The 4 by 4 and 16 by 16 Switch Fabrics Design and Implementation Draft

Daniel Gordon

16th February 1993

1 The 4 by 4 Switch Fabric

The switch fabric is a cell synchronous, crossbar switch. It is self routeing, that is, the cells have prepended to them *routeing tags* which tell the fabric which output the cell should be sent to (see figure 1). The routeing tags also contain priority information (in fact a single bit is used) which is examined when two cells contend for the same output. This tag is stripped off before the cell reaches the output stage.

Bit	Name	Use
0	Active	Always set for a cell
1	Priority	Set for high priority cell
2-3	Route	Output requested for this cell
4-7	Spare	Unused

Table 1: Fabric Route byte (4 x 4 Fabric)

Since the fabric (or indeed an output port controller) may chose not to accept a cell, an *acknowledgement* signal is provided. If the cell is blocked in the fabric, the input port on which the cell was injected will see a low on the acknowledgement line, otherwise it will see whatever the output controller chooses to respond with — in fact it is a clear return data path from output to input.

The job of the switch fabric is to detect cells, detect contention between cells, arbitrate between contentious cells, switch data between four inputs and outputs and to propagate or generate acknowledgement signals appropriately. Apart from the data and acknowledgement, the only other signal used is 'frame start'. This is generated externally and is used to synchronise cells.

Figure 1 shows a schematic view of the switch fabric.

1.1 Route Decoder

This section decodes the priority and routeing information from the fabric routeing byte. It generates 32 signals, 8 for each input — 4 indicating which output slot (if any) that input is requesting and 4 to indicate whether it is a priority request.

1.2 Priority Filter

This section just filters out each input without priority which is vying with an input which has priority for the same output. These are then put into a 16 bit register - 4 bits for each input to indicate which output slot it is requesting access to.

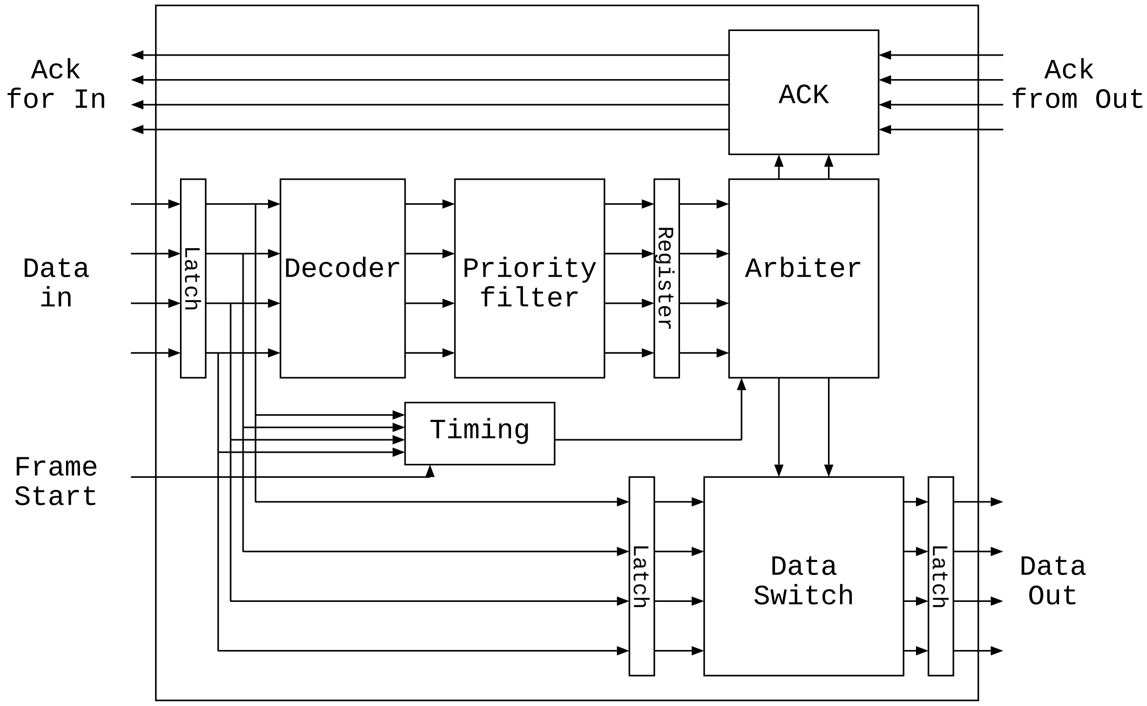


Figure 1: The 4 by 4 Fabric

1.3 Timing

This part of the chip detects an active bit from any of the inputs and enables the arbiter. It has 3 states (see figure 2) Run, Wait and Route. A frame start signal moves it from Run to Wait and an active bit from any slot moves it from Wait to Route. Any condition then moves the state machine back to Run. It remains in the 'Route' state for just one cycle which triggers the arbitration unit.

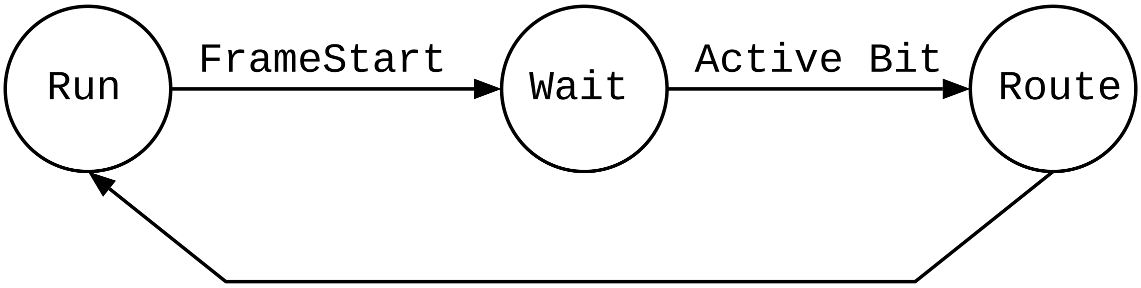


Figure 2: State Diagram for Timing Circuit

1.4 Arbitration

Whenever two or more cells contend for the same output of the switching element, arbitration takes places. Arbitration is done for each of the four outputs of a switching element. Each Arbitrater runs an independent round robin system. The inputs are numbered 0 to 3 and the last input to have a cell (of any priority) selected is remembered. If two or more cells contend, the last input which was selected is used as a base. The next input above (with suitable wrap around) the last selected is the one to be selected. Figure 2 shows which input is granted access to particular output given the last input granted access and the inputs attempting to gain access to that output.

The arbiter produces 4 two bit numbers determining which input each of the 4 outputs should copy. These are then passed to the data switch.

1.5 Data Switch

This is a simple multiplexer.

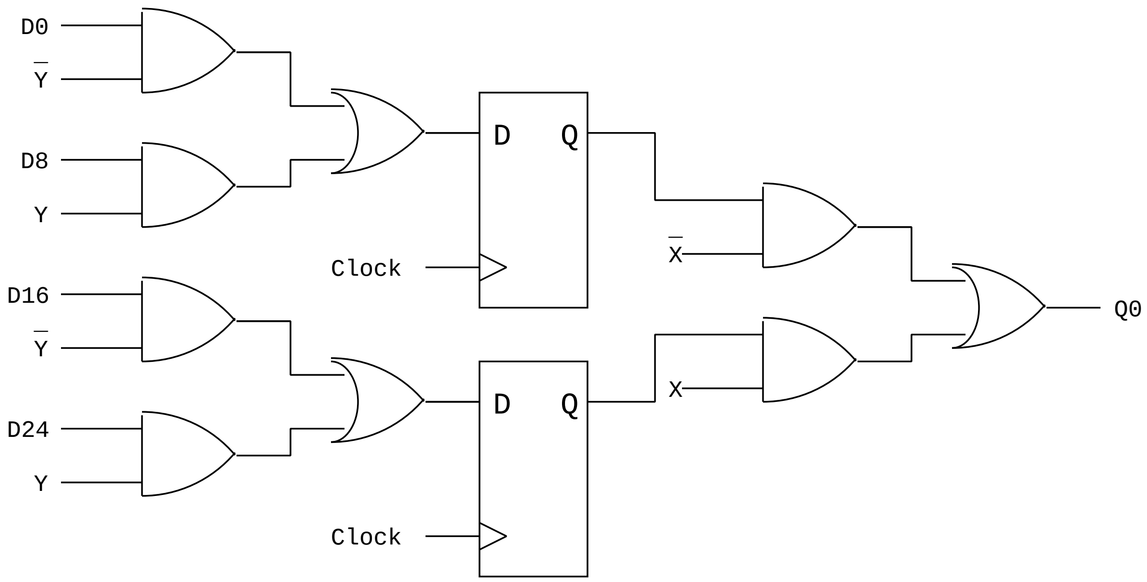


Figure 3: A Multiplexer.

1.6 Acknowledgements

As soon as the arbiter has decided which routes are to be granted, this part of the circuit transfers the acknowledgement signals from the destination to the source of each route asynchronously or generates negative acknowledgements for inputs that have not succeeded in the arbitration process.

1.7 Delay Through the Switch Fabric

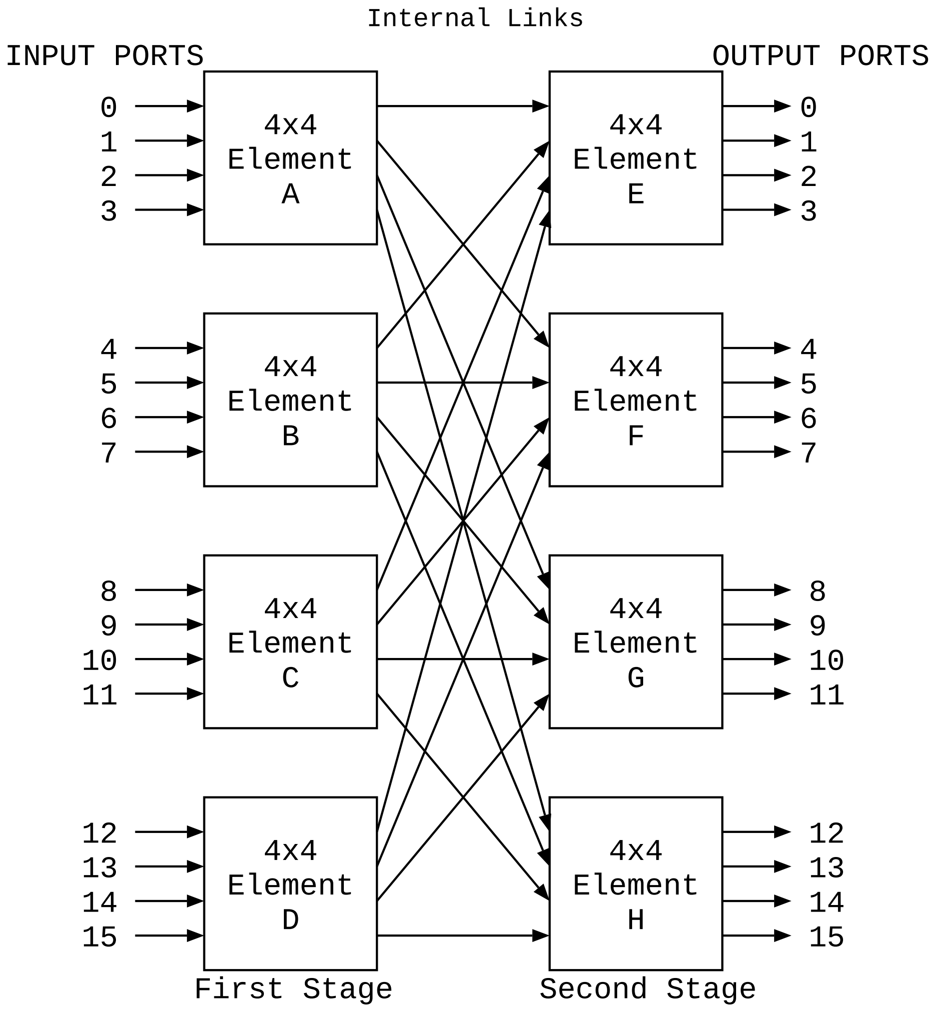
The data is latched at input and output. It is delayed one clock cycle whilst the decoding and filtering is being done and the data switch takes one cycle. The fabric routeing byte is stripped off, so the total delay from the start of a cell going into the fabric to the start of the (stripped) cell coming out is 5 clock cycles.

1.8 Implementation

The switch fabric is built on a 4200 gate equivalent Xilinx programmable gate array. This, a serial PROM and 4 SIL resister packs (which pull-down all the inputs to each switch element for when a slot has no device attached) are the only components on a single height extended eurocard which plugs directly into a fairisle backplane. It can be clocked at 20 MHz and currently frame start pulses occur every 64 clock cycles.

2 The 16 by 16 Switch Fabric

The 16 by 16 switch fabric is constructed from eight 4 by 4 crossbar switching elements arranged in a delta network as shown in figure 4. The delta network gives rise to internal blocking. (Consider when port 0 and 1 both wish to send to ports 5 and 6 respectively. The cells both wish to use the internal link from element A to F; only one of them can). A similar 8 by 8 switch can be constructed by using half (cutting it horizontally) of this switch.



Each arrow represents 8 data signals and a returning acknowledgement signal

Figure 4: A 16 Port Delta Network

Each 4 by 4 crossbar switching element is almost identical to a 4 by 4 switch fabric. The *routeing tag* (or fabric routeing byte) is slightly different (see table 3). The first stage acts in the same way as the 4 by 4 switch fabric, removing the active bit from the *routeing tag*. The second stage routes using the upper nybble and, therefore, bit 4 as its active bit.

Since the 16 by 16 switch fabric has to be constructed from 2 boards and (to keep the design simple) these two are identical, an extra signal (set to 0 on one board and 1 on the other) is needed to cope with the antisymmetry of the two halves. That is, since the lower board (with elements C, D, G and H) is identical to the upper board, output 0 from element C is connected to G and output 1 to H instead of output 2 to G and output 3 to H, and similarly for element D. The extra signal (called bd) is exclusive-ORed with the most significant routeing bit of the lower nybble of the fabric routeing byte to give the correct route from the input stage to the output stage.

2.1 Delay Through the 16 by 16 Switch Fabric

The delay through the 16 by 16 fabric equals the sum of the delays through each stage, which is 8 cycles. Since one byte is stripped from the cell as it passes through, the total delay from the start of the cell going in and the start of the cell coming out is 9 clock cycles.

2.2 Implementation

The 16 by 16 switch fabric fits onto two identical double height eurocard PCBs connected by a 96 way ribbon cable. One with elements A, B, E and F on and the other with C, D, G and H. The switching elements are built on a 6400 gate equivalent Xilinx programmable gate array. The only other components on the boards are Xilinx PROMs (to program the Xilinx chips) and SIL resister packs which pull-down all the inputs to each switch element for when a slot has no device attached.

inputs requesting output	last input granted			
	0	1	2	3
3	3	3	3	3
2	2	2	2	2
2,3	2	2	3	2
1	1	1	1	1
1,3	1	3	3	1
1,2	1	2	1	1
1,2,3	1	2	3	1
0	0	0	0	0
0,3	3	3	3	0
0,2	2	2	0	0
0,2,3	2	2	3	0
0,1	1	0	0	0
0,1,3	1	3	3	0
0,1,2	1	2	0	0
0,1,2,3	1	2	3	0

Table 2: Round-robin Arbitration

Bit	Name	Use
0	Active	Always set for a cell
1	Priority 1	Set for high priority cell at first stage
2-3	Route 1	Output requested for this cell at first stage
4	Active	Always set for a cell
5	Priority 2	Set for high priority cell at second stage
6-7	Route 2	Output requested for this cell at second stage

Table 3: Fabric Route byte (16 x 16 Fabric)