Comparative Architectures

CST Part II, 16 lectures

Lent Term 2004

Ian Pratt

Ian.Pratt@cl.cam.ac.uk

OHP slides with extra notes for supervisors

Course Outline

1. Comparing Implementations
   - Developments fabrication technology
   - Cost, power, performance, compatibility
   - Benchmarking

2. Instruction Set Architecture (ISA)
   - Classic CISC and RISC traits
   - ISA evolution

3. Microarchitecture
   - Pipelining
   - Super-scalar
     - static & out-of-order
   - Multi-threading
   - Effects of ISA on μarchitecture and vice versa

4. Memory System Architecture
   - Memory Hierarchy

5. Multi-processor systems
   - Cache coherent and message passing

Understanding design tradeoffs

Course Outline

1. *Instruction Set Architecture is the programmer-visible processor features.*
2. *ISA = Instruction Set Architecture, not Industry Standard Architecture (old PC I/O bus).*
3. *This course looks at processors and memory systems, not I/O.*
4. *Examines tradeoffs available to processor architects.*
5. *Every feature has costs. Some make chip more efficient, other less so.*
6. *Why you need to know this:*
   even if not architect, comparing performance, optimisations under the hood easily factors of 10
7. *How to make computers go fast, perhaps the sexiest part of computer science, atleast in my opinion*
8. *Don't get too bogged down in detail*
9. *But, some examples in essays are good.*
10. *This isn't all black and white. Many issues still undecided*
11. *Hard to compare the performance of* architectures*; depends on implementation. Depends on on how much money, fabrication technology, how smart..*

Reading material

- OHP slides, articles
- Recommended Book:
  John Hennessy & David Patterson,
  Computer Architecture: a Quantitative Approach
  (3rd ed.) 2002 Morgan Kaufmann
- The Web
  `http://bwrc.eecs.berkeley.edu/CIC/`
  `http://www.chip-architect.com/`
  `http://www.geek.com/procspec/procspec.htm`
  `http://www.realworldtech.com/`
  `http://www.anandtech.com/`
  `http://www.arstechnica.com/`
  `http://open.specbench.org/`
- comp.arch News Group

1. *3rd ed. out soon*
2. *First few lectures, Notes for lecturer, not necessarily for you!*
3. *I intend to say a lot more than is on notes XX write on slides*
4. *Second part of notes in a couple of lectures time*
5. *will also be a version of the slides for supervisors, with additional notes to explain what I've said in lectures.*
6. *Active learning, not sitting falling asleep*
7. *Additional material in notes: don't need to understand it all*
8. *from the leaders of the RISC movement*
9. *each feature has a cost, quantitative technique to identify benefit and cost*
10. *CPU Info Center / Use Altavista*
11. *Paul De Mone's realworldtech articles are go*
12. *Lively discussion amongst knowledgeable people with the usual interjections by morons*

Further Reading and Reference

- M Johnson
  Superscalar microprocessor design
  1991 Prentice-Hall
- P Markstein
  IA-64 and Elementary Functions
  2000 Prentice-Hall
- A Tannenbaum,
  Structured Computer Organization (2nd ed.)
  1990 Prentice-Hall
- A Someren & C Atack,
  The ARM RISC Chip,
  1994 Addison-Wesley
- R Sites,
  Alpha Architecture Reference Manual,
  1992 Digital Press
- G Kane & J Heinrich,
  MIPS RISC Architecture
  1992 Prentice-Hall
- H Messmer,
  The Indispensable Pentium Book,
  1995 Addison-Wesley
- Gerry Kane and HP,
  The PA-RISC 2.0 Architecture book,
  Prentice Hall

1. *Course uses a range of architectures including x86,ARM,Alpha,MIPS,PowerPC,PA-RISC,IA-64,Transmeta and SPARC to illustrate points made*
2. *all ISA ref manuals are on web, available off Berkeley CPU Info Centre*

Course Pre-requisites

- Computer Design (Ib)
  - Some ARM/x86 Assembler
  - Classic RISC pipeline model
  - Load/branch delay slots
  - Cache hierarchies
  - Memory Systems
- Compilers (Ib/II)
  - Code generation
  - Linkage conventions
- Structured Hardware Design
  - Critical paths
  - Memories
- (Concurrent Systems)

1. *Won't need these until next Tuesday, so time to brush up*

The Microprocessor Revolution

- Mainframe / Scalar Supercomputer
  - CPU consists of multiple components
  - performance improving at 20-35% p.a.
  - often ECL or other exotic technology
  - huge I/O and memory bandwidth

- Microprocessors
  - usually a single CMOS part
  - performance improving at 35-50% p.a.
  - enabled through improvements in fabrication technology
  - huge investment
  - physical advantages of smaller size
  - General Purpose Processors
    - desktop / server
    - SMP / Parallel supercomputers
  - Embedded controllers / SoCs
  - DSPs / Graphics Processors

The Microprocessor Revolution

1. *Scalar vs. Vector (SIMD) vs. Parallel (MIMD)*
2. *Many/some applications are unable to benefit from parallel processing, so fast scalar CPUs are still required*
3. *This course focuses on 'General Purpose' microprocessors and memory systems, e.g. as used in workstations*
4. *More I/O pins and specialist packaging techniques allow greater bandwidth for mainframe CPUs*
5. *Microprocessors have benefited over mainframe CPUs from greater financial investment, and from physical limits imposed by the speed of light benefiting smaller devices.*
6. *CMOS typically offers reduced time to market (more mature tools)*
7. *High performance embedded e.g. AMD 29K, ARM, Intel i960*
8. *Multiple embedded controllers in everything these days. 20+ in the average family car.*
9. *look at some of the technology advances that drive microprocessors*
10. *This course focuses on General Purpose CPUs and memory systems (rather than DSPs, embedded etc)*
11. *SoC : System on a Chip. Complete (simple) CPU (e.g. ARM) integrated with memory controller, peripherals and possibly even DRAM. (it is now possible to fabricate DRAM and logic on same die).*

Developments in CMOS

- Fabrication line size reduction
  - 0.8μ, 0.5, 0.35, 0.25, 0.18, 0.15, *0.13*, *0.09*
  - 10-20% reduction p.a.
  - switching delay reduces with line size

    → increases in clock speed
      * Pentium 66Mhz @ 0.8μ, 150Mhz @ 0.6μ, 233MHz @ 0.35μ
  - density increases at square of 1/line size
- Die size increases at 10-29% p.a.

⇒ Transistor count increase at 55% p.a.
  - enables architectural jumps
  - 8, 16, 32, 64, *128* bit ALUs
  - large caches
    * PA-8500: 1.5MB on-chip
  - new functional units (e.g. multiplier)
  - duplicated functional units (multi-issue)
  - whole System On a Chip (SoC)

Developments in Technology

1. *See the Semiconductor Industry Association Technology Roadmap at the back of booklet. Historically, the predictions made by this document have been surpassed by the industry!*
2. *Disclaimer: past performance is not necessarily a good indicator to the the future....*
3. *(the value of stocks and shares can go up as well as down)*
4. *'drawn line size' is the size of smallest drawn feature*
5. *drawn as opposed to effective (due to depletion) 25% difference*
6. *for comparison, wavelength of visible light is 0.6μ*
7. *people have repeatedly predicted an end to line size reduction for years, but the technology just keeps improving... Quantum effects bellow 0.1μ may put an end to CMOS's progress. Will another technology be ready to take over over?*
8. *first 0.35 Pentium parts were 133MHz, process tweaking, 'stepping', and speed-grade binning has enabled 233. Pentium end of lined at 266MHz*
9. *'stepping' is the name for creating new processor variants with minor changes e.g. to fix bugs or optimize a critical path.*
10. *continual performance leapfrogging of each of the different manufacturer's fab technology*
11. *But, wire propagation delay INCREASES with reduced line size. This is becoming increasingly a problem. IBM have developed a Cu (as opposed to Al) interconnect. 3D construction would be a better long-term solution.*
12. *PA-8500 will have greater than 150 million transistors–by far the vast majority of them will be in the huge 1.5MB L1 caches.*
13. *See the 'General Processor Information' sheet at back of booklet.*
14. *SoC means few components in a design, hence smaller, cheaper, lower power consumption.*

Developments in DRAM Technology

- DRAM density
  - increases at 40-60% p.a.
  - equivalent to 0.5-1 address bits p.a.
  - cost dropping at same rate
    - ∗ 16M, 64M, *256M*, *1G*

- Consequences for processor architectures:

→ May not be able to address whole of memory from a single pointer

  - segmentation

→ May run out of physical address bits

  - banked (windowed) memory

- DRAM performance
  - just 35% latency improvement in 10 years!
  - new bus interfaces make more *sequential b/w* available
    - ∗ SDRAM, RAMBUS, DDR, DDR2

Developments in DRAM Technology

1. *Consume DRAM at same rate as development. Mainly due to software bloat.*
2. *Always pay \$200 for the 'standard' amount of ram in a PC. Now get over 256MB*
3. *64Mb chips currently most common, but 256M chips ramping up. 1G chips working in the lab. SIA Roadmap says 256G in lab by 2012*
4. *High-end desktops are 256MB. Days of 32bit CPUS numbered? (PPro supports 36bit physical addresses, but only a 32bit flat addressing model)*
5. *segmentation: can't address all of memory in one go. Horrid.*
6. *banked memory: Switch between several banks of memory via some out-of-band signal. E.g. PC VGA cards/80286 Expanded Memory. Very, very horrid.*
7. *Fundamental density/latency tradeoff*
8. *New high clock rate synchronous interfaces (e.g. SDRAM,RAMBUS) enable much higher bandwidth to data in the same DRAM ROW.*
9. *SRAM is lower latency, but much more expensive (6T vs 1). Hence the need for caching.*
10. *CPU - Memory gap widening, becoming more of a performance issue... (iRAM etc.)*

μprocessor Development Cycle

- Fabrication technology has huge influence on power and performance

→ must use the latest fabrication process

- Full custom design *vs.* semi custom
- Keep development cycle short (3-4 years)
  - Non CMOS technology leads to complications
- Advance teams to research:
  - process characteristics
  - key circuit elements
  - packaging
  - floor plan
  - required performance
  - microarchitecture
  - investigate key problems
- Hope ISA features don't prove to be a handicap
- Keep up or die!
- Alpha architects planned for 1000x performance improvement over 25 years

μprocessor Development Cycle

1. *The performance advantage of using ECL or GaAs is often squandered through extended time-to-market.*
2. *CPUs must be fabricated on a modern process to be competitive.*
3. *Intel/IBM beginning to get ahead of the competition on fab process introduction. (Having wads of cash to through at the problem helps...)*
4. *At the start of development, CPUs are being designed for a target manufacturing process that doesn't yet exist.*
5. *Most manufacturers aim to produce a new processor 'core' every 3 years (same 'P6' core in Pentium Pro, Pentium II and Pentium III). Pentium IV uses a new core. The core gets 'shrunk' for new fab processes, and may be tinkered with to optimise paths, add new instructions, or re-balance the cache hierarchy (usually increasing sizes of on-chip caches as die space becomes available).*
6. *Costs for each generation of chip are rising almost exponentially. New fab plant costs almost 5 billion dollars.*
7. *Ideally, architectures should be designed such that they contain no features that will hamper the ability of future implementations to make use of techniques enabled by developments in fabrication technology. This requires incredible foresight.*
8. *The Alpha architects, Sites and Witek, planned a 25 year life for Alpha, over which they expect to see a >1000x increase in performance. (10x on multiple issue, 10-100x on clock, 1-10x on thread level parallelism)*
9. *Many other architectures were not planned with so much forethought. They were optimal for the implementation at the time.*
10. *look at some of the reasons for choosing a given processor for a new product*

Power Consumption

- Important for laptops, PDAs, mobile phones, set-top boxes, etc.
- 155W for Digital Alpha 21364 @ 1150MHz
- 130W for Itanium-2 @ 1500MHz
- 90W for AMD Opteron 148 @ 2GHz
- 81W for Pentium-IV @ 3GHz
- 12W for Intel Mobile Pentium M @ 1100Hz
- 420mW for Digital StrongArm @ 233MHz, 2.0V
- 130mW for Digital StrongArm @ 100MHz, 1.65V
- Smaller line size results in lower power
  - lower core voltage, reduced capacitance
  - greater integration avoids inter-chip signalling
- Reduce clock speed to scale power
  - $P = CV^2f$
  - may allow lower voltage
    * potential for cubic scaling
    * better than periodic HALTing

Performance per Watt

Power Consumption

1. *increasingly important in some markets.*
2. *batteries wouldn't last too long in a palm top with a 21264!*
3. *Fan noise from set-top boxes would be a problem*
4. *Can dissipate up to about 100W in a desktop machine using heat pipes etc.*
5. *SIA Roadmap predicts 175W processors in 2012 (running on 0.6-0.9V)*
6. *Do need significant compute power in PDAs: handwriting recognition*
7. *Both SA and 21164 fabricated in Digital's 0.35μm process. One designed for all out performance, the other for low power.*
8. *'mobile processors' are often just 'under-clocked' desktop versions, possibly with L2 cache integrated on die to save inter-chip signalling power.*
9. *Sleep modes and clock frequency scaling to reduce power. (Some fully static components enable clock to be stopped)*
10. *Lowering supply voltage reduces power consumption but requires reduction of the operating clock speed. Power is proportional to V**2, so can be very beneficial. This is what Transmeta do when there isn't enough work to warrant clocking the CPU at max.*
11. *Potential for cubic scaling, whereby a 10% reduction in clock frequency can produce a 27% reduction in power.*
12. *Need OS (or other s/w) to set CPU speed according to prevailing load. More effective than just halting CPU - most instruction sets include an instruction that halts the processor until it is awoken by interrupts. OS sets the timer to wake the CPU up in so many milliseconds unless other IO activity wakes it beforehand.*

Cost and Price

- E.g.:
  - \$0.50: 8bit micro controller
  - \$3: XScale (ARM) (400MHz, 0.18μm, 20mm^2, 2.1M[1M])
  - \$500: Pentium IV Celeron (1.2GHz, 0.13μm, 131mm^2, 28M[4M])
  - \$150: Pentium IV (3.2GHz, 0.09μm, 180mm^2, 42M[7M])
  - \$2200: Itanium2 (1Ghz, 0.18μm, 421mm^2, 221M[15M])
- Costs influenced by die size, packaging, testing
- Large influence by manufacturing volume
- Costs reduce over product life (e.g. 40% p.a.)
  - Yield improves
  - Speed grade binning
  - Fab 'shrinks' and 'steppings'

1. *logic vs. cache density. Figure in square brackets is number of core transistors.*
2. *Number of pins, packaging costs, memory bandwidth. Current BGA (solder Ball Grid Array) packages allow 600+ pins. 3000 pins by 2012...*
3. *Need to recover R&D costs, and Fab Plant construction.*
4. *Design of a new CPU costs >100M. >150 people working for 3 years*
5. *Component prices aren't static, they reduce over product life, as much as factor of 10*
6. *Speed grade binning and marketing influence. Over clocking*
7. *'Stepping': making minor changes to a design/layout to fix bugs and improve yield of higher frequency parts.*
8. *'Shrink': Re-laying out a design for a new fab process.*

Compatibility

- 'Pin' Compatibility (second sourcing)
- Backwards Binary Compatibility
  - 8086, 80286, 80386, 80486, Pentium, Pentium Pro, Pentium II/III/IV, *Itanium*
  - NexGen, Cyrix, AMD, Transmeta
  - typically need to re-optimize
- Typically hard to change architecture
  - Users have huge investment in s/w
  - Binary translators e.g. FX!32, WABI
    - typically interface to native OS
  - Need co-operation from s/w vendors
    - multi-platform support costs $'s
  - Most computer sales are upgrades
- Platform independence initiatives
  - Source, p-Code, JAVA bytecode, .NET

Compatibility is very important

Compatibility

1. *pin compatible: use same motherboard. (Though Intel have patented slot-1)*
2. *AMD K-7 uses 'Slot A', the first x86 clone NOT to be pin compatible with Intel.*
3. *Need to persuade s/w vendors to optimise their products for CPU. e.g. the Pentium IV has an entirely new set of FP instructions. Software flies if you use them, but performs horribly if you use the old 8087 instructions.*
4. *WABI is an x86 emulator available for a number of workstation platforms. FX!32 is Digital's x86 to Alpha binary translator.*
5. *binary2binary translators: hard to get software vendor to support translated product.*
6. *Transmeta's approach is to translate ALL software (including the OS) rather than just the applications. This makes sense because the x86 interface is MUCH simpler and better specified than the Windows API. Products such as Wine and FX!32 might theoretically offer better performance through native implementation of OS and API code, but keeping track of MS developments of the API is very difficult...*
7. *binary2binary optimisers: re-write old binary to produce an optimised one. Used by Digital to optimise Alpha NT versions of MS products that were compiled with a lousy compiler.*
8. *Changing arch e.g. Sun/HP/Apple from 68k, DEC from VAX and MIPS. Often loose customers.*
9. *Will we see Java bytecode / .NET versions of MSWord? How will they perform? I believe they could actually perform better, after some good virtual machine research...*

Performance Measurement

- Try before you buy! (often not possible)
- System may not even exist yet
  - use cycle-level simulation
- Real workloads often hard to characterize and measure improvements
  - especially interactive
- Marketing hype
  - MHz, MIPS, MFLOPS
- Algorithm kernels
  - Livermore Loops, Linpack
- Synthetic benchmarks
  - Dhrystones, Whetstones, iCOMP
- Benchmark suites
  - SPEC-INT, SPEC-FP, SPEC-HPC, NAS
- Application Benchmarks
  - TPC-C/H/R, SPECNFS, SPECWeb, Quake

Performance is application dependent

Performance Measurement

1. *Performance is often not particularly critical*
2. *Often not possible to try application: live system or no optimized binary available*
3. *desktop workloads are especially difficult: highly varied, idle much of the time, response time is important*
4. *cycle level simulators exist for CPUs long before they're actually fabbed. Need a good memory system model too.*
5. *simulate the 'kernel' (inner most loops) of important applications*
6. *fallacies of using MHz: 450MHz Alpha 21164 = 180MHz HP PA-8000*
7. *fallacies of using MIPS: Rarely able to occupy all functional units to achieve peak MIPS performance. PA-8000 better at doing this than Alpha 21164*
8. *livermore loops: good if that's your application, won't tell you much about anything else. (Livermore loops are an FP intensive algorithm that forms part of Gaussian elimination)*
9. *synthetic benchmarks: derived from instruction traces of real apps. They are usually too small, don't reward real compiler optimizations, and allow unrealistic ones.*
10. *eg. dhystone: 25%discarded by optimizing compiler. Function inlining had to be forbidden to enable 'fair' comparison.*
11. *whetstone: trick optimization sqrt(exp(x)) = exp(x/2)*
12. *NAS is a benchmark suite for HPC applications. It's interesting in that its a set of algorithms that you can implement however you like. (With SPEC-HPC you can't change the Fortran source) NAS is more realistic to how Supercomputers are actually used.*
13. *Application performance is not dependent on just the CPU performance, but on the compiler and memory system too.*
14. *Application performance may also depend on IO e.g. graphics, disk, network performance*
15. *TPC-C for transaction processing systems, TPC-D for 'decision support' database mining.*

Standard Performance Evaluation Corporation

- SPEC is most widely used benchmark
  - processor manufactures
  - workstation vendors
- CPU INT / FP 89, 92, 95, 2000, (2004)
- Suite updated to reflect current workloads
- CINT95/2K: 8/12 integer C programs
- CFP95/2K: 10/14 floating point in C&Fortran
- measures:
  - processor
  - memory system
  - compiler
  - NOT OS, libc, disk, graphics, network

1. *SPEC so important that it used to provide input into CPU design*
2. *SPEC-98 due to ship in '99.*
3. *execution time in non SPEC supplied code is minimal (i.e. not OS or libc)*
4. *disk,video,network performance can be very important—the CPU is often not the bottleneck.*
5. *Video performance has an important effect on how fast a system 'feels'.*

Choosing programs for SPEC2000

- More programs than SPEC95
- Bigger programs than SPEC95
  - Don't fit in on-chip caches
- Reflect some real workloads
- Run for several minutes
  - Amortize startup overhead & timing inaccuracies
- Not susceptible to trick transformations
  - Vendors invest huge s/w effort
- Fit in 256MB (95 was 64MB)
- Moving target...
- SPEC92, 95, 2K results not translatable

1. *SPEC2000 very new – only a few results available*
2. *5K dollar bounty for submitting applications*
3. *one of the SPEC92 programs was susceptible to a trick transformation*
4. *huge amount of effort goes into SPEC benchmarks by vendors*
5. *compiler optimization - some may even be useful for real progs!*
6. *I&D trace data, processors designed to execute SPEC*
7. *multiple submissions for same machine as compilers improve*
8. *Fit in 64MB: Becoming too small to really stretch modern workstation memory systems.*
9. *no simple relationship between SPEC 92 and 95 numbers - 95 more demanding on memory systems.*

CINT95 suite (C)

099.go	An AI go-playing program
124.m88ksim	A chip simulator for the Motorola 88100
126.gcc	Based on the GNU C compiler version 2.5.3
129.compress	An in-memory version of the utility
130.li	Xlisp interpreter
132.ijpeg	De/compression on in-memory images
134.perl	An interpreter for the Perl language
147.vortex	An object oriented database

CFP95 suite (Fortran)

101.tomcatv	Vectorized mesh generation
102.swim	Shallow water equations
103.su2cor	Monte-Carlo method
104.hydro2d	Navier Stokes equations
107.mgrid	3d potential field
110.applu	Partial differential equations
125.turb3d	Turbulence modelling
141.apsi	Weather prediction
145.fpppp	Quantum chemistry
146.wave5	Maxwell's equations

CINT95 suite (C)

1. *Considerable effort goes into selecting programs for SPEC. They are all (modified) real applications; Applications that someone, somewhere cares about the performance of.*
2. *126.gcc cross compiles number of programs for SUN, not native*
3. *Jpeg and Gcc have large data sets, which do exercise filesystem I/O (eg. 13% worse over NFS, 5% better with RAM disk)*
4. *Which of these apps represent my workload? What will SPEC say about the performance of emacs, Netscape, MS Word, Quake?*

SPEC reporting

- Time each program to run
- Reproduceability is paramount
  - Take mean of $\geq$ 3 runs
  - Full disclosure
- Baseline measurements
  - SPECint_base95
  - Same compiler optimizations for whole suite
- Peak measurements
  - SPECint95
  - Each benchmark individually tweaked
  - Unsafe optimizations can be enabled!
- Rate measurements for multiprocessors
  - SPECint_rate95, SPECfp_rate95
  - time for N copies to complete x N

SPEC reporting

1. *Take median of string of results*
2. *full disclosure: OS and compiler versions, system configuration*
3. *compiler and system must be for sale to be official*
4. *peak results: great long lists of cryptic compiler flags*
5. *base results: more likely what a real user would get*
6. *unsafe optimizations: but, must be correct for any input data set*
7. *→ Intel accidently cheated on one of the fp benchmarks, producing code that wouldn't work for all possible datasets.*
8. *SPECrate: copies run*reference*seconds in day / elapsed time=rate in jobs/day*
9. *SPECrate: important for batch processing systems*

Totalling Results

- How to present results?
  - Present individual results?
  - Arithmetic mean?
  - Weighted harmonic mean?
  - SPEC uses Geometric mean, normalised against a reference platform
    - allows normalization before or after mean
    - performance ratio can be predicted by dividing means
- SPEC95 uses Sun SS10/40 as reference platform

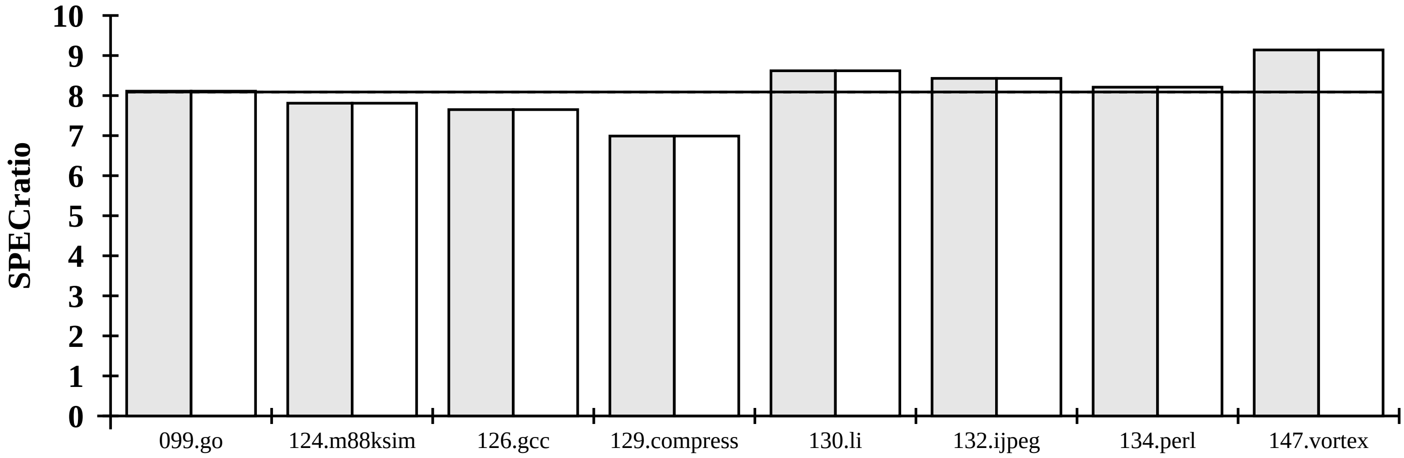
SPEC CINT95 Results

Intel Corporation
Alder System (200MHz, 256KB L2)

SPECint95 = 8.09

SPECint_base95 = 8.09

SPEC license #	14	Tested By:	Intel	Test Date:	Oct-95	Hardware Avail:	May-96	Software Avail:	Feb-96



Hardware/Software Configuration for:
Alder System (200MHz, 256KB L2)

Hardware

Model Name:	Alder
CPU:	200MHz Pentium Pro Processor
FPU:	Integrated
Number of CPU(s):	1
Primary Cache:	8KBI+8KBD
Secondary Cache:	256KB(I+D)
Other Cache:	None
Memory:	128MB (60ns fast page)
Disk Subsystem:	2GB ST32550W
Other Hardware:	AHA-2940W Controller

Software

Operating System:	UnixWare 2.0, SDK
Compiler:	Intel C Reference Compiler 2.2 Beta
File System:	ufs, vxfs (/tmp as 8MB /tmpfs)
System State:	Single user (root + killall)

Benchmark # and Name	Reference Time	Base Run Time	Base SPEC Ratio	Run Time	SPEC Ratio
099.go	4600	567	8.11	567	8.11
124.m88ksim	1900	243	7.81	243	7.81
126.gcc	1700	222	7.65	222	7.65
129.compress	1800	258	6.99	258	6.99
130.li	1900	220	8.62	220	8.62
132.ijpeg	2400	285	8.43	285	8.43
134.perl	1900	232	8.21	232	8.21
147.vortex	2700	295	9.14	295	9.14
SPECint_base95 (G. Mean)			8.09		
		SPECint95 (G. Mean)			8.09

Notes/Tuning Information

```
Base and non-base flags are the same and use Feedback Directed Optimization
Pass1: -tp p6 -ipo -xi -prof_gen -ircdb_dir /tmp/IRCDB
Pass2: -tp p6 -ipo -xi -prof_use -ircdb_dir /tmp/IRCDB
-ircdb_dir is a location flag and not an optimization flag
Portability: 124: -DSYSV -DLEHOST 130, 134, 147: -lm  132: -DSYSV  126: -lm -lc -L/usr/ucblib -lucb -lmalloc
Memory subsystem is four-way interleaved.
```

For More Information Contact:
SPEC
10754 Ambassador Drive, Suite 201
Manassas, VA 22110
(703) 331-0180
info@specbench.org
http://www.specbench.org

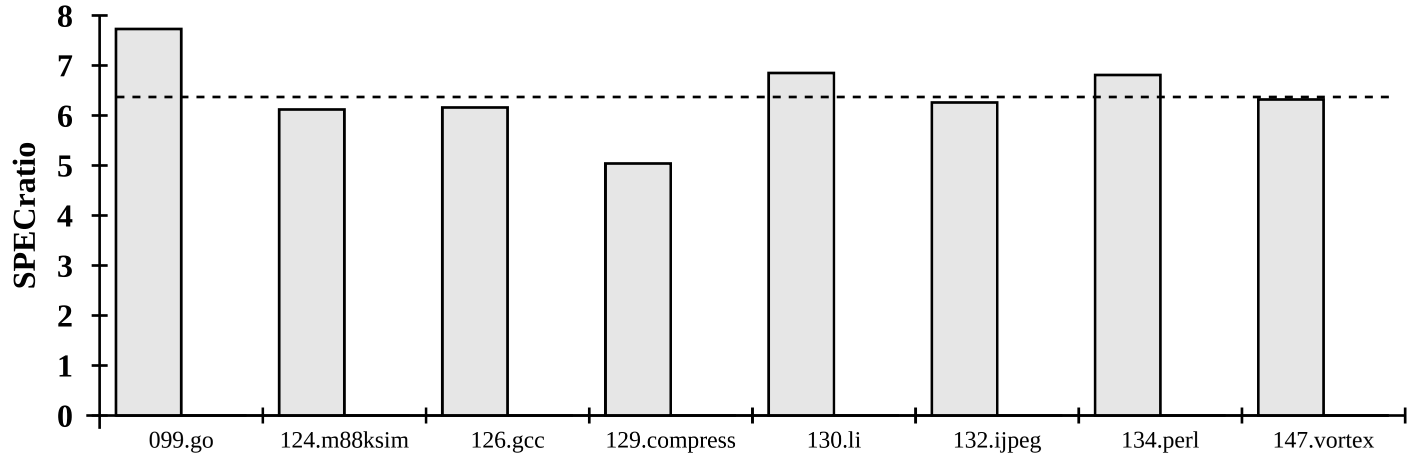
1. *full disclosure of machine / compiler configuration*
2. *compiler options used*
3. *Feedback Directed Optimization. Optimization pass uses the SPEC95 Training input set - not allowed to use the Reference input set for training.*
4. *Base SPEC ratio = Ref time / Base run time*
5. *SPECint_base = (r1*r2*r3...r8) ** 1/8*

SPEC CINT95 Results

Intel 440LX motherboard Pentium Pro 200

SPECint95	=	--
SPECint_base95	=	6.37

SPEC license # 1178	Tested By: Ian Pratt, CUCL	Test Date: Date	Hardware Avail: Date	Software Avail: Date



Hardware/Software Configuration for:
Pentium Pro 200

Hardware

Model Name: Intel 440LX
CPU: Pentium Pro 200
FPU:
Number of CPU(s): 1
Primary Cache: 8KB+8KB
Secondary Cache: 256KB
Other Cache:
Memory: 128MB
Disk Subsystem: 4GB
Other Hardware:

Software

Operating System: Linux 20.0.30
Compiler: gcc 2.7.2p
File System: ext2
System State: multiuser

Benchmark # and Name	Reference Time	Base Run Time	Base SPEC Ratio	Run Time	SPEC Ratio
099.go	4600	595	7.73	--	--
124.m88ksim	1900	310	6.12	--	--
126.gcc	1700	276	6.16	--	--
129.compress	1800	357	5.04	--	--
130.li	1900	277	6.85	--	--
132.ijpeg	2400	384	6.26	--	--
134.perl	1900	279	6.81	--	--
147.vortex	2700	427	6.32	--	--
SPECint_base95 (G. Mean)			6.37		
		SPECint95 (G. Mean)			--

Notes/Tuning Information

```
Portability flags were:
Baseline flags were: -O2 -fomit-frame-pointer
Nonbase flags were:
```

For More Information contact
SPEC c/o NCGA
2722 Merrilee Drive, Suite 200
Fairfax, VA 22031
(703) 698-9604 ext 318
spec-ncga@cup.portal.com

--
Prepared By: --

1. *Compiler does make a difference: gcc 2.7.2p 20% slower than the Intel reference compiler shipped with x86 Solaris*

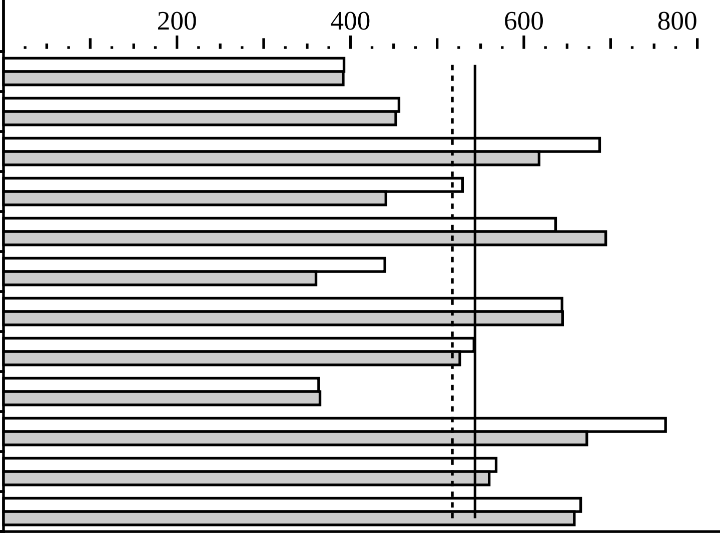
CINT2000 Result

Compaq Computer Corporation
AlphaServer ES40 Model 6/833

SPECint2000 = 544
SPECint_base2000 = 518

SPEC license #: 2	Tested by: Compaq NH	Test date: Oct-2000	Hardware Avail: Jan-2001	Software Avail: Nov-2000

Benchmark	Reference Time	Base Runtime	Base Ratio	Runtime	Ratio
164.gzip	1400	358	392	357	393
175.vpr	1400	309	452	307	456
176.gcc	1100	178	617	160	687
181.mcf	1800	408	441	340	529
186.crafty	1000	144	694	157	637
197.parser	1800	500	360	409	440
252.eon	1300	202	645	202	644
253.perlbmk	1800	342	526	332	543
254.gap	1100	301	365	303	363
255.vortex	1900	282	673	249	763
256.bzip2	1500	268	560	264	568
300.twolf	3000	456	658	451	666



Hardware

CPU:	Alpha 21264B
CPU MHz:	833
FPU:	Integrated
CPU(s) enabled:	1
CPU(s) orderable:	1 to 4
Parallel:	No
Primary Cache:	64KB(I)+64KB(D) on chip
Secondary Cache:	8MB off chip
L3 Cache:	None
Other Cache:	None
Memory:	16GB
Disk Subsystem:	1x8GB BD0096349A
Other Hardware:	Ethernet

Software

Operating System:	Tru64 UNIX V5.1 + Patch Kit 1 libc
Compiler:	Compaq C V6.3-129-44A8I Compaq C++ V6.2-033-4298H
File System:	AdvFS
System State:	Multi-user

Notes/Tuning Information

```
Baseline C  : cc  -arch ev6 -fast GEMFB ONESTEP
         C++: cxx -arch ev6 -O2 ONESTEP

GEMFB: fdo_pre0     = mkdir /tmp/pb; rm -f /tmp/pb/${baseexe}*
       PASS1_CFLAGS = -prof_gen_noopt    -prof_dir /tmp/pb
       PASS2_CFLAGS = -prof_use_feedback -prof_dir /tmp/pb
       (base uses directory /tmp/pb; peak uses /tmp/pp)

SPIKEFB: fdo_post2  = spike -feedback ${baseexe} -o tmp ${baseexe};
                      mv tmp ${baseexe}
Peak: cc [except eon: cxx] -arch ev6 ONESTEP plus:
     164.gzip: -g3 -fast -O4 +GEMFB
      175.vpr: -g3 -fast -O4 +GEMFB
      176.gcc: -g3 -fast -O4 -xtaso_short +GEMFB
      181.mcf: -g3 -fast -xtaso_short +GEMFB
   186.crafty: -g3 -fast -O4 -inline speed
   197.parser: -g3 -fast -O4 -xtaso_short +GEMFB
      252.eon: -O2
  253.perlbmk: -g3 -fast +GEMFB +SPIKEFB
      254.gap: -g3 -fast -O4 +GEMFB
```

1. *Peak results can be significantly higher than base. Note the huge list of dodgy compiler options.*

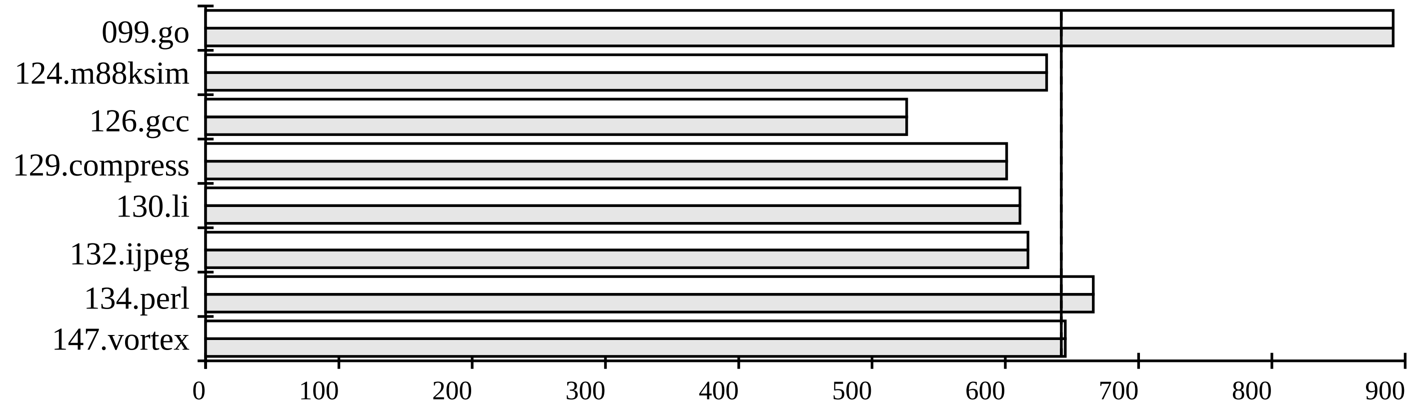
SPEC CINT95rate Results

Digital Equipment Corp.
AlphaServer 8400 5/300

SPECint_rate95 = 642

SPECint_rate_base95 = 642

SPEC license # 2	Tested By: Digital PKO	Test Date: Oct-95	Hardware Avail: Apr-95	Software Avail: Aug-95



Hardware/Software Configuration for:
AlphaServer 8400 5/300

Hardware

Model Name: AlphaServer 8400 5/300
CPU: 300 MHz 21164
FPU: Integrated
Number of CPU(s): 10
Primary Cache: 8KBI+8KBD on chip
Secondary Cache: 4MB
Other Cache: none
Memory: 1GB
Disk Subsystem: 1 x 2GB
1 x 2GB
Other Hardware: Ethernet

Software

Operating System: Digital UNIX V3.2C (Rev 148)
Compiler: DEC C V5.0-106
File System: UFS
System State: Multi User

Benchmark # and Name	Base Copies	Base Run Time	Base SPEC Ratio	Copies	Run Time	SPEC Ratio
099.go	10	464	891	10	464	891
124.m88ksim	10	271	631	10	271	631
126.gcc	10	291	526	10	291	526
129.compress	10	270	601	10	270	601
130.li	10	280	611	10	280	611
132.ijpeg	10	350	617	10	350	617
134.perl	10	257	666	10	257	666
147.vortex	10	377	645	10	377	645
SPECint_rate_base95 (G. Mean)			642			
		SPECint_rate95 (G. Mean)				642

Notes/Tuning Information

```
Baseline Optimizations: -O5 -ifo -non_shared -om
Portibility Flags: 124.m88ksim: -DLEHOST  134.perl: -DI_TIME
147.vortex: -D__RISC_64__

Compiler invokation: cc -migrate -std1 (DEC C with -std1 for strict ANSI)
```

For More Information Contact:
SPEC
10754 Ambassador Drive, Suite 201
Manassas, VA 22110
(703) 331-0180
info@specbench.org
http://www.specbench.org

1. *Running 10 copies on a 10 CPU machine*
2. *Base SPEC ratio = copies run*reference*seconds in day / elapsed time (=rate in jobs/day)*

Top SPEC2000 Results for each ISA

machine	processor	cpu MHz	cache sizes	int	fp
Intel D875	Pentium IV-X	3200	12*/8+256+2M	1620	1516
AMD/ASUS	Opteron148	2200	64/64+1M	1477	1514
Intel D875	Pentium IV	3200	12*/8+512	1330	1287
HP rx2600	Itanium2	1500	16/16+256+6M	1322	2119
IBM p690	Power4+	1700	64/32+1.5M+(128M)	1113	1699
HP Alpha ES45	21264C	1250	64/64+(16M)	928	1482
Fujitsu	SPARC64-V	1350	128+128/2M	905	1340
Apple	PPC970 (G5)	2000	64/32+512	800	840
HP	Pentium-M	1000	32/32+1024	687	552
HP c3750	PA-8700	875	768/1.5M	678	674
SGI Orgin 3200	R14000	600	32/32+(8M)	500	529
HP rx4610	Itanium	800	16/16+96+(4M)	379	701

SPEC 95 Results

1. *cache sizes L1-I/L1-D + L2 + L3*
2. *First part of table shows current top scores for a range of CPUs and architectures. Notice how some RISC chips are beginning to fall behind...*
3. *although a uniprocessor benchmark, some auto-parallelizing compilers manage to reaps benefits from multiple CPUs*
4. *futility of MHz, even MIPS or FLOPS : compare PA8000 @180 and 21164 @400, or Pentium @200 and PPro @200*
5. *"Speed demons vs brainiacs": PA8000 gets better utilization from its 2 integer pipelines, and executes more expressive instructions than Alpha.*
6. *For a given processor core, SPEC does scale reasonably with MHz, providing Memory b/w is scaled comparably.*
7. *Note lousy x86 FP performance...*

Selected SPEC95 Results

machine	processor	cpu MHz	cache sizes	int_base	fp_base
Sun SS10/40	SuprSP	40	20/16	1.00	1.00
Intel 440BX	Pentium II	300	16/16+(512)	12.2	8.4
Intel 440EX	Celeron A	300	16/16+128	11.3	8.3
Intel 440EX	Celeron	300	16/16	8.3	5.8
Compaq PC164LX	21164	533	8/8+96+(4M)	16.8	20.7
Compaq PC164SX	21164PC	533	16/16+(1M)	12.2	14.1
Intel 440BX	Pentium II	450	16/16+(512)	17.2	11.8
Intel 440BX	Pentium II	400	16/16+(512)	15.8	11.4
Intel 440BX	Pentium II	350	16/16+(512)	13.9	10.2
Intel 440BX	Pentium II	330	16/16+(512)	13.0	8.8
Intel 440BX	Pentium II	300	16/16+(512)	11.9	8.1
Intel 440BX	Pentium II	266	16/16+(512)	10.7	7.5
Intel 440BX	Pentium II	233	16/16+(512)	9.4	6.7
DEC 4100/5/400	A21164	400/75	8/8+96+4M	10.1	16.0
DEC 4100/5/400	2xA21164	400/75	8/8+96+4M	10.1	20.7
DEC 4100/5/400	4xA21164	400/75	8/8+96+4M	10.1	26.6
Intel XXpress	Pentium	200	8/8+1M	5.47	2.92
Intel Alder	PentPro	200	8/8+256	8.09	5.99

SPEC 95 Results

1. *cache sizes L1-I/L1-D + L2 + L3*
2. *First part of table shows current top scores for a range of CPUs and architectures. Notice how some RISC chips are beginning to fall behind...*
3. *although a uniprocessor benchmark, some auto-parallelizing compilers manage to reaps benefits from multiple CPUs*
4. *futility of MHz, even MIPS or FLOPS : compare PA8000 @180 and 21164 @400, or Pentium @200 and PPro @200*
5. *"Speed demons vs brainiacs": PA8000 gets better utilization from its 2 integer pipelines, and executes more expressive instructions than Alpha.*
6. *For a given processor core, SPEC does scale reasonably with MHz, providing Memory b/w is scaled comparably.*
7. *Note lousy x86 FP performance...*

Comparing Implementations Summary

- Fabrication technology has a huge influence
- Exponential improvement in technology
- Processor for a product chosen on:
  - Instruction Set Compatibility
  - Power Consumption
  - Price
  - Performance
- Performance is application dependent
  - Avoid MIPS, MHz
  - Benchmark suites

Instruction Set Architecture

- Processor s/w interface
- Externally visible features
  - Word size
  - Operation sets
  - Register set
  - Operand types
  - Addressing modes
  - Instruction encoding
- Introduction of new ISAs now rare
- ISAs need to last several generations of implementation
- How do you compare ISAs ?
  - yields ‘best’ implementation
    * performance, price, power
    * are other factors equal?
  - ‘aesthetic qualities’
    * ‘nicest’ for systems programmers

Instruction Set Architecture

1. *In the early days, new architectures were introduced on a regular basis. Today, after IA-64 we're unlikely to see the introduction of another main-stream architecture for some time.*
2. *difficult to quantitatively compare the effects of architecture upon implementation: Different implementations have different goals, different amounts of money spent on them, different people etc.*
3. *From time to time you may have to open a processor user manual in order say, to write compiler back end, optimize some code*
4. *ARM and Thumb, PPRO architecture volume, Alpha Arch volume Hold up PPro 3 volumes, daunting, complexity, only describes what it does, not how!*
5. *This section describes hoe to find your way around a processor User Manual, which sections to look at first*
6. *many of you will have only seen the ARM e.g. of what's called a 3 operand, load/store, general purpose register architecture*
7. *Tour of instrs, regs and addressing modes found in common archs*
8. *Look at the RISC decisions*

Instruction Set Architecture

- New implementations normally backwards compatible
  - Should execute old code correctly
  - Possibly some exceptions e.g.
    - Undocumented/unsupported features
    - Self modifying code on 68K
  - May add new features e.g. FP, divide, sqrt, SIMD, FP-SIMD
  - May change execution timings
  - $\rightarrow$ CPU specific optimization
  - Can rarely remove features
    - Unless never used
    - software emulation fast enough
  - $\rightarrow$ Layers of architectural baggage
    - (8086 16bit mode on Pentium IV)
- Architecture affects ease of utilizing new techniques e.g.
  - Pipelining
  - Super-scalar (multi-issue)
- But x86 fights real hard!
  - more T's tolerable unless on critical path

Instruction Set Architecture

1. *68030 separate I and D cache broke self modifying code.*
2. *PPro still has to execute self modifying code correctly - Too many old x86 programs rely on it*
3. *Can't remove features, but it may be acceptable to make them run slowly e.g. 16 bit mode on the PPro - MS assured Intel that the majority of code would be 32bit. At the time of the PPro's launch, this probably wasn't true...*
4. *utilizing new techniques made available by more transistors*
5. *older CISC architectures not ideal for pipelining -¿ more gates and gate delays*
6. *Super-scalar processors attempt to issue multiple instructions per cycle from a single threaded (scalar) program.*
7. *RISC was designed to fully utilize pipelining, but some RISC architectures made decisions which make super-scalar implementation slightly tricky*
8. *x86 (IA-32) wasn't designed to utilize these techniques, but has succeeded in doing so pretty well – just through transistors at the problem.*
9. *Mike Johnson head of AMD K6 design team: The x86 isn't all that complex - it just doesn't make a lot of sense. No show stoppers, just lots of special cases.*

Reduced Instruction Set Computers

- RISC loosely classifies a number of Architectures first appearing in the 80's
- Not really about reducing number of instructions
- Result of quantitative analysis of the usage of existing architectures
  - Many CISC features designed to eliminate the 'semantic gap' were not used
- RISC designed to easily exploit:
  - Pipelining
    - Easier if most instructions take same amount of time
  - Virtual Memory (paging)
    - Avoid tricky exceptional cases
  - Caches
    - Use rest of Si area
- Widespread agreement amongst architects

1. *Can be dangerous grouping architectures into CISC and RISC. Within each classification there are wide differences*
2. *semantic gap: 1970's, loop constructs, list and bit-field operators, procedure calls (these investigated further later on)*
3. *pre RISC, there was wide diversity in architecture. Even very new architectures have firm RISC-like foundations. e.g. IA-64*

Amdahl's Law

- Every 'enhancement' has a cost:
  - Would Si be better used elsewhere?
    - e.g. cache
  - Will it slow down other instructions?
    - e.g. extra gate delays on critical path
    - $\rightarrow$ longer cycle time

- Even if it doesn't slow anything else down, what overall speedup will it give?

- size and delay

$$speedup = \frac{execution\ \ time\ \ for\ \ entire\ \ task\ \ without\ \ using\ \ enhancement}{execution\ \ time\ \ for\ \ entire\ \ task\ \ using\ \ enhancment\ \ \boldsymbol{when}\ \ \boldsymbol{possible}}$$

1. *using enhancement* when *possible*

Amdahl's Law :2

- How frequently can we use enhancement?
  - examine instruction traces e.g. SPEC
  - will code require different optimization?
  - $Fraction_{enhanced}$
- When we can use it, what speedup will it give?
  - $Speedup_{enhanced}$
  - e.g. cycles before/cycles after

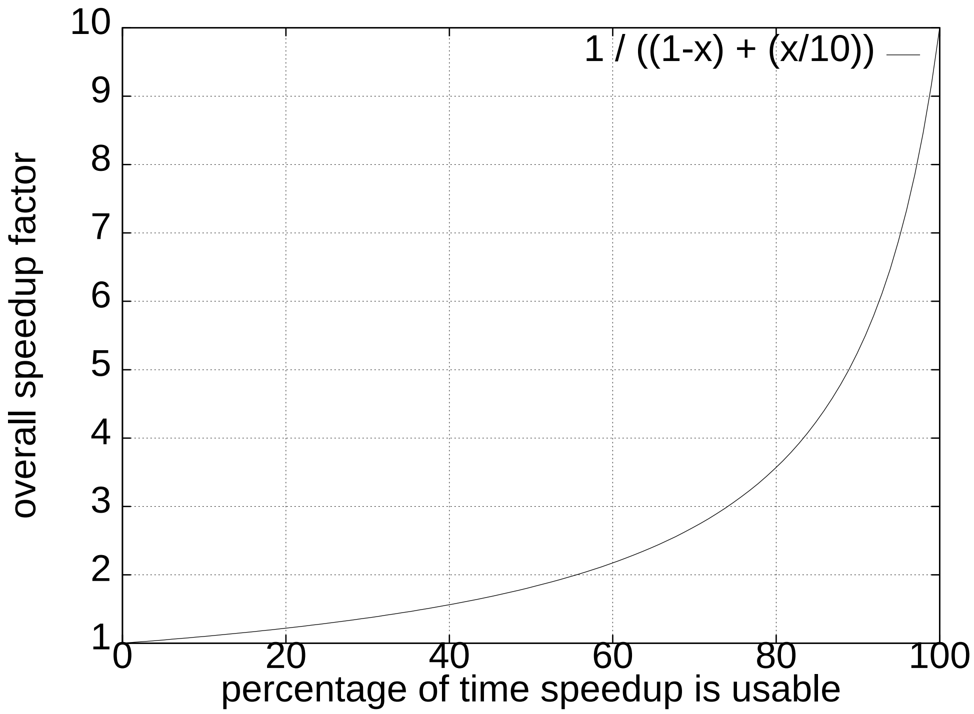
$$Speedup_{overall} = \frac{1}{(1 - Fraction_{enhanced}) + \frac{Fraction_{enhanced}}{Speedup_{enhanced}}}$$

$\rightarrow$ Spend resources where time is spent

Optimize for the common case

1. *Surprising: Speedup = 10, fraction = 0.5, speedup overall = only 1.82 : not even double*

Amdahl's Law for Speedup=10



Amdahl's Law Example

- FPSQRT is responsible for 20% of execution time in a (fictitious) critical benchmark
- FP operations account for 50% of execution time in total
- Proposal A:
  - New FPSQRT hardware with 10x performance

$$speedup_A = \frac{1}{(1-0.2)+\frac{0.2}{10}} = \frac{1}{0.82} = 1.22$$

- Proposal B:
  - Use Si area to double speed all FP operations

$$speedup_B = \frac{1}{(1-0.5)+\frac{0.5}{2}} = \frac{1}{0.75} = 1.33$$

- $\rightarrow$ Proposal B is better
- (Probably much better for other users)

Word Size

- Native size of an integer register
  - 32bits on ARM, MIPS II, x86 32bit mode
  - 64bits on Alpha, MIPS III, SPARC v8, PA-RISC v2
- NOT size of FP or SIMD registers
  - 64 / 128 bit on Pentium III
- NOT internal data-path width
  - 64bit internal paths in Pentium III
- NOT external data-bus width
  - 8bit Motorola 68008
  - 128bit Alpha 21164
- NOT size of an instruction
  - Alpha, MIPS, etc instructions 32bit
- But, 'word' also used as a type size
  - 4 bytes on ARM, MIPS
  - 2 bytes on Alpha, x86
    - longword = 4 bytes, quadword = 8

64bit *vs* 32bit words

- Alpha, MIPS III, SPARC v8, PA-RISC v2

✔ Access to a large region of address space from a single pointer
  - large data-structures
  - memory mapped files
  - persistent objects

✔ Overflow rarely a concern
  - require fewer instructions

✘ Can double a program's data size
  - need bigger caches, more memory b/w

✘ May slow the CPU's max clock speed

- Some programs gain considerably from 64bit, others get no benefit.

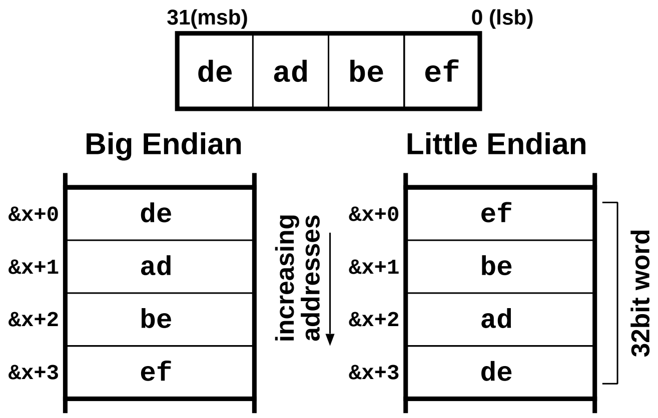
- Some OS's and compilers provide support for 32bit binaries

1. *64 bits is over 10 million terabytes (1.8E19)*
2. *There now exists true 64bit versions of a number of architectures.*
3. *Alpha has always been 64bit*
4. *32bit support: Versions of libc with both 32 and 64 bit APIs. Pointers are 32bits long.*
5. *See article on PA-RISC 64bit extensions*

Byte Sex

- Little Endian camp
  - Intel, Digital
- Big Endian camp
  - Motorola, HP, IBM
  - Sun: ‘Network Endian’, JAVA
- Bi-Endian Processors
  - Fixed by motherboard design
  - MIPS, ARM
- Endian swapping instructions

```
int   x= 0xdeadbeef;
char *p= (char*)&x;
if(*p == 0xde) printf("Big Endian");
if(*p == 0xef) printf("Little Ebdian");
```



Byte Sex

1. *There is no right or wrong...*
2. *PDP-11's used a strange combination of big and little endian*
3. *On a little-E machine, reading a byte from a word ptr gives the least significant byte. little-least.*
4. *personal view is that little is easier to wrap my head around*
5. *bit and byte endian. IBM/PowerPC use weird big bit-endian in manuals*
6. *Only programs that access the same value as both bytes and words care, or pass data to other processors.*
7. *MIPS: little-E on Digital Ultrix, big-E on SGI IRIX*
8. *processor XORs address bits A0 and A1 based on endian-mode. Motherboard data bus must also be wired accordingly. (endian mode cannot be changed on-the-fly)*
9. *NT currently only runs on Little Endian machines, and rumour has it that it is deeply ingrained within it, but JAVA...*
10. *Machines that only access data as words do not have an endian per se. e.g. Cray*
11. *Some old machines had 9bit bytes. Should really call a byte an octet...*
12. *instructions that swap the endian of a word have been added to several architectures – particluar for little endian machines to enable them to translate in and out of 'Network endian' (which is big)*

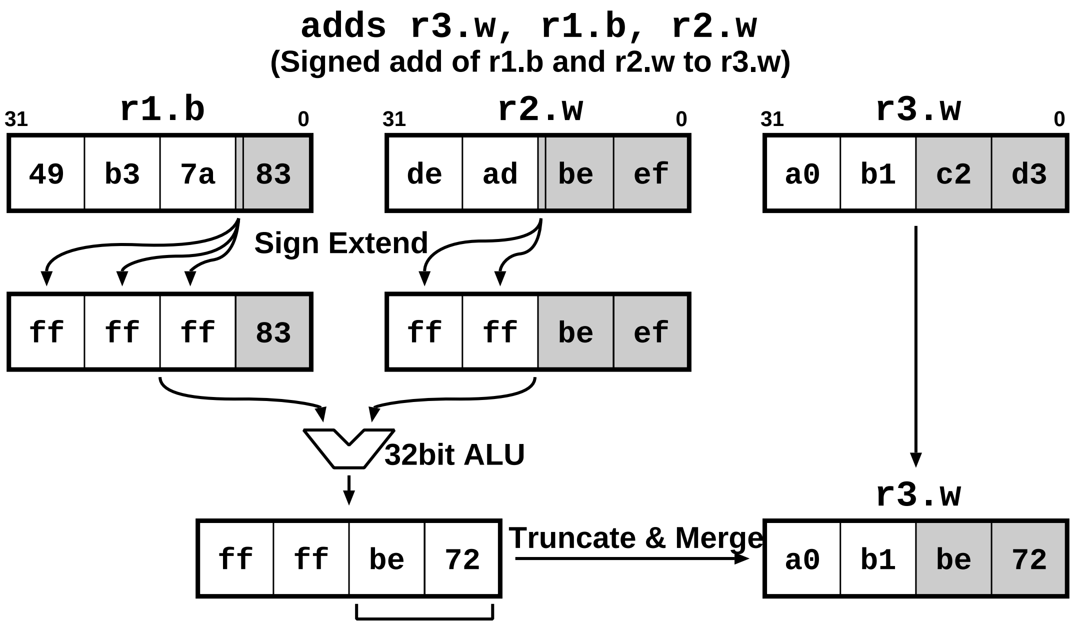
Data Processing Instructions

- 2's Complement Arithmetic
  - add, subtract, multiply, compare, multiply
  - some: divide, modulus
- Logical
  - and, or, not, xor, bic, ...
- Shift
  - shift left, logical shift right, arithmetic shift right
  - some: rotate left, rotate right

1. *add,sub: same ALU operation for signed and unsigned values*
2. *compare is a subtract in which throws result away*
3. *add overflow c: ignore, Fortran: need to know about it : exception vs test*
4. *divide: example of instruction that got added in Alpha*
5. *bic = and not*
6. *barrel shifters vs ALU (1 bit at a time shifts on old CISC e.g. 6502)*
7. *no rotate in C, but compilers can still synthesize it*
8. *rotates not in some RISC, 2 shifts and OR*
9. *arithmetic shift right replicates msb.*

Operand Size

- CISC
  - 8,16,32 bit operations
  - zero/sign extend sources
    - need unsigned/signed instrs
  - merge result into destination
  - some even allow mixed size operands



- RISC
  - Word size operations only
  - (except 64bit CPUs often support 32bit ops)
  - Pad `char` and `short` to word

Operand Size

1. *extension and merging can be a right pain on the critical path.*
2. *When doing truncation (part of merge), need to check for overflow*
3. *VAX and x86 allow operands of different sizes to be used in the same instruction.*
4. *Try to make Word size operations the common case*
5. *Compiler pads chars and shorts into words. (Can't be done for arrays).*
6. *Excluding 32bit ops from 64bit machines would result in too much performance loss on current code. But, Alpha always produces a 64bit result, avoiding merging.*

(Zero/Sign Extension)

- Unsigned values: zero extend
  - e.g. 8bit values to 32bit values

```
unsigned char a; int b;
 and  b ← a, #0xff
```

- Signed values: sign extend
  - e.g. 8bit values to 32bit values
  - Replicate sign bit

```
char a; int b;
 lsl  b ← a, #24
 asr  b ← b, #24
```

- C: 32bit to 8bit
  - Just truncate

```
 and  b ← a, #0xff
```

1. *All RISCs perform operations on 32 bit quantities: CISC*
2. *Languages other than C: Check for overflow*

CISC instructions RISC dropped

- Emulated in RISC:

```
move   r1 ← r2  e.g. or   r1 ← r2, r2
zero   r1       e.g. xor  r1 ← r1, r1
neg    r1       e.g. sub  r1 ← #0, r1
nop             e.g. or   r1 ← r1, r1
sextb  r1 ← r2  e.g. lsl  r1 ← r2, #24;
                     asr  r1 ← r1, #24
```

- Used too infrequently:
  - POLY, polynomial evaluation (VAX)
  - BCD, bit-field operations (68k)
  - Loop and Procedure call primitives
    * Not quite right for every HLL
    * Unable to take advantage of compiler's analysis
- Exceptions & interrupts are awkward:
  - memcpy/strcmp instructions

CISC instructions RISC dropped

1. *NB. Note use of $\leftarrow$ to indicate source and destination. Some assemblers use src,dest other use dest,src. The $\leftarrow$ convention is used in these notes to avoid this ambiguity.*
2. *nop, mov, zero would have had short instruction encodings on CISC*
3. *No execution time or opcode size advantage to be gained from having dedicated nop,mov,zero instructions on RISC. (would even result in more complicated decode logic).*
4. *Assembler may provide macros for mov, neg, clr etc.*
5. *Many 70's architectures attempted to close the 'semantic gap' between HLLs and machine code*
6. *Complex instructions that perform very specific functions*
7. *Although they are attractive to humans, Compilers often find them inconvenient to use.*
8. *Loop and procedure call primitives often weren't quite right*
9. *Procedure call primitives often were unable to make use of the compiler's static analysis to determine which registers to save and restore. (leaf procedures, call graph etc)*
10. *On more recent implementations, it is often slower to use the complex instruction than to use several standard ones. E.g. 'simple' instructions are executed directly by the Pentium P6 core, but complicated instructions require microcode to decode and execute.*
11. *memcpy: has intermediate state–has to be resumed after an interrupt or exception rather than restarted. VAX/x86 put intermediate state in GPRs.*
12. *However, fast memcpy would be very useful (though normally limited by memory bandwidth anyway)*

New Instructions

- integer divide, sqrt
- popcount, priority_encode
- Integer SIMD (multimedia)
  - Intel MMX, SPARC VIS, Alpha, PA-RISC MAX
  - MPEG, JPEG, polygon rendering
  - parallel processing of packed sub-words
  - E.g. 8x8, 4x16 bit packed values in 64b word
  - arithmetic ops with 'saturation'
    - s8 case: $125+4 = 127$
  - min/max, logical, shift, permute
  - RMS error estimation (MPEG encode)
  - Will compilers ever use these instrs?
- FP SIMD (3D geometry processing)
  - E.g. 4x32 bit single precision
  - streaming vector processing
  - Intel SSE, AMD 3D-Now, PPC AltiVec
- prefetch / cache hints (e.g. non-temporal)
- Maintaining backwards compatiblity
  - Use alternate routines
  - Query CPU feature set

1. *divide - Now enough Si to justify it. (though IA64 doesn't have integer multiply or divide - it relies on the FP unit)*
2. *popcount, priority - used by OS, requested on Alpha by Cray*
3. *multimedia instructions: Probably won't be generated by compilers. (Months spent hand coding MPEG routines)*
4. *x86 MMX uses FP regs so as not to create more user state (thus no changes to the OS [MS Windows] are required)*
5. *software trap mechanism is slow–better to detect the new instructions absence (e.g. via the feature register), and branch to some code providing similar functionality.*
6. *MAX- see PA-RISC 2.0 Extensions article*
7. *Is adding all these instructions still adhering to the RISC philosophy? Given the increasing importance of multimedia, the answer is probably yes. It will be interesting to see if compilers ever manage to use these instructions (some evidence that they are). In the meantime, people are putting lots of effort into coding standard libraries with MMX/KNI/AltiVec instructions e.g. the OpenGL rendering library. Another alternative is to add language extensions that enable SIMD/vector operations to be specified in a HLL.*
8. *FP Single Instruction Multiple Data (SIMD) (vector) instructions added to improve 3D geometry processing. 'Streaming SIMD' on the Pentium III adds the first new architecturally visible CPU register state since the 386 (eight new 128bit regs). This will require some modification of the OS context switch code. Now this has happened, Intel are bringing out SIMD-2, which effectively replaces the original MMX.*
9. *Vector units repeat the same operations to streams (vectors) of operands. (unlike e.g. MMX which operates on a small fixed numbers of operands in a word).*
10. *Better to have an alternate routine than to try software emulation of the missing instructions. Taking illegal instruction traps is very slow.*
11. *Finished the survey of instructions available, next, taking a look at register sets...*

Registers and Memory

- Register set types
  - Accumulator architectures
  - Stack
  - GPR
- Number of operands
  - 2
  - 3
- Memory accesses
  - any operand
  - one operand
  - load-store only

Accumulator Architectures

- Register implicitly specified
- E.g. 6502, 8086 (older machines)

```
LoadA   foo
AddA    bar
StoreA  res
```

- Compact instruction encoding
- Few registers, typically ≤ 4 capable of being operands in arithmetic operations
- Forced to use memory to store intermediate values
- Registers have special functions
  - e.g. loop iterators, stack pointers
- Compiler writers don't like non-orthogonality

1. *historical perspective*
2. *3 mem locs labelled foo, bar and res. (Offsets calculated by assembler)*
3. *registers much faster for storing intermediates. registers not being used for their special functions can be used as temporaries.*
4. *never enough of the right type of register...*

Stack Architectures

- Operates on top two stack items
- E.g. Transputer, (Java)

```
Push  foo
Push  bar
Add
Pop   res
```

- Stack used to store intermediate values
- Compact instruction encoding
- Smaller executable binaries, good if:
  - memory is expensive
  - downloaded over slow network
- Fitted well with early compiler designs

1. *Like accumulator architecture, tricky to make go fast (especially super-scalar implementations)*
2. *Stacks are easier to cache then general memory accesses, but spotting dependencies between instructions is still harder than with registers.*
3. *One advantage of stack architectures is that they do not impose a limit on the number of architecturally visible 'registers', thus allowing future scalability.*
4. *Back to something more familiar, the RISC approach*

General Purpose Register Sets

- Post 1980 architectures, both RISC and CISC
- 16,32,*128* registers for intermediate values
- Separate INT and FP register sets
  - Int ops on FP values meaningless
  - RISC: Locate FP regs in FP unit
- Separate Address/Data registers
  - address regs used as bases for mem refs
  - e.g. Motorola 68k
  - not favoured by compiler writers ($8 + 8 \neq 16$)
  - RISC: Combined GPR sets

1. *x86 32bit mode (IA-32) is basically an 8 register GPR set*
2. *AMD 29K embedded controller family (used in many laser printers) has 128 registers, as does Intel's IA-64 Merced*
3. *32 may not be enough (look at register renaming later)*
4. *FP regs are used by different functional units from INT regs. Interpreting integers as FP values and vice versa is relatively rare. Makes sense to keep them separate, enabling a dedicated register file to be located next to the appropriate functional units.*
5. *Some ISAs have instructions to move values between FP and int regs. Alpha doesn't: transfer has to be achieved via a StoreFP followed by a Load (via L1 D-cache)*
6. *New 'media processors' (DSPs) tend to have common int and FP regs. They believe that conversion between int and fp will be a common operation.*
7. *Dedicated address registers may be mildly advantageous to implementation, as address registers could be located closer to the load/store unit.*
8. *But, Separate address and data registers can be a pain, as for some loops you wish you had more address registers, for other loops you wish you had more data. End up copying values back and forth between data and address.*

Load-Store Architecture

- Only load/store instructions ref memory
- The RISC approach

→ Makes pipelining more straightforward

$$
\begin{array}{lll}
\texttt{Load} & \texttt{r1} & \leftarrow \texttt{foo} \\
\texttt{Load} & \texttt{r2} & \leftarrow \texttt{bar} \\
\texttt{Add} & \texttt{r3} & \leftarrow \texttt{r1, r2} \\
\texttt{Store} & \texttt{res} & \leftarrow \texttt{r3}
\end{array}
$$

- Fixed instruction length (32bits)
- 3 register operands
- Exception: ARM-Thumb, MIPS-16 is two operand
  - more compact encoding (16bits)

1. *Only one port required on D-cache to avoid stalls due to structural hazards*

Register-Memory

- ALU instructions can access 1 or more memory locations
- E.g. Intel x86 32bit modes
  - 2 operands
  - can't both be memory

```
Load   r1←foo
Add    r1←bar
Store  res←r1
```

- E.g. DEC VAX
  - 2 and 3 operand formats
  - fully orthogonal

```
Add  res←bar,foo
```

- Fewer instructions
  - Fewer load/stores
  - Each instruction may take longer
  - $\rightarrow$ Increased cycle time
- Variable length encoding
  - May be more compact
  - May be slower to decode

Register-Memory

1. *Variable length encoding: Intel 32bit mode 1-17bytes, 16bit mode 1-9, VAX 1-19*
2. *orthogonal: any combination is legal*
3. *each instruction may be slower, harder to pipeline*
4. *Although most post 1980 architectures use GPR sets, there are still some special registers...*

Special Registers : 1

- Zero register
  - Read as Zero, Writes discarded
  - e.g. Alpha, Mips, Sparc, IA-64
  - Data move: `add r2` $\leftarrow$ `r1, r31`
  - nop: `add r31` $\leftarrow$ `r31, r31`
  - prefetch: `ldl r31` $\leftarrow$ `(r1)`
  - Zero is a frequently used constant

- Program Counter
  - NOT usually a GPR
  - Usually accessed by special instructions e.g. branch, branch and link, jump
  - But, PC is GPR r15 on ARM

1. *zero register: r0 on MIPS, r31 on Alpha*
2. *prefetch: Allows a value to be pulled into cache without risk of generating an exception*
3. *ARM's visible PC exposes the pipeline depth. This was a pain on the StrongArm which has a 5 stage pipe instead of 3.*

Special Registers : 2

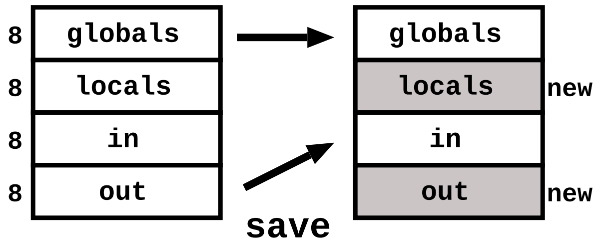
- Condition code (Flag) registers
  - Carry, Zero, Negative, Overflow
  - Used by branches, conditional moves
  - Critical for pipelining and super-scalar
  - CISC: one CC reg updated by all instructions
  - ARM, SPARC: one CC reg, optionally updated
  - PowerPC: multiple CC regs (instr chooses)
  - IA64: 64 one bit predicate regs
  - Alpha, MIPS: no special CC regs
- Link registers
  - Subroutine call return address
  - CISC: pushed to stack
  - RISC: saved to register
    - register conventions
    - only push to stack if necessary
  - Jump target/link regs (PowerPC, IA-64)
  - fixed GPR (r14, ARM) (r31,MIPS)
  - GPR nominated by individual branch (Alpha)

Special Registers : 2

1. *Revision: meanings of the various CC flags.*
2. *ARM/SPARC: Bit in opcode to determine whether CC updated.*
3. *Single CC register poses problems for super-scalar implementation due to false dependencies.*
4. *Alpha and MIPS have no CC registers. GPRs used in their place, thus using the normal dependency analysis logic. Have to test for carry/overflow etc using ALU instructions.*
5. *IA64 has 64 single bit predicate registers. Compare instructions nominate two pred regs: one will be cleared, the other set.*
6. *More on the implications of CC regs for control flow later...*
7. *register conventions: args, result, temps, caller saves, callee saves*
8. *Wheeler Jumps*
9. *push to stack: eg if recursive*
10. *The link register typically has to be moved to a callee saves register before making another procedure call. Alpha avoids this by allowing the current PC to be stored in a nominated register.*
11. *IA64 has 8 branch target registers. Special registers are used in order to give hints to the instruction prefetch unit.*

Register Conventions

- Linkage (Procedure Call) Conventions
  - Globals: `sp`, `gp` etc.
  - Args: First (4-6) args (rest on stack)
  - Return value: (1-2)
  - Temps: (8-12)
  - Saved: (8-9) Callee saves
- Goal: spill as few registers as possible in total
- Register Windows (SPARC)
  - `save` and `restore`
  - 2-32 sets of windows in ring
  - 16 unique registers per window
  - spill/fill windows to special stack



- IA-64: Allows variable size frames
  - 32 globals
  - 0-8 args/return, 0-96 locals/out args
  - h/w register stack engine operates in background

Register Conventions

1. *languages can have different conventions for the same CPU*
2. *Caller saves its own args and temps*
3. *goal: save as few as possible in total. Don't save all before procedure call as callee may not use many. Don't save all after as caller may not have used many.*
4. *Compiler should allocate variables to 'Saved' registers, then put any that are left over in 'Temps'.*
5. *Call graph optimization can be used to tune which registers are actually saved.*
6. *When SPARC was designed, the Berkeley team felt that the current compiler technology was not up to the job of assigning arguments etc to registers.*
7. *Sparc: don't have to move incoming args to safety before making another call–it's done in hardware.*
8. *Sparc: running out of windows is bad – normally need to save the full 16 registers in the window regardless of which are actually used.*
9. *Sparc: Total context is very large—painful to context switch.*
10. *Next, a look at addressing modes...*

Classic RISC Addressing Modes

- Register
  - `Mov r0 ← r1`
  - `Regs[r0] = Regs[r1]`
  - Used when value held in register
- Immediate
  - `Mov r0 ← 42`
  - `Regs[r0] = 42`
  - Constant value limitations
- Register Indirect
  - `Ldl r0 ← [r1]`
  - `Regs[r0] = Mem[ Regs[r1] ]`
  - Accessing variable via a pointer held in reg
- Register Indirect with Displacement
  - `Ldl r0 ← [r1, #128]`
  - `Ldl r0 ← 128(r1)`
  - `Regs[r0] = Mem[ 128 + Regs[r1] ]`
  - Accessing local variables

Classic RISC Addressing Modes

1. *First two not really addressing modes as don't access memory...*
2. *Different assemblers have different syntax - Even for same arch!*
3. *register indirect obviously a special case of register indirect with displacement - just a more compact encoding on CISC.*
4. *Explain arrays Regs and Mem*
5. *Pretty much all processors have the RISC addr modes*
6. *Immediate values: on most CISCs they must be prefixed with # or $*
7. *reg indirect with displacement used for accessing local vars off stack pointer, or global variables from the GP register*
8. *displacement size is often limited on RISC architectures.*

Less RISCy addr modes

- ARM and PowerPC
- Register plus Register (Indexed)
  - `Ldl r0` $\leftarrow$ `[r1,r2]`
  - `Regs[r0] = Mem[ Regs[r1] + Regs[r2] ]`
  - Random access to arrays
  - e.g. r1=base, r2=index
- Register plus Scaled Register
  - `Ldl r0` $\leftarrow$ `[r1, r2, asl #4]`
  - `Regs[r0] = Mem[ Regs[r1] + (Regs[r2]`$\ll$`4)]`
  - Array indexing
  - sizeof(element) is power of 2, r2 is loop index
- Register Indirect with Displacement and Update
  - Pre inc/dec `Ldl r0` $\leftarrow$ `[r1!, #4]`
  - Post inc/dec `Ldl r0` $\leftarrow$ `[r1], #4`
  - C `*(++p)` and `*(p++)`
  - Creating stack (local) variables
  - Displacement with post update is IA-64's only addressing mode

Less RISCy addr modes

1. *can improve instruction density*
2. *reg+reg requires an extra read port on register file*
3. *ARM: shift unit can be used to scale one operand for any instruction.*
4. *syntax for displacement and update highly variable. For some architectures, displacement is fixed to the sizeof the value being loaded.*
5. *Full descending stack use: pre dec, post inc*
6. *Empty descending stack use: post dec, pre inc*
7. *Displacement with update requires an extra write port on register file.*
8. *IA-64 ONLY supports reg+disp with compulsory post update & reg+reg with compulsory update. Accessing structure members presumably requires the compiler to deal with a 'moving base pointer' as different elements are accessed. Understanding disassembled IA-64 code is going to be a nightmare...*

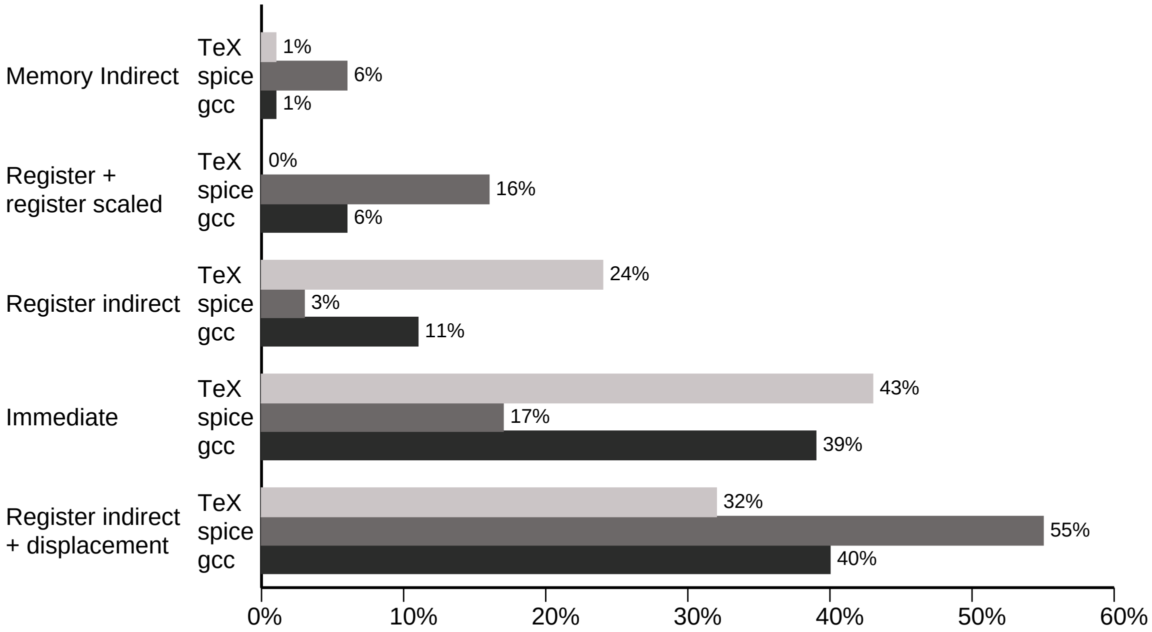
CISC Addressing Modes

- Direct (Absolute)
  - `Mov r0` ← `(1000)`
  - `Regs[r0] = Mem[1000]`
  - Offset often large
  - x86 Implicit base address
  - Most CISCs
- Memory Indirect
  - `Mov r0` ← `@[r1]`
  - `Regs[r0] = Mem[ Mem[ Regs[r1] ] ]`
  - Two memory references,
  - C `**ptr`, linked lists
- PC Indirect with Displacement
  - `Mov r0` ← `[PC, #128]`
  - `Regs[r0] = Mem[ PC + 128 ]`
  - Accessing constants

CISC Addressing Modes

1. *Absolute mode: eg. 6502 Zero page*
2. *absolute mode synthesizable on RISC with zero register*
3. *absolute mode is not compatible with object oriented programming*
4. *Memory indirect: two memory references required, thus two possibilities for exceptions to occur.*
5. *Even more complicated modes exist e.g. 68020's Mem Indirect Pre Indexed mode: two displacement, a scale, two registers and two memory accesses. Abandoned on 68040?*
6. *PC indirect useful for Position Independent code, e.g. for single address space operating systems, shared libraries etc.*
7. *branch instructions are typically PC relative – PC indirect mode enables relative access to data. Particularly useful for read-only data, which can be placed in the text segment (usually just after the function in which it is used).*
8. *Assembler can usually fills in the offset between the instruction doing the access and the data being accessed.*
9. *ARM supports PC indirect mode, as PC is just another GPR*
10. *Other RISC processors generally use other conventions to access literal constants in a position-independent manner. e.g. a register containing the 'procedure variable'*
11. *PC indirect mode can be synthesized on other RISC processors by doing a subroutine call to the following instruction, and using the PC value stored in the link register.*

Why did RISC choose these addressing modes?



Frequency of addressing modes (VAX)

- RISC
  - immediate
  - register indirect with displacement
- ARM, PowerPC reduce instruction counts by adding:
  - register + register scaled
  - index update

1. *Quantitative approach*
2. *IA-64 allows only register indirect, but forces the base register to be incremented by either a Reg or an Immediate AFTER the load.*

Immediates and Displacements

- CISC: As instructions are variable length, immediates and displacements can be any size (8,16,32 bits)
- RISC: How many spare bits in instruction format?
- Immediates
  - used by data-processing instructions
  - usually zero extended (unsigned)
    - `add` → `sub`
    - `and` → `bic`
  - For traces on previous slide:
    50-70% fit in 8bits, 75-80% in 16bits
  - IA-64 22/14, MIPS 16, Alpha 8,
    ARM 8 w/ shift
- Displacement values in load and stores
  - Determine how big a data segment you can address without reloading base register
  - usually sign extended
  - MIPS 16, Alpha 16, ARM 12, IA-64 9

1. *Immediates that can not fit in instruction must be synthesized in a register, or loaded from the literal pool code segment.*
2. *max size of displacement can limit the size of the global data area under some linkage conventions.*
3. *displacement is usually signed because negative displacements are common.*
4. *For some destination registers IA-64 allows 22 bit immediates.*
5. *IA-64 also has a hack for loading 64bit immediate values using two of the 41 bit instruction slots.*

Instruction Encoding

RISC: small number of fixed encodings of same length

Operation	Ra	Rb	Signed Displacement			load/ store
Operation	Ra	Rb	Zero SBZ	Function	Rdest	operate
Operation	Ra	Immediate Value		Function	Rdest	operate immediate
Operation	Ra	Signed Displacement				branch

RISC instruction words are 32 bit

IA-64 packs three 41 bit instructions into a 128 bit 'bundle'

VAX: fully variable. Operands specified independently

Operation and # of operands	Address specifier 1	Address field 1	••••	Address specifier N	Address field N

x86: knows what to expect after first couple of bytes

Operation	Address specifier	Address field		
Operation	Address specifier	Address field1	Address field2	
Operation	Address specifier	Extended specifier	Address field1	Address field2

1. *RISC: All instruction accesses must be to aligned addresses*
2. *Many encodings reserved so as to enable future extensions e.g. the whole swathe of instructions added to MIPS, SPARC and PA-RISC for 64 bit processing.*
3. *Read As Zero, IGNore, Must Be Zero, Should Be Zero : used to specify instruction encoding in a well defined way.*
4. *RISC: very regular format, simple to decode. (The example is Alpha. One more format exists, branch)*
5. *Alpha: 8bit immediate, 16 displ*
6. *IA-64: Needed instructions a little bigger than 32 bits, yet needed to keep alignment. Answer – collect some number of instructions together into a larger power of 2 sized bundle.*

Code Density Straw Poll

- CISC: Motorola 68k, Intel x86
- RISC: Alpha, Mips. PA-RISC
- Very rough-figures for 68k and Mips include statically linked libc

arch	text	data	bss	total	filename
x86	29016	14861	468	44345	gcc
68k	36152	4256	360	40768	
alpha	46224	24160	472	70856	
mips	57344	20480	880	78704	
hp700	66061	15708	852	82621	
x86	995984	156554	73024	1225562	gcc-cc1
alpha	1447552	272024	90432	1810008	
hp700	1393378	21188	72868	1487434	
68k	932208	16992	57328	1006528	
mips	2207744	221184	76768	2505696	
68k	149800	8248	229504	387552	pgp
x86	163840	8192	227472	399504	
hp700	188013	15320	228676	432009	
mips	188416	40960	230144	459520	
alpha	253952	57344	222240	533536	

- CISC text generally more compact, but not by a huge amount
- Alpha's 64bit data/bss is larger

1. *data is the initialised data segment, bss is allocated at load time and zeroed.*
2. *Shared libraries can make binaries much smaller—just look at an X application.*
3. *Alpha's heap is also likely to be much bigger.*

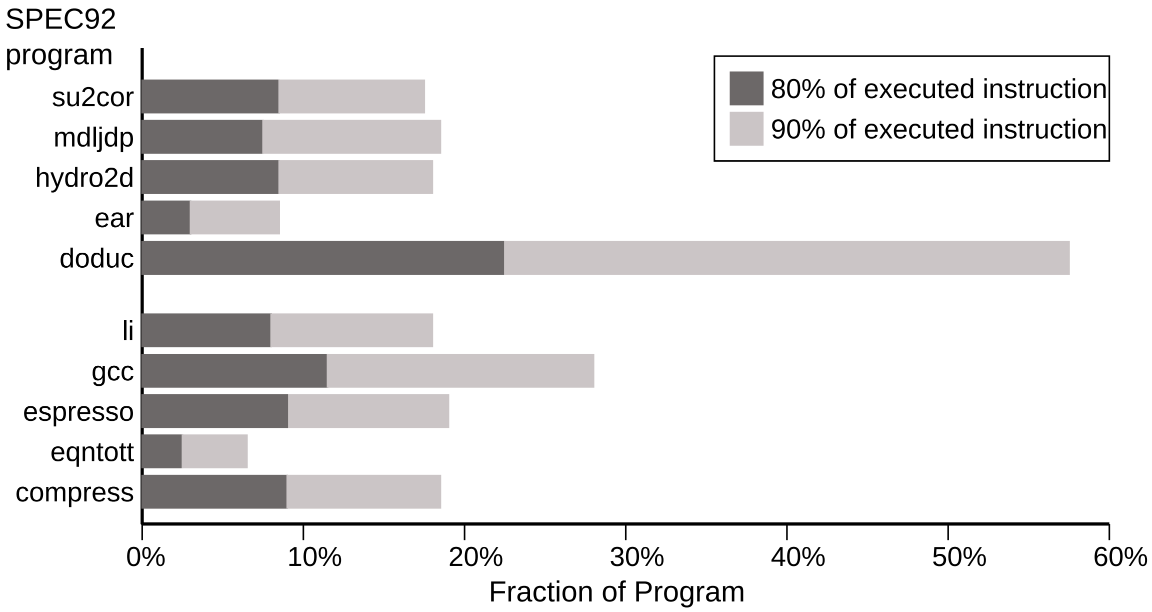
Code Density

- Important if:
  - Memory is expensive
    - * can be in embedded applications
    - * eg. mobile phones
    - ⇒ ARM Thumb, MIPS-16
  - Executable loaded over slow network
    - * Though Java not particularly dense!
- Speed vs. size optimization tradeoffs
  - loop unrolling
  - function inlining
  - brunch/jump target alignment

Code Density

1. *In an embedded application, the memory is often on the same ASIC as the processor. It must fit.*
2. *ARM Thumb decodes 16 bit instructions into normal ARM 32 bit ones which it stores in the on-chip I-cache.*
3. *Java byte code class section compacters/obfiscators (Kaffe)*

Instruction caches



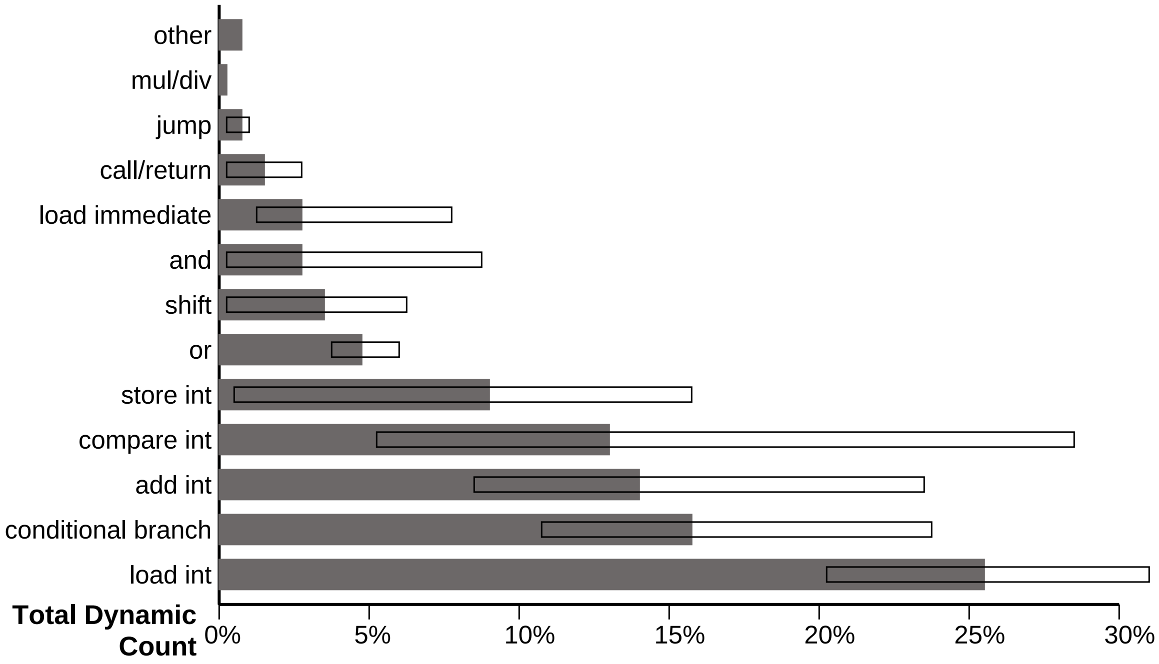
Fraction of program responsible for 80% and 90% of instruction executions

- Caches generally solve I-stream b/w requirements
  - 4bytes x 1GHz x 2-4 instrs = 8-16GB/s !
  - Loops are common! (90% in 10%)
  - Internal I-caches often get 95%+ hit-rates
  - Code density not usually a performance issue
    * assuming decent compilers and app design
    * code out-lining (trace straightening) vs. function in-lining and loop unrolling
- D-Cache generally much more of a problem

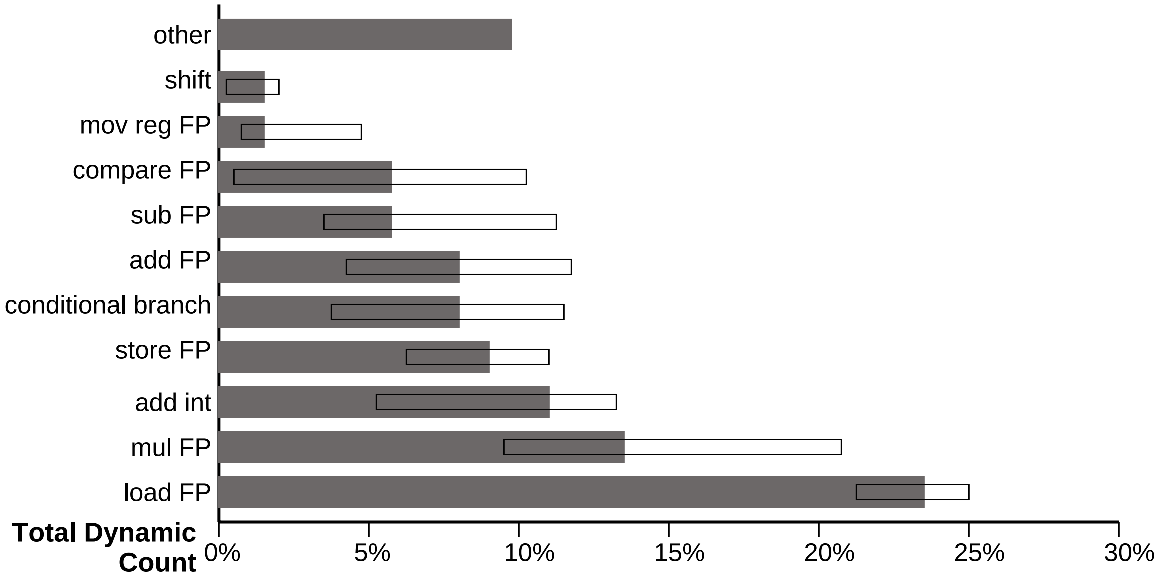
Instruction caches

1. *99% hit rate for SPEC92 with 8KB I-cache.*
2. *Some figures by Sites showing that I-caches get very poor hit rates with some commercial applications e.g. MS SQL server. These figures could be greatly improved through use of better compilers/better structured applications. Should we spend transistor budget to fix s/w inefficiencies ?*
3. *D-cache hit rates are usually much lower than I-cache hit rates, and are thus more likely to limit performance. A return to CISC-style instruction encodings is unlikely for general purpose CPUs.*
4. *function in-lining and loop unrolling are generally good optimization strategies, unless they lead to an I-stream that doesn't fit in the cache.*
5. *code out-lining is the process of moving code that is rarely executed out of the straight-line case. This avoids it being dragged into the cache as part of lines containing useful instructions. May execute better too as less wasted slots due to branches.*

Instruction Mix



Instruction mix for SPEC INT92



Instruction mix for SPEC FP92

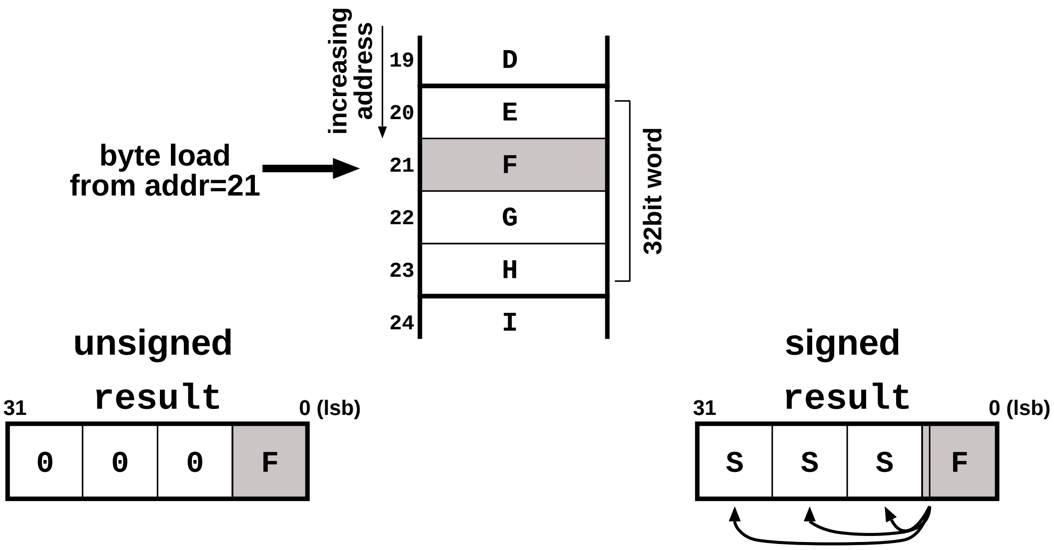
There are no 'typical' programs

Instruction Mix

1. *SPEC92 instruction mix for a MIPS processor (unknown compiler)*
2. *Dynamic vs Static mix - e.g. because compiler often moves multiplies outside of loops, the dynamic mix % will be lower than the static mix.*
3. *There is no such thing as a typical program. There exists a large degree of variance*
4. *SPEC 95's JPEG uses mul a lot - expect to see multiply getting beefed up on new processors.*
5. *loads and stores are frequent. On a CISC architecture, many would be incorporated into arithmetic instructions.*

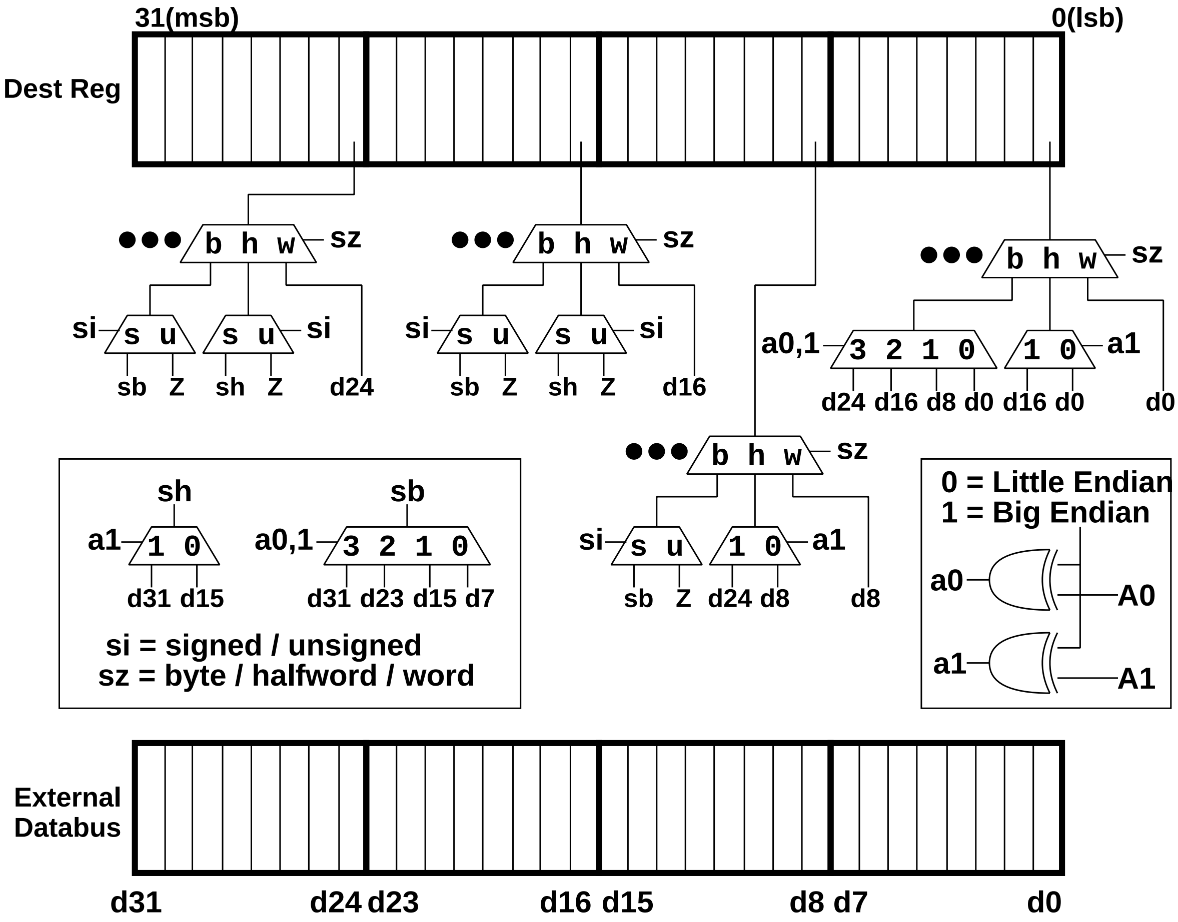
Aligned Loads and Stores

- Address mod sizeof(type) = 0
- Most ISA support 8,16,32,(64) bit loads and stores in hardware
- Signed and unsigned stores same
- Sub-word loads can be Signed and Unsigned
  - CISC: loads merge into dest reg
  - RISC: loads extend into dest reg E.g:



1. *All byte accesses are (by definition) aligned.*
2. *On a CISC, loads generally merge into destination*

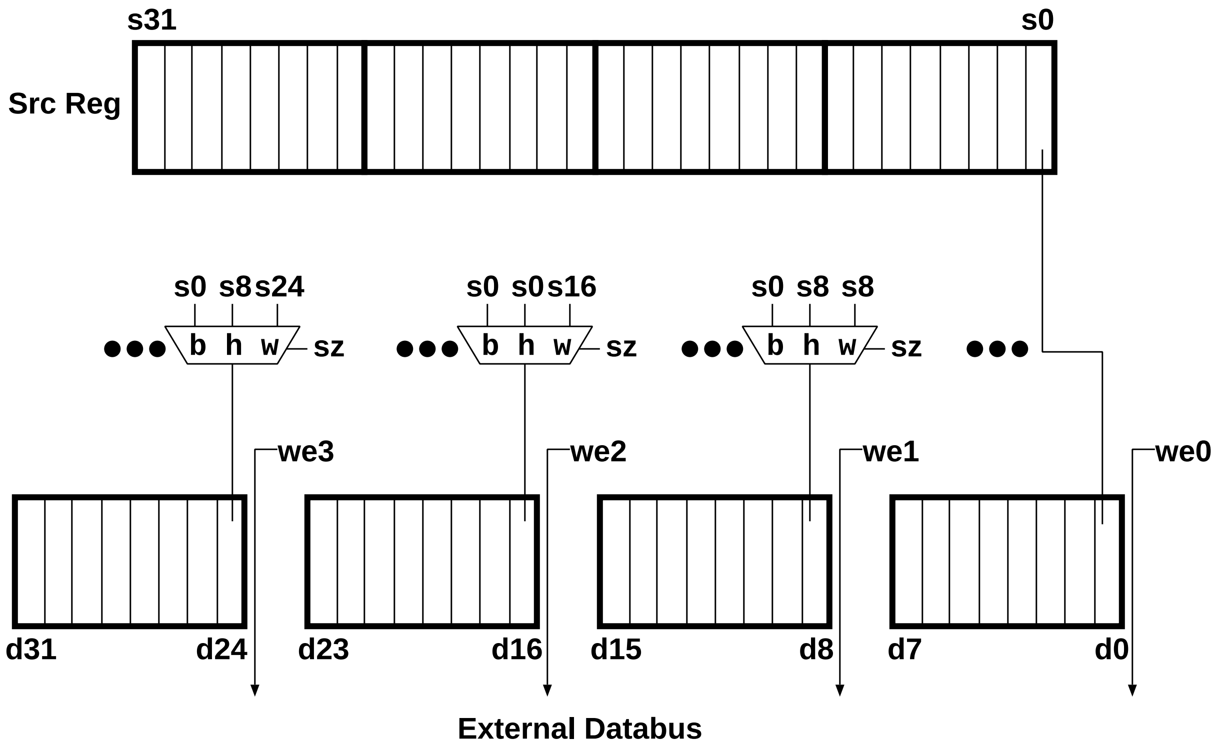
Aligned Sub-word Load Logic



- byte-lane steering
- sign/zero extension
- Big/Little endian modes

1. *These days, data value will probably actually be read from a cache line (16 or 32 bytes long), so require another set of multiplexers to select the word within the line.*
2. *Extra multiplexers will be required to perform sign extension for signed sub-word loads*
3. *Extra multiplexers will be required to perform half-word loads*
4. *In big endian mode, byte loads will capture the most significant byte. NB: I/O devices and other peripherals must be wired up appropriately for this to work correctly as a big endian machine.*

Aligned Sub-word Store Logic



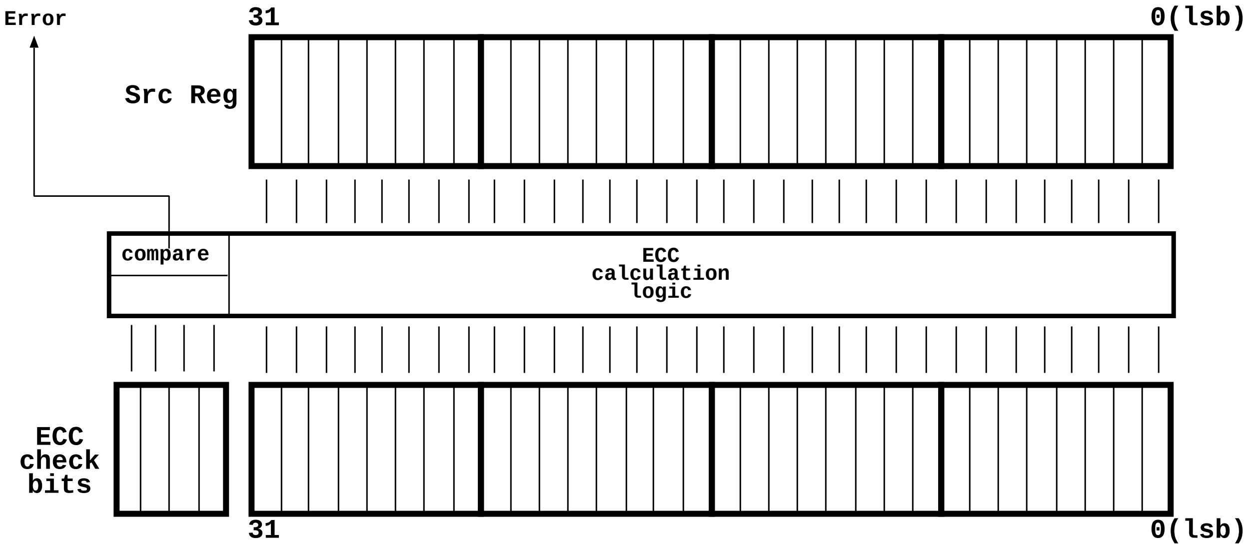
a1	a0	sz	we3	we2	we1	we0
0	0	w	1	1	1	1
0	0	h	0	0	1	1
1	0	h	1	1	0	0
0	0	b	0	0	0	1
0	1	b	0	0	1	0
1	0	b	0	1	0	0
1	1	b	1	0	0	0

- Replicate bytes/halfwords across bus
- Write enable lines tell memory system which byte lanes to latch

1. *Byte must be steered into the appropriate byte line (or duplicated across the whole bus*
2. *Big/Little Endian select can be done for stores in a similar manner to loads.*
3. *Modified bytes are merged into cache line/memory*
4. *Not all machines have individual byte write enables. Need to read-insert-write.*

Sub-Word Load/Stores

- Word addressed machines
  - Addr bit A0 addresses words
- Alpha (v1):
  - Byte addressed, but 32/64 load/stores only
  - Often critical path
  - Sub-word stores hard with ECC memory
  - So, emulate in s/w using special instructions for efficiency



1. *word addressed machines (e.g. super computers) can't directly access values smaller than a word.*
2. *Alpha has had sub-word load/store extensions added into the architecture from the 21164A (EV56) chip. It turned out that there was just too much 'legacy' code that relied upon byte ops, and the compilers couldn't make a good enough job of eliminating the accesses.*
3. *Whole word has to be written to (easily) calculate ECC bits. Generally requires a read/merge/write-back.*

Emulating Byte Loads

1. Align pointer
2. Do word load
3. Shift into low byte
4. Mask
5. (sign extend)

- e.g. 32bit, Little Endian, unsigned

```
unsigned int temp;
temp = *(p&(~3));
temp = temp >> ((p&3) *8);
reg  = temp & 255;
```

- e.g. 32bit, Big Endian, unsigned

```
unsigned int temp;
temp = *(p&(~3));
temp = temp >> ( (3-(p&3)) * 8);
reg  = temp & 255;
```

- e.g. 64bit, Little Endian, signed

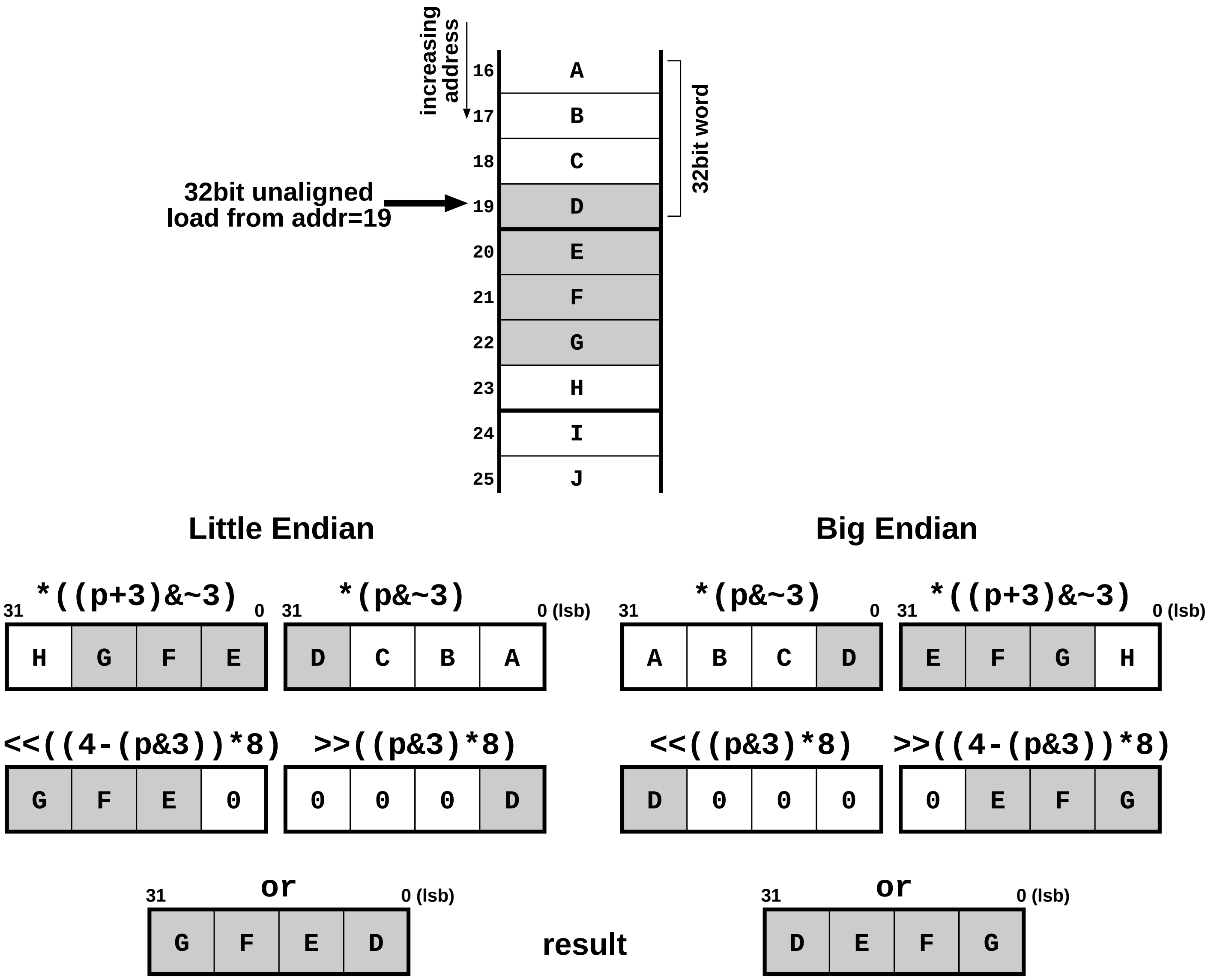
```
long temp;
temp = *(p&(~7));
temp = temp << ( (7-(p&7)) * 8);
reg  = temp >> 56;
```

Emulating Byte Loads

1. *(p & ∼3) can be done with a* `bic` *instruction*
2. *Above is pseudo code. C syntax would requires casts of* p *to (unsigned long) and then (int *)*
3. *Show how to do it using assembler. Note use of lsr and asr.*

Unaligned Accesses

- Address mod sizeof(value) $\neq$ 0
- E.g. :



1. *4 byte quantities have four possible alignments : only one is aligned.*
2. *show how this can be achieved in hardware using a 2*32 bit register and 32 four input multiplexers.*

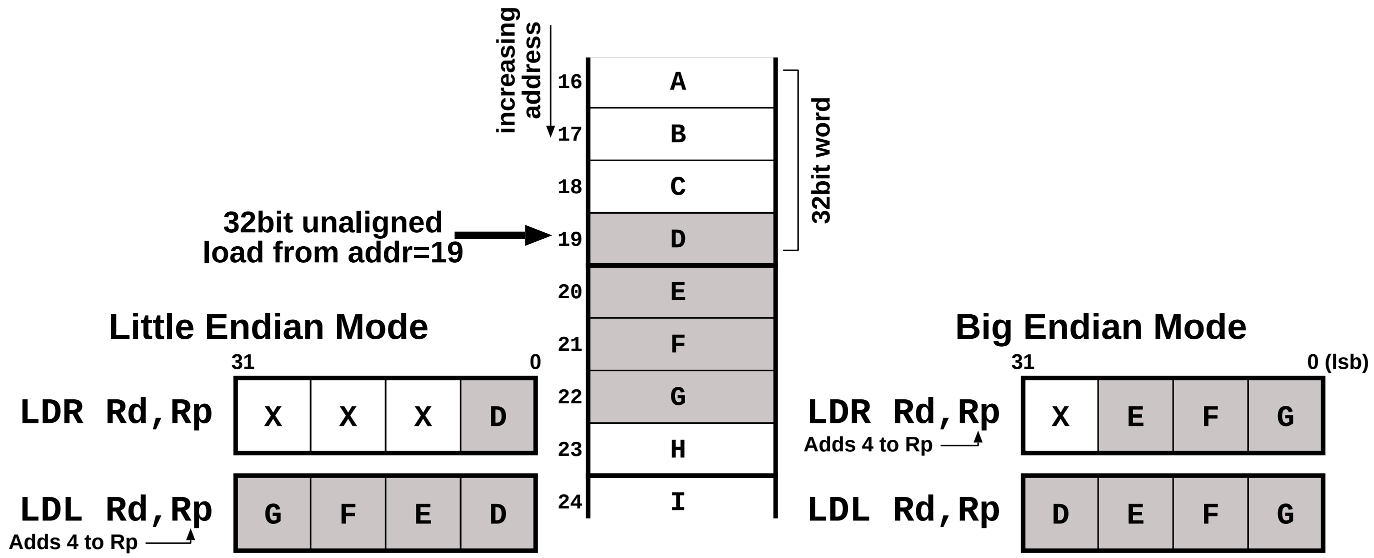
Unaligned Accesses

- CISC and Power PC support unaligned accesses in hardware
  - Two memory accesses
    - $\rightarrow$ Less efficient
  - May cross page boundaries
- Most RISCs synthesize in software
  - Provide special instructions
- Compilers try to keep data aligned
  - `struct` element padding
- Casting `char *` to `int *` dangerous

1. *PPC unusual.*
2. *Many bugs in older CISC processors to do with dealing with exceptions when crossing page boundaries*
3. *Sub word unaligned accesses may not require two memory accesses*
4. *Some operating systems automatically 'fix-up' unaligned accesses by executing a s/w handler. This is very inefficient, especially as the trap is usually logged to syslog.*

MIPS Unaligned Support

- LWR Load Word Right
- LWL Load Word Left
  - Only one memory access per instruction
  - Does shifting and merging as well as load
  - → Unaligned load in 2 instrs



- STR Store Word Right
- STL Store Word Left
- Uses byte store hardware to merge into memory/cache

1. *LWL/LRL instructions are still quite complicated. Slow the pipeline?*
2. *Normal LW/SW instructions trap if address not aligned.*
3. *See man page in supplied material.*

Alpha Unaligned Loads

- `LDQ` trap if not 8byte aligned
- `LDQ_U` ignore a0-a2
- `EXTQL Rd ← Rs, Rp`
  Shift Rs right by Rp&7 bytes and extracts quad word into Rd.
- `EXTQH Rd ← Rs, Rp`
  Shift Rs left by 8-Rp&7 bytes and extracts quad word into Rd.
- Alpha requires 5 instrs for arbitrary unaligned load
```
LDQ_U   Rd ← Rp
LDQ_U   Re ← Rp + #7
EXTQL   Rd ← Rd, Rp
EXTQH   Re ← Re, Rp
OR      Rd ← Rd, Re
```
- `EXTBL Rd ← Rs, Rp`
  Shift Rs right by Rp&7 bytes and extracts low byte into Rd.
- also `EXTLL, EXTLH, EXTWL, EXTWH`
- If alignment of pointer is known, may use optimized sequence
  E.g. load 4bytes from address 0x123
```
LDQ Rd ← -3(Rp)
EXTLL Rd ← Rd, #3
```

Alpha Unaligned Loads

1. *Simple shift-and-mask instructions.*
2. *Q stands for quad word (8 bytes), L for long word (4bytes)*
3. *Same instructions used for sub-word accesses e.g. unsigned byte load with LDQ_U Rd, Rp; EXTBL Rd, Rd, Rp*
4. *for signed byte: LDQ_U Rd, Rp; add Rp, Rp,1; EXTQL Rd, Rd, Rp; asr #56*
5. *See Alpha man pages (attached)*

Alpha unaligned stores

- No byte hardware, so load quad words, merge, and store back
- `INSQL Rd ← Rs, Rp`
  Shift Rs left by Rp&7 bytes
- `INSQH Rd ← Rs, Rp`
  Shift Rs right by 8-Rp&7 bytes
- `MSKQL Rd ← Rs, Rp`
  Zero top 8-Rp&7 bytes
- `MSKQH Rd ← Rs, Rp`
  Zero bottom Rp&7 bytes
- E.g.: Store quad word Rv to unaligned address Rp

```
LDQ_U   R1 ← Rp        Load both quad words
LDQ_U   R2 ← Rp + #7
INSQH   R4 ← Rv, Rp    Slice & Dice Rv
INSQL   R3 ← Rv, Rp
MSKQH   R2 ← R2, Rp    Zero bytes to be replaced
MSKQL   R1 ← R1, Rp
OR      R2 ← R2, R4    Merge
OR      R1 ← R1, R3
STQ_U   R2 → Rp + #7   Store back
STQ_U   R1 → Rp        Order important:aligned case
```

1. *Also Long word, word and byte versions*
2. *Unaligned stores are painful without logic to perform sub-word stores. Probably why they've since been added to Alpha...*
3. *IA-64 'architecturally' supports unaligned accesses though individual hardware implementations may not. The instruction set does have some good tools for doing software synthesising the code, such as a bit field extract instruction that extracts a 64 bit value from a concatenated pair of 64bit registers.*

Copying Memory

- Often important:
  - OS: user args, IPC, TCP/IP
  - user: realloc, pass-by-value
- `memmove`
  - Must deal correctly with overlapping areas
- `memcpy`
  - Undefined if areas overlap
  - Enables fixed direction
- `copy_aligned`
  - Source and Dest `long` aligned
  - Fastest
- Small copies ($<$ 100 bytes)
  - Avoid large start-up costs
- Medium sized copies (100–100KB bytes)
  - Use highest throughput method
- Large copies
  - Probably memory b/w limited anyway...

1. *memmove: may need to reverse copy direction e.g. dest = src+1*

copy_aligned

- E.g. for 32bit machine

```
void copy_aligned( int32 *d, const int32 *s, int n)
{
  sub n, n, #4
  blt n, return    ; if n<0 exit
loop:
  ldw tmp, (s)
  add d, d, #4
  sub n, n, #4     ; set branch value early
  add s, s, #4
  stw tmp, -4(d)   ; maximise load-to-use
  bgt n, loop      ; if n>0 branch (no delay slot)
}
```

- Use widest datapath
  - (64bit FP regs on PPro)
- Maximize cycles before tmp is used
- Update n well in advance of branch
- To further optimize:
  - Unroll loop to reduce loop overhead
  - Instruction scheduling of unrolled loop
  - (software pipelining)

copy_aligned

1. *memcpy function has src as a constant.*
2. *Utilizing delay slot*
3. *Unroll loop to reduce per iteration loop overhead*
4. *Touch the next cache line well in advance of use, thus avoiding a data dependent stall.*
5. *Re-order instructions such that they get issued to functional units without stalls due to structural hazards.*
6. *Highest throughput copy on some x86's can be achieved by using the 64bit data-paths provided by the FP regs. The dedicated MOVS instruction is actually slower! (has to deal with all the special cases...)*

copy_aligned (2)

```
void copy_8_aligned( int32 d[], const int32 s[], int n)
{
    int32 t0,t1,t2,t3,t4,t5,t6,t7;
top:
    t0 = s[0];   t1 = s[1];
    t2 = s[2];   t3 = s[3];
    t4 = s[4];   t5 = s[5];
    t6 = s[6];   t7 = s[7];
    n  = n - 32; s  = s + 32;
    d[0] = t0;   d[1] = t1;
    d[2] = t2;   d[3] = t3;
    d[4] = t4;   d[5] = t5;
    d[6] = t6;   d[7] = t7;
    d  = d + 32; if (n) goto top;
}
```

- Need to deal with boundary conditions
  - e.g. if n mod 32 != 0
- Get cache line fetch started early
  - Issue a load for the next cache line
    - OK if non-blocking cache
    - beware exceptions (array bounds)

  ⇒ prefetch or speculative load & check

  ⇒ non-temporal cache hints
- IA-64: 'Rotating register files' to assist software pipelining without the need to unroll loops

copy_aligned (2)

1. *Use of stylised C to direct compiler — can help considerably, as current product compilers are not as good as one might wish/expect.*
2. *Can issue a load for the following cache line ahead of its use. Providing the cache is non-blocking execution will continue until the destination register is actually used (non-blocking caches are pretty standard these days).*
3. *However, must be careful to ensure that the load doesn't fault by e.g. going off the end of the array. This can be avoided by the use of special prefetch or speculative load instructions that effectively suppress exceptions.*
4. *IA-64 rotating register files enable soft ware pipelining without the Icache bloat caused by loop unrolling. They also assist with dealing with the boundary conditions on loop entry and exit. I recommend wrapping a wet towel around your head before attempting to read that particular section of the IA-64 optimization manual...*
5. *Regs r0-r1 global. Rotating registers can be any multiple of 8 starting at r32. swp_branch decrements the RRBase raps at 0. e.g. 8 regs would be r32 - r39. If rrb=1 r35-¿36. rrb=6 r35-¿r33*

   ```
   L1: ld r35 = [s], 4    # post inc
       st r38 = [d], 4
       swp_branch L1
   ```

6. *non-temporal hints indicate that the data is unlikely to exhibit temporal locality. Typically this is implemented by loading the data into the line in the set that will be evicted next. Some architecture have a separate small cache to hold such data.*

Unaligned copy

- E.g. 32bit, Little Endian

```
void memcpy( char *d, const char *s, int n)
{
  uint32 l,h,k,*s1,*d1;

  /* Align dest to word boundary */
  while ( ((ulong)d&3) && n>0 ) {*d++ = *s++; n--;}

  /* Do main work copying to aligned dest */
  if( ((ulong)s & 3) == 0 ) {       /* src aligned ? */
     k = n & ~3;                    /* round n down  */
     copy_aligned(d, s, k);
     d+=k; s+=k; n&=3;              /* ready for end */
  }
  else
  {
    s1 = (uint32 *)((ulong)s & ~3); /* round s down  */
    d1 = (uint32 *) d;              /* d is aligned  */
    h  = *s1++;                     /* init h        */
    k  = (ulong)s &3;               /* src alignment */
    for(; n>=4; n-=4) {             /* stop if n<4   */
      l = *s1++;
      *d1++ = ( h >> (k*8)     ) |
              ( l << ((4-k)*8) ) ;
      h = l;
    }
    d = (char *) d1;                /* ready for end */
    s = ((char *)s1) - 4 + k;
  }

  /* Finish off if last 0-3 bytes if necessary */
  for( ; n>0; n-- ) *d++ = *s++;
}
```

1. *load and store each word once. Align to dest*
2. *use of copy_aligned if (s&3 == d&3)*
3. *big endian version: simply reverse shifts in inner loop*
4. *Consider hardware implementation of copy_aligned*

Memory Translation and Protection

- Protection essential, even for embedded systems
  - isolation, debugging
- Translation very useful
  - demand paging, CoW, avoids relocation
- Segmentation *vs.* Paging
  - x86 still provides segmentation support
  - descriptor tables: membase, limit
  - segment selectors : cs, ds, ss, fs, gs
- Page protection preferred in contemporary OSes
- Translation Lookaside Buffer (TLB)
  - translate Virtual Frame Number to PFN
  - check user/supervisor access
  - check page present (valid)
  - check page writeable (DTLB)
- Separate D-TLB and I-TLB
  - often a fully associative CAM
  - separate I-TLB and D-TLB
  - typically 32-128 entries
  - sometimes an L2 Joint-TLB e.g. 512 entry
- Hardware managed *vs.* software managed TLB

1. *often fully associative as can't rely on spatial locality*
2. *not last used replacement policy*
3. *32x4KB = 128KB not very much!*

Hardware page table walking

- Hierarchical lookup table
- E.g. x86/x86_64 4KB pages evolved over time:
  - 2-level : 4GB virt, 4GB phys (4B PTEs)
  - 3-level : [512GB] virt, 64GB phys (8B PTEs)
  - 4-level : 256TB virt, 1TB phys (8B PTEs) (48 bit VAs are sign extended to 64bit)
- 'set PT base' instruction
  - implicit TLB flush (on x86)
- Flush virtual address
- Global pages not flushed
  - special bit in PTE
  - should be same in every page table!
  - typically used for kernel's address space
  - special TLB flush all
- Superpages are PTE 'leaves' placed in higher levels of the page table structure
  - e.g. 4MB pages on x86 2-level

1. *page tables can get quite big*
2. *x86_64 finally introduces Address Space Numbers*

Software managed TLB

- OS can use whatever page table format it likes
  - e.g. multilevel, hashed, guarded, etc.
  - (generally more compact than hierarchical)
  - use privileged 'untranslated' addressing mode
- Install TLB Entry instruction
  - specify tag and PTE
  - replacement policy usually determined by h/w
    - e.g. not most recently used
- (may allow TLB contents to be read out for performance profiling)
- Flush all, flush ASN, flush specified VA
- Flexible superpage mappings often allowed of e.g. 8, 64, 512 pages.
- Notion of current Address Space Number (ASN)
- TLB entries tagged with ASN
- Try to assign each process a different ASN
  - no need to flush TLB on process switch
  - (only need to flush when recycling ASNs)
- IA-64 : s/w TLB with hardware PT walking assist
- PPC: h/w fill from larger s/w managed hash table

1. *So*

ISA Summary

- RISC
  - Product of quantitative analysis
  - Amdahl's Law
  - Load-Store GPRs
  - ALU operates on words
  - Relatively simple instructions
  - Simple addressing modes
  - Limited unaligned access support
  - (s/w managed TLB)
- Architecture extensions
  - Backwards compatibility
- Copying memory efficiently

Does Architecture matter?

1. *So far, haven't looked at branches. Different kinds of branch instruction have a big impact on control-flow, and are examined in the next section.*
2. *If ISA is hidden behind compiler, does it really matter anyway?*
3. *Next, look at how Architecture affects Implementation*

CPU Performance Equation

$$Time\ for\ task = C * T * I$$

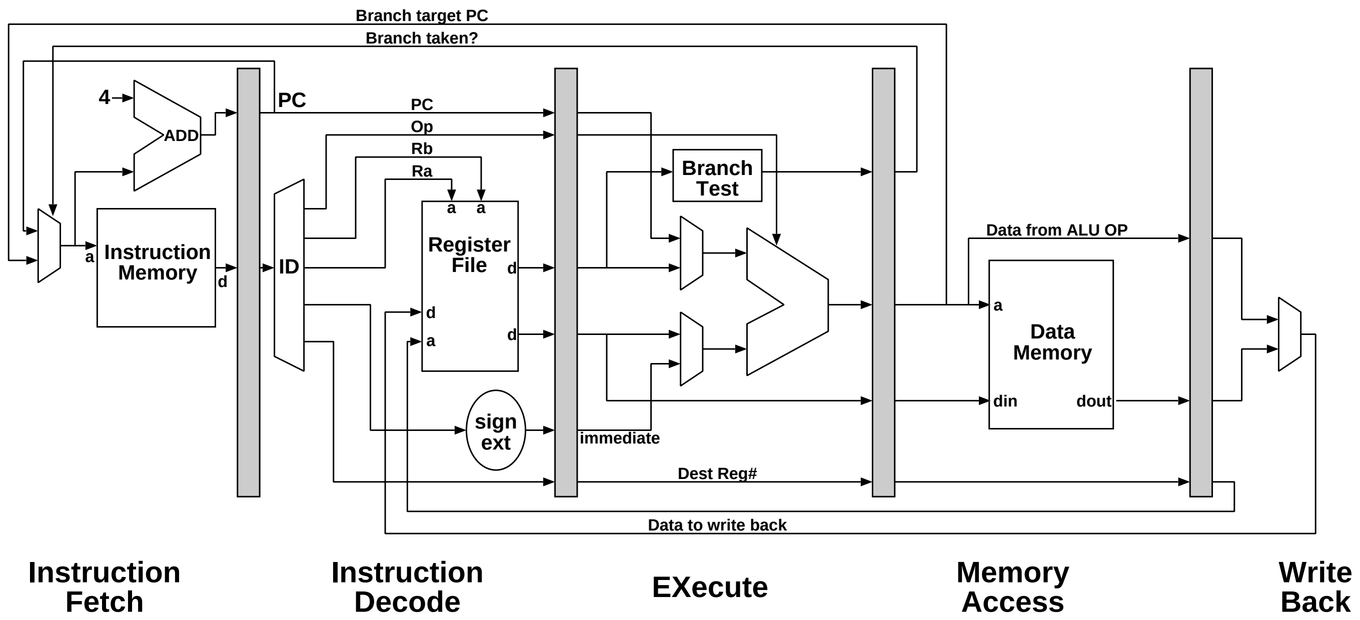
C =Average # Cycles per instruction
T =Time per cycle
I =Instructions per task

- Pipelining
  - e.g. 3-5 pipeline steps (ARM, SA, R3000)
  - Attempt to get C down to 1
  - Problem: stalls due to control/data hazards
- Super-Pipelining
  - e.g. 8+ pipeline steps (R4000)
  - Attempt to decrease T
  - Problem: stalls harder to avoid
- Super-Scalar
  - Issue multiple instructions per clock
  - Attempt to get C below 1
  - Problem: finding parallelism to exploit
    * typically Instruction Level Parallelism (ILP)

CPU Performance Equation

1. *RISC: reduce C and T. PowerPC, ARM & PA-RISC have lower I than most*
2. *super-pipelining: instruction completion latency longer, thus more forwarding paths required*
3. *super-scalar: Typically exploits Instruction Level Parallelism*
4. *other kinds of parallelism too - more later.*
5. *super-scalar: Issuing 1fp and 1int ALU instruction is easy, as there are no inter-dependencies. Issuing multiple int or fp is harder.*
6. *Most modern high-performance processors are super-scalar. The challenge is to build machines capable of searching far ahead into the instruction stream to resolve all the instruction interdependencies and seek out the available parallelism.*
7. *Will be looking closely at the issues surrounding pipelining, and then taking a look at how modern super-scalar CPUs work*

The classic RISC pipe



IF	Send out PC to I-cache. Read instruction into IR. Increment PC.
ID	Decode instruction and read registers in parallel (possible because of fixed instruction format). Sign extend any immediate value.
EX	Calculate Effective Address for Load/Stores. Perform ALU op for data processing instructions. Calculate branch address. Evaluate condition to decide whether branch taken.
MA	Access memory if load/store.
WB	Write back load data or ALU result to register file.

The classic RISC pipe

1. *Basically DLX from H&P's book, with a few mods where it suits my purposes*
2. *Very simple instruction decode. Ra and Rb get accessed regardless of whether they are needed or not.*
3. *Register file has 2 read and 1 write port*
4. *Write Back always occurs at same point for simplicity*

The cost of pipelining

- Pipeline latches add to cycle time
- Cycle time determined by slowest stage
  - Try to balance each stage
- Some resources need to be duplicated to avoid some Structural Hazards
  - (PC incrementer)
  - Multiple register ports (2R/1W)
  - Separate I&D caches

⇒ Effectiveness determined by CPI achieved

Pipelining is efficient

1. *Modern machines need many ports on register file e.g 5R/3W*

Non Load-Store Architectures

- Long pipe with multiple add and memory access stages
  - Lots of logic
  - Many stages unused by most instructions
- Or, multiple passes per instruction
  - Tricky control logic
- Or, convert architectural instructions into multiple RISC-like internal operations
  - Good for multi-issue
  - More ID stages
  - Pentium Pro/II/III (μops)
  - AMD x86 K7 (r-ops)

Pipelining easiest if all instructions do a similar amount of 'work'

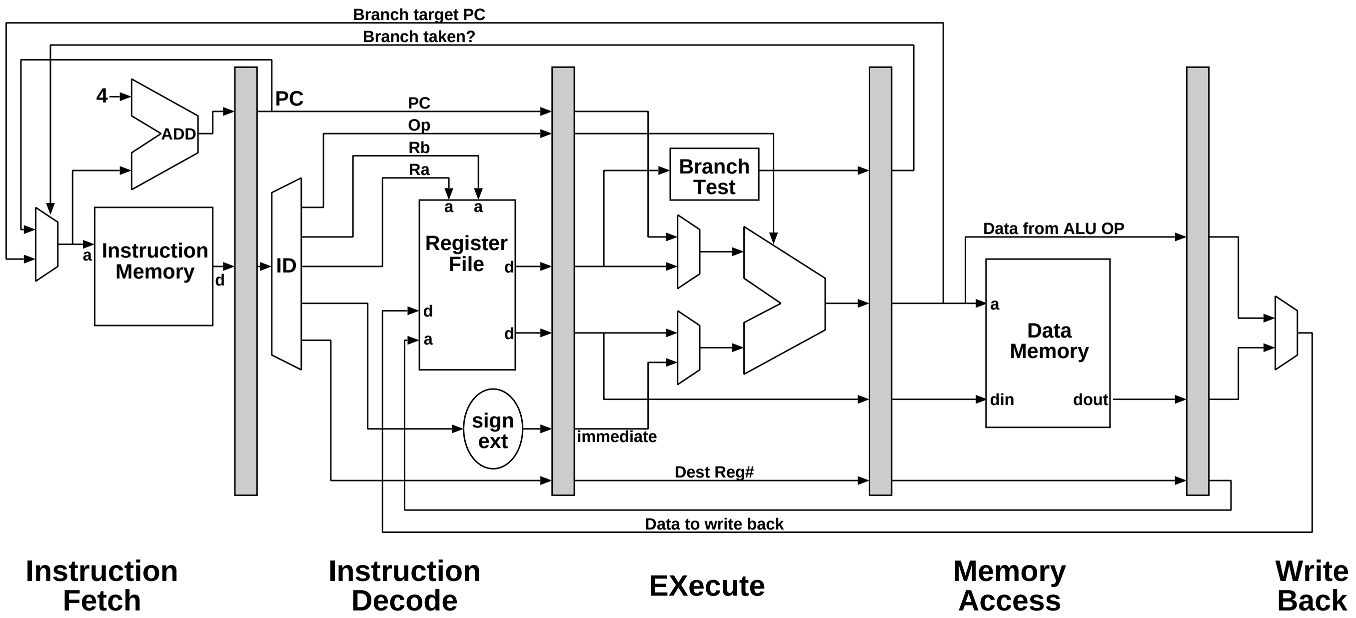
Non Load-Store Architectures

1. *A single-pass pipeline capable of executing a 3 memory operand instruction (eg VAX) would be long and inefficient*
2. *Each x86 instruction can only reference one memory location. On the Pentium, the address is calculated using a dedicated adder in one of the two instruction decode stages. The second instruction decode stage and the write back stage can access the data cache as well as the register file.*
3. *All modern x86 implementations use simpler instructions internally. A decode unit translates the architectural instructions into RISC-like internal instructions, that execute using a RISC-like execution engine. For some particularly complicated x86 instructions, microcode is used to synthesize the necessary internal instruction sequence (eg FP transcendental functions etc).*
4. *Different implementations of the same CISC architecture can use different internal instruction sets. (PPro uses 'u-ops', AMD K6 'r-ops')*
5. *The extra decode stage adds latency to instruction execution, but modern dynamic Out-Of-Order execution engines are relatively effective in hiding this.*
6. *See the Pentium Pro white paper (additional material)*

ALU Result Forwarding

- E.g. 4 forwarding paths to avoid stalls:

```
a:  add r1 ← r8, r9
b:  add r2 ← r1, r7
c:  add r3 ← r1, r2
d:  add r4 ← r1, r2
```



- Read after Write

- Doubled # mux inputs

- Deeper pipes → more forwarding
  - R4000, 8 deep pipe
    forward to next 4 instructions

ALU Result Forwarding

1. *Draw on the 4 forwarding paths*
2. *Too many MUX inputs will slow cycle*
3. *Could move WB stage forward for ALU ops, but create structural hazard*
4. *R4000 8 forwarding paths just for ALU ops*
5. *Adding too many pipeline stages can be counter productive.*

Load Data Hazards

- Impossible without a stall:

```
lw r1 ← r9(4)
add r2 ← r1, r6
```

- Read after Write (RaW) hazard
- New forwarding path saves a cycle
- Re-order code to avoid stall cycle
  - Possible for 60-90% of loads
- Software Interlocked
  - Compiler must insert `nop`
  - e.g. R2000/R3000
- Hardware Interlocked
  - Save `nop`: better for I-stream density
  - Register scoreboard
    - track location of reg values e.g.:
    - File, EX, MA, MUL1, MUL2, MemVoid
    - hold back issue until RaW hazard resolved
    - control operand routeing
  - Required for all sophisticated pipelines
- More stalls for deeper pipes
  - 2 stalls and 2 more forwarding paths for R4000

Load Data Hazards

1. *Draw on forwarding path from WB/dout to ALU a and b*
2. *Re-ordering to avoid stalls is generally more tractable on FP programs*
3. *Maximise number of cycles between load issue and use – modern CPUs won't stall until they actually need the data. If access misses L1 cache, many spare cycles will be needed. Dynamic super-scalar processors (described later) like the PPro are able to search ahead past the instruction that uses the loaded value in order to find other non-dependent instructions which can be executed.*
4. *Register scoreboard tracks the status of all the registers – which are ready to read, which can be forwarded, which are not ready to read yet (issue logic must stall an instruction requiring a register that is not ready)*
5. *Scoreboard is used to determine when instructions can be issued down the pipe, and where they will collect the operands from.*
6. *Even to get 2 stalls for R4K requires speculative use of data returned from D-cache before tag check to see if valid/correct*

Longer Latency Instructions

- Mul/Div, Floating Point
- Different functional units operating in parallel with main pipeline
- Extra problems:
  - Structural hazards
    - Unit may not be fully pipelined, eg:
      - 21264 FDiv: 16cy latency, not pipelined
      - 21164 FMul: 8cy latency, issue every 4cy

      → 21264 FMul: 4cy latency, fully pipelined
    - Multiple Write Back stages
      - more register write ports?
      - or, schedule pipeline bubble
  - Read after Write hazards more likely
    - compiler instruction scheduling
  - Instruction complete out of order
    - Write after Write hazards possible
    - Dealing with interrupts/exceptions
- Use scoreboard to determine when safe to issue
- Often hard to insert stalls after ID stage
  - synthesize NOPs in ID to create bubble
  - 'replay trap' : junk & refetch

Longer Latency Instructions

1. *Multiply latency of 6-16 cycles was common 5 years ago. Now, more silicon has been thrown at the problem...*
2. *No need to fully pipeline a unit unless it is going to be needed by a number of instructions that are in close proximity in the code. Divides occur in 'bunches' relatively infrequently, so there is no need to pipeline the unit.*
3. *21164 multiply unit has 8cycle latency, and can accept a new operation every 4 cycles.*
4. *21264 multiply unit has 4cycle latency and is fully pipelined*
5. *Each functional unit either needs its own write port on the register file (inefficient use of resources), or the issue logic should insure that the WB stage is not in use by the main pipe. This can be achieved either by stalling the pipe at the MEM stage (hard), or by dropping a NOP into the pipe after the ID stage. The NOP will pass through the pipe and will arrive at the WB stage just when the other functional unit wants to use it.*
6. *Read after Write hazards: The compiler should re-order code in order to insert cycles between when an instruction is issued to a high-latency unit, and when the result is first used.(Just as for loads). The register scoreboard will ensure the dependent instruction is held in the ID stage until the result can be made available (send NOPs down the pipe in the meantime).*
7. *Write after Write and Write after Read hazards: Issue logic has to be careful to avoid re-ordering writes to the same register, and to ensure that writes cannot overtake reads. This can be a particular problem if the pipeline contains multiple register write-back points (eg different latency functional units), or if reads can occur late in the pipe (eg. the source operand for a store). Some Write after Write hazards can be solved by aborting the first instruction.*
8. *This is not "full-blown" out-of-order execution, as all instructions are still issued in order. (dynamic execution is described later)*
9. *Replay traps are a common way of recovering after spotting that a hazard is about to occur. The newer instructions are junked and refetched, whereupon the hazard will probably have cleared (if not, repeat). replay traps are slow, but relatively easy to implement.*
10. *exceptions occurring within the pipe can be tricky.... (next slide)*

Exceptions and Pipelining

User SWI/trap	ID	Precise (easy)
Illegal Instruction	ID	Precise (easy)
MMU TLB miss	IF/MA	Precise required
Unaligned Access	MA	Precise required
Arithmetic	EX 1..N	Imprecise possible

- Exceptions detected past the point of in-order execution can be tricky
  - FP overflow
  - Int overflow from Mul/Div
- Exact arithmetic exceptions
  - Appears to stop on faulting instruction
  - Need exact PC
    - care with branch delay slots
  - Roll back state/In-order commit (PPro)
- Imprecise arithmetic exceptions
  - Exception raised many cycles later
  - Alpha: Trap Barriers
  - PPC: Serialise mode
  - IA-64: Poison (NaT) bits on registers
    - instructions propagate poison
    - explicit collection with 'branch if poison'

Exceptions and Pipelining

1. *Illegal instruction: ID stage, so still in-order*
2. *Instruction fetch TLB miss: in order, so no problem*
3. *Data TLB miss/unaligned access: easy on the 'classic RISC pipe', as no other instructions have been committed to register Write Back yet.*
4. *Divide by 0 can usually be detected quite early*
5. *Arithmetic overflow (int and fp pipes): Can be tricky, particularly if they are generated by instructions executing in functional units with long latencies—many subsequent instructions may already have been issued. INT Mul and Divide, FP instructions*
6. *Processors supporting exact exceptions need to appear as though they are executing instructions serially, one after the other. When a fault occurs, the exception handler is given the PC of the instruction that caused the fault. As far as the user is concerned, no instructions after the faulting instruction were executed. On modern processors, this is not what happens in reality — other instructions will have been executed, and there effects must be undone. The PPro achieves this by committing updates to the user-visible architectural state (the register contents) in-order. Care must also be taken to ensure that exceptions are raised in-order too. Exceptions should not be raised from instructions that should not have been executed.*
7. *Watch out for exceptions occurring in Branch delay slots! Need to re-execute the branch.*
8. *Newer RISC architectures have an imprecise exception model for dealing with arithmetic exceptions. When a fault is detected, rather than appearing to stop on the faulting instruction, such architectures simply stop issuing new instructions. Instructions already in the pipe run to completion as normal. The exception is then raised, giving the PC on which execution was stopped as the fault address. The number of instructions between the faulting instruction and the point where execution stopped is potentially undefined. Some of these instructions may have been dependent on the result of the faulting instruction. They may be executed with an undefined value as one of their inputs. This can cause havoc eg. causing an invalid memory location to be accessed, or causing a branch to go to the wrong place.*
9. *Some programs rely on being able to trap and fix-up arithmetic exceptions. This can still be achieved on Alpha and PPC despite their imprecise exception models.*

10. *Alpha has the TRAPB instruction, which prevents all following instructions from being issued until all previous instructions have advanced beyond the point where they could possibly generate an arithmetic trap. TRAPB instructions should be sprinkled around the code to ensure that exceptions are generated early enough such that the fault can be fixed-up. i.e. no destruction of important register or memory contents has occurred in following instructions, the faulting PC can still be worked out (the program may have branched or taken another exception in the meantime).*

11. *In contrast, PPC can be put into a 'serialize' mode, where it generates exact exceptions by being far less aggressive about overlapping instruction execution.*

12. *Operating out of imprecise exception mode can easily loose a factor of 10 in performance.*

13. *On a machine that ISSUES instructions out of order, all exceptions are potentially tricky. (though in actual fact get dealt with by the same h/w mechanism that allows speculative execution across multiple branches)*

14. *IA-64 stores an extra bit with each register called a 'Not A Thing'. Any faulting instruction (e.g. TLB miss, overflow) sets the NaT bit in the destination reg. All instrs that use a source register that is 'NaT' propagate the NaT to the destination register. A special branch instruction can later be used to 'collect' the exception and jump to some fixup code. Having 65 bit wide registers is a slight pain when it comes to spilling them.*

15. *Quick look at another kind of exception, interrupts....*

Interrupts

- Interrupts are asynchronous
- Need bounded latency
  - Real-time applications
  - Shadow registers avoid spilling state
    - Alpha, ARM
- Some CISC instructions may need to be interruptible
  - Resume *vs.* Restart
    - eg. overlapping memcpy
  - Update operands to reflect progress
    - VAX MOVC

1. *interrupts and SWI traps are not generally no problem on RISC – treat like a JUMP instruction being added to the very front of the pipe. Interrupt latency will be determined by how long it takes everything already in the pipe to complete.*
2. *Some CISC instructions would take too long to complete - need to interrupt. But, some instructions can not simply be re-issued from the beginning, as they have had an information destructive effect (an overlapping copy would have this effect). This can be solved if the instruction is defined to update the contents of it's source operands to reflect its progress. It can then simply be re-issued, and will continue where it left off.*
3. *Reset: drop everything, get into a well-defined state.*

Control Flow Instructions

- Absolute *jumps*
  - To an absolute address
    (usually calculated by linker)
  - Immediate / Register modes
  - usage: function pointers, procedure call/return into other compilation modules, shared libraries, switch/case statements
- PC Relative *branches*
  - Signed immediate offset
  - Limited range on RISC
    * Typically same compilation module
      (calculated by compiler)
  - Conditional
- Branch/Jump to Subroutine
  - Save PC of following instruction into:
    * CISC: stack
    * most RISC: special register
    * ALPHA: nominated register
    * IA-64: nominated Branch Reg

1. *(x86 calls all control flow changing instructions jumps)*
2. *Jump subroutine more common than branch subroutine*
3. *Some architectures have conditional jumps as well.*
4. *IA-64 uses special registers to hold PC targets as this can potentially be used to initiate prefetch etc.*

Conditional Branches

- Conditional branch types
  - most: Test condition code flags
    - Z, C, N, V
    - `Bxx label`
  - Alpha/MIPS: Test bits in named reg
    - msb (sign), lsb, 'zero'
    - `Bxx Ra, label`
  - some: Compare two registers and branch
    - `Bxx Ra, Rb, label`
    - (PA-RISC, MIPS, some CISC)
  - IA-64: Test one of 64 single bit predicate regs

- Conditional branch stats (for MIPS and SPEC92)
  - 15% of executed instructions
    - 73% forward (if/else)
    - 27% backward (loops)
  - 67% of branches are taken
    - 60% forward taken
    - 85% backward taken (loops)

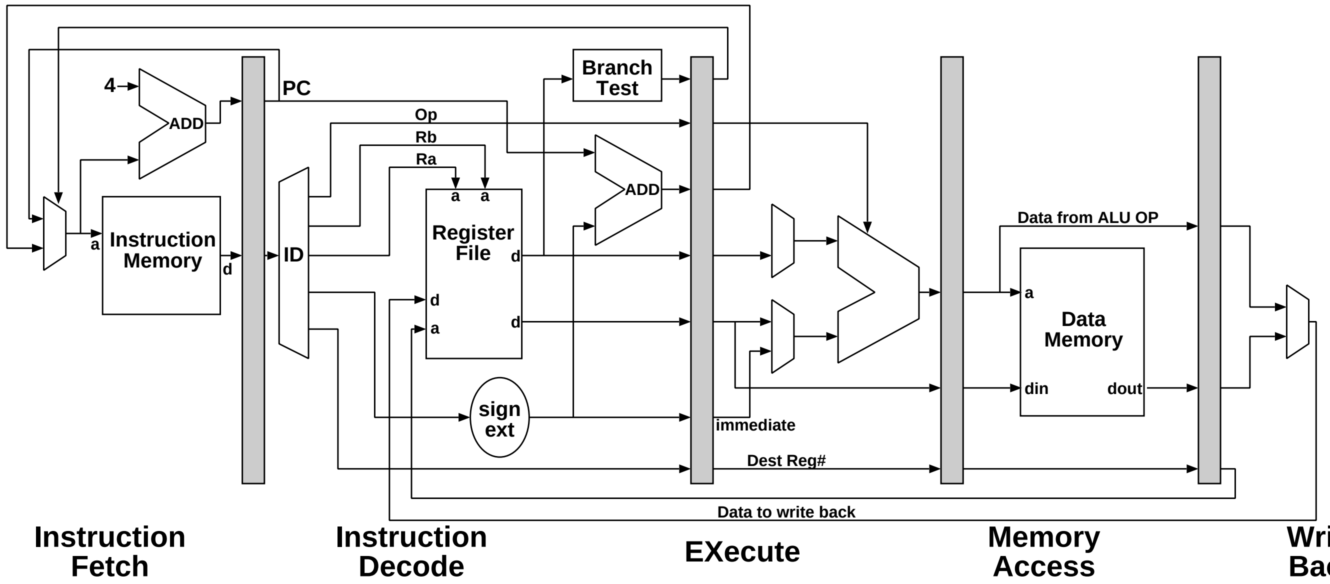
1. *Alpha/MIPS branch instructions can test for the following conditions (and combinations thereof): low bit (set/clear); high bit (positive/negative); value=0 (special zero bit stored with each register)*
2. *In fact, most IA-64 instructions are predicated.*

Control Hazards

- 'classic' evaluates conditional branches in EX
  - Identify branch in ID, and stall until outcome known
  - Or better, assume not taken and abort if wrong

  → 2 stall taken-branch penalty

- If evaluating branch is simple, replicate h/w to allow early decision
  - Branch on condition code
  - Alpha/MIPS: Test bits in named reg
    - ∗ Special 'zero' bit stored with each reg
  - Hard if `Bxx Ra, Rb, label`



1. *register bits: b0, top bit (N), Zero (Implement this by adding an extra bit to each register that indicates whether the current value is zero. This bit should be generated by the ALU and treated like any other register bit, i.e. all forwarding paths should be 33 bits wide)*
2. *On classic, need to abort the following two instructions is branch taken. (Effectively predicts all branches as not taken. However, statistics show the majority of branches are taken)*
3. *compare and branch: PA RISC, MIPS (equal)*

Control Hazards (2)

- Evaluate branches in ID (when possible)

⇒ Only 1 cycle stall if test value ready
(Set flags/reg well before branch)

  - Bad if every instruction sets flags (CISC)
  - Helps if setting CC optional (SPARC/ARM)
  - Good if multiple CC regs (PPC/IA-64), or none (Alpha/MIPS)

- Branch delay slots avoided the taken branch stall on early MIPS
  - Always execute following instruction
  - Can't be another branch
  - Compiler fills slot ∼60% of the time
  - Branches with optional slots: avoid `nop`

- Modern CPUs typically have more stages before EX, due to complicated issue-control logic, thus implying a greater taken-branch cost

- Stalls hurt more on a multi-issue machine. Also, fewer cycles between branch instructions

Control hazards can cripple multi-issue CPUs

Control Hazards (2)

1. *set flags early: compiler should be instruction scheduling.*
2. *better to have* > 1 *CC reg (PowerPC), or use GPRs instead.*
3. *IA-64 has 64 one bit predicate registers that have to explicitly be set by compare instructions. Thus, there is no opportunity for optimizing the compare out as is sometimes possible on Alpha e.g. compare equal zero, or compare less than zero etc*
4. *Branches are common. multi-issue machines will come across them more quickly, and the penalties are typically worse due to the longer front-end to the execution pipe. Control flow hazards could ruin the CPU's CPI. Need to use prediction/speculation.*

Static Branch Prediction

- Speculation should not change semantics!
- Simple prediction
  - e.g. predict backward as taken, forward not
- Branch instructions with hints
  - Branch likely/unlikely
    - strong/weak hint varients
  - Use Feedback Directed Optimization (FDO)
  - Fetch I-stream based on hint
- Delayed branch instrs with hints and annulment
  - If hint is correct execute following instruction else don't
  - e.g. new MIPS, PA-RISC
  - Compiler able to fill delay slot more easily

1. *must be careful about changing semantics: e.g. register writes, stores, exceptions, wasted mem b/w*
2. *If backward branches are predicted taken, forward not, the compiler should endeavour to structure code to assist this.*
3. *likely/unlikely: If loop, likely. Or use FDO for IF/ELSE clauses*
4. *FDO: Many branches turn out to be highly biased one way or tother. Enables miss prediction rates of less than 20%*
5. *hints with annulment: Hint hopefully is usually correct, so do useful work instead of stall. Compiler should still prefer to use non nullifying*

Dynamic Branch Prediction

- Static hints help, but need to do better
- Branch prediction caches
  - Indexed by significant low order bits of branch instruction address
  - Cache entries do not need tags (they're only hints)
  - E.g. 512-8K entries
- Bi-modal prediction method
  - ⇒ many branches are strongly biased
  - Single bit predictor
    - Bit predicts branch as taken/not taken
    - Update bit with actual behaviour
    - Gets first and last iterations of loops wrong
  - Two bit predictors
    - Counter saturates at 0 and 3
    - Predict taken if 2 or 3
    - Add 1 if taken, subtract 1 if not
    - Little improvement above two bits
    - $\geq$90% for 4K entry buffer on SPEC92

1. *Some branch prediction caches are actually implemented within the I-cache, where a number of prediction bits are stored with each cache line.*
2. *Aliasing of entries in the prediction cache will not effect correctness, but will hurt performance.*
3. *NB: The 'counters' are not actually implemented in the cache entries — they're just SRAM bits that are incremented by an adder external to the prediction cache memory array*
4. *2bit predictors: Don't work with branches that are alternately taken/not taken (e.g. odd/even iterations).*

Local History predictors

- Able to spot repetitive patterns
- Copes well with minor deviations from pattern
- E.g. 4 bit local history branch predictor
  - 4 bit shift reg stores branch's prior behaviour
  - 16 x 2 bit bi-modal predictors per entry
  - use shift reg to select predictor to use
  - perfectly predicts all patterns $<$ length 6, as well as some longer ones (up to length 16)
  - used on Pentium Pro / Pentium II
    - 512 entries x (16 x 2 + 4) = 18K bits SRAM
  - trained after two sequence reps
  - other seqs up to 6% worse than random
- An alternative approach is to use two arrays. One holds branch history, the other is a shared array of counters indexed by branch history
  - branches with same pattern share entries in 2nd array (more efficient)
  - 21264 LH predictor: 1024 entries x 10 bits of history per branch, and shared array of 1024 counters indexed by history

Local History predictors

1. *Local prediction: See x86 optimization guide in additional material*
2. *random branches are (on average) mis-predicted up to 6% more often than using chance/bi-modal approaches. This is due to the processor trying to find a sequence where none exists.*
3. *PPro 4 bit history is able to perfectly predict any sequence in which all sub-sequences of length 4 are unique. e.g.*
4. *000011 : 0000=1,0001=1,0011=0,0110=0,1100=0,1000=0*
5. *000001 : 0000=0,0000=1,0001=0,0010=0,0100=0,1000=0 : two sub-sequences that are 0000, so prediction not perfect*
6. *For an N bit history predictor, keeping 2**N 2bit counters per prediction cache entry is rather inefficient. Better to have a shared counter array and index it with a hash of (PC, history). 21264 doesn't bother with hash; they believe that 10bits of history is enough to disambiguate the branch. See section 2-4 of the 21264 manual in the additional material along with the WRL branch prediction technical report*
7. *Local History is a very effective prediction method.*

Global Correlating predictors

- Behaviour of some branches is best predicted by observing behaviour of other branches
- (Spatial locality)

⇒ Keep a short history of the direction that the last few branch instructions executed have taken

- E.g. Two bit correlating predictor:
  - 2 bit shift register to hold *processor* branch history
  - 4 bi-modal counters in each cache entry, one for each possible global history
- Rather than using branch address, some GC predictors use the processor history as the *index* into a single bi-modal counter array. Also possible to use a hash of (branch address, global history)
  - Alpha 21264 GC predictor uses a 12 bit history and 4096 x 2 bit counters
- Combination of Local History and Global Correlating predictor works well
  - $\geq$95% for 30K bit table on SPEC92
  - E.g. Alpha 21264

Global History predictors

1. *Global correlating: See WRL tech report in additional material*
2. *12 bits of global history actually serves to identify individual branches reasonably well*
3. *Multiple if statements often have the same variables in their clauses*
4. *Branch may or may not be part of its own global history*
5. *21264 uses a 'choice predictor' to choose whether the LH or GH predictor will be used for the current prediction. An array of 4096 2 bit saturating counters are used indexed by the 12 bit global branch history. I don't understand why they do this index using the GH rather than the instruction's address, but it obviously works better for some reason or other...*

Reducing Taken-Branch Penalty

- Branch predictors usually accessed in ID stage, hence at least one bubble required for taken-branches

- Need other mechanisms to try and maintain a full stream of useful instructions:

- Branch target buffers
  - In parallel with IF, look up PC in BTB
  - if PC is present in BTB, start fetching from the address indicated in the entry
  - Some BTBs actually cache instructions from the target address

- Next-fetch predictors
  - Very simple, early, prediction to avoid fetch bubbles, used on PPro, A21264
  - I-cache lines have pointer to the next line to fetch
  - Update I-cache ptr. based on actual outcome

- Trace caches (Pentium IV)
  - Replace traditional I-cache
  - Cache commonly executed instr sequences, crossing basic block boundaires
  - (c.f. “trace straightening” s/w optimization)
  - Table to map instr address to position in cache
  - Instrs typically stored in decoded form

Reducing Taken-Branch Penalty

1. *Branch prediction doesn't provide a target PC early enough to avoid a wasted cycle. Next fetch predictors are a simple mechanism to avoid bubbles in the I-fetch stream. The next-fetch predictor is updated based on the 'correct' behaviour. As such it will learn the targets of computed jumps e.g. to DLLs (shared libraries) or function pointers.*
2. *Branch target buffers: for branches that are predicted taken, store a next-fetch target. Some BTB's actually cache instruction(s) from the target address*
3. *Traces caches do a lot more than reduce taken branch penalty. They store a trace-straightened version of common 'hot paths' through the code. A table is required to map instruction addresses to lines in the trace cache as there is no direct relationship anymore. Typically, instructions are stored in a post-decoded form, making control signals available earlier in the cycle. The Pentium IV uses a trace cache to store micro-ops decoded from x86 instructions.*

Avoiding branches

- Loop Count Register (PowerPC, x86, IA-64)
  - Decrement and branch instruction
  - Only available for innermost loops
- Predicated Execution (ARM, IA-64)
  - Execution of all instructions is conditional
    - ARM: on flags registers
    - IA-64: nominated predicate bit (of 64)
  - IA-64: cmp instructions nominate two predicate bits, one is set and cleared depending on outcome
  - E.g. if([r1] && [r2] && [r3]) {...} else {...}

```
ld r4 <- [r1]
p6,p0 <= cmp( true )
p1,p2 <= cmp( r4==true )
<p1> ld r5 <- [r2]
<p1> p3,p4 <= cmp( r5==true )
<p3> ld r6 <- [r3]
<p3> p5,p6 <= cmp( r6==true )
<p6> br else
...
```

  - ✔ Transform control dependency into data dep
  - ✔ Instruction 'boosting'
    - e.g. hoist a store ahead of a branch
  - ✔ Inline simple if/else statements
  - ✘ Costs opcode bits
  - ✘ Issue slots wasted executing nullified instrs

Avoiding branches

1. *Count Register: bit of a throw back to Extended Accumulator*
2. *Count register is decremented by DECB instruction. CPU knows exactly when loop will terminate, and can thus predict the exit correctly. There is only one count register, so it can only be used for inner-most loops (where it really matters).*
3. *predicated execution consumes 4 or the ARM's opcode bits*
4. *predicated execution allows compiler to speculate e.g move instructions in predicted path ahead of the branch. (Sometimes referred to as instruction boosting)*

Avoiding branches 2

- Conditional Moves (Alpha, new on MIPS and x86)
  - Move if flags/nominated reg set
  - Just provides a 'commit' operation
    * beware side effects on 'wrong' path
  - PA-RISC supports arbitrary nullification

- Parallel Compares (IA-64)
  - Eliminate branches in complex predicates
  - Evaluate in parallel
    * (despite predicate dependancy)
  - if ((rA<0) && (rB==-15) && (rC>0)) {...}

```
cmp.eq p1,p0 = r0, r0 ;; // p1 =1
cmp.ge.and p1,p0 = rA,r0
cmp.ne.and p1,p0 = rB,-15
cmp.le.and p1,p2 = rB,10
(p1) br.cond if-block
```

Avoid hard to predict branches

Avoiding branches

1. *CMOV: provides similar sort of functionality to predicated execution, but does not cost opcode bits. Both paths of a simple if/else statement can be evaluated, and then one committed. CMOV is not as flexible as predicated execution. Care has to be taken not to cause side effects on the 'wrong' path e.g. stores etc.*
2. *CMOV instructions can avoid some control flow problems, turning them into data flow issues further down the execution pipe.*
3. *CMOV was added to the x86 architecture on the PPro.*
4. *Parallel compares enable fast evaluation of complex predicates in 'if' statements. They are an exception to the normal IA-64 rule that instructions in the same parallel group must not have data dependencies. A series of cmp instructions in the same parallel group are allowed to issue in parallel with the same destination predicate registers, with the results combined via a logical operation. e.g.*
5. *IA-64 cmp instruction will set either p1 or p2 depending on the outcome of the test. Most instructions can be prefixed with ¡pn¿ to predicate them. IA-64 also allows the compare instructions themselves to be predicated, which is handy for parallel multiway compares.*
6. *In general, predication is most useful for eliminating hard to predict branches — branches that are being predicted accurately are probably more efficient than predication since only a single path has to be evaluated.*

Optimizing Jumps

- Alpha: Jumps have static target address hint
  - A_{16-2} of target instruction virtual address
  - Enough for speculative I-cache fetch
  - Filled in by linker or dynamic optimizer
- Subroutine Call Returns
  - Return address stack
  - Alpha: Push/pop hints for jumps
  - 8 entry stack gives $\geq$ 95% for SPEC92
- Jump target registers (PowerPC/IA64)
  - Make likely jump destinations explicit
  - Buffer instructions from each target
- Next-fetch predictors / BTBs / trace caches help for jumps too
  - Learn last target address of jumps
  - Good for shared library linkage

1. *Alpha jumps: Linker fills in hint same time it fills in word containing real procedure address*
2. *Alpha return addr stack: only enough bits to address I-cache*
3. *next-fetch/BTB: work well if target doesn't change. E.g. shared library address will be a load-time calculated constant*

Super-Scalar CPUs

- # execution units (INT/FP)
  - Pentium (2/1), P6 (2+/2), P7 (4/2)
  - 21164 (2/2), 21264 (4/2)
- Units often specialised e.g 21264:
  - Int ALU + multiply
  - Int ALU + shifter
  - 2x Int ALU + load/store
  - FP add + divide + sqrt
  - FP multiply
- Max issue rate
  - Pentium (2), P6&P7 (3μops)
  - 21164 (4), 21264 (4)
  - Ideal instruction sequence
    - Right combination of INT and FP
  - Lower than number of exec units
- Two basic types
  - Static in-order issue (21164, Pentium, Merced)
  - Dynamic out-of-order execution (21264, PPro)

Super-Scalar CPUs

1. *Function units are specialised according to the expected instruction mix.*
2. *PPRO: 5 more specialised units (2 general int ALU, 2 FP, 1 load/store unit [capable of 1 load and 1 store per cycle])*
3. *The Pentium has two pipelines, U&V. The V pipeline can only execute a subset of (common) operations.*
4. *Achieving the max issue rate requires the right instruction mix, avoiding structural, control and data-flow hazards. In practice, it is rarely sustained. Dynamic execution is better at achieving this than static in-order.*
5. *The current wave of high-performance microprocessors all use dynamic execution.*

Static Scheduled

- All instructions begin execution in-order

1. Fetch and decode a block of instructions

2. 'Slot' instructions to pipes based on function unit capability and current availability

3. Issue checks:

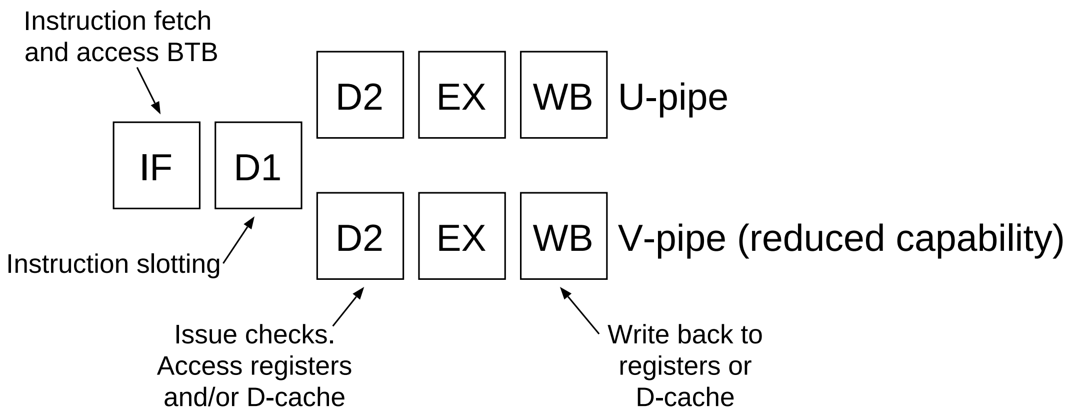
    - Data Hazards
        - Are the instructions independent?
        - Check register scoreboard
            * Are the source operands ready?
            * Will write order be preserved?
        - Non-blocking missed loads
            * Do not stall until value is used
        - *Maintain in-order dispatch*
    - Control hazards
        - Is one of the instructions a predicted-taken branch?
            * Discard instructions past branch
        - Be prepared to squash speculated instrs

4. move onto next block when all issued

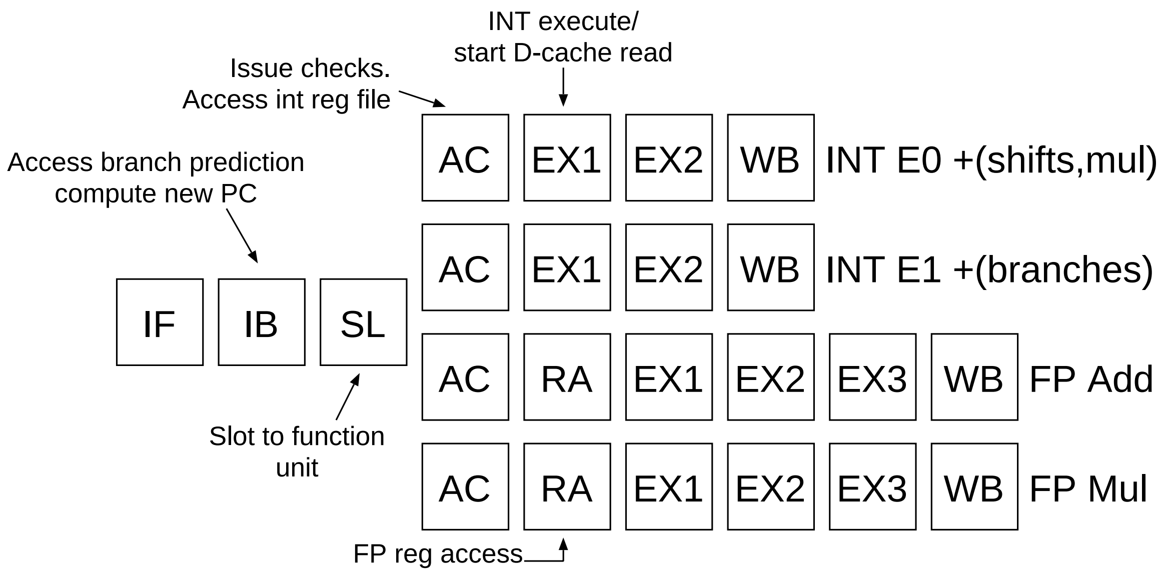
Static Scheduled Super-Scalar

1. *Fetch and decode a block of instructions e.g. the 21164 reads a block of 4 16 byte-aligned instructions.*
2. *Some function units will not be fully pipelined, so will not currently be available for use. Some units are not capable of executing all instructions e.g. the Pentium's V pipe, or the 21164's E1 pipe.*
3. *issuing an Int and FP operation simultaneously is easy, since they generally use different register sets. Doing multiple int is more challenging.*
4. *independent instructions have no producer-consumer (RaW) or producer-producer (WaW) conflicts. If such conflicts exist, issue of the later instruction must be delayed. [WaR conflicts are typically not a problem, except if a stores read their source late in the pipe.*
5. *Non-blocking loads allow instruction issue to continue until the first instruction that references the register (as a source or dest)*
6. *typically, only one control-flow instruction is dispatched in a single cycle. Branch prediction is used to drive the instruction fetch, and to decide whether to discard any subsequent instructions in the current block.*

Static Scheduled Examples



Pentium pipeline



Alpha 21164 pipeline

1. *See additional material*

Static Scheduled Super-Scalar

- Relies greatly on compiler
  - Instruction scheduling
    - slotting
    - data-dependence
  - Issue loads early (or prefetch)
  - Reduce # branches and jumps
    - unroll loops
    - function inlining
    - use of CMOV/predication
  - Align branches and targets
    - avoid wasted issue slots

- Optimization can be quite implementation dependent

- Static analysis is imperfect
  - basic blocks can be reached from multiple sources
  - compiler doesn't know which loads will miss
  - Feedback Directed Optimization can help

$\Rightarrow$ On most code, actual issue rate will be << max

1. *Compiler should try and assemble instructions into issue blocks that cause no structural hazards*
2. *Compiler should try to keep instructions with data-dependencies in different issue blocks*
3. *Branches should ideally be at the end of a fetch block. If fetch blocks need to be aligned (e.g. 16 byte aligned on 21164), branch targets should ideally be the first instruction of the block*

Helping the compiler

- Wish to issue loads as early as possible, but
  - mustn't overtake a store/load to same address
  - Stack / Global variables solvable
    - `[r12,4] != [r12,16]`
  - Heap refs harder to disambiguate
    - `[r2,8] != [r4,32] ???`
    - `C/C++` particularly bad in this respect

⇒ Data speculation (IA-64, Transmeta)

- allows loads to be moved ahead of possibly conflicting load/stores
- `ld.a r3 = [r5]` enters address into Address Aliasing Table
- any other memory reference to same address removes entry
- `ld.c r3 = [r5]` checks entry is still present else reissues load

- Predication enables load issue to be hoisted ahead of branch, but not above compare

⇒ Control speculation (IA-64)

- `ld.s r3 = [r5]` execute load before it is known if it should actually be executed
- `chk.s r3, fixup` check poison bit and branch if load generated an exception

1. *loads to the same address are a problem in multiprocessor systems, as without load-load ordering many locking strategies would break.*

Dynamic Scheduling

- Don't stop at the first stalled instruction, continue past it to find other non-dependent work for execution units
- Search window into I-stream
  - Data-flow analysis to schedule execution
  - Out-of-order execution
  - In-order retirement to architectural state
  - P6 core $\leq$ 30 μops, P7 $\leq$ 126
- Use *speculation* to allow search window to extend across multiple basic blocks
  - (Loops automatically unrolled)
  - Need excellent branch prediction
  - Track instructions so they can be aborted if prediction was wrong
  - Try to make branch result available ASAP (to limit waste caused by mis-prediction)

Dynamic Scheduling

1. *Whereas a static scheduled processor stops issuing instructions as soon as one instruction stalls, a processor with dynamic execution will continue past the stalled instruction and find other non-dependent work to do. This is why they achieve a lower CPI than a statically scheduled CPU with a similar number of function units.*
2. *On wide-issue machines, it is very difficult for the compiler to find instructions to avoid interlocks with load delays or producer-consumer latencies.*
3. *Cost of a miss-predicted branch depends on how many instructions past the branch have been issued.*
4. *Letting the hardware schedule instructions rather than compiler*
5. *May eventually come to a halt due to exhausting ILP in window*
6. *Instruction timing charts just not possible...*
7. *Speculate: 20% branch (CMOV), multiple branch, need damn good prediction, other wise we'll spend the whole time executing the wrong instructions. Worse, we might push the 'right' instructions and data out of the caches.*

Register Renaming

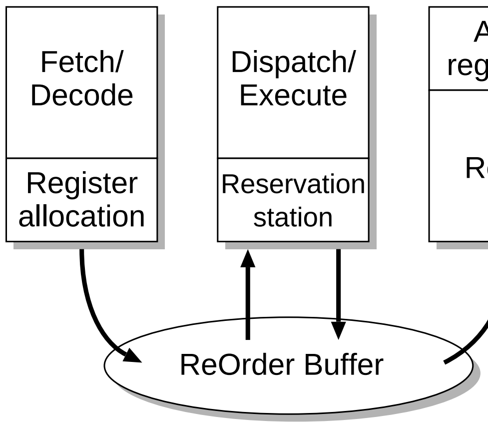
- Register reuse causes false dependencies
- (Often referred to as name-dependencies)
  - WaR, WaW: no data transfer
- Undo compiler's register colouring
- Necessary to unroll loops

⇒ Register renaming
  - Large pool of internal physical registers
  - P6 40, 21264 80+72, P7 128
  - New internal register allocated for the result operand of every instruction
  - Re-mapper keeps track of which internal registers instructions should use for their source operands
    - ∗ needs to be able to rollback upon exception or mispredict

- Architectural register state updated when instructions retire

1. *Compilers use live variable analysis to enable them to re-use registers for different variables. A dynamically scheduled processor needs to use register renaming to undo the register colouring, separating the variables into different internal registers so that the false dependency is eliminated.*
2. *Without register renaming loops could not be unrolled due to name-dependencies*

Out-of-Order Execution



1. Fetch and decode instructions
2. Re-map source operands to appropriate internal registers. Allocate a destination register from register free list. Place instruction in a free Re-Order Buffer (ROB) slot.
3. Reservation station scans ROB to find instructions for which all source operands are available, and a suitable execution unit is free (Favour older instructions if multiple ready)
4. Executed instructions and results are returned to the ROB (internal registers which are no longer needed are placed on free list)
5. Retire unit removes completed instructions from ROB in-order, and updates architecturally visible state. Detect exceptions & mis-predicted branches; Roll-back ROB contents and mapping register state, start fetch from new PC

1. *Stop issuing if there are no ROB slots, or no free registers*
2. *Try to execute older instructions first*
3. *There are lots of variations on the above algorithm dependent on what data structures are actually implemented. E.g. the 21264 still has a 'traditional' file of internal registers. In contrast, the PPro forwards results from instructions direct into the reservation stations and re-order buffer. (See H&P and addition material in notes.)*

Loads and Stores

- Dyn Exec helps hide latency of L2/L3 cache
  - Find other work to do in the meantime
  - Allow loads to issue early
- Stores cannot be undone
  - ⇒ Update memory in Retirement stage
  - Hold in Store Queue until retirement
- Loads that overtake stores must be checked to see if they refer to the same location (alias)
  - Address of store may not yet be known
  - ⇒ Speculate load and check later:
  - Load Queue stores addresses of issued loads until they retire
  - when a store 'executes' (target address is known) it checks the LQ to see whether a newer load has issued to the same address
  - if so, execution is rolled back to the store instruction (replay load)
    - ∗ 21264 has 32 entry LQ and a 1024 entry prediction cache to predict which loads to 'hold back' and thus avoid replay trap
- Loads overtaking loads treated similarly to maintain ordering rules with other CPUs/DMA

1. *A miss to main memory takes so long that its highly unlikely that the processor will be able to hide all the latency.*
2. *loads following stores can be serviced from the SQ.*
3. *Allowing loads to overtake other loads and stores is important for performance.*
4. *Current processors don't re-ordering stores, or let stores pass loads. In order to extract more ILP future processors will have to consider these issues.*

Out-of-order Execution

- Less dependency on compiler than static-sched
- Better at avoiding wasted issue slots
- But, O-o-O execution uses a lot of transistors
  - ReOrder Buffer and Reservation Stations are large structures incorporating lots of Content Addressable Memory
  - Tend to be at least $O(N^2)$ in complexity
  - Tend to be on critical path
    - diminishing returns...
  - 20%+ of chip area on 21264
- Factors effecting usable ILP
  - Window size
  - Number of renamed registers
  - Memory reference alias analysis
  - Branch and jump prediction accuracy
  - Data cache performance
  - (Value speculation performance)
- Simulation suggests the 'perfect' processor: 18-150 instructions per cycle for SPEC92
- 10 way for int progs feasible, more for FP
- Some code just exhibits very poor ILP...

Out-of-Order Execution

1. *ILP - Instruction Level Parallelism*
2. *The perfect processor assumes infinite window size, infinite number of renamed registers, perfect mem ref analysis, perfect prediction and a 100% D-cache hit rate.*
3. *A real 10 way INT processor would need something like: 100 instr window, 100 registers*
4. *Chasing down long linked lists is just very bad news...*
5. *Things we're likely to see in the future include: value prediction (predicting the outcome of a load based on PC rather than load address), and split-path execution (executing both paths of branches that are proving difficult to predict. This reduces the cost of a mispredict).*
6. *value prediction : predict instruction result based on PC. Helps reduce inter-depenedencies*

VLIW Architectures

- Very Long Instruction Word (VLIW)
- Each instruction word (or 'packet') contains fields to specify an operation for each function unit
- Compiler instruction scheduling:
  - allocates sub-instructions to function units
  - avoids any resource restrictions
  - ensures producer-consumer latencies satisfied (delay slots)

✔ CPU doesn't need to worry about issue-checks

$\Rightarrow$ High clock speed

✘ Relies heavily on compiler / assembler programmer

- loop unrolling
- trace scheduling

✘ Stall in any function unit causes whole processor to stall

- D-cache misses a big problem

✘ Often sparse I-stream (lots of nops)

✘ Exposes processor internals

- Typically no binary compatibility

1. *CPU can issue all of the sub-instructions without issue checks—the compiler has guaranteed them to be independent.*
2. *trace scheduling : select likely execution path (across multiple basic blocks) then optimise trace as a whole.*
3. *VLIW now somewhat out of fashion, except with DSPs and alike. This is perhaps due to the importance of binary compatibility, and realization that static scheduling by compilers is hard.*

Intel EPIC (VLIW-like)

- Intel: Explicitly Parallel Instruction Computer
  - Merced (Itanium) , McKinley
- Three 41 bit instrs packed into 128 bit 'bundle' with 5 template bits
- Template determines function unit dispatch
  - restricted set of possibilities simplifies instruction dispersement hardware
    - e.g. `[Mem,Int,Branch]`, no `[Int,Int,Int]`
- Stop bits: barriers between independent instructions groups
  - groups can cross multiple bundles
- Compiler collects instrs into independent groups
- Hardware interlock of longer-latency instructions as well as load-use latencies

✔ Reduces issue-check complexity for CPU

✔ Retains binary compatibility

- Need good compilers
  - hope extensive use of load speculation instructions enables hoisting of loads to avoid stalling whole CPU
- Optimization for new implementations important?

Intel EPIC (VLIW-like)

1. *IA-64 puts far greater onus on the compiler, just like in the early days of RISC. They're banking on compiler technology being good enough to schedule the instructions...*
2. *Only a very restricted set of the potential templates are valid. Memory accesses always go in slot 0, (and slot 1 if there are two). There can be at most 2 integer (or memory) ops in every bundle, and at most one FP*
3. *Optimising compilers may try and group instructions according to the number of function units its knows the current generation of processor has. If a future processor increases this number such code may execute rather poorly. Ideally the compiler should create independent groups that are as big as possible, but this may not be optima for the current hardware...*

Transmeta ‘Code Morphing’

- VLIW core hidden behind x86 emulation
- Uses combination of interpretation, translation and on-line feedback-directed optimization
- Only ‘code morphing’ s/w written for VLIW
  - Apps, OS and even BIOS are x86
- Keeps an in-memory translation cache
- Translate and optimise along frequently executed paths (trace scheduling)
  - speculative load instrs increase trace length
- Hardware features to assist translation:
  - Shadow registers with commit instruction
    - assist rollback upon x86 exceptions/mispredicts
  - hold-back stores until commit
- Performance counters assist re-optimization

✔ Binary compatibility, High clock speed, Low power

✔ Potential for more complex scheduling than h/w

✘ Overhead of performing translation

✘ Less dynamic than h/w scheduling

Transmeta 'Code Morphing'

1. *Transmeta reckon their current code morphing engine runs in 16MB, including the translation cache.*
2. *Has to detect writes to x86 code space and invalidate translations appropriately.*
3. *'gated store buffer' : hold back stores until commit instruction*
4. *Low power due to simple hardware*
5. *Apparently, the VLIW core is quite 'general purpose' and isn't particularly x86 specific.*
6. *Rather than doing this with an x86 ISA, perhaps we should be using a higher-level, typed, intermediate form???*

Beyond ILP

- Diminishing returns for further effort extracting ILP from a single thread?

- *System-level* parallelism
  - some workloads naturally parallel
    * multi-user machine
    * application plus XServer
    * application plus asynchronous I/O

- *Process/Thread-level* parallelism
  - Some applications already multithreaded
    * database, HTTP server, NFS server
    * fork, pthreads
  - may have smaller cache footprint
  - may be same Virtual address space

- *Loop-level* parallelism
  - generated by auto-parallelizing compilers
  - co-operative threads
  - need fast synchronization, communication, fork

Beyond ILP

1. *Auto-parallelisation far more attractive if locking/synchronization and communication overheads are cheap*
2. *Use of traditional SMP mechanisms is good for 'ease of use', but you may wish to have something more efficient to increase the class of applications that see benefit.*
3. *'fork' means 'create thread' in this context, not full Unix semantics.*

Exploiting Parallelism

- Multiple CPUs on a chip
  - Exploit thread/process level parallelism
  - Use traditional SMP mechanisms
  - ✘ Need correspondingly bigger caches and external memory bandwidth
  - IBM Power4 2-way SMP on a chip

- **Multi-threading**
  - Use one CPU to execute multiple threads
  - Replicate PCs, architectural register file
  - Different virtual address spaces?

- Static multi-threading
  - Round-robin issue from a large # threads
  - ✔ No instruction dependencies
  - ✔ Hides memory latency
    - ∗ No expensive caches
  - ✔ Fast synchronization / fork possible
  - ✘ Requires many register files
  - ✘ Progress of an individual thread is slow
    - ∗ Poor SPEC marks (great SPEC Rate)
  - Tera/Cray MTA, 128 threads

- Course-grained multi-threading
  - Switch between threads on a major stall
  - e.g. cache miss on Stanford SPARCLE

Beyond ILP

1. *IBM have released a 2way SMP-on-a-chip PowerPC. Other manufacturers are rumoured to be following suit.*
2. *(separate L1s, pooled L2s possible)*
3. *Multi-threading relies on having lots of thread-level parallelism in the application. Auto-parallelising compilers are supposed to seek this out and create light-weight threads. The TERA machine supports many threads and executes one instruction from each in turn. In contrast, the SPARCLE processor (not a true M-T) concentrates on a single thread until an L1 cache miss, which causes it to change thread.*

Simultaneous Multi-Threading (SMT)

- Work on a small number of threads at once, aiming to keep all function units busy
- Duplicate architectural state
- Duplicate instruction fetch units
- Need to control allocation of resources
  - priority . fair share
  - (prioritising can be counter productive)

✔ Progress of individual threads is pretty good

✔ Cooperating threads may have smaller cache footprint than independent ones

✔ Potential for register-register synchronization and communication

✔ Potential for lightweight thread create

- Pentium IV Xeon uses 2-way "hyperthreading"
  - 2 virtual CPUs per chip
  - looks like SMP - separate VM contexts
  - Staticly partitions resources if both active
  - SMT `halt` and `pause` instructions
  - OS scheduler should understand SMT

1. *Xeon uses a fixed allocation of resources between threads, but executes based on src operand availability.*
2. *5% increase in core size to store extra state. Rumoured to have been on P IV since day one put hidden...*
3. *Xeon does not have lightweight sync primitives.*
4. *See article in additional material*
5. `http://www.cs.washington.edu/research/smt/`

Other techniques

- Data-flow processors
  - Fine-grained control-flow, course-grained data-flow (opposite of standard super-scalar)
  - Begin execution of a block of sequential instructions when all inputs become available
  - ✘ Inputs are memory locations. The matching store required to figure out when all inputs are ready is large and potentially slow. (matching is easier with a small number of registers *a la* out-of-order execution)

1. x

Caching

- Caches exploit the temporal and spatial locality of access exhibited by most programs

- Cache equation:

$$Access\ Time_{Avg} = (1 - P) * Cost_{Hit} + P * Cost_{Miss}$$

  Where P consists of:

  - Compulsory misses
  - Capacity misses (size)
  - Conflict misses (associativity)
  - Coherence misses (multi-proc/DMA)

✘ Caches can increase $Cost_{Miss}$

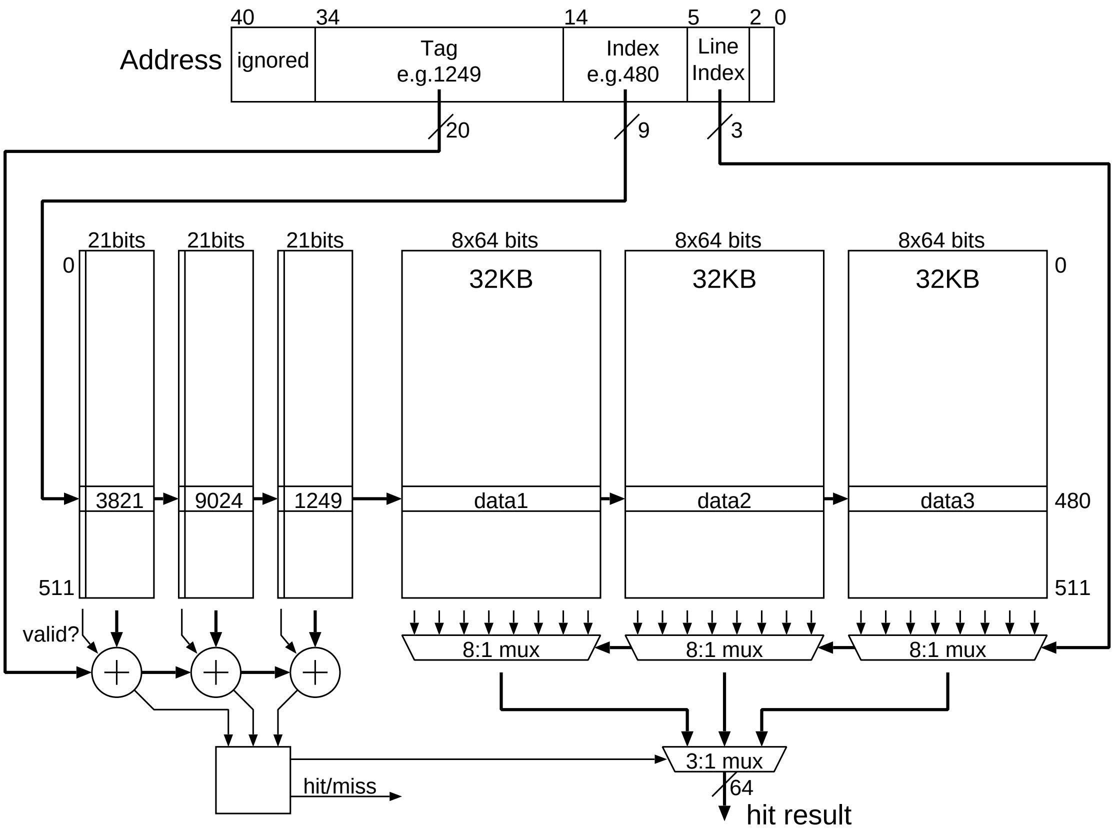
- Build using fast (small and expensive) SRAM

- Tag RAM and Data RAM

1. *Spatial locality is exploited through fetching a block of data rather than the word requested.*
2. *Compulsory misses occur when a block is requested for the first time. The number of compulsory misses is independent of cache size and associativity. However, increasing the line size may reduce the number of compulsory misses.*
3. *Capacity misses: was in cache once, but was replaced.*
4. *Conflict misses are highest for a direct mapped (1-way) cache, and zero for a fully associative cache.*
5. *Most modern system architectures demand hardware cache coherence for both multiprocessor and I/O transactions.*
6. *Tag ram and data ram are typically implemented in physically separate blocks. (e.g. see the 21164 floor plan in the additional material)*

Associativity

- Direct Mapped (1-way, no choice)
  - potentially fastest: tag check can be done in parallel with speculatively using data
- *n*-way Set Associative (choice of *n* e.g. 2/4/8)
- Fully associative (full choice)
  - many-way comparison is slow



A 96KB 3-way set associative cache with 64 byte lines
(supporting 2^{35} bytes of cacheable memory)

Associativity

1. *Beware terminology: An n-way cache has size/(n*line size) sets, not n.*
2. *Associative caches require extra gate delays due to the multiplexors to select between the different banks. Data cannot be speculatively forwarded to the CPU in parallel with the tag lookup, since the tag comparisons are necessary to select the appropriate block.*
3. *When associative caches are implemented external to the CPU, implementing the multiplexors can be troublesome due to the large number of input pins. Instead, a tristate bus is often used, but this can impair the cache access frequency.*
4. *Some systems do not have wide enough tag fields to enable the entire physical address space to be cached. Many Pentium motherboards would only cache the first 64MB of physical memory; references beyond this forced to miss.*

Replacement Policy

- Associative caches need a replacement policy
- FIFO
  - ✘ Worse than random
- Least Recently Used (LRU)
  - ✘ Expensive to implement
  - ✘ Bad degenerate behaviour
    - * sequential access to large arrays
- Random
  - – Use an LFSR counter
  - ✔ No history bits required
  - ✔ Almost as good as LRU
  - ✔ Degenerate behaviour unlikely
- Not Last Used (NLU)
  - – Select randomly, but NLU
  - ✔ $log_2 n$ bits per set
  - ✔ Better than random

1. *LFSR: Linear Feedback Shift Register (pseudo random number generator)*
2. *Random replacement requires no per line or per set, and performs rather well. NLU is potentially better, and still relatively simple—just a single pointer to the last line used in each set.*

Caching Writes

- Write-Back *vs.* Write Through
- Read Allocate *vs.* Read/Write Allocate
- Allocate only on reads and Write-Through
  - Writes update cache only if line already present
  - All writes are passed on to next level
  - Normally combined with a *Write Buffer*
- Read/Write Allocate and Write-Back
  - On write misses: allocate and fetch line, then modify
  - Cache holds the only up-to-date copy
  - Dirty bit to mark line as modified
  - ✔ Helps to reduce write bandwidth to next level
  - Line chosen for eviction may be dirty
    - *Victim writes* to next level
    - e.g. write victim, read new line, modify

Caching Writes

1. *Allocation policy and write through/back are orthogonal parameters, but in general, only Read-only Allocate combined with Write-Through and Read-Write Allocate combined with Write-Back are useful.*
2. *Write back caches reduce the number of stores seen by the next level (the next level may only ever see victim writes). This is essential in multiprocessor systems to reduce traffic on the interprocessor bus, but requires a coherence mechanism.*
3. *Write through caches avoid some of the coherence problems of write back caches, and are generally simpler. Write-through is generally the desired behaviour when communicating with memory mapped peripherals e.g. frame buffers. For this reason, many architectures allow control of cache-ability on a per page basis e.g. the x86 allows pages to be marked as un-cacheable, write-back cacheable and write-through cacheable.*
4. *Write through can actually be a win over write back for certain operations. E.g. Consider scanning over a large array initialising it to zero. The initial line fetch will be a waste of bandwidth, since all the words will be updated. Some architectures provide a write-line invalidate instruction that allocates a line in the cache without the initial fetch; the program guarantees to update all words in the line. Alternatively, some architectures provide store instructions that force write-through behaviour.*

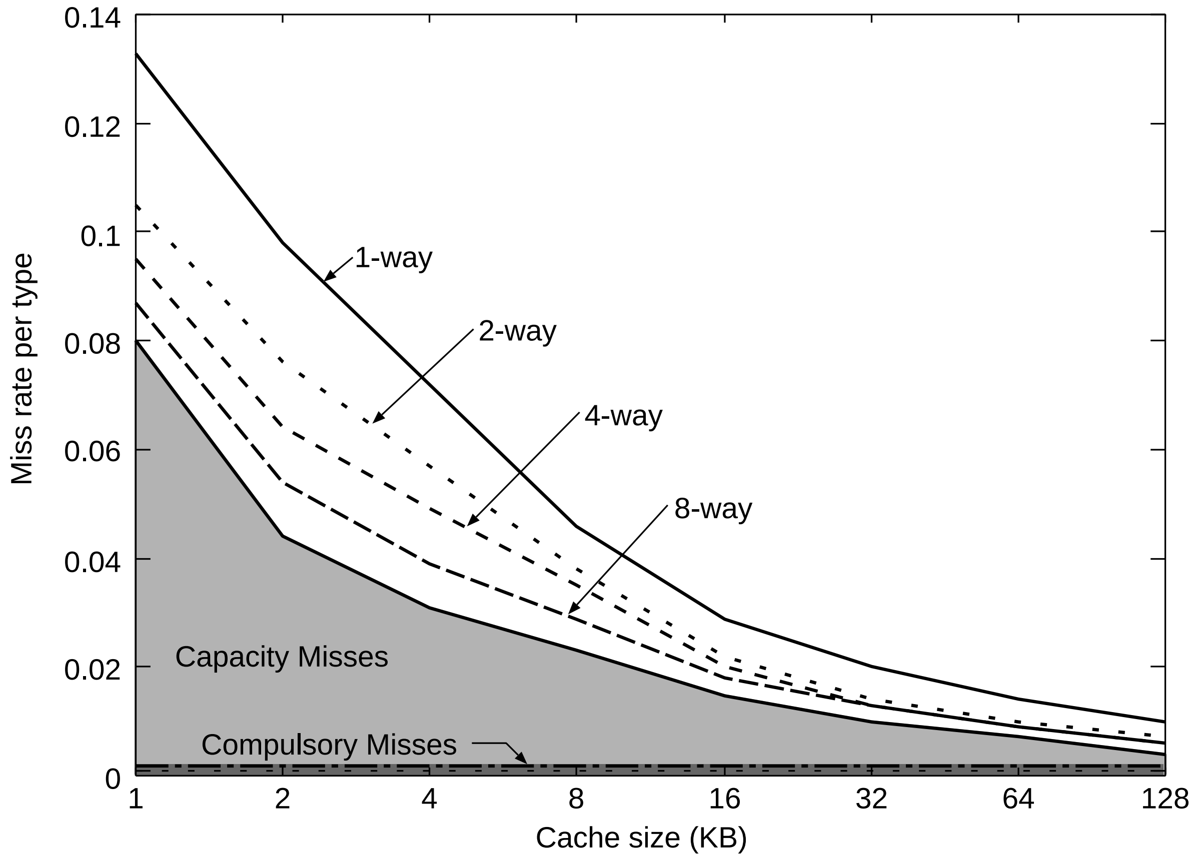
Write Buffers

- Small high-bandwidth resource for receiving store data
- Give reads priority over writes to reduce load latency
  - All loads that miss must check write buffer
  - If RaW hazard detected:
    - flush buffer to next level cache and replay
    - or, service load from buffer (PPro, 21264)
- *Merge* sequential writes into a single transaction
- *Collapse* writes to same location
- Drain write buffer when bus otherwise idle
- 21164: 6 addresses, 32 bytes per address
- ARM710: 4 addresses, 32 bytes total

1. *Write buffers accept store data from the CPU core at full-rate, and attempt to hide the lower throughput of the underlying memory system. By allowing loads to pass stores they also result in reduced average load latency.*
2. *On a cache miss, the write buffer must be examined to see whether it contains a write to the requested load address. In theory, a write buffer could service the requested load address directly, but in practise, the write buffer is normally flushed to the next level of the memory hierarchy and the load replayed.*
3. *If a write buffer becomes full, the next store will cause a stall until a slot becomes available. The 21164 write buffer has slots for six addresses and 32 data bytes per address. Other current CPUs are similar.*

Cache Miss Rate E.g.

- SPEC 92 on MIPS
- 32 byte lines
- LRU replacement
- Write allocate/write back



A direct-mapped cache of size N has about the same miss rate as a 2-way set-associative cache of size $N/2$

1. *Compulsory miss rate is independent of size or associativity.*
2. *The Capacity miss rate was calculated by considering a fully associative cache of size n.*
3. *Little advantage is gained with associativity greater than 4-way for caches >16KB (most current generation CPUs have 16KB or more L1 caches)*

L1 Caches

- L1 I-cache
  - Single-ported, read-only (but may snoop)
  - Wide access (e.g. block of 4 instrs)
  - (trace caches)
- L1 D-cache
  - Generally 8-64KB 1/2/4-way on-chip
    - Exception: HP PA-8200
  - Fully pipelined, Low latency (1-3cy), multi-ported
  - Size typ constrained by propagation delays
  - Trade miss rate for lower hit access time
    - May be direct-mapped
    - May be write-through
  - Often virtually indexed
    - Access cache in parallel with TLB lookup
    - Need to avoid virtual address aliasing
      - Enforce in OS
      - or, Ensure index size $<$ page size (add associativity)

L1 Caches

1. *All current high-performance processors have separate L1 I&D caches—It's much better to have separate caches rather a single double size cache with an extra read port.*
2. *x86 I-cache snoops D-cache (as does the rest of the pipeline)*
3. *Some new processors employ 'trace caches'. Rather than storing the original ISA instructions they store instructions in a pre-decoded internal form. Some trace caches are even laid out according to the predicted execution path i.e. the contents of cache lines may not be from sequential ISA instructions.*
4. *Dynamic execution can hide D-cache latency better, enabling larger more complex caches with longer latency but lower miss rate.*
5. *A few years ago L1 cache size may have been constrained by available die area, but now performance is generally the limiting factor.*
6. *The exception is the PA-RISC machines, which in the past have had large off-chip L1 caches. HP's assertion has been that their execution engines have been able to hide the load-use latency, and that their machines perform better on 'commercial applications' that have worse locality of reference than SPEC95.*
7. *Multi-porting L1 D-caches is often emulated by building the cache out of two or more banks that operate independently. Providing concurrent requests need to access different banks they can proceed in parallel*
8. *Cache complexity (and thus miss rate) is often traded for hit access time. Direct mapped caches are often used to enable the tag check to be done in parallel with speculatively feeding the data to the CPU (hence reducing load-use penalties). Modern dynamic-issue CPUs tend to be less sensitive to such stalls, thus the complexity of L1 caches is increasing. (see the point below also—adding associativity can help to keep the cache virtually indexed without special OS support)*
9. *Address aliasing in a virtually indexed cache occurs when two VAs which map to the same physical address map to a different set in the cache. The OS can prevent this by ensuring that the virtual to physical mappings of all pages have sufficient low-order address bits in common i.e. log2(cache size/associativity). Keeping pages so aligned also has the benefit of allowing a competent programmer to have*

more control over where his or her data is placed in the cache.

10. *Alternatively, the aliasing problem can be solved in hardware by adding sufficient associativity to the cache to reduce the number of virtual address bits used as the index until there are no more than log2(page size) (since the VA and PA must be identical within a page). E.g. x86's have a 4KB page size. The Pentium Pro has an 8KB 2-way L1 D-cache, the Pentium II a 16KB 4-way.*

Enhancing Performance :1

- Block size (Line size)
  - Most currently 32 or 64
  - ✔ Increasing block size reduces # compulsory misses
  - ✔ Typically increases bandwidth
  - ✘ Can increase load latency and # conflict misses

- Fetch critical-word-first and early-restart
  - Return requested word first, then wrap
  - Restart execution as soon as word ready
  - ✔ Reduces missed-load latency
  - Widely used. Intel order *vs.* linear wrap

- Nonblocking caches
  - Allow hit under miss (nonblocking loads)
  - Don't stall on first miss: allow multiple outstanding misses
    - merge misses to same line
  - Allow memory system to satisfy misses out-of-order
  - ✔ Reduces effective miss penalty

Enhancing Performance :1

1. *High memory system bandwidth is typically very important for FP programs. In contrast, integer programs often benefit more from reduced latency. (Sequential vector (array) operations vs. pointer chasing (e.g. scanning linked lists))*
2. *Many systems use 'linear wrap', where the word requested (signalled by the low-order address bits) is returned first, and then cache fill proceeds sequentially, wrapping at the end of the block back to the beginning. Rather than linear wrap, Intel use a bizarre ordering that is supposed to optimize the likelihood that the subsequent word requested by the CPU is the second word in the fill order.*
3. *Most memory systems now allow transactions to be serviced out-of-order. The 21264 is particularly aggressive, allowing up to 16 transactions.*

Enhancing Performance :2

- Victim caches
  - Small highly associative cache to backup up a larger cache with limited associativity
  - ✔ Reduces the cost of conflict misses
- Victim buffers
  - A small number of cache line sized buffers used for temporarily holding dirty victims before they are written to L_{n+1}
  - Allows victim to be written *after* the requested line has been fetched
  - ✔ Reduces average latency of misses that evict dirty lines
- Sub-block presence bits
  - Allows size of tag ram to be reduced without increasing block size
  - Sub-block dirty bits can avoid cache line fills on write misses
    - (would break coherence on multiprocessors)

Enhancing Performance :2

1. *Victim caches are employed on some systems with large off-chip direct-mapped L1 caches. Adding a normal L2 cache would not help much (the L1 is already big), but the extra associativity of the victim cache can help. (E.g. The HP PA-7200)*
2. *The contents of victim buffers can be written out when the memory system is otherwise idle. If the victim buffers are already full when a dirty line is evicted, the victim must be written before the fill can occur. Victim buffers don't help if large numbers of victims are being generated, e.g. by a program performing sequential writes to a very large array (memset).*
3. *As with write buffers, missed loads must check the contents of the victim buffer before proceeding.*
4. *Victim buffers are particularly useful if the latency of the Ln+1 memory system is high. (Alternatively, some systems use an address routing trick to ensure that both the victim and the line that is to be fetched must come from the same DRAM page, reducing the total transaction time.)*
5. *Another way of enhancing performance is to use a L2 (or L3) cache to reduce average miss time...*

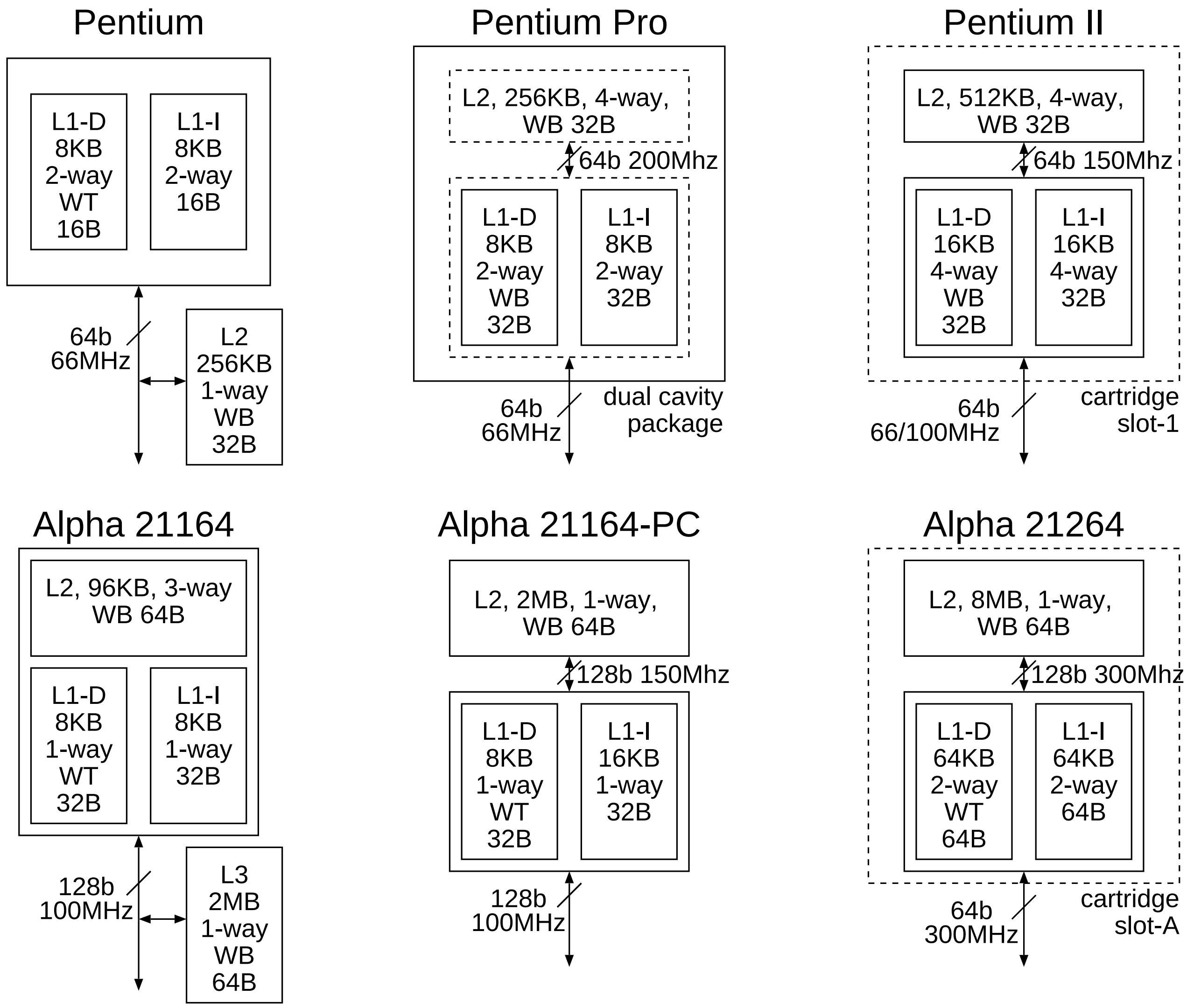
L2 caches

- L2 caches help hide poor DRAM latency
  - large write-back cache
- L2 caches used to share the system bus pins (e.g. Pentium)
  - ✘ electrical loading limits performance
- now, a dedicated ‘backside bus’ is used
- L2 on same die (21164)
  - ✔ low latency and wider bus
  - ✔ associativity easier
  - ✘ limited die size, so may need an L3
    - $*$ (e.g. 21164 has 2-16MB L3)
- L2 in CPU package (Pentium Pro)
  - ✔ lower latency than external
- L2 in CPU ‘cartridge’ (Pentium II)
  - ✔ controlled layout
  - ✔ use standard SSRAM
- L2 on motherboard
  - ✘ requires careful motherboard design
- L1/L2 inclusive *vs.* exclusive

L2 caches

1. *Large write-back L2 caches are essential for good SPEC95 results*
2. *Unless an L3 cache can be made substantially (e.g. >16x) bigger than the L2 it is unlikely to be of much benefit (it may even increase the main memory access time).*
3. *The backside bus is typically not a bus, but a point-to-point link between CPU pins and cache RAM pins. The point-to-point links typically use high speed low voltage signalling and care is taken to control skew and propagation delays.*
4. *The dual cavity packaging approach used on the Pentium Pro is expensive to manufacture, hence the Pentium II. However, PPro cache runs at core speed compared to the PII's half speed. (Xeon back to full speed)*
5. *Intel introduced cartridge to avoid the burden of cache layout falling to motherboard designers (and to maximise their own profits...)*
6. *Older caches used 'wave pipelining' to maximise throughput. This is effectively using propagation delay to provide asynchronous pipelining. Now, modern Synchronous SRAMs (SSRAMs) have inputs and outputs synchronous to a clock input, and are pipelined internally.*
7. *Some systems use an inclusive L1/L2 policy whereby all lines that are in the L1 must be in the L2. Hence, SMP coherency probes only need to look in the L2. However, the effective cache size is improved if there is no inclusivity rule.*

Hierarchy Examples



1. *Conventional system buses are 64-128 bits wide and run at 66-100MHz. Bus signals have multiple electrical loads. New CPUs (e.g. the 21264 and AMD K7) will have point-to-point system 'buses' clocked at >150MHz. 'Switch chips' will be used to provide the connections to the I/O and memory sub-systems. (This is done for the same reasons as having a dedicated backside bus for the external cache).*
2. *Next generation chips like the 21364 are implementing the memory system controller on-chip. Rambus strings will hang directly off the CPU, and CPUs will directly be connected together by virtue of four high-speed 'network connections' on each processor.*

Performance Examples

Ln	size	n-way	line	write	lat(cy)	lat(ns)	ld(MB/s)	st(MB/s)
L1	8KB	1	32	WT	1	4	1205	945
L2	96KB	3	64	WB	7	26	654	945
L3	2048KB	1	64	WB	22	83	340	315
MM	-	-	-	-	96	361	140	113

266MHz 21164 EB164 (Alcor/CIA)

Ln	size	n-way	line	write	lat(cy)	lat(ns)	ld(MB/s)	st(MB/s)
L1	8KB	2	16	WT	1	5	634	82
L2	256KB	1	32	WT	11	55	193	82
MM	-	-	-	-	28	140	123	81

200MHz Pentium 430HX

Ln	size	n-way	line	write	lat(cy)	lat(ns)	ld(MB/s)	st(MB/s)
L1	8KB	2	32	WB	2	10	695	471
L2	256KB	4	32	WB	6	30	426	426
MM	-	-	-	-	44	220	179	87

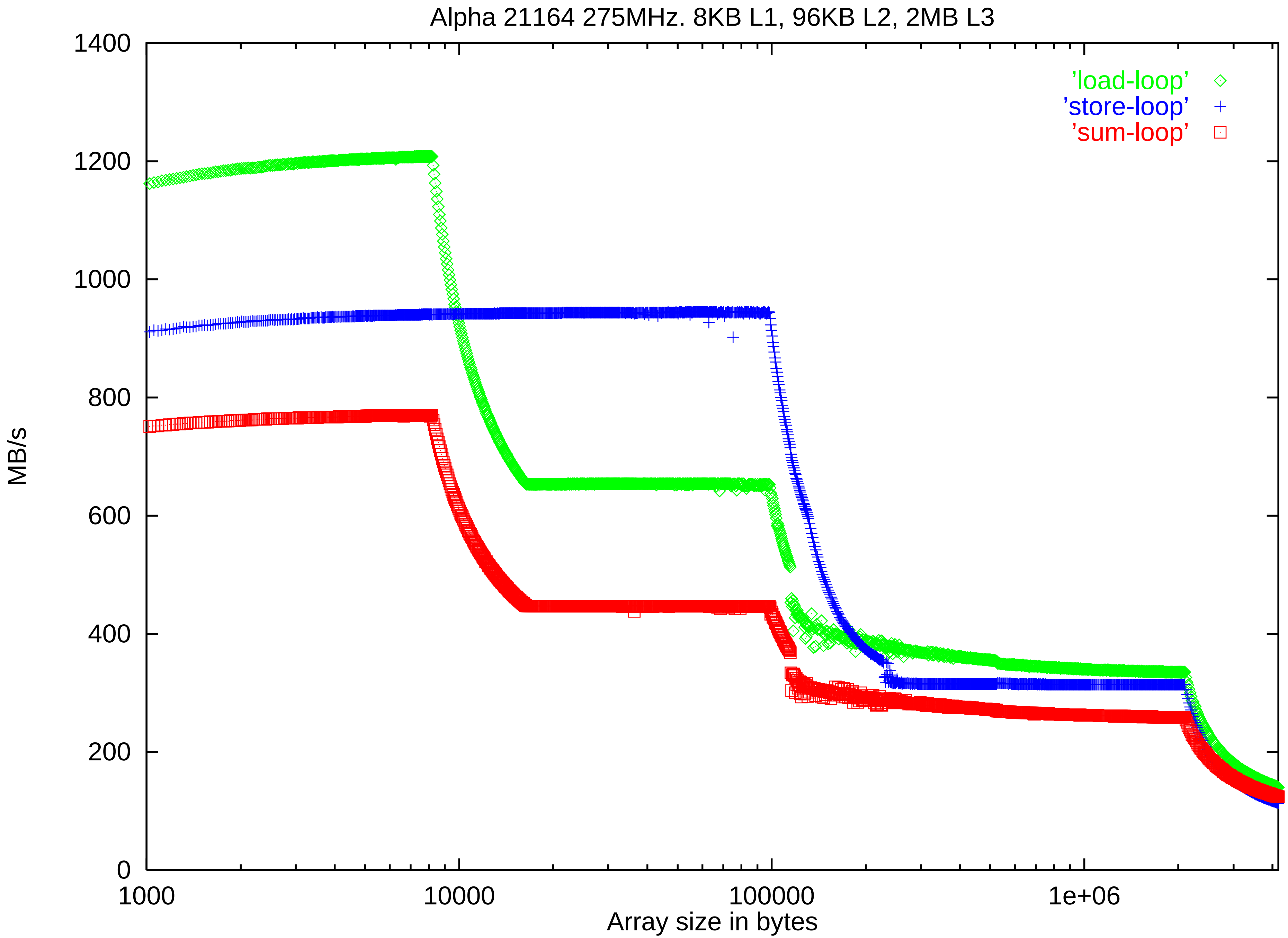
200MHz PPro 440FX

Ln	size	n-way	line	write	lat(cy)	lat(ns)	ld(MB/s)	st(MB/s)
L1	16KB	4	32	WB	2	7	1048	735
L2	512KB	4	32	WB	15	50	545	282
MM	-	-	-	-	64	213	231	116

300MHz PII 440LX

Performance Examples

1. *This table refers only to data caches. All the data was collected using a test program written by the author. Care was taken to perform 'reasonable optimization' of the loops for each of the different systems. Despite this, the numbers reported are in some cases rather lower than might be expected from reading the system documentation. Achieving the documented figures often requires 'tricks' that are not generally applicable, hence these figures represent the sort of memory system performance that real programs might hope to achieve.*
2. *The load bandwidth test simply reads words sequentially from an array of size n (where n is the size of the cache to be tested). When testing the L2 bandwidth, the L1 cache does not interfere since almost all loads will miss due to capacity misses; any useful data the cache holds will be pushed out before it is needed on the next iteration. (likewise for testing L3 and main memory)*
3. *Store bandwidth is measured by writing sequentially to an array of size n.*
4. *The latency tester creates a linked list of cache line sized elements, of total size n. It then accurately measures the time to chases down the list multiple times, and computes the load-to-use latency.*
5. *Many Pentium systems have write back L2 caches; I'm not sure why the BIOS hadn't enabled write-back on this system.*
6. *Note how the 21164's L1 cache is write through, and thus how the L1 store bandwidth is the same as the L2. (Likewise for both caches on the Pentium system).*
7. *The Pentium Pro and PII have longer latency L1 D-caches (enabled by the dynamic execution engine. Despite the two cycle latency, both are fully pipelined)*
8. *The Pentium II's off-chip L2 cache has longer latency than the Pentium Pro's L2 in the 2nd die. However, The PII's cache is made from off-the-shelf SSRAM parts, and is thus cheaper. (The use of SSRAM enables the throughput to be the same despite the latency being greater).*
9. *Larger caches sometimes benefit from larger line sizes (compare 21164 L1 and L2, or Pentium L1 and L2).*


Alpha 21164 275MHz. 8KB L1, 96KB L2, 2MB L3
'load-loop'
'store-loop'
'sum-loop'
MB/s
Array size in bytes
0
200
400
600
800
1000
1200
1400
1000
10000
100000
1e+06

Main Memory

- Increasing sequential bandwidth
  - Wide memory bus
  - Interleave memory chips

  $\Rightarrow$ DDR SDRAM or RAMBUS

- Access latency can impair bandwidth
  - Larger cache block sizes help

- Reducing average latency
  - Keep memory banks 'open'
    - ∗ Quick response if next access is to same DRAM Row
  - Multiple independent memory banks
    - ∗ Access to an open row more likely
    - ∗ SDRAM/RAMBUS chips contain multiple banks internally
  - System bus that supports multiple outstanding transaction requests
    - ∗ Service transactions out-of-order as banks become ready

Main Memory

1. *Wide memory systems (e.g. 256 bits on Alpha EB164) typically require a large number of memory modules (e.g. SIMMs/DIMMs) to be installed. This results in a large, expensive memory configuration which is often hard to upgrade due to a limited number of motherboard SIMM slots.*

2. *Wide memory systems are expensive. The higher clock rate of SDRAM has helped to reduce the need for wide memory systems in many applications. Pentium II and Pentium Pro systems make do with 64 bit memory systems. Alpha SDRAM systems are 128-256 bits wide.*

3. *Intel are committed to using RAMBUS on future processors. This will result in main memory connecting directly to the processor rather than system support chips. Each CPU will probably support 4-16 separate RAMBUS buses (known as 'strings')operating concurrently.*

4. *With a simple memory system, DRAM latency often results in the system bus laying idle much of the time, waiting for the row access to complete. Increasing the cache line size increases the percentage of cycles where useful work is done.*

5. *With independent memory banks, each cache line is serviced from a single memory bank (no word-level interleaving). The memory system typically accepts multiple transactions from the CPU, and services them out-of-order depending on the availability of each bank. Each bank is kept 'open' after an access, enabling subsequent accesses to the same DRAM page to continue with low latency. This type of scheme is becoming increasingly popular due to the fact that individual SDRAM/RAMBUS parts are now able to deliver relatively high sequential bandwidth—it's just the latency that's the real problem. Having multiple banks helps to reduce the average latency. Individual SDRAM and RAMBUS parts now contain multiple banks internally as well; the memory controller must keep track of which rows are currently latched and thus available for rapid access.*

6. *The STREAM benchmark is often used to measure memory system bandwidth. lmbench is often used to measure latency.*

Programming for caches

- Design algorithms so working set fits in cache
  - Large lookup tables may be slower than performing the calculation
- Organise data for spatial locality
  - Merge arrays accessed with the same index
- Fuse together loops that access the same data
- Prefer sequential accesses to non-unit strides
  - innermost loop should access array sequentially
- If row and column access to 2D arrays is necessary, use *cache blocking*
  - divide problem into sub-matrices that fit cache
  - e.g. matrix multiply $C = C + A \times B$

```
for (kb=0;kb<N;kb+=b){
  for (jb=0;jb<N;jb+=b){
    for (ib=0;ib<N;ib+=b){
      for(k=kb;k<kb+b;++k){
        for(j=jb;j<jb+b;++j){
          for(i=ib;i<ib+b;++i){
            C[k][i] = C[k][i] + ( A[k][j] * B[j][i] );
} } } } } }
```

- Avoid access patterns that are likely to cause conflict misses (aliasing)
  - e.g. large powers of 2
- Large strides can thrash the TLB

Programming for caches

1. *Caches can lead to large performance variances on different machines—possibly even on the same machine due to different physical page placement for different runs.*
2. *Organise data for spatial locality, e.g. collect elements of different arrays that are accessed together into a single* `struct` *array. This will also reduce the possibility of aliasing.*
3. *Fuse loops that access the same data in the same pattern into a single loop.*
4. *Large strides can cause cause thrashing of the TLB*
5. *Large strides that are powers of 2 are more likely to cause aliasing.*
6. *Two good articles on programming for caches are included in the additional material. (one for the 21164, one for x86). H&P is also very good.*
7. *Big Wins for cache optimizations*

Special Instructions

- Prefetch
  - fetch data into L1, suppressing any exceptions
  - enables compiler to speculate more easily
    e.g. Alpha: `ld r0 ← [r1]`
- 'Two-part loads' (e.g. IA-64)
  - speculative load suppresses exceptions
  - 'check' instruction collects any exception
  - enables compiler to 'hoist' loads to as early as possible, across multiple basic blocks
  - `ld.s r4 ← [r5]`
    `chk.s r4`
- Load with bypass hint
  - indicates that the load should bypass the cache, and thus not displace data already there
  - e.g. random accesses to large arrays
- Load with spatial-locality-only hint
  - fetch line containing the specified word into a special buffer aside from the main cache
    - or, into set's line that will be evicted next
- Write invalidate
  - allocate a line in cache, & mark it as modified
  - avoids mem read if whole line is to be updated

1. *Cache manipulation instructions have been added as part of most architecture's 'Multimedia' extensions*
2. *spatial-locality-only (non-temporal) loads are often implemented by loading the line into the set's line that will be evicted next.*

Multiprocessor Systems

Two main types:

1. Cache Coherenet

   ⇒ implicit shared memory communicationbetween processes/threads on single OS instance

   - Symmetric MultiProcessor (SMP) (Uniform Mem Access)
   - Non-Uniform Memory Access (ccNUMA)
   - 2-256+ processors

2. Message Passing

   ⇒ explicit communication between processes on multiple OS instances

     – May appear as 'single system image' cluster

   - Conventional networking (send/receive)
     – RPC
   - Remote DMA (read/write)
     – requires more trust & co-ordination
   - Gigabit Ethernet or specialist low-latency network
   - Avoid OS latency and overhead
     – zero-copy user-level accessible interface
     – (still need OS for blocking RX)
   - Highly scalable

⇒ Big supercomputers typically use a combination of both

1. *Beowulf ckusters*

Cache Coherent Systems

1. Single shared memory (SMP)
   - shared bus (2-4 CPUs)
   - switched interconnect (2-8 CPUs)
2. Distriubted Memory (ccNUMA)
   - memory per group of CPUs (e.g. 4)
   - or, memory hangs off each CPU
   - addressable as single physical memory
     - e.g. top bits identify 'node'
     - accesses to remote memory slower
   - (physical memory may be sparse)
   - interconnect via: crossbar, mesh, hypercube

Desire large WB caches to reduce memory traffic

1. x

Cache Coherence

- Coherence
  - Write serialization : all writes to the same location are seen in the same order
- Consistency
  - Behaviour of reads and writes wrt other memory locations

⇒ Implement *shared* & *exclusive* cache line ownership states (and *invalid*)

  - CPU must get exclusive access to cache line before updating it
  - Must inform all CPUs that have the line in ‘shared’ state
  - On bus systems, snoop ‘write upgrade request’
  - On switched interconnect, broadcast, or use directory stored at ’home node’

1. *x*

Programming considerations

- False sharing
  - variables with different ‘owners’ placed in same cache line
  - line ‘exclusive’ ownership thrashes between CPUs
  - ⇒ pack variables according to owner
  - ⇒ consider padding to cache line boundaries
- Page allocation - ccNUMA
  - physical address determines which node pages ’live’ on
  - ideally, same node they’ll be accessed from
  - different placement policies:
    - ∗ local
    - ∗ random
  - OS may employ page migration
    - ∗ copy page to different node, update all page tables

1.

Course Conclusions

- Modern CPUs are impressive feats of engineering
- Rely heavily on 'common case' assumptions
  - potential for large performance variance
  - need better compilers
  - on-line feedback-directed optimization
- Extracting loop and thread-level parallelism will be increasingly important
  - auto-parallelising compilers
  - functional / declarative languages
- ISA compatibility is currently very important
  - MSIL / Java byte code

Questions

- How much ILP will be economically exploitable?
- Will we succeed in exploiting other forms of parallelism?
- Hardware vs. software scheduling?
- What will we do with 1 billion transistors?
- Will x86 ever die?

1. *bugs surprisingly rare in modern processors...*
2. *I expect to see 8-way integer SMT systems in 5 years time.*
3. *I believe multi-processor SMP motherboards (2 and 4 way) will become more common place too.*
4. *Will other architectures be sufficiently better than x86 to finally kill it?*
5. *SIA Roadmap for 2010: 1B logic transistors, 27x27mm chips, 5GHz local clock speeds, 64GB main memories*