i++) {
        lines[i][x] = lines[i + 1][x];
    }
    lines[15][x] = pxVal;
    uint8_t result = lines[1][x];
    x = (x + 1) == WIDTH ? 0 : (x + 1);
    return result;
}
```

Why does this fail? When we convert from a 2-D to a 1-D array by multiplying up the indices we lose information that the subscripts are manifestly distinct.

### 7.2.1 Modulo Scheduling

The high-level expression of a basic block may contain some number of arithmetic operators, such as 9 additions and 10 multiplies for a 10-stage FIR filter. But we may wish to render this in hardware using fewer FUs. For instance, to do this in 3 clock cycles, 3 adders and 4 multipliers are nominally sufficient as this will meet the required basic operations-per-second budget.

But the FUs will often be pipelined and the output from one may not be ready for the input to the next in time.

If data is streaming, then the problem should be solvable with the minimal number of FUs given sufficient pipelining. The pipeline initiation interval (II) will be 3, but the latency larger than 3.

Streaming hardware designs with low II are surprisingly complex and difficult to find when the ALUs available are heavily pipelined. Floating point ALUs tend to have multi-cycle delay (latency of four to six cycles is common for add and multiply in FPGA).

A scheduler program will solve the problem of making a static mapping of operations to FUs. It can be phrased nicely as an ILP problem (Sittel/Koch 'ILP-based Modulo Scheduling and Binding for Register Minimization' FPL2018). The result is called a **modulo schedule**. The modulo basis will be the II, 3. Once the schedule is computed it is then fairly simple to render a sequencer circuit and additional holding registers as needed.

A design with initiation interval of one is said to be *fully-pipelined* whereas a higher initiation interval can generally be supported with less silicon area using *modulo scheduling*.

#### Example 1: The Running Sum

If the adder has unity latency the running sum example is simple. For a two-cycle adder with three inputs, it is also trivial. The minimal circuit for latency-of-two adders with only two inputs is surprisingly complex.

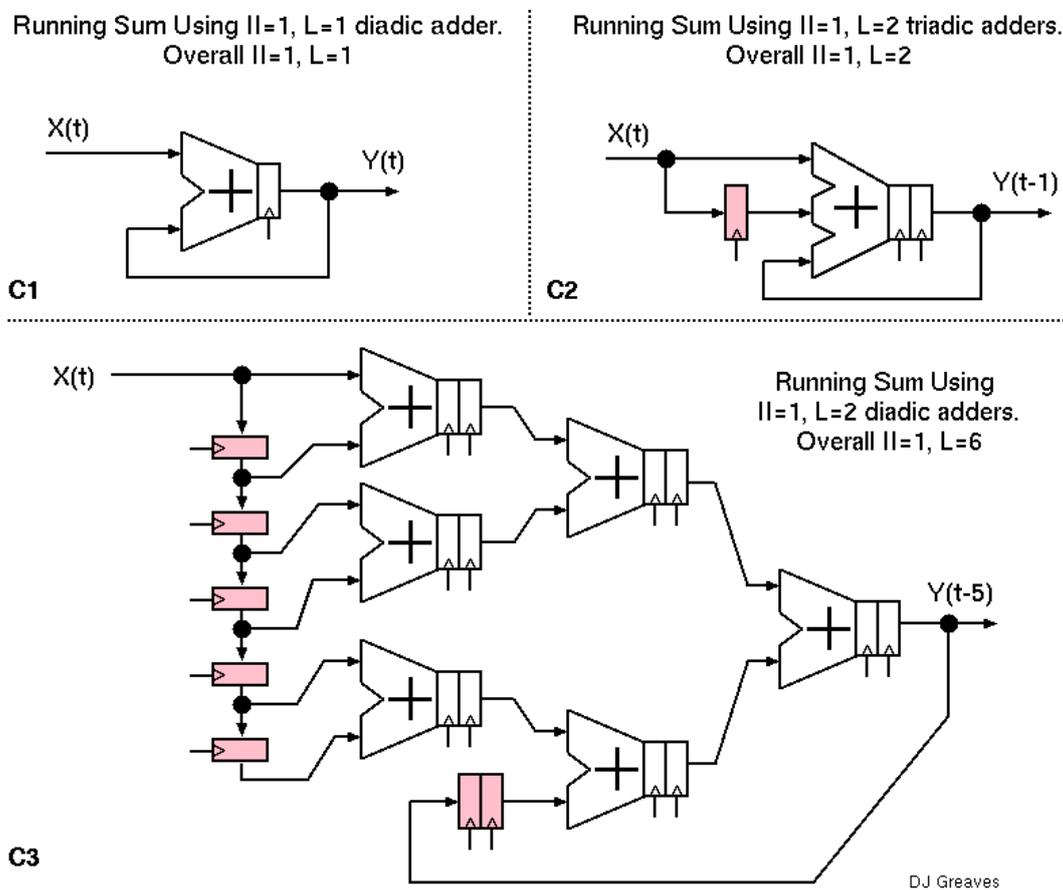


Figure 7.12: Various circuits that perform a running sum.

#### Example 2: Bi-quad Filter Element

Rearranging the operator association can greatly help with producing a good schedule.

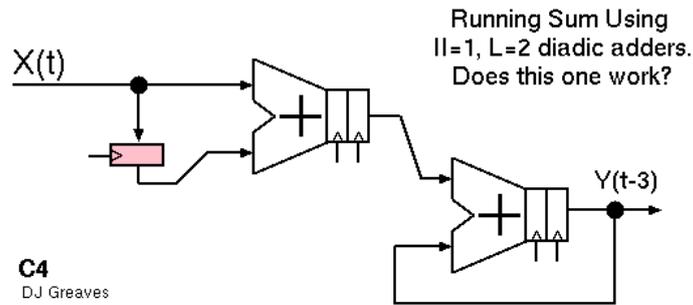


Figure 7.13: A slightly odd-looking possible solution. *Ex: check it works.*

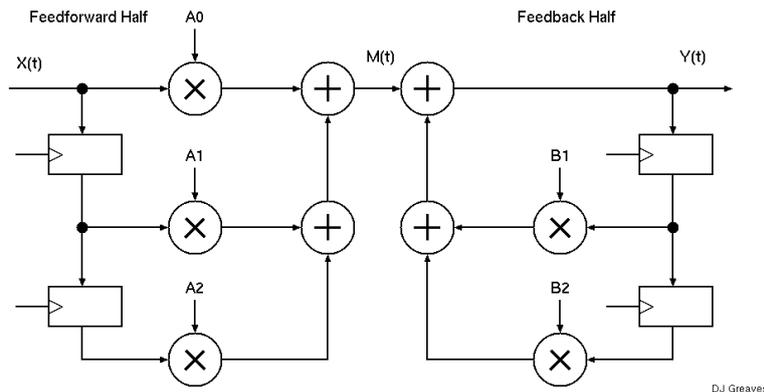


Figure 7.14: Standard circuit for a bi-quad filter element.

## 7.2.2 Memory Banking and Widening

Whether computing on standard CPUs or FPGA, memory bandwidth is often a main performance bottleneck. Given that data transfer rate per-bit of read or write port is fixed, two solutions to memory bandwidth are to use **multiple banks** or **wide memories**. Multiple banks (aka channels) can be accessed simultaneously at different locations, whereas memories with a wider word are accessed at just one location at a time (per port). Both yield more data for each access. Both also may or may not need lane steering or a crossbar routing matrix, depending on the application and allowable mappings of data to processing units.

The best approach also depends on whether the memories are truly random access. SRAM is truly random access, whereas DRAM will have different access times depending on what state the target bank (i.e. bit plane) is in.

With multiple RAM banks, data can be arranged randomly or systematically between them. To achieve ‘random’ data placement, some set of the address bus bits are normally used to select between the different banks. Indeed, when multiple chips are used to provide a single bank, this arrangement is inevitably deployed. The question is which bits to use.

Using low bits causes a fine-grained interleave, but may either destroy or leverage spatial locality in access patterns according to many details.

Ideally, concurrent accesses hit different banks, therefore providing parallelism. Where data access patterns are known in advance, which is typically the case for HLS, then this can be maximised or even ensured by careful bank mapping. Interconnection complexity is also reduced when it is manifest that certain data paths of a full cross-bar with never be used. In the best cases (easiest applications), we need no lane-steering or interconnect switch and each processing element acts on just one part of the wide data bus. This is basically the GPU architecture.

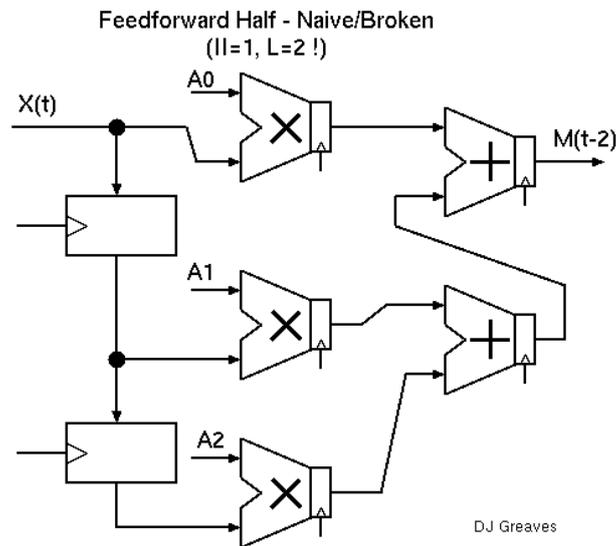


Figure 7.15: Feed-forward part of the bi-quad possible design?

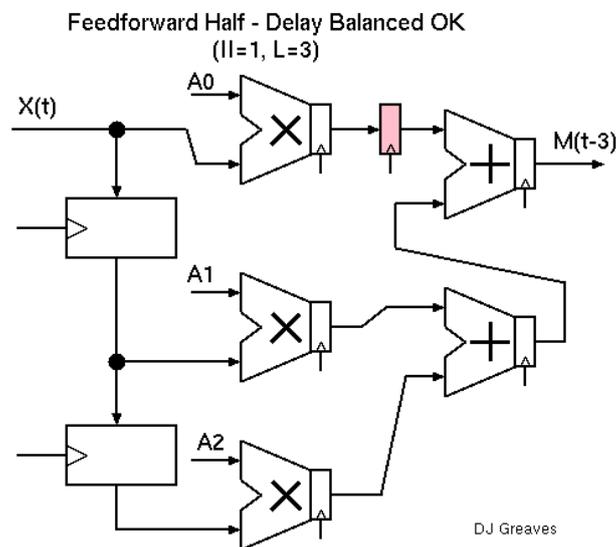


Figure 7.16: Corrected feed-forward section of bi-quad filter element.

### 7.2.3 Data Layout for Burrows-Wheeler Transform

The BWT of a string is another string of the same length and alphabet. According to the definition of a transform, it is information preserving and has an inverse. We shall not define it in these notes. The following code makes efficient perfect string matches of needles in a given haystack using the BWT. It uses lookup in the large Rank array that is pre-computed from BWT which itself is a precomputed transform of the haystack. It also uses the Tots\_before array, but this is very small and easily fits in BRAM on an FPGA. The Rank array is 2-D, being indexed by character, and contains integers ranging up to the haystack size (requiring more bits than a character from the alphabet).

The problem can be that, for big data, such as Giga-base DNA genomes, the Rank array may be too big for available the DRAM. The solution is to decimate the Rank array by some factor, such as 32, and then only store every 32nd row in memory. When lookup does not fall on a stored row, the row's contents are interpolated on-the-fly. This does require the original BWT-transformed string is stored, but this may well be useful for many related purposes anyway.

Access patterns to the Rank array will exhibit no spatial or temporal locality (especially at the start of the search

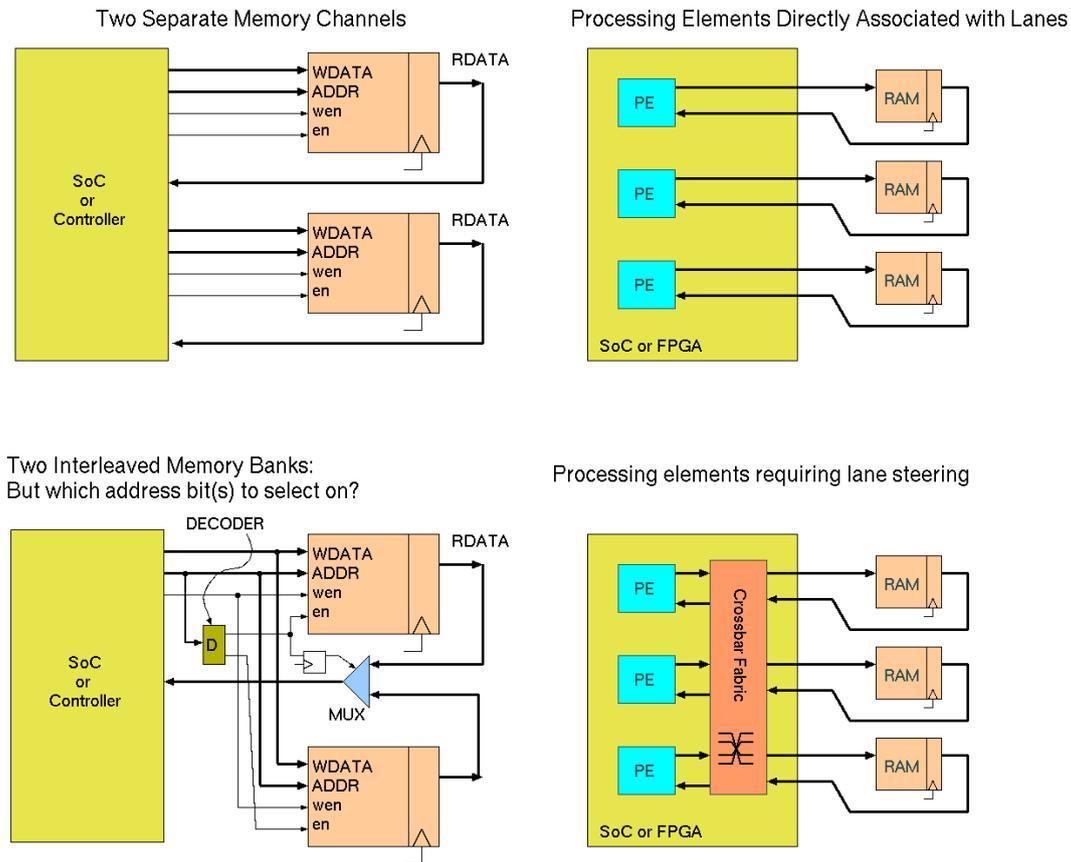


Figure 7.17: Alternative Memory Bank Structures giving wider effective bus width.

when *start* and *end* are well separated. If only one bank (here we mean channel) of DRAM is available, then random access delays dominate. (The only way to get better performance is task-level parallelism: searching for multiple needles at once, which at least overcomes the round-trip latency to the DRAM, but not the low performance intrinsic to no spatial locality). However, one good idea is to store the BWT fragment in the same DRAM row as the ranking information. Then only one row activation is needed per needle character. For what factor of decimation is the interpolator not on the critical path? This will depend mainly on how much the HLS compiler chooses (or is commanded via pragmas) to unwind it. In general, when aligning data in DRAM rows, sometimes a pair of items that are known to both be needed can be split into different rows. In which case storing everything twice may help, with one copy offset by half a row length from the other, since then it is possible to manually address the copy with the good alignment.

Pico Computing White PaperKiwi Implementation

### 7.2.4 Smith-Waterman D/P Data Dependencies

The Smith-Waterman algorithm has become an icon for FPGA acceleration. Two strings are matched for edit distance.

A quadratic algorithm based on dynamic programming is used. The maximum score needs to be found in a 2-D array where each score depends on the three immediate neighbours with lower index as shown in figure 7.20. Zeros are inserted where subscripts would be negative at the edges.

Acceleration is achieved by computing many scores in parallel. There is no simple nesting of two for loops that can work. Instead, items on a diagonal frontier can be computed in parallel. Normally one string behaves as a needle with perhaps 1000 characters and the other is a haystack streamed from a filesaver.

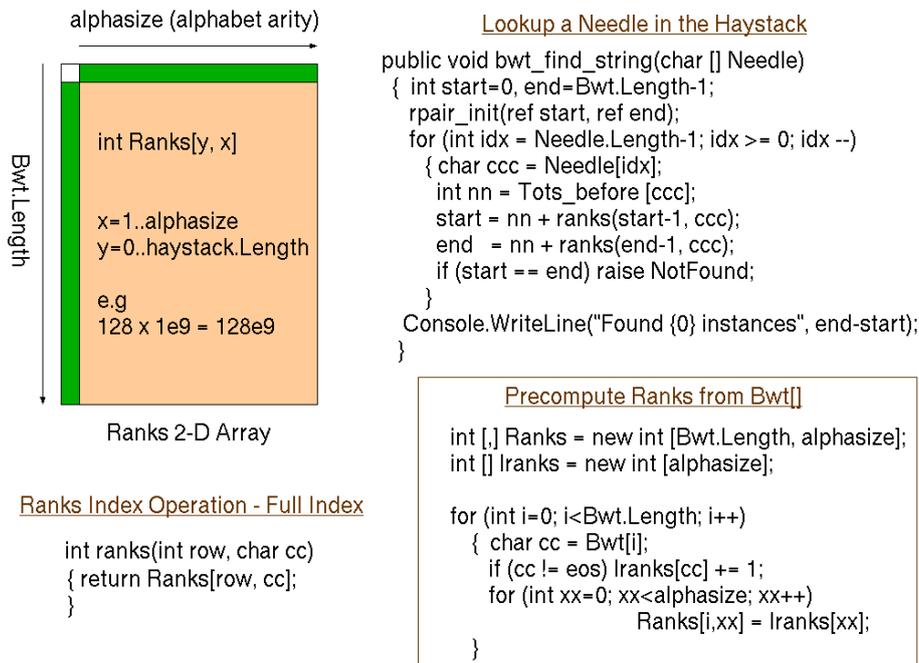


Figure 7.18: Lookup procedure when string searching using BWT.

We shall discuss suitable hardware architectures on the example sheet.

### 7.2.5 Polyhedral Address Mapping

A number of restricted mathematical systems have useful results in terms of decidability of certain propositions. A restriction might be that expressions only multiply where one operand is a constant and a property might be that an expression always lies within a certain intersection of  $n$ -dimensional planes. There is a vast literature regarding integer linear inequalities (linear programming) that can be combined with significant results from Presburger (Preburger Arithmetic) and Mine (The Octagon Domain) as a basis for optimising memory access patterns within HLS.

We seek transformations that:

1. compact provably sparse access patterns into packed form or
2. where array read subscripts can be partitioned into provably disjoint sets that can be served in parallel by different memories, or
3. where data dependencies are sufficiently determined that thread-future reads can be started at the earliest point after the supporting writes have been made so as to meet all read-after-write data dependencies.

#### Improving High Level Synthesis Optimization Opportunity Through Polyhedral Transformations

A set of nested loops where the bounds of inner loops depend on linear combinations of surrounding loop (aka induction) variables defines a polyhedral space or polytope. This space is scanned by the vector consisting of the induction variables. Under polyhedral mapping, the loop nesting order may be changed and affine transformations are applied to many of the loop variables with the aim of exposing parallelism and/or re-packing array subscripts to use less overall memory.

In the example, both the inner and outer loop bounds are transformed. The skewing of block 1 enables it to be parallelised with computable data dependencies. The following block, that runs on its output, can be run in

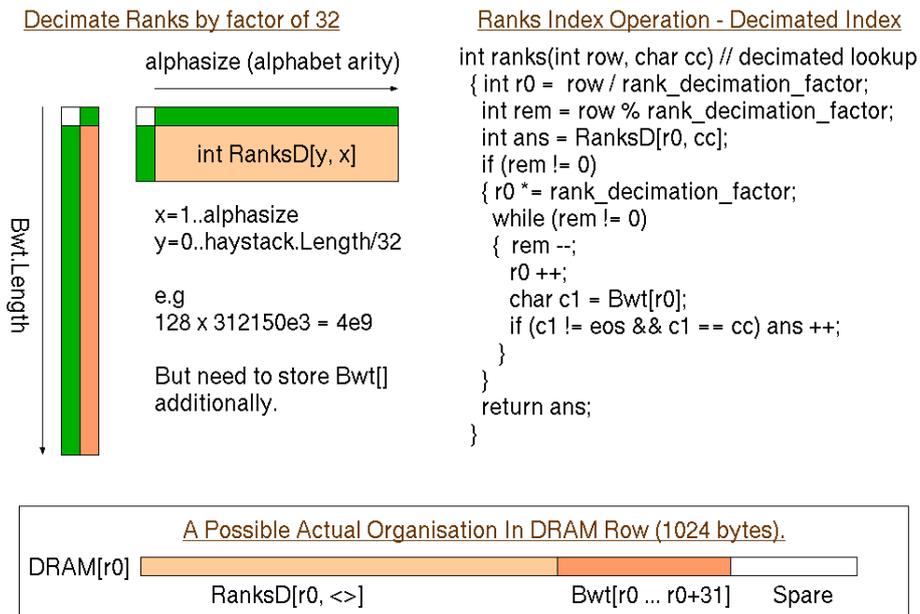


Figure 7.19: Compacted Rank Array for BWT and a sensible layout in a DRAM row.

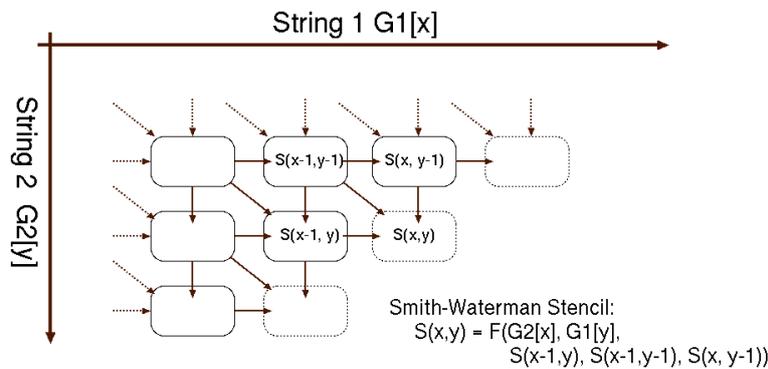


Figure 7.20: Data dependencies (slightly simplified) in Smith-Waterman Alignment Matcher.

parallel after an appropriately delayed start.

Wikipedia:Polyhedral

For big data, at least one of the loop bounds is commonly an iteration over a file streamed from secondary storage. (The field of automatic parallelisation of nested for loops on arrays has been greatly studied over recent decades for SIMD and systolic array synthesis ...)

Q. Does the following have loop interference ?

```
for (i=0; i<N; i++) A[i] := (A[i] + A[N-1-i])/2
```

A. Yes, at first glance, but we can recode it as two independent loops. ('Loop Splitting for Efficient Pipelining in High-Level Synthesis' by J Liu, J Wickerson, G Constantinides.)

“In reality, there are only dependencies from the first  $N/2$  iterations into the last  $N/2$ , so we can execute this loop as a sequence of two fully parallel loops (from  $0 \dots N/2$  and from  $N/2+1 \dots N$ ). The characterization of this dependence, the analysis of parallelism, and the transformation of the code can be done in terms of the instance-wise information provided by any polyhedral framework.”

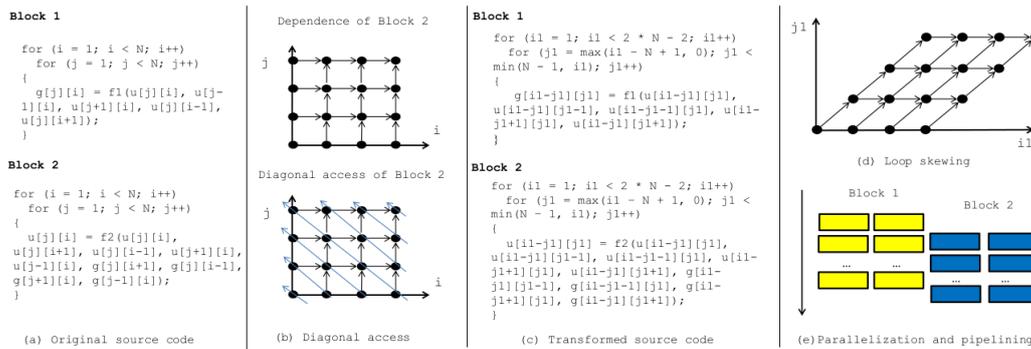


Figure 7.21: Affine transformations (Zuo, Liang et al).

Q. Does the following have loop body inter-dependencies ?

```

for (i=0; i<N; i++) A[2*i] = A[i] + 0.5f;

```

A. Yes, but can we transform it somehow?

### 7.2.6 The Perfect Shuffle Network - FFT Example

A number of algorithms have columns of operators that can be applied in parallel. The FFT is one such example. The operator is commonly called a 'butterfly'. The operator composes two operands (which are generally each complex numbers) and delivers two results. The successive passes can be done in place on one array whose values are passed by reference to the butterfly code in the software version.

Our diagram shows a 16-point FFT, but typically applications use hundreds or thousands. (The code fragment shows a single-threaded implementation that does not attempt to put butterflies in parallel.)

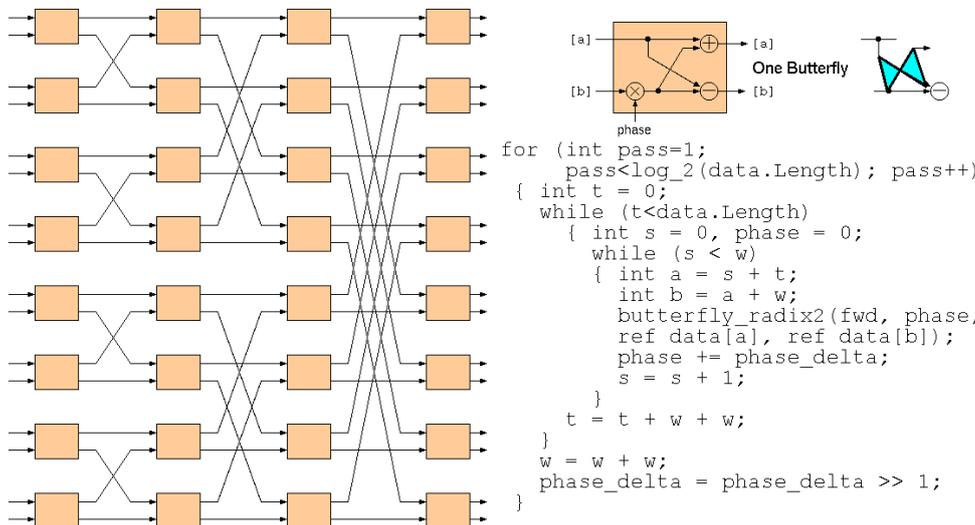


Figure 7.22: Shuffle Data Flow in the Fast Fourier Transform.

The pattern of data movement is known as a shuffle. It is also used in switching and sorting algorithms. The downside of shuffle computations for acceleration is that there is **no packing of data** (structural partitioning of the data array) into spatially-separate memories **that works well for all of the passes**: what is good at the start is poor at the end.

### 7.2.7 Classical HLS: Pros and Cons

Classical HLS performs ‘auto parallelising’ where as much concurrency as possible is found from a single thread of control. The amount of parallelism discoverable in ‘legacy code’ such as C programs is often severely limited by unintentional artefacts of the way the program is expressed ...

Legacy code has a certain amount of parallelism: M. S. Lam and R. P. Wilson. ‘Limits of control flow on parallelism’ by (Lam+Wilson) . In Nineteenth International Symposium on Computer Architecture 1992.

Much of the limit is artificial, as explored in: Limits of Parallelism Using Dynamic Dependency Graphs (Mak+Mycroft)

But today, the best way to achieve parallel performance is to write either in an explicitly parallel language or certainly one that is more declarative or ...

An alternative recent approach can be found in Rust ...

Sometimes **fully-pipelined** HLS is required, where there is no scheduler and new data is input every cycle. Or there may be a simple scheduler with no control flow: e.g. for a small **initiation interval** with new input every, say, 4 cycles.

### 7.2.8 Kiwi: Compiling Concurrent Programs to Hardware

Advert: Current project led by David Greaves and Satnam Singh: [Web Site](#)

[Times Table Very Simple Demo](#)

Kiwi is developing a methodology for hardware design using the **parallel programming constructs** of the CSharp language. Specifically, Kiwi consists of a run-time library for native simulation of hardware descriptions within CSharp and a compiler that generates RTL from stylised .net bytecode. The designer uses more concurrency than ‘natural’ for software. This is mapped to concurrent hardware by the Kiwi tools. Each thread is subject to classical HLS (that generates a static schedule for that thread) but then threads interact dynamically with arbiters and queues.

A number of systems have implemented the ‘parallel-For’ or ‘parallel-ForEach’ loop construct. It is normally a mark-up applied to a standard FOR loop that can be safely ignored such that standard sequential execution will return the correct result, albeit, by taking more time.

For example, in C++ OpenMP, one puts

```
#pragma omp parallel for
```

and in CSharp one can map a delegate using

```
Parallel.For(0, matARows, i => ...)
```

The programmer must be aware of adverse interactions between the loop bodies. Where one body reads a value written by another, the order of scheduling is likely to make a difference. Commutable effects such as increment and bit-set are ok if they are atomic.

Q. What is the role of a parallel-For inside an HLS tool that is searching for parallelism anyway?

A. Classical HLS tools normally approach FOR-loops using unwinding. Such unwinding typically leads to structural hazards on memory arrays and hence there has emerged a vast literature on memory banking and polyhedral mapping that we have already discussed. Such unwinding typically assumes that there is little control-flow variation between the bodies and is generally applied to inner loops only. A parallel-For is useful for outer loop

unwinding or where the same static schedule is unlikely to hold for each iteration owing to data-dependent control flow or variable-length operations (cache misses, divide etc.).

This example shows outer-loop parallelisation for matrix multiply. The inner loops should be unwound to some suitable extent by the HLS tool's normal behaviour.

```
// https://docs.microsoft.com/en-us/dotnet/standard/parallel-programming/
//   how-to-write-a-simple-parallel-for-loop
// A basic matrix multiplication.
// Parallelize the outer loop to partition the source array by rows.
Parallel.For(0, matARows, i =>
{
    for (int j = 0; j < matBCols; j++)
    {
        double temp = 0;
        for (int k = 0; k < matACols; k++)
        {
            temp += matA[i, k] * matB[k, j];
        }
        result[i, j] = temp;
    }
}); // Parallel.For
```

But is there an ideal memory layout or banking model for matrix multiplication ?

Note that hardware acceleration of matrix multiplication is currently (2017) receiving a lot of attention as it is the primary cost in convolutional neural networks.

- **Automatic Parallel Inference:** has had limited success in the past but the field is still evolving.
- **Manual Parallel Markups:** have been added to most imperative high-level languages recently, and they also allow the programmer to assert that there are no name aliases which sometimes increases the available parallelism by a factor of 10 or more.

By 'effects' we mean side effects: ie. imperative mutations of the surrounding state (such a write to file).

A number of operations fall into classes of *atomic commutable effects*. These typically operate on a single word, where a load or store is intrinsically an atomic action, but involve both a load and a store and so are potentially non-atomic in some implementations.

```
counter += 3;
flags |= (1<<9);
counter -= 21;
flags |= (1<<10);
```

Classic examples of atomic operations are test-and-set, increment, and bit-set. A pair of increments may be commuted in order without altering the final result stored in the word. Clearly increments and decrements of any amount can be commuted provided the range of the underlying word is not exceeded. Bit-sets of the same or different bits within the word may also be commuted with each other. But bit-set and increment to the same word cannot be commuted and are said to be in different atomic effects classes (by DJG at least). But note that test-and-set (or compare-and-swap) operations, despite being (and needing to be) atomic, are **not** commutable: they are not merely side-effecting, they return a result that generally affects the caller's behaviour.

Where all commutable atomic effects on a shared variable within the bodies of parallel code are in the same effects class, then the resulting composition is *race free*. In other words, the system will accumulate the same result in the shared variable regardless of scheduling order.

```

class primesya
{
    static int limit = 1 * 1000;
    static bool [] PA = new bool[limit];

    // This input port, vol, was added to make some input data volatile.
    [Kiwi.OutputWordPort(31, 0)][Kiwi.OutputName("count")] static uint count = 0;
    static int count1 = 0;
    [Kiwi.OutputWordPort(31, 0)][Kiwi.OutputName("elimit")] static int elimit = 0; // The main parameter (abscissa).

    // The evariant_master is also edited by a sed script that runs an individual experiment.
    // For fair comparison with mono native this needs to be a compile time constant.
    const int evariant_master = 0;
    [Kiwi.OutputWordPort(31, 0)][Kiwi.OutputName("evariant")] static int evariant_output = evariant_master;

    [Kiwi.HardwareEntryPoint()]
    public static void Main()
    {
        bool kpp = true;
        elimit = limit;
        Kiwi.KppMark("START", "INITIALISE"); // Waypoint
        Console.WriteLine("Primes Up To " + limit);
        // Clear array

        count1 = 2; count = 0; // RESET VALUE FAILED AT ONE POINT: HENCE NEED THIS LINE
        for (int woz = 0; woz < limit; woz++)
        {
            PA[woz] = true;
            Console.WriteLine("Setting initial array flag to hold : addr={0} readback={1}", woz, PA[woz]); // Read back.
        }

        Kiwi.KppMark("wp2", "CROSSOFF"); // Waypoint
        int i, j;

        for (i=2;i<limit;i++) // Can our predictor cope with the standard optimisations?
        {
            // Cross off the multiples - optimise by skipping where the base is already crossed off.
            if (evariant_master > 0)
            {
                bool pp = PA[i];
                Console.WriteLine(" tnow={2}: scanning up for live factor {0} = {1} ", i, pp, Kiwi.tnow);
                if (!pp) continue;
                count1 += 1;
            }
            // Can further optimise by commencing the cross-off at the factor squared.
            j = (evariant_master > 1) ? i*i : i+i;
            if (j >= limit)
            {
                Console.WriteLine("Skip out on square");
                break;
            }
            for (; j<limit; j+=i)
            {
                Console.WriteLine("Cross off {0} {1} (count1={2})", i, j, count1);
                Kiwi.Pause(); PA[j] = false;
            }
        }
        Kiwi.KppMark("wp3", "COUNTING"); // Waypoint
        Console.WriteLine("Now counting");
        // Count how many there were and store them consecutively in the output array.
        for (int w = 0; w < limit; w++)
        {
            if (PA[w]) count += 1;
            Console.WriteLine("Tally counting {0} {1}", w, count);
        }
        Console.WriteLine("There are {0} primes below the natural number {1}.", count, limit);
    }
}

```

Black arcs indicate state trajectories. Blue arcs indicate operations on structural resources such as ALUs and RAMs.

### Kiwi Primes Demo Sieve of Eratosthenes

#### Misc notes on Classical HLS:

- Works well when there is little or no cycle time variation. Not so good with DRAM+cache or floating point.
- Creates a precise schedule of addresses on register file and RAM ports and ALU function codes.
- Typically unwinds inner loops by some factor.
- Can cope with data-dependent control flow. But relies on predicate hoisting when datapath is heavily pipelined (rather than speculation).

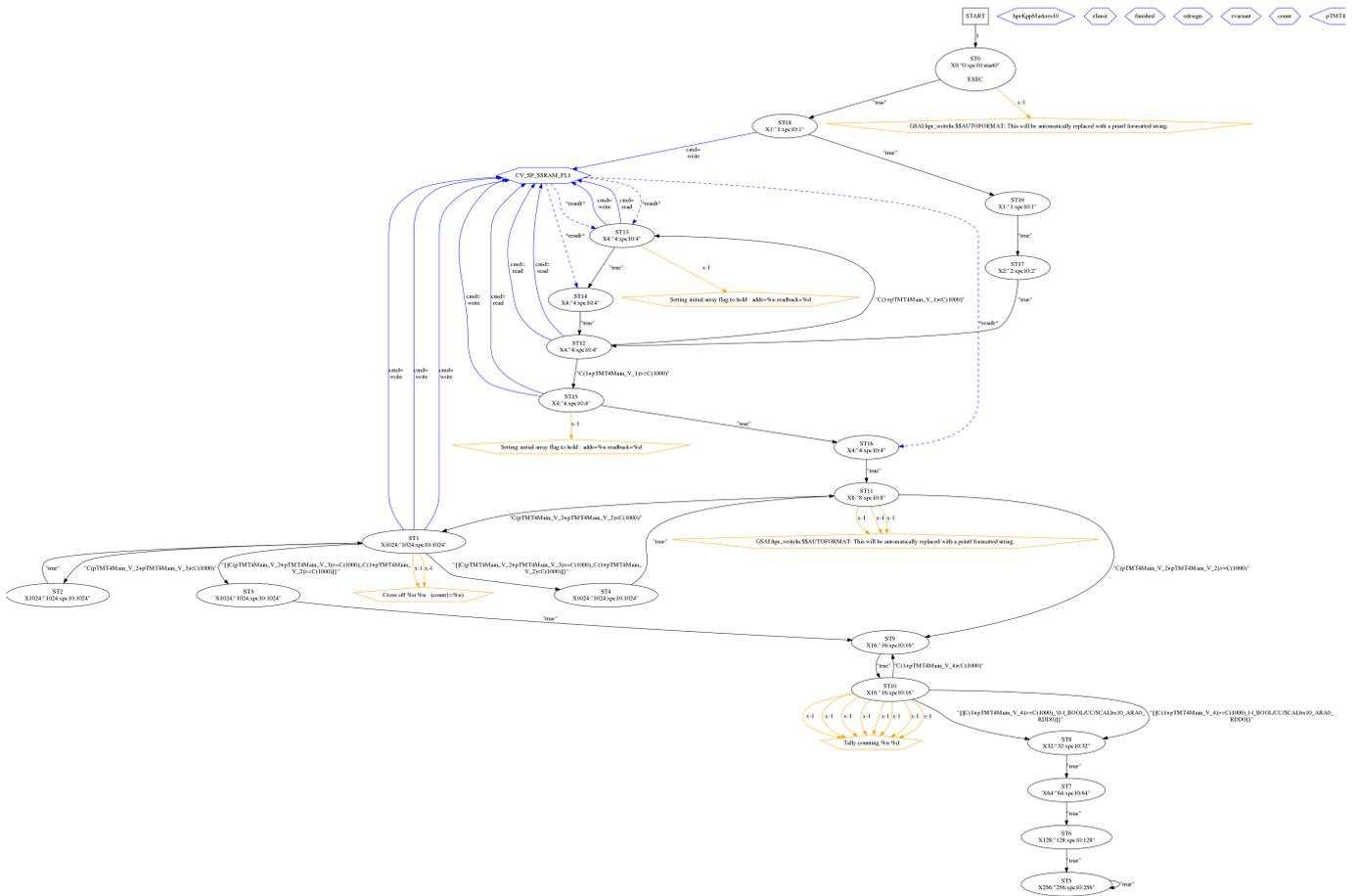


Figure 7.23: Sieve-based primes program: sequencer states.

- Data-dependent control flow and RAM bandwidth ultimately limit parallelism.

We do not want to deploy a lot of resource to support seldom-used data paths. **Profile-directed guidance** or explicit datapath description hints, such as **unroll** pragmas, are needed to select a good RAM arrangement and datapath structure. In the absence of profile information from actual runs, we can either assume sequencer states are equi-probable or better solve balance equations based on the flowgraph entry node being visited exactly once.

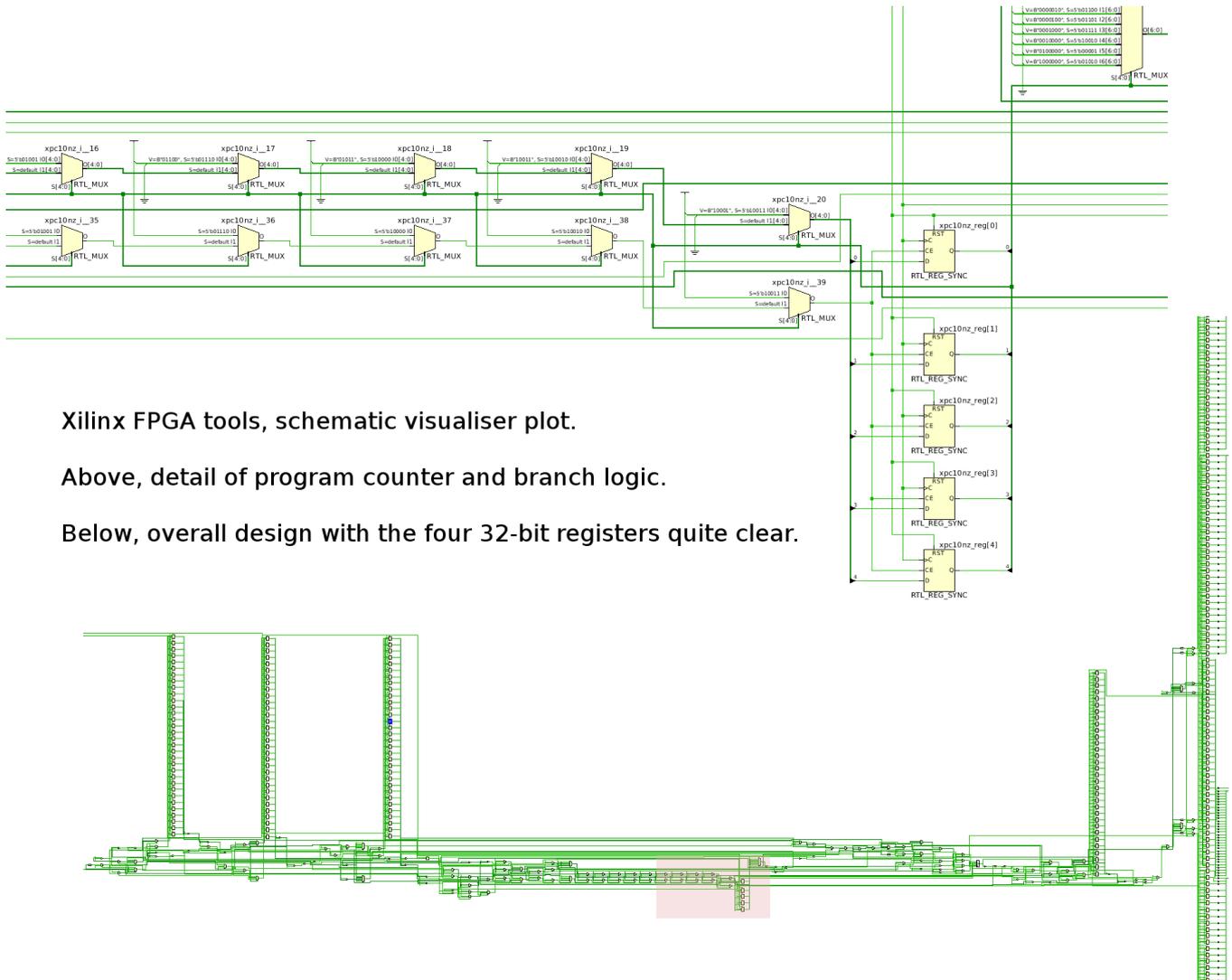
For example, the best mapping of the record fields *x* and *y* to RAMs is different in the two foreach loops:

```

class IntPair
{
    public bool c;    public int x, y;
}

IntPair [] ipairs = new IntPair [1024];

void customer(bool qcond)
{
    int sum1 = 0, sum2 = 0;
    if (qcond) then foreach (IntPair pp in ipairs)
    {
        sum1 += pp.x + pp.y;
    }
    else foreach (IntPair pp in ipairs)
    {
        sum2 += pp.c ? pp.y : pp.x;
    }
    ...
}
    
```



Xilinx FPGA tools, schematic visualiser plot.

Above, detail of program counter and branch logic.

Below, overall design with the four 32-bit registers quite clear.

Figure 7.24: Sieve primes program: FPGA schematic visualisation with inset detail.

The fields  $x$  and  $y$  could be kept in separate RAMs or a common one. If  $qcond$  rarely holds then a skinny RAM will serve but the lookup of  $c$  will then add a pipeline delay (assuming read latency of one clock). Whereas if  $qcond$  holds most of the time then keeping  $x$  and  $y$  in separate RAMs or a common wide RAM will boost performance and it hardly matters where  $c$  is stored. The skinny solution in the Depending on the implementation technology, wide RAMs may or may not be better than skinny ones. Also, RAMs whose word size is not a power of 2 might be readily available in FPGAs where a RAM is aggregated from small tiles. For instance, the Virtex 7 FPGA has, at the lowest level, an 18 Kb BRAM tile that can be configured as a 16K x 1, 8K x 2, 4K x 4, 2K x 9, 1K x 18 ...

Whatever is locally best is not necessarily globally best. For instance, the fragment in the example might only be executed once at start up, in which case it should be ignored when considering data layout performance.

Execution profiles must be used to guide the datapath structure and RAM layout. A simple profile will give the relative execution frequency of each operator in the source code. When balancing loads a more complex profile based on the critical path through the value flow graph of the program needs to be used.

The details of data layout in memory are generally very important, especially for DRAM that is non-uniform in access time owing to its banks, rows and its cache lines when cached. Even for a single array in the user's input code, a permutation of the address lines used by the indexing subscripts may well lead to quite a big difference in performance given banked and non-uniform underlying storage. Fortunately, programmers for high-performance scientific computing (big data) are aware of this since the effects of caches and so on are just

as pronounced in pure software implementations.

### 7.2.9 Static versus Dynamic Scheduling

As mentioned in the RTL section of these notes, RAM ports, ALUs, non fully-pipelined components and other shared resources can cause *Structural Hazards*.

Recall the **Structural Hazard**: Cannot proceed with an operation because a resource is in use. To overcome hazards we must use scheduling and arbitration:

- **Scheduling**: deciding the operation order in advance,
- **Arbitrating**: granting access dynamically, as requests arrive.

One scheduling decision impacts on another: ideally need to find a global optimum.

The scheduling and arbitration operations can often be done at **compile-time**, (e.g. for constant-time operations performed by a single behavioural thread). Remainder must be done at **run-time** according to actual input data since some operations may be vari-time and the relative interleaving of different threads is often unpredictable.

Q. If you are going to do something lots of times, is it always more efficient to invest heavily in planning to get a rapid execution ?

A1. Following that approach lead Intel to Itanium VLIW processor. But has that has sunk?

A2. Classical HLS has saved a lot of energy based on static schedules. (Have you tried compressing MPEG on your laptop and then wondered how your mobile manages it so easily ?)

Even without data-dependent control flow, variable-latency operations are incompatible with a completely static schedule. Keeping a large system in global synchronisation is bound to miss opportunities locally available, but overly-fine-grain dynamic scheduling has a lot of management overhead.

The Kiwi HLS flow uses classical HLS on each thread in turn to generate a static schedule for that thread, but these interact dynamically. E.g. using FIFO queues between components. Combined with a server farm we get localised static schedules and global dynamics.

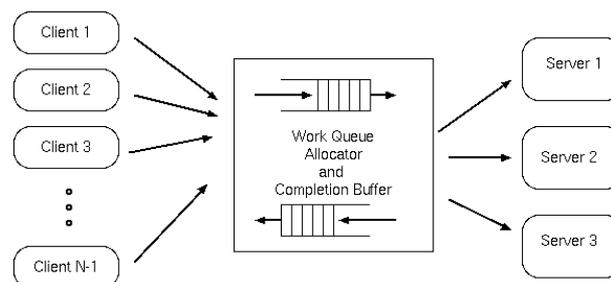


Figure 7.25: Dynamic Load Balancing using Server Farms.

A static computational graph takes less management than approaches that dynamically map work to processing nodes.

The **work stealing** scheduling approach is widely used for dynamic scheduling of workload that is already statically divided into parallel tasks that may dynamically vary in execution time. Wikipedia

A **hardware server** can be shared (contended for) by multiple clients. For example, Bluespec's rich library contains a **Completion Buffer** and other flexible structures for easy creation of pools of servers for dynamic load sharing.

Alternatively: current research, (Ali Zaidi) discards input control flow and compiles to locally to dataflow hardware: The VSFG-S Approach. KiwiC should have a plugin for this soon...

Many hardware designs call for memories, either RAM and ROM. Small memories can be implemented from gates and flip-flops (if RAM). For larger memories, a customised structure is preferable. Large memories are best implemented using separate off-chip device where as sizes of hundreds of kilobytes can easily be integrated in ASICs. Having several smaller memories on a chip takes more space than having one larger memory because of overheads due mainly to address decoding, but, where data can be partitioned (i.e. we know something about the access patterns) having several smaller memories gives better bandwidth and less contention and uses less power for a given performance.

In an imperative HDL, memories readily map to arrays. A primary difference between a formal memory structure and a bunch of gates is the I/O bandwidth: it is not normally possible to access more than one location at a time in a memory. Consider the following Verilog HDL

```
reg [7:0] myram [1023:0]; // 1 kbyte memory
always @(posedge clk) myram[a] = myram[a+1] + 2; // Addresses different - not possible in one cycle.
```

If myram is implemented as an off-the-shelf, single-ported memory array, then it is not possible to read and write it at different addresses in one clock cycle. Compilers which handle RAMs in this way either do not have explicit clock statements in the user code, or else interpret them flexibly. An example of flexible interpretation, is the 'Superstate' concept introduced by Synopsys for their Behavioural Compiler, which splits the user specified clock intervals into as many as needed actual clock cycles. With such a compiler, the above example is synthesisable using a single-ported RAM.

When multiple memories are used, a scheduling algorithm must be used by the compiler to determine the best order for reading and writing the required values. Advanced tools (e.g. C-to-Gates tools and Kiwi) generate a complete 'datapath' that consists of various ALUs, RAMs and register files. This is essentially the execution unit of a custom VLIW (very-long instruction word) processor, where the control unit is replaced with a dedicated finite-state controller.

The decisions about how many memories to use and what to keep in them may be automated or manual overrides might be specified.

### 7.2.10 Shortcomings of Verilog and VHDL as Algorithmic Expression Languages

Verilog and VHDL are languages focused more on simulation than logic synthesis. The rules for translation to hardware that define the 'synthesisable subset' were standardised post the definitions of the language.

Circuit aspects that could readily be determined or decided by the compiler are frequently explicit or directly implicit in the source Verilog text. These aspects include the number of state variables, the size of registers and the width of busses. Having these details in the source text makes the design longer and less portable.

Perhaps the major shortcoming of Verilog (and VHDL) is that the language gives the designer no help with concurrency. That is, the designer must keep in her head any aspect of handshaking between logic circuits or shared reading of register resources. This is ironic since hardware systems have much greater parallelism than software systems.

Verilog and VHDL have allowed vast ASICs to be designed, so in some sense they are successful. But improved languages are needed to meet the following EDA aims:

- Speed of design: time to market,
- Facilitate richer behavioural specification,
- Readily allow time/space folding experiments,

- Greater freedom and hence scope for optimisation in the compiler,
- Facilitate implementation of a formal specification,
- Facilitate proof of conformance to a specification,
- Allow rule-based programming (i.e. a logic-programming sub-language),
- Support modern synchronisation primitives (e.g. join patterns)
- Portability: can be compiled into software as well as into hardware.
- Analysts, such as A DeHon, extrapolate that reconfigurable spatial computing will become better than Von Neumann in general.

New motivation: today, **energy conservation** is often more important than performance.

The **Dark Silicon** future makes custom and reconfigurable hardware more important than ever.

### 7.2.11 Other Models of Computation: Channel Communication

Using shared variables to communicate between threads is a low-level model of computation that:

- requires that the user abide by self-imposed protocol conventions and hence can be hazard prone,
- may not be apparent to the toolchain and hence optimisations may be missed,
- requires cache consistency and sequential consistency,
- has become (unfortunately) the primary parallel communications paradigm in today's chip multiprocessors (CMPs),
- is generally better avoided (so say many at least)!

CSP and Kahn-like process networks are an important model of computation based on channels. In computation theory terms, they might be viewed as a set of Turing machines connected via one-way tapes. CSPKahn Process Networkspn tool for Process Networks.

Disadvantage: deadlock is possible.

Some languages, such as Handel-C, Occam, Erlang and the best Bluespec coding styles completely ban shared variables and enforce use of CSP-like channels (LINK: Handel-C.pdf)

Handel-C uses explicit Kahn/Occam/CSP-like channels ('!' to write, '?' to read):

```

// Generator (src)    // Processor          // Consumer (sink)
while (1)            while(1)              while(1)
{
  ch1 ! (x);          {
    ch2 ! (ch1? + 2)  {
      $display(ch2?);
    }
  }
}

```

#### Using channels makes concurrency explicit and allows synthesis to re-time the design.

In CSP and Kahn-like networks, all communication is via blocking read and lossless write to/from unbound FIFOs. A variation on the basic paradigm is whether or not a reader can peek into the FIFO and make random access removal. Another variation: a chordal dequeue may typically be supported, where an atomic (pattern-matching) read of multiple input channels is made at once, as in the join calculus. Atomic write multiple is also sensible to support at the same time.

A basic system based on this paradigm requires the user code in this manner and renders RTL circuits correspondingly. Each channel has some physical FIFO or one-place buffer manifestation with the handshake wires being automatically synthesised. But advanced compilers can:

- Automatically convert standard code to this form (every shared variable or array becomes a process with explicit read and write messages)
- Automatically determine the FIFO depth needed at each point (or remove the FIFO entirely if deadlock will provably not arise),
- Automatically conglomerate and collapse such forms during code generation, with handshaking wires internal to a module or that are always ready disappearing during synthesis.

A **systolic array** is similar to a CSP/Kahn network, but the processing elements operate in lock-step instead of having FIFO queues between the nodes. A number of HLS compilers target systolic arrays instead of (or as well as) the classical sequencer approach.

*Exercise for the sheet:* It is argued that the systolic array approach is superior to a Kahn network when there will be no deviations from a static schedule. Consider a matrix multiplication or CNN application: is the FIFO between nodes needed for CNN accelerators?

### 7.2.12 Other Expression forms: Hardware Construction Languages

The **generate** statements in Verilog (Section 4.2.1) and VHDL are clunky imperative affairs. How much nicer it is to print out your circuit using higher-order functional programs! That's the approach of Chisel, a DSL embedded in Scala. Lava was the first HCL of this nature: 'Lava: Hardware Design in Haskell (1998)' by Per Bjesse, Koen Claessen, Mary Sheeran.

## A Varadic Priority Arbiter in Chisel

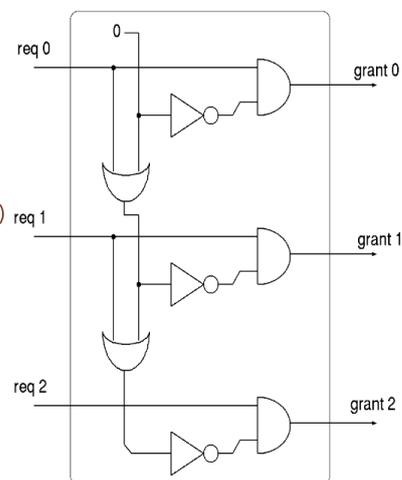
```
class genPriEncoder(n_inputs : Int) extends Module
{
  val io = new Bundle {
    val terms = (0 until n_inputs).map
      (n => ("req" + n, "grant" + n))

    terms.foldLeft (Bool(false))
    { case (sofar, (in, out)) =>
      val (req, grant) = (Bool(INPUT), Bool(OUTPUT))
      io.elements += ((in, req))
      io.elements += ((out, grant))
      grant := req & !sofar
      val next = new Bool
      next := sofar | req
      next
    }
  }
}
```

H/W components extend Module.  
They do their I/O via a Bundle.

All the standard operators & | ! are overloaded for h/w generation.

David J Greaves – Computer Lab Cambridge



Memocode 2015, Austin Texas.

Figure 7.26: An Example Chisel Module.

### 7.2.13 Other Expression forms: Logic Synthesis from Guarded Atomic Actions (Bluespec)

Using guarded atomic actions is an old and well-loved design paradigm. Recently Bluespec System Verilog has successfully raised the level of abstraction in RTL design using this paradigm.

- Every leaf operation has a guard predicate: says when it CAN be run.
- A Bluespec design is expressed as a list of rules where each rule is a guarded atomic action (hence declarative),
- Operations are embodied in the rules for atomic execution where the rule takes on the conjunction of its atomic operation guards and the rule may have its own additional guard predicate.
- Shared variables are ideally entirely replaced with one-place FIFO buffers with automatic handshaking,
- All communication to and from registers, FIFOs and user modules is via transactional/blocking ‘method calls’ for which argument and handshake wires are synthesised according to a global ready/enable protocol,
- Rules are allocated a static schedule at compile time and some that can never fire are reported,
- There is a strict mapping and packings of rules so that none is spread over a clock cycle (time/space folding) implemented by the compiler from Bluespec Inc but this, in principle, could be relaxed in other/future compilation strategies.
- Rules have the expectation they WILL be run (fairness).
- The wiring pattern of the whole design is generated from an embedded functional language (rather than embedding the language as a DSL in the way of Chisel and Lava).

The term ‘wiring’ above is used in the sense of TLM models: binding initiators to target methods.

The intention was that a compiler can direct scheduling decisions to span various power/performance implementations for a given program. But designs with an over-reliance on shared variables suffer RaW/WaR hazards when the schedule is altered. [LINK: Small Examples Toy BSV Compiler \(DJG\)](#)

First basic example: two rules: one increments, the other exits the simulation. This example looks very much like RTL: provides an easy entry for hardware engineers.

```

module mkTb1 (Empty);

  Reg#(int) x <- mkReg (23);

  rule countup (x < 30);
    int y = x + 1;           // This is short for int y = x.read() + 1;
    x <= x + 1;             // This is short for x.write(x.read() + 1);
    $display ("x = %0d, y = %0d", x, y);
  endrule

  rule done (x >= 30);
    $finish (0);
  endrule

endmodule: mkTb1

```

The problem with this example is that nice atomic rules are acting on a nasty mutable shared variable (the register). In general, RAMs and registers cannot be shared by freely-schedulable rules owing to RaW hazards and the like. It is much nicer if rules communicate with FIFOs, like the CSP process networks.

Our second example shows a FIFO-like pipe that is acted on by two rules. This is immune from scheduling artefacts/hazards. The example interface is for a pipeline object that could have arbitrary delay. The sending process is blocked by implied handshaking wires (hence far less typing than Verilog) and in the future would allow the programmer or the compiler to re-time the implementation of the pipe component.

```

module mkTb2 (Empty);

  Reg#(int) x    <- mkReg ('h10);
  Pipe_ifc pipe <- mkPipe;

  rule fill;
    pipe.send(x);
    x <= x + 'h10; // This is short for x.write(x.read() + 'h10);
  endrule

  rule drain;
    let y = pipe.receive();
    $display ("    y = %0h", y);
    if (y > 'h80) $finish(0);
  endrule
endmodule

```

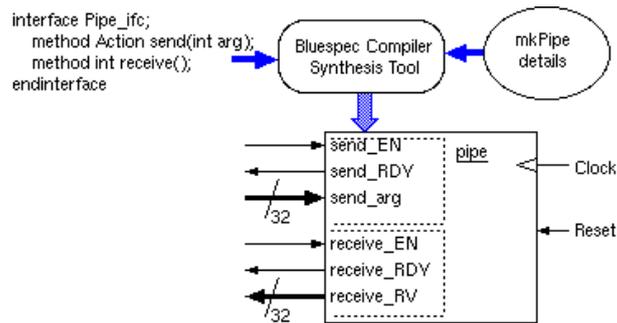


Figure 7.27: Synthesis of the ‘pipe’ Bluespec component with handshake nets.

Bluespec RTL was intended to be declarative, both in the elaboration language and with the guarded atomic actions for actual register transfers. Its advanced generative elaborator is a functional language and a joy to use for advanced/functional programmers. So it is/was much nicer to use than pure RTL. It has a scheduler (cf DBMS query planner) and a behavioural-sub language for when imperative is best. Like Chisel, it has good support for **valid-tagged data** in registers and busses. Hence compiler optimisations that ignore dead data are potentially possible.

As said, the main shortcoming of Bluespec is/was that the nice guarded atomic actions normally operate on imperative objects such as registers and RAMs where WaW/RaW/WaR bites as soon as transaction order is not carefully controlled. Also, imperative expression using a conceptual thread is also much loved by programmers, so Bluespec has a behavioural sub-language compiler built in that generates state machines.

#### 7.2.14 Classical Imperative/Behavioural H/L Synthesis Summary

Logic synthesisers and HLS tools cannot synthesise into hardware the full set of constructs of a general programming language. There are inevitable problems with:

- unbounded recursive functions,
- unbounded heap use
- other sources of unbounded numbers of state variables,
- many library functions: access to file or screen I/O.

And it is not currently sensible to compile seldom-used code to the FPGA since conventional CPUs serve well.

A Survey and Evaluation of FPGA High-Level Synthesis Tools, Nane et al, IEEE T-CAD December 2015]

Generating good hardware requires global optimisation of the major resources (ALUs, Multipliers and Memory Ports) and hence automatic time/space folding. New techniques are needed that note that wiring is a dominant power consumer in today's ASICs

The major EDA companies, Synopsys, Cadence and Mentor all actively marketing HLS flows. Altera (Intel) and Xilinx, the FPGA vendors, are now also promoting HLS tools.

Many people remain highly skeptical, but with FPGA in the cloud as a service in 2017 onwards, a whole new user community is garnered.

Synthesis from formal spec and so on: This is currently academic interest only? Except for glue logic? Success of formal verification means abundance of formal specs for protocols and interfaces: automatic glue synthesis seems highly-feasible.

### 7.2.15 High-Level Synthesis Survey

HLS removes instruction-fetch and decode entirely. Custom register lengths also commonly used.

A Survey and Evaluation of FPGA High-Level Synthesis Tools, Nane et al, IEEE T-CAD December 2015]

Kiwi (Greaves/Singh) Scientific Accelerator: CSharp programs are implemented on FPGA for high-performance with low energy.

```
public static class test4
{
    static int limit = 10;
    [Kiwi.OutputBitPort("dout")] static bool dout;
    [Kiwi.InputBitPort("din")] static bool din;

    static public void arraypart()
    {
        int odd = 0, even = 0; // These are V_0 and V_1 in the ast.cil file.
        int[] arr = new int [] {0, 1, 222221, 5, 7, 8, 1121, 2021, 2048}; // arr=V_6

        arr[2] = 2;
        // V_3 is the current element of the array
        // V_4 is a copy of V_2
        int pr = 0;
        foreach (int i in arr) // i=V_6
        {
            if (i%2 == 0)
                even++;
            else
                odd++;
            Kiwi.Pause();
            Console.WriteLine("{1} i={0}: ", i, pr++);
            Console.WriteLine("so far {0} Odd Numbers, and {1} Even Numbers.", odd, even);
        }
        Console.WriteLine("Found {0} Odd Numbers, and {1} Even Numbers.", odd, even);
        Kiwi.Pause();
    }
}
```

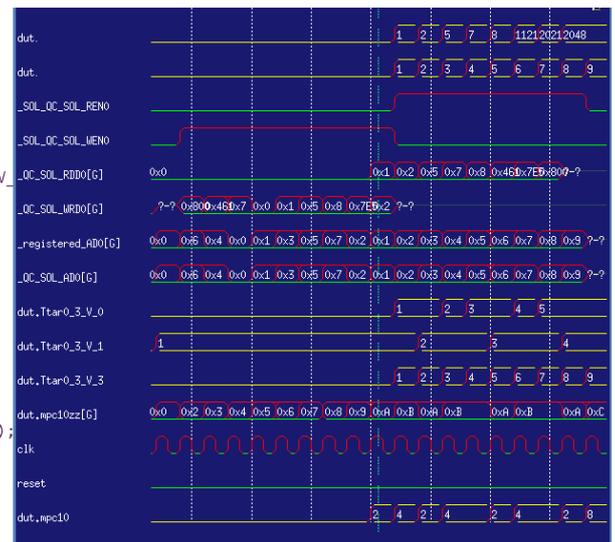


Figure 7.28: Input and Hardware Waveforms from a tiny Kiwi HLS example.

Mentor Catapult

Microsoft Catapult

Synopsys SymphonyC

Cadence -

Altera/Intel - OpenCL SDK and A++ and Spectra-Q

LegUP

Xilinx Vivado HLS

Reconfigure I/O - HLS from Go

A large number of failed startups ...

### 7.2.16 Maxeler SPL Open Stream/Spatial Processing (OpenSPL)

Key observation (from DJ Greaves): **Much high-performance computing on FPGA is limited by I/O bandwidth, so let's make sure we use 100 percent of that and factor everything else around it.**

No program counter - Stream Processing processes the stream in stream order.

Data flows as fast as possible to/from DRAM bank or fileserver - good. Programs need to be linearised to wrap around the stream.

- Transmit (T) the entire input to the program,
- Compute (C) sophisticated functions on large pieces of input at the rate it is presented,
- and Store (S), capture temporarily or archive all of it long term generator and reducer converge from and to scalars (or small vectors) respectively.

[github.com/maxeler/maxpower](https://github.com/maxeler/maxpower)

Trivial example : blurrer: compute the local 3-average:

```
HWType flt = hwFloat(8,24);
HWVar x = io.input("x", flt );
HWVar x_prev = stream.offset(x, -1);
HWVar x_next = stream.offset(x, +1);
HWVar cnt = control.count.simpleCounter(32, N);
HWVar sel_nl = cnt > 0;
HWVar sel_nh = cnt < (N - 1);
HWVar sel_m = sel_nl & sel_nh;
HWVar prev = sel_nl ? x_prev : 0;
HWVar next = sel_nh ? x_next : 0;
HWVar divisor = sel_m ? 3.0 : 2.0;
HWVar y = (prev+x+next)/divisor;
io.output("y", y, flt );
```

### 7.2.17 VSFG Value State Flow Graph - Zaidi

Imperative code hides too much parallelism. Static analysis is overly conservative on RaW and name alias avoidance by a factor of 10 or more.

Dark Silicon means having a lot of largely passive logic 'is a good idea'.

Current aim: Use a **dataflow** internal representation in compiler tools.

Final aim: Lets design a reconfigurable array that can directly execute dataflow programs.

```
Example
\begin{quote}
    for ( i = 0 ; i < 100; i ++ )
    {
        if (A[ i ] > 0 ) foo();
    }
    bar();
\end{quote}
```

Figure 1. Example C Code.

Exposing ILP in Custom Hardware with a Dataflow Compiler IR, Zaidi+Greaves|

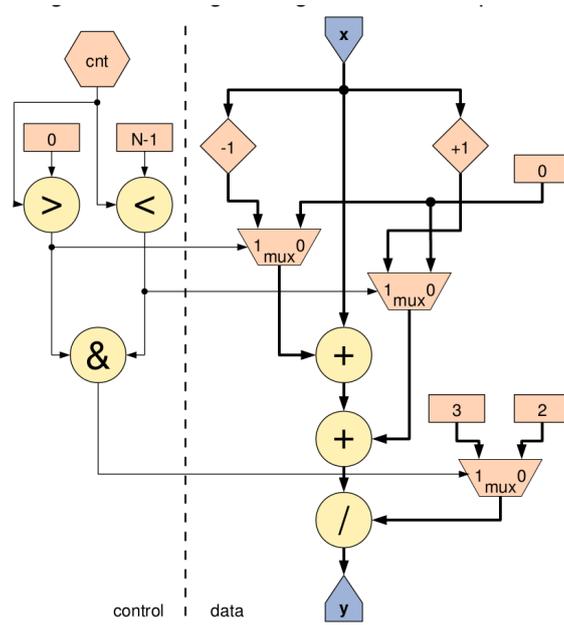


Figure 7.29: Maxeler Stream Flow - Running Average Fragment.

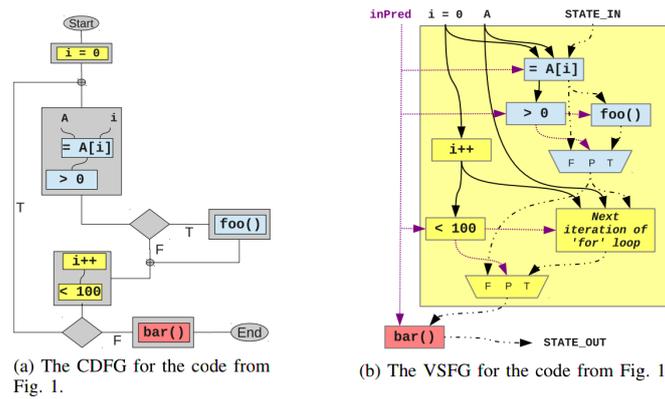


Figure 7.30: Comparing Control Flow and VSFG

## KG 8 — SystemC: Hardware Modelling Library

Topics: SystemC - (net-level SystemC not lectured to part IB 2015/16 onwards).

### 8.1 SystemC: Hardware Modelling Library for C++

SystemC is a free library for C++ for hardware SoC modelling. Download from [www.accelera.org](http://www.accelera.org) SystemC was developed over the last ten years. There have been two major releases, 1.0 and 2.0. Also of importance is the TLM sub-library, TLM 2.0. (SystemC using transactional-level modelling (TLM/ESL) is covered later).

Greaves developed the TLM\_POWER3 add-on library for power modelling.

It can be used for detailed net-level modelling, but today its main uses are :

- Architectural exploration: Making a fast and quick, high-level model of a SoC to explore performance variation against various dimensions, such as bus width and cache memory size.
- Transactional level (TLM) models of systems, where handshaking protocols between components using hardware nets are replaced with subroutine calls between higher-level models of those components.
- Synthesis: RTL is synthesised from from SystemC source code using High-Level Synthesis (HLS) (commonly also called 'C-to-gates' compiler.) SystemC Synthesis

SystemC includes (at least):

- A module system with inter-module channels: C++ class instances are instantiated in a hierarchy, following the circuit component structure, in the same way that RTL modules instantiate each other.
- An eventing and threading kernel that is non-preemptive and which allows user code inside components to run either in a trampoline style, returning the thread without blocking, or to keep the thread and hold state on a stack.
- Compute/commit signals as well as other forms of channel for connecting components together. The compute/commit signals are needed in a zero-delay model of hardware to avoid 'shoot-thru': i.e. the scenario where one flip-flop in a clock domain changes its output before another has processed the previous value.
- A library of fixed-precision integers. Hardware typically uses all sorts of different width busses and counters that wrap accordingly. SystemC provides classes of signed and unsigned variables of any width that behave in the same way. For instance the user can define an `sc_int` of five bits and put it inside a signal. The provided library includes overloads of all the standard arithmetic and logic operators to operate on these types.
- Plotting output functions that enable waveforms to be captured to a file and viewed with a program such as `gtkwave`.
- A transactional modelling sub-library: TLM 1.0 provided separate blocking and non-blocking interfaces prototypes that a user could follow and in TLM 2.0 these are rolled together into 'convenience sockets' that can convert between the two forms.

Problem: hardware engineers are not C++ experts but they can be faced with complex or advanced C++ error messages when they misuse the library.

Benefit: General-purpose behavioural C code, including application code and device drivers, can all be modelled in a common language.

```

SC_MODULE(mycounter) // An example of a leaf module (no subcomponents).
{
    sc_in < bool      > clk, reset;
    sc_out < sc_int<10> > myout;

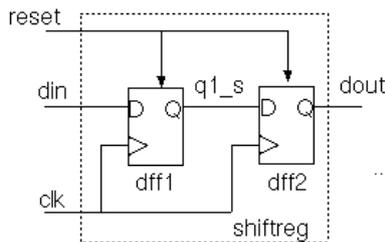
    void m() // Internal behaviour, invoked as an SC_METHOD.
    {
        myout = (reset) ? 0: (myout.read()+1); // Use .read() since sc_out makes a signal.
    }

    SC_CTOR(mycounter) // Constructor
    { SC_METHOD(m); //
      sensitive << clk.pos();
    }
}
// Complete example is on course web site and also on PWF.

```

SystemC enables a user class to be defined using the the `SC_MODULE` macro. Modules inherit various attributes appropriate for an hierarchic hardware design including an instance name, a type name and channel binding capability. The `sensitive` construct registers a callback with the EDS kernel that says when the code inside the module should be run. An unattractive feature of SystemC is the need to use the `.read()` method when reading a signal.

### 8.1.1 SystemC Structural Netlist



```

//Example of structural hierarchy and wiring between levels:
SC_MODULE(shiftreg) // Two-bit shift register
{
    sc_in < bool > clk, reset, din;
    sc_out < bool > dout;

    sc_signal < bool > q1_s;
    dff dff1, dff2; // Instantiate FFs

    SC_CTOR(shiftreg) : dff1("dff1"), dff2("dff2")
    {
        dff1.clk(clk);
        dff1.reset(reset);
        dff1.d(din);
        dff1.q(q1_s);

        dff2.clk(clk);
        dff2.reset(reset);
        dff2.d(q1_s);
        dff2.q(dout);
    }
};

```

A SystemC templated **channel** provides general purpose interface between components. We rarely use the raw channels: instead we use the derived forms `sc_in`, `sc_out` and `sc_signal`. These channels implement compute/commit paradigm required for **delta cycles**. This avoids non-determinacy from races in zero-delay models (see earlier).

Other provided channels include the buffer, fifo, mutex, semaphore and clock (non-examinable). Users can overload the channel class to implement channels with their own semantics if needed. A user-defined channel type can even contain other SystemC components but the importance of this is reduced when using the TLM libraries.

```

// A signal is an abstract (templated) data type that has a current and next value.
// Signal reads are of the current value. Writes are to the next value.
// This example illustrates:
int nv; // nv is a simple c variable (POD)
sc_out < int > data; // data and mysig are signals (non-POD)
sc_signal < int > mysig; //
...
    nv += 1;
    data = nv;
    mysig = nv;
    printf("Before nv=%i, %i %i\n", nv, data.read(), mysig.read());
    wait(10, SC_NS);
    printf("After nv=%i, %i %i\n", nv, data.read(), mysig.read());
...
Before nv=96, 95 95
After nv=96, 96 96

```

When the scheduler blocks with no more events in the current time step, the pending new values are committed to the visible current values.

For faster system modelling, we do not want to enter EDS kernel for every change of every net or bus: so is it possible to pass larger objects around, or even send threads between components, like S/W does ?

Yes, it is possible to put any datatype inside a signal and route that signal between components (provided the datatype can be checked for equality to see if current and next are different and so on). Using this approach, a higher-level model is possible, because a complete Ethernet frame or other large item can be delivered as a single event, rather than having to step through the cycle-by-cycle operation of a serial hardware implementation.

Even better: SystemC 2.0 enabled threads to be passed along the channels, allowing intermodule thread calling, just like object-oriented software. This will enable TLM modelling (described later). Hence we have three intermodule communication styles:

1. **Pin-level modelling:** an event is a change of a net or bus,
2. **Abstract data modelling:** an event is delivery of a complete cache line or other data packet,
3. **Transactional-level modelling:** avoid events as much as possible: use intermodule software calling.

### 8.1.2 SystemC Abstracted Data Modelling

Here we raise the modelling abstraction level by passing an abstract datatype along a channel. the abstract data type must define a few basic methods, such as the equality operator overload this is shown:

```

sc_signal < bool > mywire; // Rather than a channel conveying just one bit,

struct capsule
{ int ts_int1, ts_int2;
  bool operator==(struct ts other)
  { return (ts_int1 == other.ts_int1) && (ts_int2 == other.ts_int2); }

  int next_ts_int1, next_ts_int2; // Pending updates
  void update()
  { ts_int1 = next_ts_int1; ts_int2 = next_ts_int2;
  }

  ...
  ... // Also must define read(), write() and value_changed()

};

sc_signal < struct capsule > myast; // We can send two integers at once.

```

For many basic types, such as `bool`, `int`, `sc_int`, the required methods are provided in the SystemC library, but clearly not for user-defined types.

```

void mymethod() { ... }
SC_METHOD(mymethod)
sensitive << myast.pos(); // User must define concept of posedge for his own abstract type.

```

### 8.1.3 Threads and Methods

SystemC enables a user module to keep a thread and a stack but prefers, for efficiency reasons if user code runs on its own upcalls in a trampoline style.

- An SC\_THREAD has a stack and is allowed to block.
- An SC\_METHOD is just an upcall from the event kernel and must not block.

Comparing SC\_THREADS with trampoline-style methods we can see the basis for two main programming TLM styles to be introduced later: blocking and non-blocking.

The user code in an SC\_MODULE is run either as an SC\_THREAD or an SC\_METHOD.

An SC\_THREAD has a stack and is allowed to block. An SC\_METHOD is just an upcall from the event kernel and must not block. Use SC\_METHOD wherever possible, for efficiency. Use SC\_THREAD where important state must be retained in the program counter from one activation to the next or when asynchronous active behaviour is needed.

The earlier ‘mycounter’ example used an SC\_METHOD. Now an example using an SC\_THREAD: a data source that provides numbers using a net-level four-phase handshake:

```

SC_MODULE(mydata_generator)
{
  sc_out < int > data;
  sc_out < bool > req;
  sc_in < bool > ack;

  void myloop()
  {
    while(1)
    {
      data = data.read() + 1;
      wait(10, SC_NS);
      req = 1;
      do { wait(10, SC_NS); } while(!ack.read());
      req = 0;
      do { wait(10, SC_NS); } while(ack.read());
    }
  }

  SC_CTOR(mydata_generator)
  {
    SC_THREAD(myloop);
  }
}

```

A SystemC thread can block for a given amount of time using the `wait` function in the SystemC library (not the Posix namesake). NB: If you put ‘`wait(4)`’ for example, you will invoke the unix system call of that name, so make sure you supply a SystemC time unit as the second argument.

**Additional notes:**

Waiting for an arbitrary boolean expression to hold is hard to implement on top of C++ owing to its compiled nature:

- C++ does not have a reflection API that enables a user's expression to be re-evaluated by the event kernel.
- Yet we still want a reasonably neat and efficient way of passing an uninterpreted function.
- Original solution: the delayed evaluation class:

```
waituntil(mycount.delayed() > 5 && !reset.delayed());
```

Poor user had to just insert the **delayed** keyword where needed and then ignore it when reading the code. It was too unwieldy, now removed. So today (pre C++11) use the less-efficient:

```
do { wait(10, SC_NS); } while(!(mycount > 5 && !reset));
```

Within SystemC, there is no direct equivalent to the continuous assignment of Verilog. But the 'fully-supported' sensitivity list always @(\*) or always\_comb can be reproduced with an SC\_METHOD if you manually list the supporting nets. More-efficient performance gained by putting the continuous assignment behaviour in a method and remembering to call just after whenever the support is changed in other parts of the model. Fortunately, for TLM designs, there will be very little continuous assignment needed (e.g. perhaps just for interrupt routing).

### 8.1.4 SystemC Plotting and GUI

We can plot to industry standard VCD files and view with gtkwave (or modelsim).

```
sc_trace_file *tf = sc_create_vcd_trace_file("tracefile");

// Now call:
// sc_trace(tf, <traced variable>, <string>);

sc_signal < int > foo;
float bar;
sc_trace(tf, foo);
sc_trace(tf, bar, "bar"); // Give name if anon constructor

sc_start(1000, SC_NS); // Simulate for one microsecond
sc_close_vcd_trace_file(tr);
return 0;
```

VCD can be viewed with **gtkwave** or in **modelsim**. There are various other commercial interactive viewer tools...

Try-it-yourself on PWF

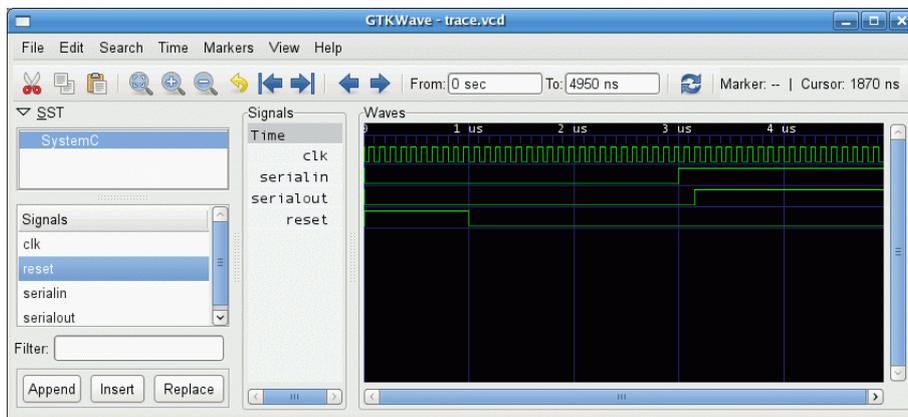


Figure 8.1: Waveform view plotted by gtkwave.

## KG 9 — SoC Engineering and Associated Tools

In this section we look at engineering aspects and associated tools used in SoC design and modelling. A lot of the effort is dedicated to maximising performance and minimising power dissipation.

### 9.0.5 RAM Macrocell Compiler Tool

The average SoC is 71 percent RAM memory. The RAMs are typically generated by a RAM compiler. The input parameters are:

- Size: Word Length and Number of Words.
- Port description: Each port has an address input and is one of r, w, r/w.
- Clocking info: Frequency, latency, or access time for asynchronous RAM.
- Resolution: What to do on write/write and write/read conflicts.

The outputs are a datasheet for the RAM, high and low detail simulation models and something that turns into actual polygons in the fabrication masks.

```
// Low-level model (RTL) for a RAM. Example 1.
module R1W1RAM(din, waddr, clk, wen, raddr, dout);
  input clk, wen;
  input [14:0] waddr, raddr;
  input [31:0] din;
  output [31:0] dout;

  // Mem array itself: 32K words of 32 bits each.
  reg [31:0] myram [32767:0];
  always @(posedge clk) begin
    dout <= myram[raddr];
    if (wen) myram[waddr] <= din;
  end
end
```

```
// Low-level model (RTL) for a RAM. Example 2.
module R1W1RAM(din, addr, clk, wen, dout);
  input clk, wen;
  input [14:0] addr;
  input [31:0] din;
  output [31:0] dout;

  // Address register: latency of 1 one cycle.
  reg [14:0] addr1;
  // Mem array itself: 32K words of 32 bits each.
  reg [31:0] myram [32767:0];
  always @(posedge clk) begin
    addr1 <= addr;
    if (wen) myram[addr1] <= din;
    else dout <= myram[addr1];
  end
end
```

```
// Example high-level model for both RAMs
SC_MODULE(R1W1RAM)
{
  uint32_t myram [32768];
  int read_me(int A) { return myram[A]; }
  write_me(int A, int D) { myram[A] = D; }
  tlm_utils::simple_target_socket<R1W1RAM> port0;
  ...

  // This RAM model has a pair of entry points
  // for reading and writing.
  // It also has a TLM convenience socket
  // which would decode a generic payload and
  // call one or other of those entry points
  // for each transaction.
}
```

Sometimes self test modules are also generated. For example Mentor's MBIST Architect(TM) generates an SRTL BIST with the memory and ARM/Artisan's Generator will generate a wrapper that implements self repair of the RAM by diverting access from a fault row to a spare row. ARM Artisan

Other related generator tools would be similar in use: e.g. a FIFO generator would be similar and a masked ROM generator or PLA generator.

### 9.0.6 Dynamic Clock Gate Insertion Tool

Clock trees consume quite a lot of the power in an ASIC and considerable savings can be made by turning off the clocks to small regions. A region of logic is idle if all of the flip-flops are being loaded with their current contents, either through synchronous clock enables or just through the nature of the design (see later slides).

Instead of using synchronous clock enables, current design practice is to use a clock gating insertion tool that gates the clock instead.

Care must be taken not to generate glitches on the clock as it is gated and transparent latches in the clock enable signal can re-time it to prevent this.

How to generate clock enable conditions ? One can have software control (additional control register flags) or automatically detect. Automatic tools compute 'clock needed' conditions. A clock is 'needed' if any register will change on a clock edge. A lot of clock needed computation can get expensive, resulting in no net saving, but it can be effective if computed once at head of a pipeline.

*Beyond just turning off the clock or power to certain regions, in another LG we look at further power saving techniques: dynamic frequency and voltage scaling.*

### 9.0.7 Test Program Generator Tool

*Lectured if time permits:* A test program generator works out a short sequence of tests that will reveal 'stuck-at' and other faults in a subsystem.

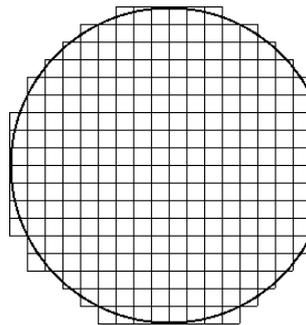


Figure 9.1: The wafer (six to ten inches diameter) is diced into chips (1cm on a side or so).

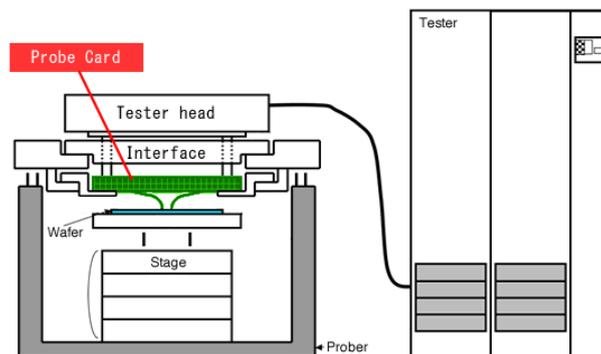


Figure 9.2: General configuration of a wafer probe testing machine (not showing the wafer conveyers that load a new undiced wafer every few minutes).

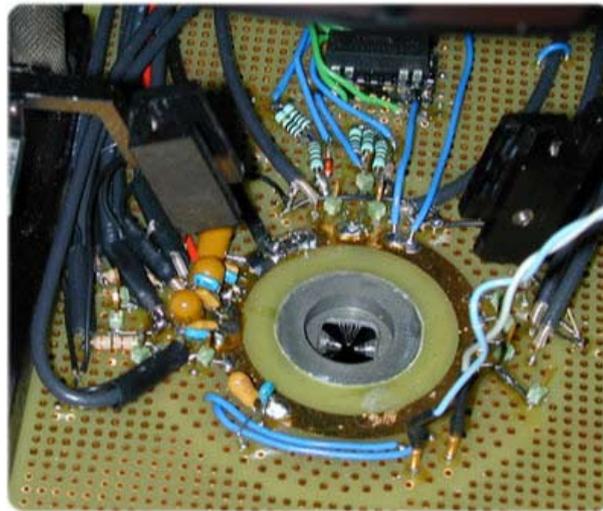


Figure 9.3: Probe needles just visible inside testing load card.

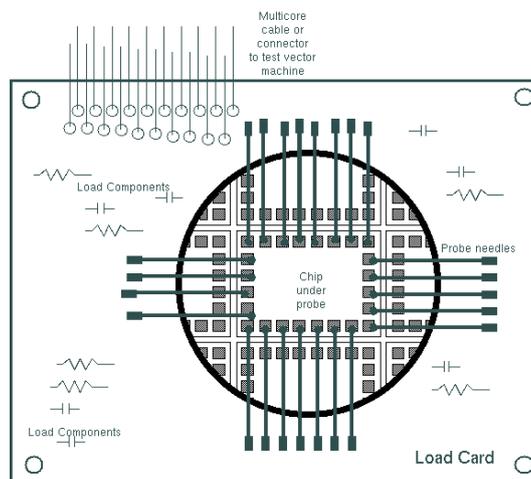


Figure 9.4: Drawing of the test load card.

### 9.0.8 Scan Path Insertion and JTAG standard test port.

*Lectured if time permits:* A scan path insertion tool replaces the user's D-type flip-flops with a scan path, connected to the external JTAG test access port for post-fabrication testing.

## KG 10 — Custom Accelerator Structures

Perhaps the first hardware accelerators added to alongside the integer execution units of early computers were for floating-point arithmetic. But accelerators can serve many different purposes and sit elsewhere within the architecture. In this section we note the design considerations.

### 10.1 H/W to S/W Interfacing Techniques

The term ‘Programmed I/O’ can refer to either MMIO or PMIO. These are the main alternative to I/O performed by DMA. A note on terms:

- Port-mapped I/O (PMIO) refers to a special address space outside of normal memory that is accessed with instructions such as IN and OUT.
- Memory-mapped I/O (MMIO) refers to I/O devices being allocated addresses inside the normal Von Neumann address space that is primarily used for program and data. Such I/O is done using instructions such as LOAD and STORE.

PMIO was very useful on A16 microprocessors since valuable address space was not consumer by the I/O devices, but A32 architectures generally provide no PMIO instructions and hence use MMIO.

An accelerated system is divided into some number of hardware and software blocks with appropriate means of interconnection. The primary ways of connecting hardware to software are:

- CPU coprocessor and/or custom instructions,
- Packet channel connected as coprocessor or mapped to main register file,
- Programmed I/O to pin-level GPIO register,
- Programmed I/O to FIFOs,
- Interrupts (hardwired to one core or dynamically dispatched),
- Pseudo-DMA: processor generates memory addresses or network traffic and the accelerator simply snoops or interposes on the data stream,
- DMA - Autonomous unit, much like a CPU in its own right.

Another design point is to do everything in hardware with no CPUs, but a CPU in a supervisory role is normally sensible.

### 10.2 Custom Accelerators on SoC or ASIC

Suppose something like the following fragment of code is a dominant consumer of power in a portable embedded mobile device:

```
for (int xx=0; xx<1024; xx++)
{
    unsigned int d = Data[xx];
    int count = 0;
    while (d > 0) { if (d&1) count ++; d >>= 1; }
    if (!xx || count > maxcount) { maxcount = count; where = xx; }
}
```

This kernel tallies the set bit count in each word: such bit-level operations are inefficient using general-purpose CPU instruction sets. This is about the smallest possible example for which a hardware accelerator might be worthwhile. Do you think it could be worthwhile?

A dedicated **hardware accelerator** avoids instruction fetch overhead and is generally more power efficient. Analysis using Amdahl's law and high-level simulation (SystemC TLM) can establish whether a hardware implementation is worthwhile. There are several feasible partitions:

1. Extend the CPU with a **custom datapath** and custom ALU (Figure 10.1a) for the inner tally function controlled by a **custom instruction**.
2. Add a tightly-coupled **custom coprocessor** (Figure 10.1b) with fast data paths to load and store operands from/to the main CPU. The main CPU still generates the address values *xx* and fetches the data as usual.
3. Place the whole kernel in a **custom peripheral unit** (Figure 10.2) with operands being transferred in and out using programmed I/O or pseudo-DMA.
4. As 3, but with the new IP block having bus master capabilities so that it can fetch the data itself (DMA), with polled or interrupt-driven synchronisation with the main CPU.
5. Use an FPGA or bank of FPGAs without a conventional CPU at all ... (see later).

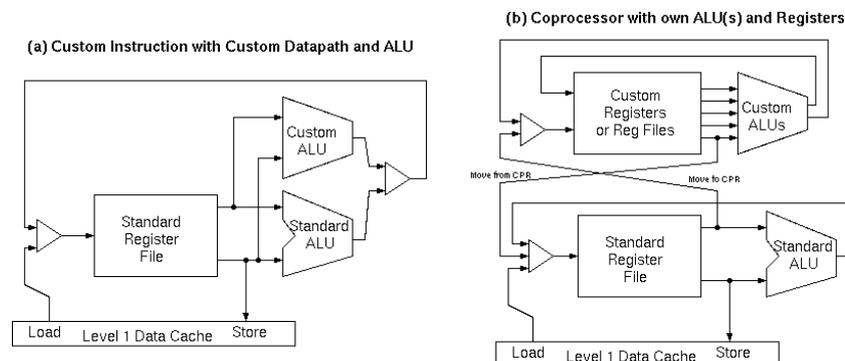


Figure 10.1: A custom ALU operation implemented in two similar ways: as a custom instruction or as a coprocessor.

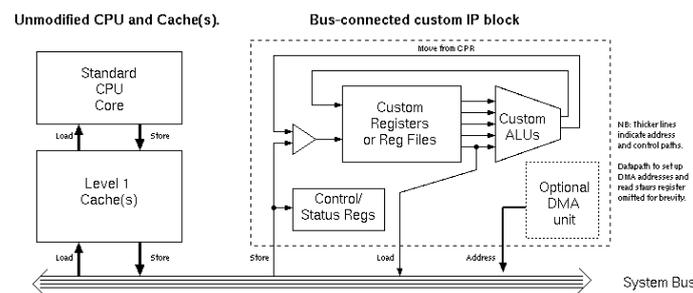


Figure 10.2: A custom function implemented as a peripheral IP block, with optional DMA (bus master) capability.

The special hardware in all approaches may be manually coded in RTL or compiled using HLS from the original C implementation.

In the first two approaches, both the tally and the conditional update of the maxcount variable might be implemented in the custom ALU, but most of the gain would come from the tally function itself and the detailed design might be different depending on whether custom instruction or coprocessor were used. The custom instruction operates on data held in the normal CPU register file. The bit tally function alone reads one input

word and yields one output word, so it easily fits within the addressing modes provided for normal ALU operations. Performing the update of both the maxcount and word registers in one custom instruction would require two register file writes and this may not be possible in one clock cycle and hence, if this part of the kernel is placed in the custom datapath we might lean more towards the co-processor approach.

Whether to use the separate, bus-connected IP block depends on whether the processor has something better to do in the meantime and that there is sufficient bus bandwidth for them both to operate. Whether the data is likely to be in or be needed in a given L1 data cache is also an important factor.

The separate IP block may or may not use a DMA controller. Given that the ARM now has a ones-tally instruction in its normal ALU, getting an ARM to move the data into the separate IP block may be a really poor design point.

With increasing available transistor count in the form of **dark silicon** (ie. switched off most of the time) in recent and future VLSI, implementing standard kernels as custom hardware cores is a possible future trend for power conservation. The **conservation cores** project Venkatesh considered implementing the inner loops of a 'mature computations' such as a selection of popular Android applications in silicon on future mobile phones.

### 10.3 Bump-in-Wire Reconfigurable Accelerator Architectures

FPGA is increasingly seen as a computing element alongside CPU and GPU. Energy savings of two orders of magnitude are often seen when a suitable application is accelerated on FPGA. Execution speed can also commonly increase, although this is hampered by the order-magnitude reduction in clock frequency compared with CPU (e.g 200 MHz instead of 2 GHz).

Historically, many hardware accelerator projects have ultimately been unsuccessful because: either

- The hardware development takes too long and general-purpose CPUs meanwhile progress and overtake (their development teams are vastly more resourced)
- The overhead of copying the data in and out of the accelerator exceeds the processing speed up.
- The hardware implementation is out of date, such as when the requirements or a protocol standard is changed.

But by implementing accelerators on FPGA at a place where the data is moving already, these problems can be largely mitigated. Also, until recently, FPGAs have not had hardened DRAM controllers and consequently been short of DRAM bandwidth.

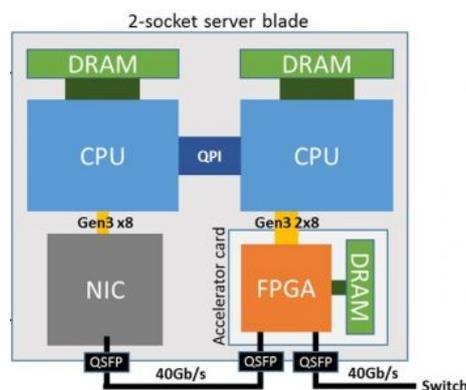


Figure 10.3: Bump-in-Wire design for Microsoft Catapult Accelerator (2016).

Microsoft have had several generations of blade design for their data centres. Recent ones have placed the FPGA in series with blade's network connection, thereby enabling copy-free pre- and post-processing of data. For instance, an index hash can be computed on database fields.

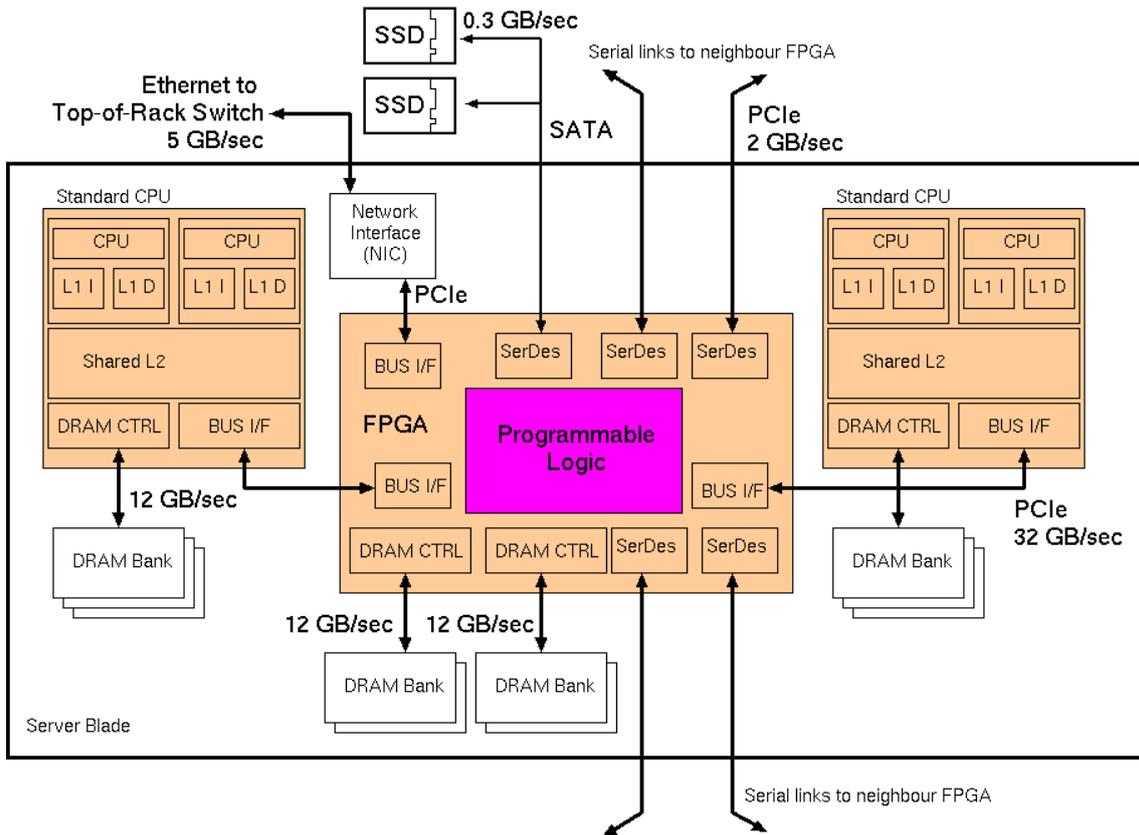


Figure 10.4: Representative ‘bump-in-wire’ server blade architecture that has the FPGA in series with Network Interface and Disk Drives.

The FPGAs on neighbouring cards are also locally interconnected with a high-speed ring or mesh network, enabling them to be pooled and managed independently of the blade’s CPUs. This enables systolic sorting networks and the like to be formed; e.g. for keeping the k-best Bing search results.

The QPI interconnection between CPUs is cache-consistent. Some FPGA-accelerated blade designs connect the FPGA to such a cache-consistent interconnect.

On the Zynq platform (Figure 10.7) a number of methods for connecting to the reconfigurable logic are available - they are mostly via AXI ports. They vary in cache-consistency and bandwidth and initiator/target polarity. Of the initiating ports, both provide connection to the on-chip SRAM and the single DRAM bank that is also shared with the ARM cores, But one form is cache-coherent with the ARMs and the other is not, but has higher bandwidth.



Figure 10.5: Catapult Blade - FPGA is at the lower right, its heatsink visible above its blue DRAM DIMMs.

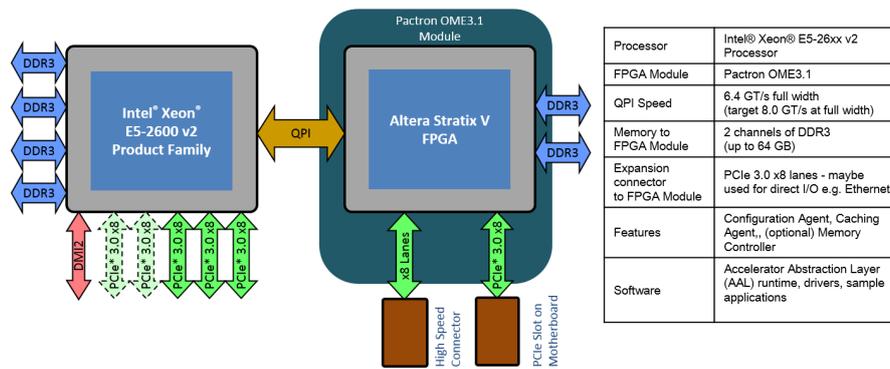


Figure 10.6: Cache-consistent interconnection between CPU and FPGA.

### 10.3.1 FPGA Computing to replace Von Neumann Dominance?

The Von Neumann computer hit a wall in terms of increasing clock frequency. It is widely accepted that Parallel Computing is the most energy-efficient way forward. The FPGA is intrinsically massively-parallel and can exploit the abundant transistor count of contemporary VLSI. Andre DeHon points out that the Von Neumann architecture no longer addresses the correct problem: he writes ‘Stored-program processors are about compactness, fitting the computation into the minimum area possible.’

Why is computing on an FPGA becoming a good idea? Spatio-Parallel processing uses less energy than equivalent temporal processing (ie at higher clock rates) for various reasons. David Greaves gives nine:

1. Pollack’s rule states that energy use in a Von Neumann CPU grows with square of its IPC. But the FPGA with a static schedule moves the out-of-order overheads to compile time.
2. To clock CMOS at a higher frequency needs a higher voltage, so energy use has quadratic growth with frequency.
3. Von Neumann SIMD extensions greatly amortise fetch and decode energy, but FPGA does better, supporting precise custom word widths, so no waste at all. Standard computers support fixed data sizes only and only two encodings (integer and floating point), both of which can waste a lot of energy compared with better encodings [Constantinidies].
4. FPGA can implement massively-fused accumulate rather than re-normalising after each summation.
5. Memory bandwidth: FPGA has always had superb on-chip memory bandwidth but latest generation FPGA exceeds CPU on DRAM bandwidth too.
6. FPGA using combinational logic uses zero energy re-computing sub-expressions whose support has not changed. And it has no overhead determining whether it has changed.
7. FPGA has zero conventional instruction fetch and decode energy and its controlling micro-sequencer or predication energy can be close to zero.
8. Data locality can easily be exploited on FPGA — operands are held closer to ALUs near-data-processing (but the FPGA overall size is x100 times larger (x100 area) owing to overhead of making it reconfigurable.  
So
9. The massively-parallel premise of the FPGA is the correct way forward, as indicated by asymptotic limit studies [DeHon].

Programming an FPGA has been a problem. As we shall discuss in a later section, end users cannot be expected to be hardware or RTL experts. Instead, new compiler techniques to port software-style programming to the FPGA are developing. The main approaches today are OpenCL and HLS.

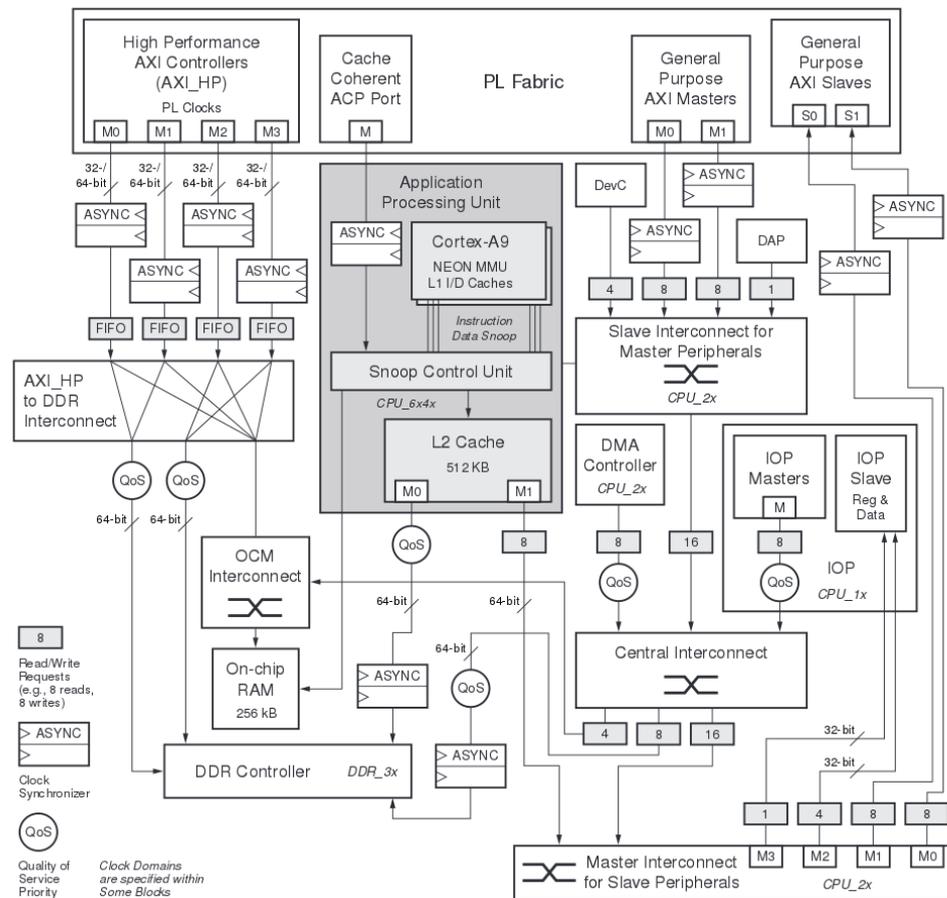


Figure 10.7: Block Diagram of the Xilinx Zynq Platform.

### 10.3.2 FPGA As The Main Or Only Processor ?

The FPGA generally needs connecting to mainstream computing resources, if only for management or file system access. Figure 10.9 shows a typical setup using a Xilinx Zynq chip (from the Kiwi HLS project) KiwiThe so-called ‘substrate’ provides basic start/stop and debugging control, together with access to a programmed I/O (PIO) register file. It also provides services for a network-on-chip between the processing blocks that provides an ‘FPGA Operating System’ API for file server input and output. Finally, it also provides access to DRAM banks and platform-specific inter-FPGA links (not shown).

Will FPGA totally replace Von Neumann? Currently many problems remain: the static-schedule does not work well for all applications, the debugging support is in its infancy and compile times are often hours. The length overhead of programmable wiring adds a lot of capacitance, so only if the data ends up moving far less physical distance will energy be saved. Nonetheless, there are many success stories in areas such as automated trading, database indexing, deep packet inspection, genomics and cryptology. There are many further papers and PhDs to be written ...

### 10.3.3 Virtualising the FPGA

FPGA in the datacentre was first deployed by Microsoft for Bing acceleration. A user-programmable FPGA fabric as part of the Cloud offering was made available by Amazon at the start of 2017. Where customer designs are loadable into the FPGA, security issues arise. Overheating, denial of service and data theft must be avoided and a means to virtualise (share) the FPGA resources is needed.

The FPGA has much more user state to context switch than a conventional CPU, so a sub-Hertz context switch

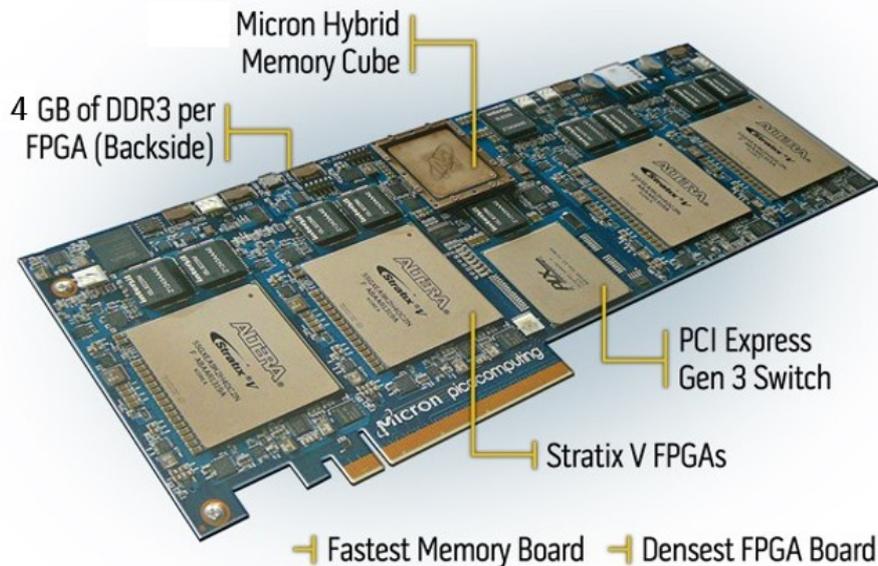


Figure 10.8: A CPU-less cloud server blade from Pico Computing (4 FPGAs + Lots of DRAM).

rate is likely. This is at least 3-orders coarser than typical CPU time sharing. All recent FPGA families support dynamic loading of selected parts of the die while the remainder continues operating. The reconfiguration time is typically 10 to 100 ms per zone. This perhaps opens the way to finer-grained sharing, if needed.

One approach to virtualisation is to use dynamically-loaded modules under the ‘Shell’ and ‘Role’ approach to FPGA real-estate (Figure 10.10) where the customer only has access to the part of the die via partial-reconfiguration. The service shim or shell provides a protected service layer. Disadvantages of this technique are that quite a lot of real estate (and especially the multiplier resource therein) is consumed with overhead. Also there are design fragmentation inefficiencies. And many of the FPGA resources, such as most of the I/O pins and hardened peripherals go unused (not a problem for a Dark Silicon future!).

### 10.3.4 Future FPGA Architectures for the Cloud

It is possible that a future generation of FPGA devices will be developed specialised for server blades in the cloud. See Figure 10.11. The user logic will connect to the outside world only through a service network. Kiwi HLS uses such a NoC for O/S access, as does LEAP. But reconfigurable interconnection between adjacent zones that jointly host an application (blue) might be provided. Having the service network hardened makes it about 100-fold more area efficient, allowing much more complexity, security and monitoring. The local memory devices would also provide virtualisation assistance in their hardened controllers.

A simple NoC does not take up much FPGA resources, but the ideal NoC would probably support credit-based flow control and also include support for virtualisation.

### 10.3.5 Overlays and Coarse-Grain Reconfigurable Array (CGRA)

FPGA virtualisation is also achievable using overlays. An overlay typically consists of an initialisation pattern for FPGA RAM that then behaves as ROM and which is served by a fixed design loaded into the rest of the configurable logic, regardless of the overlay selected. Security is achieved in the same way that conventional O/S security is achieved: no direct/unchecked access to anything sensitive.

FPGA advantages: Massively parallel. No fetch/execute overhead.



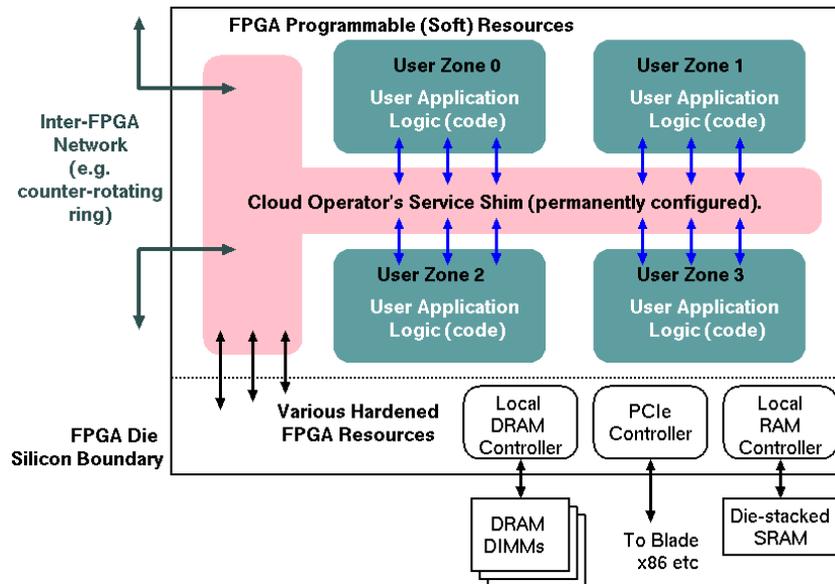


Figure 10.10: The reconfigurable logic resources of a conventional FPGA can be manually partitioned into user reconfigurable areas and a shell/shim for service access.

Michael Dales Phd Thesis addressed this first: Managing a Reconfigurable Processor in a General Purpose Workstation Environment - Date 2003

### 10.3.8 PALs and CPLDs

PALs are Programmable Array Logic and CPLDs (Complex Programmable Logic Devices) achieve very low delay in return for simple, nearly fixed, wiring structure. All expressions are expanded to SOP form with limited number of products. Expanding to sum-of-products form can cause near-exponential area growth (e.g. ripple carry converted to fast carry).

```

pin 16 = o1;
pin 2 = a;
pin 3 = b;
pin 4 = c

o1.oe = ~a;
o1 = (b&o1) | c;

-x- - - - - (oe term)
--x- x- - - - (pin 3 and 16)
- - - - - x- - - - (pin 4)
xxxx xxxx xxxx xxxx xxxx xxxx
x (macrocell fuse)
    
```

A PAL is programmable array logic device. Figure 10.13 shows a typical device. Such devices have been popular since about 1985. They are really just highly structured gate arrays. Every logic function must be multiplied out into sum-of-products form and hence is achieved in just two gate delays. The illustrated device has 8 product terms per logic function, and so can support functions of medium complexity. Such devices were very widely used in the 1980' because they could support clock rates of above 100 MHz. Today, FPGA speeds of 200 MHz are common and they also provide special function blocks, such as PCI-e interfaces, so the need for PALs has diminished.

Programmable *macrocells* (Figure 10.14) enable the output functions to be either registered or combinatorial. Small devices (e.g. with up to 10 macrocells) offer one clock input; larger devices with up to about 100 macro-

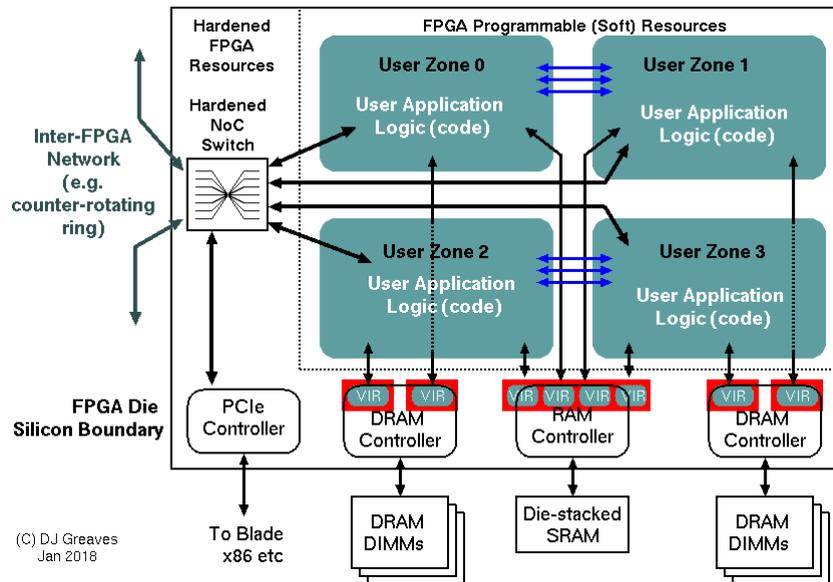


Figure 10.11: Putative block diagram for a cloud-oriented FPGA for scientific acceleration in the future.

cells are also available, and generally offer several clock options. Often some macrocells are not actually associated with a pin, providing a so called *buried state* flip-flop.

Mini design example: As entered by a designer in a typical PAL language, and part of the fuse map that would be generated by the PAL compiler. Each product line has seven groups of four fuses and produces the logical AND of all of the signals with intact fuses. An 'x' denotes an intact fuse and all of the fuses are left intact on an unused product lines in order to prevent the line ever generating a logical one (a gets ANDed with  $\bar{a}$ bar etc.). The fuse map is loaded into a programming machine (in a file format known as JEDEC), an unused PAL is placed in the machine's socket and the machine programs the fuses in the PAL accordingly.

PALs achieve their speed by being highly structured. Their applicability is restricted to small finite state machines and other *glue logic* applications.

### 10.3.9 HLS Expression forms: Behavioural using a Thread or Threads

Using the first of these, behavioural expression, we express the algorithm and steps to be performed as an executable program

- using an **imperative** program (containing loops and assignments), or
- a **functional** program (where control flow is less-explicit).

Either way, the tool chain may:

- **re-order** the operations while preserving semantics, and/or
- **re-encode** the state and modify memory layouts.

Examples:

- Hardware Construction Languages (Lava, Chisel 2.0),
- Synopsys Behavioural Compiler for RTL (now defunct),

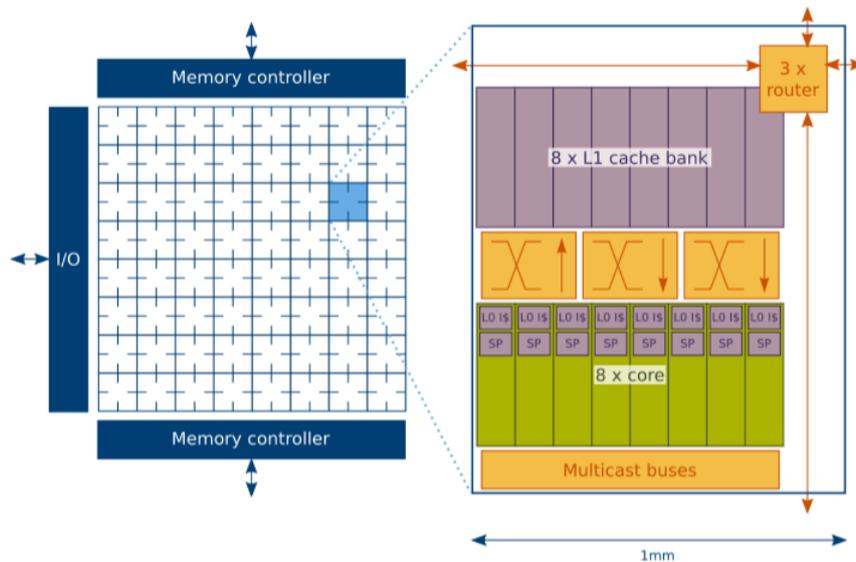


Figure 10.12: Loki tiled processors (Mullins/Cambridge).

- Classical HLS: C-to-Gates : C-To-Verilog, SystemCrafter, LegUp, Catapult,
- Bluespec's Imerative (FSM) sublanguage,
- Handel-C (based on OCCAM),
- Object oriented: Liquid Metal (Lime Language, IBM), Kiwi (Univ Cambridge/Microsoft),
- Using Join Patterns or other parallel process algebra (C-omega).

The Kiwi, C-omega and Handel-C approaches start with parallel programs and exploit the scheduler non-determinacy to allow variations in implementation.

### 10.3.10 Join Calculus

Many software languages now include primitives for parallelism. (For instance, the CSharp language has the 'asynch' key word that can wrap up a method invocation via a delegate.)

The Join Calculus is an elegant concurrent programming language. It is globally asynchronous but supports local imperative programming (including a functional subset as usual).

In Join Calculus, a function has more than one signature - all need actual parameters available, generally provided from different sources, before an instance of the body is run. The set of signatures is called a chord.

```
//
//A simple join chord:
//
public class Buffer
{
    public asynch Put(char c);

    public char Get(bool f)&Put(char c)
    { return (f) ? toupper(c):c; }
}
```

New concurrency primitives - asynch dispatch - now builtin to C Sharp and F Sharp.

Has been compiled to hardware by a couple of research projects.

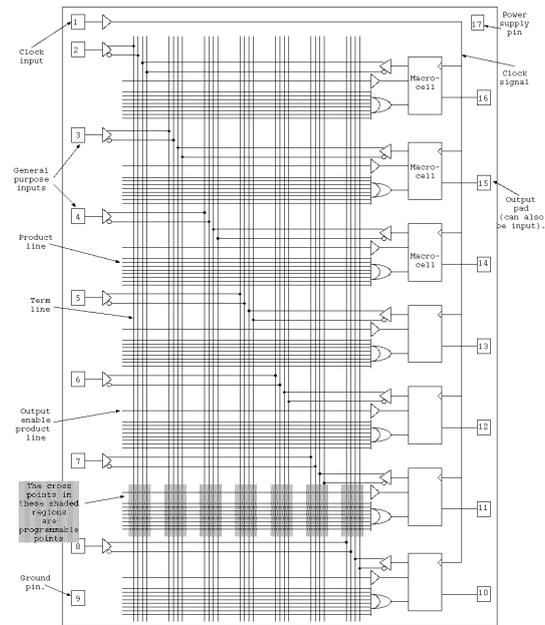


Figure 10.13: A typical PAL with 7 inputs and 7 I/Os.

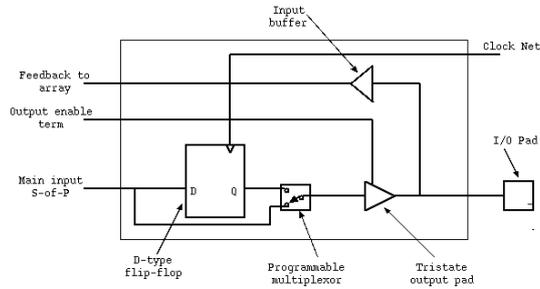


Figure 10.14: Contents of the example PAL macrocell.

### 10.3.11 Synopsys Behavioural Compiler

... was an advanced (for the late 90's) compiler that extended RTL synthesis semantics. Synopsys Behavioural Compiler Tutorial

- It implemented compile-time loop unrolling of loops containing event control (hence not like a generate construct that only elaborates structure),
- Arithmetic operations were freely moved between clock cycles,
- Additional cycles were inserted to overcome hazards (user's clock is called a 'super state'),
- Provided temporally-floating I/O with compiler-chosen pipelining between ports.

Problem: Existing RTL paradigms **not** preserved within the same source file: existing syntax has new meaning. Expert RTL users potentially lost their control over detailed structure in critical places.

Nonetheless, manual design of complex DSP accelerators using conventional RTL is very time consuming and error prone where the ALUs are heavily pipelined. So application-specific DSP compilers (such as Silage/Catherdral II) were widely used until general-purpose HLS tools matured.

**Additional notes:**

Citations:

- Understanding Behavioral Synthesis, A Practical Guide to High Level Design by John P Elliott; Kluwer Academic Publishers ISBN 0-7923-8542-X
- Behavioral Synthesis, Digital System Design Using the Synopsys Behavioral Compiler by David W. Knapp, Prentice Hall, ISBN 0-13-569252-0

## 10.4 Expression forms: Declarative Specifications

Rather than specify the algorithm (behaviour) we specify the required outcome. Rather like constraint-based linear programming, the design is a piece of hardware that satisfies a number of simultaneous assertions.

Examples:

- Synthesis using **Stepwise Refinement** from Formal Specs (Dijkstra 69),
- SAT-based logic Synthesis (Greaves 04),
- Rule-based hardware generation (Bluespec),
- Automatic Synthesis of Glue, Transactors and Bus Monitors (Greaves/Nam 10).

### 10.4.1 Synthesis from Formal Specification

It is desirable to eliminate the human aspect from hardware design and to leave as much as possible to the computer. The idea is that computers do not make mistakes, but there are various ways of looking at that!

A holy grail for CAD system designers is to restrict the human contribution towards a design to the top-level entry of a specification of the system in a formal language. By 'formal' we tend to mean a declarative language based on set theory and typically one in which it is easy to prove properties of the system. (The Part II course on hardware specification shows how to use predicate logic to do this.) The detailed design is then synthesised by the system from the specification.

There are many ways of implementing a particular function and the number of ways of implementing a complete system is infinite. Most of these are silly, a few are sensible and one, perhaps, optimum. Research using expert systems to select the best implementation is ongoing, but human input is needed in practical systems. But the human input should only be a guide to synthesis, choosing a particular way out of many 'formally correct' ways. Therefore errors cannot be introduced.

For instance, an inverter with input A and output B, expressed declaratively as predicates of time, can be specified as

$$\forall t. A(t) \leftrightarrow \neg B(t)$$

Here the logic levels of the circuit have the same notation as the logic values in the proof system, but an approach where they are separate might be typically needed when don't care states are encompassed.

$$\forall t. A(t) == 1 \leftrightarrow B(t) == 0$$

When time is quantised in units equal to a tick of the global clock then a D-type flip-flop can be expressed:

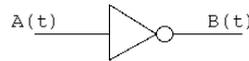
$$Q(t+1) == x \leftrightarrow D(t) == x$$

Here we have dropped the implied, leading  $\forall t$ .

Refinement outline:

Designs can be specified using predicate calculus.

Example, an inverter

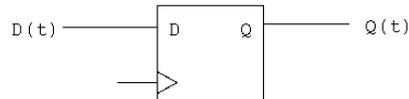


$$\forall t. A(t) \Leftrightarrow \sim B(t)$$

Above, the digital logic values are the truth values of the proof system, but they may be separated as follows:

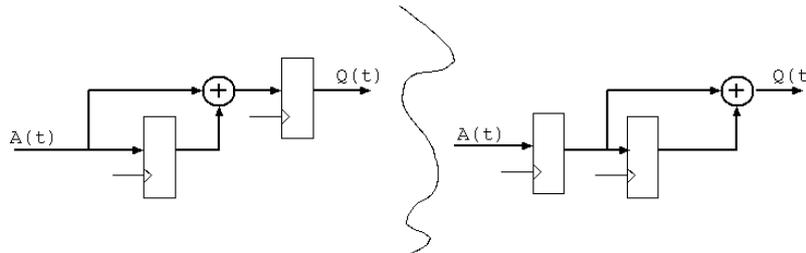
$$\forall t. A(t)==1 \Leftrightarrow B(t)==0$$

For synchronous machines with a single global clock, time may be quantised and time steps mapped to clock steps:



$$\forall t. D(t)==x \Leftrightarrow Q(t+1)==x$$

Of course, when D and Q are busses, multiple flip-flops are used forming a broadside register.



$$\forall t. Q(t)==A(t-1) + A(t-2)$$

Using such logic-based, formal specification, it is easy to specify systems that cannot be made, for instance, systems which are  
 - Non causal (future input affects current output).

Figure 10.15: Fragments: compilation from formal specifications.

1. Start with a formal spec plus a set of refinement rules,
2. Apply a refinement rule to some part of the spec,
3. Repeat until everything is executable.

A complex formal specification does not necessarily describe the algorithm and hence does not describe the logic structure that will be used in the implementation. Therefore, synthesis from formal specification involves a measure of inventiveness on the part of the tool.

Wikipedia: program refinement. Conversion from specification to implementation can be done with a process known as *selective stepwise refinement*. This chips away at bits of the specification until, finally, it has all been converted to logic. Some example rules for the conversion are given in Figure 10.15.

There are a vast number of refinement rules available for application at each refinement step and the quality of the outcome is sensitive to early decisions. Therefore, it is hard to make this fully automated.

Perhaps a good approach is for much of the design to be specified algorithmically by the designer (as in the above work) but for the designer to leave gaps where he is confident that a refinement-based tool will fill them. These gaps are often left by designers in their first pass at a design anyway; or else they are filled with some

approximate code that will allow the whole design to compile and which is heavily marked with comments to say that it is probably wrong. These critical bits of code are often the hardest to write and easiest to get wrong and are the bits that are most relevant to meeting the design specification. Practical examples are the handshake and glue logic for bus or network protocols.

Systems that can synthesise hardware from formal specifications are not in wide commercial use, but there is a good opportunity there and, in the long run, such systems will probably generate better designs than humans.

The synthesis system should allow a free mix of design specifications in many forms, including behavioural fragments and functional specifications. and only complain or fail when:

- the requested system is actually impossible: e.g. the output comes before the input that caused it,
- the system is over-specified in a contradictory way,
- the algorithm for implementing the desired function cannot be determined afterall.

### 10.4.2 Synthesis from Rules (SAT-based idea).

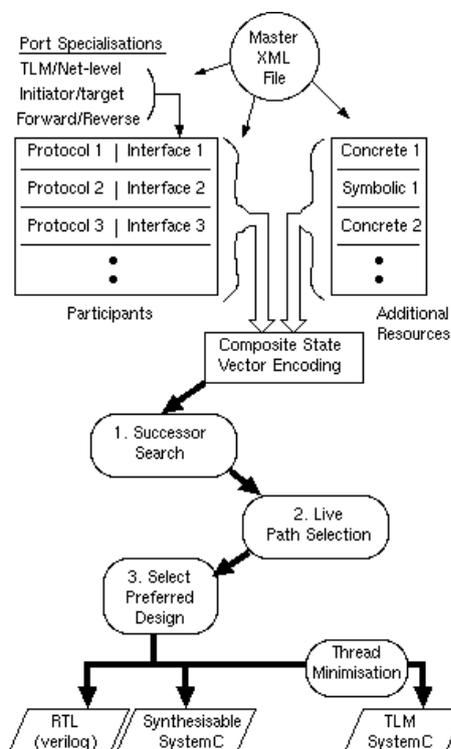
Crazy idea? If we program an FPGA we are generating a bit vector. SAT solvers produce bit vectors that conform to a conjunction of constraints.

Let's specify the design as a set of constraints over a fictional FPGA... We can also convert structural and behavioural design expressions to very-tight constraints and add those in.

The SAT solution wires up the FPGA and we can then apply logic trimming. LINK: SAT Logic Synthesis (Greaves)

Main problem: how large an FPGA to start with? Redundant logic might need a bi-simulation erosion to remove it.

Seems to work for generating small custom protocols.



Envisioned as an IP-XACT Eclipse Plugin:

1. XML file pulls protocols and interfaces from library.
2. Interfaces are parameterised with their direction and bus widths.
3. XML file also contains glue equations (e.g. filter predicates).
4. Additional resources added by human.
5. Then an automatic procedure...

### 10.4.3 Synthesis from Cross-Product (Greaves/Nam).

Can we automatically create RTL glue logic from port specifications ? Can the same method be used for joining TLM models ? Can the same method be used for making ESL-to-RTL transactors ?

Yes: [www.cl.cam.ac.uk/research/srg/han/hprls/orangepath/transactors](http://www.cl.cam.ac.uk/research/srg/han/hprls/orangepath/transactors) and Bus Monitors

Method is:

- List participating interfaces and their protocols,
- Specify the function needed: commonly just need data conservation, but sometimes need other operations:
  - Filtering
  - Multiplexing
  - Demultiplexing
  - Buffering
  - Serialising
  - Deserialising
- Add in additional resources that can be used by the glue (e.g. holding register or FIFO),
- Form protocol cross-product of all participants and resources,
- Trim so still fully-reactive and with no deadlocking trails,
- Emit resultant machine in SystemC or RTL of choice.

### 10.4.4 Expression forms: State charts and Graphical ‘languages’

Synthesis from diagrams (especially UML/SysML) embodies guarded actions:

- Full schematic entry at the gate level was once popular,
- Still popular for high-level system block diagrams,
- Also popular for state transition diagrams.

The stategraph general form is:

```

stategraph graph_name()
{
  state statename0 (subgraph_name, subgraph_entry_state), ... :

    entry: statement;
    exit:  statement;
    body: statement;

    statement;
    ...      // implied 'body:' statements
    statement;

    c1 -> statename1: statement;
    c2 -> statename2: statement;
    c3 -> exit(good);
    ...

    exit(good) -> statename3: statement;
    exit(bad)  -> statename4: statement;
    ...

  endstate

  state statename2:
  ...
  ...
  endstate

  state abort: // A special state that can be
               // forced remotely (also called disable).
  ...
}

```

There have been attempts to generate hardware systems via graphical entry of a finite state machine or set of machines. The action at a state or an edge is normally the execution of some software typed into a dialog box at that state, so the state machine tends to just show the top levels of the system. An example is the 'Specharts' system [IEEE Design and Test, Dec 92]. The Unified Modeling Language (UML) is promoted as *'the industry-standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems'* [Rational] for hardware too. Takeup of new tools is slow, especially if they are only likely to prove themselves as worth the effort on large designs, where the risk of using brand new tools cannot normally be afforded.

Schematic entry of netlists is now only applicable to specialised, 'hand-crafted' sub-circuits, but graphical methods for composing system components at the system-on-a-chip level is growing in popularity.

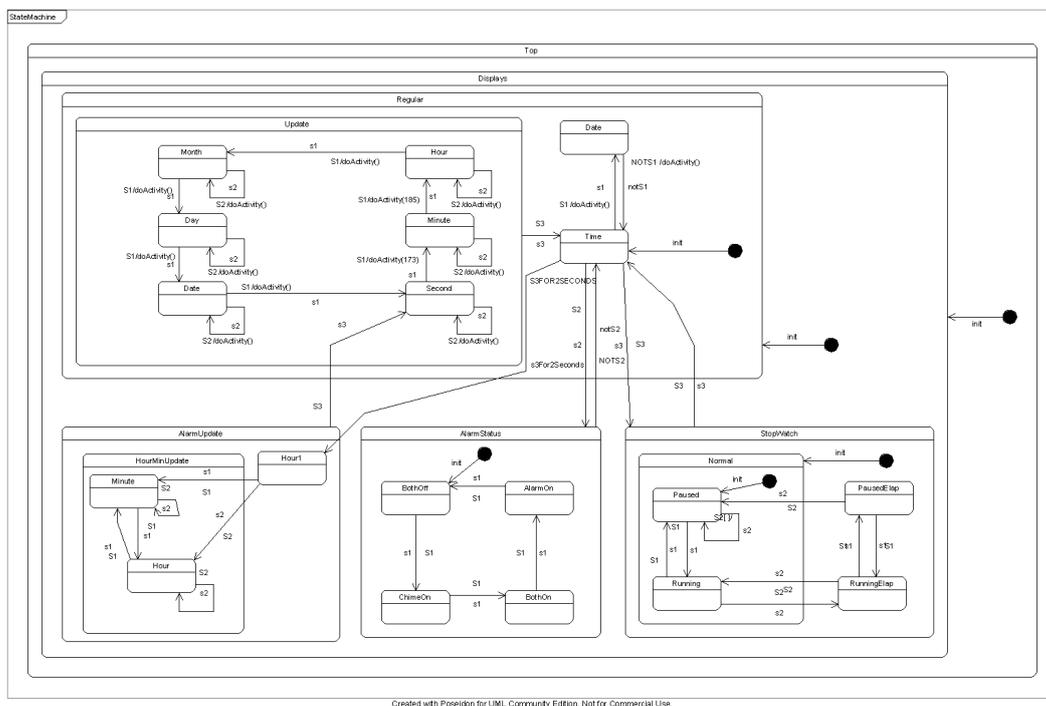


Figure 10.16: A statechart for a stopwatch (primoridon.com)

### 10.4.5 Statechart Details (from my experimental H2 Language).

#### Additional notes:

A state may contain tagged statements, each of which may be a basic block if required. They are distinguished using three tag words. The 'entry' statement is run on entry to the state and the 'exit' statement is run on exit. The 'body' statement is run while in the state. A 'body' statement must contain idempotent code, so that there is no concept of the number of times it is run while in the state. Statements with no tag are treated as body tagged statements. Multiple occurrences of statements with the same tag are allowed and these are evaluated as though executed in the textual order they occur or else in parallel.

A state contains transition definitions that define the successor states. Each transition consists of a boolean guard expression, the name of one of the states in the current stategraph and an optional statement to be executed when taking the transition. In situations where multiple guard expressions currently hold, the first holding transition is taken.

The guard expressions range over the inputs to the stategraph, which are the variables and events in the current textual scope, and the exit labels of child stategraphs.

When a child stategraph becomes active, it will start in the starting state name is given as an argument to the instantiation, or the first state of no starting name is given.

A child stategraph becomes inactive when its parent transitions, even if the transition is to the current state, in which case the child stategraph becomes inactive and active again and so transitions to the appropriate entry state.

A child stategraph can cause its parent to transition when the child transitions to an exit state. There may be any number, including zero, of exit states in a child stategraph but never any in a top-level stategraph. The parent must define one or more transitions to be taken for all possible exit transitions of its children. An exit state is either called 'exit' or 'exit(id)' where 'id' is an exit tag identifier. Exit tags used in the children must all be matched by transitions in the parent, or else the parent must transition itself under the remaining exit conditions of the child or else the parent must provide an untagged exit that is used by default.

### 10.4.6 All-forms High-level Synthesis Summary

This is the same slide as before!

The major EDA companies, Synopsys, Cadance and Mentor all heavily pushing C-to-Gates flows.

Altera (Intel) and Xilinx, the FPGA vendors, are now also promoting HLS tools.

Many people remain highly skeptical, but with FPGA in the cloud as a service in 2017 onwards, a whole new user community is garnered.

Success of formal verification means abundance of formal specs for protocols and interfaces: automatic glue synthesis seems highly-feasible.

Synthesis from formal spec - academic interest only ? Except for glue logic.

### 10.4.7 HLS to replace Von Neumann?

#### Certain maniacs predict FPGA may replace Von Neumann!

Spatio-Parallel processing uses less energy than equivalent temporal processing (ie at higher clock rates) for various reasons. David Greaves gives nine:

1. Pollack's rule states that energy use in a Von Neumann CPU grows with square of its IPC. But the FPGA with a static schedule moves the out-of-order overheads to compile time.
2. To clock CMOS at a higher frequency needs a higher voltage, so energy use has quadratic growth with frequency.
3. Von Neumann SIMD extensions greatly amortise fetch and decode energy, but FPGA does better, supporting precise custom word widths, so no waste at all.
4. FPGA can implement massively-fused accumulate rather than re-normalising after each summation.
5. Memory bandwidth: FPGA has always had superb on-chip memory bandwidth but latest generation FPGA exceeds CPU on DRAM bandwidth too.
6. FPGA using combinational logic uses zero energy re-computing sub-expressions whose support has not changed. And it has no overhead determining whether it has changed.
7. FPGA has zero conventional instruction fetch and decode energy and its controlling micro-sequencer or predication energy can be close to zero.
8. Data locality can easily be exploited on FPGA — operands are held closer to ALUs, giving near-data-processing (but the FPGA overall size is x10 times larger (x100 area) owing to overhead of making it re-configurable).
9. The massively-parallel premise of the FPGA is the correct way forward, as indicated by asymptotic limit studies [DeHon].

### 10.4.8 Embedded System

THIS SLIDE IS ON HOLD: WE WILL FINISH THIS AFTER THE ESL/TLM MATERIAL

The hardware and software components are entered in a common language.

A processor that is permanently connected to a single ROM which contains all of the software that the processor will ever execute is called an **embedded processor**. Here we consider when to use an embedded processor, and then, when using one, the designer has to choose whether to design a new one or use an existing one.

**The main difference between a hardware solution and a software solution is the degree of parallelism.** Processors typically execute one instruction at a time, re-using the same hardware components again and again. Hardware solutions tend to have dedicated circuits for each function. If most of the hardware is likely to be idle for most of the time, a processor is preferred, but if a processor cannot achieve the throughput required, increased parallelism using hardware is preferred.

Complex functions normally require a processor, but CAD tools are evolving, allowing complex functions to be expressed algorithmically and then automatically converted to a logic gate implementation.

High-speed processing normally requires dedicated hardware. For instance, consider the error correction performed by a CD player to overcome dirt and scratches. When CD players first came out, this error correction was done with dedicated hardware, but today, microprocessors have increased in speed, and so the function can be done using the processor that is already there to provide other complex functions (e.g. track skip). However, on the latest, x52 speed CD/DVD drives the error correction must be performed that much faster, and so dedicated hardware is re-introduced.

#### 10.4.9 Processors can be placed on an ASIC.

Processors give flexibility: if the design is likely to be changed in minor ways to form new models, or to implement field upgrades, then using a processor that may be given a new software release is a good technique.

Standard processor chips can be very cheap for a given performance. This is because a very great deal of effort is put into their design and they are sold in large quantities, thereby reducing price. Part of the cost of any product is in testing it. A standard processor not only comes with its own test qualification programme, it is able to execute software to perform tests on the rest of the system.

The decision to design and use a custom processor for an application should not be taken lightly. It can be useful, however, when the application is fairly simple or highly specialised. Examples are found in signal processing, where a particular algorithm needs to be executed at high speed; for instance, to track a missile or in the steady-cam function of a camcorder.

#### 10.4.10 Bus Mux

```
class busmux:sc_module
{
public:
// Use tagged sockets to be able to distinguish incoming backward path calls
tlm_utils::multi_passthrough_target_socket<busmux>   targ_socket;
tlm_utils::multi_passthrough_initiator_socket<busmux> init_socket;

// Constructor
busmux(sc_module_name name, uint32_t threshold);

// FORWARD PATH
// TLM-2 blocking transport method
virtual void b_transport(int id, tlm::tlm_generic_payload &trans, sc_time &delay);
uint32_t threshold;
};
```

```

busmux::busmux(sc_module_name name, uint32_t threshold):
    sc_module(name),
    targ_socket("targ_socket"),
    init_socket("init_socket"),
    threshold(threshold)
{
    // Register callbacks for incoming interface method calls
    targ_socket.register_b_transport(this, &busmux::b_transport);
}

// TLM-2 blocking transport method
void busmux::b_transport(int id, tlm::tlm_generic_payload &trans, sc_time &delay)
{
    uint32_t adr = (uint32_t) trans.get_address();
    if (adr < threshold) init_socket[0]->b_transport(trans, delay);
    else
    {
        init_socket[1]->b_transport(trans, delay);
    }
}

```

## 10.5 Synthesis & Simulation Algorithms:

Toy versions of two RTL processing algorithms are described in this section:

- (EDS) Event-driven simulation,
- Gate-level synthesis.

Gate-level generator code for multipliers, adders, shifters, checkers and other low-level operators is given elsewhere.

```

datatype m_t = M_AND | M_OR | M_INV | M_XOR | M_DFF | M_BUFIF | M_TLATCH | M_CLOCK;

datatype internal_state_t =
    IS_NONE
  | IS_DFF of value_t ref
  | IS_CLOCK of int ref
;

and
    model_t = MODEL of string * int * m_t * net_t list * internal_state_t
;

```

This form of model requires the leaf components (gates) to be built in to the simulator with hard-coded behaviour.

Another form of model, derived originally from Simula in the 1960s, uses user-coded leaf models to be implemented using an imperative language that is interpreted by the simulator.

Verilog RTL supports both built-in gate models and modelling with behavioural threads. SystemC has no built-in models.

### 10.5.1 Toy implementation of EDS RTL Simulator.

### 10.5.2 Non Determinism.

If we have the following:

```
always @(posedge clk) v <= 1;
always @(posedge clk) v <= 0;
```

we have a non-deterministic design (banned for synthesis).

*Exercise:*

Here's a more interesting example:

```
always @(posedge clk) begin
    v1 = v1 + 10;
    v2 = v3;
end

assign v3 = v1 + 1;
```

What happens here: does the continuous assignment get executed between the assignments of v1 and v2 ?

### 10.5.3 Basic RTL Synthesis Algorithm

Compiling RTL to gates involves three main steps:

- convert to a pure RTL form where each variable is assigned only once,
- convert each assignment of a vector variable to a list of assignments to individual bits,
- convert the r.h.s. of each bit assignment to a network of gates.

Firstly, to convert to 'pure RTL', for each register we need exactly one hardware circuit for its input, regardless of however many times it is assigned, so we need to build a multiplexor expression that ranges over all its sources and is controlled by the conditions that make the assignment occur. In other words, we need a list for each clock domain that holds pairs of the form (register name, value to be assigned on clock edge).

There are two varieties of the conversion to pure RTL algorithm depending on whether non-blocking signal assigns are used or normal variable assignment is used. The difference is simply whether we need to look up variables occurring on the right-hand side of expressions in the list of already assigned variables. The two techniques can be mixed when both forms of assignment are present.

Conversion to 'pure RTL' list form, ML fragment

Secondly, for each register that is more than one bit we generate separate assignments for each bit. This is colloquially known as 'bit blasting'. Logic and sub-expressions can be shared between variables and bit lanes of a given variable. This stage removes arithmetic operators and leaves only boolean operators.

Conversion to Bit Blasted Form, ML fragment

Thirdly, we produce gate-level circuits for each of the expression trees with a gate builder function that recurses to the leaves of the expression and emits the gates, returning their output net name as it returns up the stack. gatebuilder, ML fragment

Assignments to arrays are slightly more problematic.

The **name alias problem** is that at compile time we might not be able to determine whether a pair of subscripts are going to be the same or not at run time, and hence, for blocking variable assigns we cannot always do a lookup. Secondly, the restricted number of ports leads to hazards that may need the design to be re-timed.

Of course, this is a simplified approach to logic synthesis and real tools must consider sub-expression sharing and replication depending on whether they are aiming for speed, area, power or some composite performance goal. Also, not all arithmetic units should be converted to gates: it is better to implement by instantiating special-purpose components. If these components are not-fully pipelined then we get further hazards.

### 10.5.4 Examples of conversion to binary (bit lane) form

Sign extend an arg to width n:

```
fun sex n nil = if n<=0 then nil else raise "cannot do sex on an empty list"
| sex n [item] = if n<=1 then [item] else item :: sex (n-1) [item]
| sex n (h::t) = h :: (sex (n-1) t);
```

Example: integer constant:

```
| pandex w (Num n) = if n = 0 then [ xi_false ]           else
  let fun k 0 = nil (* lsb first *)
      | k n = (if (n mod 2)=1 then xi_true else xi_false) :: k (n div 2)
      fun q 0 = [xi_true] (* final negative sign bit *)
      | q n = (if (n mod 2)=0 then xi_true else xi_false) :: q (n div 2)
      in if (n >= 0) then k n else sex w (q (0-1-n)) end
```

Example: -4 in a 6 bit field is 111100.

Example: conditional expression: a broadside multiplexor:

```
| pandex w (Query(g, t, f)) =
  let val t' = pandex w t
      val f' = pandex w f
      fun k([], []) = []
      | k(a, nil) = k(a, [ false ])
      | k(nil, b) = k([ false ], b)
      | k(a::at, b::bt) = gen_mux2(g, a, b) :: k(at, bt)
      in k(t', f') end
```

See other examples in the multiplier/adder additional material.

### 10.5.5 Automatic Synthesis of Glue and Interface Automata

If two (or more) interfaces obey protocols each defined by an automaton

and if we have some over-arching design goal (e.g. data conservation)

then we can take the cross-product of the constituent automata and trim it to form our design.

Required trims:

- Non-deadlocking (live path),
- Data conserving (shovels data according to our intentions).

See Convertibility verification and converter synthesis: two faces of the same coin. Passerone et al

May wish to consider various point-to-point data path patterns:

- plain old data conserving, or
- filtering, or
- serialising, or
- deserialising, or
- with internal multi-stage FIFO, or

Can consider data patterns for more than two ports:

- multiplexor, or
- demultiplexor, or
- router.

Greaves current research is to do apply this at the TLM level and between levels (with MJ Nam).

Certain design patterns require extra state to be 'thrown in', giving problems like the over large fictional FPGA problem for the SAT synthesis idea.

So that components can be interconnected easily, automatic synthesis of glue logic is a critical future development. A standardised interconnect format should allow bus creation to be fully automated. The bandwidth budget of a bus depends on the number of data wires used and the number of devices attached. It may be pipelined, giving various latencies of access. In the future, the user should just express the bandwidth budget and the interconnections should be synthesised, provided each component has been designed to a standardised interface paradigm. The paradigm could be a first-class component of the HDL.

A standard EDA interconnection networking system can also usefully serve as the partition between different modelling styles. For instance, a system simulation may involve a mixture of gate-level modelling, behavioural models and cycle-accurate models. Each style of simulation might have variations in how accurately it represents time. A standard interconnection network would support flow control so that variations in timing can be accommodated, but the total number and ordering of 'transactions' between models would be preserved.

### 10.5.6 Automatic Synthesis of Transactors and Glue Logic

In a previous section we mentioned Automatic Synthesis of Transactors and Bus Monitors. Temporal logics, such as PSL and SVA can be compiled to automata that serve as:

- Checkers for Dynamic Validation (e.g. a bus protocol monitor),
- Transactors (for converting from TLM to net-level) in ESL.

#### Transactor Synthesis

A behavioural program that drives outputs can be 'reflected' to become a checker by interpreting the assignments as assertions.

If we have a formal specification of participating interfaces, we aim to synthesise the 'glue'

- RTL glue logic between components for use on target SoC,
- TLM glue between components for modelling,
- Complex switch/router components with filtering might also be made,

- Cross-layer glue: TLM to RTL transactors.

Net and method definitions could come from IP-XACT but it does not define protocols, so this worked example we use a separate mini-language for the protocols:

```

type protocol_spec_t =
| Set of string * string
| Seq of protocol_spec_t list
| Conjunction of protocol_spec_t list
| Disjunction of protocol_spec_t list
;

type net_spec_t =
| Concrete_enum of string list
| Symbolic_bitvec of int
;

(*
 * In a reversed port, the inputs are swapped with the outputs, apart from Always_x forms
 * that do not swap direction (e.g, master clock).
 *)
type netdir_t = Input | Output | Local | Always_input
;

```

### 10.5.7 Formal Specification of the Four Phase Handshake

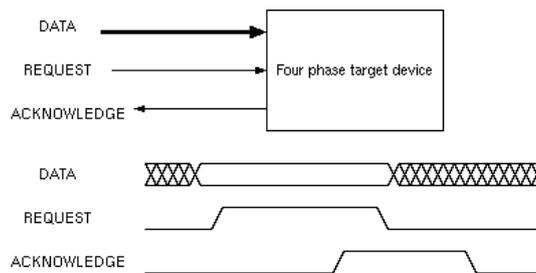


Figure 10.17: Four phase handshake nets and protocol.

```

let four_phase_handshake_nets =
[ ("Strobe", Output, Concrete_enum ["true"; "false"]);
  ("Ack", Input, Concrete_enum ["true"; "false"]);
  ("DL", Output, Symbolic_bitvec (8) )
]

let four_phase_handshake_idle =
[ ("dead", "DL");
  ("false", "Strobe");
  ("false", "Ack")
]
;

let four_phase_handshake_protocol =
Seq[
  Set("live", "DL");
  Set("true", "Strobe");
  Set("true", "Ack");
  Set("dead", "DL");
  Set("false", "Strobe");
  Set("false", "Ack")
]
;

(* Overall composition of the protocol in a direction-agnostic form *)
let fourphase = ("4P", four_phase_handshake_nets, four_phase_handshake_idle, four_phase_handshake_protocol)

```

### 10.5.8 Formal Specification of the TLM Call with one argument

```

(* The net directions are for the male (data sending) side but the protocol specification is unsexed. *)
let tlm1_handshake_nets =
  [ ("call",   Input,   Concrete_enum ["idle"; "active"]);
    ("DH",    Input,   Symbolic_bitvec (8) )
  ]

let tlm1_handshake_idle =
  [
    ("dead", "DH");
    ("idle", "call");
  ]
;

(* This is for a TLM entry point that is a data sink.  If it were a source, the data would be live as the call goes idle. *)
let tlm1_handshake_protocol =
  Seq[
    Set("live", "DH");
    Set("active", "call");
    Set("dead", "DH");
    Set("idle", "call");
  ]
;

```

The product machine has a state transition when the TLM caller puts its data live, but of course the glue logic will not see this until the call state goes live, so it cannot really make a transition at that point. Therefore a pre-processing of the participating state machines is live that elides the two operations into one is appropriate.

This is for a TLM entry point that is a data sink. If it were a source, the data would be live as the call goes idle.

In the example live and active occur at once, whereas the dead transition is not elided, since the glue will make an active transition to read out the contents of the register. In a complementary transactor where the TLM is a write upcall, that has the same overall pattern, it is the going dead transition that cannot be seen by the glue, and the going dead and call going idle are elided in the preprocessing.

```

(* Overall composition of the protocol in a direction-agnostic form *)
let tlm1 = ("TLM1", tlm1_handshake_nets, tlm1_handshake_idle, tlm1_handshake_protocol)

```

### 10.5.9 Formal Specification of a Holding Register

We are free to throw any other interfaces or resources into the mixture, such as a holding register or two...

```

let holder8_nets =
  [
    ("H8",    Local, Symbolic_bitvec (8) )
  ]

let holder8_idle =
  [
    ("dead", "H8");
  ]

let holder8_protocol =
  Seq[ ] (* Null protocol - can be freely used *)

let holder8 = ("HOLDER8", holder8_nets, holder8_idle, holder8_protocol)

```

### 10.5.10 Composing a Mixture

Perhaps using drag and drop under IP-XACT/Eclipse, we generate a mixture of our components:

```
form_product [ fourphase; tlm1; holder8 ]
```

A fragment of the output code:

```
if (reset) state = 215; else
switch(state)
{
case 215: //4P_state_0/PC_4P    TLM1_state_0/PC_TLM1 idle/call
// 634,4P_state_1/PC_4P Ack TLM1_state_1/PC_TLM1 idle/call DH : Ok E=0 A=0 Rank=-10
if ((DH==live)&&(Ack==true)&&1) { state = 634;}
break;
// 874,4P_state_1/PC_4P Ack TLM1_state_1/PC_TLM1 active/call DH : Ok E=0 A=0 Rank=-19
if ((DH==live)&&(call==active)&&(Ack==true)&&1) { state = 874;}
break;
// 646,4P_state_1/PC_4P    TLM1_state_1/PC_TLM1 idle/call DH : Ok E=0 A=0 Rank=-1
if ((DH==live)&&1) { state = 646;}
break;
...

case 646: //4P_state_1/PC_4P    TLM1_state_1/PC_TLM1 idle/call DH
// 826,4P_state_1/PC_4P Ack TLM1_state_2/PC_TLM1 active/call DH : Ok E=0 A=0 Rank=-14
if ((call==active)&&(Ack==true)&&1) { state = 826;}
break;
// 838,4P_state_1/PC_4P    TLM1_state_2/PC_TLM1 active/call DH : Ok E=0 A=0 Rank=-5
if ((call==active)&&1) { state = 838;}
break;
// 634,4P_state_1/PC_4P Ack TLM1_state_1/PC_TLM1 idle/call DH : Ok E=0 A=0 Rank=-9
if ((Ack==true)&&1) { state = 634;}
break;
// 874,4P_state_1/PC_4P Ack TLM1_state_1/PC_TLM1 active/call DH : Ok E=0 A=0 Rank=-18
if ((call==active)&&(Ack==true)&&1) { state = 874;}
break;
// 886,4P_state_1/PC_4P    TLM1_state_1/PC_TLM1 active/call DH : Ok E=0 A=0 Rank=-9
if ((call==active)&&1) { state = 886;}
break;
}
```

### 10.5.11 Working Output

Wrap this up in a suitable wrapper and it works!

```
SC_MODULE(xactor)
{
public:
int state;
bool call;
bool DH;
uint8_t DH_d, H8_d;
void advance();

void TLM_call(uint8_t d)
{
DH_d = d;
DH = live;
call = active;
do { sc_wait(); } while (call != idle);
}

void process()
{
#include "generated_state_machine_code"
printf("State Strobe=%i Ack=%i DL_d=%i, state=%i\n", Strobe, Ack, DL_d, state);
}

SC_CTOR(xactor)
{
SC_THREAD(process) ;
...
}
}
```

Here's the working output:

```

State Strobe=0 Ack=0 DL_d=0, state=215
State Strobe=0 Ack=0 DL_d=0, state=215
State Strobe=0 Ack=0 DL_d=0, state=215
State Strobe=0 Ack=0 DL_d=0, state=646
State Strobe=0 Ack=0 DL_d=0, state=838
State Strobe=1 Ack=0 DL_d=10, state=327
netlevel4psink sink0: received data 10
State Strobe=1 Ack=1 DL_d=10, state=807
State Strobe=1 Ack=1 DL_d=10, state=794
State Strobe=0 Ack=1 DL_d=10, state=775
State Strobe=0 Ack=0 DL_d=10, state=786
State Strobe=0 Ack=0 DL_d=10, state=790
State Strobe=0 Ack=0 DL_d=10, state=502
State Strobe=0 Ack=0 DL_d=10, state=646
State Strobe=0 Ack=0 DL_d=10, state=838
State Strobe=1 Ack=0 DL_d=10, state=327
netlevel4psink sink0: received data 10
State Strobe=1 Ack=1 DL_d=10, state=807
State Strobe=1 Ack=1 DL_d=10, state=794
State Strobe=0 Ack=1 DL_d=10, state=775
State Strobe=0 Ack=0 DL_d=10, state=786
State Strobe=0 Ack=0 DL_d=10, state=790
State Strobe=0 Ack=0 DL_d=10, state=790
State Strobe=0 Ack=0 DL_d=10, state=790

```

If two (or more) interfaces obey protocols each defined by an automaton

and if we have some over-arching design goal (e.g. data conservation)

then we can take the cross-product of the constituent automata and trim it to form our design.

Required trims:

- Non-deadlocking (live path),
- Data conserving (shovels data according to our intentions).

See Convertibility verification and converter synthesis: two faces of the same coin. Passerone et al

May wish to consider various point-to-point data path patterns:

- plain old data conserving, or
- filtering, or
- serialising, or
- deserialising, or
- with internal multi-stage FIFO, or

Can consider data patterns for more than two ports:

- multiplexor, or
- demultiplexor, or
- router.

Greaves current research is to do apply this at the TLM level and between levels (with MJ Nam).

Certain design patterns require extra state to be 'thrown in', giving problems like the over large fictional FPGA problem for the SAT synthesis idea.

So that components can be interconnected easily, automatic synthesis of glue logic is a critical future development. A standardised interconnect format should allow bus creation to be fully automated. The bandwidth budget of a bus depends on the number of data wires used and the number of devices attached. It may be pipelined, giving various latencies of access. In the future, the user should just express the bandwidth budget and the interconnections should be synthesised, provided each component has been designed to a standardised interface paradigm. The paradigm could be a first-class component of the HDL.

A standard EDA interconnection networking system can also usefully serve as the partition between different modelling styles. For instance, a system simulation may involve a mixture of gate-level modelling, behavioural models and cycle-accurate models. Each style of simulation might have variations in how accurately it represents time. A standard interconnection network would support flow control so that variations in timing can be accommodated, but the total number and ordering of 'transactions' between models would be preserved.

### 10.5.12 Transactor Configurations (2nd)

Four possible transactor's are envisionsable for the 4/P handshake.

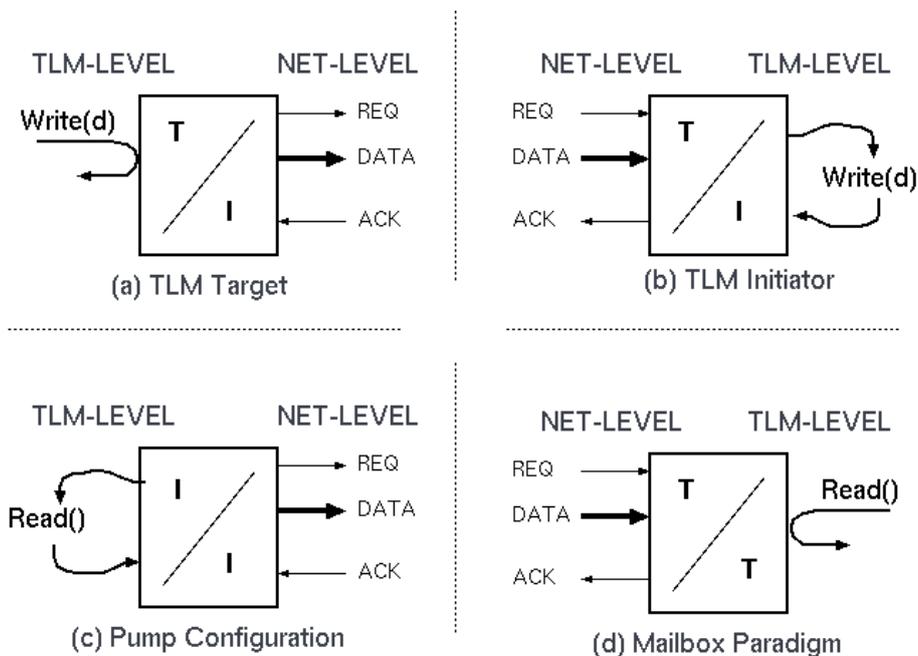


Figure 10.18: Possible configurations for basic TLM to net-level transactors.

The form that is an initiator on boths sides will never do anything!

The form that is a target on boths sides follows the *mailbox* design pattern.

An (ESL) Electronic System Level *transactor* converts from a hardware to a software style of component representation. A hardware style uses shared variables to represent each net, whereas a software style uses callable methods and up-calls. Transactors are frequently required for busses and I/O ports. Fortunately, formal specifications of such busses and ports are becoming commonly available, so synthesising a transactor from the specification is a natural thing to do.

A transactor tends to have two ports, one being a net-level interface and the other with a thread-oriented interface defined by a number of method signatures. The thread-oriented interface may be a target that accepts

calls from an external client or it may be an initiator that make calls to a remote client. The calls may typically be blocking to implement flow control.

There are four forms of transactor for a given bus protocol. Either side may be an initiator or a target, giving four possibilities.

The initiator of a net-level interface is the one that asserts the command signals that take the interface out of its starting or idle state. The initiator for an ESL/TLM interface is the side that makes a subroutine or method call and the target is the side that provides the entry point to be called.

. come EOF

## 10.6 Adder & Multiplier Structures.

### 10.6.1 Adder Build (Synthesis)

Adding a pair of bit lists, lsb first.

Ripple carry adder:

```
fun add c (nil, nil) = [c]
| add c (a::at, b::bt) =
  let val s = gen_xor(a, b)
      val c1 = gen_and(a, b)
      val c2 = gen_and(s, c)
      in (gen_xor(s, c))::(add (gen_or(c2, c1)) (at, bt))
  end
```

Faster adder: use wide gates: use functions like gen\_add1

Carry argument is replaced with a list of generate and propagate pairs from the earlier stages.

The ripple carry adder is generated by the ML fragment above. The kogge stone is frequently used as a practical, synthesisable adder that is fast and not critical over its layout.

### 10.6.2 Kogge Stone adder

Kogge-Stone is very fast and the area is not too bad, but the wiring is not regular.

Synthesises well. Hard to understand!

For FPGA: Just use RTL '+' and FPGA tools instantiate special paths.

**Ex (long):** Write a Kogge-Stone generator.

### 10.6.3 Subtractor, Equality, Inequality, Shifts

Using negation and swapping, all of these operators boil down to just two (addition and equality): + - < > ?= <= != == << >> .

Subtractor: use an adder but instead pass in a one as the leading borrow-bar and complement each bit from the second operand.

A subtractor will generate a borrow output. If  $a < b$  then  $a-b$  will need a borrow, hence the raw subtractor implements less-than.

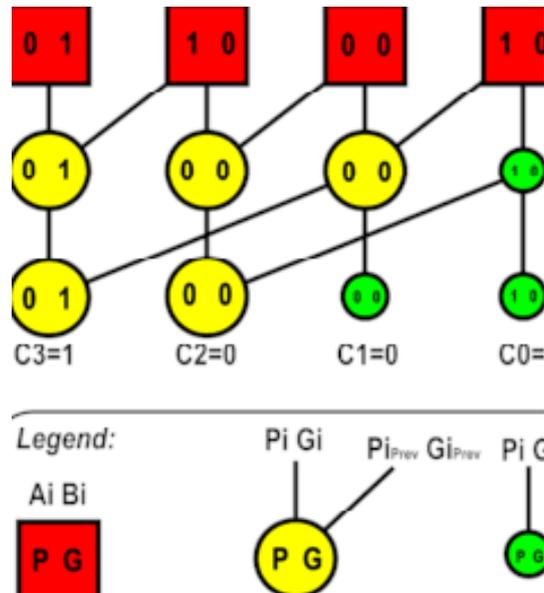


Figure 10.19: Recursive structure of the the Kogge Stone Adder.

Greater than or equal is just the complement of less than. For the other two inequalities: just swap the operands.

Equality test: one could use a subtractor and check the output is zero, but it's simpler to use a disjunction of XORs (i.e. we don't need to consider carry operations).

### 10.6.4 Long Multiplication

Flash multipliers use special wallace tree cells (not lectured) whereas long multiplication using Booth's algorithm is a nice example of a non-fully pipelined multiplier.

Flash multiplier - combinatorial implementation (e.g. a Wallace Tree).

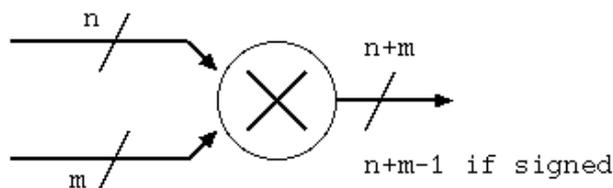


Figure 10.20: Multiplier schematic symbol.

#### Sequential Long Multiplication

```

RA=A
RB=B
RC=0
while(RA>0)
{
  if odd(RA) RC=RC+RB;
  RA = RA >> 1;
  RB = RB << 1;
}
    
```

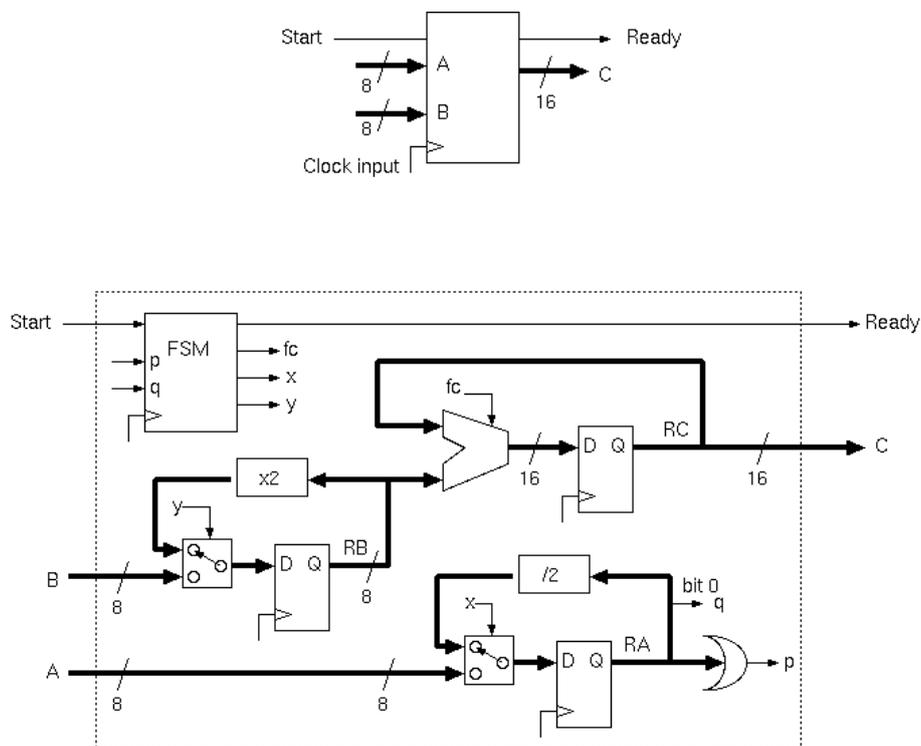


Figure 10.21: A suitable data path for a long multiplier.

### 10.6.5 Micro-Architecture for a Long Multiplier

Figure 10.21 implements conventional long multiplication.

Certainly not fully-pipelined.

**Exercise:** Write out the complete design, including sequencer in RTL or SystemC.

### 10.6.6 Booth's Multiplier

Booth does two bits per clock cycle:

```
(* Call this function with c=0 and carry=0 to multiply x by y. *)
fun booth(x, y, c, carry) =
  if(x=0 andalso carry=0) then c else
let val x' = x div 4
    val y' = y * 4
    val n = (x mod 4) + carry
    val (carry', c') = case (n) of
      (0) => (0, c)
    |(1) => (0, c+y)
    |(2) => (0, c+2*y)
    |(3) => (1, c-y)
    |(4) => (1, c)
  in booth(x', y', c', carry')
end
```

**Exercise:** Design a micro-architecture consisting of an ALU and register file to implement Booth. Design the sequencer too.

### 10.6.7 Shifters

Constant shifts: these are trivial in bit lane form.

## KG 11 — SoC D/M Course Practicals 2: Toy ESL Using TLM 1.0

These toy ESL classes start with an RTL style and then use a TLM 1 style of modelling. Both use a small processor ISS call 'nominalproc'.

They are numbered 3 to 6, with 1 and 2 being the previously-covered SystemC get started LINK

### 11.0.8 Toy ESL Class 3: RTL-style Wiring

The files for class 3 are may be found in SOCDAM /toyclasses/class3.

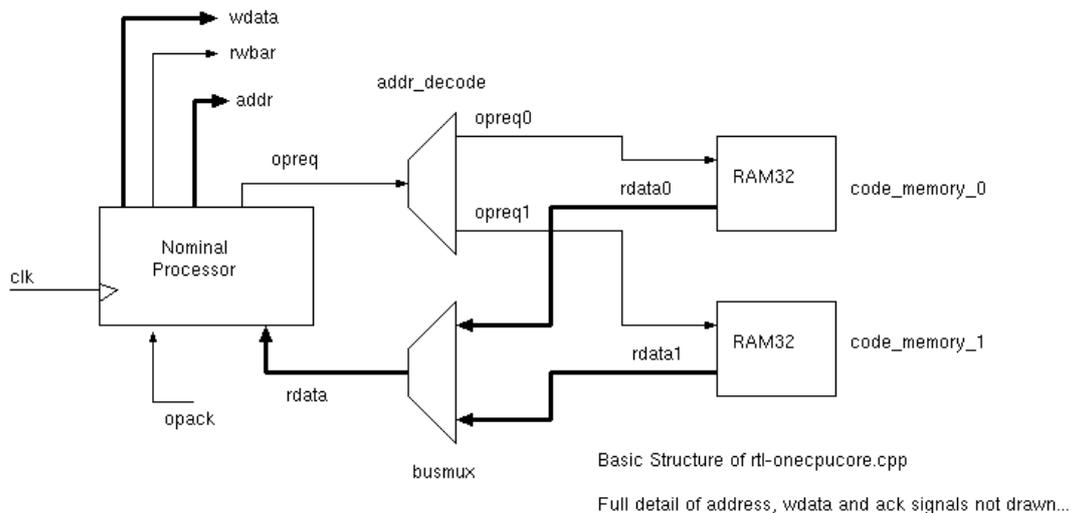


Figure 11.1: Basic structure of the RTL-style, single core, nominal processor configuration.

This includes an abstract ISS for a really trivial microprocessor called nominalproc as well as a concrete ISS and an RTL-style bus initiator in SystemC.

Some simple IP blocks (address decoder and response mux) for building a tree-structure can be examined.

**Exercises:** using the RTL-style blocks provided, please experiment with various configurations and understand how you would make a more complex system using further `addr_decode` and `busmux` components to address the components.

### 11.0.9 Class 3: Provided Compilation Targets

- Makefile
- nonsysmain.cpp
- rtl-onecpucore.cpp
- rtl-twocpucores.cpp(for class 5).

### 11.0.10 DJIP Blocks - Toy IP Blocks.

A high-level model of the nominal processor:

- CBG Nominal Processor - ISS CORE

The MPhil course work will use the OpenRISC or ARM processors and not this nominal processor.

A library of SystemC IP blocks that can be used by various targets in different classes:

- Address Decoder
- Bus Bridge (fragment of)
- busmux.h
- clock100.h
- DMA Controller
- Nominal Processor - Transacted to RTL
- ram32.h - RTL style memory model

TLM (transactional-level modelling) versions of example IP blocks:

- Example transport port (simpler than TLM2.0)
- Nominal Processor - TLM version
- *t1m<sub>b</sub>usmux.h – TLM version*
- t1mram32.h - TLM style memory model

For simplicity, these TLM blocks do not use the TLM2 library, but they do have a TLM2-like standard payload that is passed between the TLM components.

### 11.0.11 Toy Class 4 : ESL-style 'Wiring'

**Aim:**In class 4 we look at the TLM version of nominalproc and the TLM version of the RAM t1mram32 . h.

We also look at the tiny **Quantum Keeper** embedded in the TLM version of nominal processor.

We look at how the memory models can be called in both a TLM style and an RTL style from a TLM style implementation of nominalproc.

### 11.0.12 Class 4: Provided Compilation Targets

- Makefile
- t1m-onecpucore.cpp

Figure 11.2: Structure of the second file.

The t1m-onecpucore.cpp is unmodified at the start, still instantiating the RTL memory components.

Then it instantiates a TLM processor, memory and busmux.

It is left as Exercise 1 below to write a **transactor** that connects the TLM processor to the instantiated RTL parts so that they too can be used.

### 11.0.13 Class 4: Optional Exercises

#### Exercise 1:

Write and instantiate a transactor that connects the TLM processor to RTL-style components of the last class. This gives us a mixed abstraction simulation. The transactor is best implemented as its own SystemC module instantiated alongside the components it connects.

#### Exercise 2:

Introduce a high-level model of a new peripheral of your own design and a device driver for it into the system. Each can be entirely trivial: it will still demonstrate the coding style.

You may replace one of the memory blocks with your peripheral or else add a further bus multiplexor and connect the peripheral along side the existing memory blocks. Or, to make it really easy, you can just pretend one of the memory blocks is your peripheral.

Then modify the test program loaded into the nominal processor to run a cycle or two on your new peripheral.

Finally, sketch some typical device driver code for your new peripheral and show how you could make a 'mid-level model' where the device driver calls the code of the peripheral directly. Ideally you can reuse the same body of code in various places with conditional compilation using the C++ preprocessor to adjust the level of detail.

### 11.0.14 Toy Class 5 : Multiple Bus Masters

The new component for this class is the bus arbiter in TLM/ESL form `tlm_busarb.h`.

**Aim:** The aim of this class is to see that the bus arbiter is trivial in TLM form, especially when untimed.

Two instances of the nominal processor are connected to one bus.

The ESL-style bus arbiter implements no queueing and allows more than one thread to block in the target at once, but exclusion should be added to model more-realistic sharing.

**Exercise:** Introduce some timing or other accounting into the TLM bus arbiter and see if similar answers are as predicted by fluid-flow/spreadsheet style analysis. Experienced RTL designers may wish to implement an RTL-style arbiter and compare answers with that.

### 11.0.15 Class 5: Provided Compilation Targets

- Makefile
- `tlm-dualmasters.cpp`

### 11.0.16 Toy Class 6 : DMA Controller - Partial

There is an implementation the DMA controller and other IP blocks to look at.

- DMA Controller
- `rtl-twocpucores.cpp`(for class 6).

**Aim:**To complete a full working example with a running program that performs the DMA.

We look at a simple, one-channel DMA unit, but they can be as quite complex in their following of linked lists and so on. So what's the difference from a processor then what: no instruction fetch, functionality fixed at tape out.

**Exercise:** Complete the example, putting code on the processor to initiate a DMA block copy.