VLSI PHOTONICS

USC SWITCH IC SPECIFICATIONS

AND

DESIGN CHOICES

Panduka Wijetunga, Jeff Sondeen

A.F.J. Levi

University of Southern California

8/17/99

Rev 0.1

Table of Contents

Appendix

1.0 NG TALKING POINTS

1.1 Summary of WIDE SLOW and NARROW FAST

This section summarizes the discussions with NG including those held in April 1999 at NG on the WIDE SLOW design and the USC proposed FAST NARROW design.

- NG introduced a "WIDE SLOW" design as a non-integrated version of the original IC. Non-integration means that the optical receiver array, optical transmitters array and electronic switch core are implemented as separate chips. This increases the number of electrical I/O to 5000 per Switch IC with a GHz clock rate. For comparison current state of the art in high performance packaging is less than 1500 I/O at sub-GHz clock rate. For reference, Table 1 shows the Semiconductor Industry Association (SIA) performance of packaged integrated circuits road map for CMOS technology. Currently we are designing in 0.35 μm and 0.25 μm CMOS technology. Chip I/O of 5000 occurs in year 2009 in high performance packages. Circuits for such packages make use of sub-0.1 μm CMOS technology.

Table 1: Performance of Packaged Chips

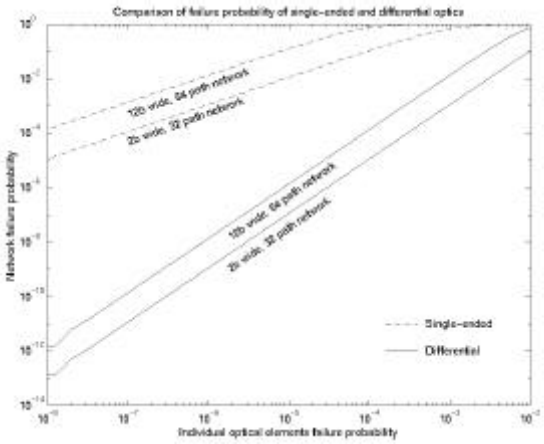
YEAR OF FIRST PRODUCT SHIPMENT	1997	1999	2001	2003	2006	2009	2012
TECHNOLOGY GENERATIONS DENSE LINES (DRAM HALF-PITCH) (nm)	250	180	150	130	100	70	50
ISOLATED LINES (MPU GATES) (nm)	200	140	120	100	70	50	35
Number of Chip I/Os							
Chip-to-package (pads) high-performance	1450	2000	2400	3000	4000	5000	7300
Chip-to-package (pads) cost-performance	800	975	1195	1460	1970	2655	3585
Number of Package Pins/Balls							
ASIC (high-performance)	1100	1500	1800	2200	3000	4100	5500
MPU/controller, cost-performance	600	8100	900	1100	1500	2000	2700
Cost-performance package cost (cents/pin)	1.40-2.80	1.25-2.50	1.15-2.30	1.05-2.05	0.90-1.75	0.75-1.50	0.65-1.30
Chip Frequency (MHz)							
On-chip local clock, high-performance	750	1250	1500	2100	3500	6000	10000
On-chip, across-chip clock, high-performance	750	1200	1400	1600	2000	2500	3000
On-chip, across-chip clock, cost-performance	400	600	700	800	1100	1400	1800
On-chip, across-chip clock, high-performance ASIC	300	500	600	700	900	1200	1500
Chip-to-board (off-chip) speed, high-performance (reduced-width, multiplexed bus)	750	1200	1400	1600	2000	2500	3000
Chip-to-board (off-chip) peripheral buses	250	480	785	885	1035	1285	1540
Maximum number of wiring levels	6	6-7	7	7	7-8	8-9	9

- The "NARROW FAST" design was suggested by USC as a way of reducing the number of electrical I/O, reducing power consumption, and increasing the testability of the final electrical Switch IC.
- The NARROW FAST approach reduces the number of optoelectronic paths per electrical Switch IC to 32, reduces the path width to 2 bits, and increases the signaling data rate to 2.0 Gb/s. The ability to pass data through the switch at 2.0 Gb/s was tested using the TE02 test IC.
  - There are 4 electrical ports per Switch IC. For the case of 16 nodes and 4 Switch ICs per node, the network can support a maximum of 128 processors.
  - To increase the reliability of the network, complementary (differential) signaling is used in the optoelectronics.

    If the failure probability of an optoelectronic element is *p*, the number of optoelectronic signal channels is *n*, and number of nodes in the network is *m*, assuming a differential channel can operate single-ended (i.e. in a complimentary mode) the network failure probability is

$$f_{single} = 1-(1-p)^{nm}$$

$$f_{differential} = 1-(1-p^2)^{nm}$$



e.g. If the mean time to failure peaks strongly at 10 years for a single element $p = 10^{-5}$ (1/(24*10*364)), the number of signal channels is n= 70, and the number of nodes in the network is m = 16, then the mean to failure is 4 days for single-ended and over 390 years for differential.

1.1.1 Estimates for NARROW-FAST

- 16 Busy-Bits serialized into one signal line and one FRAME

  If each processor had a dedicated TOKEN, 32 processors require 32 optical lines (channels). On the other hand 16:1 serialized BUSY BIT scheme require only 2 optical lines for the BUSY BITs and 2 lines for the FRAME and the spare FRAME.

- Reduces the power consumption by a factor of 6.

  The path width is reduced from 12 bits to 2 bits, but the data rate is increased from 1.0 Gb/s to 2.0 Gb/s. The number of optical path is reduced from 64 to 32. These changes combine to reduce the power consumption of the IC by a factor of six to 6.25 W in 0.35 µm CMOS.

- Node pass-through latency for BUSY-BIT and data is 4 cycles (4 ns for 1.0 GHz clock) plus time of flight.

  At the optoelectronic input interface the data is retimed and demuxed (1 cycle latency). At the input to the Switch core, data is retimed (1 cycle). At the output end of the Switch core data is again retimed (1 cycle). At the optoelectronic output interface the data is retimed and muxed 2:1 (1 cycle). Which gives a pass-through latency of 4 cycles. If an additional optoelectronic to internal clock data hand-off stage is included instead of using the switch core input interface for the for the data hand-off, then the pass-through latency (node latency) will increase by one cycle. The BUSY BIT circuitry shares the clocks with the datapath, therefore BUSY BIT circuitry is design to have the same pass-through latency as the data. These calculations do not include the latencies incurred outside the Switch IC (external delays)

- Maximum round-trip latency through 16 nodes is 64 cycles (64 ns for 2.0 Gb/s, 1.0 GHz clock)

  This does not include time of flight and optoelectronic conversion delays external to the Switch (transmit/receiver arrays). Estimates also assumes optoelectronic to internal clock data-hand off is integrated into the switch core input interface.

- Maximum grab latency, if busy-bit is free, and there is only one node requesting

  $(n.t_{hop} + \#serial\_bits) = (16 \times 4 + 16) \times 1$ ns = 80 ns. Where n is the number of nodes in the network (16 for our case), t_{hop} is the hop delay and #serial_bits is the number of bits in the bit stream.

  Only one node is requesting the destination, and the corresponding BUSY BIT has just passed the node, and the requested BUSY BIT is BUSY BIT <0>. The node has to wait for the FRAME (BUSY BIT stream) to traverse the whole network, and then has to wait till the last bit in the stream.

- If the owner of a particular path is unavailable to receive data, it can request and grab its own BUSY-BIT.

- For increased reliability, the NARROW FAST approach assumes VCSELs and detectors are used in a differential signaling configuration. Example: two 70-element VCSEL and detector arrays {125 μm pitch: (4 x 18 = 72, 0.5 x 2.25 mm^2) or 250 μm pitch: (8 x 9 = 72, 2.0 x 2.25 mm^2)} or one 140 element VCSEL and detector arrays {125 μm pitch: (8x18=144, 1.0 x 2.25 mm^2) or 250 μm pitch: (16x9=144, 4.0 x 2.25 mm^2)}.

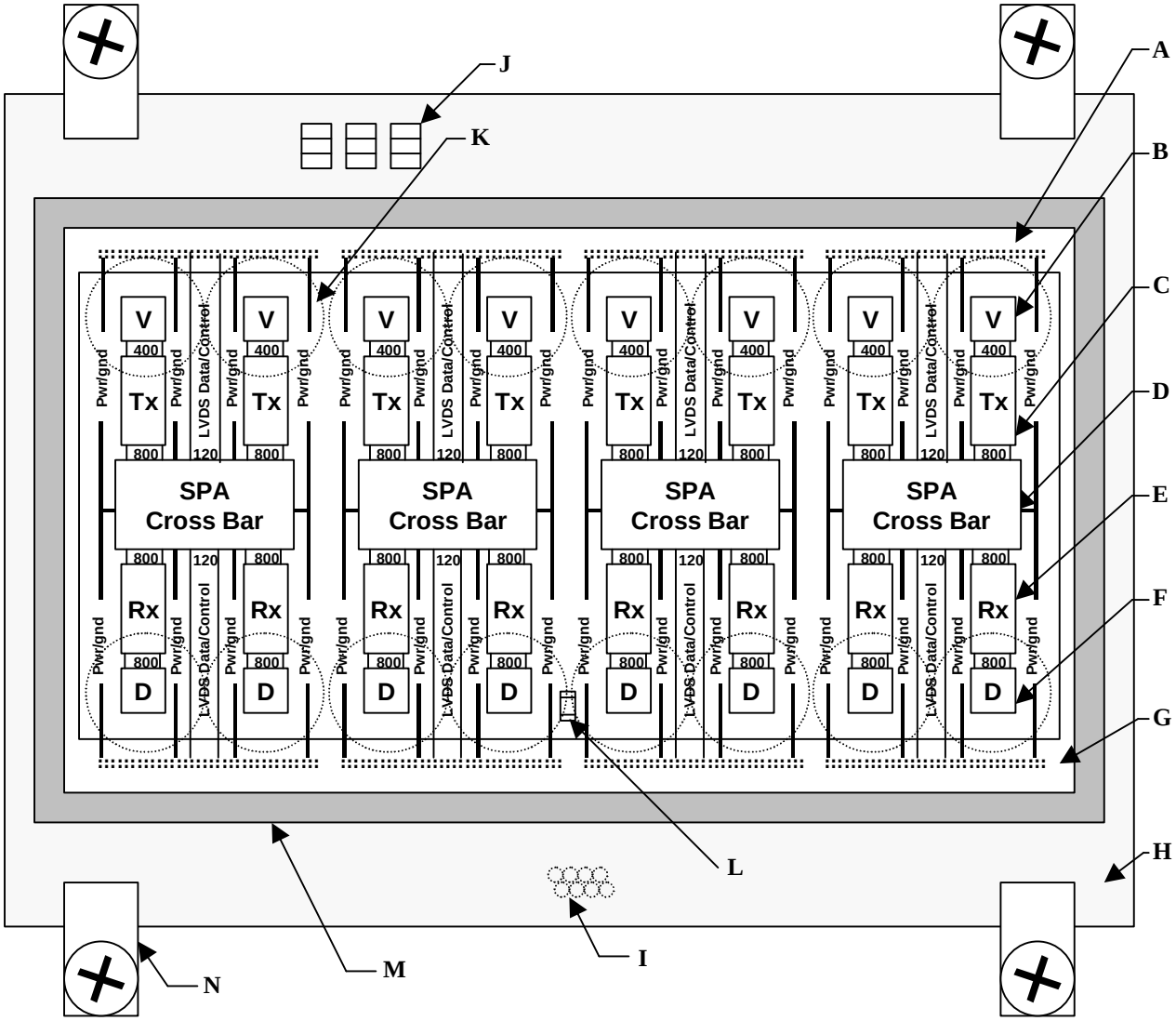
1.1.2 WIDE-SLOW initial NG design (1Gb/s/signal, 64x12 IC)

- Base level design features
  - The SPA crossbar is partitioned into 4 4x64 ICs instead of a single 16x256 IC.
  - The SPA crossbar IC has one internal clock which is common to all optical paths. This clock is selected from either the electrical or the optical clock
  - The IC has a 64:1 mux at each electrical output port.

- Base level design limitations
  - 3392 electrical lines for I/O per IC which must be delivered to the IC deskewed, and will probably require termination.
  - Control lines for busy-bit request, and the switch setup decode logic do not have a register with separate load enables.

- 1696 differential I/O signals ((800 = 48) x 2) which must be delivered to the IC de-skewed and properly terminated
  - 1696 differential signals (3392 lines) require about 800 ground and 800 power pads, totaling 5000 pads (excluding control)
  - The I/O pads are distributed over the chip area (flip-chip bump bonding), and need to be delivered de-skewed at the pad boundary
  - Due to the distance between Switch IC and the optical Rx/Tx, destination (load) terminated LVDS may have to be used, increasing the power consumption.
  - With termination and I/O distribution of the optical I/O, the power consumption of the IC can be as high as 40 W [20 W (original estimate) + 7W (additional re-timing) + 13 W(LVDS termination, 0.25 Vpp signal into 50 Ω using 3.3V rail, 5 mA x 3.3 V x 64 x 12)]

"WIDE SLOW" initial NG design (1.0 Gb/s/signal, 64x12 IC)



A Substrate to Package Solder Bump Connection
2 Rows x 75 @0.006" C_L, 8 places
1200 Total (960 Signal, 240 P/G)

B 400 Element VCSEL
~2.7mm x 2.5mm
400 Signal I/O
Plus Common Cathode I/O

C 400 Element VCSEL Driver
~2.7mm x 5mm
1200 Signal I/O
Plus Power/Ground/Bias I/O

D 4 x 64 Crossbar
~5mm x 10mm (est)
3392 Signal I/O
Plus Power/Ground/Control I/O

E 400 Element Receiver
~2.7mm x 5mm
1600 Signal I/O
Plus Power/Ground/Bias I/O

F 400 Element Diff. Detector
~2.7mm x 2.5mm
800 Signal I/O
Plus Common Cathode I/O

G Sapphire Substrate
First Metal: 1.5μm Min Line/Space
Second Metal: 2μm Min Line/Space
Third Metal: Pwr/Gnd/Bumps

H Cofired Ceramic Package
~2.75" x 2.00" x 0.10"
(Driven by SPA-to-Board IO)

I SFI Pads @ 0.100 in. C_L
0.050 Staggered Rows
1080 Pads Total
(960 Signal, 120 Power/Ground)

J De-coupling Capacitors

K 7mm dia. Lens/Housing Allowance

L High Frequency Caps

M Solder Seal Ring

N SFI Spring Clip

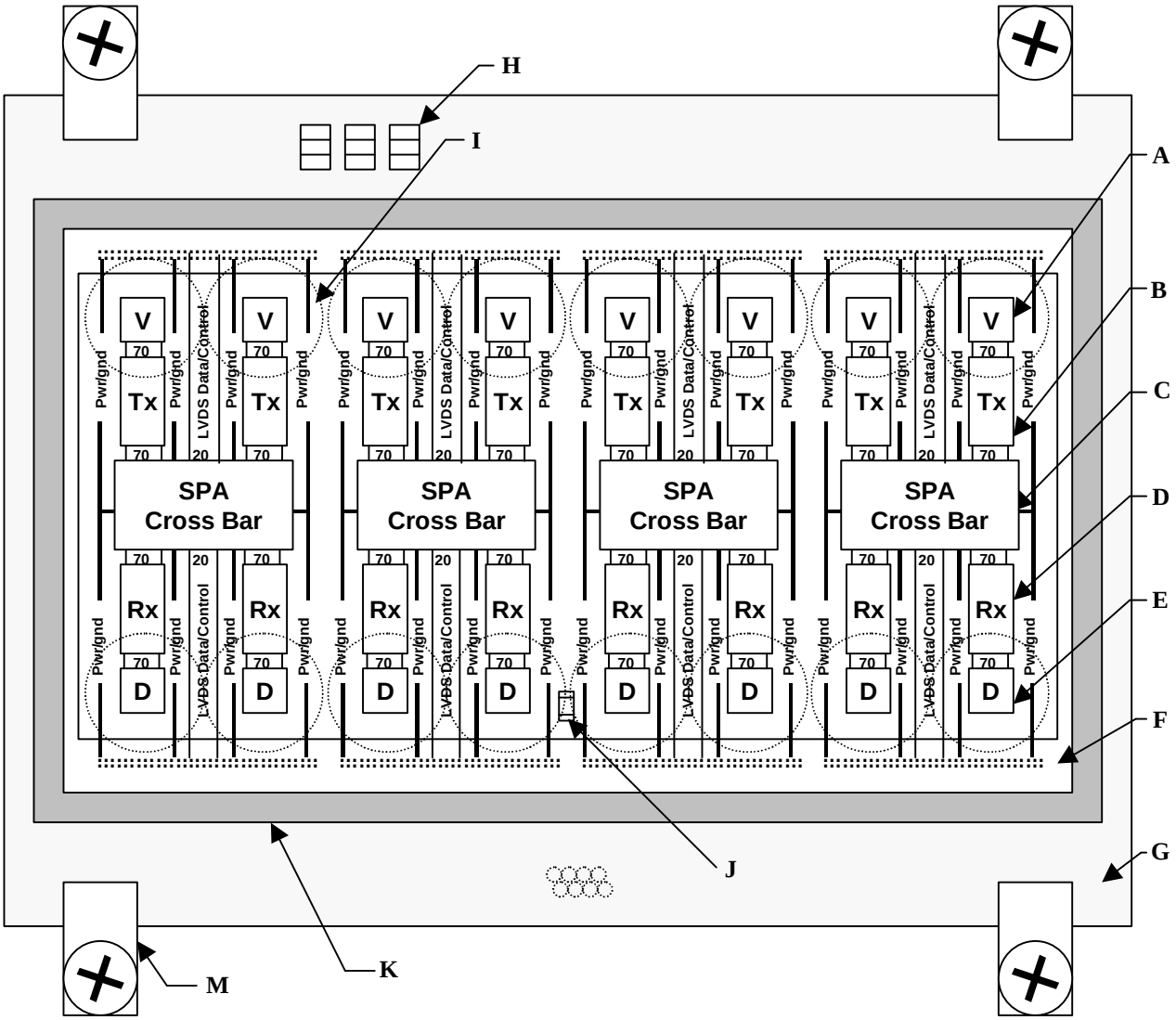
1.1.3 NARROW-FAST design (2 Gb/s/signal, 32x2 IC)

- 156 differential I/O signals ((70 + 8) x 2) which must be delivered to the IC deskewed and correctly terminated (32 data + 2 control + 1 clock + 1 spare clock + 2 BUSY BITs + 1 FRAME + 1 spare FRAME, times 2 input and output = differential I/O signals).
  - 156 differential signals (312 lines) require about 70 grounds 70 power pads, totaling 420 pads (excluding control).
  - The optoelectronic I/O can be placed at the top and bottom of the IC (3 rows at the top, 3 rows at the bottom on a 135 μm pitch totaling 9.0 mm (leaving 1 mm for data/control) each long-side, and must be de-skewed at the chip boundary.
  - Due to distance between IC and optical Rx/Tx destination (load) terminated LVDS may have to be used, increasing the power consumption.
  - With termination and I/O distribution of the optical I/O, the power consumption of the IC will be around 6.25 W (4.0 W (original estimate) + 1.25 W (additional re-timing) + 1.0 W [LVDS termination, 0.25 V signal into 50 Ω using 3.3V rail, 5 mA x 3.3 V x 64 x 2)]

- If BUSY BIT request and switch setup have separate load enables. The BUSY BIT grab and release latencies are dependent on $T_{request}$ and $T_{release}$ and can be masked to improve BUSY BIT grab latency
  - $T_{request}$ is the time it takes to identify a new request and activate its corresponding BUSY BIT request (time it takes to request a BUSY BIT)
  - $T_{release}$ is the time it takes to release the BUSY BIT, which includes the realization of end of data transfer, and deactivation of the BUSY BIT request.
  - If there are separate enables for loading the controls, pre-computation can be done so that a grabbed BUSY BIT can be released and a new BUSY BIT can be requested immediately at the end of transmission
  - This then masks the $T_{request}$ and $T_{release}$ from the grab latency, but requires additional external circuitry to do computation in parallel with data transfer

- If BUSY BIT is serialized then the BUSY BIT optical line usage can be reduced
  - Having an individual dedicated TOKENs for each processor requires 64 lines (32 signals x 2)
  - The BUSY BIT protocol uses a serialized stream of bits, where each bit in the stream corresponds to a processor in the network, and reduces the signal line usage and helps relieves the pad constraints discussed earlier.
  - Serialization of the busy-bits requires a FRAME signal to identify the beginning of the serial bit-stream
  - If m busy-bits are serialized into one signal the number of lines used is ((32 x 2)/m+2), which is nearly a factor of m improvement in the busy-bit signal line usage
  - The serialization should be done carefully to ensure that the busy-bit grab latency is kept constant or is improved compared to the single signal per busy-bit

"NARROW FAST" design (2.0 Gb/s/signal, 32x2 IC)



A 70 Element VCSEL in differential configuration
~2.7mm x 2.5mm
70 Signal lines
Plus Common Cathode I/O

B 70 Element VCSEL Driver
~2.7mm x 5mm
70 Signal lines
Plus Power/Ground/Bias I/O

C 4 x 32 Crossbar
~5mm x 10mm (est.)
312 Signal lines
Plus Power/Ground/Control I/O

D 70 Element Receiver
~2.7mm x 5mm
70 Signal lines
Plus Power/Ground/Bias I/O

E 70 Element Diff. Detector
~2.7mm x 2.5mm
70 Signal lines
Plus Common Cathode I/O

F Sapphire Substrate
First Metal: 1.5㎛ Min Line/Space
Second Metal: 2㎛ Min Line/Space
Third Metal: Pwr/Gnd/Bumps

G Cofired Ceramic Package
~2.75" x 2.00" x 0.10"
(Driven by SPA-to-Board IO)

H De-coupling Capacitors

I 7mm dia. Lens/Housing Allowance

J High Frequency Caps

K Solder Seal Ring

L SFI Spring Clip

1.2 Possible use of the Vitesse VSC852 as the Switch IC

This section summaries the possible use of the Vitesse VSC852 as the Switch IC.

1.2.1 Introduction

The VSC852 is a commercial GaAs 64x64 asynchronous crosspoint switch from Vitesse semiconductor with 1.6 Gb/s signaling rate per pin. It is assumed that serialization of parallel data, clock and data recovery, and link protocol are implemented using other ICs. Though there are drawbacks to using this component as the Switch IC for the upcoming backplane demonstration, it also brings the component support, flexibility, and reliability associated with packaged commercial components. The VSC852 has 64 PECL compatible inputs, 64 PECL compatible outputs, a serial input port for switch setup, and a 3.3 V single power supply (excluding the termination voltage V_{tt}), and consumes 28 W (without the power of the V_{tt} current). A 13-wide, TTL-level, serial bit-stream is used to identify a specific input-output combination. The LOAD signal is used to load the serial bit-stream to the program register and CONFIG is used to load the program register to the switch control registers. The serial control input port can be clocked using a 25 MHz clock. The high-speed outputs have peak-peak jitter of 100 ps, and the typical skew between two lines is 1.0 ns. The pin-to-pin delay of the VSC852 is currently unavailable (although the VSC850, a 1.25 Gb/s 16x32 crosspoint switch has a delay of 1.8 ns).

The following sections will discuss approaches to incorporate this product into the datapath and control-path of a distributed control network, as well as the advantages, disadvantages, and expected performance of such a network. The last section compares the use of VSC852 with the proposed USC "Narrow Fast" approach.

1.2.2 Performance estimates for using the VSC852 for the datapath

The VSC852 is a non-blocking crosspoint switch, thus it can be used instead of a custom Switch IC in fixed receiver path optical backplanes. We will examine two example systems: an asynchronous 1-bit wide, 4-processor per node, 16-node network, requiring end-to-end clock and data recovery, and a 2-bit wide, 4- processor per node, 8-node network (can be synchronous or parallel asynchronous). An asynchronous network with a path/port width of greater than one is defined as a parallel asynchronous network.

In an asynchronous network there will be no retiming at each node, and will contain end-to-end clock and data recovery. Since there is no retiming at each node, the skew between signals as well as signal jitter will accumulate as the data propagates through the network. If the worst case end-to-end total uncertainty (total skew + jitter) between any two bits of the same path/port are greater than the bit period, then the network will require extra data aligning circuitry to take care of the uncertainty between the two signals. If you assume that jitter accumulates linearly, this gives us a relationship between skew (skew between two signals of the same path at each node t_{skew}), peak-to-peak jitter (δo), destination latch setup time (t_s), number of nodes (n), uncertainty of the clock recovery circuit (Δs) and bit period (tp) (Figure 1).

$$(\delta_0 + t_s) \le (t_p - \Delta s - t_s) \cdot \frac{1}{n}$$

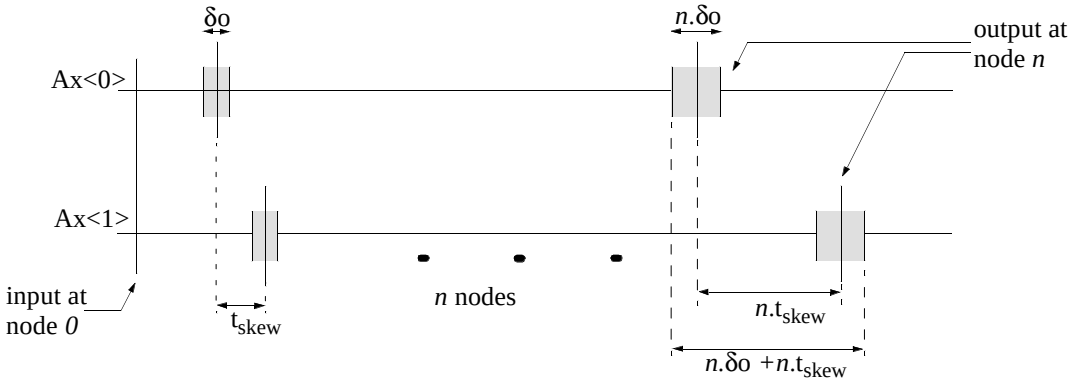


FIGURE 1: Jitter and skew accumulation in a asynchronous parallel the network (no retiming at each node)

This means for a n = 16, tp = 625 ps (800 MHz or 1.6 Gb/s NRZ) and a latch setup time and uncertainty of 0.0 ns

$$(\delta o + t_s) < 39 \text{ ps}$$

This shows that the VSC852 with 1.0 ns typical skew and 100 ps jitter can not be used to design a parallel asynchronous network without end-to-end logic to compensate for the errors due to skew and jitter (without additional data aligning circuitry). At each node external retiming circuitry can be used to compensate for skew and jitter (the network is no longer asynchronous). But since the typical 1.0 ns skew between two outputs is much larger than the 625 ps bit period (1.6 Gb/s), a 1-bit wide (serial), 16-node network is preferable to a parallel network. In the case of a serial network, the VSC852 can be used to connect 64 processors arranged in 16 nodes with 4 processors per node, and the switch backplane will be asynchronous with end-to-end clock and data recovery requiring additional ICs and protocols.

The power consumption of the VSC852 is 28 W excluding the termination and 33 W including termination. Thus, the use of the VSC852 in the datapath increases the power consumption at each node, compared to the "Narrow Fast" approach. However, the primary disadvantage of the VSC852 stems from its low-speed (25 MHz), 13-bit, serial configuration bit-stream. This can be demonstrated by noting the control setup time of the following example:

Assume that processors A_1, A_2, A_3 and A_4 are of the same node and that processors A_2, A_3 and A_4 are dormant (only A_1 require communication at the present time). Assume that A_1 is currently communicating with processor B_x and now wishes to communicate with processor C_x, and that A_1 has already reserved the path for C_x. Note that our notation for indicating a flow-through path for some processor X is X' $\rightarrow$ X, while we indicate processor Z talking to processor Y as Z $\rightarrow$ Y (Figure 2).

First A_1 must recreate the flow-through path for B_x (B_x' $\rightarrow$ B_x), and then must create the connection $A_1 \rightarrow C_x$. Assuming that the bit-stream for B$\rightarrow$B was already loaded to the program register, and the bit-stream for $A_x \rightarrow C_x$ is currently present in the serial registers, then the reconfiguration of the switch will consume less than 150 ns (excluding the latencies associated with reserving C_x).

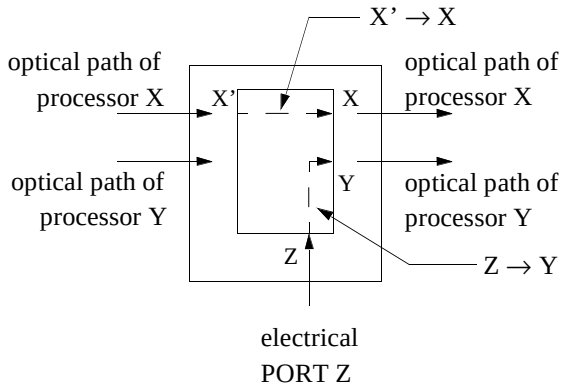


FIGURE 2: VSC852 switch setup notations

1.2.3 Performance estimates for using the VSC852 for the control-path

To implement the original NG TOKEN, the VSC852 can be utilized to reduce some of the external TOKEN grab, inhibit and release complexity. In this design there would be 2 VSC852 ICs per node, one dedicated for data transmission and the other for control. The VSC852 for control would switch the TOKEN that needs monitoring to the external TOKEN processing logic. By doing so, the external TOKEN processing logic can be reduced by a factor of 16 (now the external circuitry only needs to monitor 4 TOKENs instead of all 64).

The disadvantage of using the VSC852 for control also stems from the long slow-speed serial bit-stream required for the VSC852 setup. Assume that the node wishes to monitor $TOKEN_i$. First the switch should create a connection between $TOKEN_i$' (Figure 3) and the node electrical output. Then it needs to create a connection between $TOKEN_i$ and the node

electrical input. For TOKEN releasing the switch should be resetup for $TOKEN_i$ flow-through (connection $TOKEN_i' \rightarrow TOKEN_i$). If only the current node is requesting $TOKEN_i$, and if $TOKEN_i$ had just past the node, even if we assume that the node can release currently held TOKENs soon after data transmission, the TOKEN grab will still consume 1.2 μs (excluding ring round-trip delay).

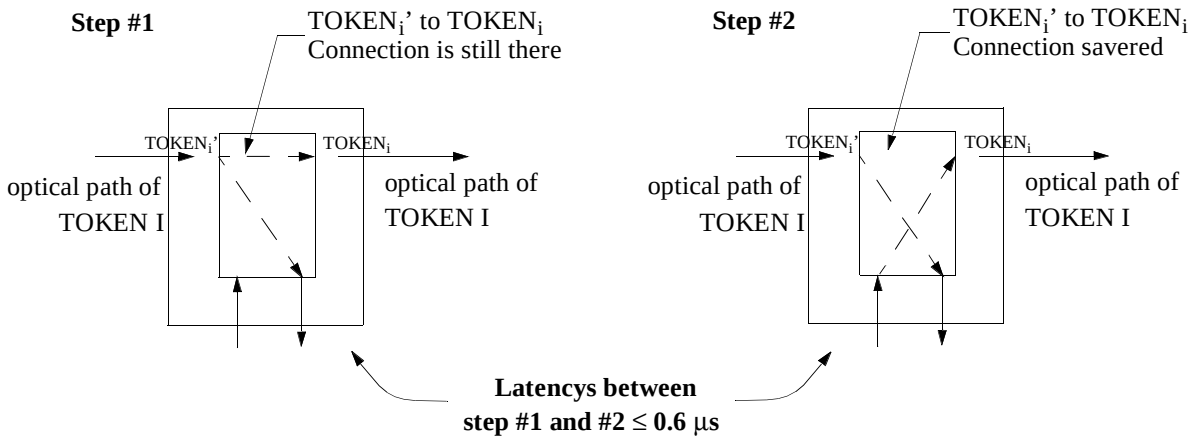


FIGURE 3: Two steps of the TOKEN grab process

In case of a loaded network (loads greater than 10% of maximum network load), the TOKEN latencies will be cumulative and can dramatically increase the average latency and reduce the network throughput. For example if the packet size is 128 Bytes, then the data transmission will consume 625 ns, while the grab latency will consume a minimum of 1.2 μs (even if only one processor is requesting the destination). In the above example less than 40% of the complete data transmission cycle is used for actual data transmission.

Note that there is a hazard concerning the integrity of a TOKEN because it may be passing through the switch during the switch setup. Seperate logic is required to avoid this.

1.2.4 Comparison and summary

Comparison between the VSC852 and the NARROW FAST implementations of a 1-bit wide, 4-processor per node, 16-node network.

Table 2: Comparison between VSC852 and Switch IC implementations

	VSC852 for data path and control	Switch ICs
Power consumption	33 W	< 8 W
Speed	1.6 Gb/s	2 Gb/s
Minimum switch setup delay	140 ns	10 ns
Path reservation latency:	1.2 μs (1200 ns)	80 ns
Pass-through latency	not known	4 ns

Moving to a smaller fabrication feature size (0.25 μm), and designing the control and datapaths as two separate ICs will further reduce the grab latency, by allowing higher data rates and lower flow-through latencies.

As seen above, designing a custom ICs for distributed control has speed, power and network latency/throughput advantages over commercial crosspoint switches such as VSC852. A network built using non-custom ICs VSC852, suffers most performance degradation when implementing media access control. In a network, overall performance is inversely proportional to the amount of time requests spend on control compared to the amount of time spent on data transmission. Thus, it is important to reduce the control latency and overhead. If commercial components are used, they should be used for the datapath and custom ICs should be built for media access control.

USC Program

VLSI Photonics Switch IC for a modular 64 Processor Architecture

2.0 Timeline

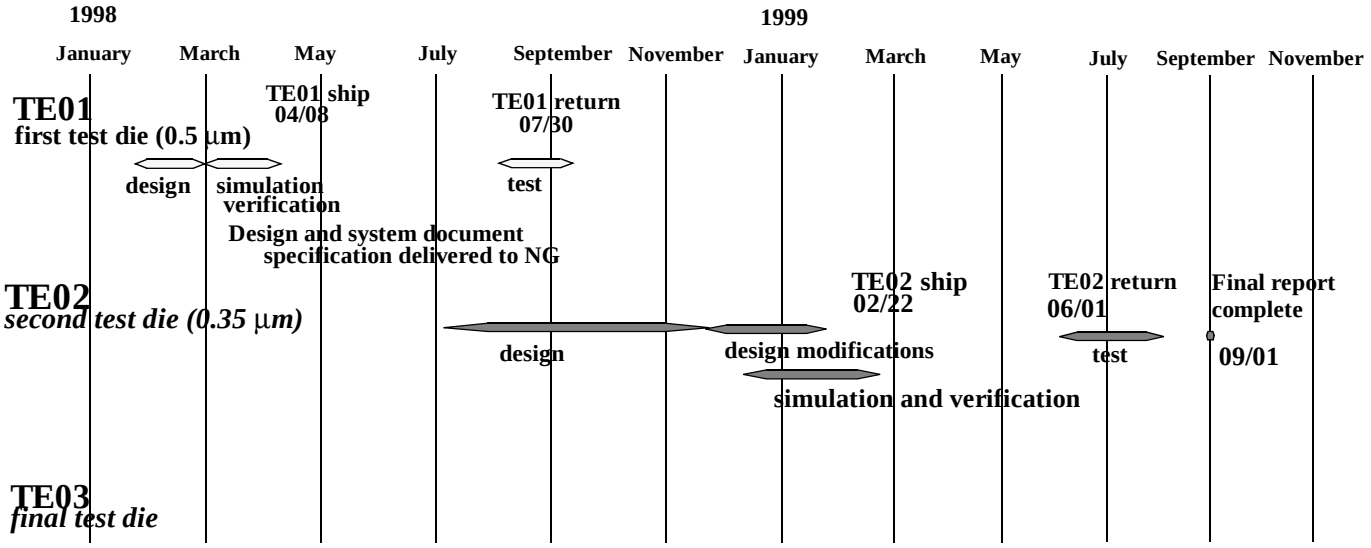


FIGURE 4: Timeline

3.0 Introduction

Microprocessors continue to increase in performance, bringing us ever closer to the era of tera flop computing through parallel processing. To exploit the advantages of parallel processing, high performance interconnection networks are needed. Fixed-path rings and buses, where each destination has a dedicated path, increases the network data bandwidth and reduces node latencies (compared to static and dynamic networks), by avoiding complex routing logic at each node. However, all electrical fixed path rings and buses are not built due to high wiring, packaging and interface costs (I/O bottleneck).

This bottleneck is overcome by using the improvements in integration of CMOS and optics, thus making fixed-path rings a viable alternative to more traditional all-electrical static and dynamic networks. An added advantage of having no internal routing logic is that the network can be customized using the external interface.

The successful implementation of such networks requires a reduction in receiver/transmitter power consumption to few milli-watts and an increase in the optoelectronics reliability. Also, switches need to be low-power, low-latency, low-cost, and scalable in both the number of nodes supported and the incremental cost of increasing network size. Finally we require low-latency, high-throughput and low-hardware overhead media access protocols that are scalable and expandable.

Compared to single-ended optics, differential or complementary optical signaling can improve network reliability. The adoption of logic styles such as low-swing pass-transistor logic together with current steering logic can improve the scalability, data rates and area and power consumption. A BUSY BIT media access protocol can be used to reduce the optoelectronic overhead while retaining the latency and throughput performance of dedicated tokens. Because of its low node latency (pin-to-pin) and reduced optoelectronic overhead, BUSY BIT protocol is scalable and can be expanded to support ACK/NACK, multi-cast and guaranteed reply bits.

4.0 Summary and design goals:

The USC VLSI-Photonics Switch IC (TE03) forms the basic block of a larger scalable switch network that is capable of supporting interconnect bandwidths in the Tb/s regime. This document presents the initial specifications for the 4-processor, 64 path Switch IC to be designed in the 0.35 µm HP C10 CMOS process. Each Switch IC will be placed on a processor board with 4 processors and the complete network will contain up to 16 processor boards.

The processors of a processor board interface to the Switch IC through the NG Formatter interface. The boundary between the Switch ICs and the NG Formatter board (electrical I/O interface) uses unidirectional (full-duplex) signaling. At the electrical I/O interface, a group of data signal lines assigned to a processor is called a port. Both at the optoelectronic interfaces and within the Switch IC a group of data signal lines assigned to a processor is called a path. The optoelectronic input interface interfaces the optical detector array to the Switch IC using 250 mVpp minimum differential signaling (unterminated). The optoelectronic output interface interfaces the VCSEL driver array to the Switch IC using 250 mVpp minimum differential signaling. Finally, the host control interface provides the host's control signals to the Switch IC using static CMOS signal levels.

4.1 Switch IC Overview

The USC VLSI-Phontonics Switch IC performs a distributed muxing operation. The Switch IC supports sixty-four 12b-wide optoelectronic input/output paths and four 12b-wide electrical I/O ports. Section 3 of this document discusses this architecture in detail. 16 of these Switch ICs, when configured in a distributed optoelectronic network can fully interconnect 64 processors in a non-blocking fashion. All 4 electrical ports can deliver data to the 64 optoelectronic output paths. On the other hand the electrical ports can only receive data from the first 4 optoelectronic input paths (optoelectronic input PATH 0 delivers data to PORT A, PATH 1 delivers data to PATH B, etc.). The Switch IC does not provide internal physical loop back and NG Formatter and the Race200 handle connectivity between processors of the same processor board.

The signaling between the electrical I/O interface (NG Formatter) and the Switch IC is unidirectional (full-duplex) LVDS, and is either destination terminated or source and destination terminated with 50 Ω resistors. The on-chip resistors are polysilicon with ± 6% (1-sigma) tolerance. The optoelectronic input interface has a maximum per line load capacitance of 100 fF. Also, the optoelectronic output interface is rated for a 100 fF output load. The host control interface uses CMOS signal levels.

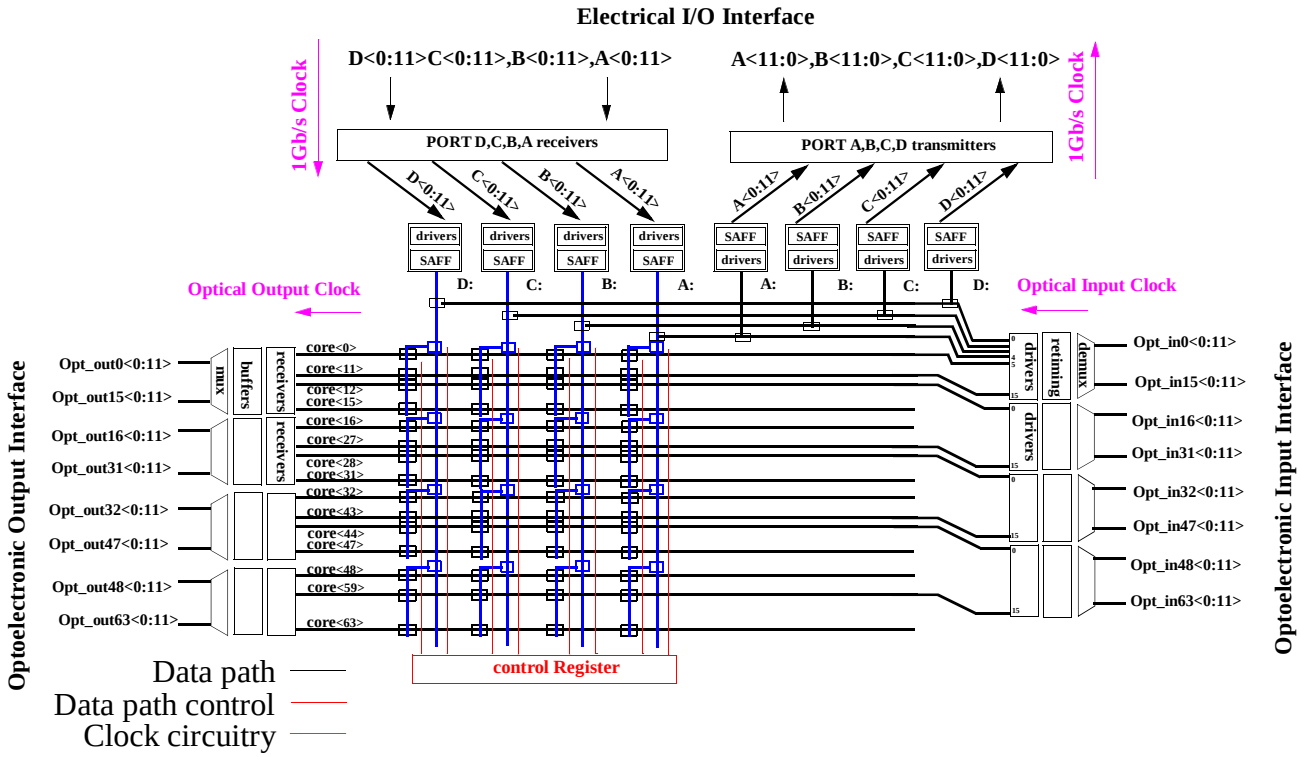


FIGURE 5: Top level view of the Switch IC designed using soft partitioning (clock circuitry, BUSY BIT circuitry and optical elements are not shown)

The minimum core height can be found by assuming a spacing of 3λ between signal wires

Height of core	= 64*12*2*2*1.2	= 3.7 mm
Width of core	= 2*12*2*2*1.2	
	+ 2*32*2*1.2	= 0.3 mm

The actual core height and width will be greater than the minimum height and width given above. The actual height and width is bounded by the sense amplifier flip-flop (SAFF), driver and 2x2 switch layouts.

The source of the Switch IC internal clock can be either the optical input clock or the electrical input clock. Because this is a reference-clocked system, the reference Switch IC receives its clock electrically from the reference host. All other Switch ICs in the network receives their clock optically (Section: 5.0). The nominal clock rate is 1 Gb/s for a 1 Gb/s data rate (i.e. data on both edges). The internal clock also runs at 1 Gb/s.

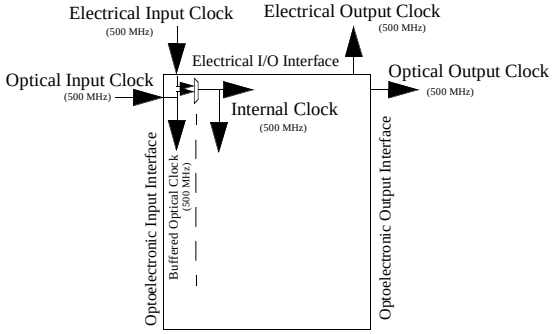


FIGURE 6: Different clock regimes

The optoelectronic input interface will provide data synchronized to the optical input clock. This data is internally retimed to the internal clock. Section: 5.0 of this document discusses clock synchronization issues. Both the electrical and optical output data are synchronized to the internal clock. The internal clock is buffered and provided as the electrical output clock at the electrical I/O interface (NG Formatter), as well as the optical output clock at the optoelectronic output interface. The electrical I/O interface will provide its input data synchronized to this electrical output clock

At the electrical and optoelectronic input interfaces the received 1 Gb/s data is 1:2 demuxed to a 500 Mb/s by the receivers. The internal switch data rate is 500 Mb/s and is 2:1 muxed and buffered at the output. To reduce power consumption and noise generation, the switch core is implemented in complementary low-swing pass-transistor logic.

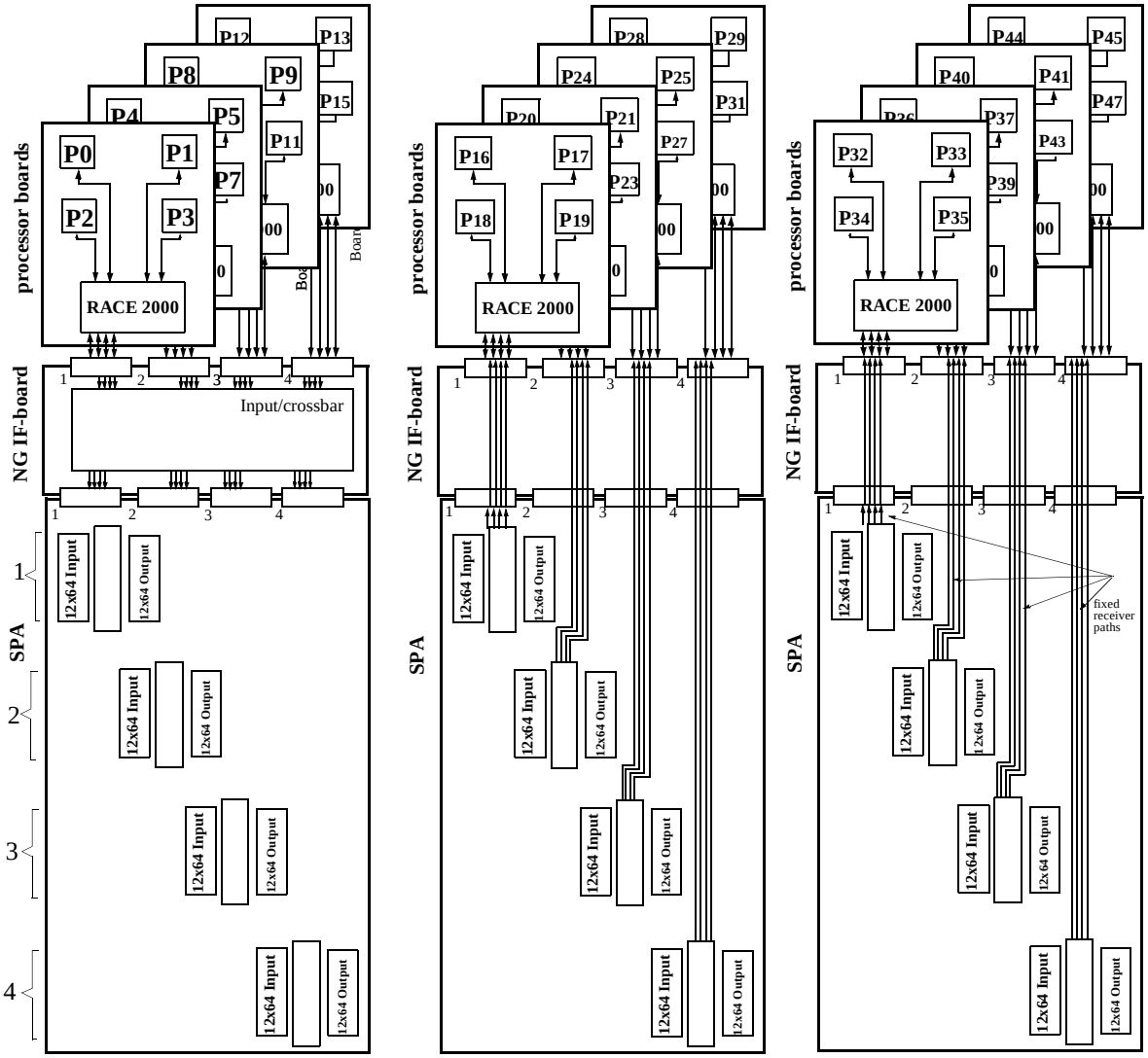


FIGURE 7: VLSI Photonics interconnect network using TE03 to create a 256 processor architecture

4.2 Specifications

- **Process technology 0.35 μm CMOS from HP, Power rail 3.3 V**

Table 3: Electrical characteristics

			Min	Typ	Max	Units
ELECTRICAL I/O INTERFACE - (LVDS)						
DIFFERENTIAL DRIVER CHARACTERISTICS						
V_{OD}	Output differential voltage	R_L=50 Ω, Figure 8	250	300	400	mV
V_{CM}	Output common mode voltage		1.1	1.25	1.6	V
V_{TT}	Biasing voltage for T_X		1.0		1.2	V
DIFFERENTIAL RECEIVER CHARACTERISTICS						
V_{TH}	Input threshold high	V_{CM}=0V to 2.4V			+125	mV
V_{TL}	Input threshold low		-125			mV
I_{IN}	Input current				10	μA
HOST CONTROL INTERFACE TTL/CMOS						
V_{TH}	Input high voltage		3.0			V
V_{IL}	Input low voltage				0.5	V
I_{IN}					10	μA
V_{OH}	Output high voltage		3.0			V
V_{OL}	Output low voltage				0.5	V
OPTOELECTRONIC INPUT AND OUTPUT INTERFACES						
DRIVER CHARACTERISTICS						
V_{OD}	Output differential voltage	C_L=100 fF	250	300	400	mV
ΔV_{OD}	V_{OD} magnitude change				50	mV
V_{CM}	Output common mode voltage		1.1	1.25	1.6	V
RECEIVER CHARACTERISTICS						
V_{TH}	Input threshold high	V_{CM}=0 V to 2.4 V			+125	mV
V_{TL}	Input threshold low		-125			mV
I_{IN}	Input current	V_{IN}=0 V to 2.4 V			10	μA
C_{IN}	Input capacitance	excluding external wiring	30	40	100	fF

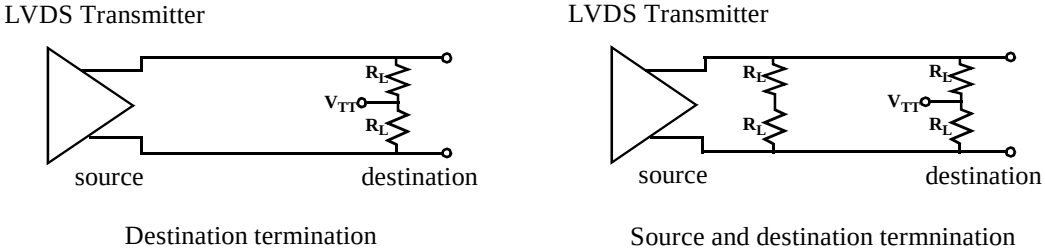


FIGURE 8: Electrical LVDS signal line with destination termination.

Table 4: Switching characteristics

			Min	Typ	Max	Units
ELECTRICAL I/O INTERFACE - (LVDS)						
DIFFERENTIAL DRIVER CHARACTERISTICS						
t_{RISE}	Output rise time	R_L=50 Ω, 10%–90%			0.3	ns
t_{FALL}	Output fall time				0.3	ns
t_{SKD}	Output data skew				0.2	ns
DIFFERENTIAL RECEIVER CHARACTERISTICS						
t_{RISE}	Input rise time	10%–90%			0.3	ns
t_{FALL}	Input fall time				0.3	ns
t_{SKD}	Input data skew	Relative to falling edge of clock_out (Figure 7)			0.2	ns
t_{HOLD}	Input hold time		0.3	0.34		ns
t_{SETUP}	Input setup time		0.0	0.1	0.3	ns
HOST CONTROL INTERFACE TTL/CMOS						
t_{RISE}	Input rise time				10	ns
t_{FALL}	Input fall time				10	ns
OPTOELECTRONIC INPUT AND OUTPUT INTERFACES (200 mVpp, 1.2V bias)						
DRIVER CHARACTERISTICS						
t_{RISE}	Input rise time	C_L=100 fF, 10%-90%			0.3	ns
t_{FALL}	Input fall time				0.3	ns
t_{SKD}	Data skew	C_L=100 fF			0.2	ns
t_{CHD}	Clock to output high delay		0.2		0.4	ns
t_{CLD}	Clock to output low delay		0.2		0.4	ns
RECEIVER CHARACTERISTICS						
t_{RISE}	Input rise time	10%–90%			0.3	ns
t_{FALL}	Input fall time				0.3	ns
t_{SKD}	Input data skew	Relative to falling edge of opto_clock (Figure 7)			0.2	ns
t_{HOLD}	Input hold time		0.3	0.34		ns
t_{SETUP}	Input setup time		0.0	0.1	0.3	ns

Switch IC input/output specifications

- **Electrical I/O interface (NG Formatter) to the Switch IC**

  Input/Output (LVDS)

  Total of 196 signal lines

  Total of 48 directional input data lines (96 signal lines, four-12b-wide ports)

  Total of 48 directional output data lines (96 signal lines, four-12b-wide ports)

  1 LVDS input clock (electrical input clock) (2 signal lines)

  1 LVDS output clock (electrical output clock) (2 signal lines)

The CMOS level full-swing signals interfacing the host control interface to the Switch IC are unidirectional.

- **Host control interface to the Switch IC**

  Input (CMOS/TTL)

  Data path control lines, total of 12 control lines

  8 Control lines to select the control latch for a specific electrical input port and its corresponding optical output path (electrical input optical output: EIOO)

  1 Control line to select synchronization loop back latches.

  1 Control line to identify a data path control latch load (data path control register load: DPCRL)

  BUSY BIT monitoring controls, total of 18 control inputs

  6 Control bits to identify the BUSY BIT that is being requested (BUSY BIT identify: BBI)

  1 Control bit for resetting the request bit (BUSY BIT reset: BBR)

  1 Control bit for setting the request bit (BUSY BIT set: BBS)

  6 Control bits to identify which grab interface register bit is selected (grab interface select: GIS)

  1 Control bit to load grab register to the grab interface register (load grab interface: LGI)

  1 Control bit to identify the valid input phase for FRAME and BUSY_IN (FRAME/BUSY_IN phase shift: TBPS)

  1 Control bit to identify Master IC (master IC: MIC)

  1 Control bit to identify FRAME creation (FRAME creation input: FC_IN)

  BUSY BIT monitoring outputs, total of 3 output bits

  1 Bit for grab interface register output (grab interface register output: GIRO)

  2 Bits for FRAME (decoded) output to the processor board (FRAME electrical out: TEO±)

  Clock control, total of 3 control lines

  1 clock select line for selecting the source of the internal clock between the electrical input clock and the optical input clock (internal clock select: ICS)

  1 clock select line for selecting the source of the optical output clock between the buffered electrical input clock and the internal clock (optical output clock select: OOCS)

  1 control line for selecting 0 or 180 degrees phase shift (clock phase shift: CPS)

  Analog inputs

  Total 6 analog input lines

  1 analog differential input for optical input clock delay element

  1 analog differential input for electrical input clock delay element

  2 analog inputs for adjusting the duty cycle of the optical input clock, and the internal clock

- **Optoelectronic selection control for spare clock, FRAME, and BUSY BIT optical lines**

  Input (CMOS/TTL)

  Main versus spare optical line control, total of 3 control lines
  - 1 Control line to identify which optoelectronic clock line should be selected for optical input clock (optical input clock select: OICS)
  - 1 Control bit to identify which optoelectronic input line should be selected for FRAME (FRAME input select: TIS)
  - 1 Control bit to identify which optoelectronic input line should be selected for BUSY_IN (BUSY_IN input select: BIS)

NOTE: The above CMOS/TTL controls can also be implemented as LVDS, in which case the number of control line inputs will increase by a factor of 2.

The 250 mVpp minimum, 1.2 V bias (unterminated LVDS) signals interfacing the optoelectronic input interface and the Switch IC are unidirectional.

- **Optoelectronic input interface to the Switch IC**

  Assume a 12x64 logical array optical device

  Total of 774 LVDS voltage level inputs, with a 250 mVpp minimum at 1.2 V bias
  (1548 signal lines)

  12b-wide, 64 input data paths go into the switch core (1536 signal lines, Figure 14).

  4 BUSY BIT lines (8 signal lines)

  1 FRAME line (2 signal lines)

  1 clock input for optical clock (2 signal lines)

  The interface between the optoelectronic input interface and the Switch IC consists of 12x64x2 = 1536 signal lines on a 4.0 μm pitch. With respect to the optical input clock the signals are deskewed to better than 200 ps.

  Input requirements for optoelectronic input interface to the Switch IC

Input load (for demux 2:1)	< 100 fF not including external wire loads	
Maximum skew	< 0.20 ns	
Rise/fall time (10%-90%)	< 0.30 ns	
Nominal input voltage swing	> 250 mV_{pp}	(250 mV_{pp} minimum at 1.2 V bias)

The 250 mVpp minimum, 1.2 V bias signals interfacing between the optoelectronic output interface and the Switch IC are unidirectional.

- **Optoelectronic output interface to the Switch IC**

  Assume 12x64 logical array optical device

  Total of 774 LVDS voltage level outputs with a 250 mVpp minimum at 1.2 V bias
  (1548 signal lines)

  12b-wide, 64 output ports (1536 signal lines)

  4 BUSY BIT lines (8 signal lines)

  1 FRAME line (2 signal lines)

  1 clock output for internal clock (2 signals)

  The interface between the optoelectronic input interface and the Switch IC consists of 12x64x2 = 1536 signal lines on a 4.0 μm pitch. With respect to the optical input clock the signals are deskewed to better than 200 ps.

  Each output is retimed using the on-chip 1 Gb/s internal clock.

- **Switch IC, switch core specifications**

  The switch performs a muxing operation between the electrical input ports and optical output paths.

  PORTs A, B, C and D as well as the electrical input and output clocks connect to the Switch IC from the top.

  The optoelectronic inputs are received from the right and are demuxed before entering the switch core.

  The output of the switch core is muxed (2:1) and delivered to the optoelectronic output interface.

  Each PORT/PATH is 12 bits wide

- 8 bits are data
- 1 bit FRAME
- 2 bits are parity
- 1 bit is reserved for control.

The data path and BUSY BIT control as well as switch control inputs connect from the bottom of the IC. The power is primarily received from the top and the ground is primarily received from the bottom
500 Mb/s internal data rate, full speed clocking.
1 Gb/s internal clock.

- **Even though the input and output arrays are logically 12 x 64, any physical arrangement is allowed.**
- **BUSY BIT monitoring is discussed in Section: 8.2.2 under TE02 IC Functionality.**

Table 5: HOST CONTROL switch configuration inputs (excluding any address control lines)

Control line						Description
Data path control						
SLL	DPCRL	RESET				Description
High	-	-				Select synchronization loop-back latches
Low	-	-				Select the electrical in, optical output connect latches
	High	-				Load values to selected data path latch
-	-	High				Reset Switch IC (Initialize the IC)
-	-	Low				Normal operating mode
Clock control						
Reset	ICS	OOCS	CPS			Description
-	High	-				Select electrical input clock for internal clock
-	Low	-				Select optical clock for internal clock
-	-	High				Select internal clock as the optical output clock
-	-	Low				Select buffered electrical input clock as the optical output clock
-	-	-	High			Phase shift clock by 180 degree
BUSY BIT Monitoring						
BBR	BBS	MIC	FC_IN	TBPS	LGI	Description
High	Low	-	-		-	Reset request bit
High	High	-	-		-	Set request bit
High	High	-	-		-	Not allowed
-	-	High			-	Identify the IC as the master/reference IC
-	-	Low	-		-	Identify the IC as a slave IC
-	-	High	High		-	Inhibit the flow of the FRAME in the network (if MIC)
-	-	High	Low		-	Reset the FRAME
-	-	-	-	High	-	Select negative edge of clock as the valid phase
-	-	-	-	Low	-	Select positive edge of clock as the valid phase
-	-	-	-	-	Low	Load grab register to the grab interface register

Above table does not include the analog controls for the two delay elements.

On-Chip Terminations.

On-Chip 50 Ω LVDS termination resistors are included in the SA receivers receiving data from the electrical input interface (NG Formatter). These resistors are 28.3 squares of salicided polysilicon having a nominal sheet resistance of 2.0 Ω per square. Although HP specifies the process limits for polysilicon sheet resistance from 1 - 12 Ω per square, the mean and standard deviation of poly sheet resistance for wafers selected by MOSIS to represent each run is 2.0 ± 0.103 Ω per square for the 35 0.5 micron runs since 1995 (from on-line run summary data (*.prm files)). The minimum and maximum poly sheet resistance of these representative wafers is 1.8 - 2.2 Ω per square. (MOSIS only publishes data for a single selected wafer per run chosen to be representative of all the wafers in a run). MOSIS internal test data of a recent run give the standard deviation of poly sheet resistance within the run as 0.12 Ω per square (run N81D, 1.76 ± 0.12 Ω per square, 10 wafers, 108 sites). With this standard deviation, we expect the following sheet resistance tolerances:

Table 6: Poly-resistor sheet resistance tolerance

# sigmas	Tolerance(%)
1	± 6.82
2	± 13.64
3	± 20.45

HP specifies the temperature coefficient of poly resistance as 0.0042 Ω per degree C.

No terminations are used at the interface between the Switch IC and the optoelectronic input interface or the optoelectronic output interface, since the short wire lengths do not warrant use of transmission lines.

4.3 Area and power consumption estimates

- **Layout estimates(0.35 μm process)**

Switch IC area (y-x of Figure 19)	3.0 mm x 6.00 mm
Optical input array (excluding pads)	5.0 mm x 10.0 mm
Optical output array (excluding pads)	5.0 mm x 10.0 mm
Total area:	13.0 mm x 10.0 mm

- **Power consumption estimates (rms., calculated for 0.5μm)**

Electrical I/O interface (NG Formatter)

Transceiver (Tx portion)	1.40 W (x 2 for source and destination termination)
Transceiver (Rx portion)	0.27 W
Clocking	0.15 W
Muxing + buffering	0.30 W
SAFF + drivers before after switch	0.40 W
Total	2.52 W

Optoelectronic input interface (can be reduced by 25% or more if a larger input swing is received at the optoelectronic interface)

Rx (SAFF + Drivers)	4.20 W
Rx clock	1.50 W
Retiming registers	0.20 W
Total	5.90 W

Optoelectronic output interface

SAFF + mux circuits	4.20 W
Buffers + LVDS level drivers	2.00 W
Clock tree	2.00 W
Total	7.20 W

TOTAL POWER	15.62 W (17.0 W for source & destination termination)

The above estimates exclude power consumed by the optoelectronic input/output arrays. We also exclude the power consumed by the control logic, assuming that the time between switch configuration setups is greater than 100 of the 1 Gb/s clock cycles. In addition the power consumption of delay elements are also not included in the above power estimate.

5.0 Clocking and data synchronization

Globally clocked synchronous switching is adopted for the Switch IC, and a reference clocking strategy is adopted for hte network. The Appendix: A discusses the issues favoring the Switch IC's clocking strategy, globally clocked synchronous switching (GCSS), over the alternatives of unclocked asynchronous switching (UAS) and locally clocked asynchronous switching (LCAS). This section discusses the operation of the GCSS scheme and the refernce clocking strategy.

lobal clocked synchronous switching (GCSS) in a reference clocked network

Because the UAS and CLAS clocking strategies aren't suitable for large Switch ICs, a globally clocked synchronous switching (GCSS) strategy is adopted for the Switch IC. This scheme uses one central clock (internal clock) distributed throughout the switch. The use of a clock allows the switch architecture to take advantage of complementary low-swing pass-transistor logic and other dynamic techniques to reduce power and noise. The GCSS strategy avoids the need for a 1 Gb/s switch, multiple clock receivers, transmitters, buffer trees and clock delay elements required in LCAS strategy.

However the GCSS clocking scheme must deal with two types of synchronization boundaries. That between the buffered version of an input clock and its associated data valid timing window (data synchronization) and that between the internal clock (when its source is the electrical input clock - i.e. in the reference Switch IC) and the latched optoelectronic input data (clock domain synchronization).

In a reference clocked network, all Switch ICs are clocked by a derivative of a central reference clock, received electrically by the reference Switch IC and transmitted optically to the slave ICs. As a result, the reference Switch IC has both types of synchronization boundaries while slave Switch ICs only have a data synchronization boundary (Figure 9).

Since data is retimed at the input as well as the output optoelectronic interfaces, the network jitter and skew is limited to that of a single Switch IC and does not accumulated across Switch ICs. To reduce the internal clock skew, a balanced H-Tree structure is used to distribute the clock inside the Switch IC.

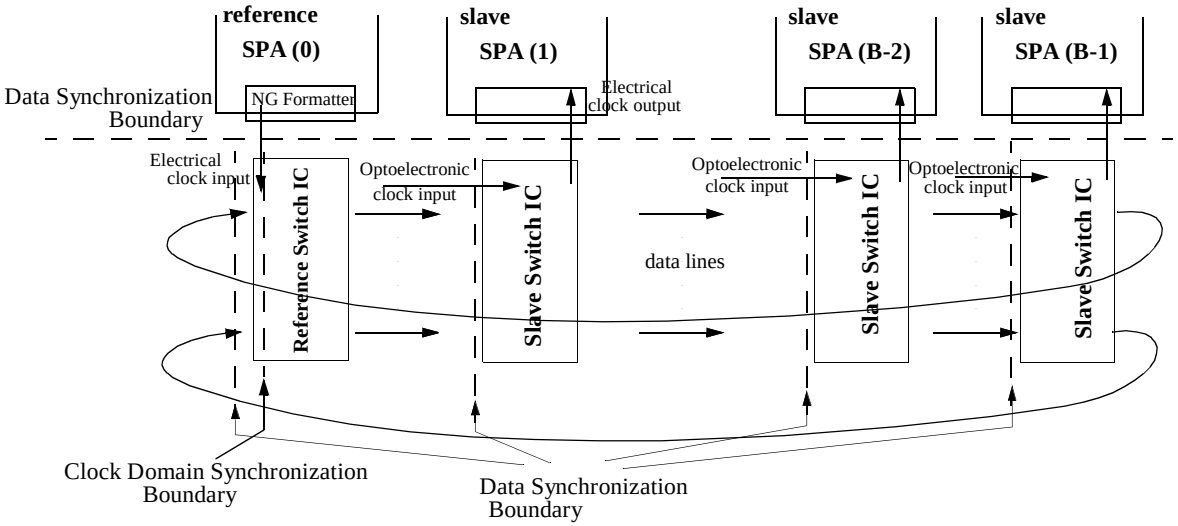


FIGURE 9: Synchronization boundary where all but reference Switch receives clocked optically

Figure 10 outlines the clock domains in the Switch IC. Two variable-delay elements as well as a selectable 180-degree phase shifter are provided in order to synchronize clock latching edges to their data during the initialization process.

The initialization process has 4 steps: Step 1 is reference electrical input data to internal clock synchronization. Step 2 is slave optoelectronic input data to buffered optical clock synchronization. Step 3 is reference optoelectronic input data to buffered optical clock synchronization. Step 4 is reference electrical input data to internal clock synchronization.

- **Set #1**
  The reference board adjusts the reference Switch IC's electrical input clock (EIC) delay element until the electrical data is properly received at the electrical I/O interface by the internal clock (as determined by proper data being transmitted by the Switch IC at the electrical I/O interface).

- **Set #2**
  Each slave board adjusts each slave Switch IC's optical input clock (OIC) delay element until the optical data is properly received at the optical input interface by the buffered optical clock (as determined by proper data being transmitted by the Switch ICs at the electrical I/O interface). This is possible because the Switch IC is designed to delay-match the buffered optical and internal clocks within 100ps, when the source of the internal clock is the optical input clock.

- **Set #3**
  The reference board switches the reference Switch ICs internal clock source to be the optical input clock. It also sets the source of the optical output clock (OOC) to be the buffered electrical input clock (BEIC). Then the reference board adjusts the reference Switch IC's OIC delay element until the optical data is properly received at the optical input interface by the buffered optical clock (as determined by proper data being transmitted by the Switch IC at the electrical I/O interface).

- **Set #4**
  The reference board switches the internal clock source back to the electrical input clock and then can enable a 180-degree phase-shift to select the clock edge that best latches the outputs of the optical input latch.

For further details on the initialization process, refer to Appendix: B.

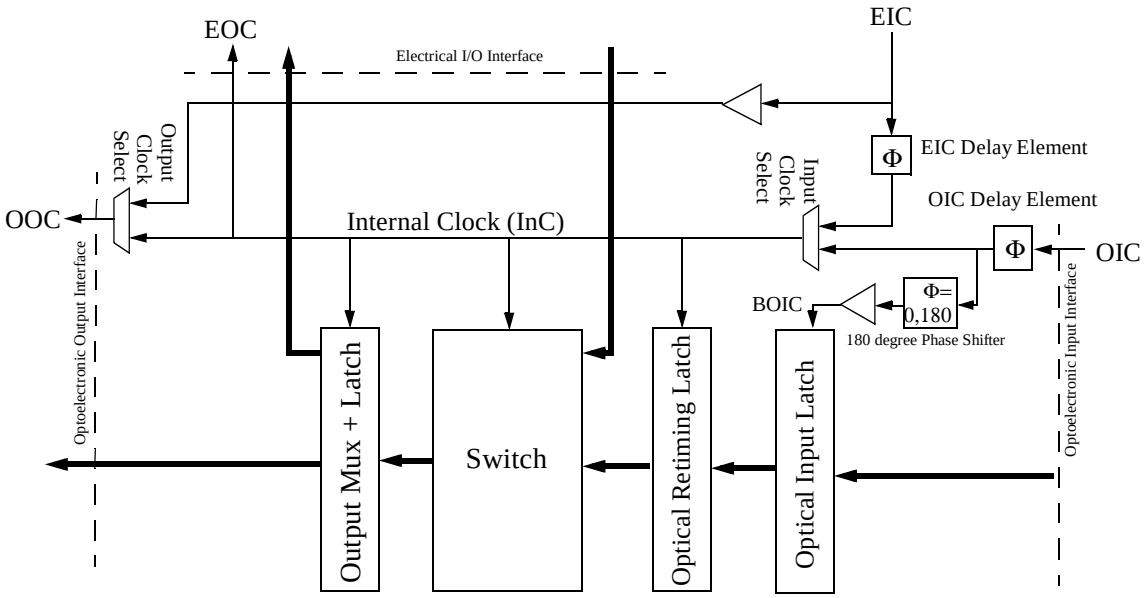


FIGURE 10: Abstract diagram of the Switch IC clocking domains. OIC = Optical Input Clock, EIC = Electrical Input Clock, EOC = Electrical Output Clock, OOC = Optical Output Clock, BOIC = Buffered Optical Input Clock.

In all Switch IC's except the Reference, the NG Formatter provides electrical input data synchronized to the electrical output clock (EOC).

5.1 Internal clock routing

The internal clock can be generated and synchronized to either the optoelectronic input clock or to the electrical input clock. A mux (1 Gb/s) is used to select between the two possible input clocks and the resulting internal clock is then buffered to drive the internal circuitry. The optoelectronic input clock is also buffered independent of the internal clock and is used to drive the optical input latch (demux-SAFF) at the optoelectronic input interface (Figure 13). A tree structure is adopted for both clocks to reduce the skew.

Low-power, low-noise, differential inverters (based on current switch logic) and static CMOS clocking are being considered for the clock tree buffers. The clock tree buffers designed using current switch logic (CSWL) generates less noise compared to traditional static CMOS buffers. Since the SAFF are designed using pre-discharge CSWL logic, the clock load is dependent on the magnitude of the clock swing. A clock input with a larger clock swing (full-rail 3.3 V) require smaller pull-down transistors, while a reduced swing clock (2.0 V) require larger pull-down transistors. Thus, a leaf node with a 2.0 V swing will see 20 - 30% more clock load against a 3.3 V clock.

The question that remains to be answered is whether the clock tree buffers could be designed using CSL logic and whether the actual leaf nodes can be designed using static CMOS. The noise and power of the current switch logic inverters will scale down with Vdd almost as much as static CMOS, but with a slightly higher lower bound for Vdd since the biasing transistor raises the Vdd requirement a few tenths of a Volt.

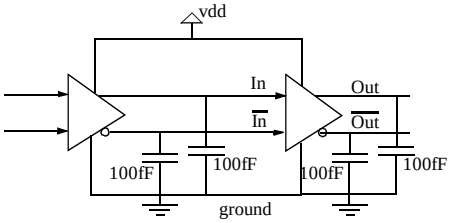


FIGURE 11: Simulation test setup (depth of chain = 2, wire load = 100fF)

Table 7: Comparison between Inverters for 3.3V for 0.5 μm HP CMOS

	Current switch logic	Current steering logic	static CMOS	units
Wire load (fig)	100	100	100	fF
Input high/low	1.6/0.0	1.6/0.0	3.6/0.0	V
Output high/low	1.6/0.0	1.6/0.0	3.6/0.0	V
Switching noise(ground)	0.13	0.14	0.92	mA
Normalized switching noise	1	1	9.2	
Power consumption	1.9	3.0	2.8	mW
Delay	0.22	0.2	0.2	ns
Rise/Fall (10%-90%)	0.45	0.45	0.5	ns

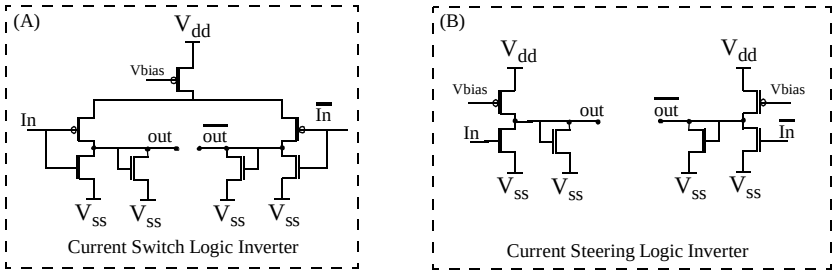


FIGURE 12: Illustration of (A) current switch and (B) current steering logic inverters

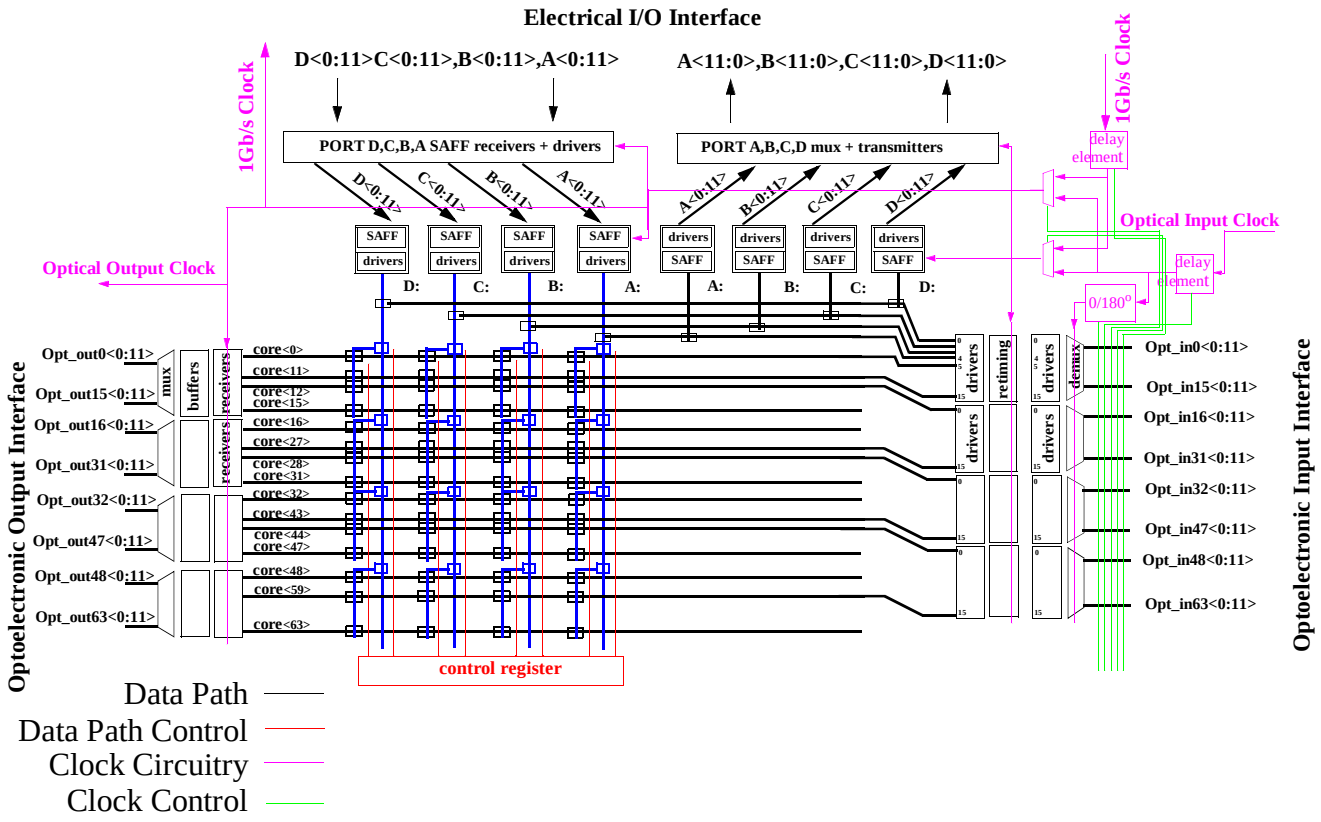


FIGURE 13: Clock routing inside the Switch IC (excluding BUSY BIT and optical arrays)

6.0 Major blocks of the Switch IC datapath

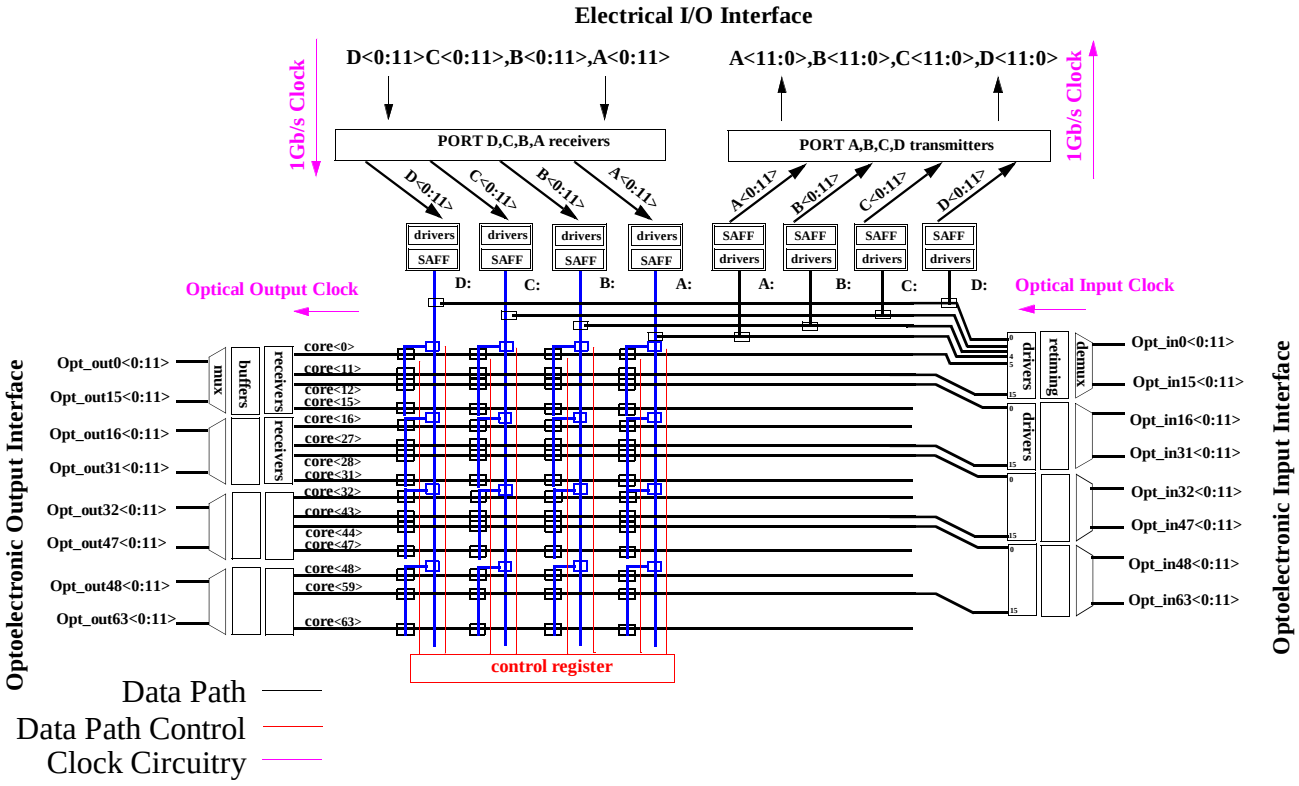


FIGURE 14: Major blocks of the Switch IC data path for the rigid partitioning option proposed in section "Scaling issues for TE02 (preliminary TE03 design considerations)" on page 74

Figure 14 shows the major blocks of the Switch IC datapath. The 6 major blocks of the Switch IC datapath are the input stage (1:2 demux) and SAFF, current-steering logic driver, pass-transistor switch core, output stage (2:1 mux), driver for optoelectronic output and the LVDS transmitter. It also shows the signal levels at the interface between each of the blocks. Current steering logic is used in the input stage (1:2 demux), output stage (2:1 mux) sense-amps, internal low-swing drivers. Low-swing complementary pass-transistor logic is used in the switch core. The drivers of the LVDS transmitter as well as the drivers for optoelectronic outputs are differential circuits driven by static CMOS buffers.

6.1 Input stage (demux) and SAFF

In conventional designs most receivers are either analog receivers or are limiting amplifiers. Such receivers are unclocked and the output is equal to the input data rate. The output of the receiver is then latched/demuxed by additional circuitry. However, by incorporating the demuxing function in the receiver, and by building a receiver using sense-amps, it is possible to reduce the area and power consumption of the receiver stage. In case of large number of inputs even a few milli Watts difference in power consumption can add up to couple of Watts of power savings (per bit reduction of 1 mW is equal to 0.8 W power reduction for the complete Switch IC).

In order to reduce the noise generated by the receiver, a new sense-amp has been designed using current steering logic (CSL-SA). The output of the CSL-SA is latched using a CSL SR flip-flop. The sizes for the SA is chosen based on the operating frequency and the internal clock amplitude (TE02 clock amplitude is a 0.7 Vdd). The CSL-SAFF generates less than 1/8 of the noise of conventional SAFF schemes, while its power consumption is no more than 1.5x to 2x larger than the conventional SAFF schemes (for 0.5 µm HP CMOS).

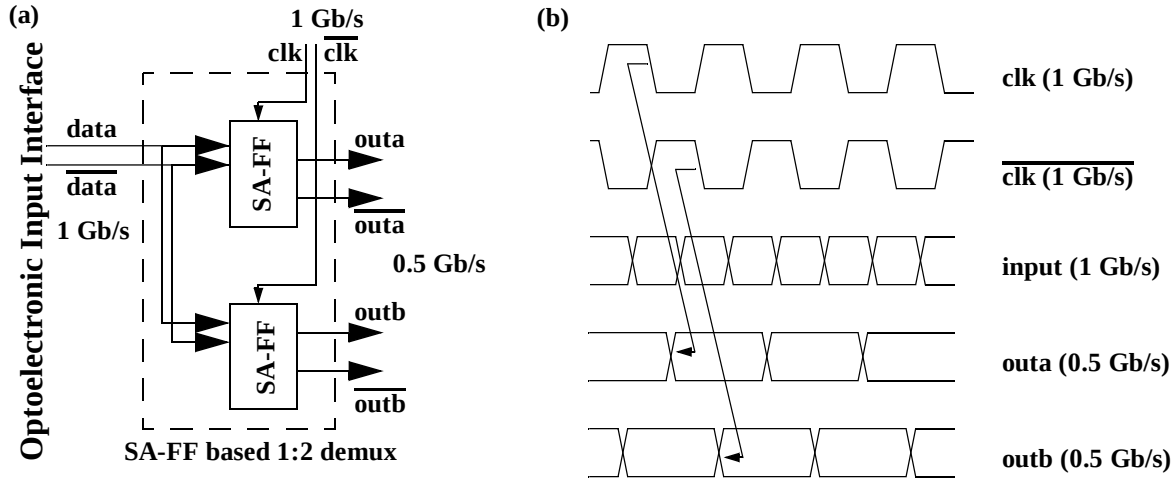


FIGURE 15: Demuxing using SA-FF

Simulations using a model of the 0.5 μm process slow-corner indicate that a 1.18 Gb/s CSL-SAFF consumes less than 5 mW per SAFF (10 mW under a demuxing receiver configuration). For 0.35 μm process this value is expected to reduce to 3.5 mW per bit (7 mW for a demuxing receiver), and 14 mW per SAFF for 2.0 Gb/s. This is a significant power reduction over a traditional receiver-demux-latch combination. Also, unlike conventional analog receivers, the CSL-SAFF receiver can handle input common-mode voltage ranging from 0.0 V - 2.4 V. A comparison of different receivers are given in Table 8.

Table 8: Comparison of receivers (with or with out 50 Ω termination, *estimates for 0.5 μm HP CMOS*)

	CSL-SAFF (new)	Tr.Receiver + FF	SA (new) + TSPC latch	SAFF (new)	SAFF Traditional	
Differential input	200	200	200	200	200	mV
Common-mode input	0.0 - 2.4	0.8 - 2.0	0.0 - 2.4	0.0 - 2.4	1.0 - 3.0	V
Output high	2.0		3.6	3.6	3.6	V
Output low	0.0		0.0	0.0	0.0	V
Precharge (SA)					x	
Predischarge (SA)	x		x	x		
Clock to high delay	0.3	0.8	0.7	0.3	0.4	ns
Clock to low delay	0.3	0.8	0.7	0.3	0.4	ns
Relative clock load	1	1	4	2	3	
Power consumption	10	20 - 30	5	5	9	mW
Total power per port	0.10	2.2 - 3.3	0.06	0.06	0.10	W
Relative noise	1	1 - 2	6	7	7	
Relative area	1	10	0.8	0.8	1.2	

6.2 Dynamic driver

To reduce the Switch IC clock loading and noise generation, a current steering logic unclocked driver is used instead of domino logic or any other clock logic scheme. (In another implementation of low-swing differential logic using pass-transistor chains, the driver was implemented in domino logic [4]. Even though such a driver generates less rms power, it can generate a considerable amount of noise and clock loading).

The current steering driver has no clock load and generates less than 1/4 of the noise of a conventional dynamic driver. The reduction in generated noise comes at the cost of higher power consumption. A current steering driver consumes 2x times more power compared to a domino driver. Comparison between current steering and domino drivers is included in Table 9.

Table 9: Comparison between CS drivers and Domino drivers, driving a 22b-wide 8x8 switch

	Current-steering driver		Domino driver		Units
	Max	Min	Max	Min	
Input voltage	0.3 - 2.4	0.2 - 1.9	0.3 - 2.4	0.2 - 1.9	V
Clock input	1.0 - 3.0	1.1 - 2.9	1.0 - 3.0	1.1 - 2.9	V
Setup time	-	-	0.1		ns
Output voltage (50 MHz)	0.2 - 2.4		0.0 - 3.6		V
Output voltage high (1.18 Gb/s)	800		800		mV
Output voltage low (1.18 Gb/s)	50		50		mV
Power/bit (200 fF wire load)	2.5		1.5		mW
Power/bit (20 fF wire load)	1.75		1.0		mW
Total power	340		187		mW
Relative power	1		0.55		
Noise/bit (200 fF wire load)	0.05		0.35		mA
Noise/bit (20 fF wire load)	0.075		0.2		mA
Total noise	11		40		mA
Relative noise	1.0		3.6		
Clock load/bit	0	0	26		fF
Relative area	1.0		0.7		
Power rail	3.60		3.60		

Even though each individual CSL driver consumes 2x more power, when the power consumption of the clock is taken into account both CSL and domino schemes consume about the same power. This is seen in Table 9, where the domino driver has 26 fF of clock load against the 0fF load of CSL drivers.

6.3 Output stage (2:1 mux)

The 200 mVpp output of the pass-transistor chain is sensed using a CSL-SA (Figure 16). The outputs of the SA are then fed to a CSL NOR gate. Since pre-discharge type SAs are used for sensing the pass-chain output, the CSL NOR gate acts as a multiplexer. An added advantage of the above scheme is that it only requires a clock input for the sense-amp, because the multiplexing stage (CSL-NOR gate) does not require a clock.

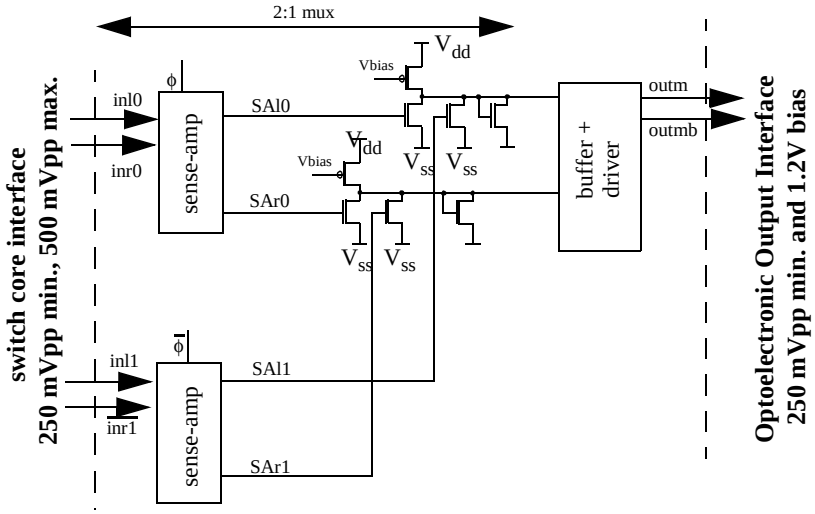


FIGURE 16: 2:1 Mux using SA and current-steering logic NOR gates

Abstract timing diagrams of the outputs of the switch core, CSL-SA, and the final mux are shown in Figure 17.

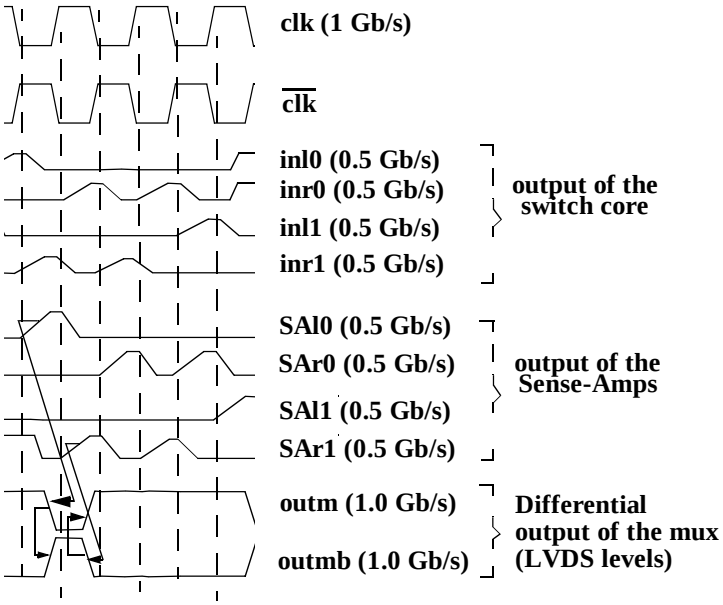


FIGURE 17: Data timing diagram for switch core through output mux

6.4 Buffer and driver for optoelectronic output

The current steering logic mux is designed for a load of less than 50 fF. Also, the output logic 0 of the mux is 0.0 V, and the output logic 1 is 2.0 V. Since a driver producing differential outputs needs a greater input swing, the output of the mux is buffered to full-rail prior to the differential circuitry. The differential output driver can provide a 250 mVpp minimum differential output (LVDS levels), at 1.2 Gb/s, into a 100 fF load. The common-mode voltage of the driver can be between 0.8 V and 1.6 V, thereby accommodating process variations.

Note that no 50 Ω terminations are used at either end of the optoelectronic input and output interfaces.

6.5 LVDS Transmitter

A simplified version of a conventional current mode transmitter is shown in Figure 18.

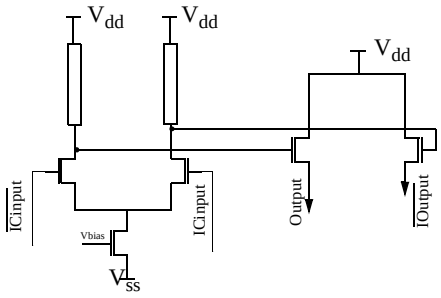


FIGURE 18: Abstract diagram of a conventional current mode transmitter.

The current mode LVDS transmitter used will be modified version of the transmitters designed and tested by Bindu Madhavan. The transmitters are designed for destination termination, and the power consumption of the transmitter is dominated by current necessary to swing 250 mVpp across 50 Ω termination resistors (the power consumption > 5xVdd mW).

As shown in Figure 8 the electrical LVDS transmitters are 50 Ω terminated at the destination. If the transmitters were both source and destination, the power consumption will be increased by a factor of two.

7.0 Ceramic package and cavity

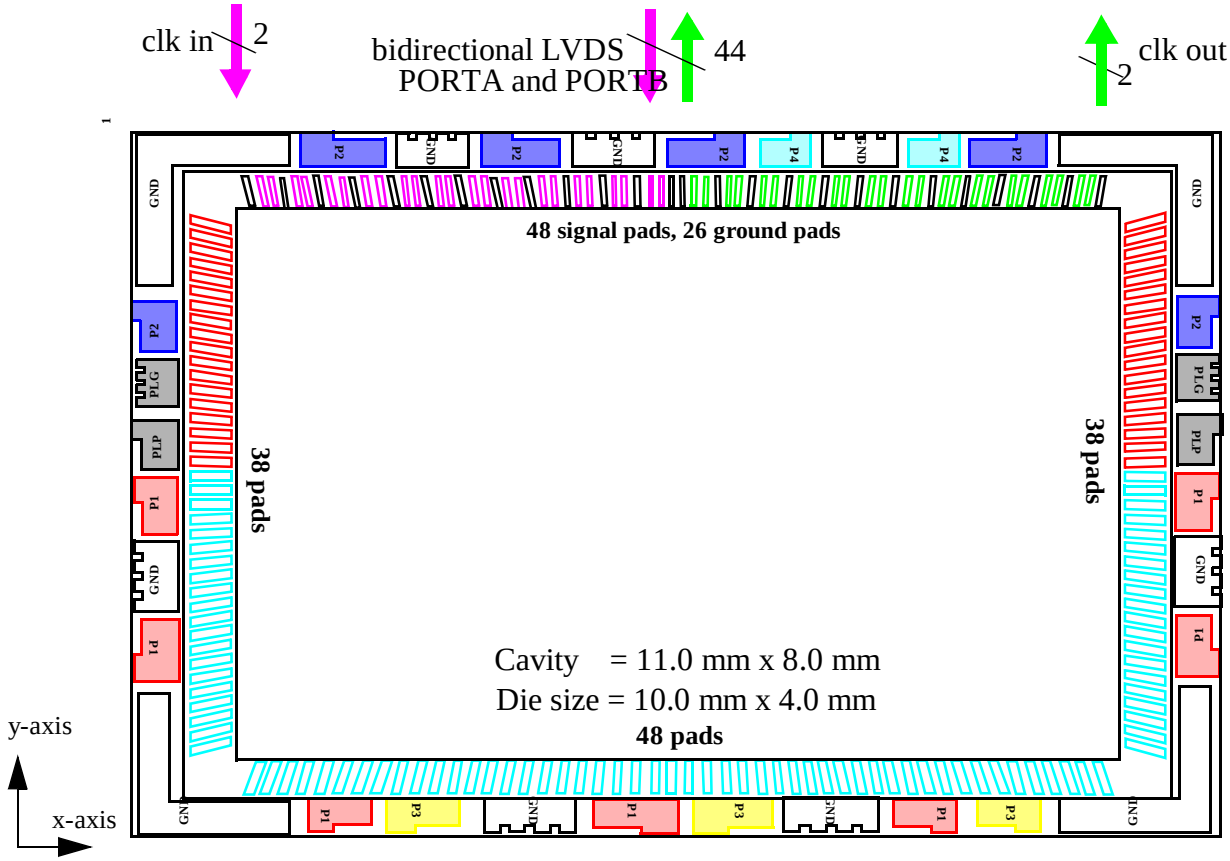


FIGURE 19: PAD OUT (cavity up) view of proposed package for the CROSSBAR IC electrical/optical test. LVDS signal pads arrive at the top and sides. Signal pads in red are control and test signals. Signal pads at the top are coupled 50 Ω differential pairs and every pair is separated by a ground line. The signals on the other three sides are 50 Ω single-ended lines. On the upper terrace, cyan is Power 4, yellow is Power 3, blue is Power 2, red is Power 1.

- **Even though the input and output arrays are logically 12 x 8, any physical arrangement is allowed, such as a 10 x 10 array.**
- **The minimum horizontal spacing between the arrays must be 2.4 mm to allow room for vertical routing of the 24x2x2 electrical I/O interface signals, power and clock lines, as well as the internal Switch circuitry**
- **The optoelectronic outputs of the Switch IC are deskewed to the point of delivery, which is the left edge of the VCSEL Array (to the VCSEL driver circuitry inputs).**
- **The optoelectronic inputs are deskewed by the detector circuitry and delivered to the Switch IC at the right edge of the receiver array.**

Even though this package will be used for testing TE02, at the moment we have not determined whether it will be used for TE03 electrical testing. Even though the switch core and related circuitry will require a height less than 7 mm, the actual height of the IC will be dependent on the optical arrays.

Further information on the package and cavity is given in Appendix: C.

8.0 Testing the Switch IC test chips

8.1 TE01 design and test

8.1.1 Introduction

The TE01 (te01_ln) test IC is designed to evaluate the feasibility of using sense-amp receivers and low-swing complementary pass transistor logic for a Switch IC. The IC is designed for a data rate of 1.25 Gb/s, a data input voltage of 250 mVpp and a common mode input of 1.2 V.

Sense-amp receivers are used instead of the traditional receivers to reduce the total electrical I/O interface power. The transmit circuitry and the clock receiver are from Bindu Madhavan, and have been tested on a previous run.

Chip size	3.75 mm^2 (2499 µm x 1491 µm)
Ship date	04/08/98
Return date	07/30/98

8.1.2 IC diagrams

Test die

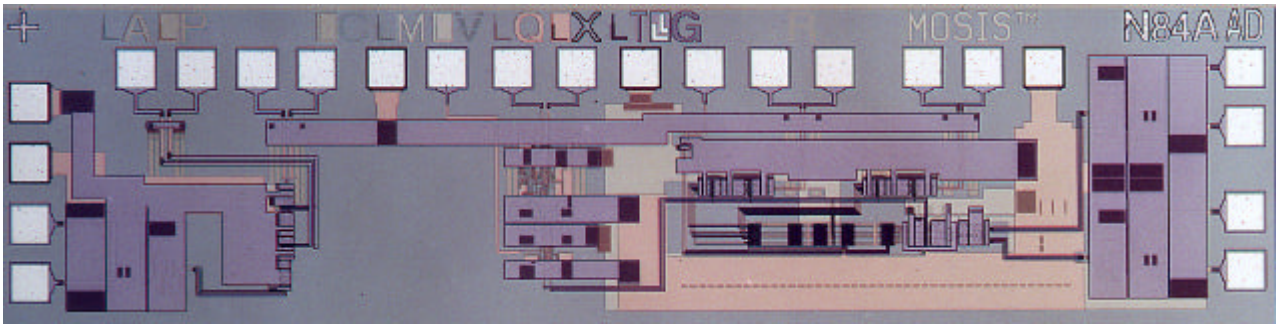
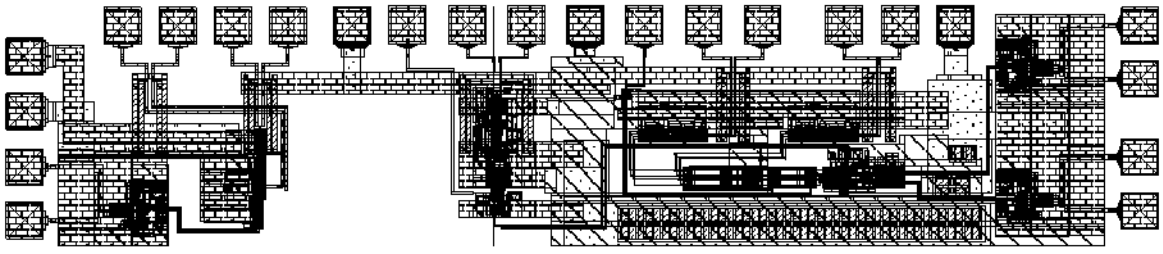


FIGURE 20: Photograph of the test die (te01_ln)

Layout:



SECTION 01 **SECTION 02**

FIGURE 21: Layout of TE01 showing sections 1 and 2

Bonding diagram

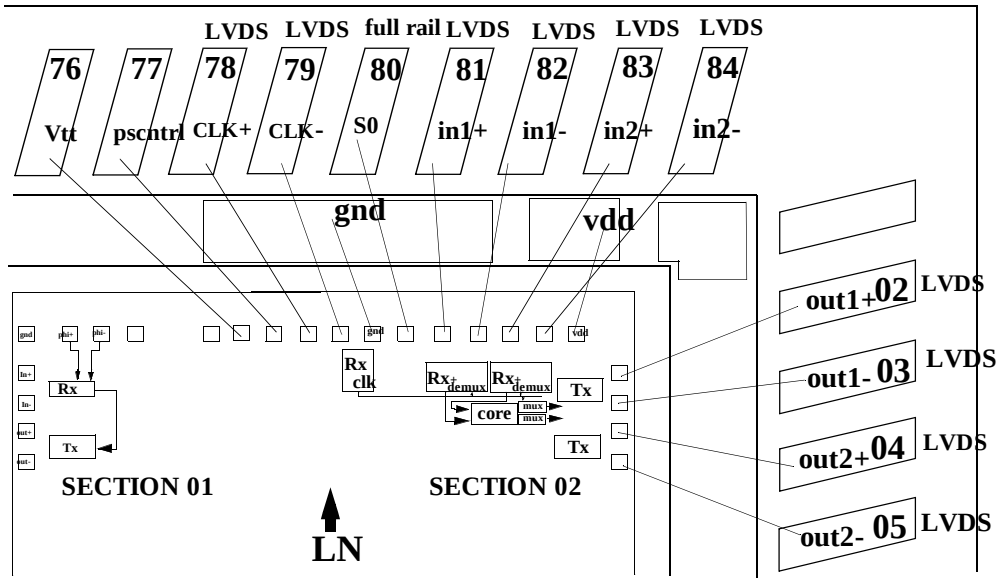


FIGURE 22: Bond wire connections for testing the switch core, receivers and transmitters

8.1.3 Test setup

For testing te01_ln, a bit-error tester is used to generates a 2^{31}-1 pseudorandom bit pattern, with a 1.25 Gb/s data rate, and a 1.25 GHz clock. The clock is divided by 2 using the Motorola MC100EL32. The output of MC100EL32 is received by TE01 on pins clk± and is buffered internally.

General settings:

MC100EL32

Table 10: MC100EL32 pinout and voltage settings

		PINOUT		
Node name		*IC*	*Triquint fixture*	*Voltage*
Reset		1	6	-1.7
CLK	Input	2	7	
$\overline{\text{CLK}}$	Input	3	8	
VBB		4	1	
VEE		5	2	-4.6
$\overline{\text{Q}}$	Output	6	3	-1.3*
Q	Output	7	4	-1.3*
VCC		8	5	0.0

The output power of MC100EL32 is divided by 2 using a power splitter, and the output of the power splitter is 380 mVpp.

Table 11: TE01 PINOUT and voltage settings

Node name	*PINOUT Triquint fixture*	*flight time (ps)*	*Voltage*
Vtt	76	-	1.2
pscntrl	77	-	2.5 - 2.6
CLK+	78	740	1.8*
CLK-	79	560	1.8*
S0	80	860	0.0V/3.6
in1+	81	0	
in1-	82	310	
in2+	83	830	
in2-	84	130	
out1+	02	840	1.74*
out1-	03	320	1.74*
out2+	04	0	1.74*
out2-	04	880	1.74*

Basic test:

Output connections:

Bit error tester:

Data outputs $\rightarrow$ in1±

Clock outputs $\rightarrow$ CLK/$\overline{\text{CLK}}$ of MC100EL32 and the oscilloscope

MC100EL32:

Q± $\rightarrow$ CLK± (through power splitter)

Switching test:

Output connections

Bit error tester:

Data output (DATA OUT) $\rightarrow$ S0

Clock outputs $\rightarrow$ CLK/$\overline{\text{CLK}}$ of MC100EL32 and the oscilloscope

MC100EL32:

Q± $\rightarrow$ CLK± (though power splitter)

$\rightarrow$ in1± through a delay element (flight time = 380 ps)

8.1.4 Results

All results shown in this report were measured under the following conditions:

V_{DD} = 3.60 V

V_{TTout} = 1.20 V

V_{TTin} = 1.20 V

1.0 1.25 Gb/s eye measurement

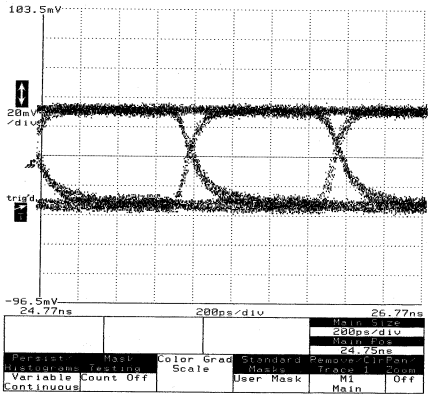


FIGURE TE01.1: Output eye-diagram at 1.25 Gb/s (800 ps bit period, 208 mVpp input, 200 ps/div, 200 mV/div)

DATA RATE: 1.25 Gb/s

V_{DD} = 3.60 V

V_{TTin} = 1.20 V

V_{TTout} = 1.74 V

V_{in} = 208 mV_{pp}

Data window = 800 ps

PRBS: 2^{31}-1 NRZ

BER < 10^{-12} at center of the eye

EYE-WIDTH = 640 ps

EYE-HEIGHT = 438 mV

measured at BER = 10^{-3}

Wave forms are attenuated by 20 dB prior to display on the scope.

2.0 1.25 Gb/s output jitter measurements

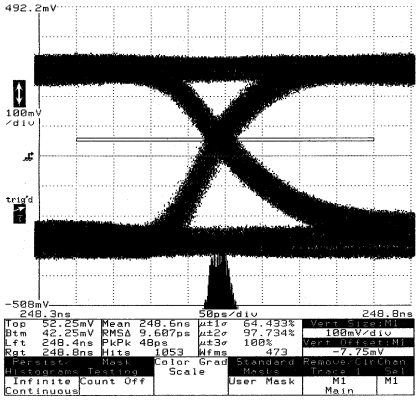


FIGURE TE01.2: Output jitter at 1.25 Gb/s (800 ps bit period, 208 mVpp input, 50 ps/div, 100 mV/div)

DATA RATE: 1.25 Gb/s

V_{DD} = 3.60 V

V_{TTin} = 1.20 V

V_{TTout} = 1.74 V

V_{in} = 208 mV_{pp}

Peak to Peak Jitter= 48 ps

RMS jitter = 9.6 ps

PRBS: 2^{31}-1 NRZ

3.0 1.25 Gb/s output crosstalk measurements

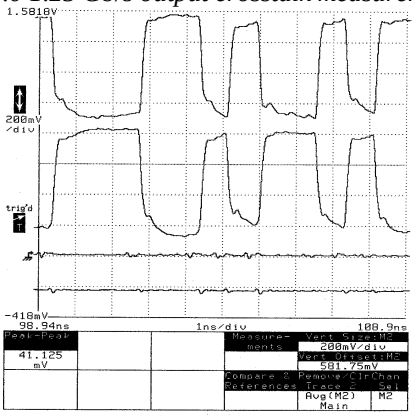


FIGURE TE01.3: Measured output at 1.25 Gb/s. Cross-talk is less than -24 dB

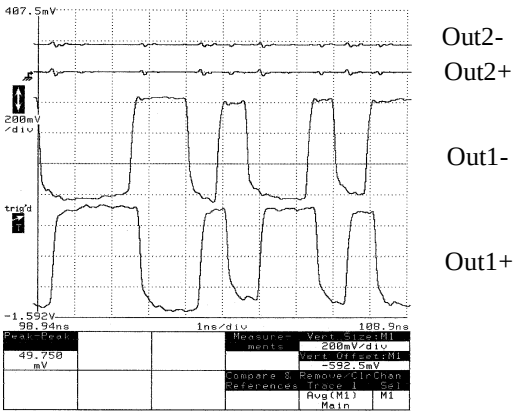


FIGURE TE01.4: Measured output at 1.25 Gb/s. Cross-talk is less than -23 dB

DATA RATE: 1.25 Gb/s

V_{in} = 208 mV_{pp}
Vout = 670 mVpp on Out2-
= 670 mVpp on Out2+
Noise = 38 mVpp on out1+
41 mVpp on out1-
Signal to noise ratio
= 17.6 for out1+
= 16.0 for out1-

DATA RATE: 1.25 Gb/s

V_{in} = 208 mV_{pp}
Vout = 692 mVpp on Out1+
= 692 mVpp on Out1-
Noise = 38 mVpp on out2-
50 mVpp on out2+
Signal to noise ratio
= 18.0 for out2-
= 14.0 for out2+

4.0 *1.25 Gb/s phase margin* measurement

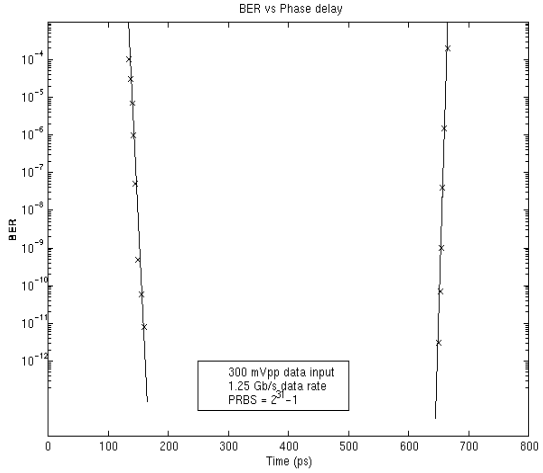


FIGURE TE01.5: Measured phase margin of TE01 at 1.25 Gb/s (800 ps/bit, with 300 mVpp data input)

500 ps phase margin at BER = 10^{-11} NRZ 2^{31}-1 PRBS

5.0 1.25 Gb/s input *sensitivity measurement*

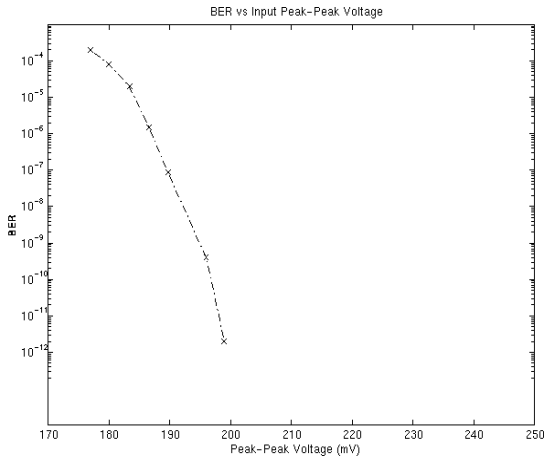


FIGURE TE01.6: Input sensitivity of TE01at 1.25 Gb/s (800 ps/bit)

200 mVpp at BER = 10^{-12} NRZ 2^{31}-1 PRBS

6. Output switching

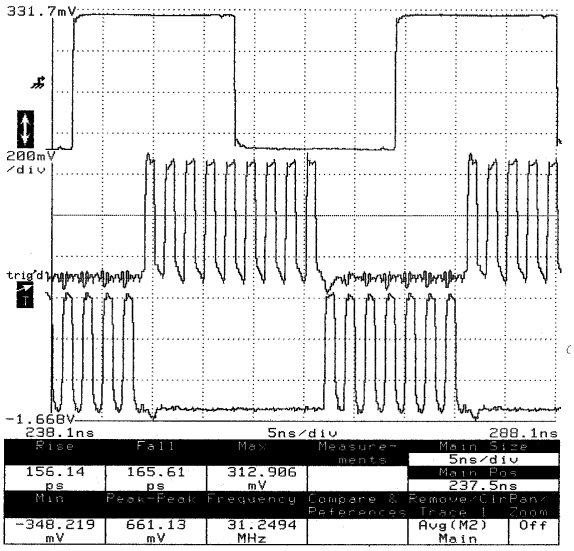


FIGURE TE01.7: Output switching at 1.25 Gb/s input data rate and 40 MHz control input.

DATA RATE:
1.25 Gb/s input
80 Mb/s Control

Input to output delay for when control input is changed is 12 ns.
Of the 12ns, 10ns is due to control voltage buffering, and 2ns is due to switch setup.
Therefor there is a 2-3ns data invalid window.

The inputs are received through in1+/in1-.
The control input is switched at 40MHz (80 Mb/s).
The outputs are out1+, out2-.

7. Power consumption:

The majority of the power consumption is expected to be at the transmitters and the clock receiver. The sense-amp receivers, latches and complementary drivers were expected to consume less than 10% of the total power consumption of the IC.

Total power	: 200 mW
Transmitters	: 96 mW total (48 mW per transmitter)
Clock receiver	: 42 mW
Receivers (demux), mux and buffers	: 60 mW

8.1.5 Discussion

The measurements revealed that the test chip frequency response and BER measurement are sensitive to the sense-amp clock common mode voltage. We noted a drop in frequency response and an increase in BER whenever the common mode voltage is raised or lowered from the target value of 2.0 V. Also, the noise in the power supply can raise and lower this common mode voltage.

If the common mode voltage of the clock increases from the target value (2.1 V compared to the expected 2.0 V), then when the clock is at its lower value (1.1 V compared to the expected 1.0 V), the sense-amp pull-down transistors will be sinking more current than expected. This causes the charging node of the differential pair to charge slower than expected. Similarly, if the common mode voltage of the clock decreases from the target value (1.9 V compared to the expected 2.0 V), then when the clock is at its higher value (2.9 V compared to the expected 3.0 V), then the pull-down transistors will be sinking less current than expected, causing a residue voltage at the differential pair nodes.

Later layout extractions showed that there is more capacitance on sense-amp clock lines than expected (15%). Due to this increase in capacitance there is a 250 mV reduction in the clock swing. This reduction in the clock would have increased the sense-amp circuits sensitivity to clock, and amplitude of the input data. This in turn would have reduced the frequency response, PSRR, phase margin and the sensitivity.

This problem can be eliminate by either resizing the sense-amp transistors to be less sensitive to actual amount of pull-down transistor currents (for the desired operating frequencies), or by modifying the sense-amp circuitry to operate on a clock with a low value that is much less than the threshold voltage of the pull-down transistors.

The TE01 had a phase margin of 50% for 1.25 Gb/s input data rate (60% phase margin at 1.0 Gb/s). In the later designs the setup and hold time of the sense-amps will be reduced to achieve better phase margins.

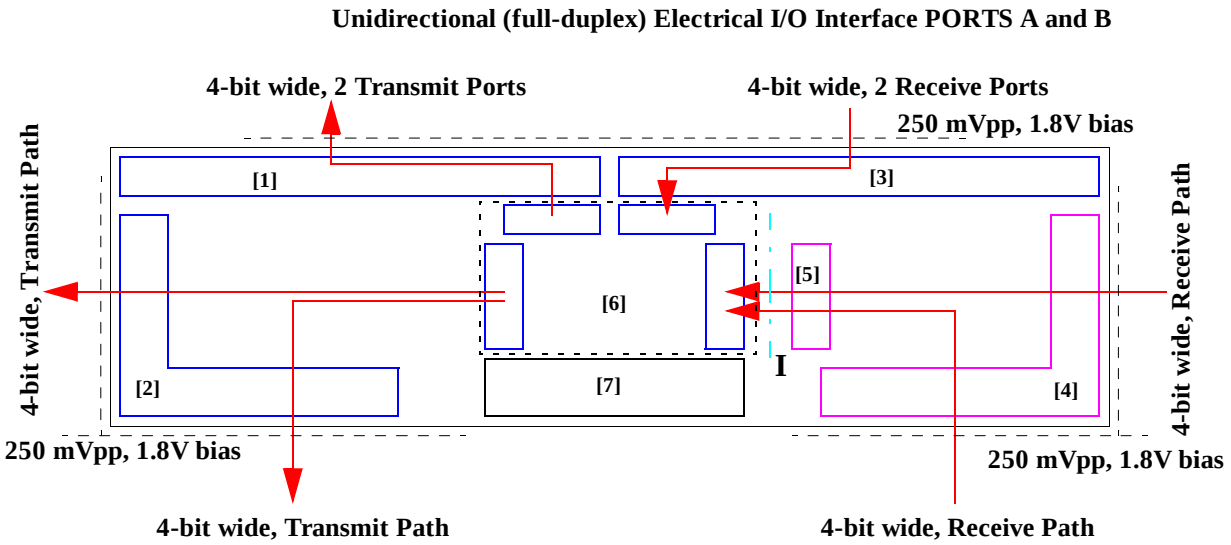
8.2 TE02 (Second test die in 0.35 μm CMOS)

8.2.1 Introduction

The TE02 Switch IC is implemented in 0.35 μm HP CMOS process and operates at 2.0 Gb/s to fully connect 4 processors with a peak throughput of 32 Gb/s. The TE02 IC contains two LVDS 4-bit electrical I/O interface ports representing the electrical I/O interface of a scaled-up Switch IC. It also contains two separate LVDS 4-bit electrical input and output paths representing the optoelectronic interface of such a scaled-up IC. TE02 also includes BUSY BIT monitoring circuitry to monitor the processor status.

The Rx portion of the IC contains a higher speed, more robust version of the sense-amp circuitry used in TE01. The IC operation includes physical loop back as well as a 4-b wide, 2-path all-electrical interconnection network (Figure 36).

Layout organization:



1. 4-bit wide, 2 Port Transmit Circuitry (Electrical Output Interface)
2. 4-bit wide, 2 Path Transmit Circuitry (Represents the Optical Output Interface)
3. 4-bit wide, 2 Port Receive Circuitry (Electrical Input Interface)
4. 4-bit wide, 2 Path Receive Circuitry (Represents Optical Input Interface)
5. Optical Clock Domain to Switch Internal Clock Domain hand off point
6. Switch core
7. BUSY BIT Circuitry

Switch Internal Clock Domain
Optical Clock Domain
Synchronization Boundary Between Switch Internal Clock Domain and Optical Clock Domain
Data flow

FIGURE 23: Layout organization of TE02 showing the major blocks (9.0 mm x 2.2mm)

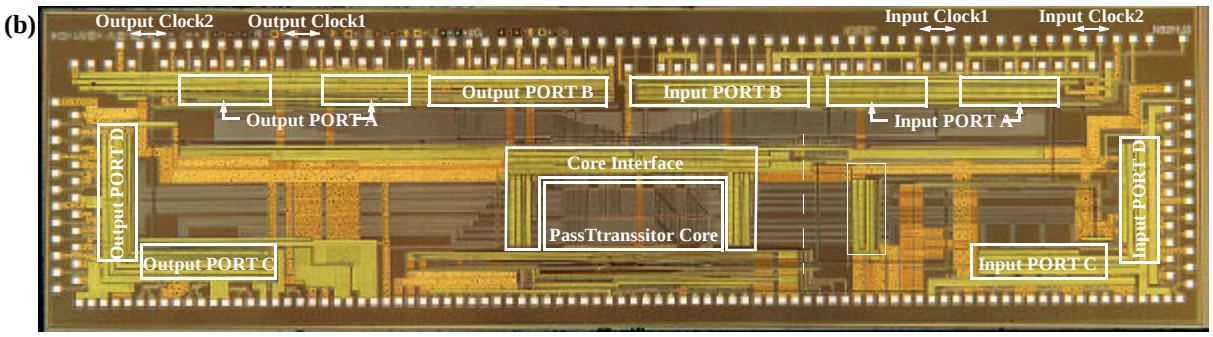


FIGURE 24: Photograph of TE02

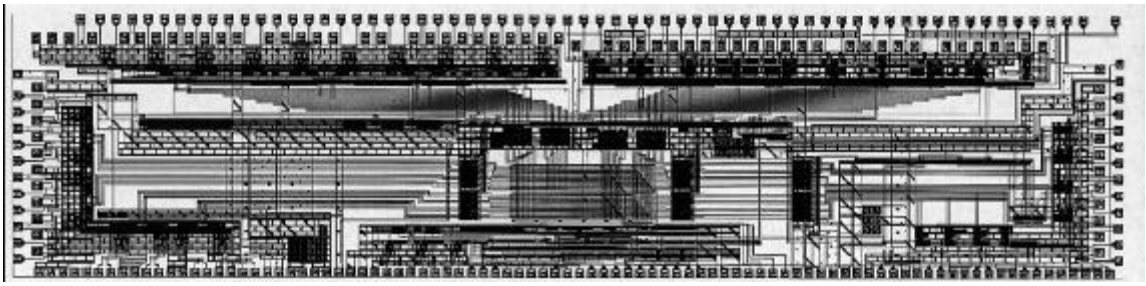


FIGURE 25: TE02 layout plot (9.0 mm x 2.2 mm)

Ship date: 02/22/99
Return date: 06/01/99
Process: 0.35 µm HP CMOS C10

8.2.2 IC Functionality

Data path:

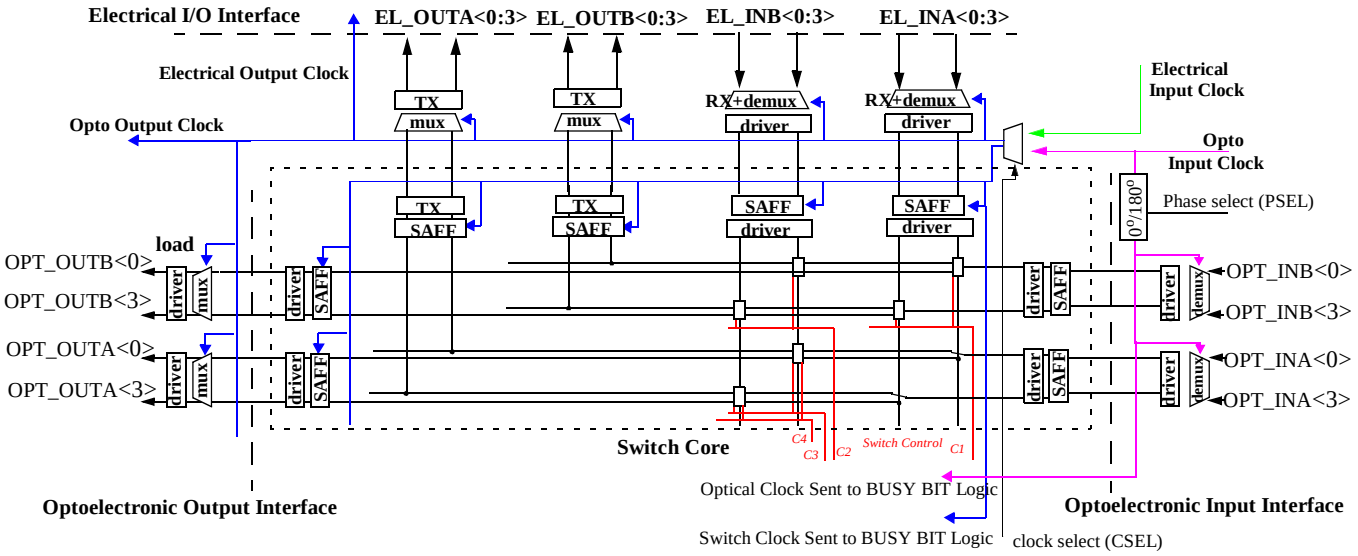


FIGURE 26: Abstract of the TE02 IC, without BUSY BIT monitoring circuitry

Figure 26 shows an abstract block diagram of the data path. PORT A and PORT B are electrical I/O interface. OPT_INA and

OPT_INB represent the optoelectronic inputs, while OPT_OUTA to OPT_OUTB represent the optoelectronic outputs. Note that the high-speed inputs to the IC are on the right and the high-speed outputs are on the left.

Input data are demuxed by 2 at the input, and are muxed again at the output. A half-speed clock is used for receiving and transmitting the data. The IC has two clock domains, and two possible clock inputs, both of which are received electrically at the right top of the IC. One clock is the electrical input clock, and the other is the mock optical clock which represents the optoelectronic input clock.

The switch core contains SAFFs and drivers at its boundary (Figure 26), and the internal switching is implemented using pass-transistor logic. Four control signals are provided to set up the Switch IC. Figure 27 shows the possible switch configurations and their corresponding control signal levels.

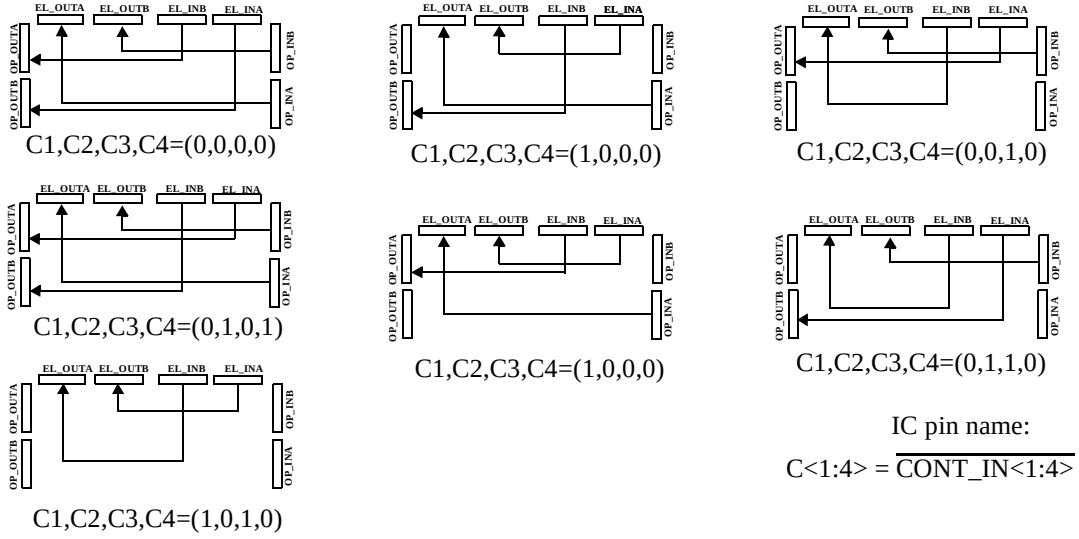


FIGURE 27: Possible switch configurations

Clocking:

In Figure 26, the two clock domains have different colors. The switch internal clock is blue, while the mock optical clock is magenta. Both clocks are received using traditional LVDS receivers, and are level-shifted and buffered internally. The signal clock select "CSEL" specifies which clock is used as the switch internal clock.

In addition, the mock optical clock is buffered and delivered to the optical input interface. A $0^o/180^o$ degree phase selection circuit is used to control the relative phase of the mock optical clock to the switch internal clock at the boundary between them. The signal phase select "PSEL" controls the $0^o/180^o$ phase selection circuit.

Two analog signals, "BIASCON_OP" and "BIASCON_SW," are provided to control the duty cycle of the mock optical clock and the switch internal clock.

In the Switch IC there are no feedback loops between stages, and data has a right-to-left flow. This data flow pattern is used to simplify the IC clock routing.

Even though the BUSY BIT logic can be designed to have its own separate clock, in TE02, the BUSY BIT logic shares the same mock optical and switch internal clocks with the data path.

BUSY BIT Circuit

In addition to the data path, the TE02 also implements a BUSY BIT scheme for network access arbitration. The BUSY BIT provides information about the status of the processors in the network. A high value indicates that a processor is busy, while a low value indicates that it is available. If individual processors in network do not share the paths and BUSY BITs, then a

network with *N* processors (optical paths), requires *N* BUSY BITs for network access arbitration.

The BUSY BITs can be implemented using a dedicated optical channel per BUSY BIT (a dedicated TOKEN per processor), or by serializing the BUSY BIT information on one optical channel. In case of a dedicated TOKEN, an active low TOKEN pulse circulating in the network indicates that a processor is free. A processor P_j can grab a free processor P_i, if the TOKEN pulse corresponding to P_i is currently at P_j's Switch IC. P_j grabs the free TOKEN by inhibiting the propagation of the pulse. In this case the pulse itself indicates the time frame at which the TOKEN can be grabbed. Thus, there is no need for an additional FRAME bit. Note that a static low voltage can not be used to indicate a processor is free, since more than one processor may grab it simultaneously.

On the other hand in the BUSY BIT scheme information is serialized into one or more optical channels, and there is a stream of 1's and 0's (stream of pulses) flowing in the network. In this stream of information, we need to know exactly where BUSY BIT *i* is located. Thus, there needs to be scheme to identify the starting and an ending point of the data stream. An additional FRAME signal is used for this purpose.

The FRAME constantly circulates around the network on its own dedicated optical channel. Another dedicated optical channel caries the processor BUSY BITs, which are encoded serially starting with the BUSY BIT for processor *n*, at the 0→1 transition of the FRAME. The total FRAME/BUSY BIT latency is 5 cycles:

- 2 cycle latency due to FRAME receive and buffering
- 1 cycle latency due to clock domain synchronization (Section: 5.0).
- 2 cycle latency due to decode/encode required for AC coupling.

Figure 28 shows an abstract block diagram of the BUSY BIT circuit. It consists of a request interface register, a request register, a grab register, and a grab interface register. The request interface register is the external interface to the request register. The request register holds which processor BUSY BITs are being requested by this node. The grab register saves which BUSY BITs have been obtained by this node, and the grab interface register is the external interface to the grab register.

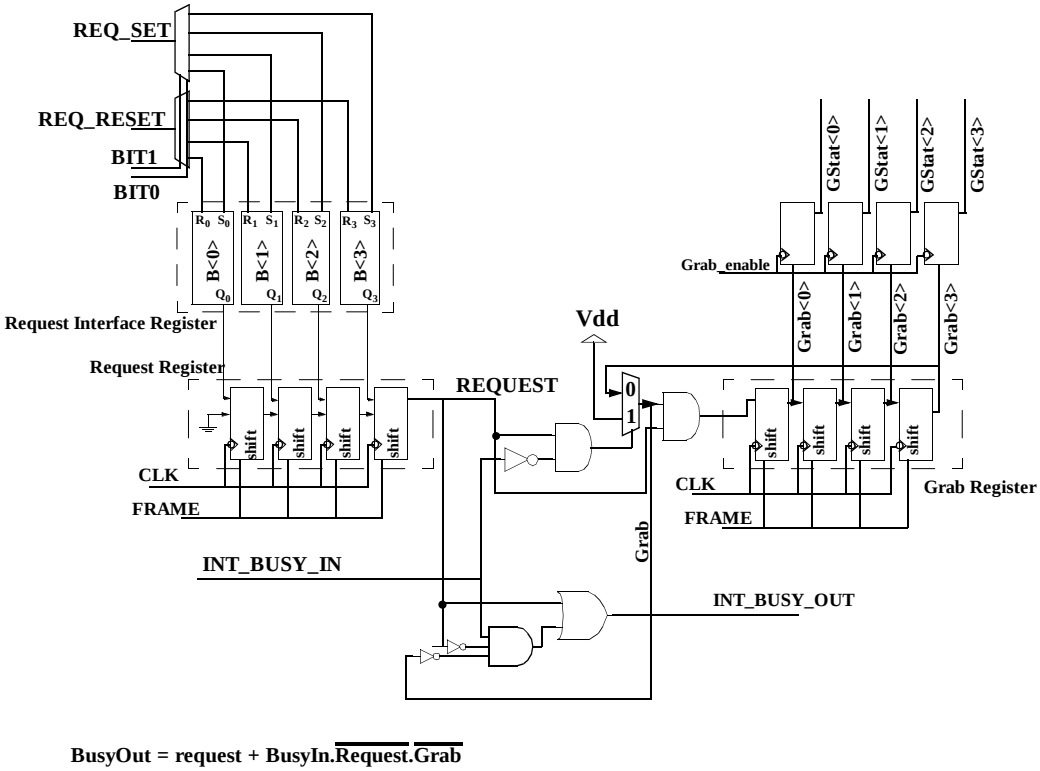


FIGURE 28: Abstract of the BUSY BIT monitoring circuitry

Request interface register:
The request interface register is the interface between the request register and the external world. It consists of RS latch type FFs, and is controlled by external inputs BIT<1:0>, REQ_SET and REQ_RESET. The bits BIT<1:0> specifies the appropriate request interface register, and REQ_SET specifies a set operation while REQ_RESET specifies a reset operation. BIT1 and BIT0 must remain stable when REQ_SET or REQ_RESET is active. The truth table for the request control interface is given in Table 12.

Request register:
The request register holds which processor BUSY BITs are being requested by this node, and consists of sense-amp FFs which load at the negative clock edge. When FRAME is low (inactive FRAME), the request interface register values are continuously loaded into the request register. During FRAME the values stored in the request register are right-shifted. During this right-shift, the left most bit of the request register loads a low value. In addition to handing BUSY BIT requests during normal operation, the request register is also used by the master node to set the length of the FRAME during FRAME creation.

Grab register:
The grab register saves which BUSY BITs have been obtained by this node. During FRAME low, the grab register retains its values. During FRAME, the grab register is right-shifted, with the input at the left-most SAFF being a function of the current output of the right-most SAFF, the input BUSY BIT value, and the REQUEST signal:
GRAB = REQUEST.($\overline{\text{BUSY\_IN}}$ + GRAB<3>)

Grab interface register:
The grab interface register is the interface between the grab register and the external world. When the input grab_enable is active (set to low), the grab register values are loaded into the grab interface register. Grab_enable should only be activated when the FRAME is not present at the node, preferably right after a FRAME has past the node. Grab interface register signal GSTAT<3:0> show which processor BUSY BITs were grabbed by this node.

Table 12: Truth table for the request interface register

REQ_SET	REQ_RESET	BIT1	BIT0	Reset				Set				Request interface register output
				R_3	R_2	R_1	R_0	S_3	S_2	S_1	S_0	
0	0	X	X	0	0	0	0	0	0	0	0	Inactive set/reset
1	0	0	0	0	0	0	1	0	0	0	0	reset #0, $Q_0 = 0$
1	0	0	1	0	0	1	0	0	0	0	0	reset #1, $Q_1 = 0$
1	0	1	0	0	1	0	0	0	0	0	0	reset #2, $Q_2 = 0$
1	0	1	1	1	0	0	0	0	0	0	0	reset #3, $Q_3 = 0$
0	1	0	0	0	0	0	0	0	0	0	1	set #0, $Q_0 = 1$
0	1	0	1	0	0	0	0	0	0	1	0	set #1, $Q_1 = 1$
0	1	1	0	0	0	0	0	0	1	0	0	set #2, $Q_2 = 1$
0	1	1	1	0	0	0	0	1	0	0	0	set #3, $Q_3 = 1$
1	1	X	X	-	-	-	-	-	-	-	-	Not allowed

Figure 29 shows an abstract timing diagram of a processor REQUEST and GRAB operation. First the requesting processor sets the request interface register bit corresponding to the desired processor BUSY BIT (when FRAME is not present). If the corresponding BUSY BIT is free, when FRAME arrives the node sets the BUSY BIT and record the fact that it grabbed the bit in the grab register. To read the Grab Register, the processor activates the grab_enable signal, which loads the grab register values into the grab interface register (when FRAME is not present).

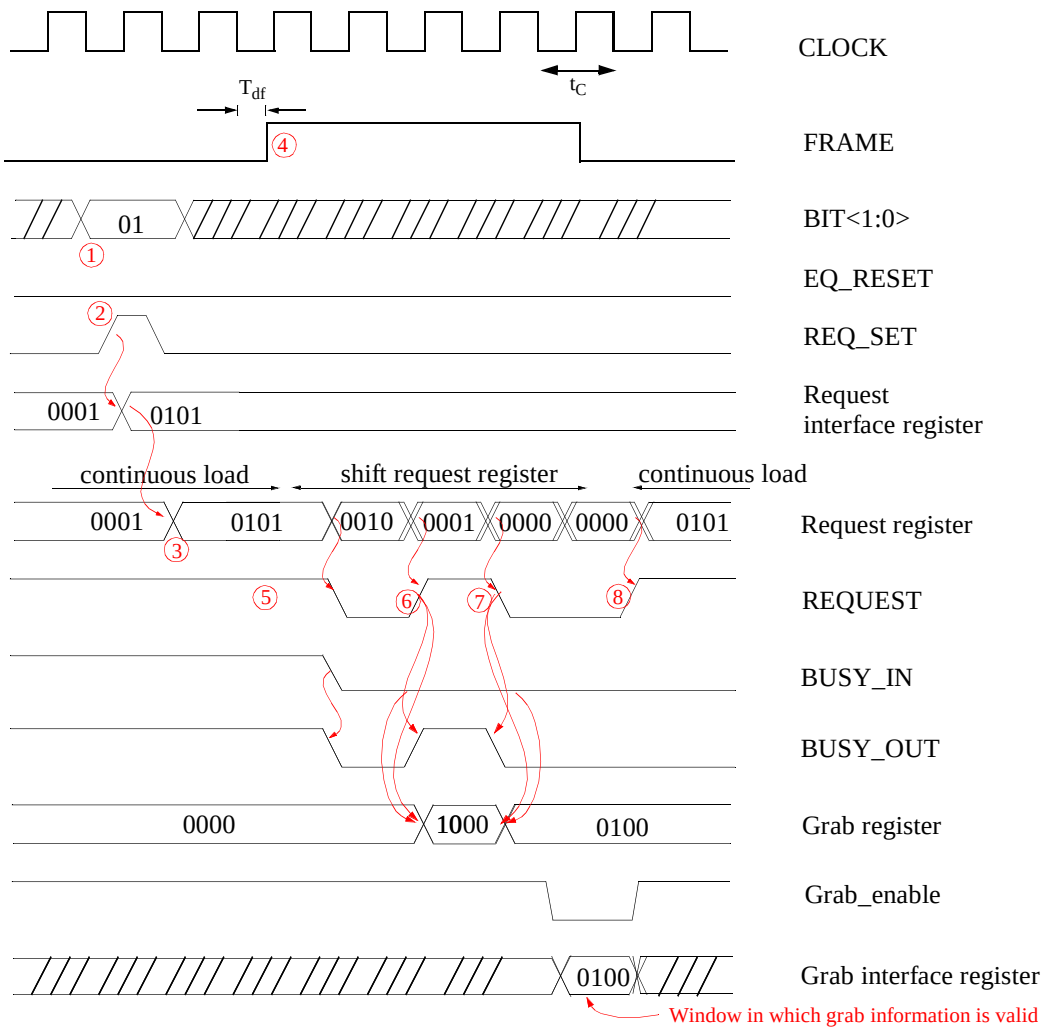


FIGURE 29: Abstract timing diagram showing the request and grabbing of BUSY BIT for P_1

The following steps trace the operation of the external control and internal signals for the case of requesting the BUSY BITs for processors P_1 and P_3 and successfully grabbing the BUSY BIT for P_1.

Step #1
Select the request interface register bit corresponding to the BUSY BIT being requested or released, in this case the BUSY BIT for P_1. BIT<1:.0> = 01decodes to BUSY BIT #1. Note that in this example both the request interface and request registers already have the bit for processor P_3 set.

Step #2
Activate the SET signal. This sets the #1 bit of the request interface register (bits are numbered from #0-#3).

Step #3.
Since FRAME is not present, at the negative edge of the clock, the new value of request interface register is loaded to the Request Register.

Step #4
FRAME arrives at the node and starts the shifting of both the request and grab registers.

Step #5
Though REQUEST (for P_3) is high, since BUSY_IN is high indicating that P_3 is already busy, the node is unable to grab P_3. Therefore a zero is be shifted in to the grab register indicating that P_3 was not grabbed, and BUSY_OUT is high.

Step #6
Since REQUEST (for P_1) is high and BUSY_IN is low indicating P_1 is free, the node is able to grab the BUSY BIT for P_1. Thus, a '1' (high) is shifted into the grab register, and BUSY_OUT is high.

Step #7
Since REQUEST (for P_0) is low, indicating that the node is not requesting P_0, and since BUSY_IN is low, indicating P_0 is free, BUSY_OUT is low.

Step #8
FRAME has left the node, ending the shifting of the request and grab registers. Therefore the request register returns to continuously loading of the request interface register.

FRAME_OUI = FRAME
FRAME = $\overline{MS}$.FRAME_IN
+ MS.$\overline{FC\_IN}$.(HIF.REQUEST + $\overline{HIF}$.FRAME_IN)
HIF = MS.(FC_IN + HIF.REQUEST)

FIGURE 30: Abstract of the FRAME circuit

Figure 30 shows both the FRAME and HIF (Hide Incoming Frame) logic. The FRAME circuit is used to create and to receive the FRAME. TE02 uses a static chip input MS+ to identify which processor node is the master and can therefore create a new FRAME. An additional chip input (FC_IN) tells the master's Switch IC when to create a new FRAME. The FRAME logic has three possible states:

1. FRAME initialization (Can only be done by the master, and sets the outgoing FRAME to low, regardless of the incoming FRAME). When both MS and FC_IN are high, the internal signal HIF is activated (is high).
2. FRAME creation (Can only be done by the master, and creates the FRAME, only after the FRAME has been properly initialized before entering this stage). During FRAME creation, the signal HIF is active and FRAME_OUT = FRAME = REQUEST. REQUEST is the output of the request register, which right-shifts during FRAME creation. The length of the FRAME is determined by the number of 1's in the request register. If there are *m* number of 1's in the request register (right to left), then the length of the FRAME is *m+1*. When REQUEST finally goes low, the FRAME and HIF goes low completing the FRAME creation.
3. FRAME pass-through (All processor nodes except the master are always in this state, and after FRAME creation, the master node comes to this stage). During pass-through the internal signal HIF is inactive (low).

Note that when FRAME is first created, the master grabs all the BUSY BITs except its own. The master then subsequently release the BUSY BITs. This scheme allows the master to communicate with each of the nodes of the network during initialization.

State diagram in Figure 31 illustrates the three possible states of the FRAME logic.

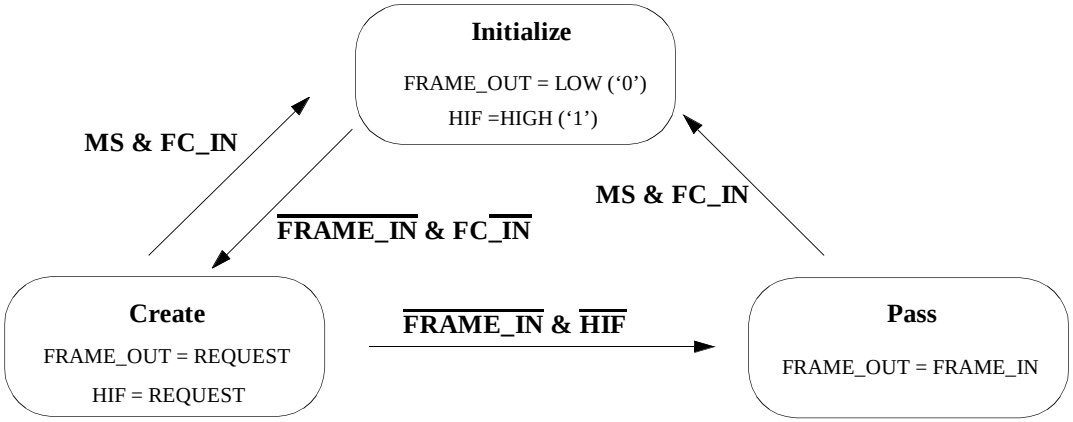


FIGURE 31: State diagram of the FRAME logic

Figure 32 shows an abstract timing diagram of the FRAME creation process. When MS and FC_IN are both active, FRAME_OUT goes low. FC_IN should only be released after enough time has elapsed for the FRAME to circulate around the network and arrive at the master node.

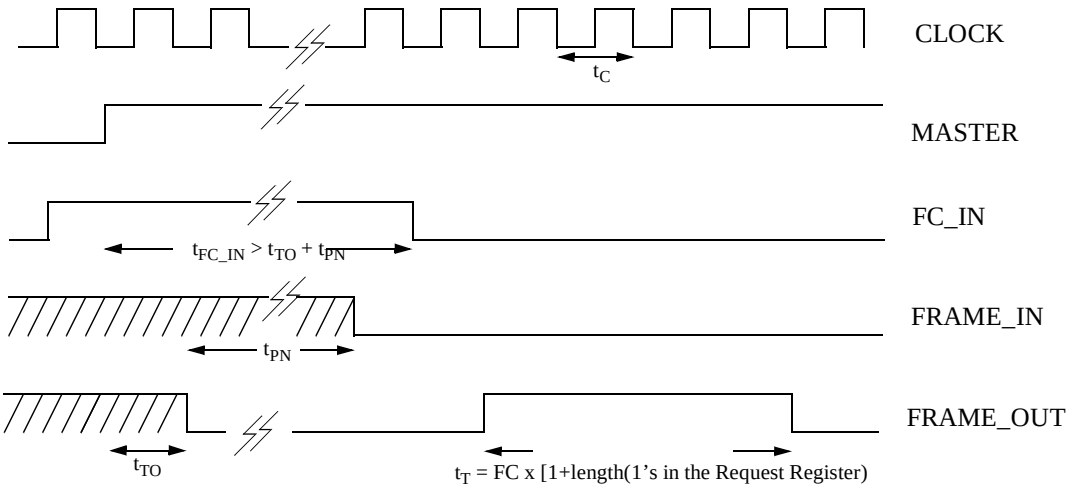


FIGURE 32: Abstract timing diagram of the FRAME creation operation

Additional characteristics of the BUSY BIT operation are as follows:

1. The request interface register, the request register, and the grab register are *n* bits wide, where *n* is the maximum number of processors that can be monitored by the IC. (4 in the case of TE02).
2. If the input BUSY BIT corresponding to a particular processor that we are requesting is reset (processor available), then the outgoing BUSY BIT and the corresponding GRAB bit is set.
3. If the incoming BUSY BIT corresponding to a particular processor that we are requesting is already set (processor busy), the outgoing BUSY BIT remains high, but the corresponding GRAB bit won't be set.
4. On the other hand, in case #5, if we had previously grabbed the BUSY BIT (so the grab register bit corresponding to it is set), and if we are no longer requesting that bit, then the outgoing BUSY BIT and the corresponding grab register bit is reset.

This BUSY BIT scheme can also support a two-way communication protocol (such as request/reply), in which whenever processor P_i connects to P_j, P_j connects to P_i.

To create such a connection pattern, the protocol can allow P_i to grab both P_i's and P_j's BUSY BITs. In this case there are cyclic

dependencies, and as a result network is not deadlock free. To recover from deadlock, one can implement a time out strategy. The impact of a time out scheme on the latency and throughput of the network is proportional to the networks deadlock probability.
The above explained deadlock condition can be avoided by requiring P_j (reply node) grab the BUSY BIT for P_i (Request nodes) to complete the communication. In this case the reply latency dependents on the number of processors that are trying to communicate with P_i. Here the worse case latency is equal to the total number of processors times the transmission time of a data packet.
To over come this problem, the proposed BUSY BIT scheme can be extended to limit the worse case reply latency to the transmission time of a data packet.

We have adopted a simple encoding to support AC coupling of the BUSY BIT and FRAME signals. An internal signal of "1" (HIGH) is encoded as "10", and an internal signal "0" is encoded as "01". As seen in Figure 33, this encoding strategy switches at the data rate of the IC, except during a transition.

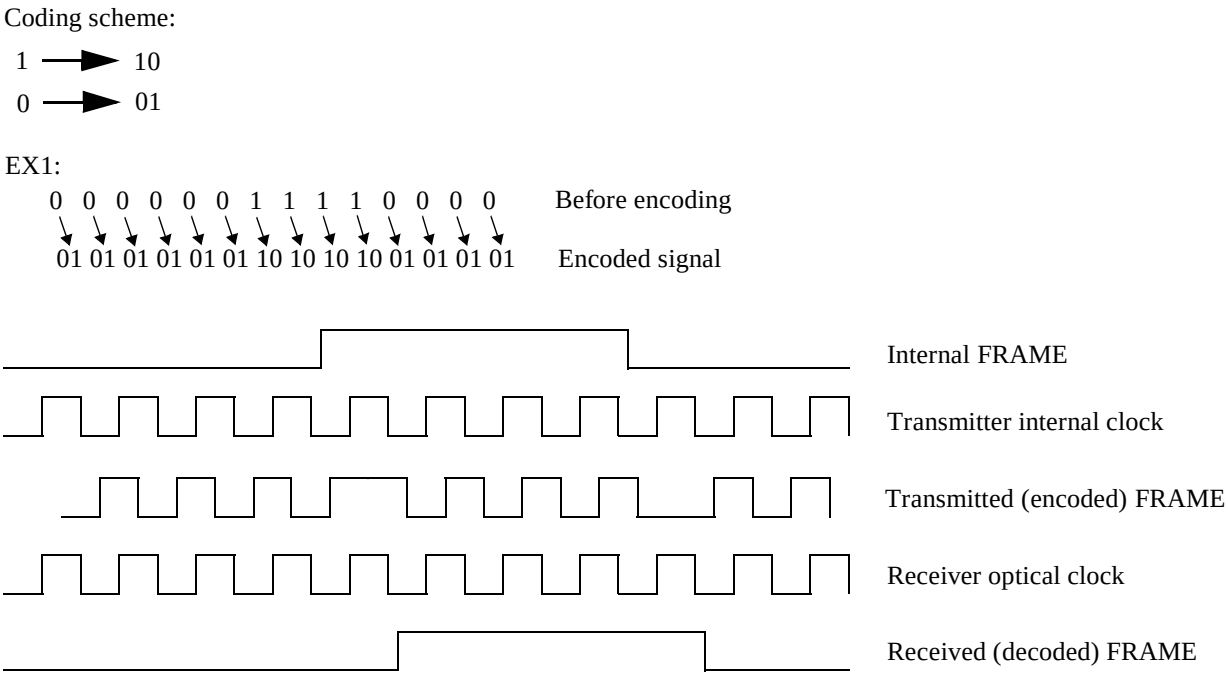


FIGURE 33: Encoding strategy employed for BUSY BIT and FRAME

8.2.3 Power consumption of TE02

The TE02 Switch IC is partitioned into 6 major blocks (Electrical receive, optical receive, switch core, electrical transmit, optical transmit and BUSY BIT circuitry). Simulation results shows that the SA-Demux circuit consumes 11mW, and the latch and driver circuits together consume 13 mW of power. The total power consumption of the IC is estimated to be 2.9 Watts. Further optimizations of the SA_Demux and the low voltage driver circuits are being investigated.

Electrical receive	200 mW
Optical receive + hand off	400 mW
Switch core	800 mW
Electrical transmit	600mW
Optical transmit	600 mW
BUSY BIT circuitry	300 mW
Total power consumption	2.90 W

NOTE: The above numbers are rounded off to the nearest 100 mW. The simulation results are for the 0.35 μm HP CMOS C10 process using the vendor slow corner library, and a 3.3V power rail.

8.2.4 Package and pad description

Ceramic package and cavity

Figure 34 shows the chip package and the locations of the PIN names designated in Table 13. $PIN_T<n>$ designates pins across the top of the package, $PIN_L<n>$ designates pins across the left side of the package, $PIN_B<n>$ designates pins across the bottom of the package, and $PIN_R<n>$ designates pins across the right side of the package.

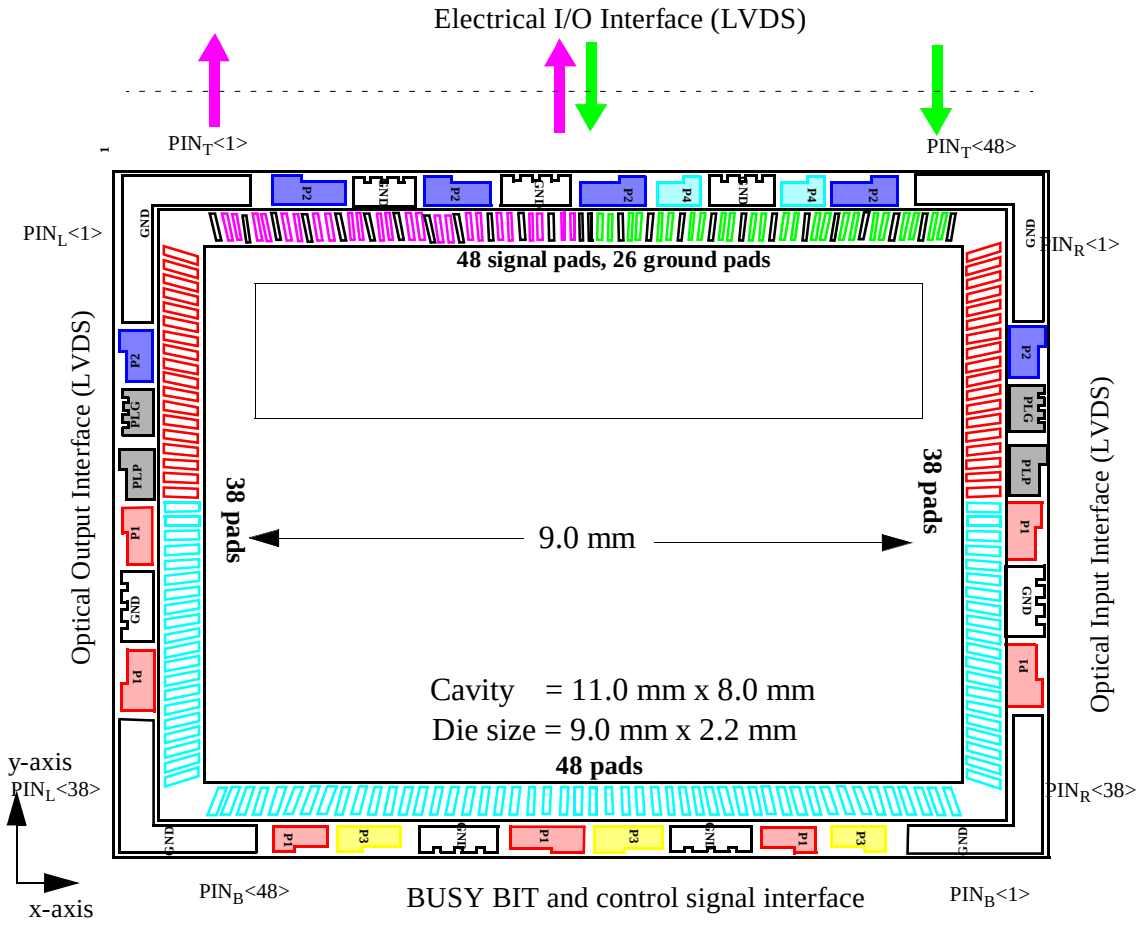


FIGURE 34: Ceramic package and cavity

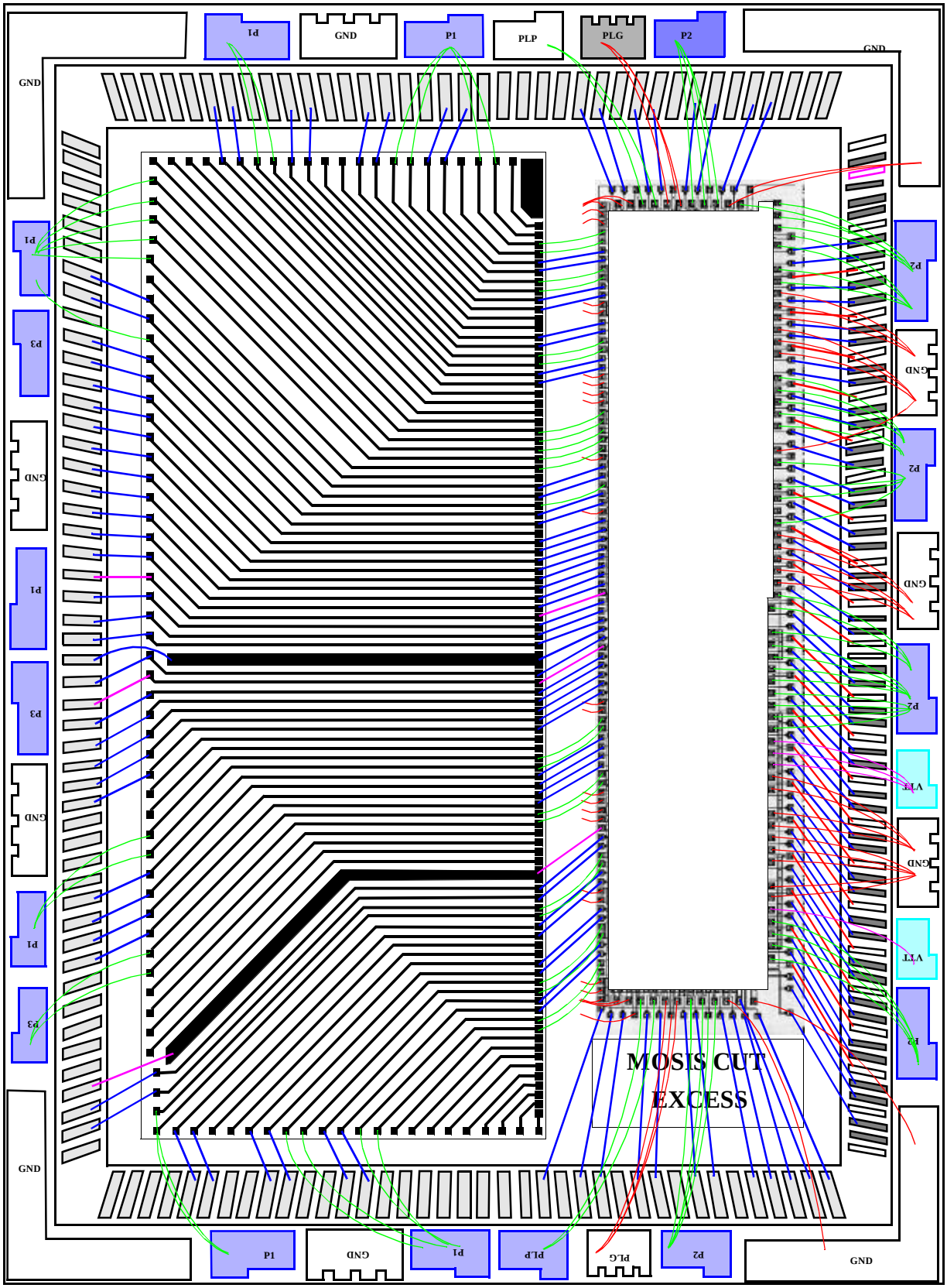


FIGURE 35: Bonding diagram of TE02

Pad description

Table 13: Pin-out for TE02

PIN	Signal name	Type	Description
PIN_T<1>	CONT_SWCLK	Input	(CMOS) Switch primary clock selection
PIN_T<2>	BIASCON_OP	Input	(Analog) Duty cycle control for the optical receive clock
PIN_T<3>	CONT_OPCLK	Input	(CMOS) 180 Phase shift control
PIN_T<4>	PSCNTRL	Input	(Analog) Slew rate control of the clock receivers
PIN_T<5:6>	OPCLK_IN ±	Input	(LVDS) Optical input clock
PIN_T<7:8>	ELEC_IA1 ±	Input	(LVDS)Input electrical PORT A, Bit 1
PIN_T<9:10>	ELEC_IA2 ±	Input	(LVDS) Input electrical PORT A, Bit 2
PIN_T<11:12>	ELCLK_IN ±	Input	(LVDS) Electrical input clock
PIN_T<13:14>	ELEC_IA3 ±	Input	(LVDS) Input electrical PORT A, Bit 3
PIN_T<15:16>	ELEC_IA3 ±	Input	(LVDS) Input electrical PORT A, Bit 4
PIN_T<17:18>	ELEC_IB1 ±	Input	(LVDS) Input electrical PORT B, Bit 1
PIN_T<19:20>	ELEC_IB2 ±	Input	(LVDS) Input electrical PORT B, Bit 2
PIN_T<21:22>	ELEC_IB3 ±	Input	(LVDS) Input electrical PORT B, Bit 3
PIN_T<23:24>	ELEC_IB4 ±	Input	(LVDS) Input electrical PORT B, Bit 4
PIN_T<25:26>	ELEC_OB1 ±	Output	(LVDS) Output electrical PORT B, Bit 4
PIN_T<27:28>	ELEC_OB2 ±	Output	(LVDS) Output electrical PORT B, Bit 3
PIN_T<29:30>	ELEC_OB3 ±	Output	(LVDS) Output electrical PORT B, Bit 2
PIN_T<31:32>	ELEC_OB4 ±	Output	(LVDS) Output electrical PORT B, Bit 1
PIN_T<33:34>	ELEC_OA1 ±	Output	(LVDS) Output electrical PORT A, Bit 4
PIN_T<35:36>	ELEC_OA2 ±	Output	(LVDS) Output electrical PORT A, Bit 3
PIN_T<37:38>	ELCLK_OUT ±	Output	(LVDS) Electrical output clock
PIN_T<39:40>	ELEC_OA3 ±	Output	(LVDS) Output electrical PORT A, Bit 2
PIN_T<41:42>	ELEC_OA4 ±	Output	(LVDS) Output electrical PORT A, Bit 1
PIN_T<43:44>	OPCLK_OUT ±	Output	(LVDS) Optical output clock
PIN_R<1>	BIASCON_SW	Input	(Analog) Switch primary clock duty cycle control
PIN_R<2>	VTTIN	Bias	1.6 V Bias voltage for LVDS inputs
PIN_R<3:4>	OP_IB1 ±	Input	(LVDS) Optical PATH B, Bi4
PIN_R<5>	VDD	Power	VDD
PIN_R<6:7>	OP_IB1 ±	Input	(LVDS) Optical PATH B, Bit3
PIN_R<8>	GND	Ground	GND
PIN_R<9:10>	OP_IB1 ±	Input	(LVDS) Optical PATH B, Bit2
PIN_R<11>	GND	Ground	GND
PIN_R<12:13>	OP_IB1 ±	Input	(LVDS) Optical PATH B, Bit1
PIN_R<14>	MS	Input	(CMOS) Master identifier signal
PIN_R<15:17>			NC
PIN_R<18:20>	GND	Ground	GND
PIN_R<21:22>	VDD	Power	VDD

Table 13: Pin-out for TE02

PIN	Signal name	Type	Description
PIN_R<23:24>	OP_IA4 ±	Input	(LVDS) Optical PATH A, Bit4
PIN_R<25:26>	VDD	Power	VDD
PIN_R<27:28>	OP_IA3 ±	Input	(LVDS) Optical PATH A, Bit3
PIN_R<29:30>	GND	Ground	GND
PIN_R<31:32>	OP_IA2 ±	Input	(LVDS) Optical PATH A, Bit2
PIN_R<33:34>	VDD	Power	VDD
PIN_R<35:36>	OP_IA1 ±	Input	(LVDS) Optical PATH A, Bit1
PIN_R<37>	VTTIN	Bias	1.6 V Bias voltage for LVDS inputs
PIN_R<38>	VTTIN	Bias	1.6 V Bias voltage for LVDS inputs
PIN_B<1:4>			NC
PIN_B<5:8>	CONT_IN4:1	Input	(CMOS) Switch core control
PIN_B<9:10>			NC
PIN_B<11:12>			NC
PIN_B<13>	PHASE_SEL	Input	(CMOS) 180 phase shift for BUSY clock
PIN_B<14:15>	BUSY_IN ±	Input	(LVDS) Input BUSY BIT
PIN_B<16:17>	FRAME_IN ±	Input	(LVDS) Input FRAME
PIN_B<18>	VTT	Bias	1.6 V LVDS Bias voltage used in destination termination
PIN_B<19>	FC_IN	Input	(CMOS) FRAME create
PIN_B<20:21>	BIT<1:0>	Input	(CMOS) Request bits
PIN_B<22>	REQ_SET	Input	(CMOS) Set the request bit specified by BIT<1:0>
PIN_B<23>	REQ_RESET	Input	(CMOS) Reset the request bit specified by BIT<1:0>
PIN_B<24>	VTT_OUT	Bias	1.6 V LVDS Bias voltage for source termination
PIN_B<26:29>	GRAB<1:4>	Output	(CMOS - 2.6V) Grab status bit <0:4>
PIN_B<30>	Grab_enable	Input	(CMOS) Grab bit interface register clock
PIN_B<31>	RESET	Input	(CMOS) Reset grab register
PIN_B<32>	SET	Input	(CMOS) Set grab register
PIN_B<33>	SOUT	Output	(LVDS, source terminated) FRAME output to the board
PIN_B<34:35>			NC
PIN_B<36:38>			NC
PIN_B<39:41>			NC
PIN_B<42:44>			NC
PIN_B<45:48>			NC
PIN_L<1>			NC
PIN_L<2:3>	OP_OB4 ±	Output	(LVDS) Optical PATH B4 output
PIN_L<4>			NC
PIN_L<5:6>	OP_OB3 ±	Output	(LVDS) Optical PATH B3 output
PIN_L<7>			NC
PIN_L<8:9>	OP_OB2 ±	Output	(LVDS) Optical PATH B2 output
PIN_L<10>			NC

Table 13: Pin-out for TE02

PIN	Signal name	Type	Description
PIN_L<11:12>	OP_OB1 ±	Output	(LVDS) Optical PATH B1 output
PIN_L<13:17>			NC
PIN_L<20:21>			NC
PIN_L<22:23>	OP_OA4 ±	Output	(LVDS) Optical PATH A4 output
PIN_L<24:25>			NC
PIN_L<26:27>	OP_OA3 ±	Output	(LVDS) Optical PATH A3 output
PIN_L<28:29>			NC
PIN_B<30:31>	OP_OA2 ±	Output	(LVDS) Optical PATH A2 output
PIN_B<32:33>			NC
PIN_B<34:35>	OP_OA1 ±	Output	(LVDS) Optical PATH A1 output
PIN_B<36:38>			NC

NOTE: NC - Not connected

8.2.5 Testing of the Switch IC

Functional testing of the data path

The first test of the IC determined the sensitivity and phase margin of each individual channel of both the ports and paths at BER = 10^{-12} for 2^{31}-1 PRB NRZ, 2.0 Gb/s.

Summary of figures:

- Figure TE02.1 and Figure TE02.2 show eye measurements at 2.0 Gb/s.
- Figure TE02.3 shows the output jitter and Figure TE02.4 shows the worst case output crosstalk between data lines.
- Figure TE02.5 shows the eye measurements at 1.0 Gb/s and Figure TE02.6 shows the jitter measurements at 1.0 Gb/s.
- Figure TE02.7 and Figure TE02.8 respectively show the best and worst case phase margin at 2.0 Gb/s and 1.0 Gb/s
- Figure TE02.9 shows the best and worst case input sensitivity at 3.3 Gb/s.
- Figure TE02.10, Figure TE02.11 and Figure TE02.12 are examples of the switch control.
- Figure TE02.13 and Figure TE02.14 respectively show the clock to data crosstalk and source-referred jitter of the electrical clock referenced to the input electrical clock.

Table 14 shows the best and worst case measured values for direct input and output signals (those signals not using the insert).

Table 14: The best and worst case measurement results for direct I/O

Measurement	Best		Worst	
	Value	Signal line	Value	Signal line
Input sensitivity	100 mVpp		200 mVpp	
Output phase margin				
Input phase margin				
Adjacent line crosstalk	-22 dB		-19.9 dB	
Clock to data crosstalk*	-28 dB		-20 dB	

* Only electrical output B was tested for clock to data crosstalk

1.0 2.0 Gb/s eye measurements

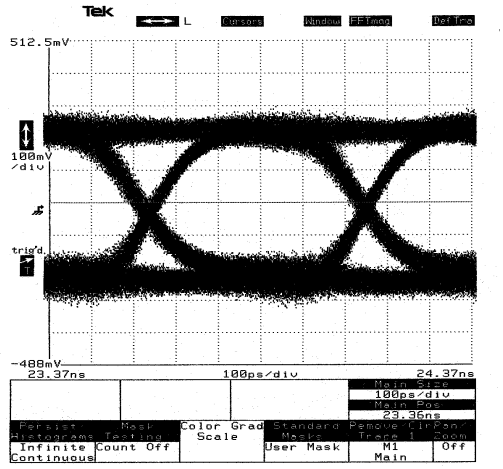


FIGURE TE02.1: Output eye diagram at 2.0 Gb/s (500 ps bit period, 250 mVpp input, 100 ps/div, 100 mV/ div)

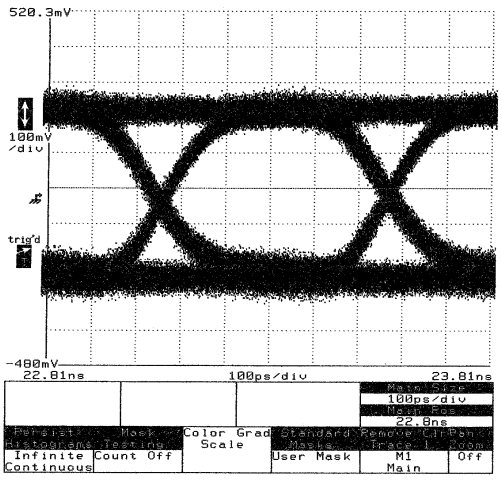
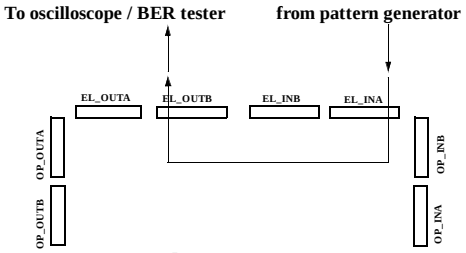


FIGURE TE02.2: Output eye diagram at 2.0 Gb/s, for the case of multiple-pass through the switch. (500 ps bit period, 250 mVpp input, 100 ps/div, 100 mV/ div)

DATA RATE: 2.0 Gb/s

V_{DD} = 3.30 V
V_{TTin} = 1.80 V
V_{TTout} = 1.20 V
V_{in} = 250 mV_{pp}



Eye diagram setup

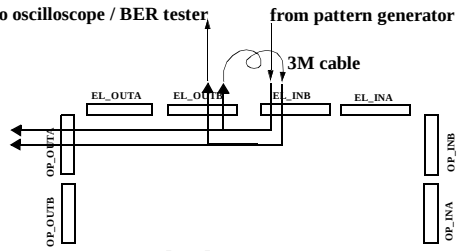
Input node = electrical A<0>
Output node = electrical B<0>

Data window = 500 ps
PRBS: 2^{31}-1 NRZ
BER < $3x10^{-13}$

EYE-WIDTH = 365 ps
EYE-HEIGHT = 176 mV
measured at BER = 10^{-3}

DATA RATE: 2.0 Gb/s

V_{DD} = 3.30 V
V_{TTin} = 1.80 V
V_{TTout} = 1.20 V
V_{in} = 250 mV_{pp}



Multiple pass test setup

The switch is setup for physical loop back. Electrical PORT B → electrical PORT A.

The signal is first received at electrical input B<3>, is sent through the switch to electrical output A<3>, runs through a 30 cm 3M cable to electrical input B <2>, and is sent through the switch again to electrical output A<2>.

Data window = 500 ps
PRBS: 2^{31}-1 NRZ
BER < $5x10^{-13}$

EYE-WIDTH = 357 ps
EYE-HEIGHT = 193 mV
measured at BER = 10^{-3}

2.0 2.0 Gb/s output jitter measurement

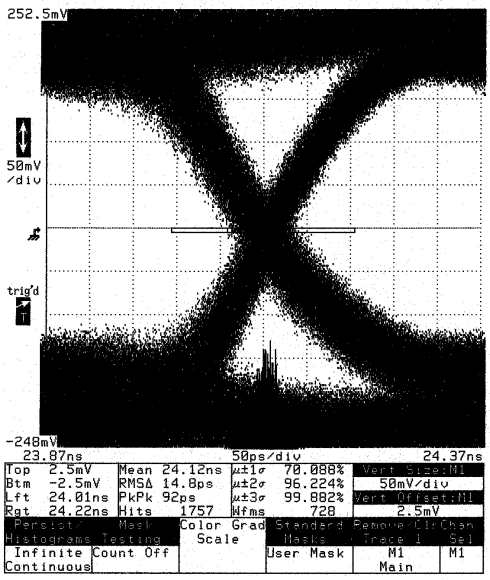


FIGURE TE02.3: Output jitter at 2.0 Gb/s (500 ps bit period, 250 mVpp input, 50 ps/div, 50 mV/div)

DATA RATE: 2.0 Gb/s

V_{DD} = 3.30 V

V_{TTin} = 1.80 V

V_{TTout} = 1.20 V

V_{in} = 250 mV_{pp}

Input node = electrical A<0>

Output node = electrical B<0>

Peak to Peak Jitter = 98 ps

RMS Jitter = 15 ps

PRBS: 2^{31}-1 NRZ

3.0 2.0 Gb/s data output crosstalk measurement

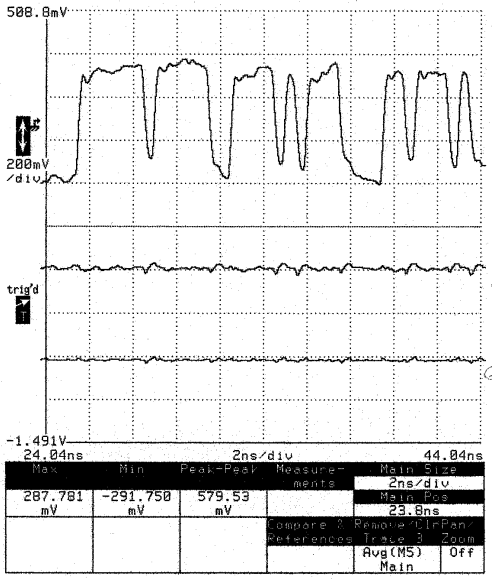


FIGURE TE02.4: Measured output at 2.0 Gb/s. Worst case crosstalk is less than -20 dB

DATA RATE: 2.0 Gb/s

V_{in} = 250 mV_{pp}

Vout = 543 mVpp on electrical A<2>

Noise = 55 mVpp on electrical A<3>

Signal to noise ration (SNR)

≥ 9.87 (19.9 dB)

4. 1.0 Gb/s eye measurement

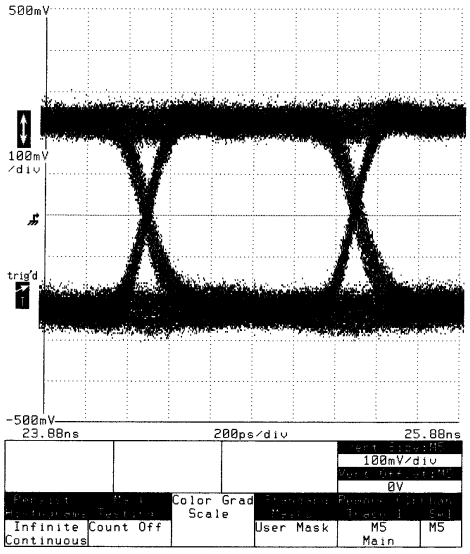
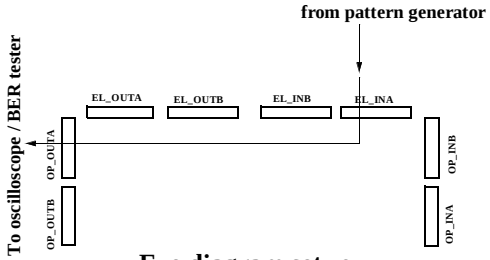


FIGURE TE02.5: Output eye diagram at 1.0 Gb/s (1.0 ns bit period, 200 mVpp input, 200 ps/div, 100 mV/div)

DATA RATE: 2.0 Gb/s

V_{DD} = 3.30 V
V_{TTin} = 1.80 V
V_{TTout} = 1.20 V
V_{in} = 200 mV_{pp}



Eye diagram setup

Input node = electrical A<3>
Output node = optical B<3>

Data window = 500 ps
PRBS: 2^{31}-1 NRZ
BER < $1x10^{-13}$

EYE-WIDTH = 835 ps
EYE-HEIGHT = 257 mV
measured at BER = 10^{-7}

5.0 1.0 Gb/s output jitter measurement

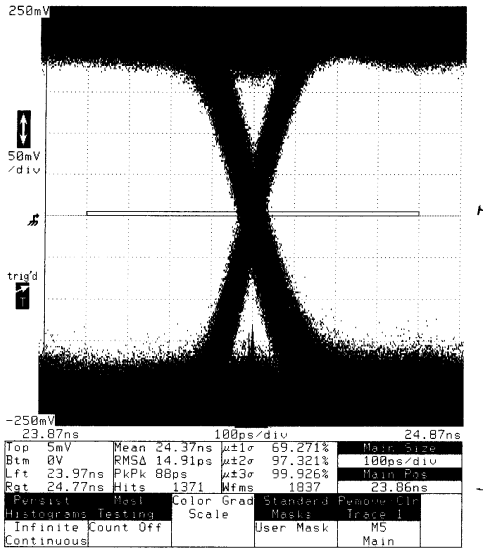


FIGURE TE02.6: Output jitter at 1.0 Gb/s (1.0 ns bit period, 200 mVpp input, 100 ps/div, 50 mV/div)

DATA RATE: 2.0 Gb/s

V_{DD} = 3.30 V
V_{TTin} = 1.80 V
V_{TTout} = 1.20 V
V_{in} = 250 mV_{pp}

Input node = electrical A<3>
Output node = optical B<3>

Peak to peak jitter = 88 ps
RMS Jitter = 15 ps
PRBS: 2^{31}-1 NRZ

6.1 2.0 Gb/s phase margin measurement

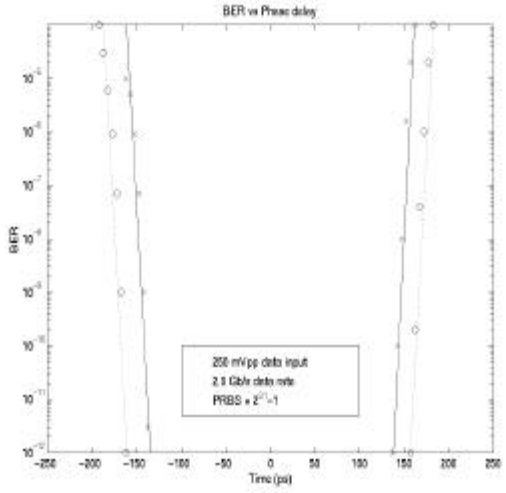


FIGURE TE02.7: Measured best and worst case phase margin of TE02 at 2.0 Gb/s (500 ps/bit, with 250 mVpp data input)

Better than 285 ps phase margin (57% of bit period) at BER = 10^{-11} NRZ 2^{31}-1 PRBS. With insert the worst case phase margin was 240 ps at BER = 10^{-11}

Better than 293 ps phase margin (60% of bit period) at BER = 10^{-11} NRZ 2^{31}-1 PRBS.

6.2 1.0 Gb/s phase margin measurement

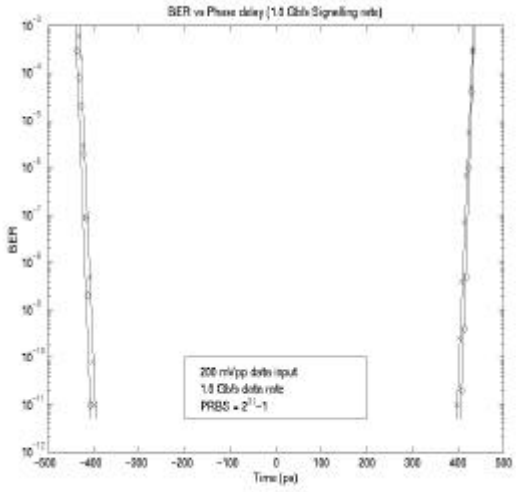


FIGURE TE02.8: Measured best and worst case phase margin of TE02 at 2.0 Gb/s (1.0 ns/bit, with 200 mVpp data input)

Better than 800 ps phase margin (80% of bit period) at BER = 10^{-11} NRZ 2^{31}-1 PRBS (With and without insert).

Better than 820 ps phase margin (82% of bit period) at BER = 10^{-11} NRZ 2^{31}-1 PRBS.

7.0 2.0 Gb/s input sensitivity

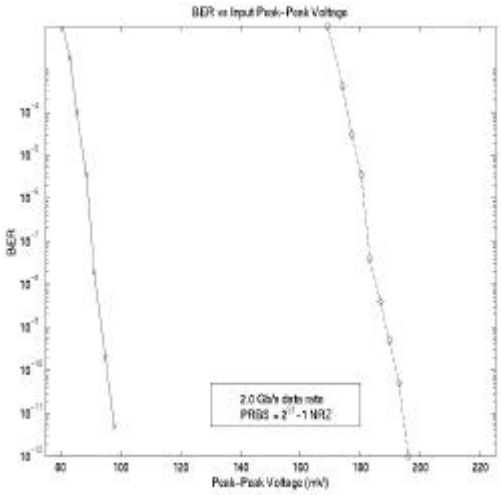


FIGURE TE02.9: Measured best and worst case input sensitivity of TE02 at 2.0 Gb/s (400 ps/bit)

Better than 200 mVpp input sensitivity at BER = 10^{-12} NRZ 2^{31}-1 PRBS. Input common-mode voltage is 1.8 V. The graph does not include inputs that require the insert.

8.1 Switch control (operation of control input C1)

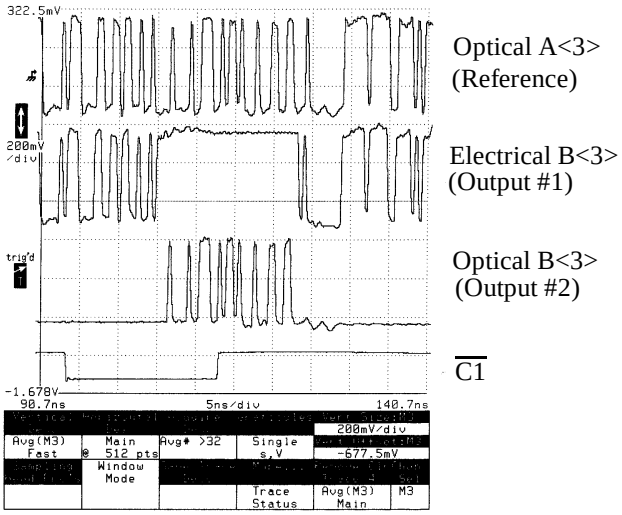
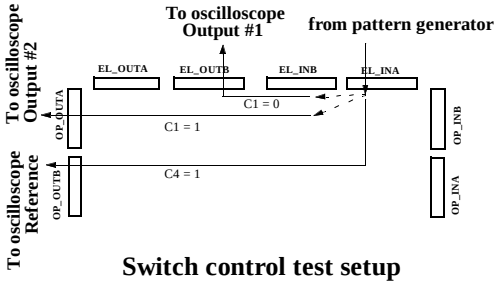


FIGURE TE02.10: TE02 Switch setup test, C<2:4> kept constant while C1 is switched.

Control inputs (C2,C3) = (0,0), C4 = 1.



Switch control test setup

2^7-1 PRBS NRZ data pattern is sent through electrical PORT A<3>.

Using control signal C1 the input is switched between optical PATH A<3> (C1=1) and electrical PORT B<3> (C1=0).

The input C1 is generated by the pulse generator and is synchronized to the data pattern.

The control input C4 is set to 1, so that the input at electrical PORT A<3> is also continuously switched to opitcal PATH A<3>, and is used as a reference.

The output of optical A<3>, the output of electrical B<3>, output of optical B<3> and the complement of control input C1 are shown above.

8.2 Switch control (operation of control input C2)

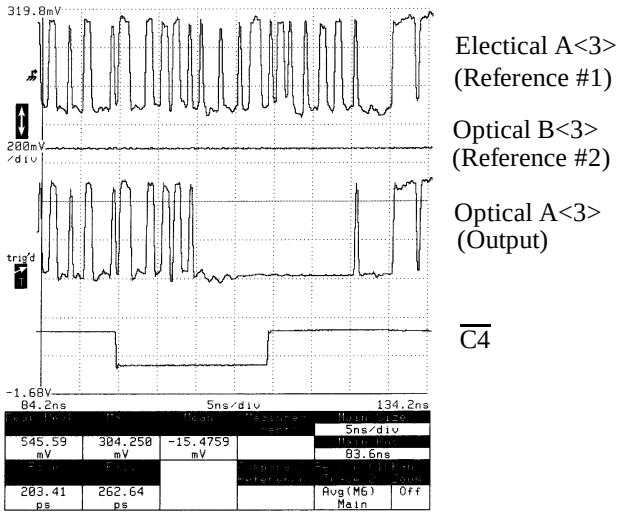
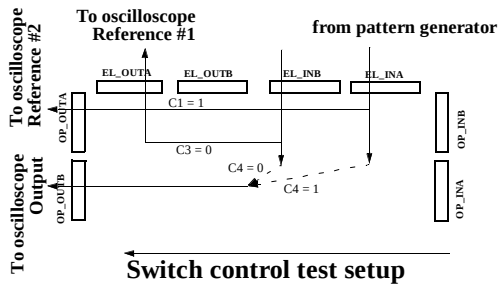


FIGURE TE02.11: TE02 Switch setup test, C<1:3> kept constant while C4 is switched.

Control inputs (C2,C3) = (0,0), C1 = 1.



2^7-1 PRBS NRZ data pattern is sent through electrical PORT B<3>. A static zero is sent through electrical PORT A<3>.

Using control signal C4 the output optical PATH A<3> is switched between inputs electrical PORT B<3> (C4=0) and electrical PORT A<3>(C4=1).

The input C4 is generated by the pulse generator and is synchronized to the data pattern.

The control input C1 is set to 1 and C3 = 0, so that the input at electrical PORT A<3> is also continuously switched to opitcal PATH B<3>, and input at electriacal PORT B<3> is continuously switched to electrical PORT A<4> and are used as a references.

The output of electrical A<3>, the output of optical B<3>, the output of optical A<3>, and the complement of control input C4 are shown above.

8.3 Switch control (operation of control input C3 & C4)

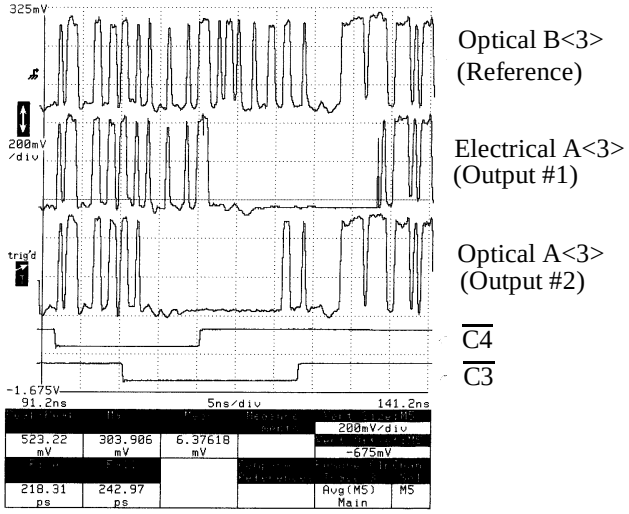
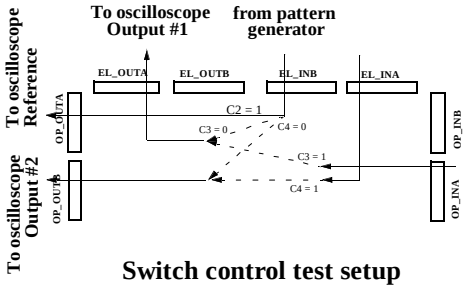


FIGURE TE02.12: TE02 Switch setup test, C<1:2> kept constant while C3 and C4 are switched.

Control inputs (C1,C2) = (1,1).



2^7-1 PRBS NRZ data pattern is sent through electrical PORT B<3>.

Using control signal C3 the output electrical PORT A<3> is switched between inputs electrical PORT B<3> (C3=0) and optical PATH A<3> (C3=1). Using control signal C4 the output optical PATH A<3> is switched between inputs electrical PORT B<3> (C3=0) and electrical PORT A<3> (C3=1).

The input C3 and C4 are generated by the pulse generator and are synchronized to the data pattern.

The control input C2 is set to 1, so that the input at electrical PORT B<3> is also continuously switched to opitcal PATH B<3> which is used as a reference output.

The output of optical B<3>, the output of electrical A<3>, output of optical A<3> and complement of control inputs C3 and C4 are shown above.

9.0 2.0 Gb/s output clock coupling to data

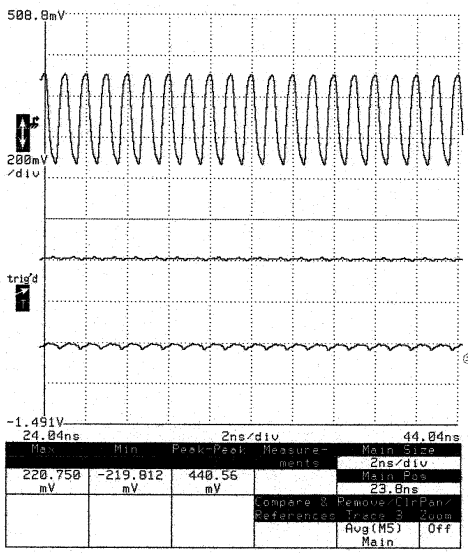


FIGURE TE02.13: Measured clock output at 1.0 GHz clock. Worst case cross-talk is less than -26 dB

Clock frequency 1.0 GHz

V_{DD} = 3.30 V

V_{TTin} = 1.80 V

V_{TTout} = 1.2 V

10.0 2.0 Gb/s clock Jitter (source referred)

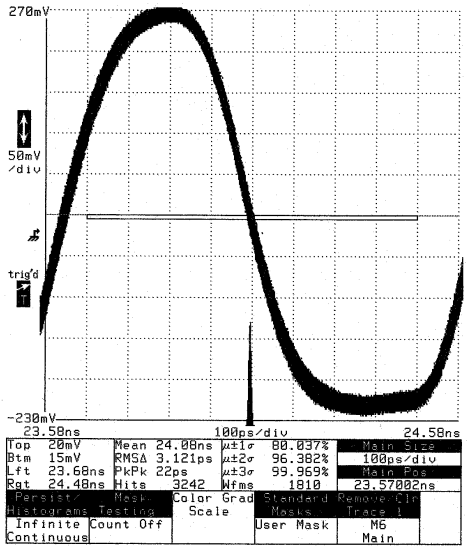


FIGURE TE02.14: Electrical output clock jitter referenced to the electrical input clock.

Clock frequency 1.0 GHz

V_{DD} = 3.30 V

V_{TTin} = 1.80 V

Switch clock select = 0.0 V

Pscntrl = 2.6 V

Duty cycle control <0:1>

= 1.5 V

Peak to peak jitter = 22 ps

RMS jitter = 3.5 ps

When the switch was clocked using the input electrical clock (Switch IC clock select = 0.0 V), the jitter of the electrical output clock and optical output clock referenced to the input clock were 4.7 ps and 4.0 ps RMS and 28 ps and 22 ps peak-to-peak.

BUSY BIT testing

The testing of the BUSY BIT circuitry is broken into four categories

- ***Frame logic test***

  The FRAME logic test includes testing the three different modes of FRAME operation.

  Pass-through mode (Figure TE02.15 through Figure TE02.17): If the current node (switch) is not the master (MS+ = 0), then regardless of the value of the FRAME creation input (FC_IN+), the switch always remains in pass-through mode. If the current node is the master (MS+ = 1), then if the FRAME creation input is not set (FC_IN+ = 0), then the switch remains in pass-through mode. In pass-through mode the output FRAME mirrors the input FRAME.

  Reset mode (Figure TE02.19, Figure TE02.20): During reset mode, the output FRAME is zero regardless of the input FRAME. If the current node is the master (MS+ = 1), and if the FRAME creation input is set (FC_IN+ = 1), then the switch is in reset mode.

  Creation mode (Figure TE02.21): During creation mode, if the input FRAME is low, then it creates a new FRAME depending on the values in the BUSY BIT registers. If the current node is the master (MS+ = 1), and if the input FRAME = 0, and if the FRAME creation input changes from one to zero (FC_IN+ 1 $\rightarrow$ 0), the switch enters into creation mode.
  Since creation mode requires that the input FRAME be zero (“0”), it is important that the master resets the FRAME for a long enough time so that the FRAME is reset throughout the network (ring). In order to create a FRAME, before the transition of the FRAME create input, one or more of the request interface registers should be set (starting with request interface register <3> (BS<3>).
  The length of the created FRAME is set to the number of request bits set starting from B<3> + 1. Thus, if only one request bit (B<3>) is set, then the length of the created FRAME is 2, corresponding to a 2-processor network.

- ***Encode, decode, and inversion control logic test***

  Successful encoding of 2, 3 and 4-bit wide FRAME’s are shown in Figure TE02.15 and Figure TE02.16. Figure TE02.17 shows the case where an unencoded FRAME is sent through FRAME_IN, gets encoded using the FRAME encoding circuitry, and is then re-input as the BUSY BIT, and gets decoded and finally re-encoded as the BUSY BIT output. To indicate a successful decoding and encoding of the BUSY BIT, the encoded BUSY_OUT signal should mirror the encoded FRAME_OUT signal.

- **BUSY BIT logic test**

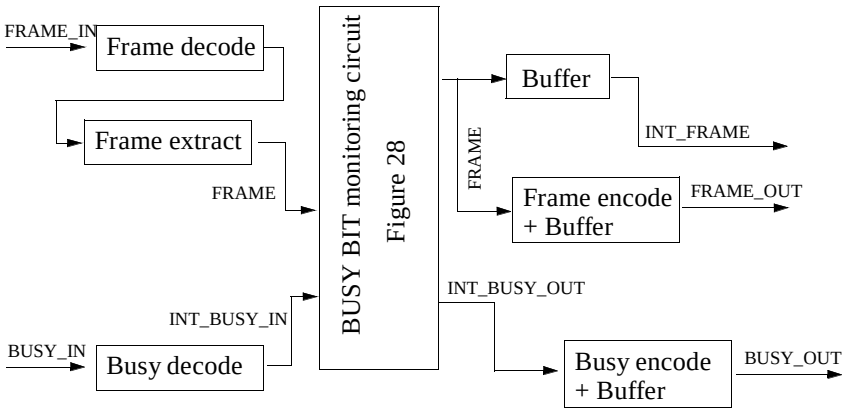
  BUSY BIT testing includes BUSY BIT pass-through, and BUSY BIT grab and release.

- External interface (busy interface registers, grab interface registers and clocking them)

The tests revealed that the current-steering RS latches were unable to hold values when they were set. This hindered the complete test of the BUSY BIT circuitry in TE02, but was corrected in TSMC_BUSY chip. Please refer to the next section (section 8.3) for complete test results of the BUSY BIT circuitry in the TSMC_BUSY chip. By going to a smaller feature size we were able to also reduce the power consumption and increase the BUSY BIT operating frequency to 1.5 GHz (3.0 Gb/s signaling rate).

This set of measurements demonstrates correct encoding and FRAME pass-through mode operation.

- If the master is low (MS+ = 0), or if master and FRAME control are both low (FC_IN+,MS+ = 0,0), the FRAME extraction circuitry will be in pass-through mode.
- If master and FRAME control are both high (FC_IN+,MS+ = 1,1) and one or more Busy Requests including B<3> are set and FC_IN+ has a 1 → 0 transition, the circuit will be in pass-through mode after FRAME creation
- The example of pass-through of a 4-bit wide FRAME was used earlier as a successful encoding example (Figure TE02.17).



11.0 Encoding and pass-through of a 2-bit wide FRAME

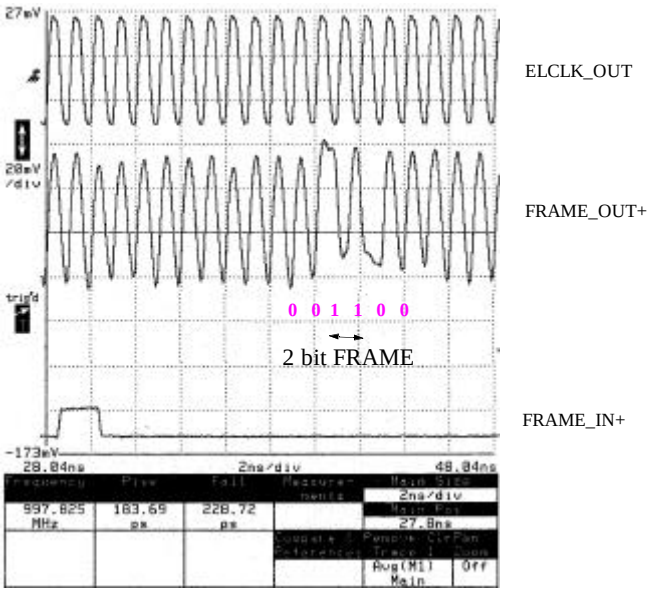


FIGURE TE02.15: Encode example using 2-bit wide FRAME

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example the switch is set for FRAME pass-through, and an unencoded 2-bit wide FRAME is sent through FRAME_IN±, and FRAME_OUT+ shows the correctly encoded output FRAME.

12.0 Encoding and pass-through of a 3-bit FRAME

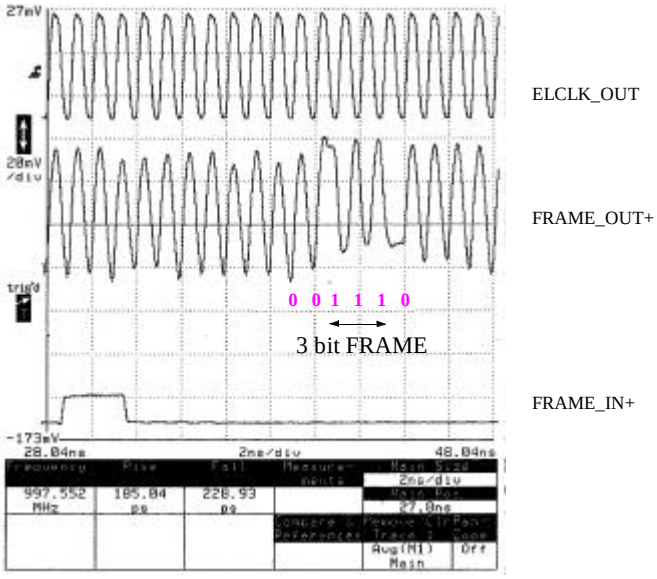


FIGURE TE02.16: Encode example using 3-bit wide FRAME.

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example the switch is set for FRAME pass-through, and an unencoded 3-bit wide FRAME is sent through FRAME_IN±. FRAME_OUT+ shows the correctly encoded output FRAME.

13.0 Encoding and pass-through of a 4-bit FRAME

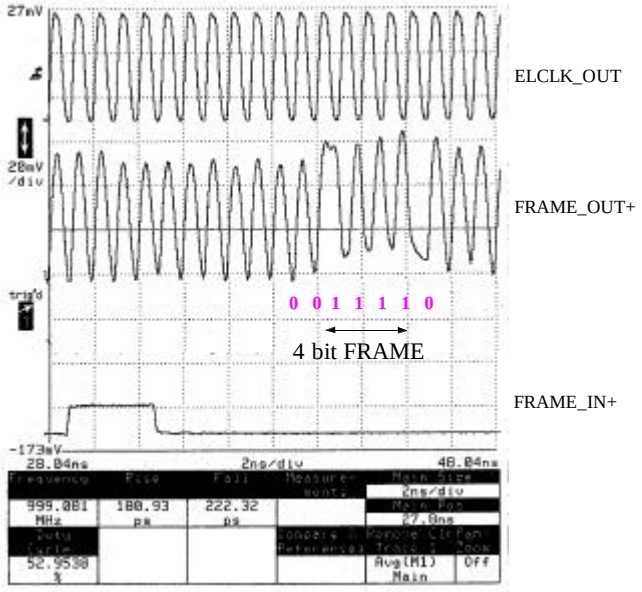


FIGURE TE02.17: Encode example using 4-bit wide FRAME

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example the switch is set for FRAME pass-through, and an unencoded 4-bit wide FRAME is sent through FRAME_IN± (the circuit allows either encoded or unencoded input).
- FRAME_OUT+ shows the encoded output FRAME.
- ELCLK_OUT (electrical clock output) is shown for reference.

14.0 Successful encode / decode / encode

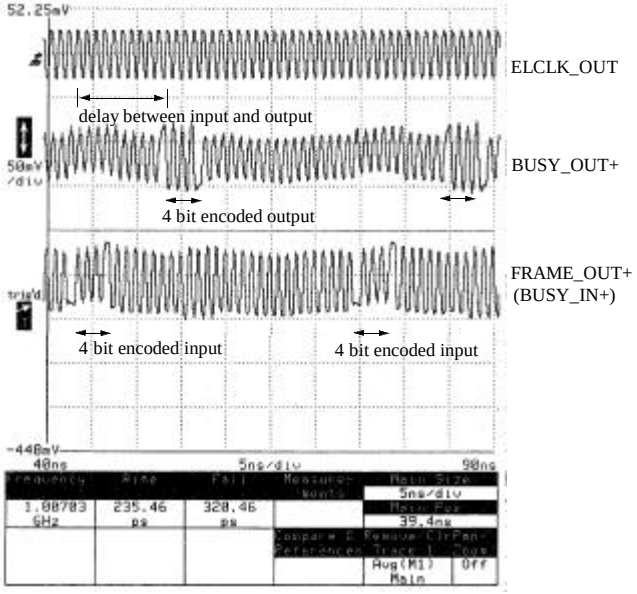


FIGURE TE02.18: Example of a successful encode/ decode / encode.

Clock frequency 1.0 GHz (2.0 Gb/s)

- As in the previous example the switch is set for pass-through. The encoded 4-bit wide output FRAME is then sent in through BUSY_IN± as a mock input BUSY BIT.
- The switch is not requesting any BUSY BIT, so the output BUSY BIT should mirror the BUSY BIT input (in this case the encoded output FRAME). As expected, the BUSY BIT output BUSY_OUT+ mirrors the encoded FRAME output.
- The 10 cycle delay between the input and output is due to the 3M cable and the internal BUSY BIT circuit logic (4.5 cycles).

This set of measurements demonstrates the functionality of the input inversion and the FRAME reset logic.

- MS+ = 0 identifies this as the master switch with the right to reset/create the FRAME.
- If IN_INV = 0, the input FRAME and BUSY BIT will not be inverted. On the other hand if IN_INV=1, then the inputs will be inverted (that is, if IN_INV=1 then FRAME_IN =1 will generate INT_FRAME = 0).
- If the master and FRAME control are both high (FC_IN+,MS+ = 1,1), the internal FRAME will be zero regardless of the input FRAME.

15.0 Example of inversion control

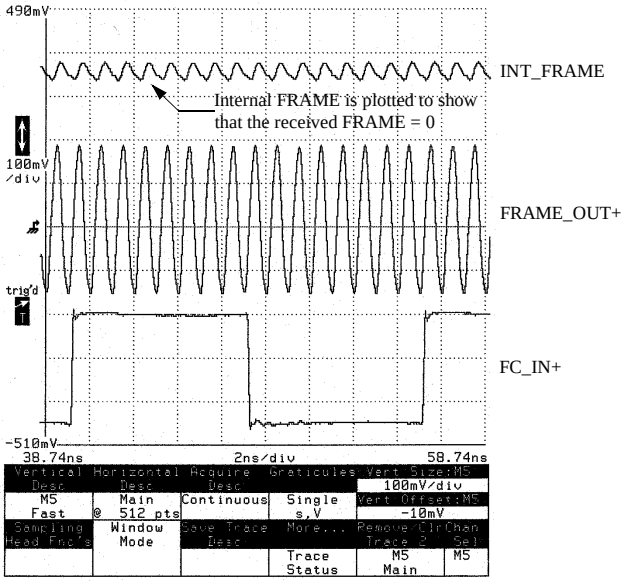


FIGURE TE02.19: Frame reset example with IN_INV = 0 (no input inversion).

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example, a static zero is sent on FRAME_IN while MS+ is continuously high.
- Request interface register <3> = 0 (B<3> = 0), otherwise, in the above example when FC_IN+ does a 1 → 0 transition, a FRAME would be created.
- Since no inversion of the received FRAME is requested, the internal FRAME and output FRAME are always zero.

16.0 Example of FRAME reset, and inversion control

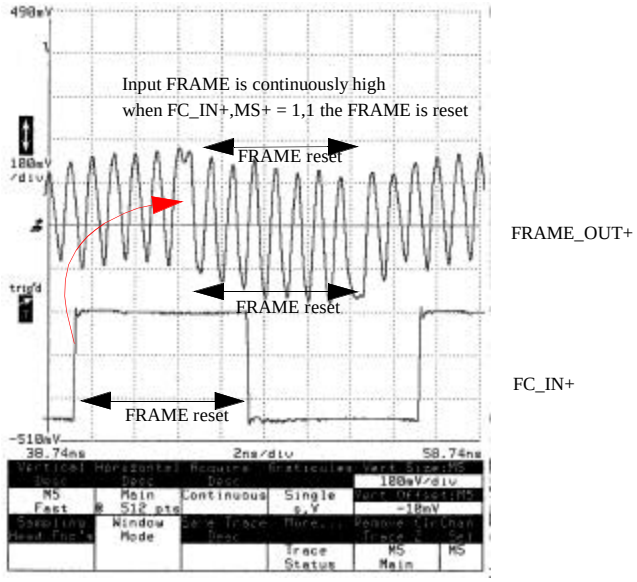


FIGURE TE02.20: Frame reset example with IN_INV = 1 (with input inversion).

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example also a static zero is sent on FRAME_IN while MS+ is continuously high.
- Since inversion is selected, the received FRAME is continuously one.
- When FC_IN+ is high, the FRAME is reset. When FC_IN+=1, both the internal FRAME (INT_FRAME) and the output FRAME (FRAME_OUT) are zero.
- When FC_IN+ is low the logic goes to FRAME pass-through mode. When FC_IN+=0, both the internal FRAME (INT_FRAME) and the output FRAME (FRAME_OUT) are one, mirroring the received FRAME.

This set of measurements demonstrates the functionality of the FRAME creation logic.

- Before FRAME creation, the master resets the FRAME and ensures a FRAME = 0 continuously flows through the network. Thus, for this example we also set MS+=0 (to identify the switch as the master switch with the right to reset/ create the FRAME), IN_INV=0 (no inversion of the inputs), and FRAME_IN=0 (FRAME input is zero).
- The length of the created FRAME = # of request bits set starting from B<3> + 1 (so if only one request bit is set, the length of the created FRAME = 2)
- When FC_IN+ changes form 1 → 0, the FRAME will be created.

17.0 Example of FRAME creation

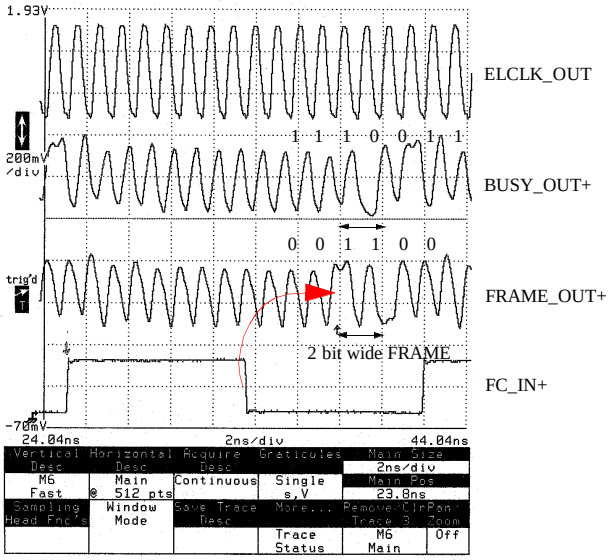


FIGURE TE02.21: Frame creation (creation of 2 bit wide FRAME).

Clock frequency 1.0 GHz (2.0 Gb/s)

- In this example a static zero is sent on FRAME_IN while MS+ is continuously high.
- Since B<3> is set, BUSY_OUT should remain at one while the FRAME is being reset.
- When FC_IN+ does a 1 → 0 transition, a FRAME is internally created, and since only B<3> is set the length is one.
- Since B<2> =0, after the first shift, BUSY_OUT should go low, and when the FRAME passes, there will be no more shifting, so the BUSY_OUT should now again reflect the value of B<3> which is high (set).
- As expected when B<3> was not set and one or more of the remaining request interface registers were set, a FC_IN+ 1 → 0 transition did not create a FRAME

This set of measurements demonstrates the functionality of the BUSY BIT circuit when no requests are being made.

- The switch is set for FRAME pass-through (MS+,FC_IN+=0).
- A 4-bit wide FRAME is input.
- Since there are no requests, the output BUSY BIT should mirror the input BUSY BIT, regardless of the FRAME.
- If B<3> is not set, then when there is no FRAME, the output BUSY BIT should mirror the input BUSY BIT. On the other hand if B<3> is set, when there is no FRAME, the output should be high (one) before encoding.

18.0 Examples of BUSY BIT pass-through (no requests at current node)

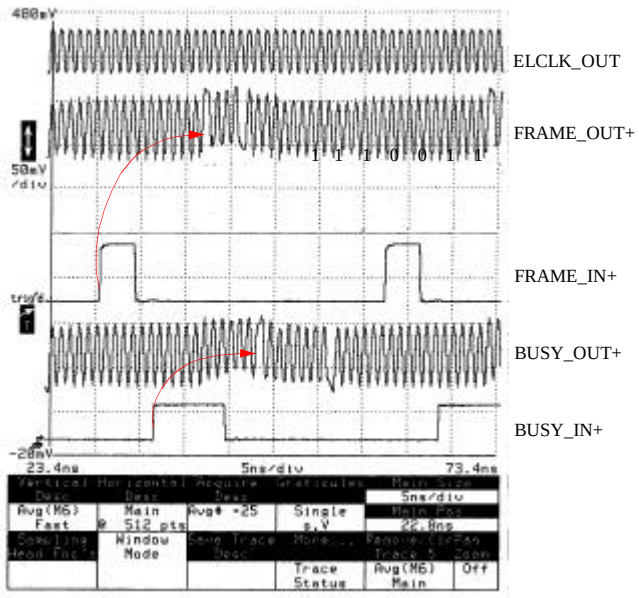


FIGURE TE02.22: BUSY BIT pass-through, node is not requesting any BUSY BITs.

Clock frequency 1.0 GHz (2.0 Gb/s)

- The example shows that the FRAME was received and transmitted properly.
- The input shows that currently all the input BUSY BITs are free (BUSY_IN is zero during FRAME).
- After the FRAME the BUSY BIT input is made high for a short period.
- As expected the output BUSY BIT mirrors the input BUSY BIT regardless of the FRAME

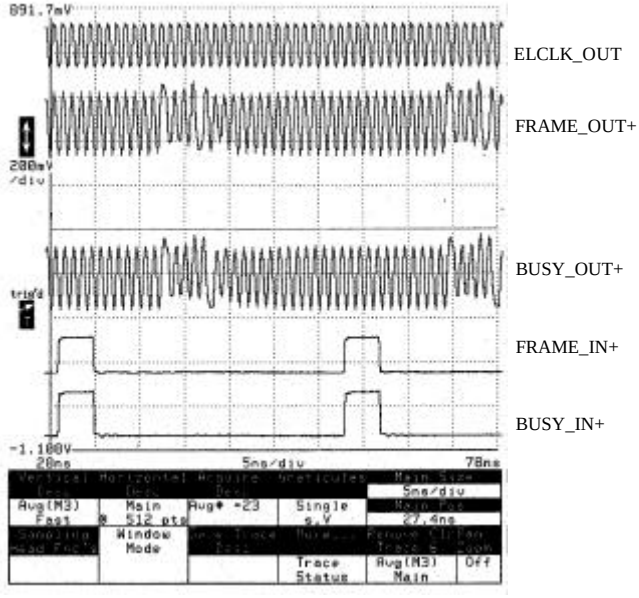


FIGURE TE02.23: BUSY BIT pass-through, node is not requesting any BUSY BITs.

Clock frequency 1.0 GHz (2.0 Gb/s)

- This example shows that the FRAME was received and transmitted properly.
- The input shows that currently all the input BUSY BITs are grabbed (BUSY_IN is high during FRAME).
- As expected the output BUSY BIT mirrors the input BUSY BIT regardless of the FRAME, except for the bit right after the FRAME.

This set of measurements demonstrates the BUSY BIT grab functionality.

- The switch is set for FRAME pass-through (MS+,FC_IN+=0).
- A 4-bit wide FRAME in input.
- A BUSY BIT request is made using the request interface registers.
- When a request for a BUSY BIT is made, if the corresponding input BUSY BIT is free (zero), then the BUSY BIT will grabbed, and the corresponding output BUSY BIT will go high (one).
- This test does not test the grab registers or the BUSY BIT release.

19.0 Examples of request and grab of BUSY BIT (BO<3>)

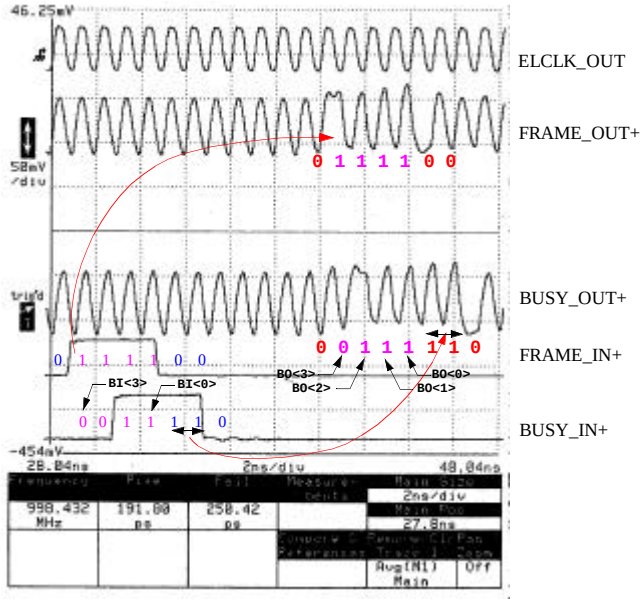


FIGURE TE02.24: BUSY BIT grab example, where part of the bits are already reserved.

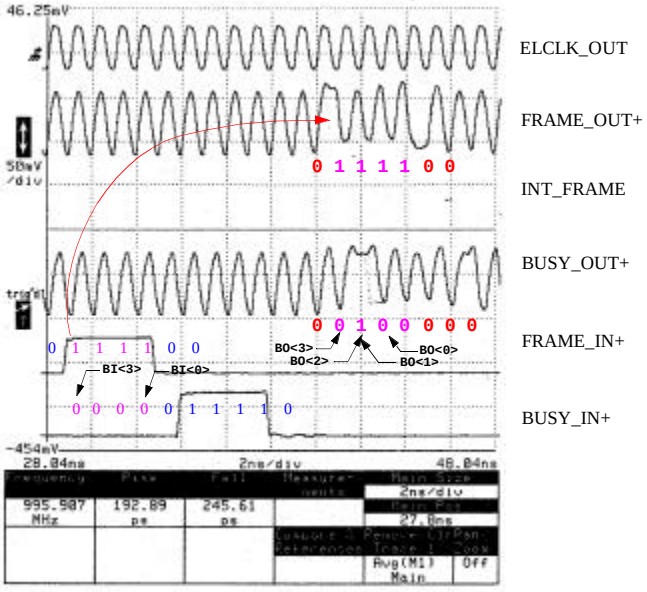


FIGURE TE02.25: BUSY BIT grab example, where all the bits are initially free.

Clock frequency 1.0 GHz (2.0 Gb/s)

- In the example the FRAME is 4 bits wide.
- Input BUSY BITs BI<0> and BI<1> are high, BUSY BITs <1> and <0> are busy, having already been grabbed. The decoded input during the FRAME is 0011.
- The current node is requesting BUSY BIT <2>.
- As expected, the output BUSY BIT (decoded) is 0111, which indicates that the current node has grabbed BUSY BIT <2>, and the only BUSY BIT that is free is BUSY BIT <3>.

Clock frequency 1.0 GHz (2.0 Gb/s)

- The input FRAME is 4-bit wide.
- All of the input BUSY BITs are zero, indicating that all the bits are free.
- The current node is requesting BUSY BIT <2>.
- As expected, the output BUSY BIT (decoded) is 0100, which indicates that the current node has grabbed BUSY BIT <2>, and all others are still free.

System level testing

Following the data path and BUSY BIT testing, two ICs will be interconnected to simulate a network of four processors. This will verify the basic data transfer between ICs.

Due to the constraints of the test equipment, test setup, required time and cost a final test using two boards and 4 processors (Figure 36) will not performed.

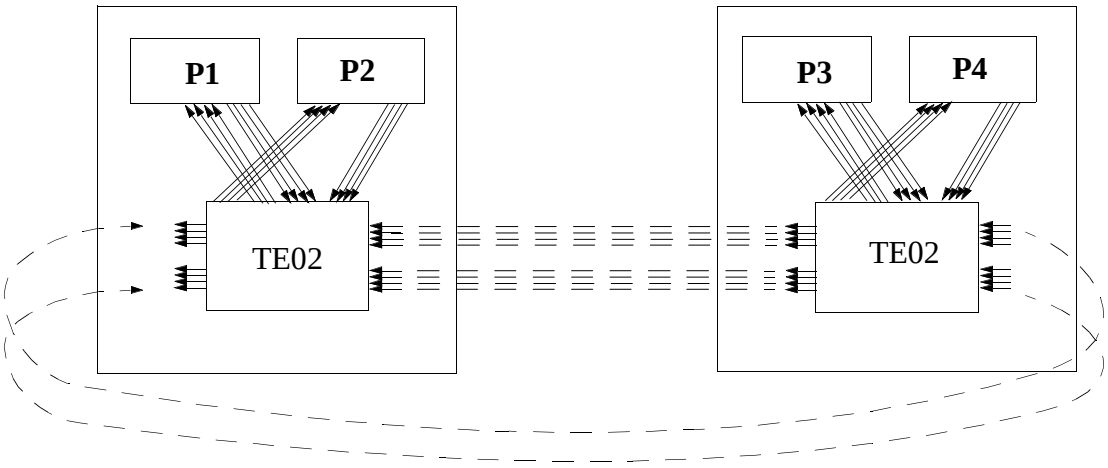


FIGURE 36: TE02 connection supporting 4 processors, grouped 2 processors per board

Discussion

Due to pad frame constraints, there were long power and ground connections, and power and ground were shared between blocks. Reduction in the power rails (> 10%) due to shared power rails and long power lines, as well as coupling between bits as well as blocks due to the shared power rails are expected to reduce the phase margins and the input sensitivity of the receivers. The common-mode input termination voltage V_{ttin} is also shared by the data and clock receivers. The on-chip termination resistors were measured to be between 41 and 43 Ohms, corresponding to a reflection coefficient of -0.09.

The test results of TE02 shows 100 ps peak-to-peak output jitter for 2.0 Gb/s, 2^{31}-1 PRBS NRZ signaling rate, and a output phase margin of greater than 275 ps for BER = 10^{-12}. 2^{7}-1 PRBS NRZ data pattern had a worst case BER = 10^{-12} phase margin of greater than 325 ps. The PRBS length dependance in phase margin can be attributed to a drive problem at the output interface muxes.

In TE02, there were 3.5 mm long on-chip wires between the electrical input/output interfaces and the switch core interface. At present most literature models such wires as pure RC transmission lines under the assumption that RC delay is large compared to the signal flight time. Which is true for lower level metal wiring, but is not true for third or fourth layer metal wiring. Even in a Switch IC integrating optics and CMOS such wire lengths can be expected, and thus needs to be modeled better.

Even though most of the basic operations of the BUSY BIT logic were tested, the test results suggest that the NMOS transistors used in the current steering RS latches need to be resized. Due to sizing problems of the NMOS transistors in the RS latches, the latches had difficulty retaining the voltages when they were set. The strength problem of the NMOS pull-down transistors in the current-steering latches of the BUSY BIT interface, grab registers and interface were identified and modified in the TSMC-BUSY circuit designed and tested in 0.25 µm TSMC CMOS process. The test results for the TSMC_BUSY is given in section 8.3. A re-sized version of the TE02 IC is required to ensure robus operation.

8.3 TSMC_BUSY (0.25 μmCMOS test die on TSMC)

8.3.1 Introduction

The BUSY BIT circuitry that was implemented in 0.35 μm HP CMOS process was directly scaled to the 0.25 μm TSMC CMOS process and operates at 3.0 Gb/s. During simulation of the BUSY BIT circuitry an NMOS strength problem in the RS latches was detected and fixed.

Except for this fix, the BUSY BIT circuitry present in the TSMC_BUSY IC is a direct scale of the TE02 BUSY BIT circuitry and the BUSY BIT discussion presented in section 8.2.2 also applies to the TSMC_BUSY test IC. The test IC was tested at 2.5 V (1.5 Gb/s) and 2.7 V (1.75 Gb/s) power rails.

BUSY BIT testing

Similar to the BUSY BIT testing in TE02, the tests are broken into four categories

- ***Frame logic test***
  Frame logic test includes the testing the three different modes of FRAME operation.

  A discussion on the three different modes of FRAME operation were given in section 8.2_BUSY BIT testing. Example test results for pass-through and reset are not explicitly shown. Figure TSMC.3and Figure TSMC.4 show exmples of a creation of a 4-bit wide FRAME, for 3.0 Gb/s signaling rate (2.5 V rail, 1.5 GHz clock), and 3.5 GB/s signaling reate (2.7 V, 1.75 GHz clock).

- ***Encode, decode, and inversion control logic test***
  Figure TSMC.1 (3.0 Gb/s, 2.5 V) and Figure TE02.2 (3.5 Gb/s, 2.7 V) show the case where an unencoded FRAME is sent through FRAME_IN, gets encoded using the FRAME encoding circuitry, and is then re-input as the BUSY BIT, and gets decoded and finally re-encoded as the BUSY BIT output. To indicate a successful decoding and encoding of the BUSY BIT, the encoded BUSY_OUT signal should mirror the encoded FRAME_OUT signal.

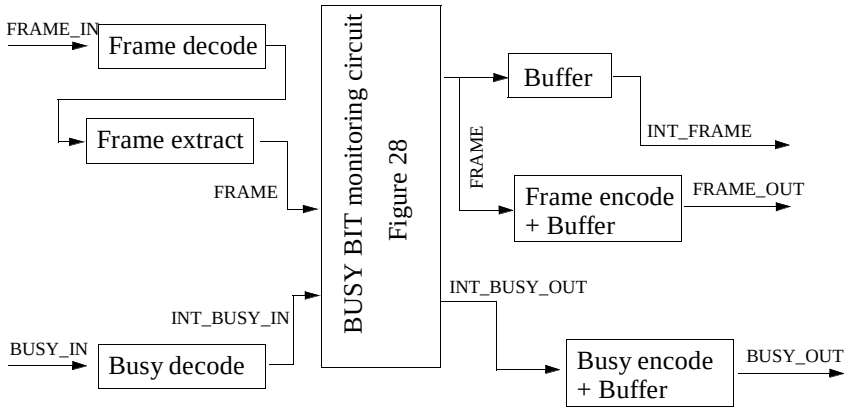
- **BUSY BIT logic test**
  BUSY BIT testing includes BUSY BIT pass-through, and BUSY BIT grab and release (Figure TSMC.5, Figure TSMC.6). The proper operation of BUSY BIT pass-through, grab and release proves that the BUSY BIT logic and the grab registers are working properly. Also this test will include the proper operation of grab register set and reset operation. The grab register set signal, sets the 4-bit wide grab register to all ones, and the grab register reset signal resets the grab register to all zeros (Figure TSMC.7, Figure TSMC.8).

- **External interface** (busy interface registers, grab interface registers and clocking them)
  The external interface includes the request interface and grab interface registers. The request interface registers can be individually set and reset (set - request, reset - release). The proper operation of FRAME creation and BUSY BIT grab reset proves the operation of the request interface registers. The grab interface register operation is shown in Figure TSMC.5, and Figure TSMC.6.

NOTE: All aspects of the BUSY BIT circuitry that could be tested using the available test equipment were tested at 3.0 Gb/s GHz (2.5 V) and 3.5 Gb/s (2.7 V rail). For an operating frequency of 1.5 GHz (3.0 Gb/s signaling rate) the IC consumed 80 mA at 2.5 V power rail and 95 mA at 2.7 V power rail.

This set of measurements demonstrates correct encoding, decoding and FRAME pass-through mode operation.

- If the master is low (MS+ = 0), or if master and FRAME control are both low (FC_IN+,MS+ = 0,0), then the FRAME extraction circuitry will be in pass-through mode. For the test MS- = 0 (MS+ = 1) and FC_IN+ = 0.
- If master and FRAME control are both high (FC_IN+,MS+ = 1,1) and one or more of the BUSY BIT requests including B<3> are set and FC_IN+ has a 1 → 0 transition, the circuit will be in pass-through mode only after FRAME creation
- For the example a 4-bit wide unencoded signal is sent through FRAME_IN. The output FRAME_OUT is resent through the BUSY_IN and the output BUSY_OUT is monitored to show the proper operation of encode/decode.



1.0 3.0 Gb/s (2.5 V) successful encoding / decode / encode

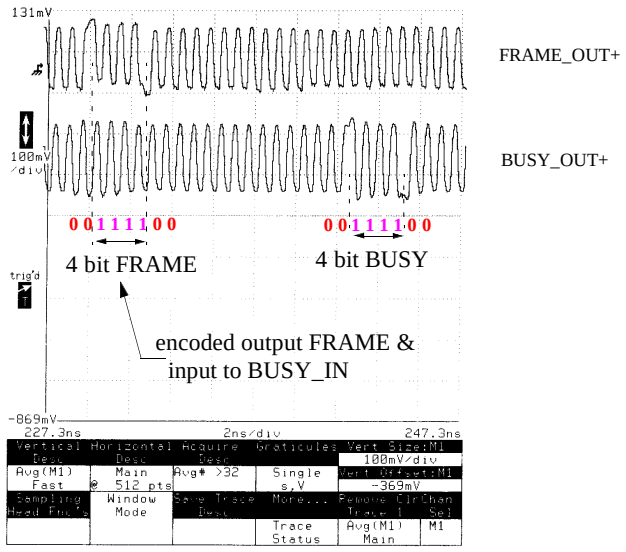


FIGURE TMSC.1: Successful encode / decode / encode example.

Clock frequency 1.5 GHz (2.0 Gb/s, 2.5 V rail)

- All of the request registers are reset, therefore the output BUSY BIT should mirror the input BUSY BIT which is also the FRAME_OUT.
- 4 Cycles of the delay between the input and output of BUSY are due to BUSY BIT logic and the rest is external delay due to cables etc.

2.0 3.5 Gb/s (2.7 V) successful encoding / decode / encode

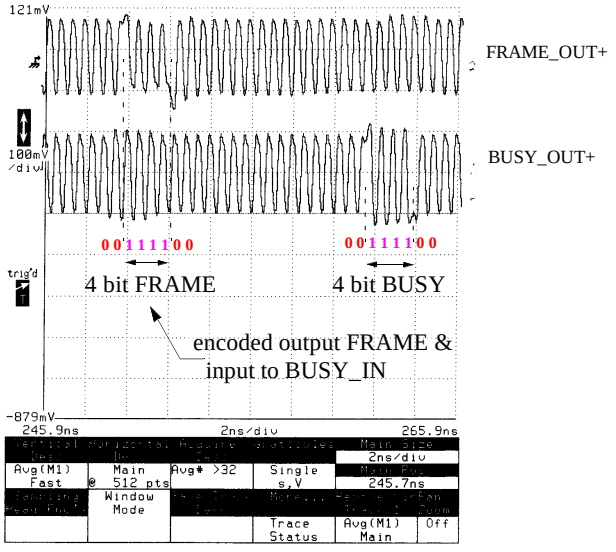


FIGURE TMSC.2: Successful encode / decode / encode example.

Clock frequency 1.5 GHz (2.0 Gb/s, 2.5 V rail)

- All of the request registers are reset, therefore the output BUSY BIT should mirror the input BUSY BIT which is also the FRAME_OUT.
- 4 Cycles of the delay between the input and output of BUSY are due to BUSY BIT logic and the rest is external delay due to cables etc.

This set of measurements demonstrates the functionality of the FRAME creation logic.

- Before FRAME creation, the master resets the FRAME and ensures a FRAME = 0 continuously flows through the network. Thus, for this example we also set MS+=0 (to identify the switch as the master switch with the right to reset/ create the FRAME), IN_INV=0 (no inversion of the inputs), and FRAME_IN=0 (FRAME input is zero).
- The length of the created FRAME = # of request bits set starting from B<3> + 1.
- When FC_IN+ changes form 1 → 0, the FRAME will be created.

3.0 3.0 Gb/s (2.5 V) example of FRAME creation

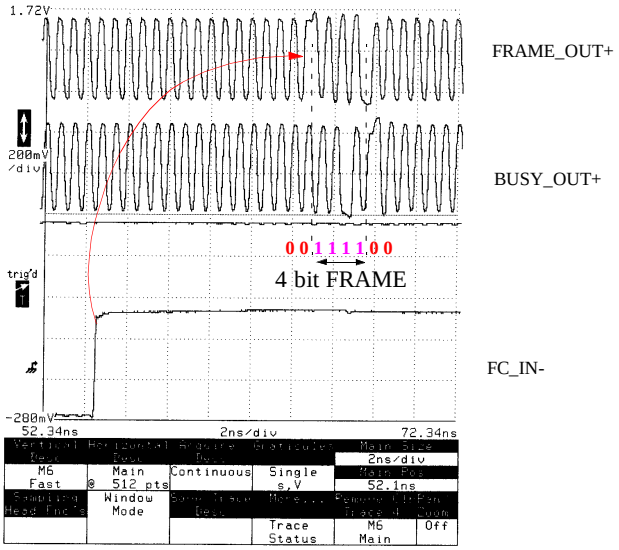


FIGURE TMSC.3: Frame creation (creation of 2 bit wide FRAME).

Clock frequency 1.5 GHz (3.0 Gb/s, 2.5 V)

- The test board --- has close to 300 ps skew between FRAME_OUT+ and BUSY_OUT+ signals. The 300 ps skew is significant compared to the 333 ps phase of the clock.
- In this example a for a portion of the time a one is sent, during this time period MS and FC_IN are activated, thus the FRAME is reset.
- Since B<3> is set, BUSY_OUT should remain at one while the FRAME is being reset.
- When FC_IN+ does a 1 → 0 transition (note that the figure shows FC_IN-), a FRAME is internally created, and since B<3:1> are set, the the FRAME is 4-bits wide.
- As expected when B<3> was not set and one or more of the remaining request interface registers were set, a FC_IN+ 1 → 0 transition did not create a FRAME

4.0 3.5 Gb/s (2.7 V) example of FRAME creation

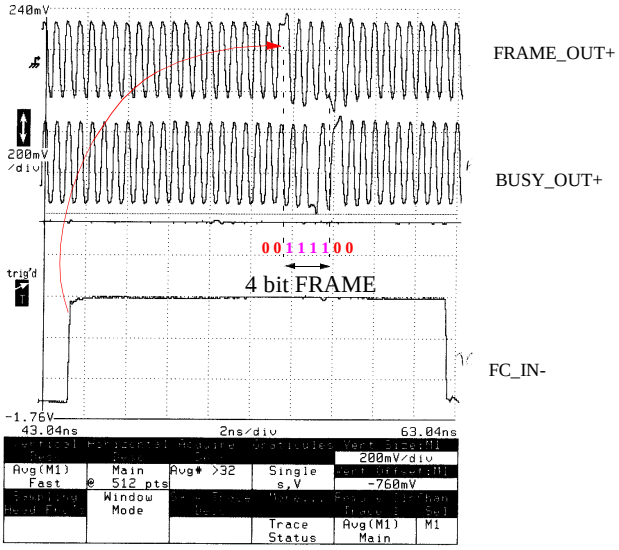


FIGURE TMSC.4: Frame creation (creation of 2 bit wide FRAME).

Clock frequency 1.75 GHz (3.5 Gb/s, 2.7 V)

- The test board --- has close to 300 ps skew between FRAME_OUT+ and BUSY_OUT+ signals. The 300 ps skew is significant compared to the 286 ps phase of the clock.
- In this example a for a portion of the time a one is sent, during this time period MS and FC_IN are activated, thus the FRAME is reset.
- Since B<3> is set, BUSY_OUT should remain at one while the FRAME is being reset.
- When FC_IN+ does a 1 → 0 transition (note that the figure shows FC_IN-), a FRAME is internally created, and since B<3:1> are set, the the FRAME is 4-bits wide.
- As expected when B<3> was not set and one or more of the remaining request interface registers were set, a FC_IN+ 1 → 0 transition did not create a FRAME

This set of measurements demonstrates the BUSY BIT request / grab / release and grab interface register functionality.

- The switch is set for FRAME pass-through (MS+,FC_IN+=0).
- To test the complete operation of the BUSY BIT (grab, release and grab interface registers), the input FRAME should be 4-bits wide.
- BUSY BIT requests are made using the request interface registers.
- The setup inputs two 4-bit wide pulses as the FRAME, and synchronized with it before the first FRAME B<3> is set, and after the first FRAME it is reset. This simulates a request and then a release. Also after each FRAME the grab interface registers are loaded (active low). Since BUSY BIT <0> is continuously requested we expect GRAB_OUT<0> to remain at zero. Since BUSY BITs <2> and <1> are not requested the corresponding grab interface register outputs should remain at one.
- The first FRAME is shown in Figure TSMC.5 and second FRAME is shown in Figure TSMC.6.

5.0 3.0 Gb/s (2.5 V) example of request, grab, release and grab interface register operation.

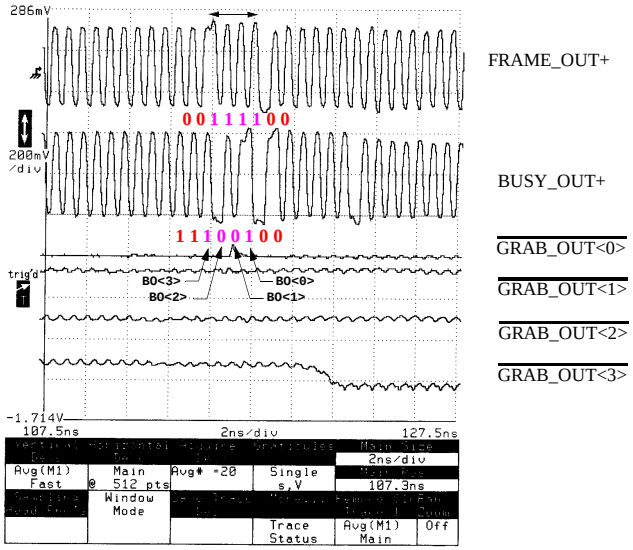


FIGURE TMSC.5: BUSY BIT grab example.

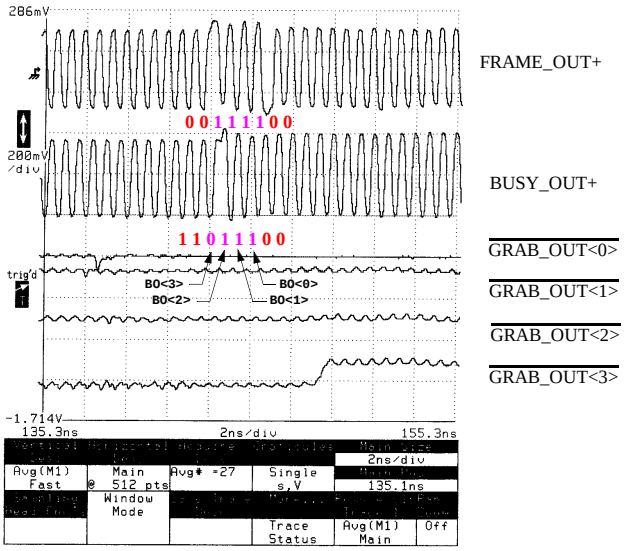


FIGURE TMSC.6: BUSY BIT release example.

Clock frequency 1.5 GHz (3.0 Gb/s, 2.5 V)

- In the example the FRAME is 4 bits wide.
- All the input BUSY BITs are all zero, indicating that all the bits are free.
- The current node is requesting BUSY BIT <3> and BUSY BIT<0>
- As expected, the output BUSY BIT (decoded) is 1001, which indicates that the current node has grabbed BUSY BIT <3> and BUSY BIT <0>.
- As expected GRAB_OUT<2:0> = 110, and GRAB_OUT<3> changes from 1 to 0. BUSY BITs <3> and <0> were grabbed by the node.

Clock frequency 1.5 GHz (3.0 Gb/s, 2.5 V)

- The input FRAME is 4-bit wide.
- All of the input BUSY BITs are one, indicating that all the BUSY BITs are grabbed.
- Since this node had requested and grabbed bits <3> and <0>, if they are no longer requested by the node they should be released.
- The node is no longer requesting <3> (release BUSY BIT <3>).
- As expected, the output BUSY BIT (decoded) is 0111, which indicates that the current node has released BUSY BIT <3>, while the node is still holding BUSY BIT <0>.
- Also as expected GRAB_OUT<2:0> remains at 110, and GRAB_OUT<3> changes from 0 to 1 indicating that <3> was released.
- A BUSY BIT can only be released if the corresponding grab bit was one (grab register was working properly). Thus this test also tests the proper operation of the grab register.

This set of measurements demonstrates the grab register reset and set signal functionality.

- The switch is set for FRAME pass-through (MS+,FC_IN+=0).
- Since BUSY BIT request / grab and release is used in the test, here also the inut FRAME should be 4-bits wide.
- The test setup inputs two 4-bit wide pulses as the FRAMEs. During the first FRAME the input BUSY BITs are all zero, and a request is made for BUSY BIT <0> (the request interface register bit <0> is set). Thus the IC should grab the BUSY BIT <0> during the first FRAME.
- This request is then reset before the second FRAME (the request interface register bit <0> is reset). During the second FRAME the input BUSY BITs are all one. Thus the IC should release the BUSY BIT <0> during the second FRAME, and the output BUSY BITs should be 1110. This is because the IC is the one holding the BUSY BIT <0> yet no longer wants it.
- If after the first FRAME all the grab register bits had been set, then during the second FRAME the output BUSY BITs should be all zeros (all the bits will be released). This example is shown in Figure TSMC.7.
- If after the first FRAME all the grab register bits had been reset, then during the second FRAME the output BUSY BITs should be all ones (none of the bits will be released). This example is shown in Figure TSMC.8.

6.0 3.0 Gb/s (2.5 V) example of grab register set.

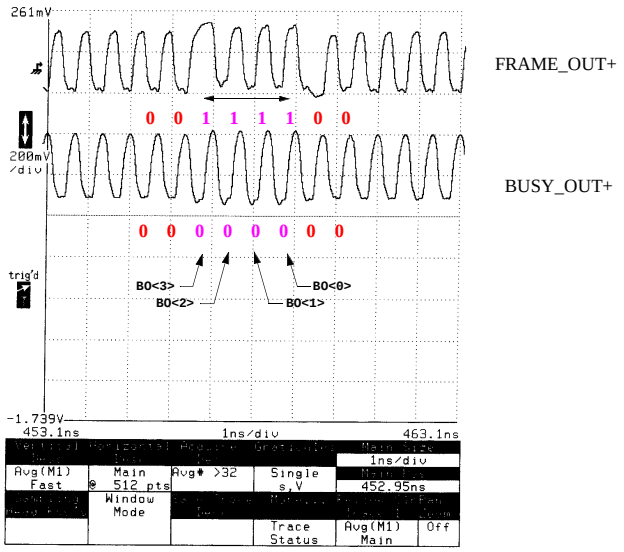


FIGURE TMSC.7: 3.0 Gb/s example of grab register set operation.

Clock frequency 1.5 GHz (3.0 Gb/s, 2.5 V)

- In the example the FRAME is 4 bits wide.
- All the input BUSY BITs are one, indicating that all the bits are grabbed (busy).
- As expected, since the grab register was set before the second FRAME, even though this node is only supposed to release BUSY BIT <0>, all the BUSY BITs are released.

7.0 3.0 Gb/s (2.5 V) example of grab register reset.

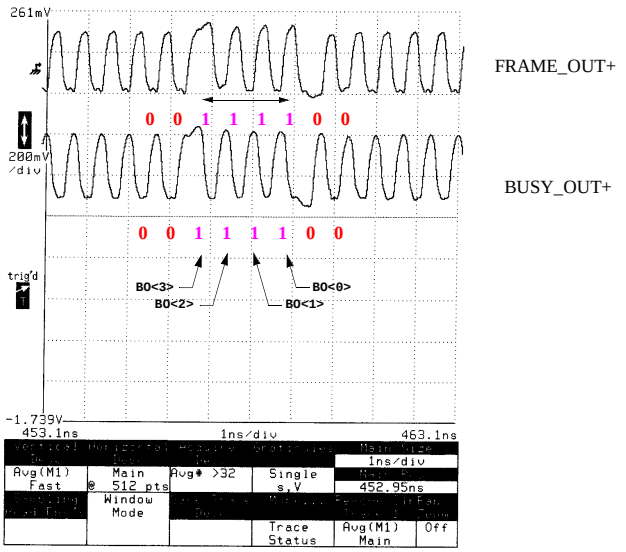


FIGURE TMSC.8: BUSY BIT release example.

Clock frequency 1.5 GHz (3.0 Gb/s, 2.5 V)

- The input FRAME is 4-bit wide.
- All the input BUSY BITs are one, indicating that all the BUSY BITs are grabbed (busy).
- As expected, since the grab register was reset before the second FRAME, even though this node is supposed to release BUSY BIT <0>, none of the BUSY BITs are released.

Discussion

The BUSY BITcircuitry implemented in the TSMC 0.25 µm process is similar to that in TE02 except for the corrections made to the current-steering RS latches. At a 3.0 Gb/s signaling rate, the BUSY BIT circuitry consumed 85 mA for 2.5 V rail and 95 mA for a 2.7 V power rail (16 % reduction in power).

9.0 Scaling issues for TE02 (preliminary TE03 design considerations)

Selection of the Switch IC data path architecture

Figure 37 shows the TE03 data path architecture resulting from a straight scale-up of the TE02 floor plan. Four 64-bit registers are required to hold the 1:64 demux control lines. Each optical path connects to a single 2x2 switch for each electrical port, but each electrical port connects to 64 2x2 switches. To reduce the load of the electrical ports, and to share vertical control line routing space, the 64-bit register can be partitioned into sections and can be distributed vertically, as shown in Figure 38 for the case of four 1:16 demux blocks.

If the Race-2000 4x4 crossbar is used to accommodate a rigid partitioning of the electrical paths to optical ports, large reductions in chip area and delay can be achieved. Figure 39 shows one example where each electrical port has connectivity to only 16 optical paths. Locating the electrical ports on both the top and bottom of the chip are particularly suited to this example.

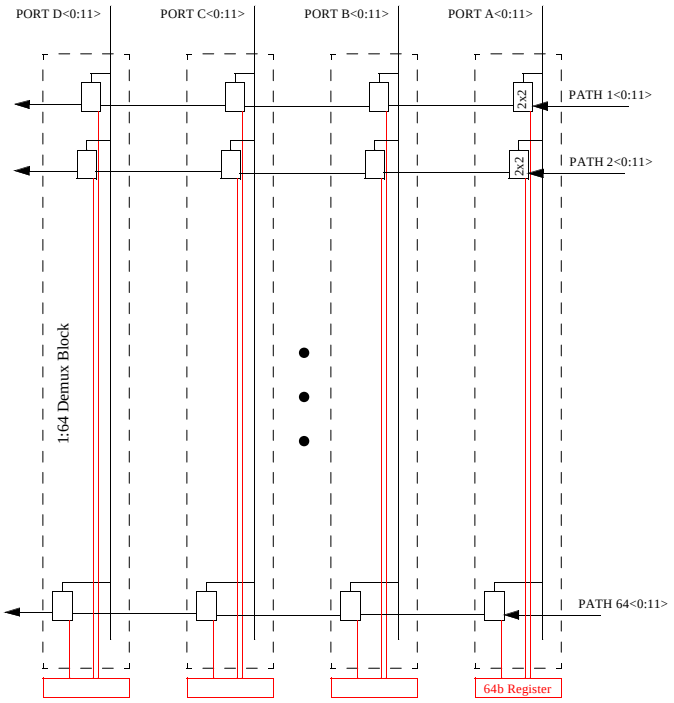


FIGURE 37: Straight scale-up of the TE02 to support 64 optical paths, and 4 electrical ports

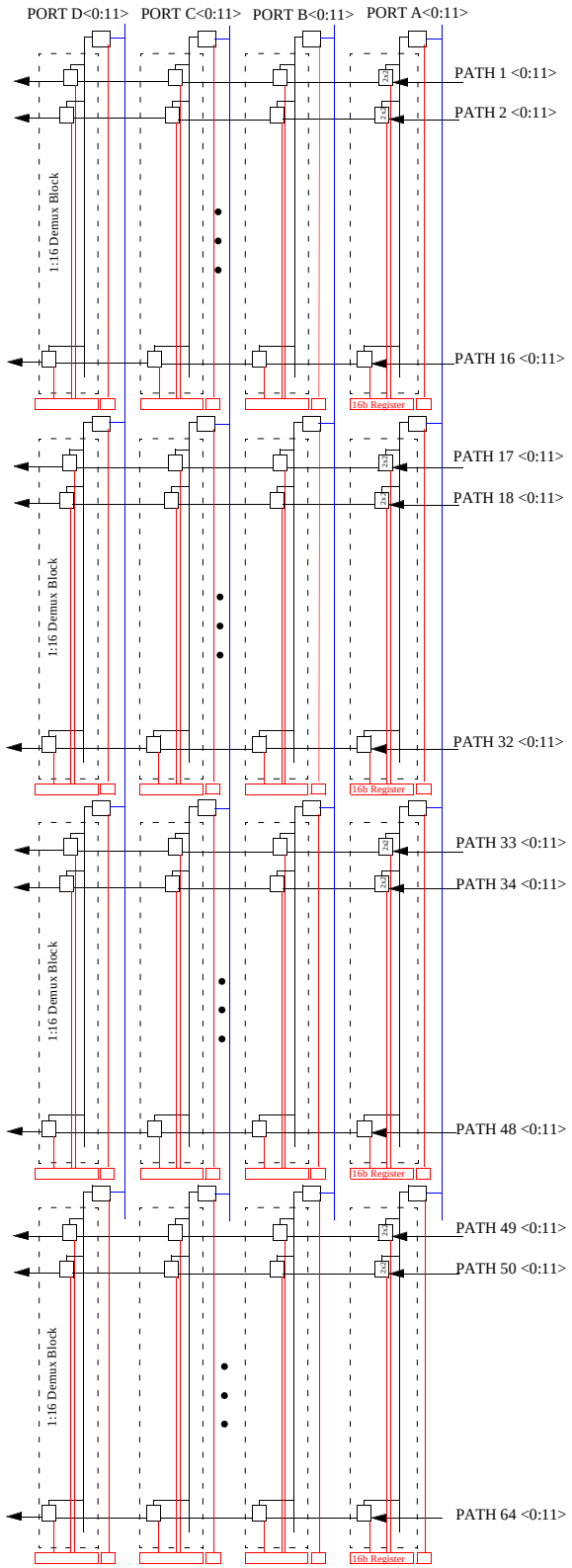


FIGURE 38: Scale-up of the TE02 using partitioning to support 64 optical paths, and 4 electrical ports

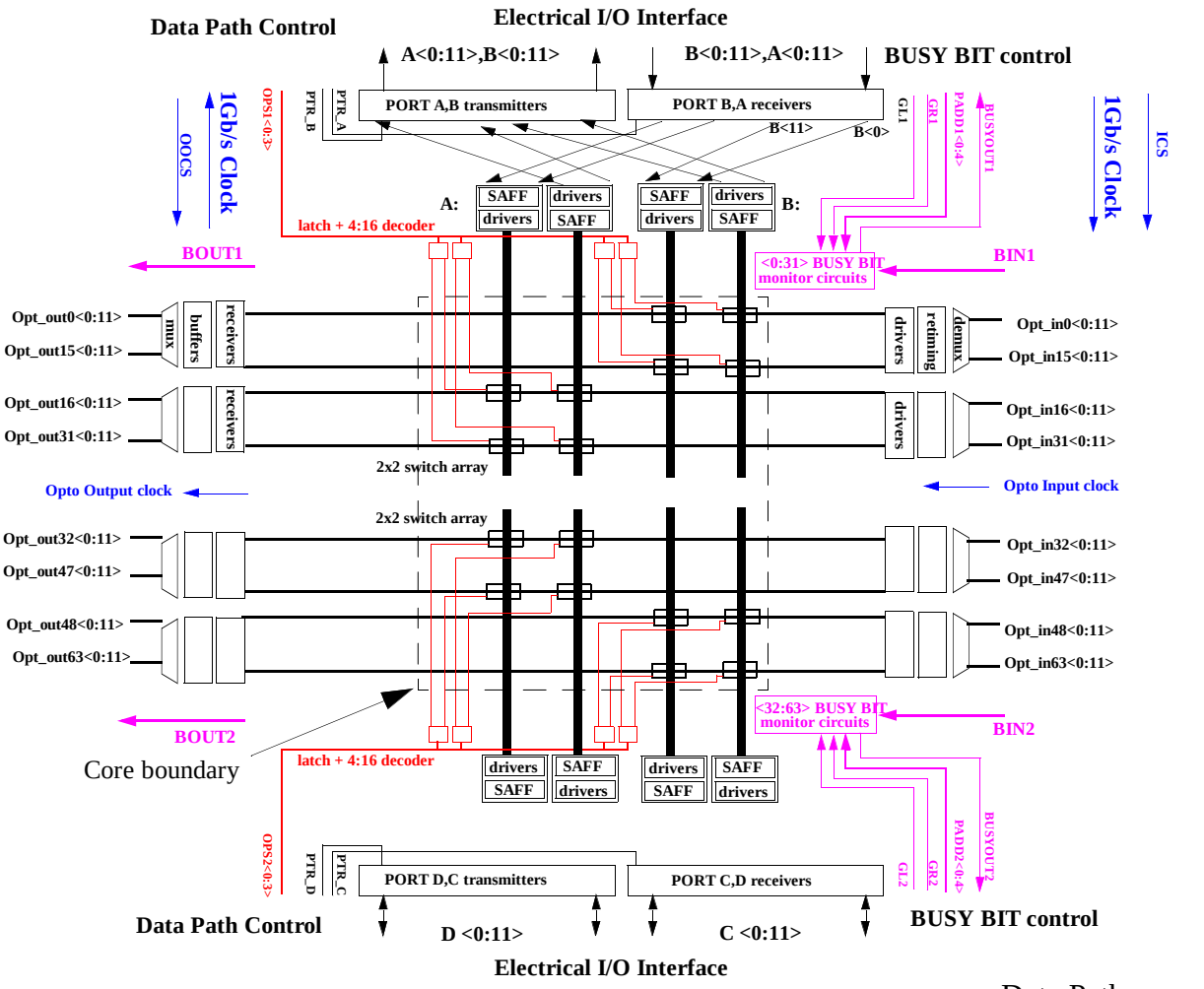


FIGURE 39: Example of a rigid partitioning scheme

Selection of the Switch IC clocking strategy

The recommended clocking strategy for the TE03 is the global clock strategy referred to as “globally clocked synchronous switching (GCSS)” in the section “Clocking and data synchronization” on page 24 and in Appendix A, “Selection of the Switch IC’s clocking strategy” on page 79. To summarize, the advantages of GCSS are

- the reduction of the number of synchronization boundaries to only one boundary per IC, so that the Switch IC would require only two phase-locked loops or delay elements
- the avoidance of the extra power, area and noise required to switch per-channel individual full-speed clocks

10.0 References

1. S. K. Bhogavilli, H. Abu-Amara, "Design and Analysis of High Performance Multistage Interconnection Networks," IEEE Trans. Computers Vol. 46, No. 1, pp. 110-117, Jan. 1997.
2. D. Nassimi, S. Sahni, "Parallel Algorithms to Set Up the Benes Permutation Network," IEEE Trans. Computers Vol. C-31, No. 2, pp. 148-154, Feb. 1982.
3. D. Nassimi, S. Sahni, "A self-routing Benes network and parallel permutation algorithms," IEEE Trans. Computers Vol. C-30, No. 5, pp.148-154, May 1981.
4. M. Matsui, et. al. "200MHz Video Compression Macrocells Using Low-Swing Differential Logic," Proceedings of ISSCC, pp. 118-119, 1994.
5. Kai Hwang, "Advance Computer Architecture, Parallelism Scalability Programmability," McGraw Hill, New York, 1993
6. D.H.Lawrie, "Access and Alignment of Data in a Array Processor," IEEE Trans. Computers Vol. C-24, No. 12, pp. 1145-55, Dec. 1975.
7. C.L. Wu, T.Y. Feng," On a Class of Multistage Interconnection Networks," IEEE Trans. Computers Vol. C-29, No. 8, pp 694-702, Aug. 1980.
8. Y. Taur, et. al. "CMOS scaling into the 21st century: 0.1 μm and beyond," IBM J. Res. Development Vol. 39, No. 12, pp. 245-258, Jan./March 1995.

11.0 Glossary of Terms

• BOIC	Buffered optical input clock
• CMOS	Complementary metal oxide semiconductor
• Crossbar Circuit:	Includes the crossbar switch, mux/demux, clock tree, and control logic
• Crossbar Switch	Includes the switch core, the dynamic drivers, and the SAFFs. When one includes all three blocks, the logic scheme of the switch is defined as low swing differential pass transistor logic.
• CSL	Current steering logic
• CSL-SA	Current steering logic sense amp
• CSL-SR	Current steering logic set reset flip-flop
• CSL-SAFF	Current steering logic sense-amp flip-flop
• CSWL	Current switch logic
• DEMUX	De-multiplexer
• EIC	Electrical input clock
• FF	Flip-Flop
• GCSS	Global clocked synchronous switching
• HCI	Host control interface
• I/O	Input/Output
• LCAS	Local clocked asynchronous signaling
• LVDS	Low voltage differential swing
• Reference Switch IC	The Switch IC connected to the reference processor board
• MIN	Multistage interconnect network
• MUX	Multiplexer
• Network Reference:	Reference controller or processor for the total network
• NG Formatter	Northrop Grumman Formatter chip
• NOR	Negation of OR
• OIC	Optical input clock
• OOC	Optical output clock
• Path	At the optoelectronic interfaces, and within the Switch IC, a group of data signal lines assigned to a processor is called a path
• Port	At the electrical I/O interface a group of data signal lines assigned to a processor is called a port.
• SA	Sense-amp
• SAFF	Sense-amp flip-flop
• SRFF	Set reset flip-flop
• Switch core:	The five deep pass transistor block
• TSPC	True single phase clocking
• UAS	Unclocked asynchronous switching

Appendix: A

Selection of the Switch IC's clocking strategy

There are three competing clocking strategies: Unclocked asynchronous switching (UAS), Locally clocked asynchronous switching (LCAS), and globally clocked synchronous switching (GCSS).

A.I Unclocked asynchronous switching (UAS)

A complete asynchronous switch may be built as a single stage unclocked crossbar network or as an unclocked multistage interconnection network (MIN). Clock routing, clock buffering and retiming is unnecessary in an unclocked asynchronous network, and any incoming clocks from the processor boards are considered as data. The responsibility of clock and data recovery is handed down to the destination. Since the switching is independent of the individual clocks of different processor boards, the switch can effectively connect processor boards of different clock frequencies.

Even though the data skew of an unretimed single stage may be a few tens of pico seconds, in the case of a multiple stage switch, the total skew through the switch can be hundreds of pico seconds.

Due to the unavailability of a clock, the switch (periphery as well as switch core) needs to be implemented using static logic, differential logic or some other unclocked logic scheme. Out of the competing logic schemes, static logic may be the most area efficient and the fastest. But due to the noise generated by static logic, an optoelectronic switch using static logic will not scale-up efficiently. At the same time, differential and current steering logic are also unattractive candidates for large switch cores due to their larger circuit area requirements. Hence, unclocked asynchronous switching should be avoided in large switch networks (e.g. 64x64).

A.II Locally clocked asynchronous switching (LCAS)

Similar to the unclocked asynchronous switch, here also data recovery and synchronization are handled by the destination. The clock is received with the data and is buffered and then used to demux, latch, and retime its own data. In parallel with the data the clock is routed to the output stage using a 1 Gb/s unclocked switch. At the output, each individual clock is again buffered and used to mux, latch and retime its own data.

The advantage of LCAS is that instead of deskewing 11*64 inputs to a global clock one only needs to skew 11 bits (one path) to its own clock. While retiming using a local clock helps to reduce the data skew, it causes an excessive number of clock trees in the switch. The number of clock trees is proportional to the number of input/output paths (or ports). Since there are now 64 different optoelectronic clocks the IC require 64 traditional receivers at the optoelectronic input interface. Also, in case of higher clock frequencies, the delay of a clock tree (including clock receive and buffer circuitry) can be greater or equal to the clock period. In such cases individual delay elements may be needed at the optoelectronic input interface, and the number of synchronization boundaries at the optoelectronic interface will increase to 64.

Appendix: B

Clock initialization

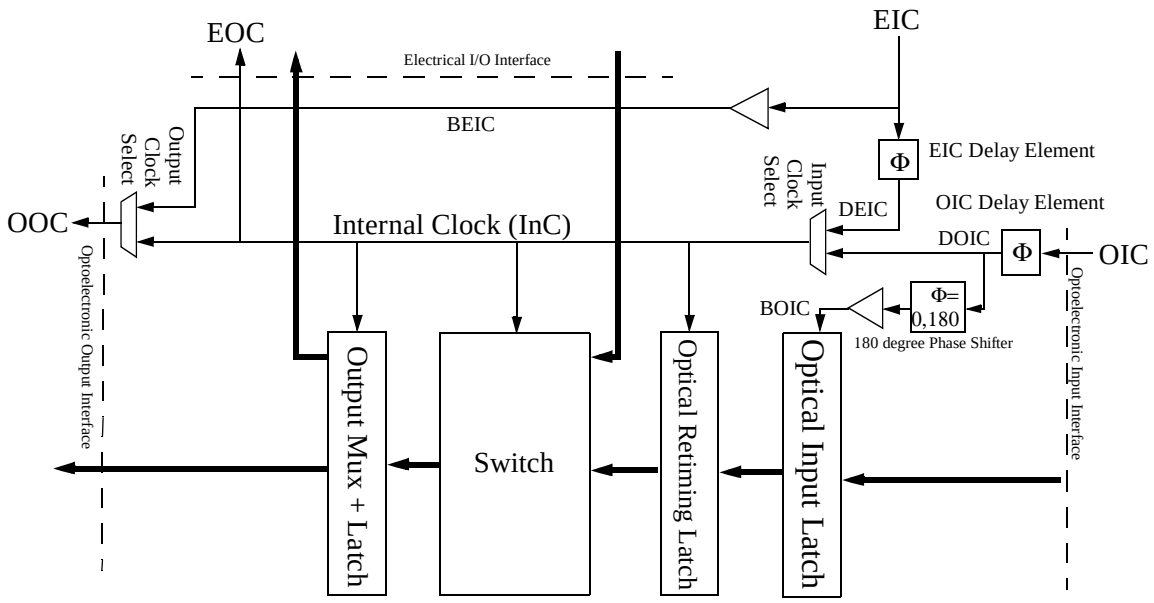


FIGURE 40: Abstract diagram of the Switch IC clocking domains. OIC = Optical Input Clock, EIC = Electrical Input Clock, EOC = Electrical Output Clock, OOC = Optical Output Clock, DOIC = Delayed Optical Input Clock, DEIC = Delayed Electrical Input Clock, BEIC = Buffered Electrical Input Clock, BOIC = Buffered Optical Input Clock.

Need for data synchronization

Despite the low timing skew (200 ps) between the data and clock delivered at each of the input interfaces, the total delay through the clock tree is expected to be greater than one cycle, and will change with variation in process parameters by up to 20%. Also, the phase difference between each input clock and its own buffered version varies with frequency. This phase variation can cause the clock latching edge to coincide with the period when data is changing, failing to latch the data. A means to control this phase variation is by use of the variable-delay delay elements within the buffer chain (clock tree).

Need for clock domain synchronization

Even though the electrical and optical input clocks are frequency-matched, the reference Switch IC receives the optical clock with an unknown phase relationship to its internal clock. Because of this, an additional set of input latches is needed to retime the optical data to the internal clock.

Both of these required synchronizations are performed during the initialization procedure.

Initialization procedure

The clock initialization procedure uses the input clock select control, the output clock select control, the 180-degree phase shifter control, and sets the analog voltages controlling the electrical and optical delay elements, to achieve clock and data synchronization in a four-step procedure. To ease implementation and debug, we assume an electrical communication channel exists between the reference and slaves.

In the first 2 steps, the reference Switch IC provides the data used to synchronize the electrical input interface of the reference

Switch IC (step 1) as well as the slave ICs (step 2). In steps 3 and 4, the final slave ahead of the reference Switch IC supplies the data used to synchronize the optical input data to buffered optical input clock of the reference (step 3) as well as its buffered optical input clock to internal clock (step 4).

The source of the internal clock for the reference Switch IC is the electrical input clock, except in step 3, when the optical input clock is used. Even during this step, however, the reference Switch IC continues to provide the optical clock to the network derived from the electrical input clock. The source of the internal clock for all slave ICs is the optical input clock.

Step 1

First, the reference Switch IC selects the electrical input clock (EIC) as the source of its internal clock, and selects its internal clock to drive the optical output clock (OOC). It receives data from the electrical I/O interface PORT A and transmits on PORT B (Figure 41.(a)). The reference host adjusts the reference Switch IC's EIC delay element analog control until the proper data is transmitted by the Switch IC at the electrical I/O interface. After the reference is synchronized, it receives data on PORT A and transmits on optical PATH 0 (Figure 41(b)).

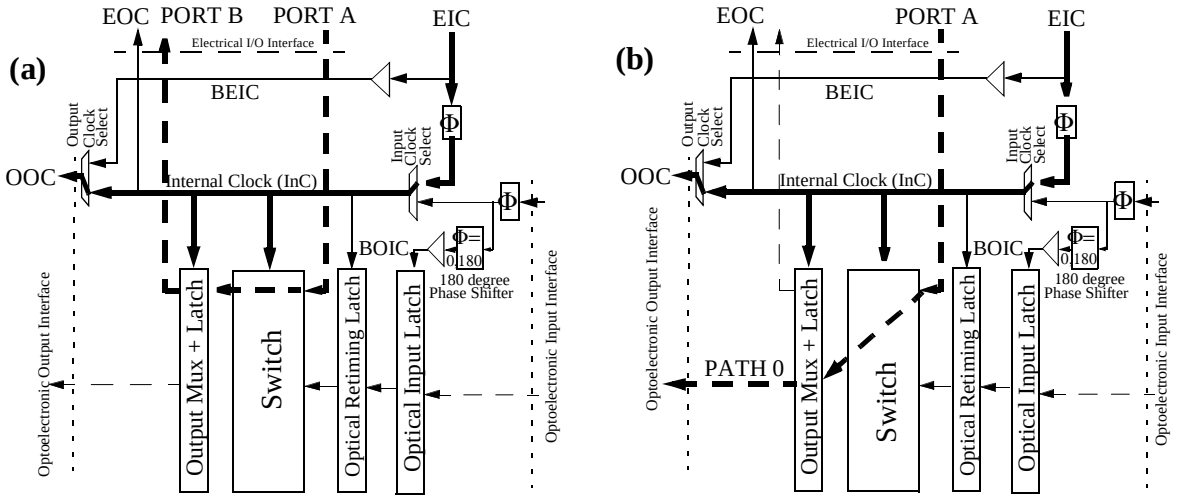


FIGURE 41: Clock and data flow (step 1)

Step 2

Next, each slave Switch IC has the optical input clock (OIC) selected as the source of its internal clock, which is also transmitted as its optical output clock (OOC). The first slave IC receives the optical data on PATH 0 and transmits on PORT B, while the slave host adjusts the OIC delay element until the proper data is transmitted (Figure 42). After each slave is synchronized, it receives data on PORT A and transmits on optical PATH 0. Each subsequent slave IC is likewise adjusted.

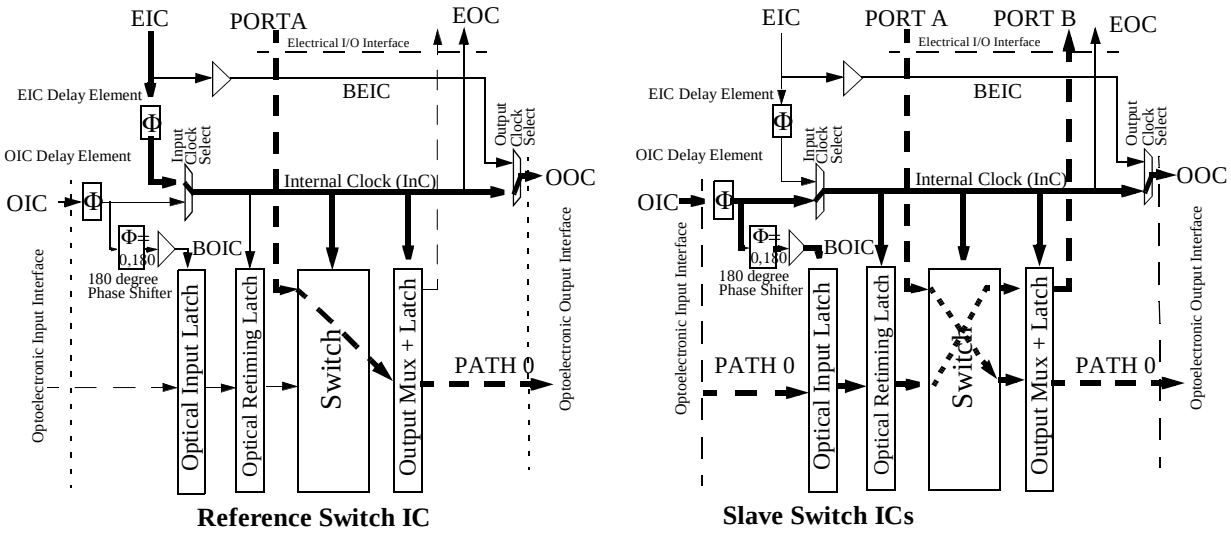


FIGURE 42: Clock and data flow (step 2)

Step 3

The electrical communication channel informs the reference that the final slave is up. Next, the reference Switch IC selects the optical input clock (OIC) as the source of its internal clock, but continues to transmit the electrical input clock (EIC) as the optical output clock, by changing the output clock select control. It receives data from optical PATH 0 and transmits on PORT B (Figure 43). The reference host adjusts the reference Switch IC's OIC delay element analog control until the proper data is transmitted.

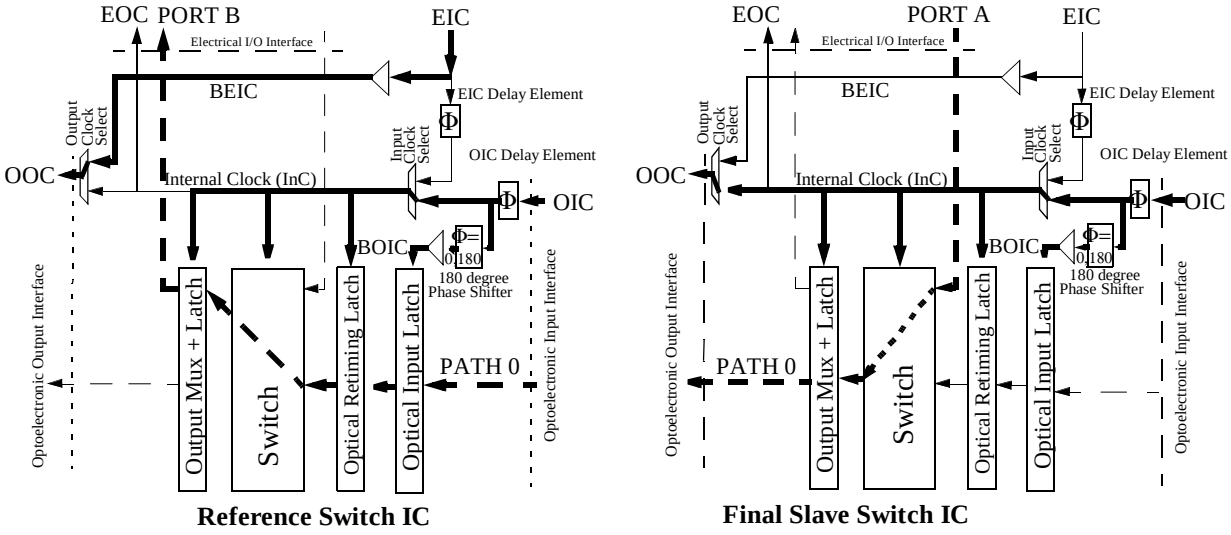


FIGURE 43: Clock and data flow (step 3)

Step 4

Finally, the reference Switch IC switches the source of its internal clock back to the electrical input clock (Figure 45). After this clock source change, the received data may be corrupted. This may happen if the phase difference between the buffered optical input clock and the internal clock is such that the output of the optical input latch is invalid at the internal clock latching edge, as shown in case (b) of Figure 44.

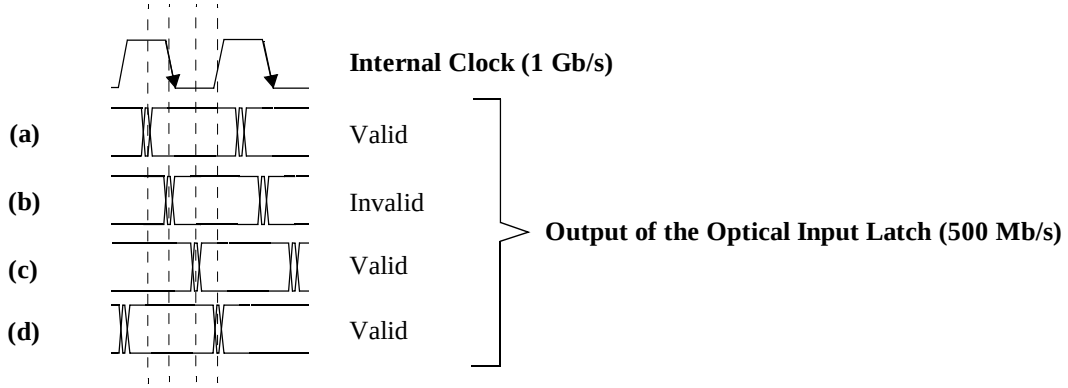


FIGURE 44: Possible phase relationships between the internal clock and optical input latch data

In this case, the reference host can enable the 180-degree phase-shift select control so that the output of the optical input latch is delayed by 1 phase to achieve the relationship shown in case (d).

If the original relationship of the internal clock and the optical input latch data is one of the other cases (a,c,d), then proper data will be received without requiring a phase-shift. However, for cases (a) and (c), enabling a phase-shift may increase the tolerance of the latching circuitry to delay variations in clock and data. This could be determined by comparing the amount of EIC delay element variation tolerated by the latching circuitry with and without a 180-degree phase-shift.

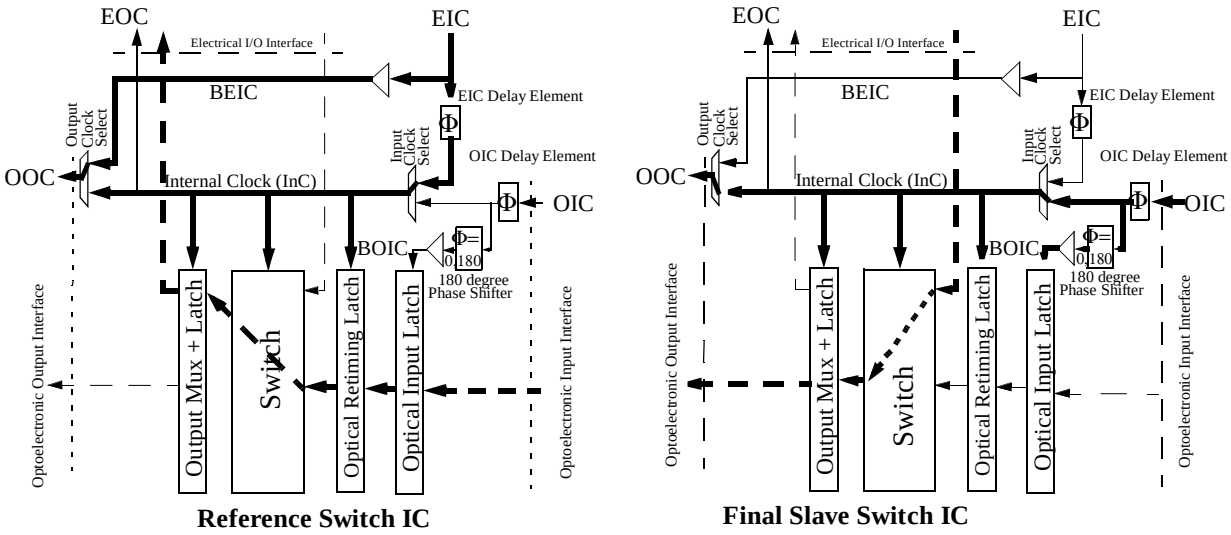
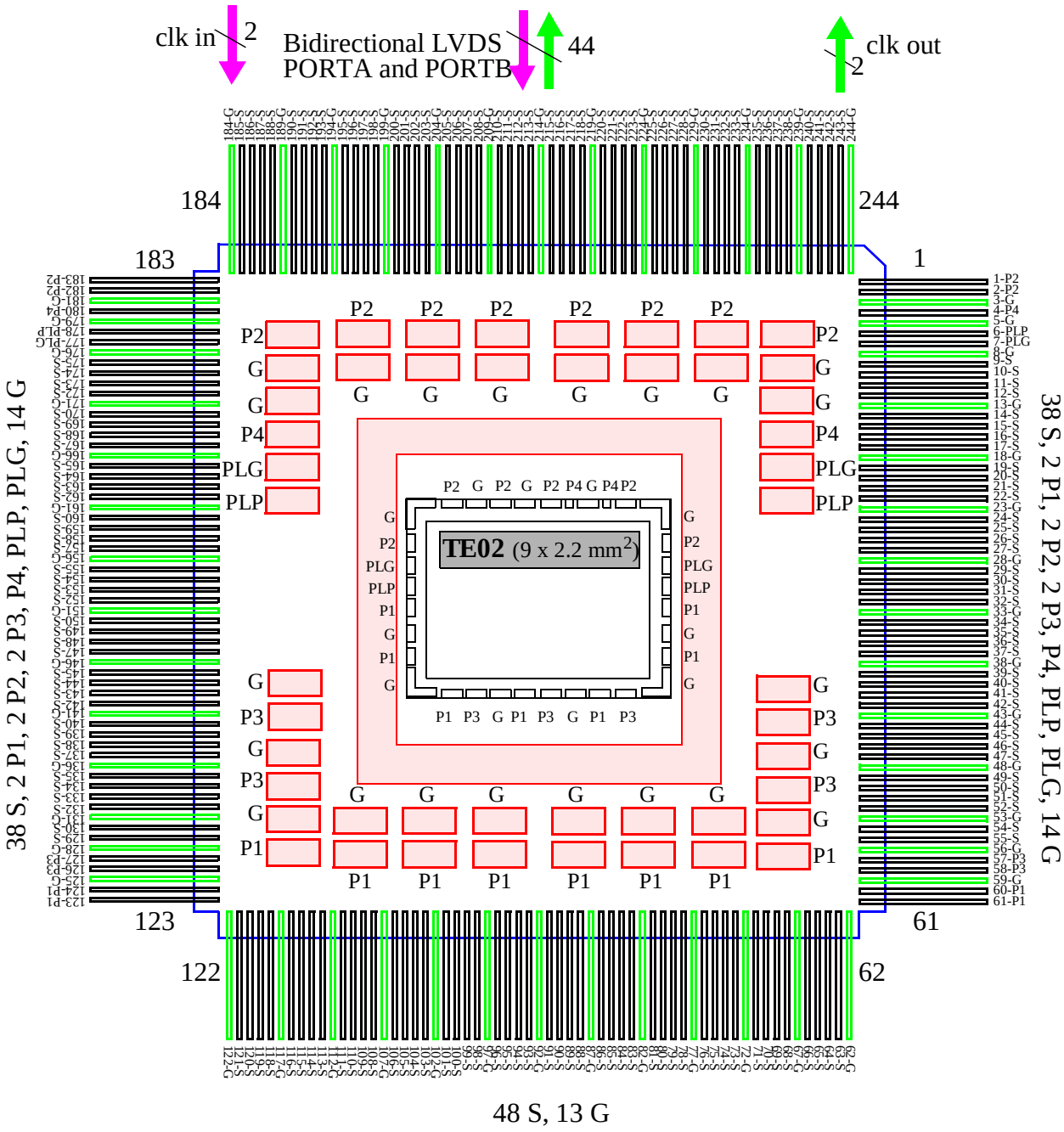


FIGURE 45: Clock and data flow (step 4)

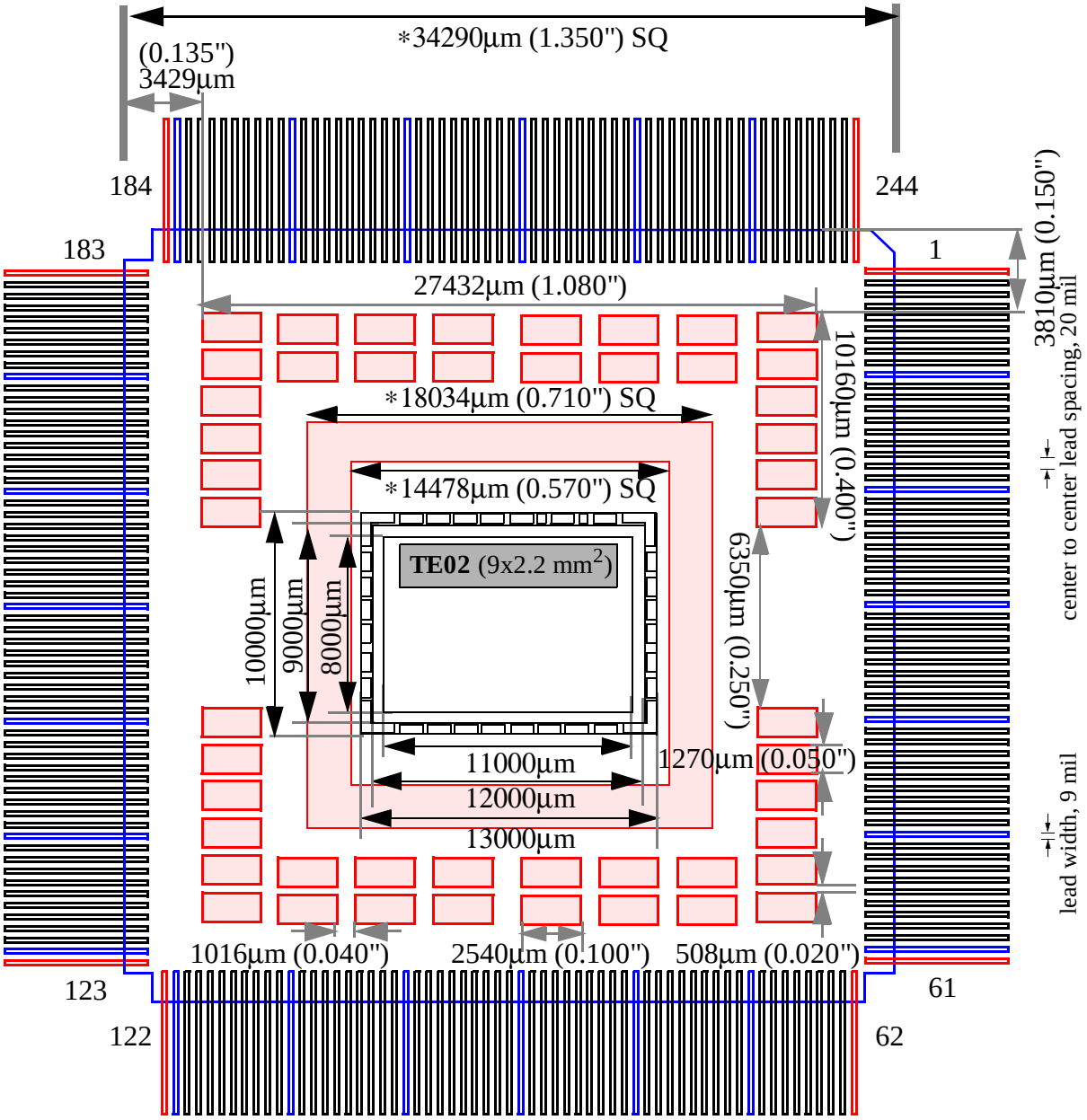
Appendix: C

Further details of QFP:

Cavity up view lead assignments



CAVITY UP VIEW



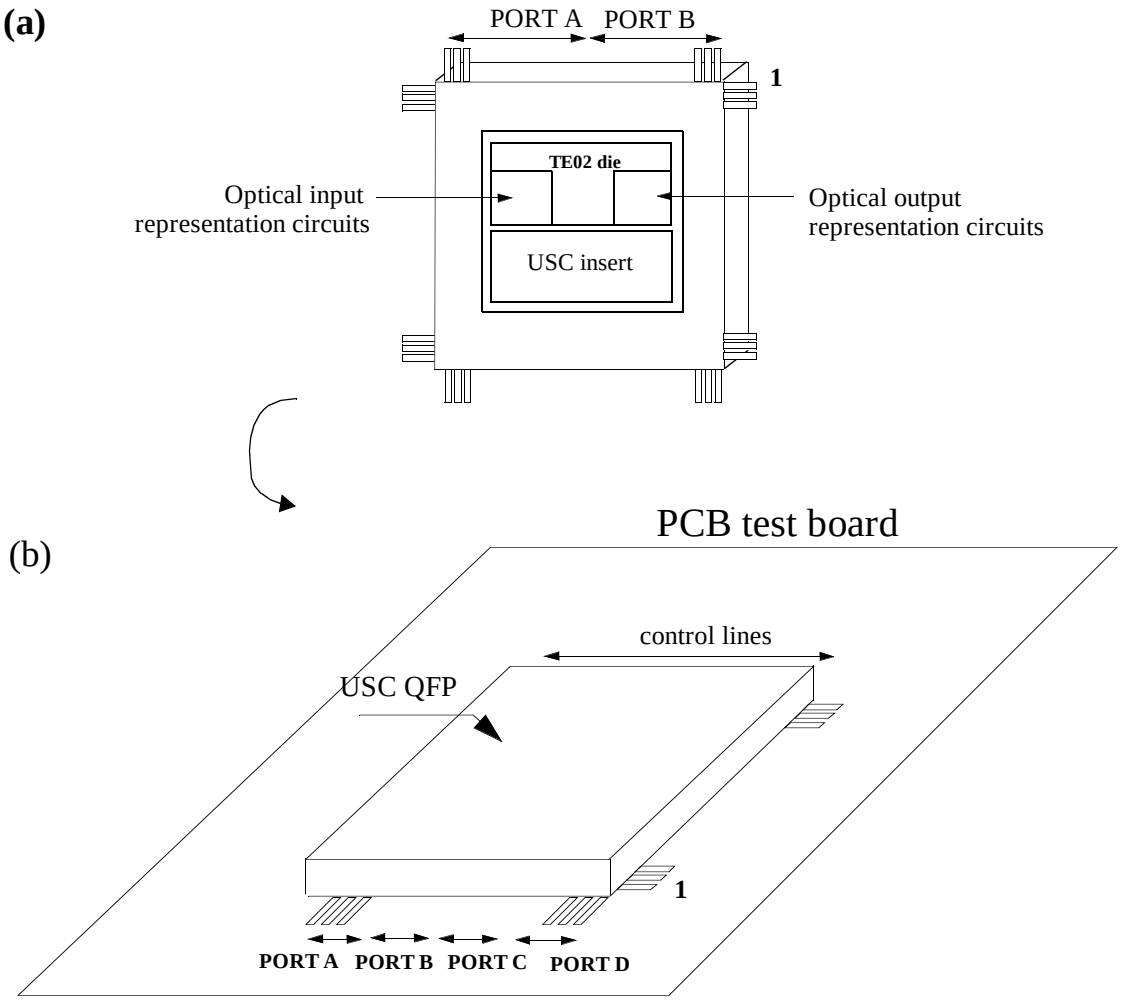


FIGURE 46: (a) Cavity up view. (b) Cavity down view on PCB