Switch Concept & Architecture

What types of architecture makes sense ?

How to reduce the effect of blocking ?

Example architectures

Switching Principles

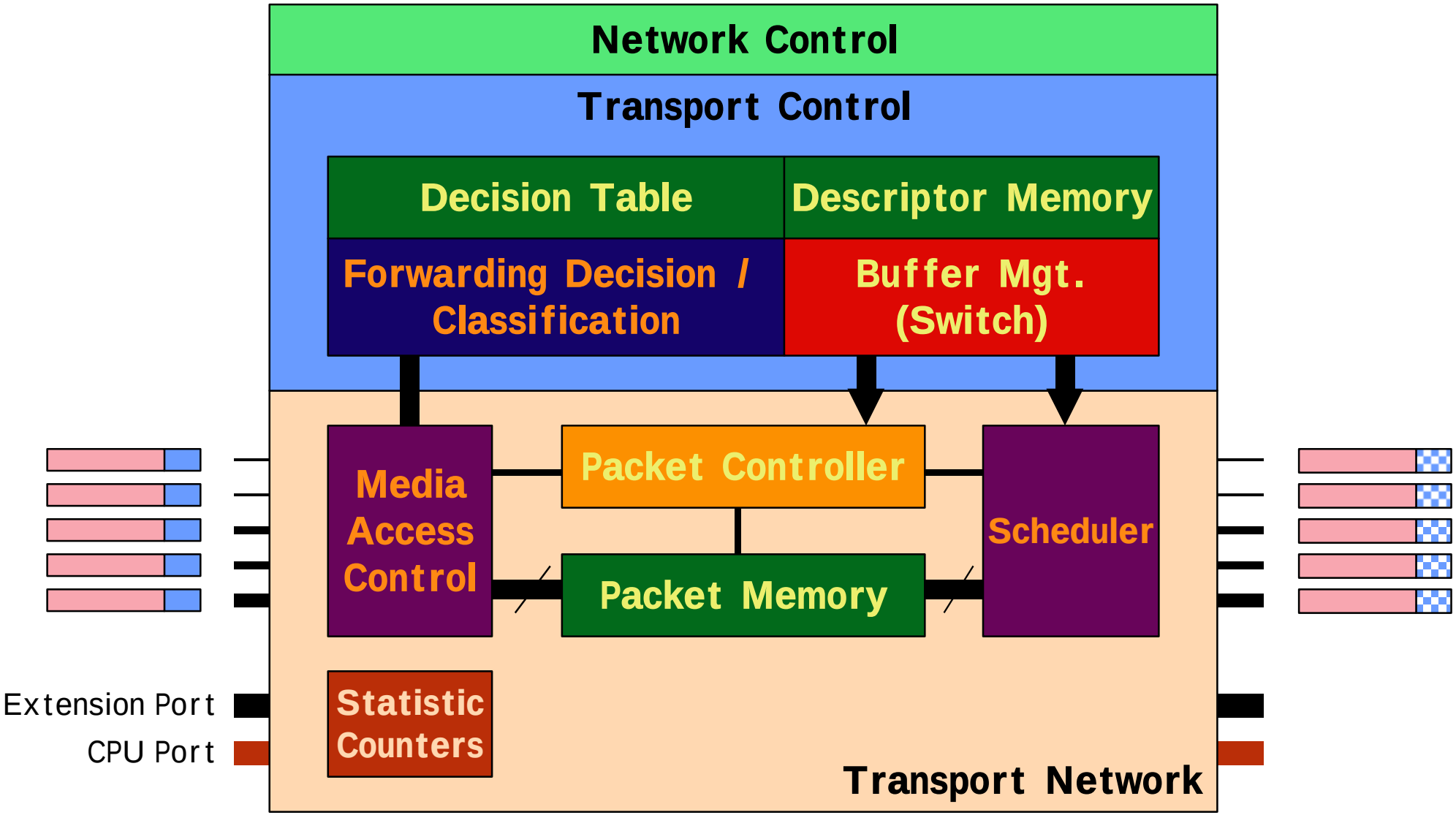
- Switching is the transport of information/data from an *incoming logical channel* to an *outgoing logical channel*.
- Logical channels are characterized by
  - » physical inlet / outlet identified by a physical port number
  - » logical channel on the physical port identified by
    - – ATM: VPI / VPI
    - – L2: DA / SA
    - – L3: IP Address
    - – Multi-layer: Flow, TCP port, etc.
- Switch operations
  - » cells / packets arrive at the input port
  - » header lookup, switching decision and translation is performed
  - » packets are routed through or queued in the switch to the appropriate output port.

Switching Principles

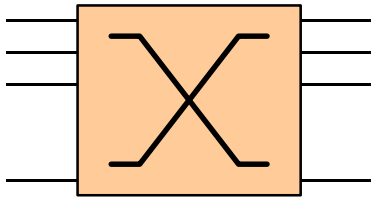
Switches have three main part.

- Transport Network:
  - » Physical means for switching cells from one input port to one (or more) output port(s) in the switch
  - » Performs functions of the User Plane
- Transport Control
  - » Controls the transport network based on analysis of signalling information
    - – decides which inlet connects to which outlet via routing, switching, etc. …
- Network Control
  - » Sets *transport control* parameters / tables.
    - – Routing table, resource mgt, call admission control (CAC), etc.
  - » Performs functions in the Control Plane
    - – Routing, network mgt., signalling, signalling AAL (SAAL), etc.

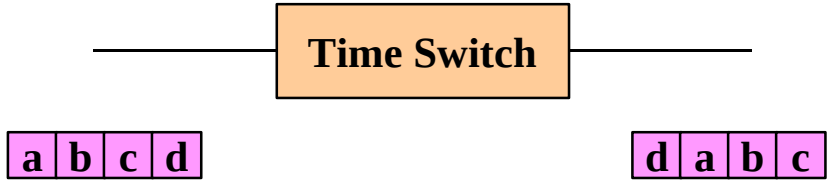
Switching Components

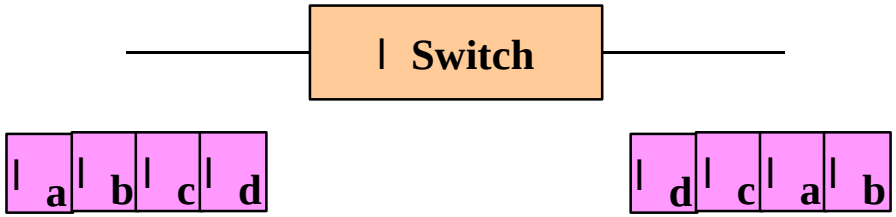


Classical Switching Systems

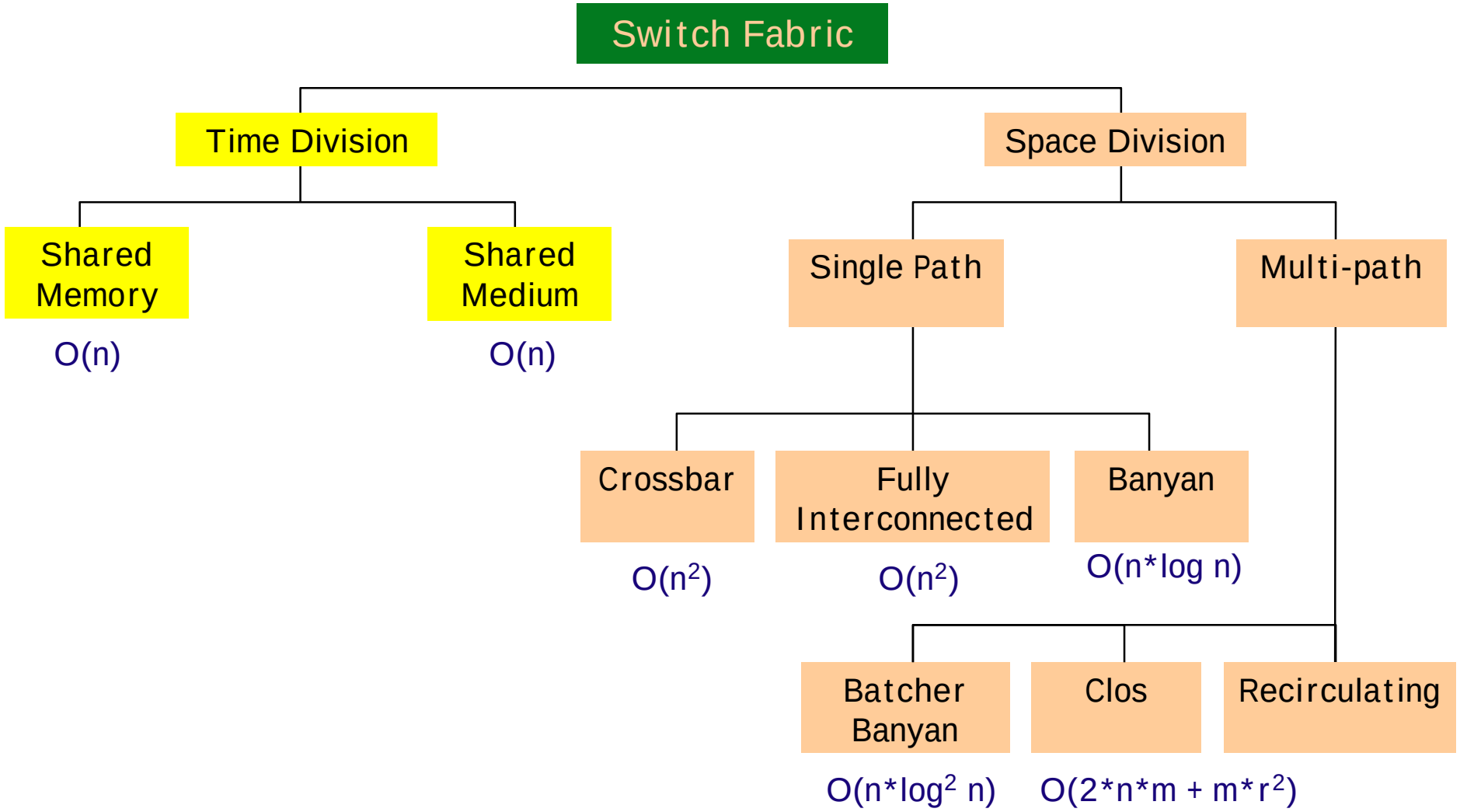


Space Switch

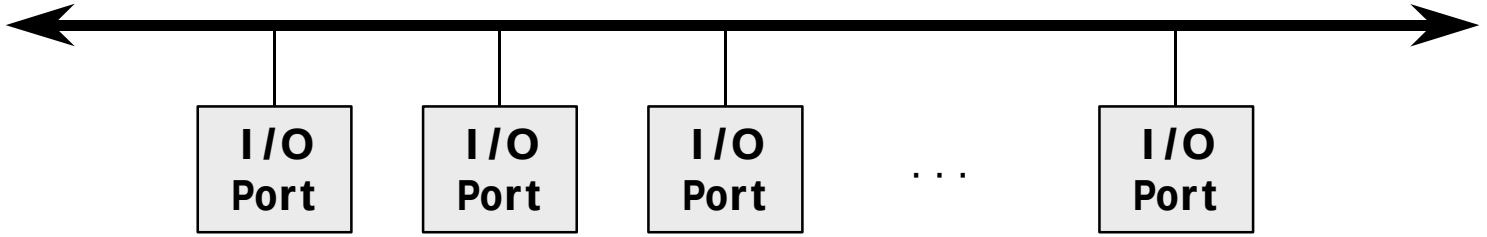




Switching Fabric Classification

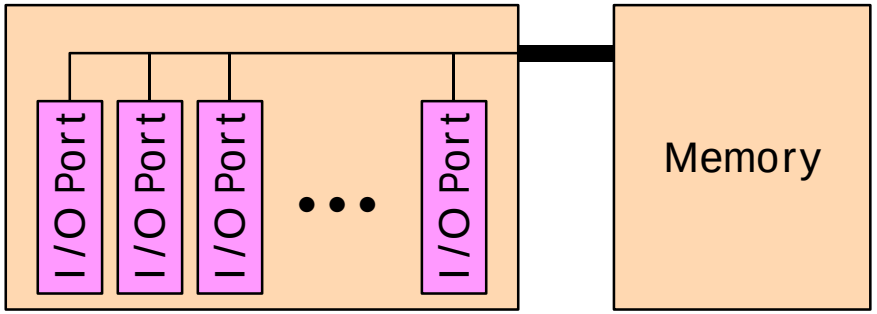


Backplane Switching System



- The Bus is used to connect the inlet / outlet ports together.
  - » Cells are transported via the *Bus.*
  - » Outlet ports accepts cells/packets based on their address
- Non-blocking bus: Σ [Port speed] < Bus throughput rate
  - » Blocking bus effect can be reduced via statistical muxing effect.
- Advantages:
  - » Multicasting is easily supported
  - » Bandwidth limited to about 2 Gbps.
  - » Easy integration with existing LAN equipment based on backplane technology, such as hubs.

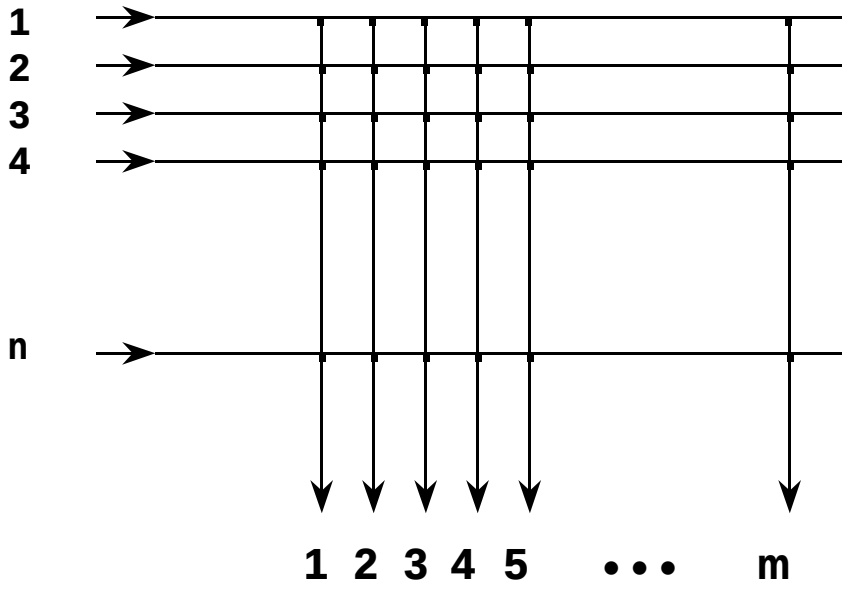
Shared Memory Switching System



- The Shared Memory is used to connect the inlet / outlet ports together.
  - » Cells are temporarily buffered
  - » Outlet ports pulls cells/packets based on their queue
- Non-blocking memory bandwidth required:
  - ≈ Σ [Port speed] < Memory throughput rate
- Advantages:
  - » Memory usage reduced via sharing
  - » Multicasting is easily supported
  - » Bandwidth limited by memory technology.

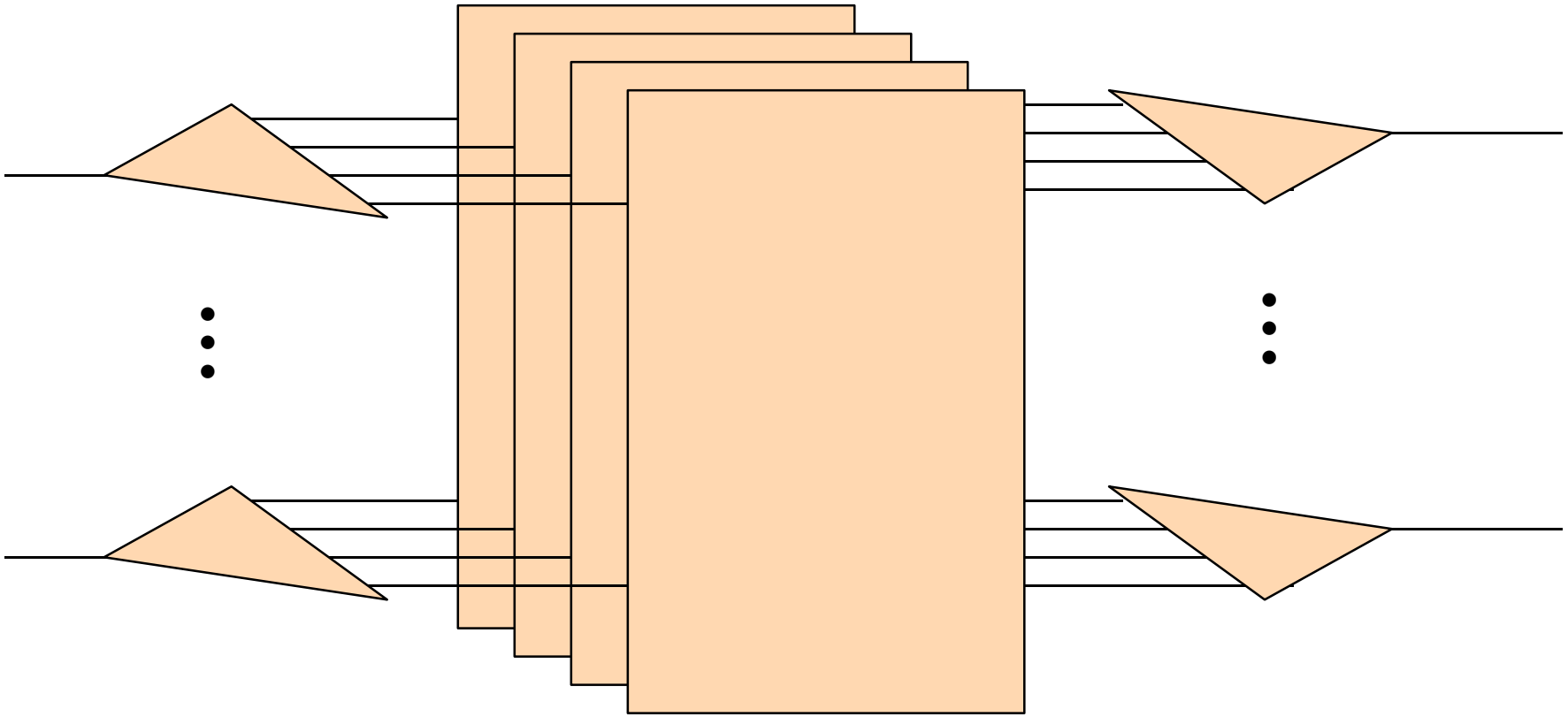
ITRI
CCL

Crossbar Switching System



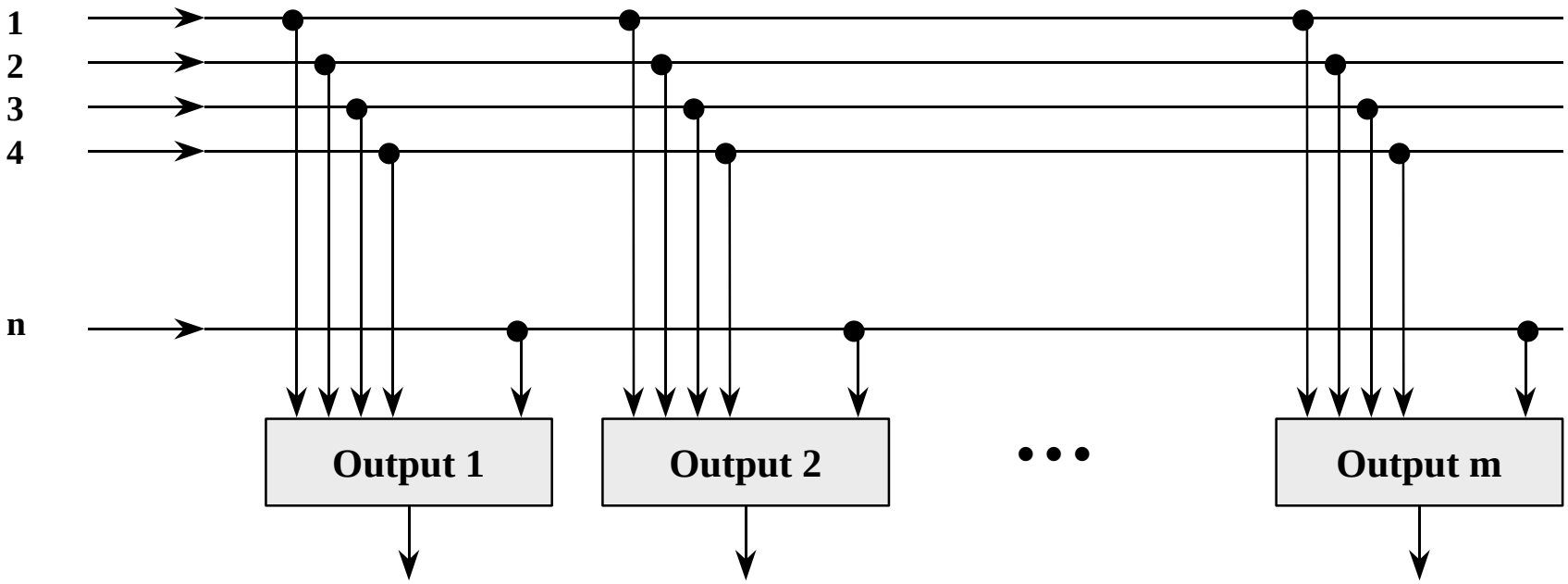
- Internally non-blocking:
  - » each I/O pair has an unique path.
  - » externally blocking at the output.
- Advantage:
  - » high speed internal paths
  - » multicast ready
  - » simple
- Disadvantage:
  - » difficult to scale
  - » input queueing
  - » switching elements, $O(n^2)$

Multiplane Network



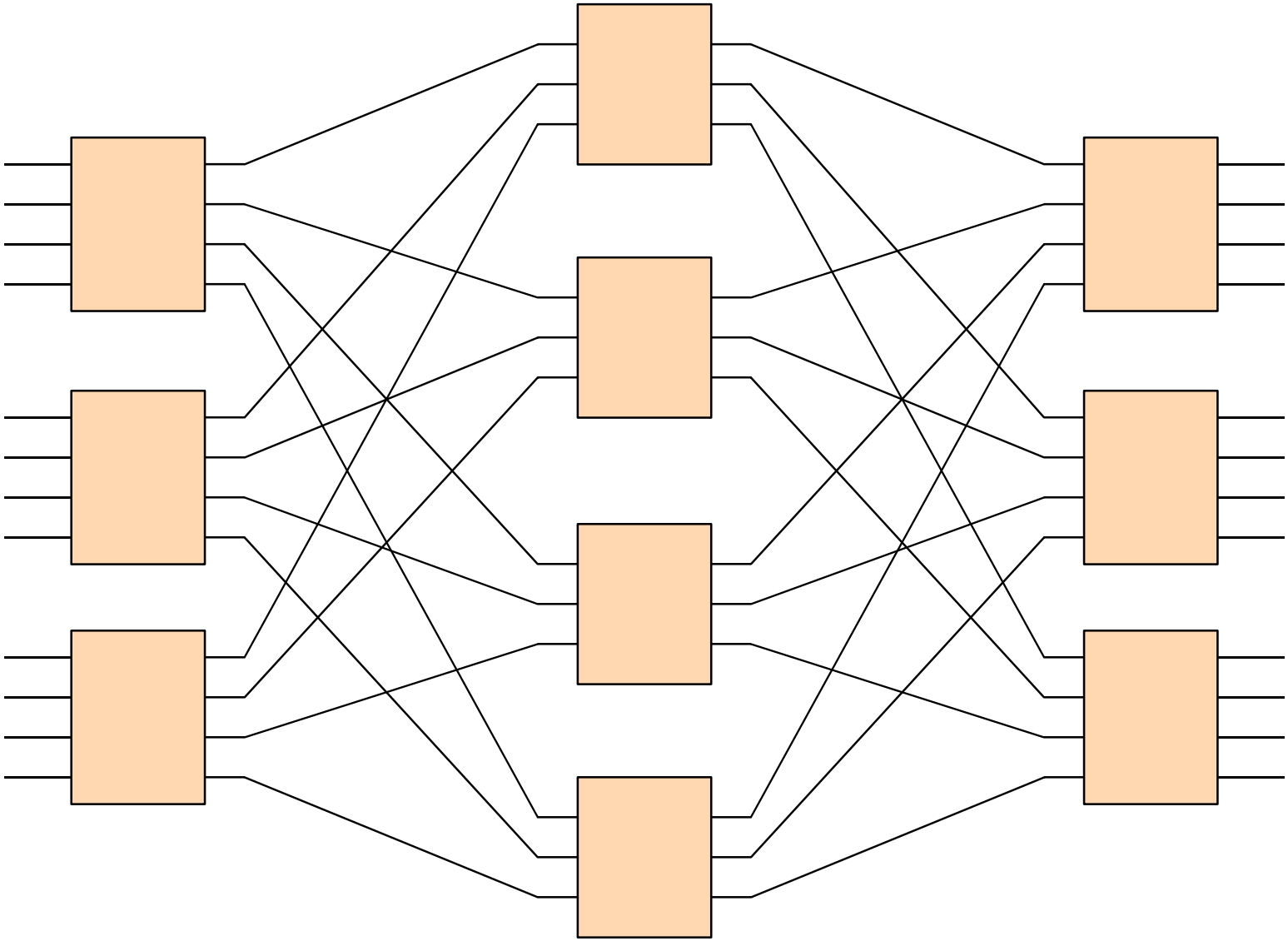
ITRI
CCL

Knockout Switching System



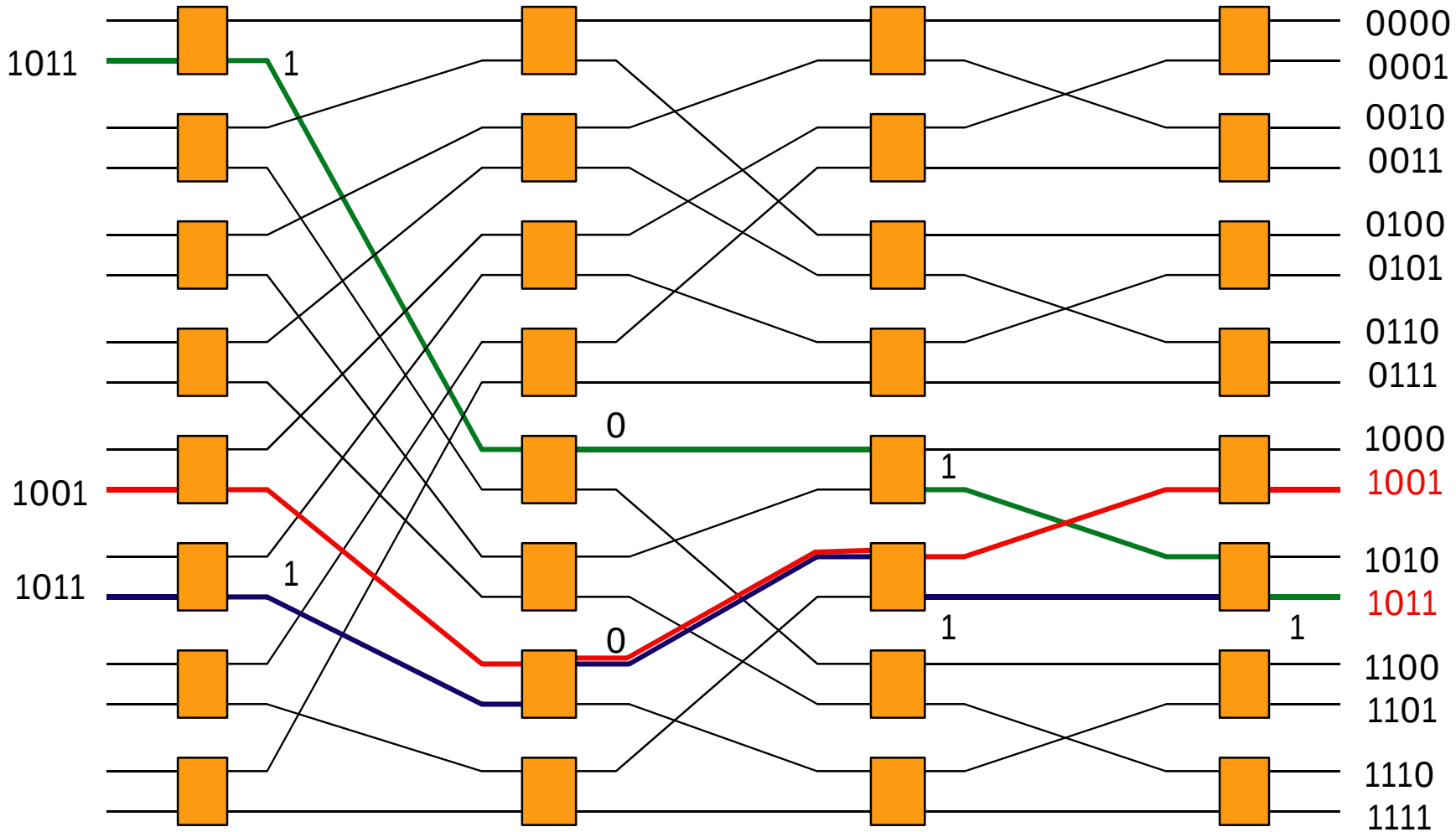
- Internally & externally non-blocking: *"m"* multiple paths between I/O pair.
- Each output port consists of a concentrator and a shared buffer.
- Advantage: high performance, multicast-ready
- Disadvantage: costly to implement, switch elements $O(n^2)$, difficult to scale

Clos Network



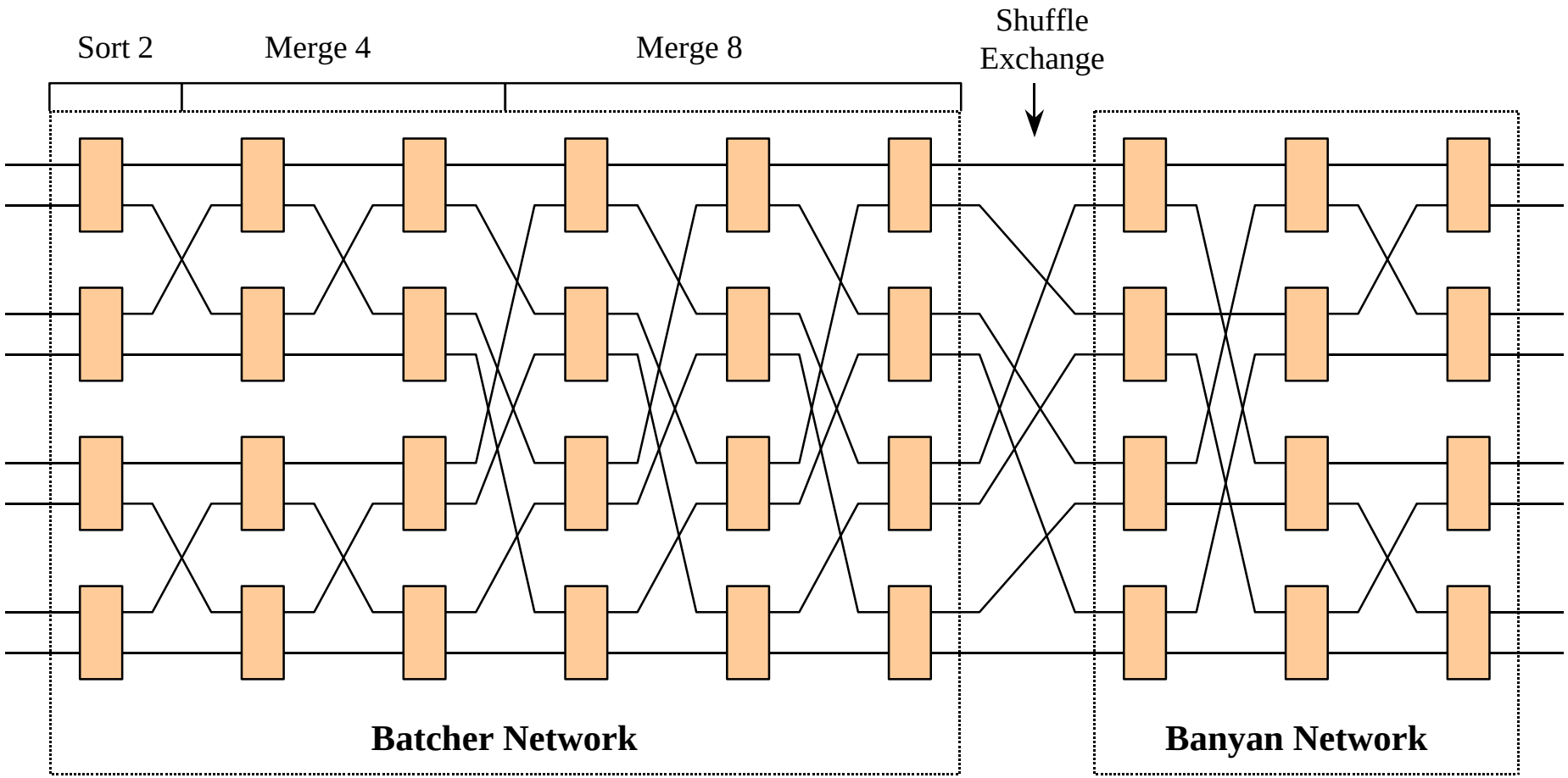
ITRI
CCL

Delta Network Switching System



ITRI
CCL

Batcher-Banyan Switching System



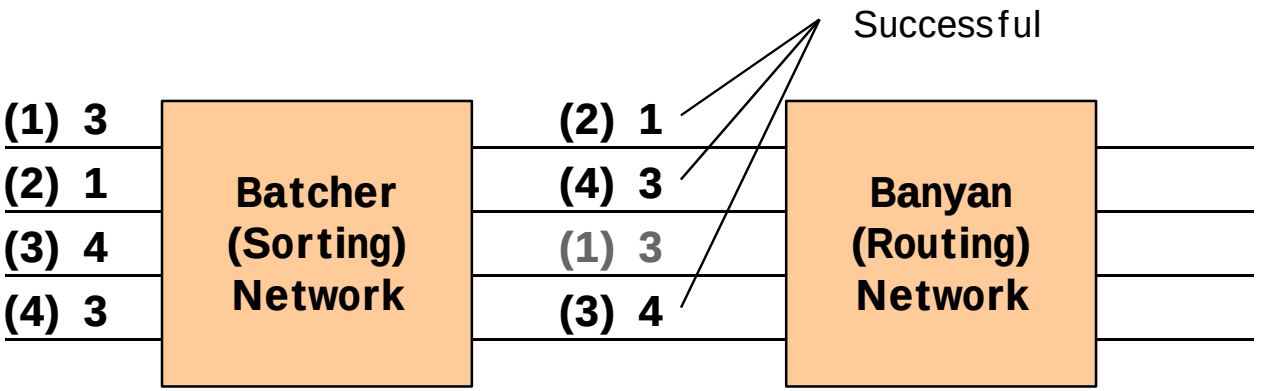
Batcher-Banyan Switching System

- Category:
  - Space switching
  - Internally non-blocking Multiple Interconnection Network (MIN)
  - No internal buffering required
  - Output contention still a possibility, can be solved by using
    - additional arbitration logic
    - input buffering
    - output buffering with fabric speed up.

Batcher-Banyan Switching System

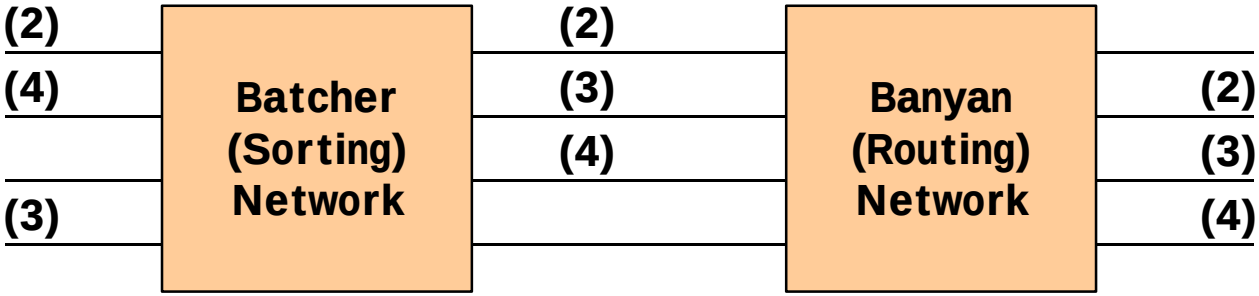
- Batcher Network = sorting network
  - » Sorts all cells according to destination
  - » Cells to same destination are adjacent.
- Banyan Network = self-routing network
  - » No internal blocking, if cells are sorted at inlet
  - » No external blocking, if only one cell for each outlet.
- Output port contention solved using 3 phase algorithm
  - » Arbitration phase
  - » Acknowledgement phase
  - » Sending phase

Batcher-Banyan Switching System



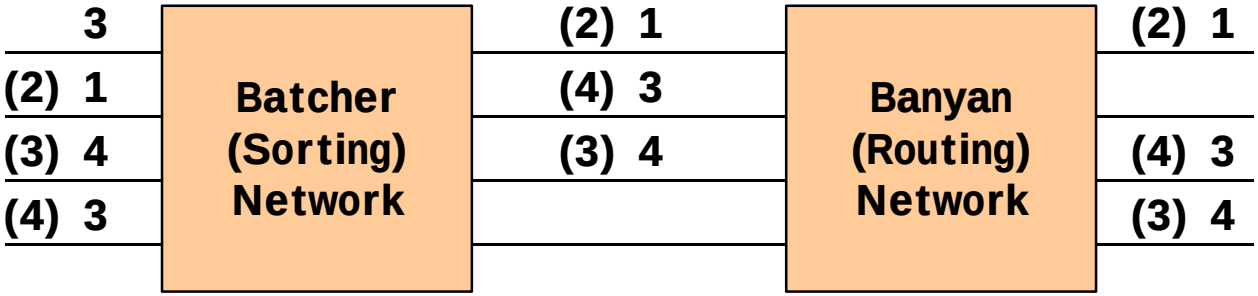
- Phase I: Send and Resolve Request
  - » Send source-destination pair through sorting network
  - » Sort destination in non-decreasing order
  - » Purge adjacent requests with same destination.

Batcher-Banyan Switching System



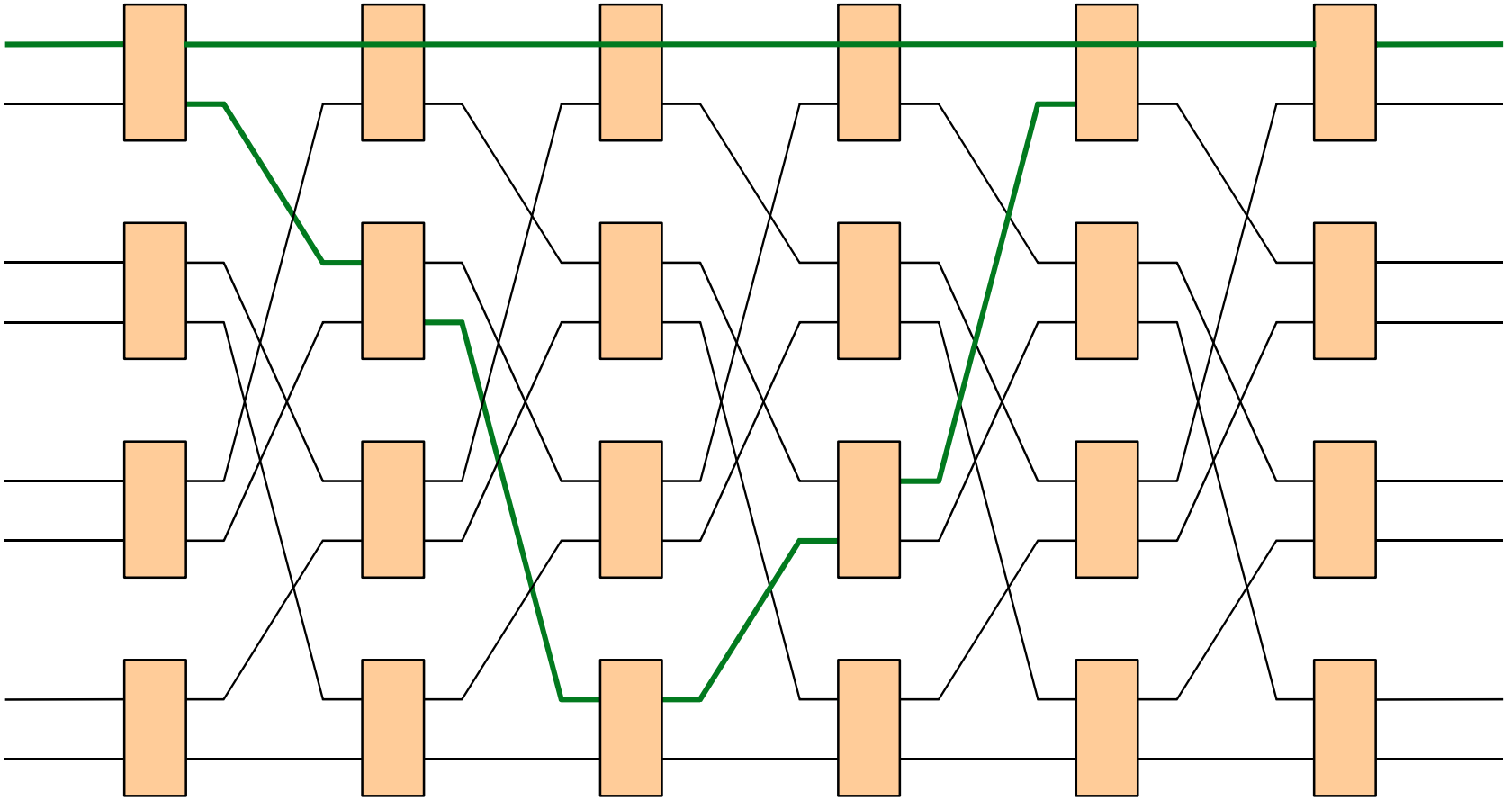
- Phase II: Acknowledge winning port
  - » Send ACK with destination to the port with winning contention
  - » Route ACK through Batcher-Banyan network

Batcher-Banyan Switching System

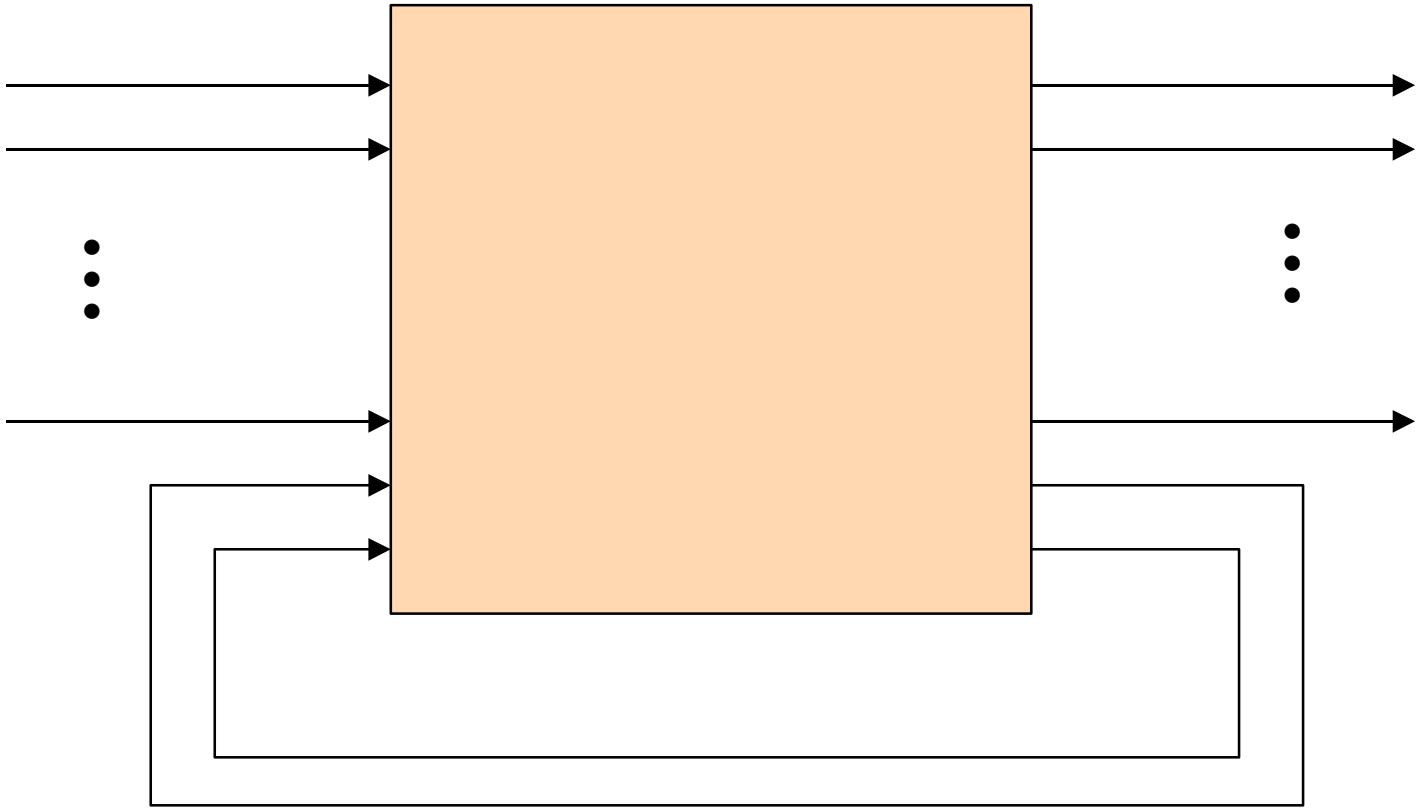


- Phase III: Send Packet
  - » Acknowledged ports send packet through Batcher Banyan network
  - » Buffers at port controller

Augmented Banyan



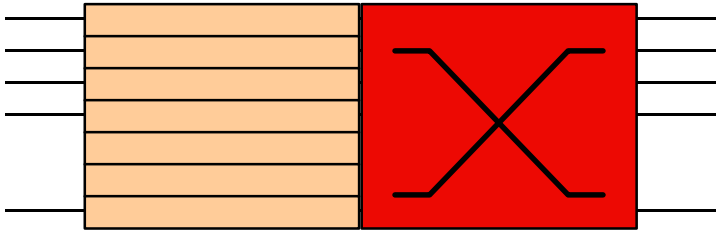
Recirculation Network



Blocking Effects in Switches

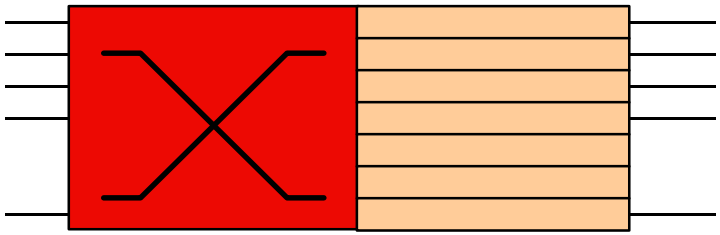
- *Blocking* is a fundamental problem of switching
  - » occurs when two or more cells contend for the same resources
    - – paths through the switching fabric (internal blocking)
    - – output ports (external blocking / output contention)
- *Queueing* is the solution to the blocking effect
  - » cells are temporarily stored in buffer memory until it can be safely transported to the intended output port.
  - » basic queueing techniques
    - – input, output, internal, shared buffering.

Input Buffering



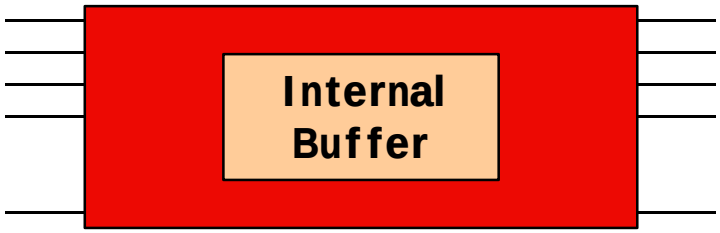
- All arriving cells are queued at the input
  - » cells are queued until the switch indicates that the cell has been successfully switched.
  - » cells that fail to reach the output port stay in the buffer and can be tried again on the next round of arbitration.
- Advantage: Simple to implement.
- Disadvantage: Head of Line (HOL) blocking.

Output Buffering



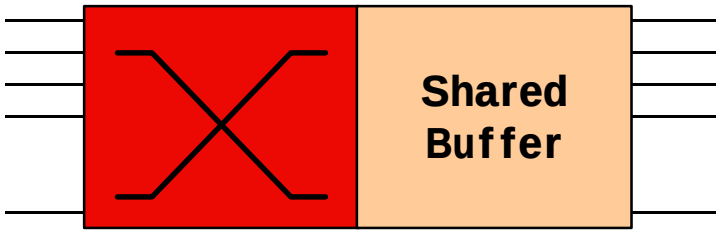
- Assumption: All arriving cells must be switched to the output for buffering
  - » When two or more cells are switched to the same output port in a cycle, extra cells are buffered until it can be transmitted.
  - » Switching transfer speed must be performed at N times the inlet speed. If transfer speed < N * inlet speed, then internal cell loss will occur.
- Advantage: Avoids HOL blocking, no arbitration required
- Disadvantage: Switch fabric may be difficult to implement

Internal Buffering



- Buffering of cells is done within the internal fabric of the switch.
- Advantage: Hides the concept of buffering
- Disadvantage: Difficult to implement efficiently.
- This technique is hardly used.

Shared Buffering



- Assumption: All arriving cells are switched to the shared output buffer.
- When two or more cells are switched to the same output port in a cycle, extra cells are buffered until it can be transmitted.
- Advantage: Memory requirement is greatly reduced, no HOL blocking
- Disadvantage: Limited memory bandwidth.

Other Queueing Strategies

- Multiple priorities may require more complicated queueing strategies.
- Partial buffer sharing
  - » Buffer is "divided" for high and low priority cells.
  - » High priority cells may only occupy high priority buffers.
- Push out buffer
  - » Buffer is shared
  - » High priority cells may overwrite low priority cells.

Queueing: Implementation Parameters

- Implementation complexity influenced by
  - » Queue Size
    - – Performance requirements
    - – Queueing discipline
  - » Memory Speed
    - – Queueing discipline
    - – Link speeds
    - – Number of ports
    - – Memory width
  - » Memory Control
    - – Queueing discipline

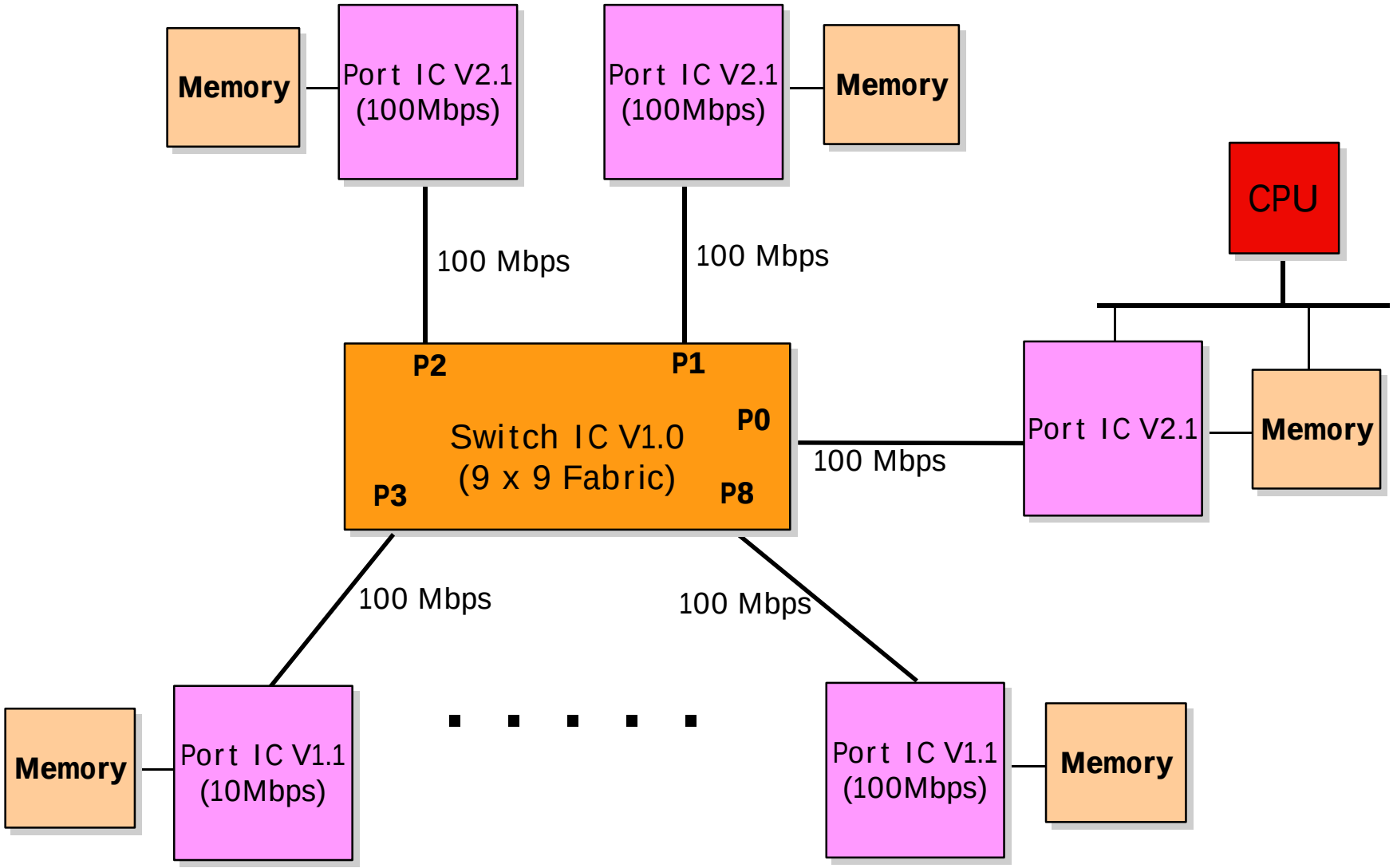
Queueing: Comparison

	Input Queueing	Output Queueing	Central/Shared Queueing
Single Port Memory	W/2F	W/(N+1)F	W/2NF
Example (ns)	53.3	6.3	3.8
Dual Port Memory	W/F	W/NF	W/NF
Example (ns)	106.6	6.7	6.7
Memory Speed	Low	High	High
Control Logic	FIFO	FIFO	Complex
Memory size	Very high	High	Low
Performance	Low	High	High
Multicast	Difficult	Easy	Medium

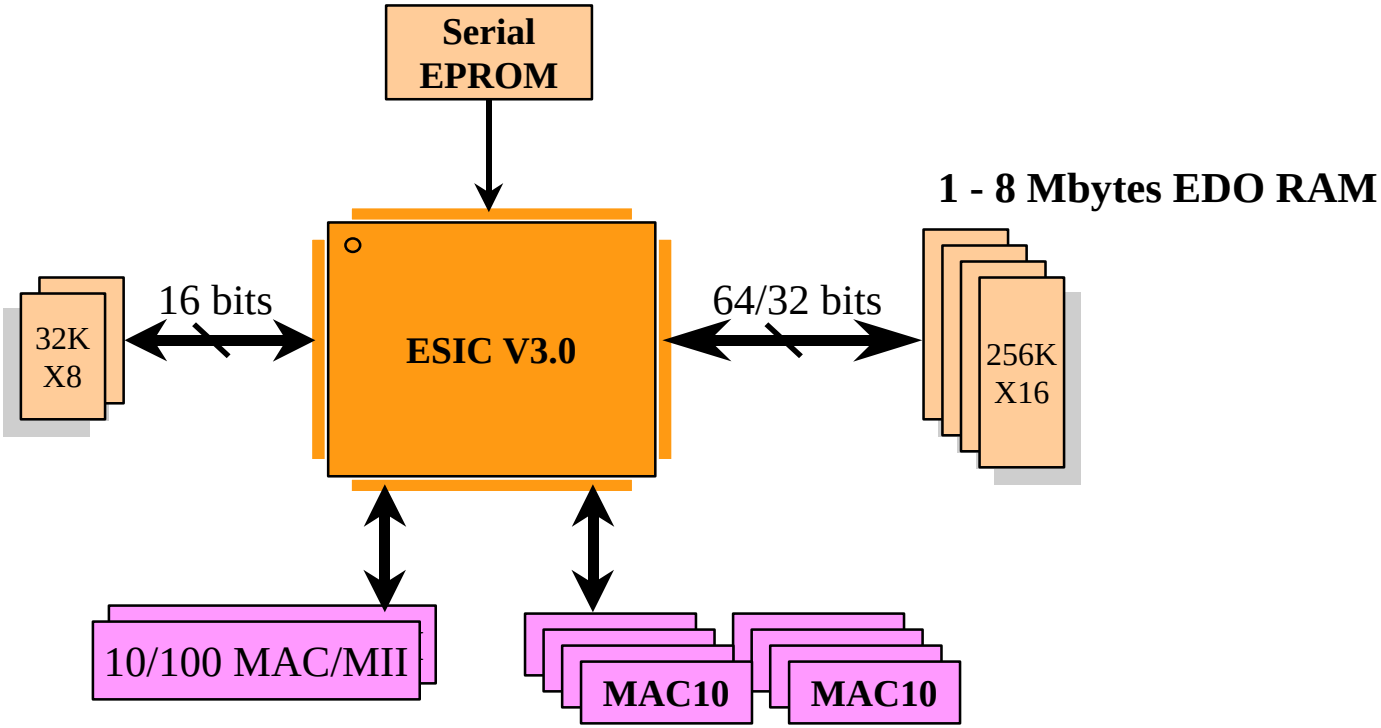
Cell size = 53 bytes
W = 16 bits
F = 150 Mbps
N = 16

ITRI
CCL

Ethernet Switch v1.x/v2.x Architecture

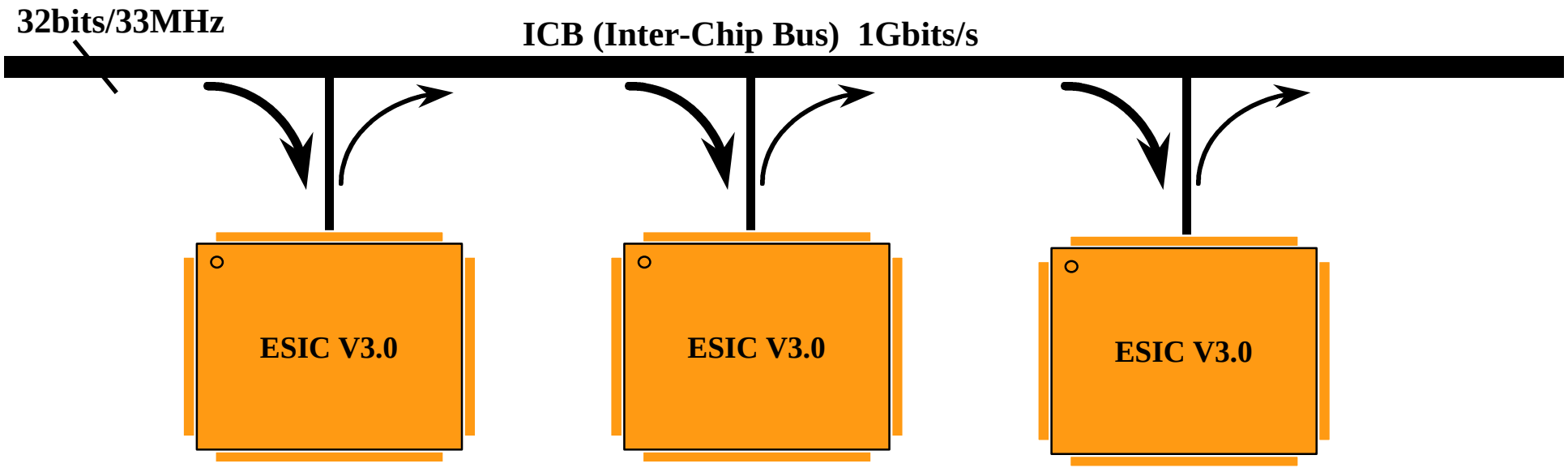


Ethernet Switch v3.0 Architecture



ITRI
CCL

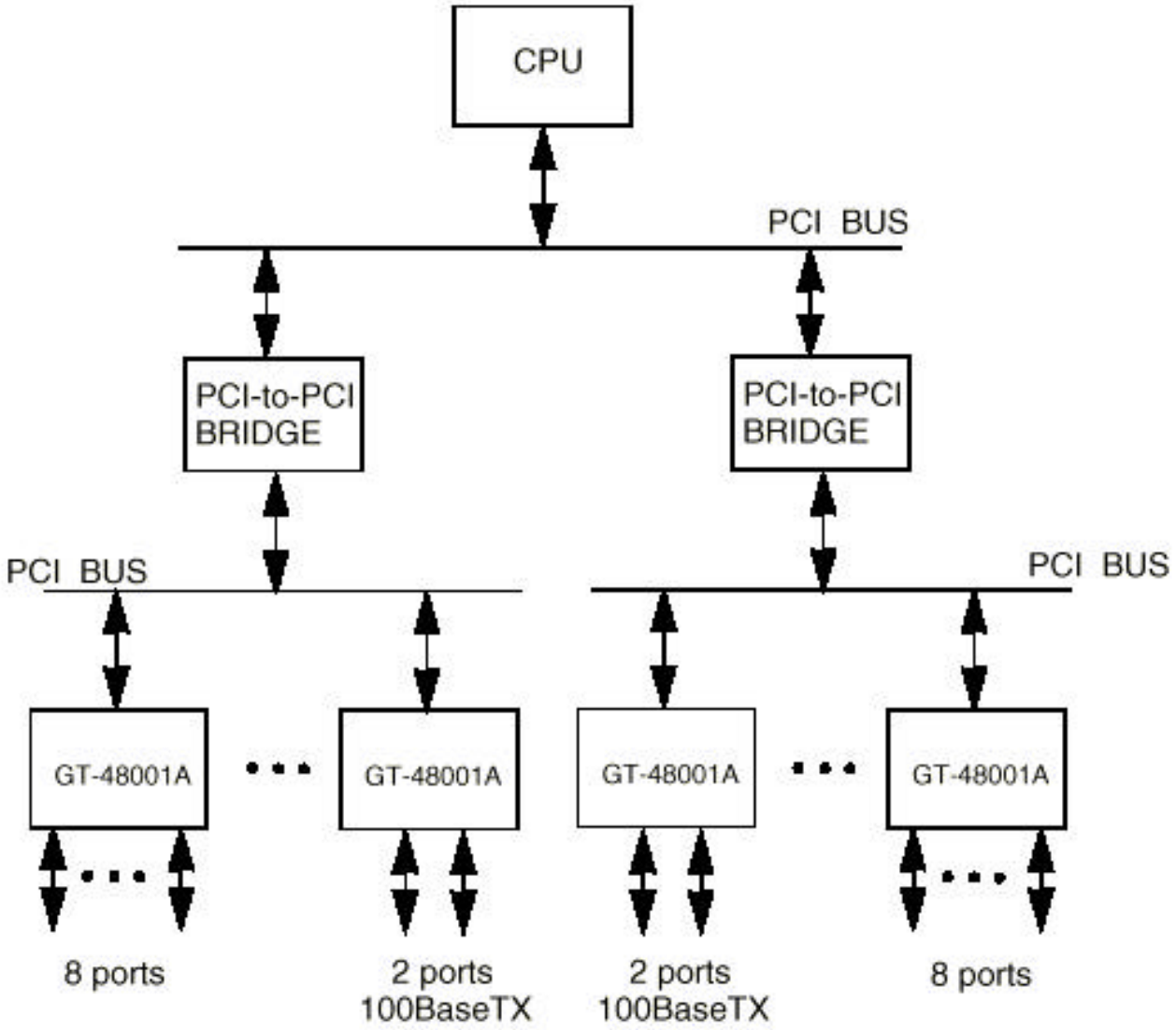
ESIC v3.0 Architecture: Scalable



- Non-Blocking for 3 Chips
- Maximum Connections for 8 Chips

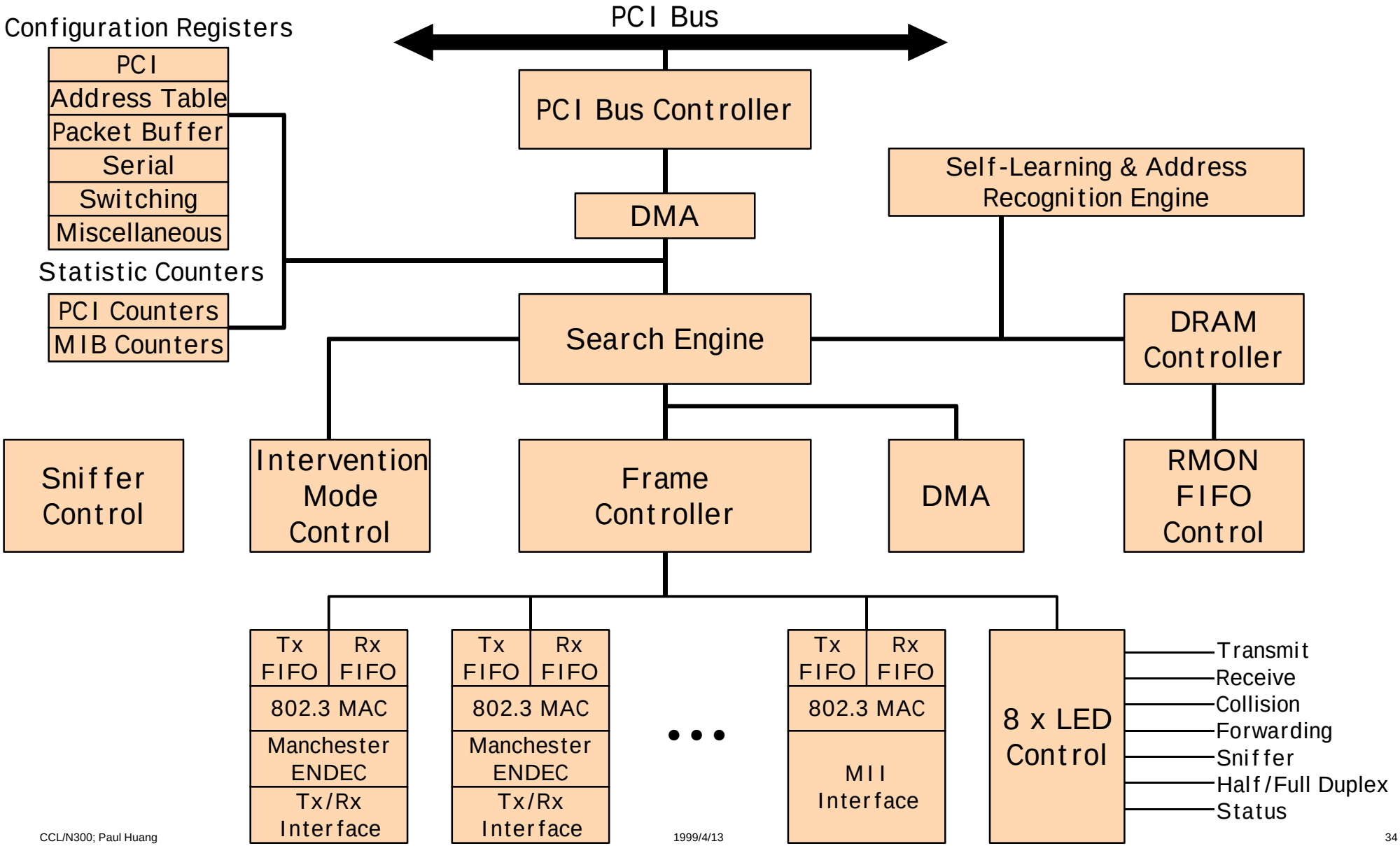
ITRI
CCL

Galileo System Architecture

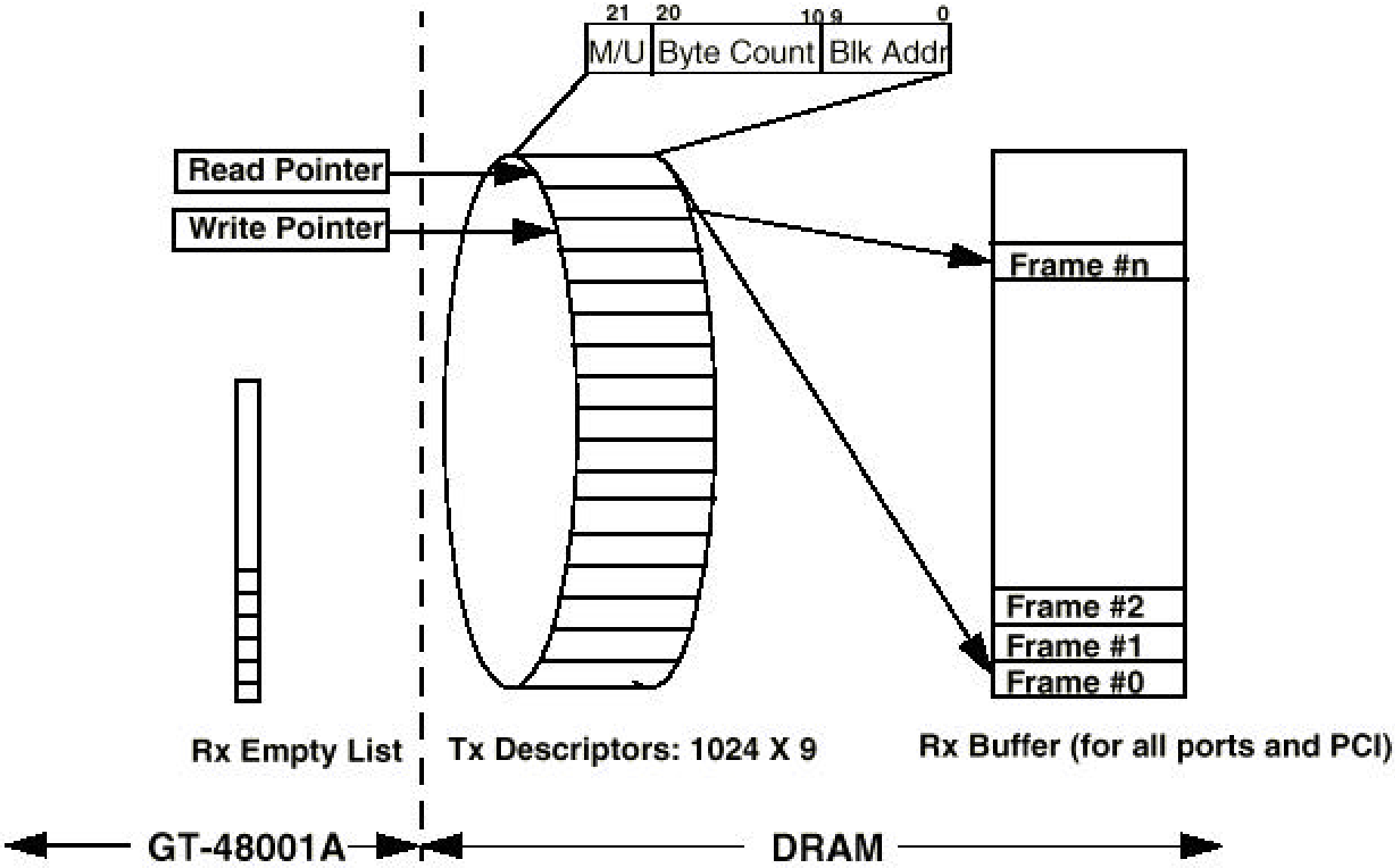


ITRI
CCL

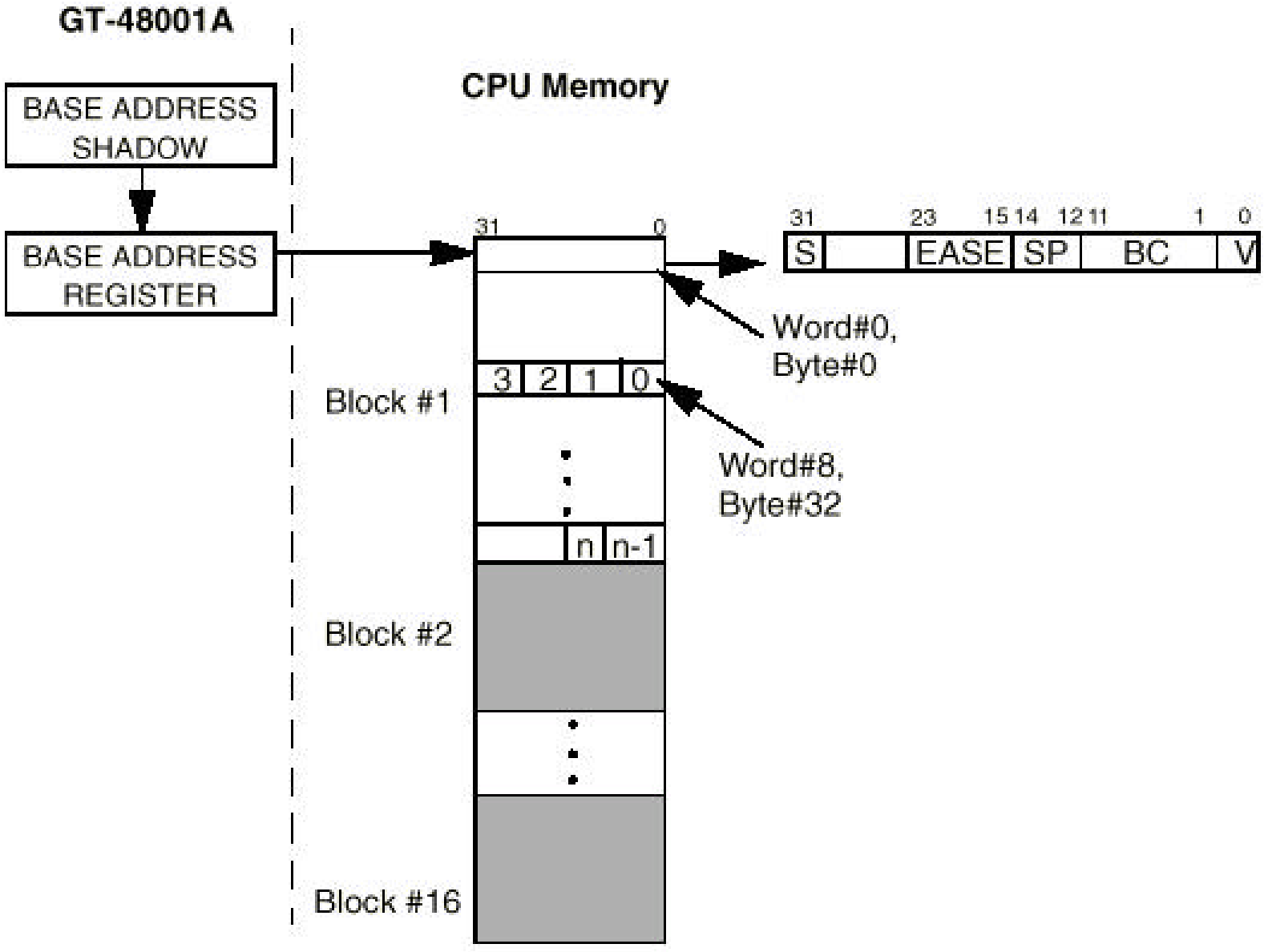
Galileo Chip Architecture



Galileo Buffer Management

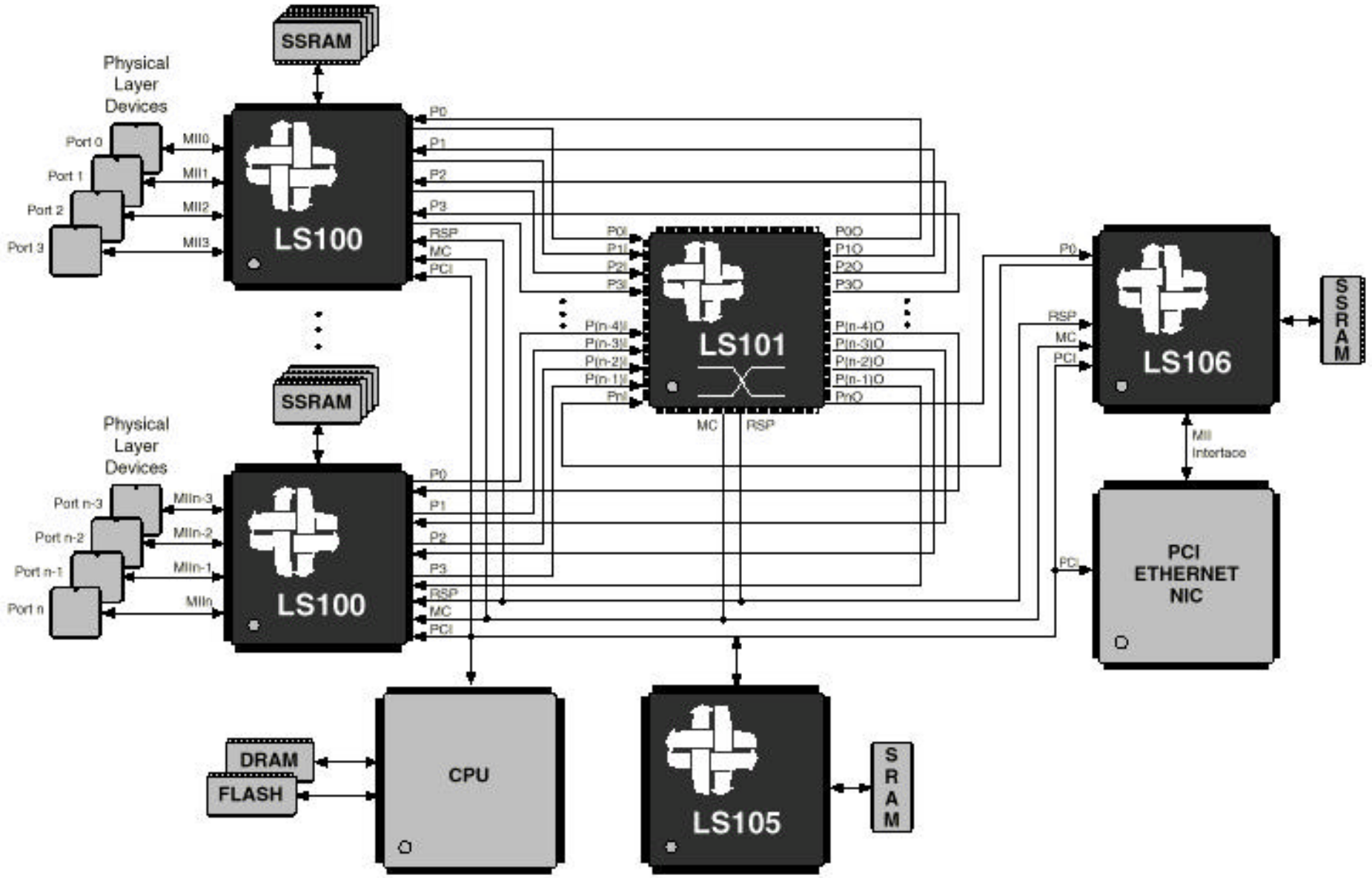


Galileo Memory Structure

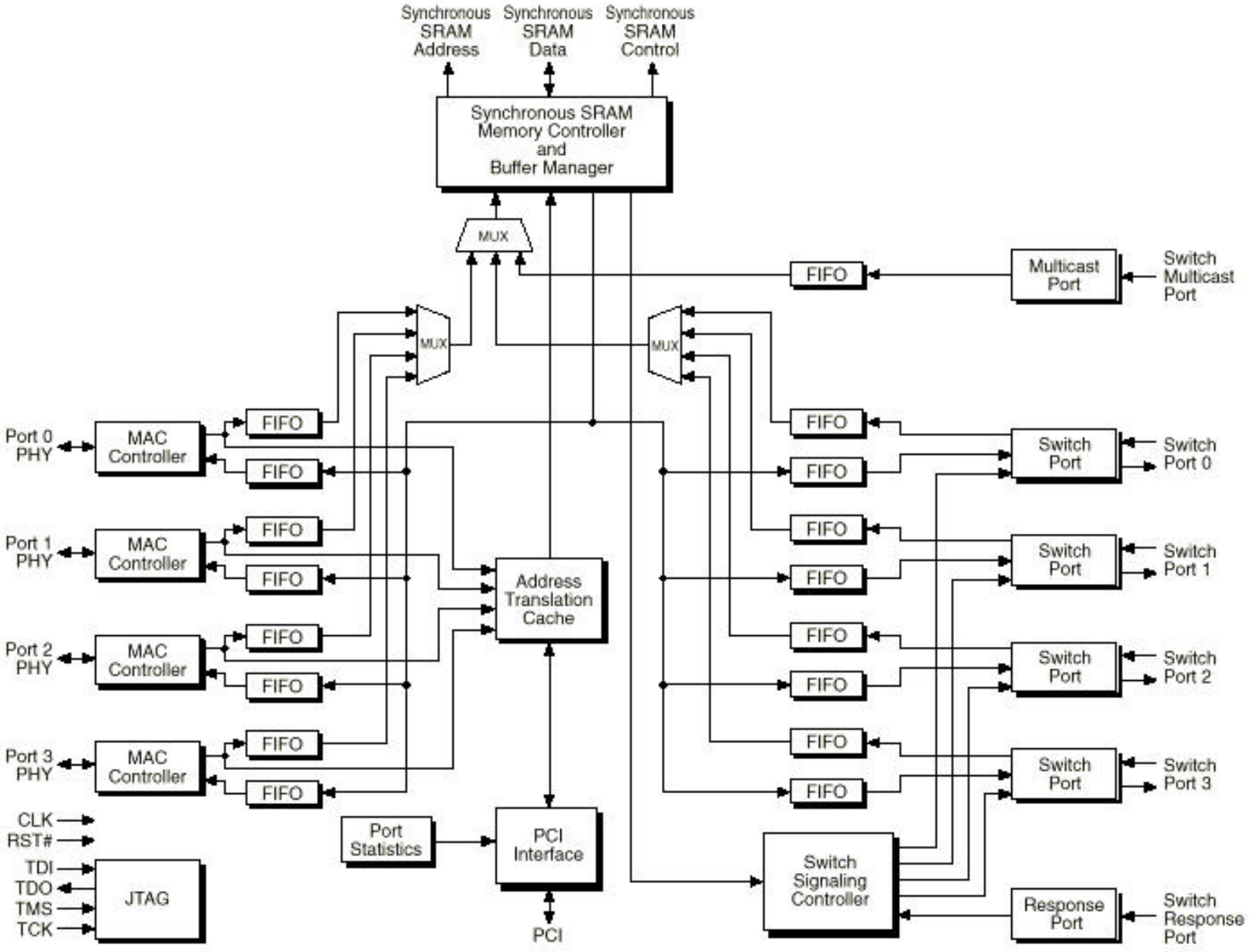


ITRI
CCL

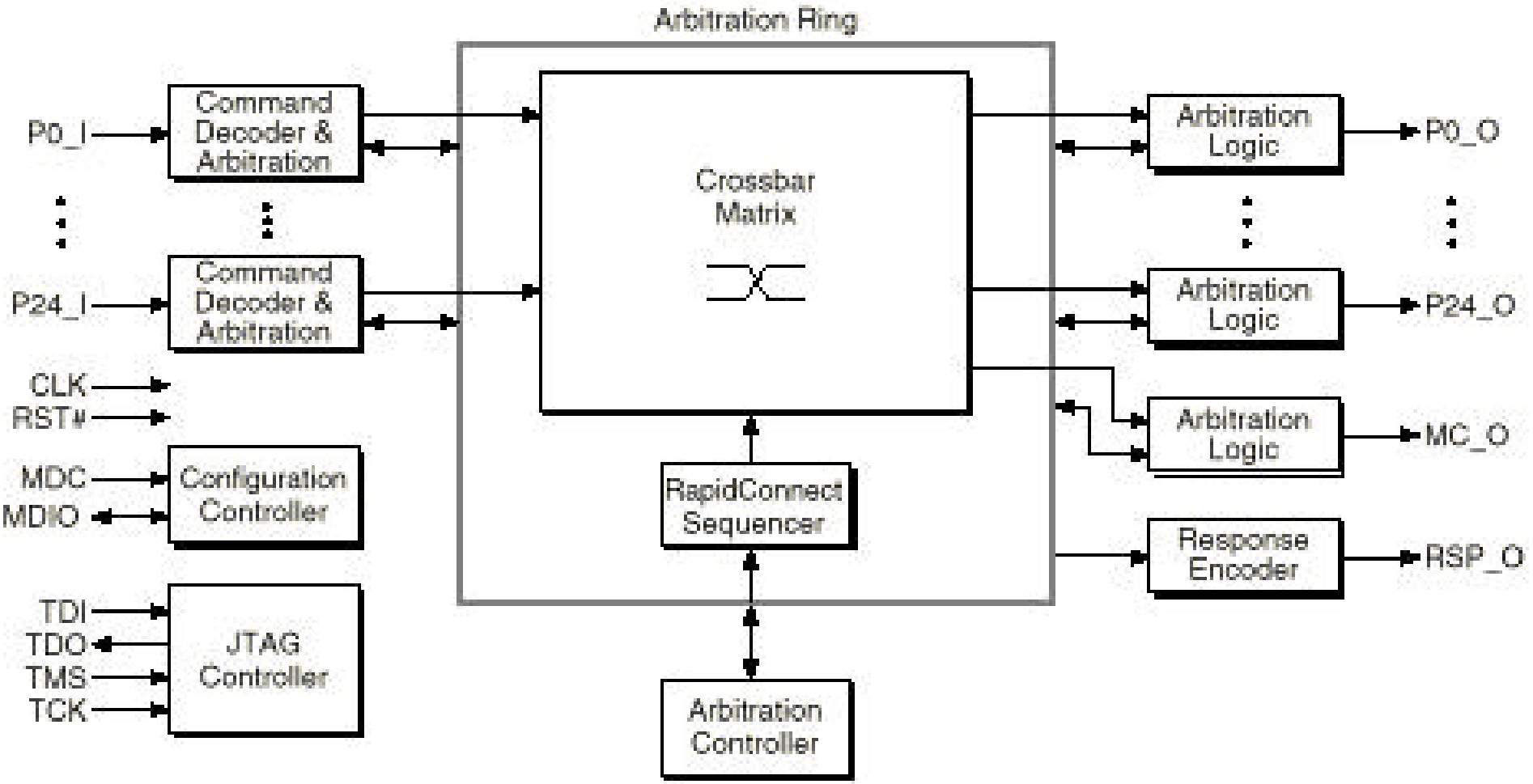
I-Cube Switch Architecture



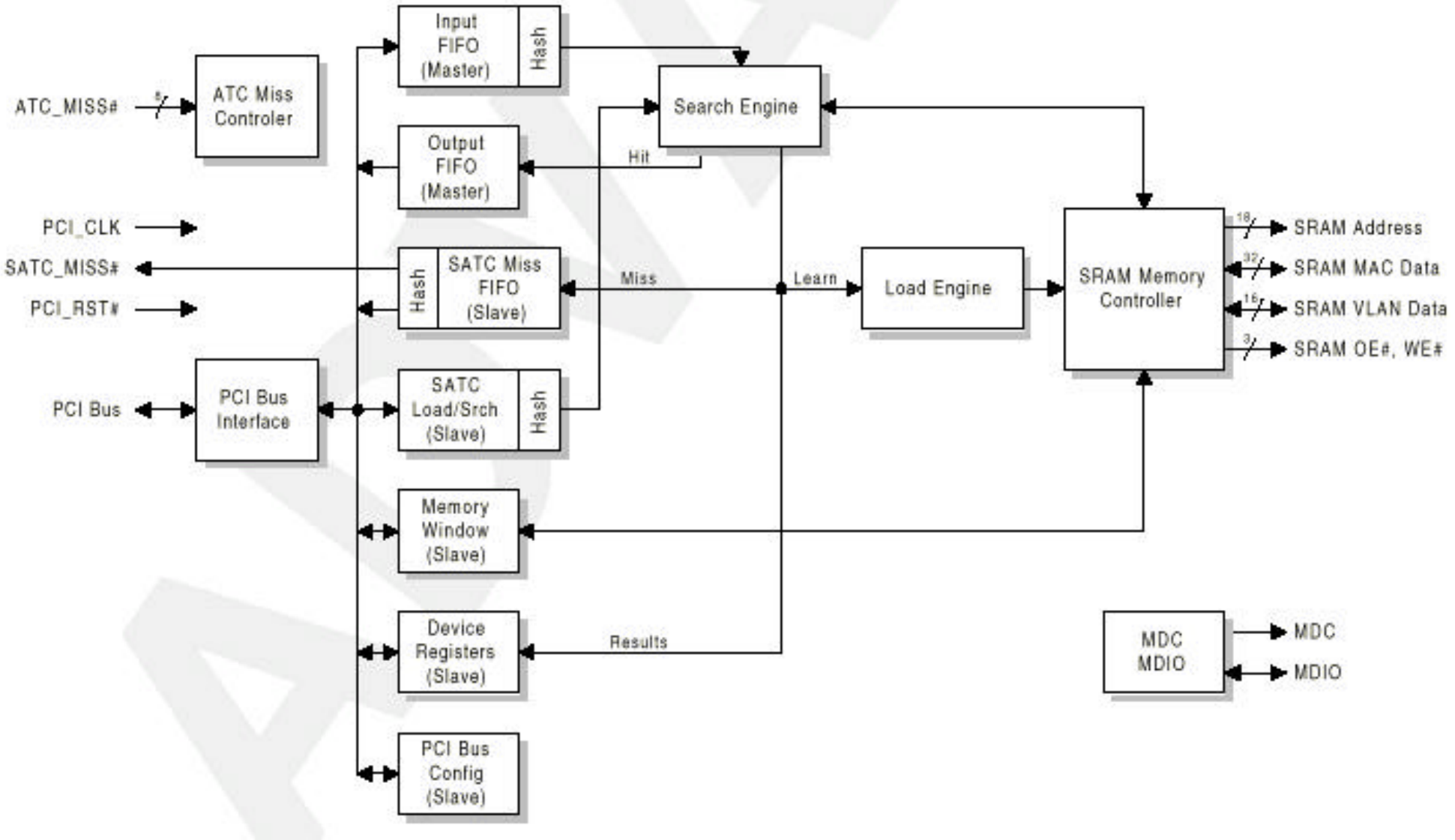
I-Cube LS100


Synchronous SRAM Address
Synchronous SRAM Data
Synchronous SRAM Control
Synchronous SRAM Memory Controller and Buffer Manager
MUX
MUX
MUX
FIFO
Multicast Port
Switch Multicast Port
Port 0 PHY
MAC Controller
FIFO
FIFO
Switch Port
Switch Port 0
Port 1 PHY
MAC Controller
FIFO
FIFO
Switch Port
Switch Port 1
Port 2 PHY
MAC Controller
FIFO
FIFO
Switch Port
Switch Port 2
Port 3 PHY
MAC Controller
FIFO
FIFO
Switch Port
Switch Port 3
Address Translation Cache
CLK
RST#
TDI
TDO
TMS
TCK
JTAG
Port Statistics
PCI Interface
PCI
Switch Signaling Controller
Response Port
Switch Response Port

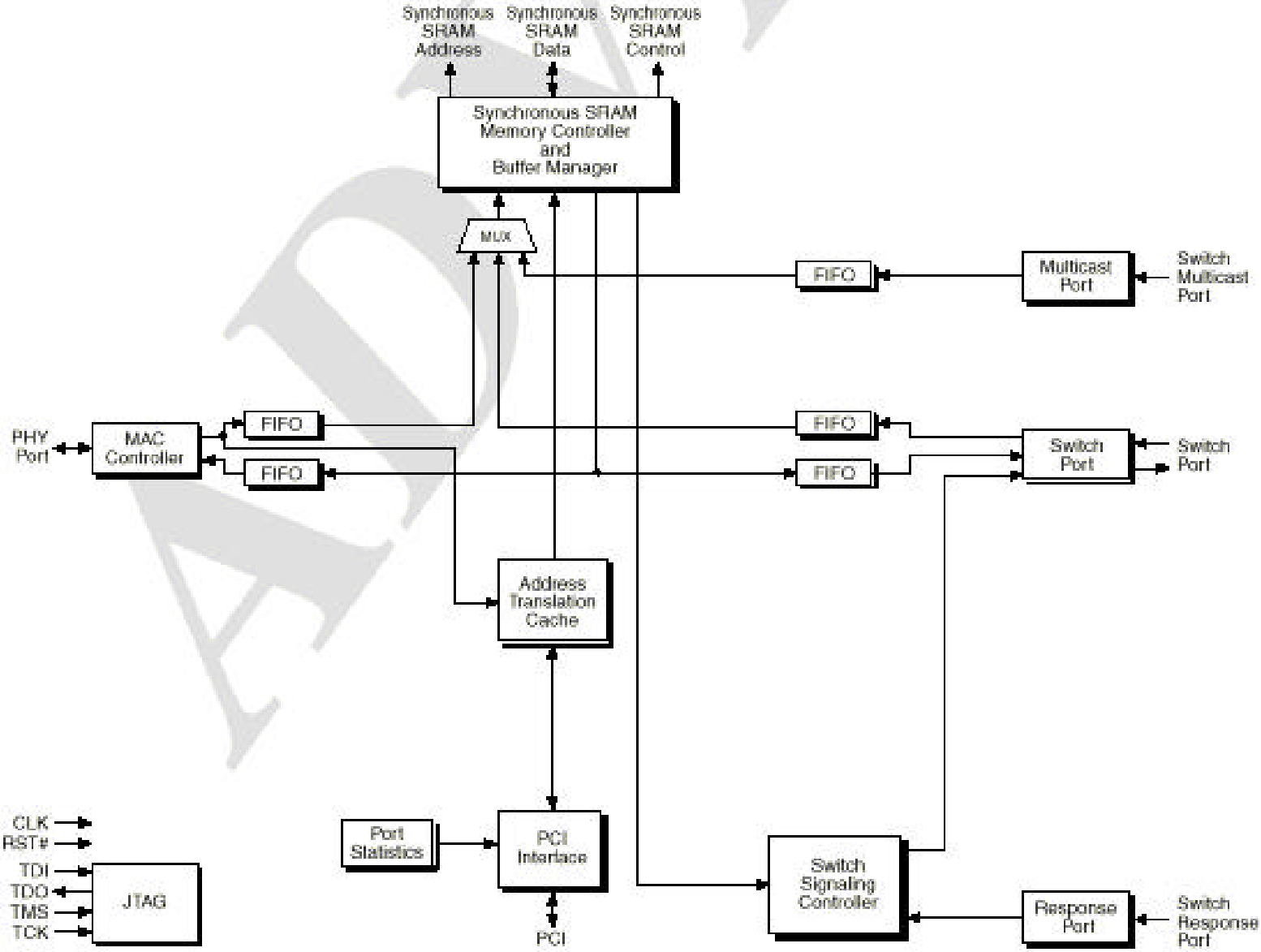
I-Cube LS101


Arbitration Ring
P0_I
Command Decoder & Arbitration
P24_I
Command Decoder & Arbitration
CLK
RST#
MDC
MDIO
Configuration Controller
TDI
TDO
TMS
TCK
JTAG Controller
Crossbar Matrix
RapidConnect Sequencer
Arbitration Controller
Arbitration Logic
Arbitration Logic
Arbitration Logic
Response Encoder
P0_O
P24_O
MC_O
RSP_O

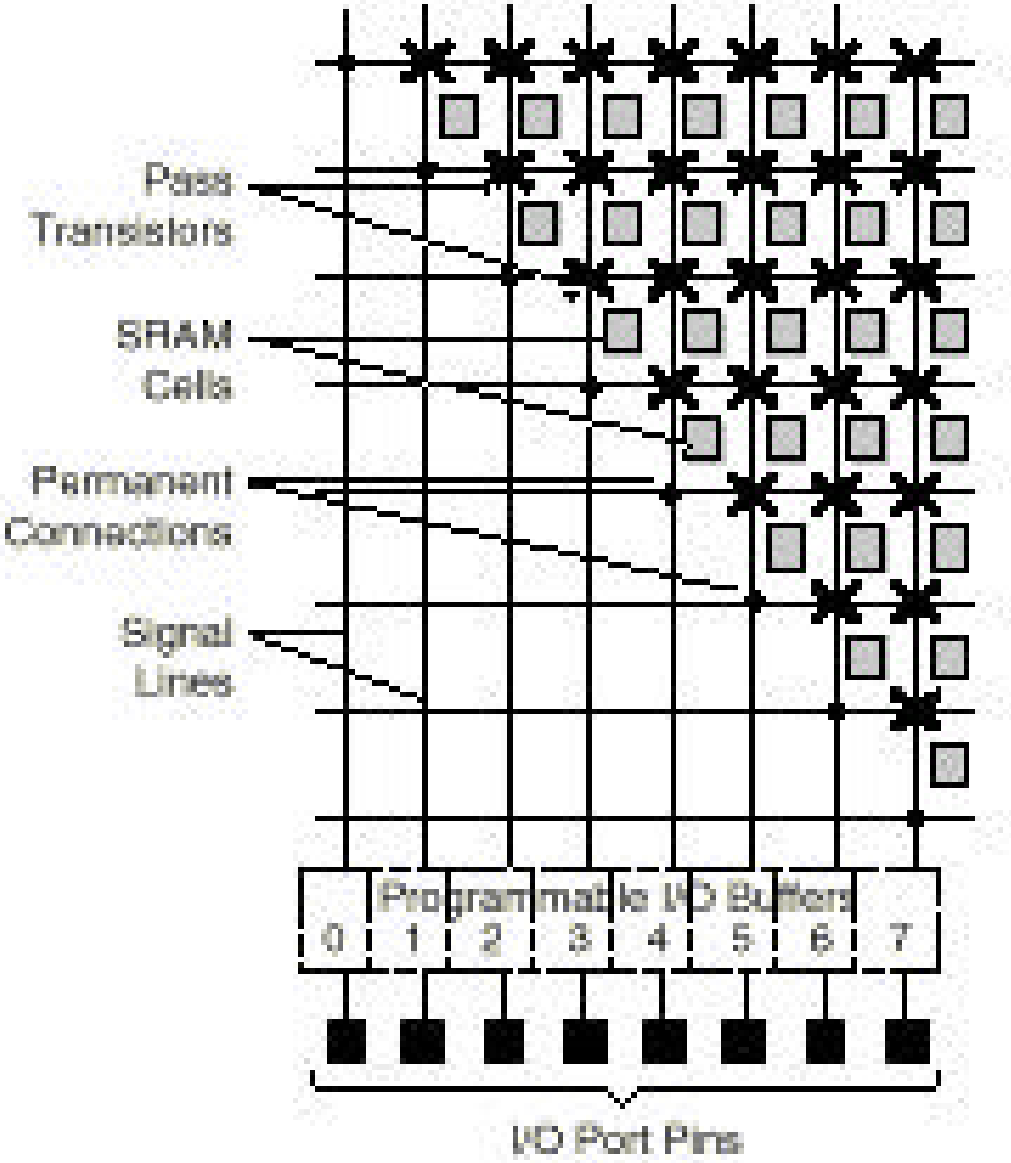
I-Cube LS105


ATC_MISS#
5
ATC Miss Controler
PCI_CLK
SATC_MISS#
PCI_RST#
PCI Bus
PCI Bus Interface
Input FIFO (Master)
Hash
Output FIFO (Master)
Hash
SATC Miss FIFO (Slave)
SATC Load/Srch (Slave)
Hash
Memory Window (Slave)
Device Registers (Slave)
PCI Bus Config (Slave)
Search Engine
Hit
Miss
Learn
Load Engine
Results
SRAM Memory Controller
18
SRAM Address
32
SRAM MAC Data
16
SRAM VLAN Data
3
SRAM OE#, WE#
MDC
MDIO
MDC
MDIO

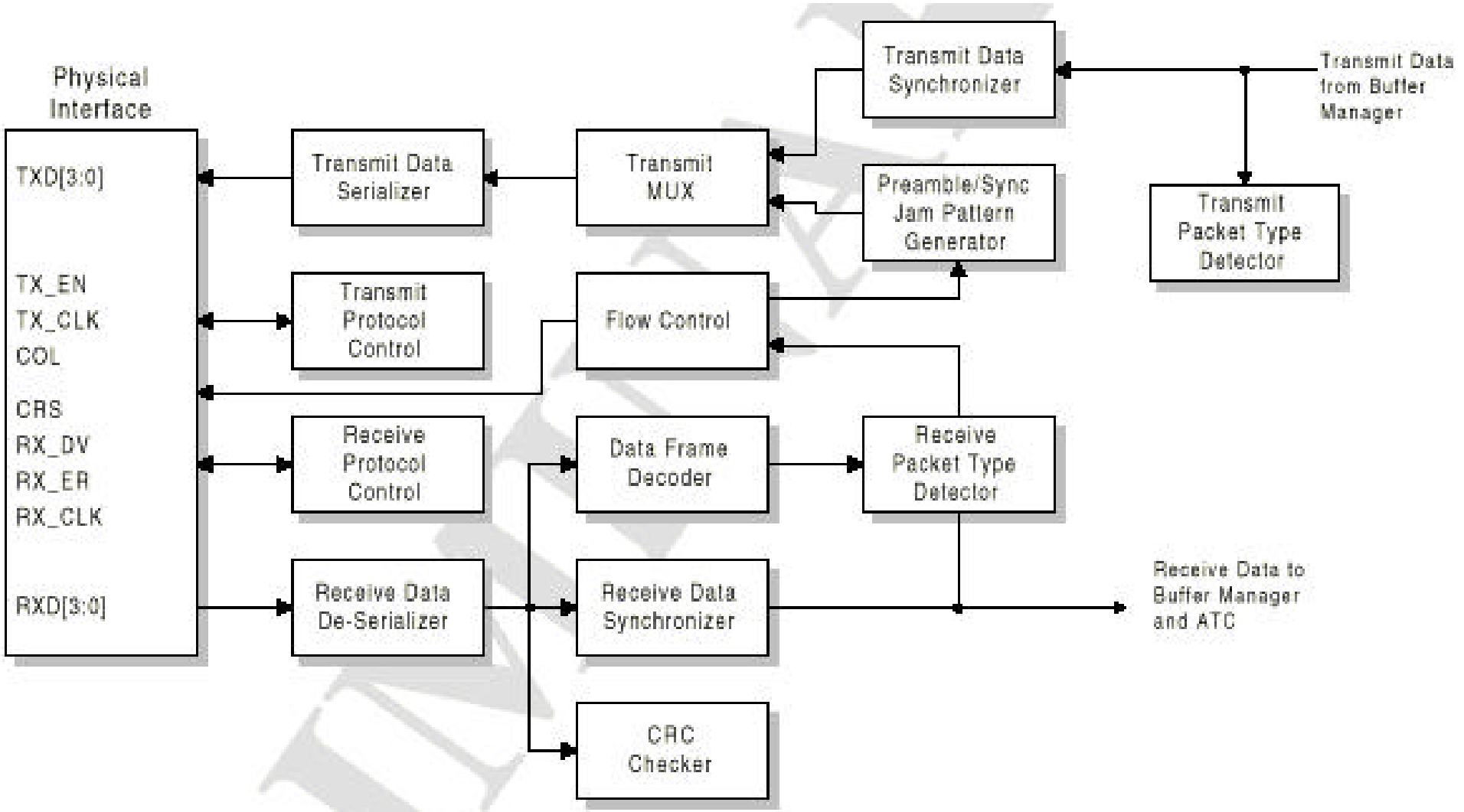
I-Cube LS106


Synchronous SRAM Address
Synchronous SRAM Data
Synchronous SRAM Control
Synchronous SRAM Memory Controller and Buffer Manager
MUX
FIFO
Multicast Port
Switch Multicast Port
PHY Port
MAC Controller
FIFO
FIFO
FIFO
FIFO
Switch Port
Switch Port
Address Translation Cache
CLK
RST#
TDI
TDO
TMS
TCK
JTAG
Port Statistics
PCI Interface
PCI
Switch Signaling Controller
Response Port
Switch Response Port

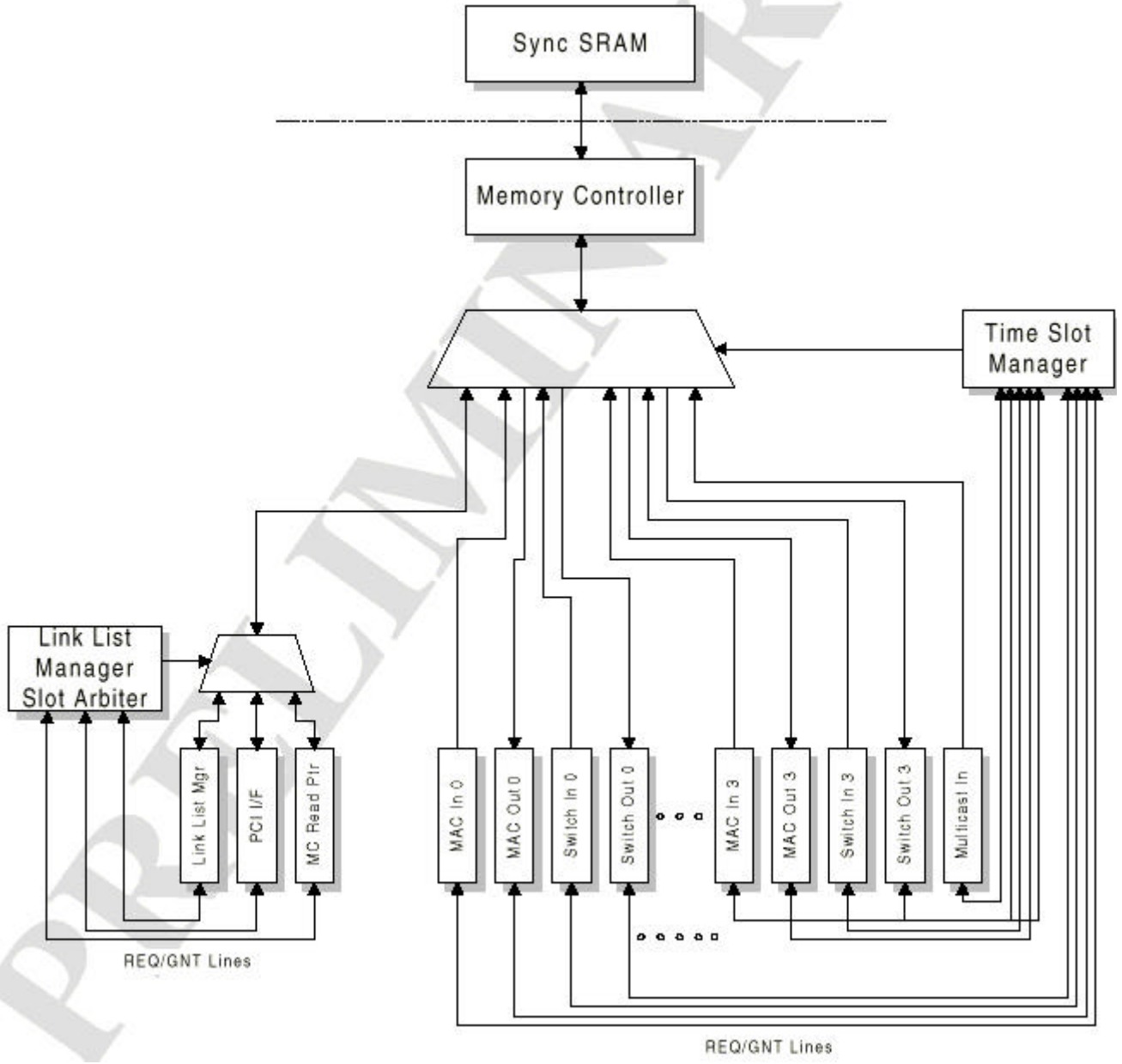
I-Cube LS101 Switch Matrix



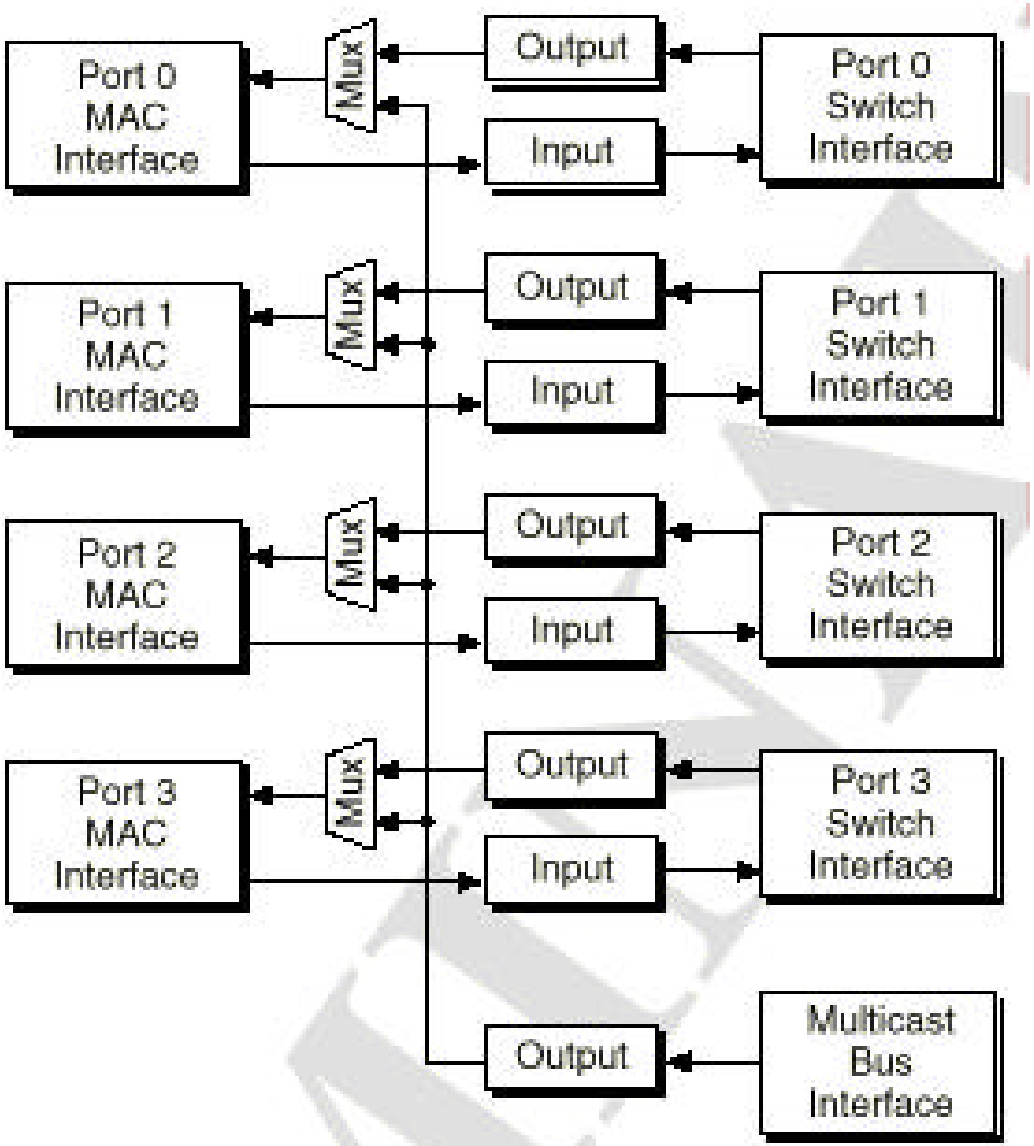
MAC Layer Controller



Memory & Buffer Mgt.

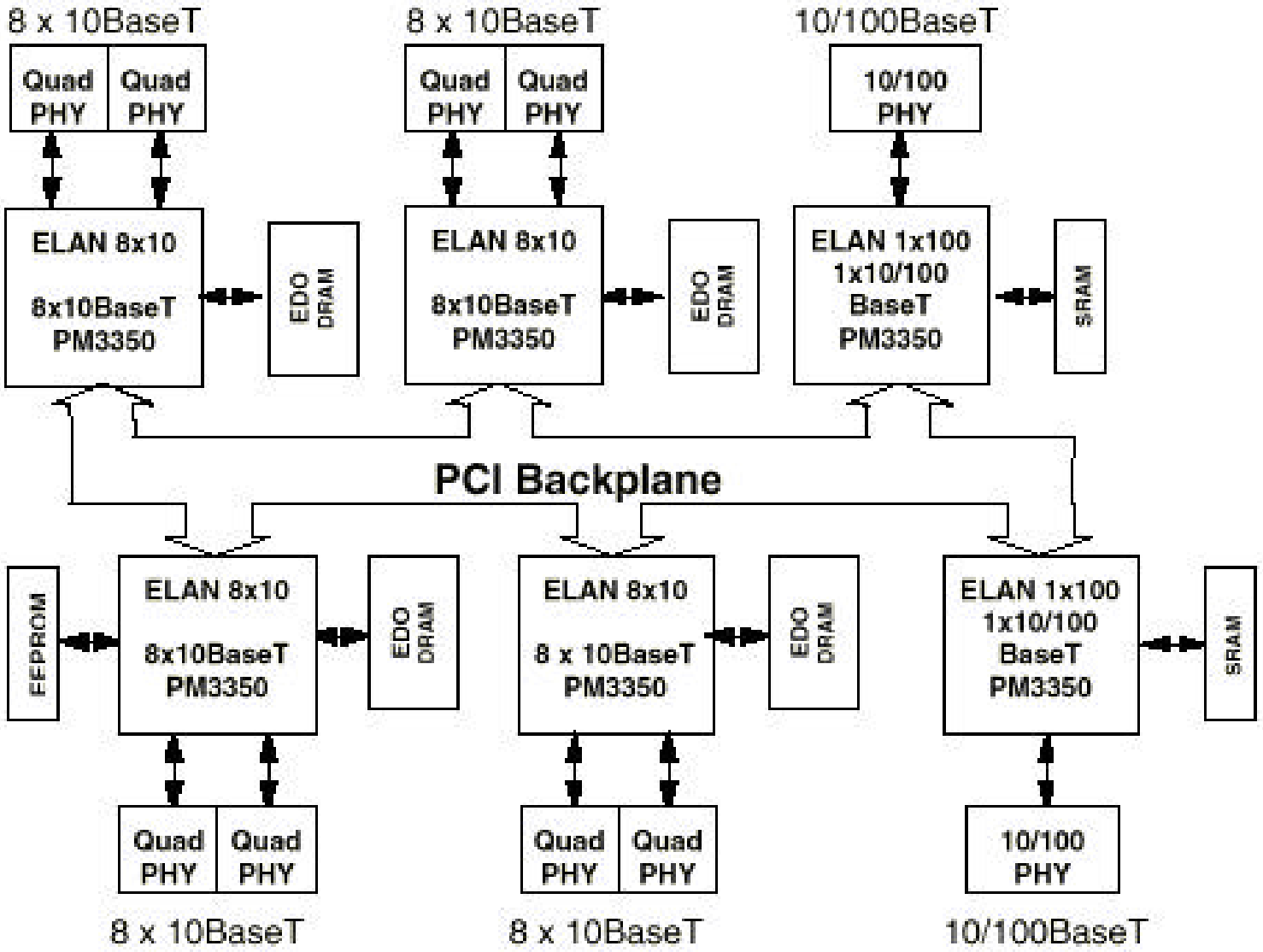


Queue Dataflow

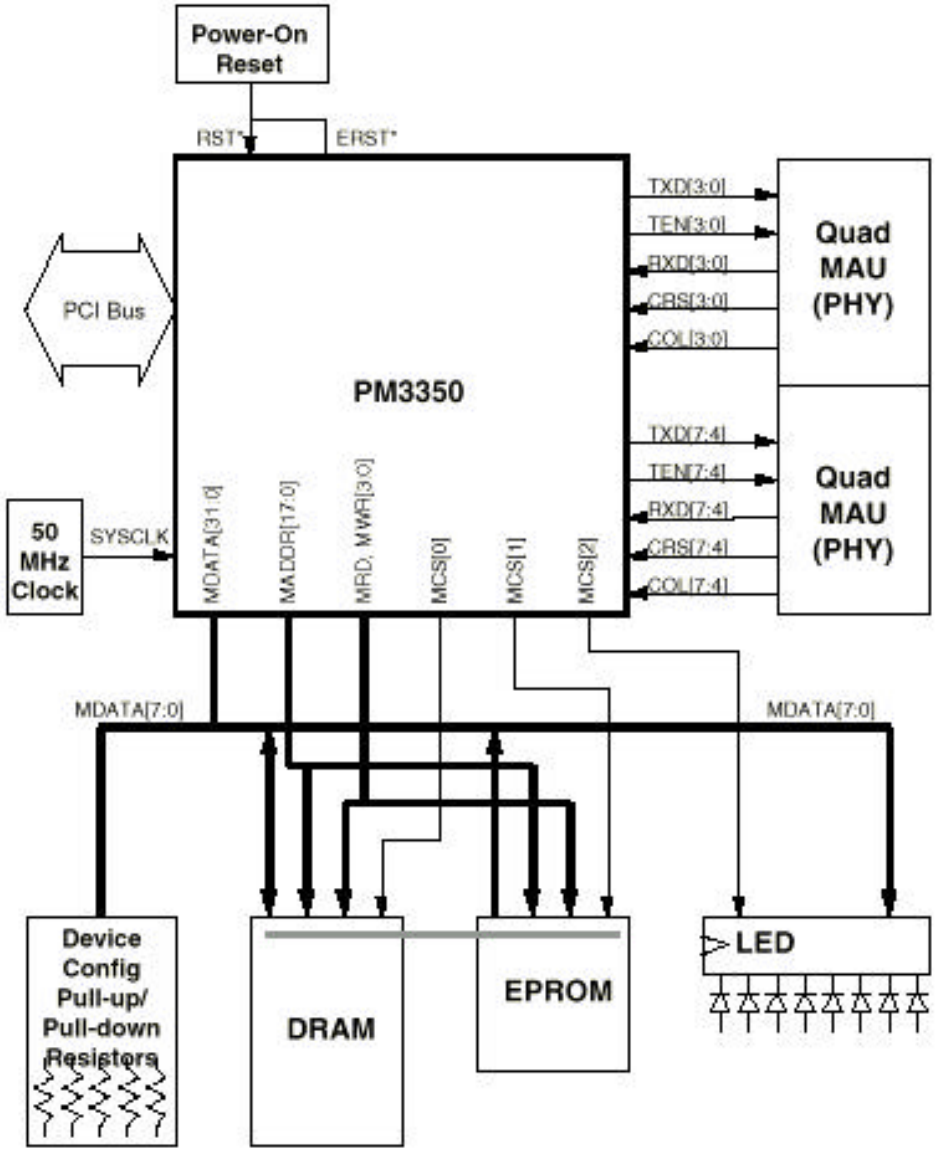


ITRI
CCL

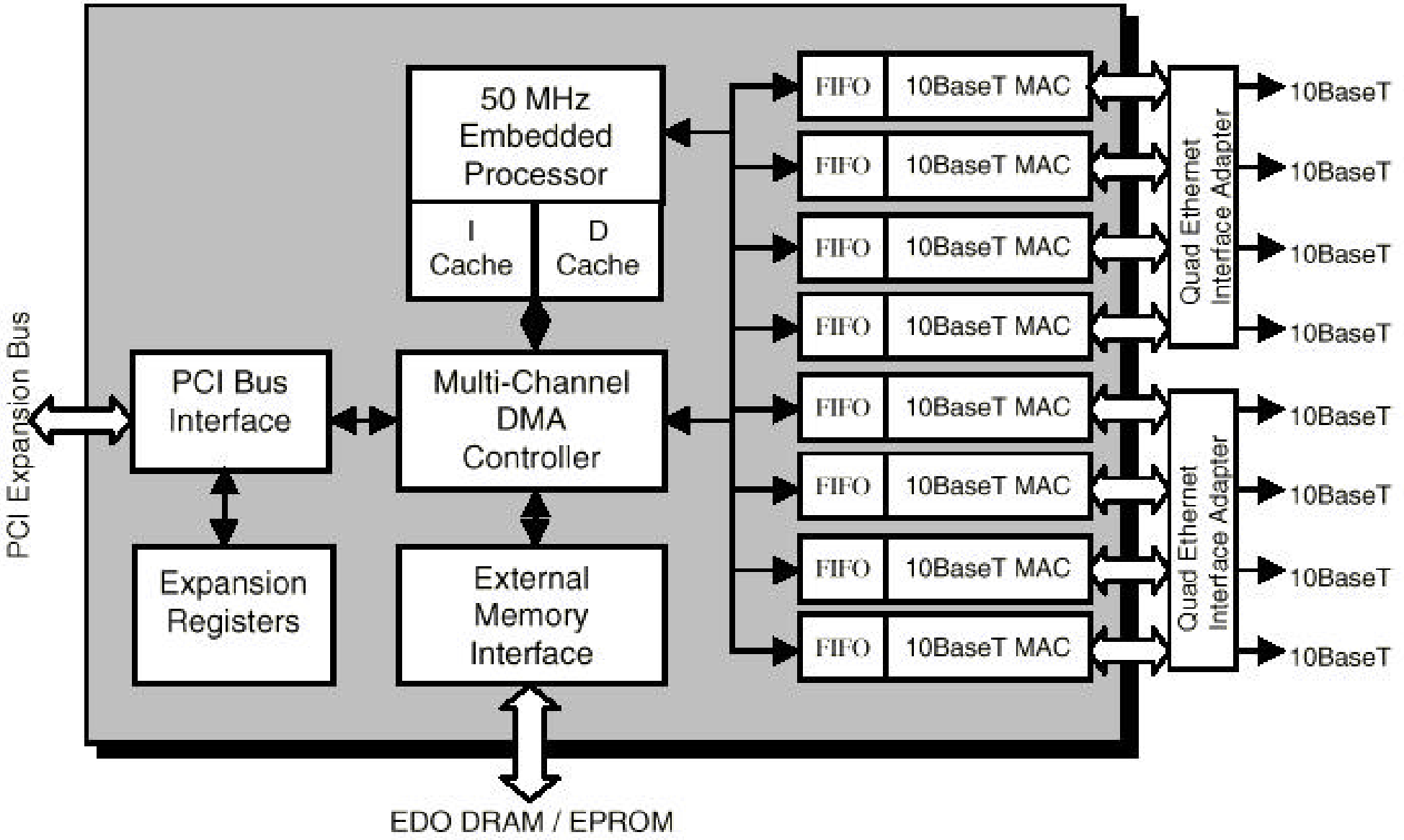
PMC Sierra System Architecture



PMC Sierra Chip Architecture

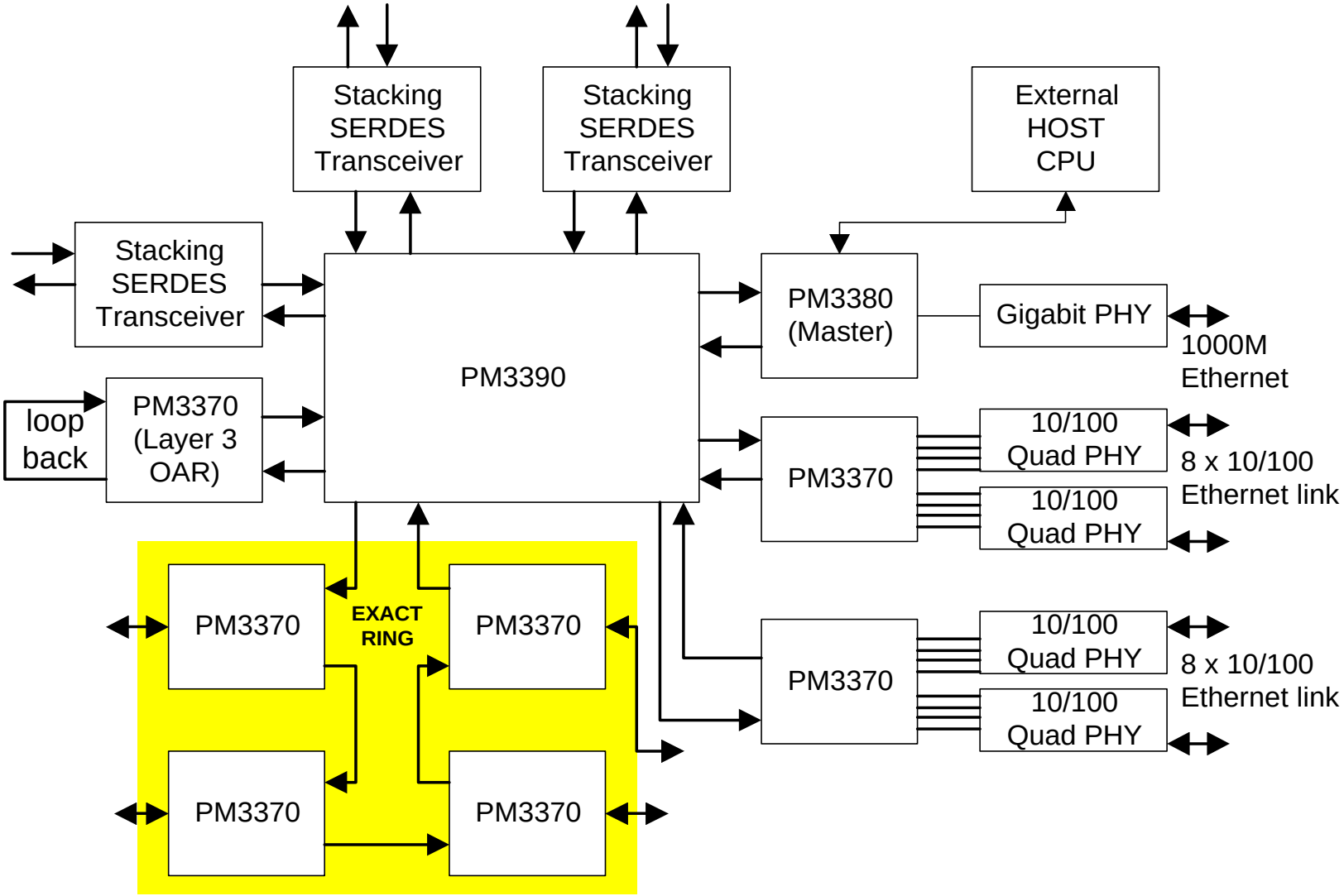


PMC Sierra Chip Architecture



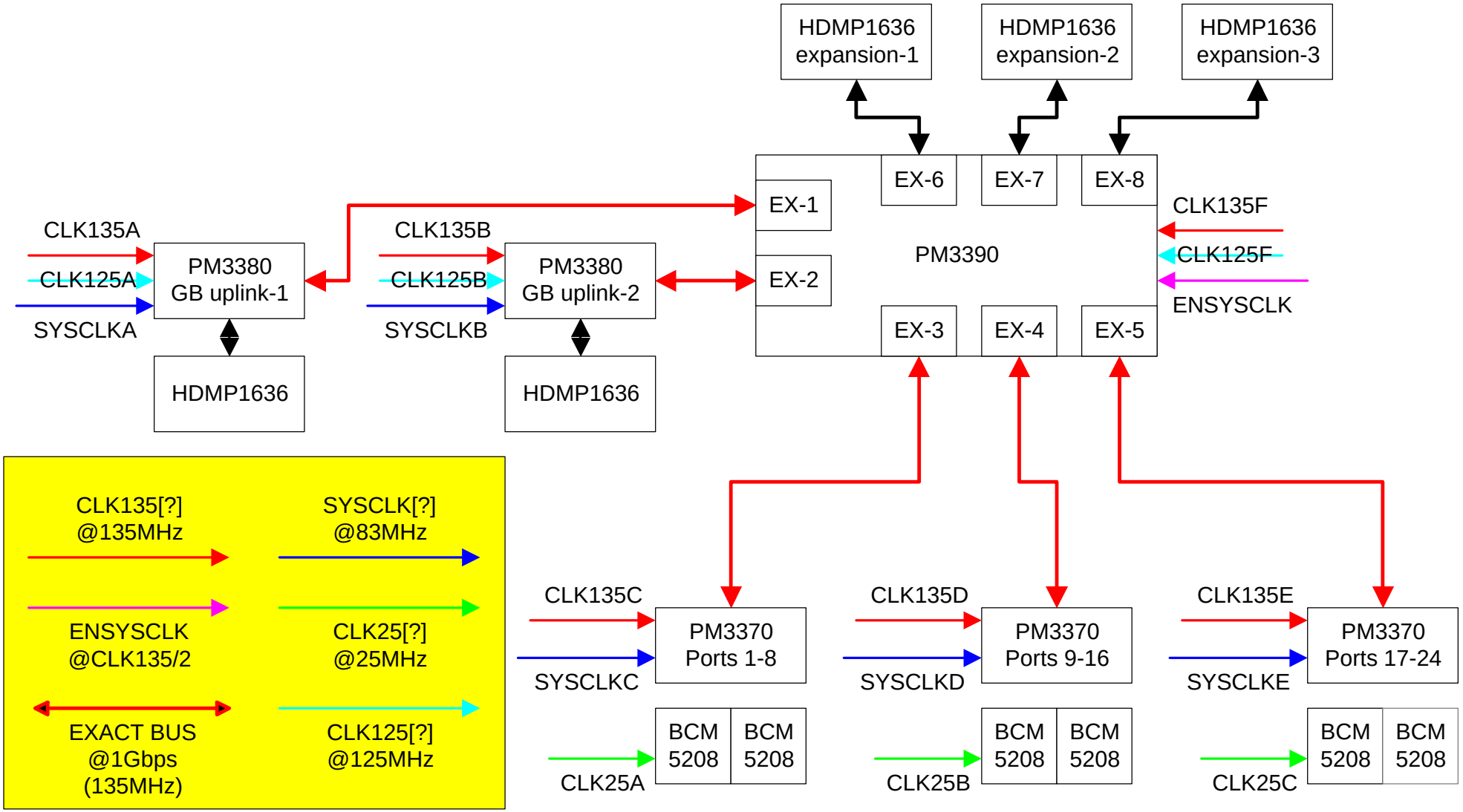
ITRI
CCL

PMC Sierra L3 System Architecture



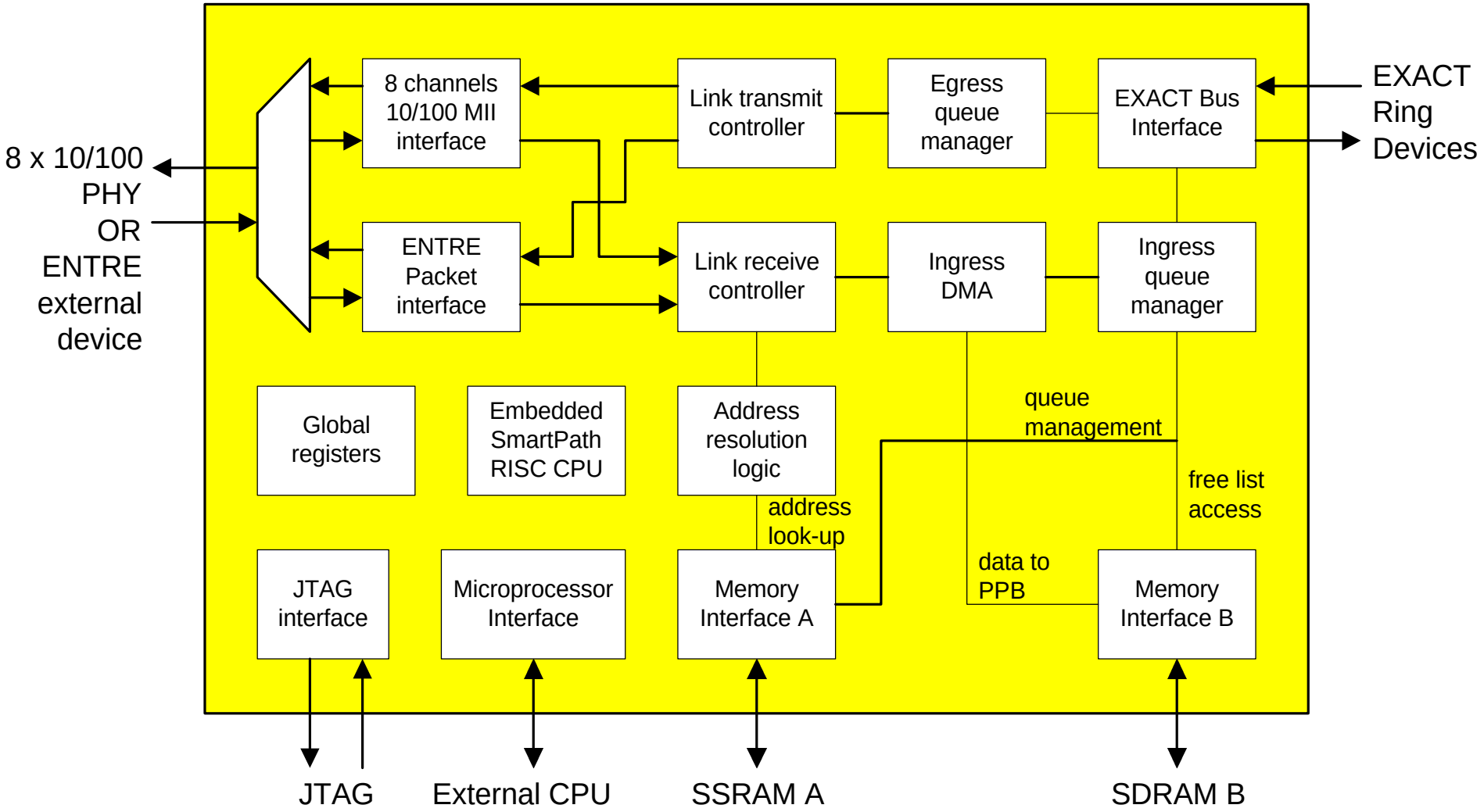
ITRI
CCL

PMC Sierra PM3390



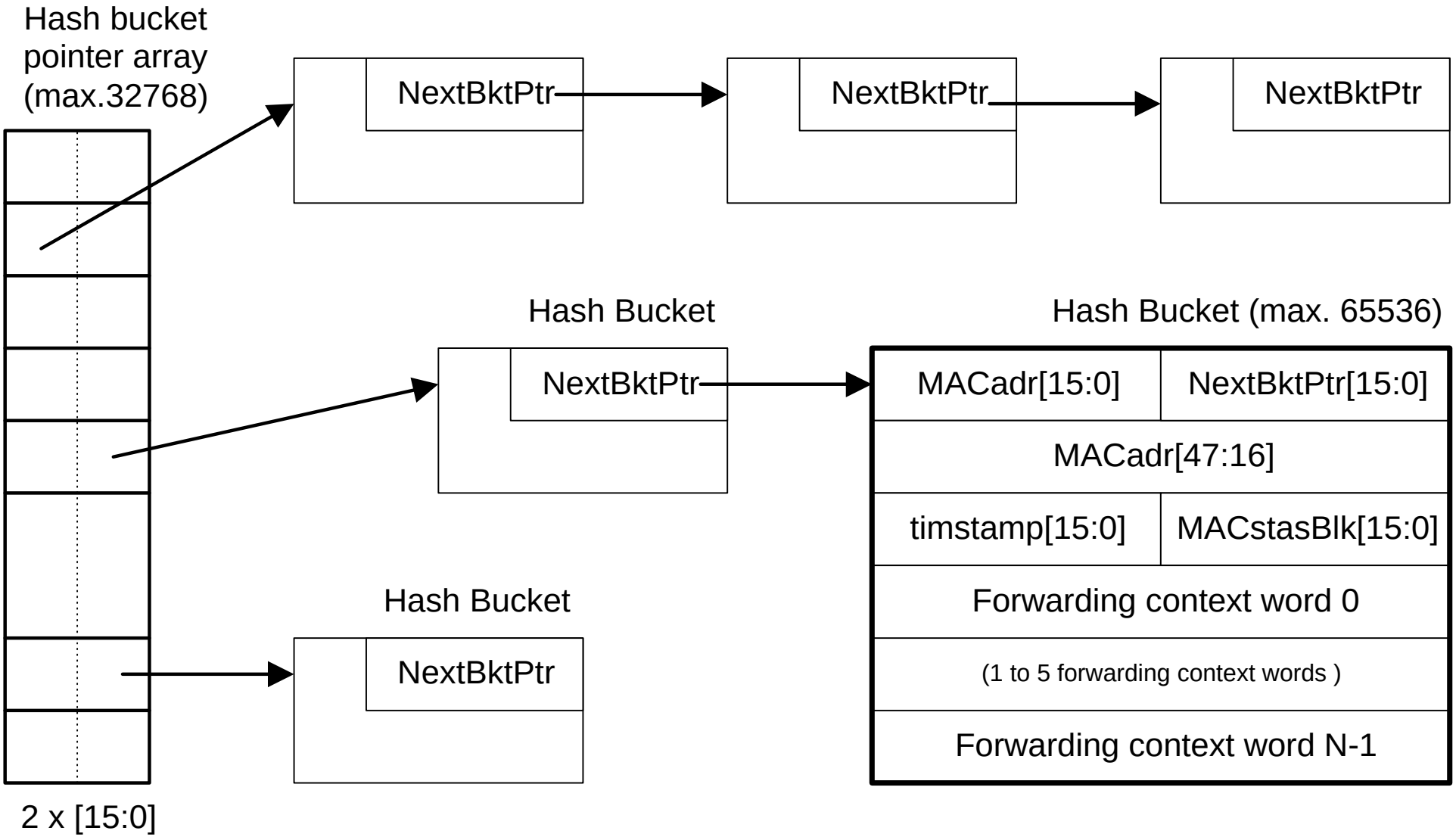
ITRI
CCL

PM3370 Octal Fast ES Port Controller

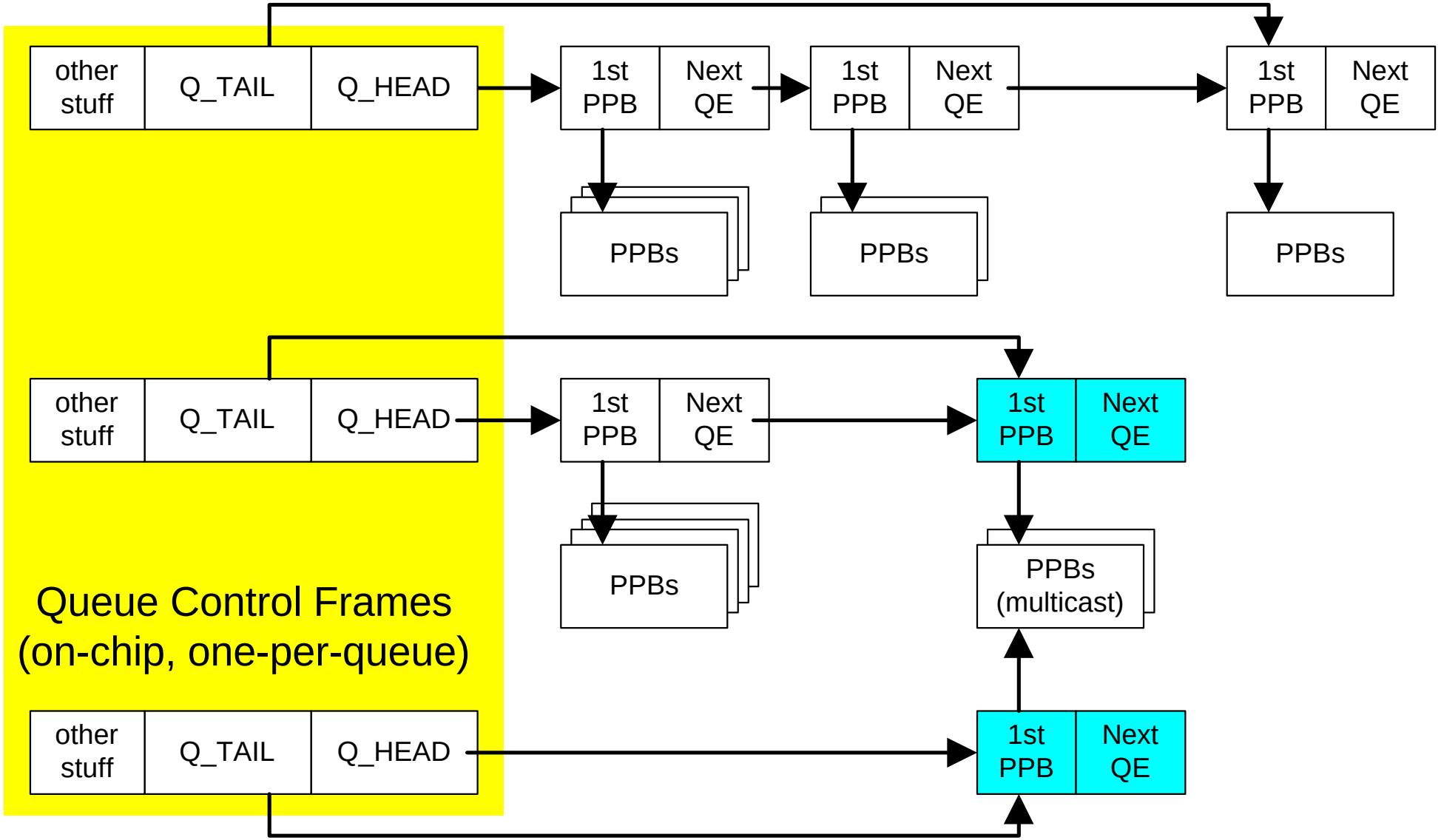


ITRI
CCL

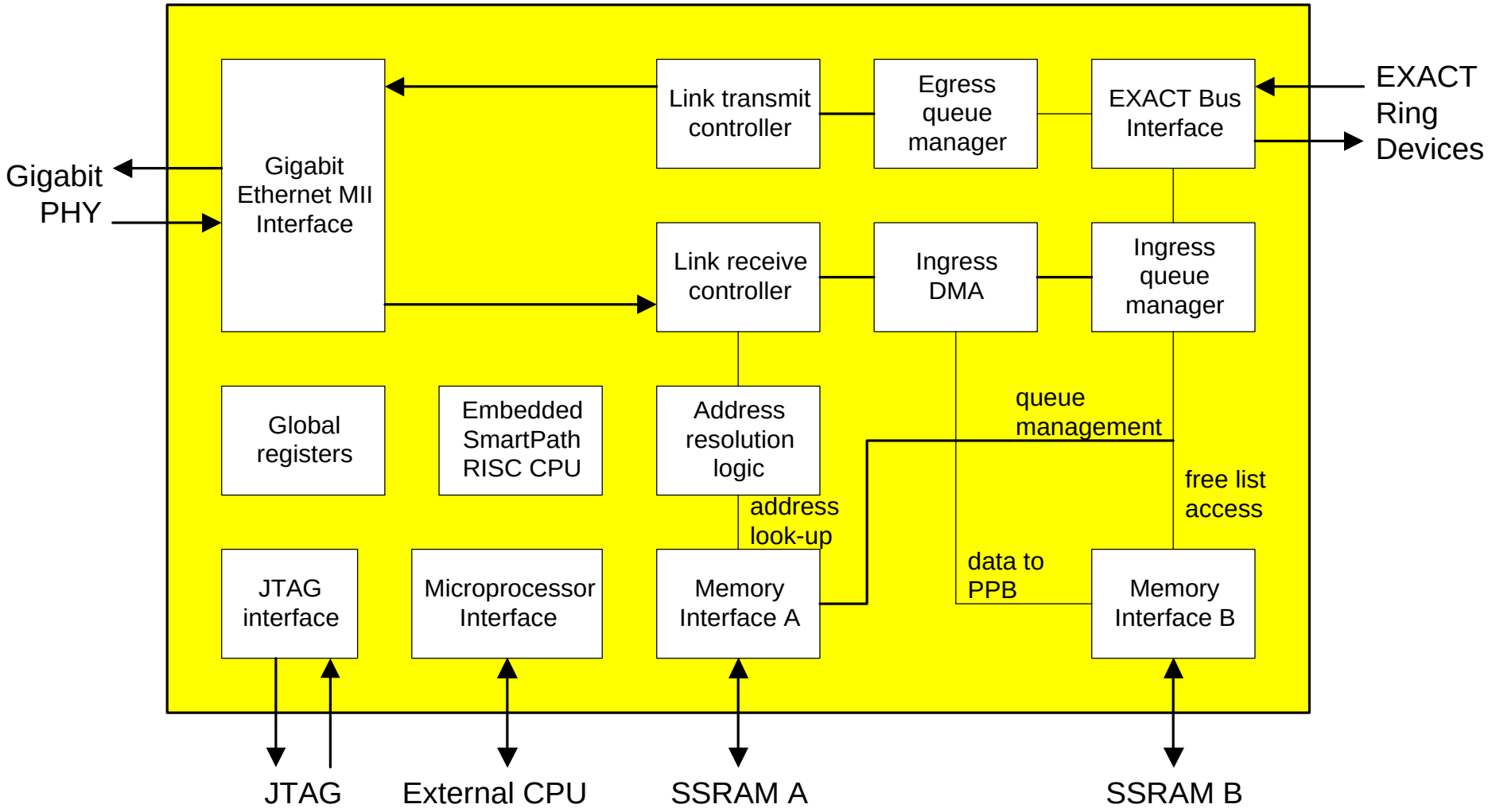
Address Table: Hash Bucket



Output Queue Structures



PM3380 GE Switch Port Controller



ITRI
CCL

Output Associated Input Queueing

- Packets segmented into 240-byte blocks of Partial Packet Buffer (PPB).
- Input packets are classified by destination and priority (256-port x 4-class).
- IQM issues Queue Allocated (QA) to destination if queue is not empty.

(1-a): DA,SA extracted to ARL

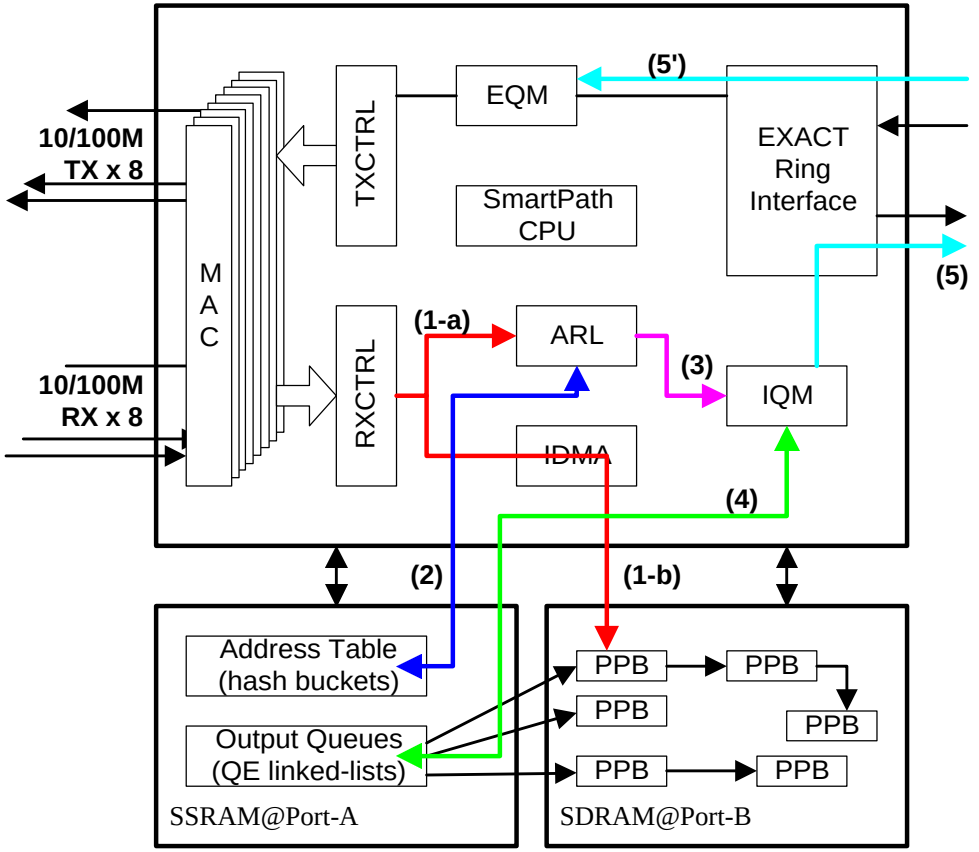
(1-b): packet segmented into PPB

(2): DA/SA address lookup

(3): routing result to IQM

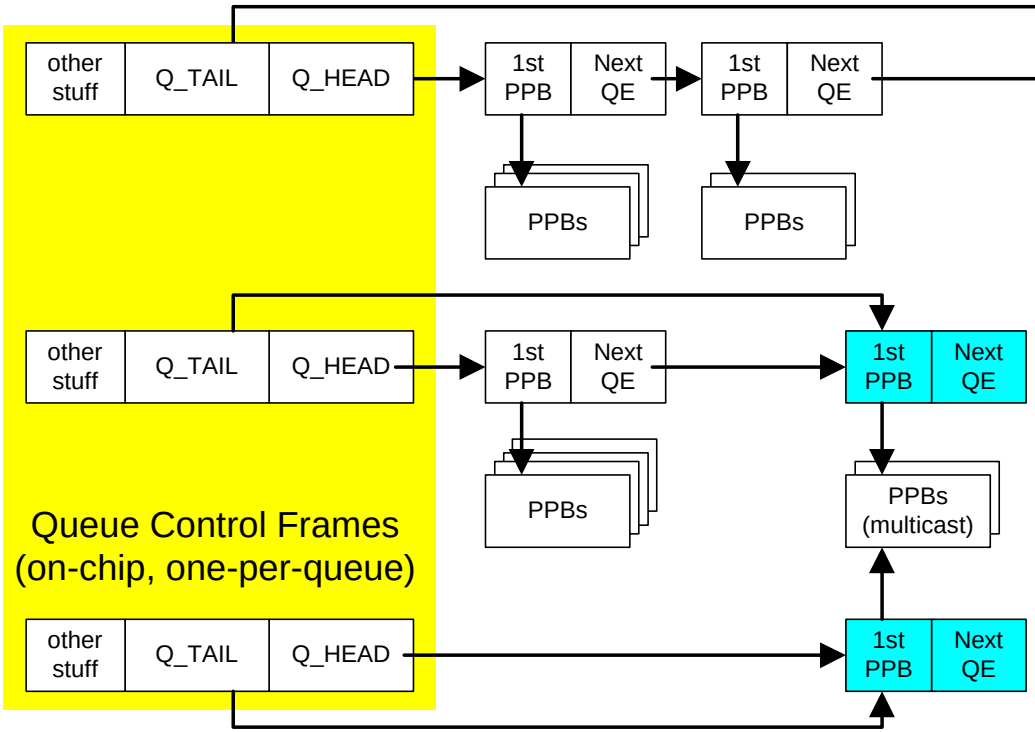
(4): Classify packets into output queues

(5): Issue QA to EQM if queue not empty



ITRI
CCL

Multicast Packets: Duplicate QEs



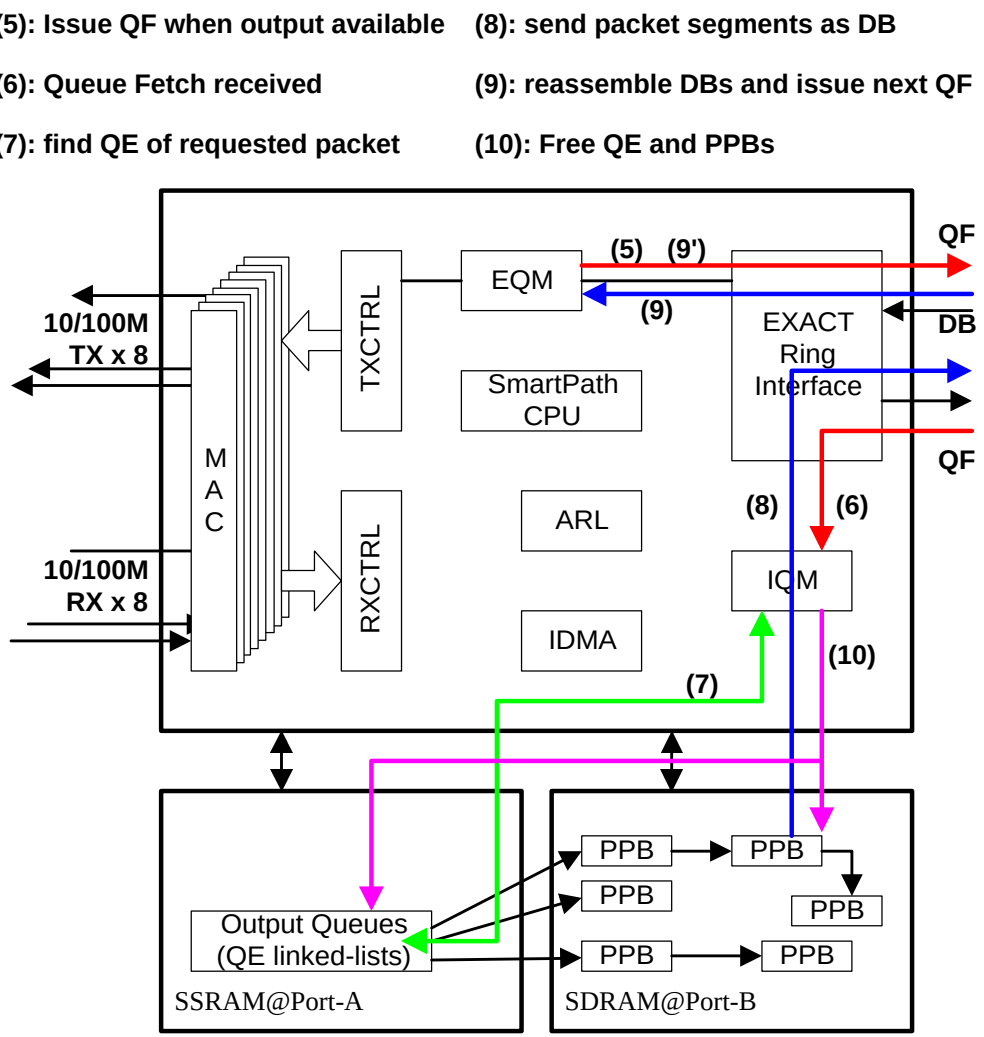
PSIZE (8)	STAT (4)	L	M	P	CHANS (6)	QID/GID (10)
TX_STAT (16)					RX_STAT (16)	
TIMESTAMP (16)					PINDEX (6)	**REF_COUNT (10)**
PAYLOAD (up to 240 bytes)						
RESERVED (16)					NEXT_PPB (16)	

PPB format

- Multicast is achieved by duplicating Queue Elements (QE).
- PPB is released after transmission if REF_COUNT =1.

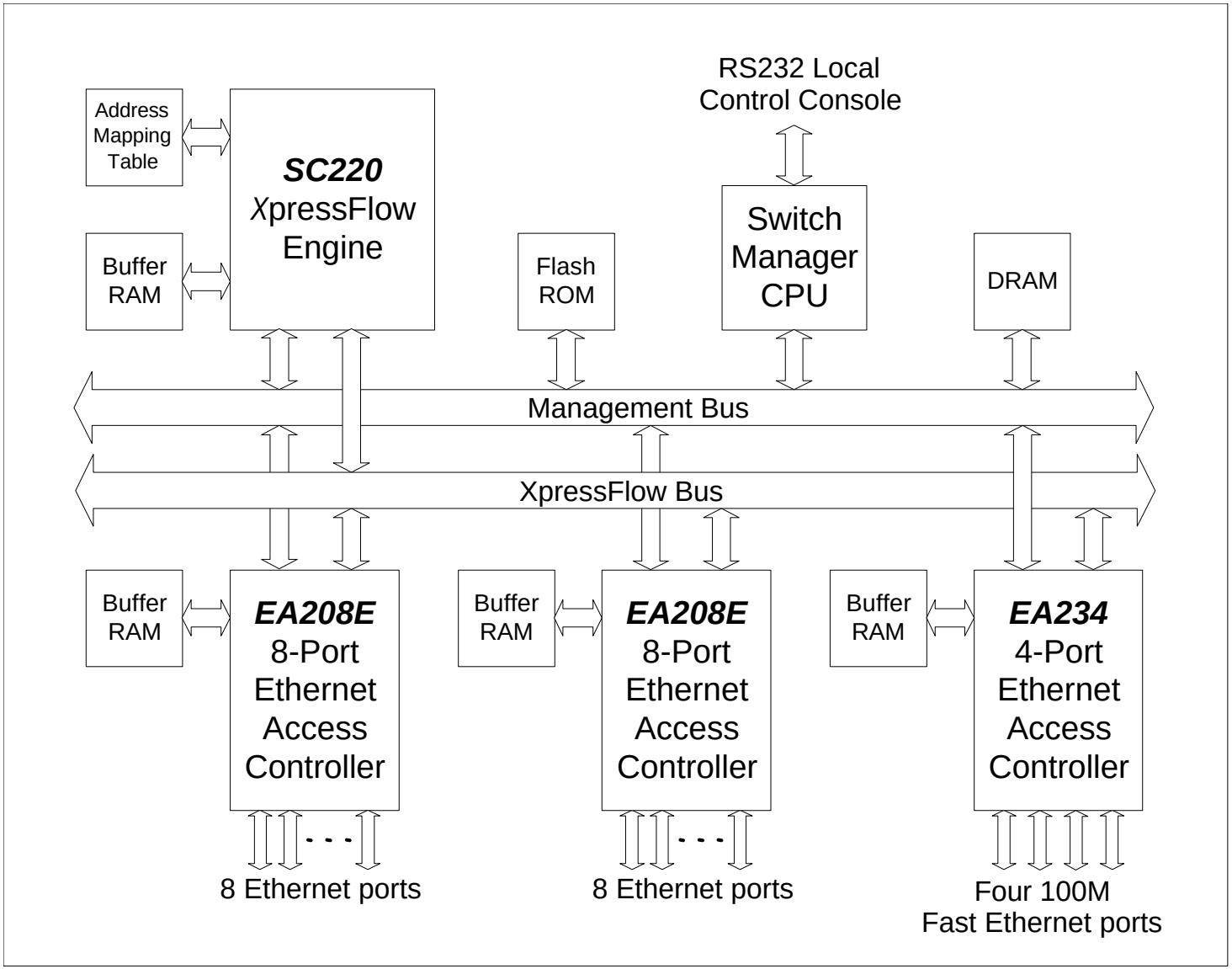
"PULL-ed" by Output

- When output FIFO becomes available for pending requests, EQM issues Queue Fetch (QF)
- When received QF, IQM response with Data Block (DB).
- Packets are "pulled" from input by output.
- IQM frees PPBs to IDMA after packets are delivered (REF_COUNT=1) or discarded.
- Data Flow:
  QA - QF - (DB - QF)^{n} - DB_{last}



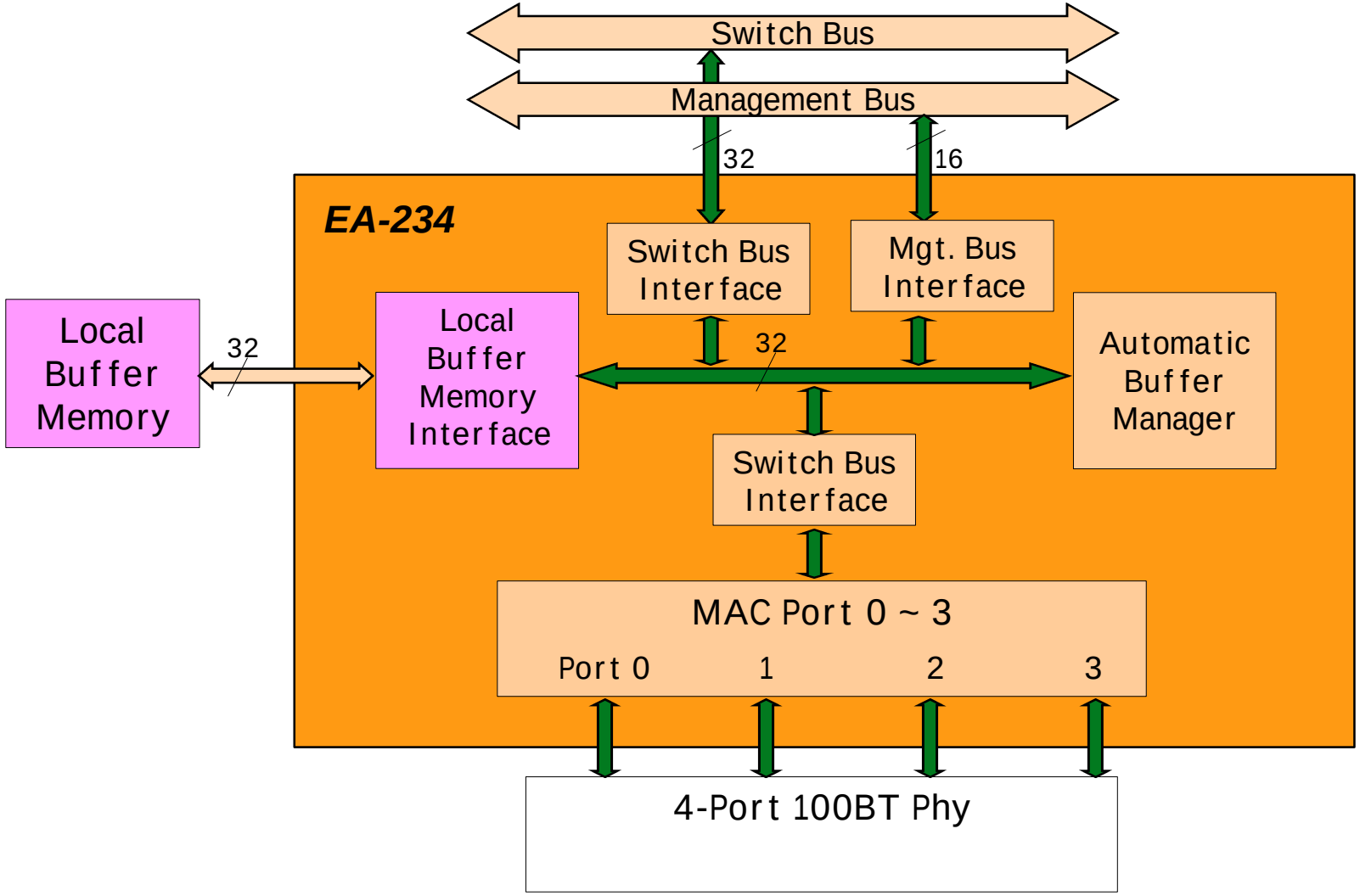
ITRI
CCL

Vertex L3 Switch Architecture



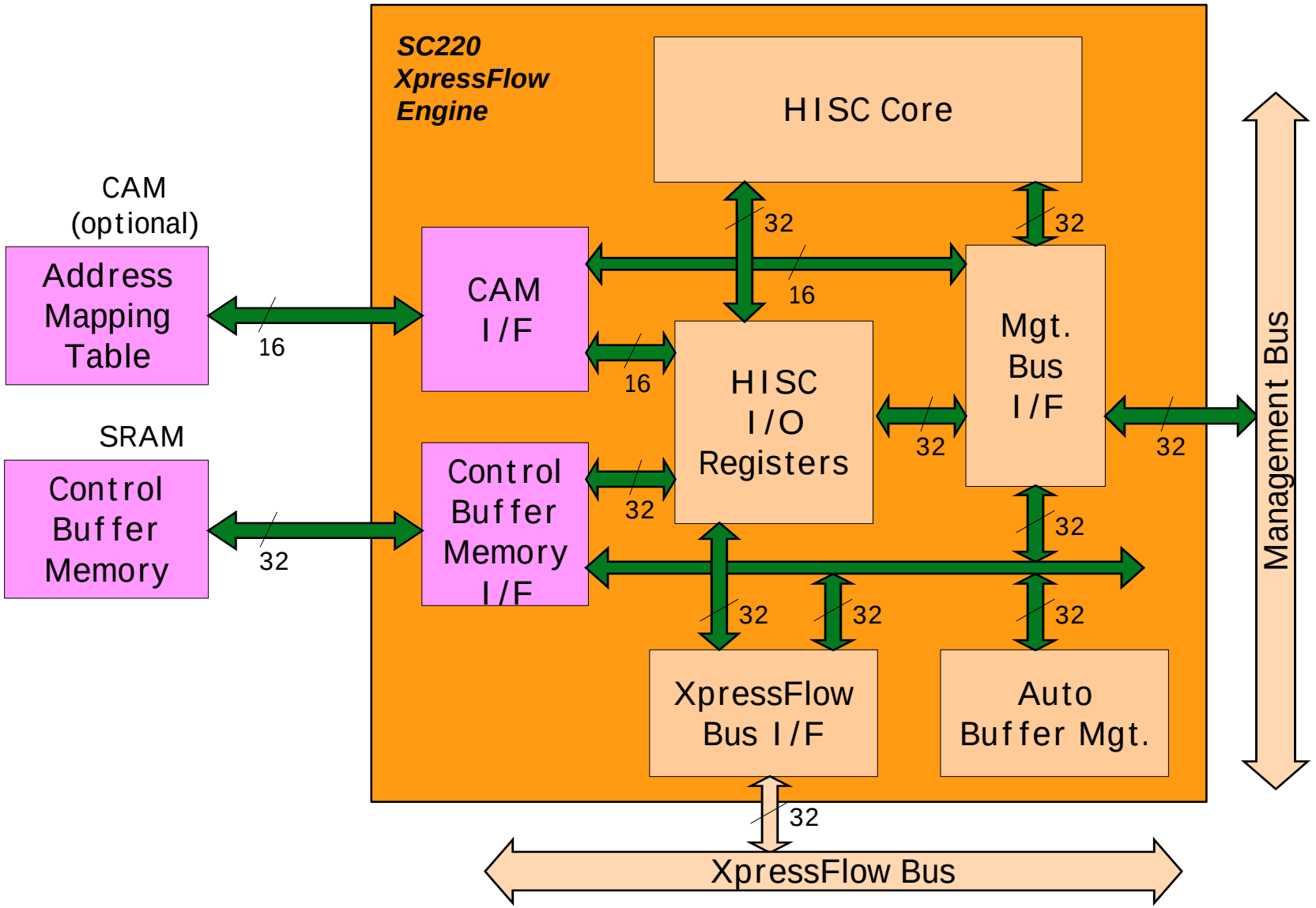
ITRI
CCL

Vertex L3 Access Controller

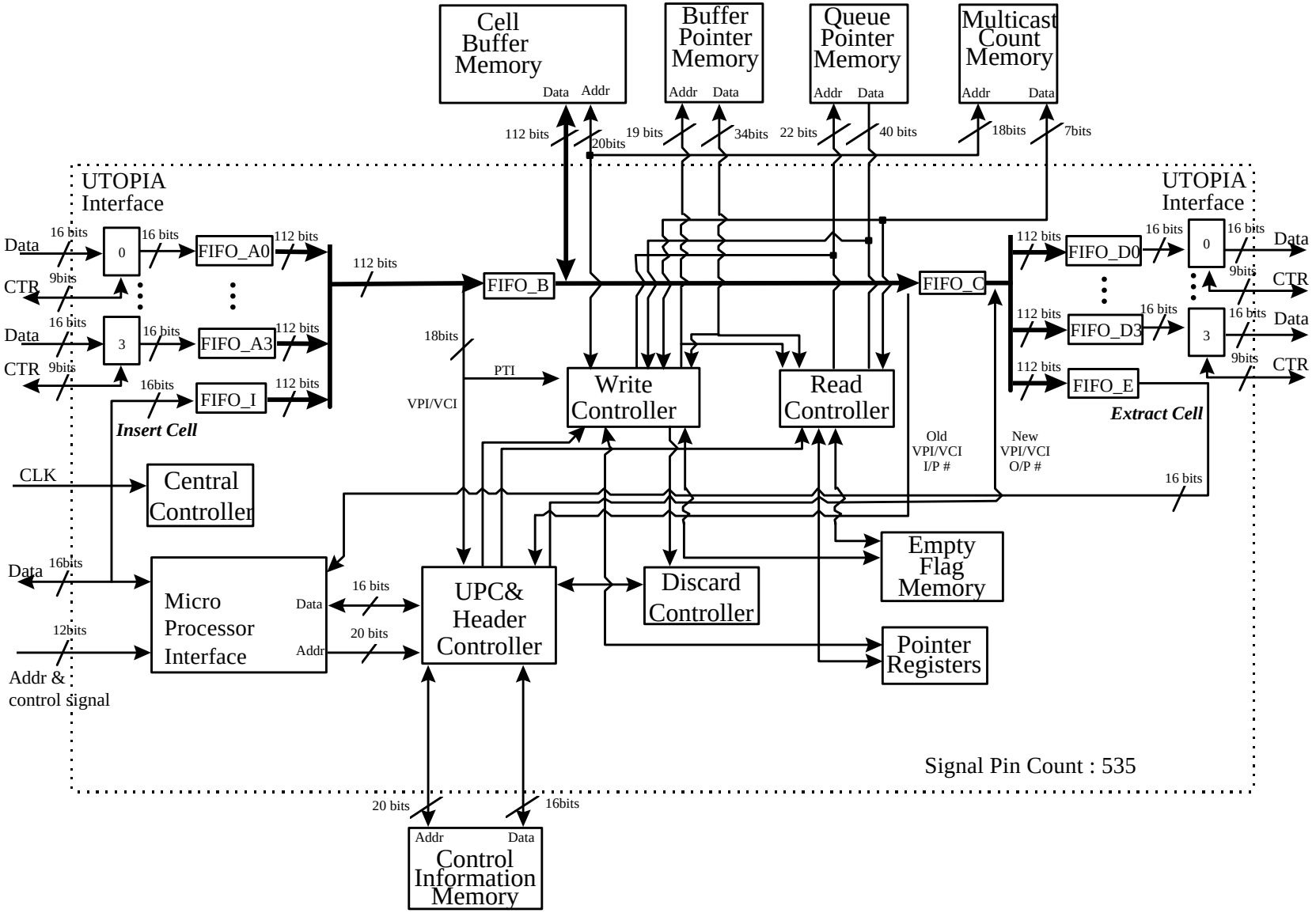


ITRI
CCL

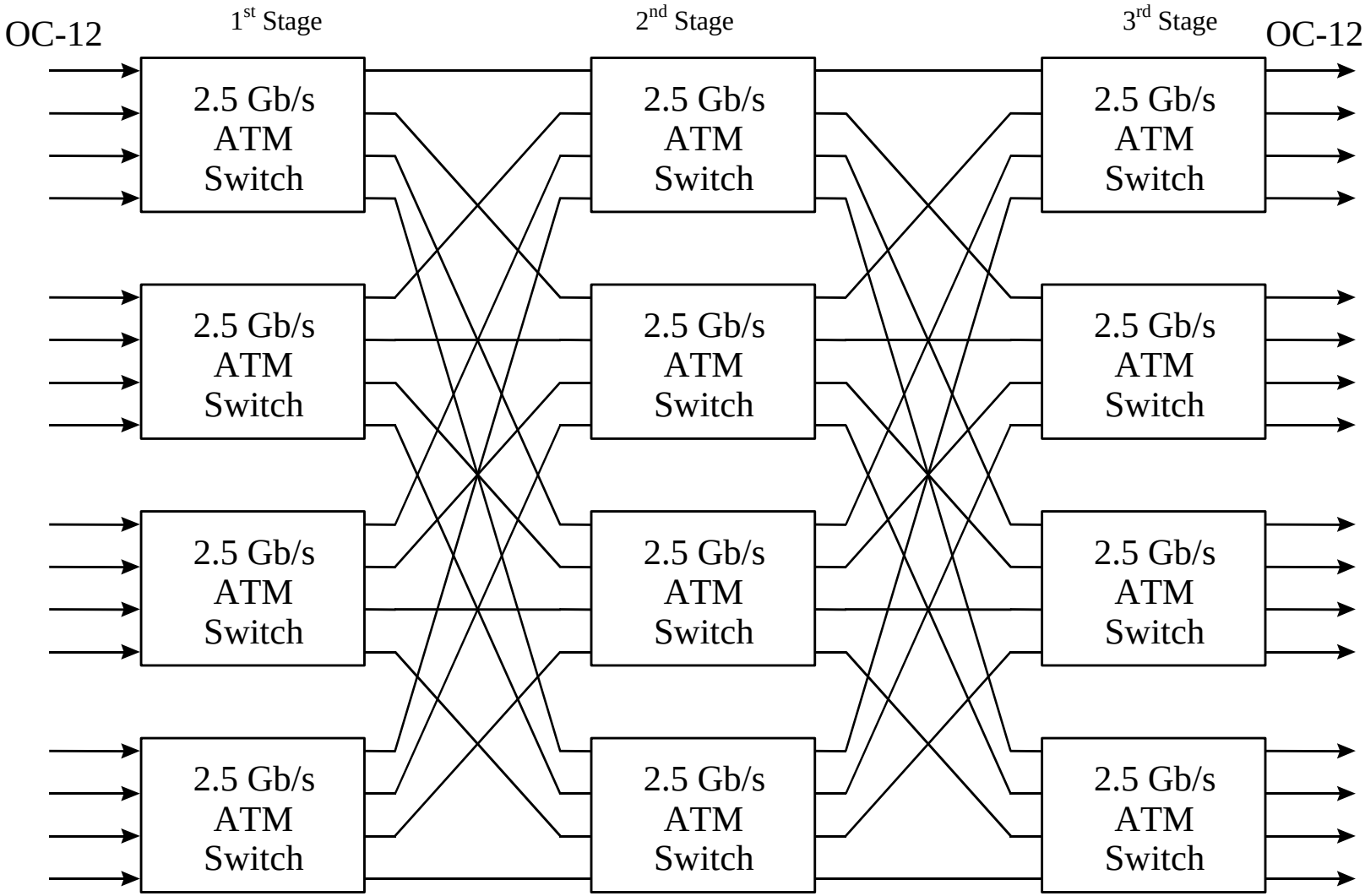
Vertex XpressFlow Engine



SMAS Chip Architecture

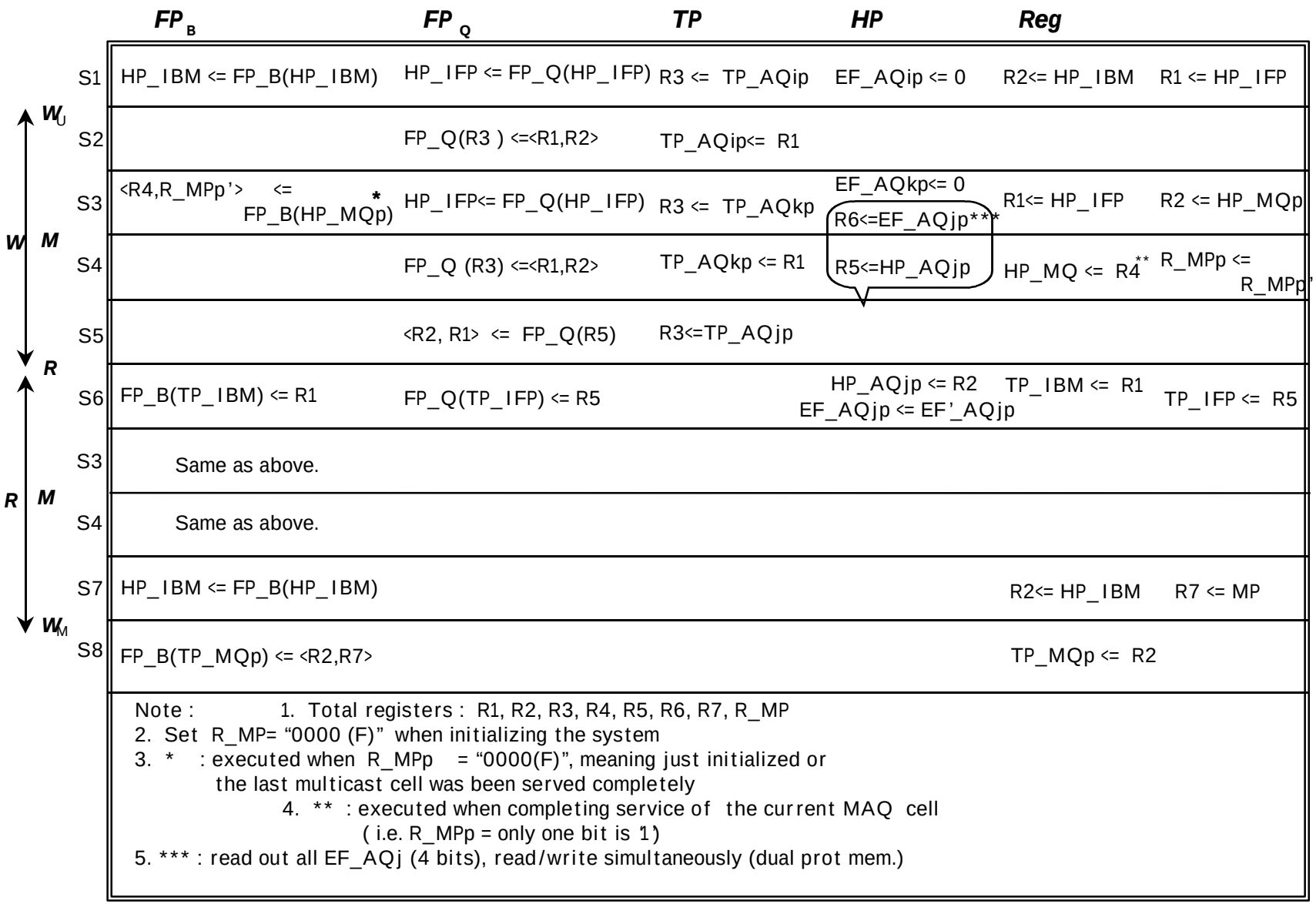


SMAS Switch

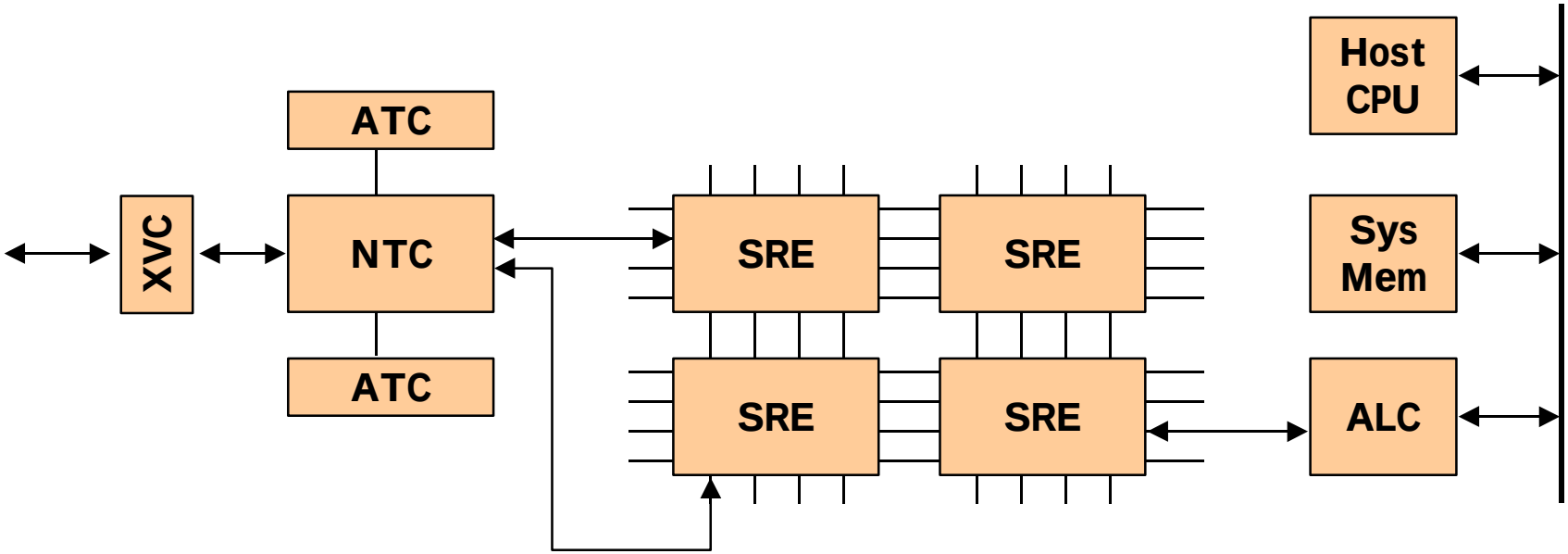

OC-12
1st Stage
2nd Stage
3rd Stage
OC-12
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch
2.5 Gb/s ATM Switch

ITRI
CCL

SMAS Timing Diagram

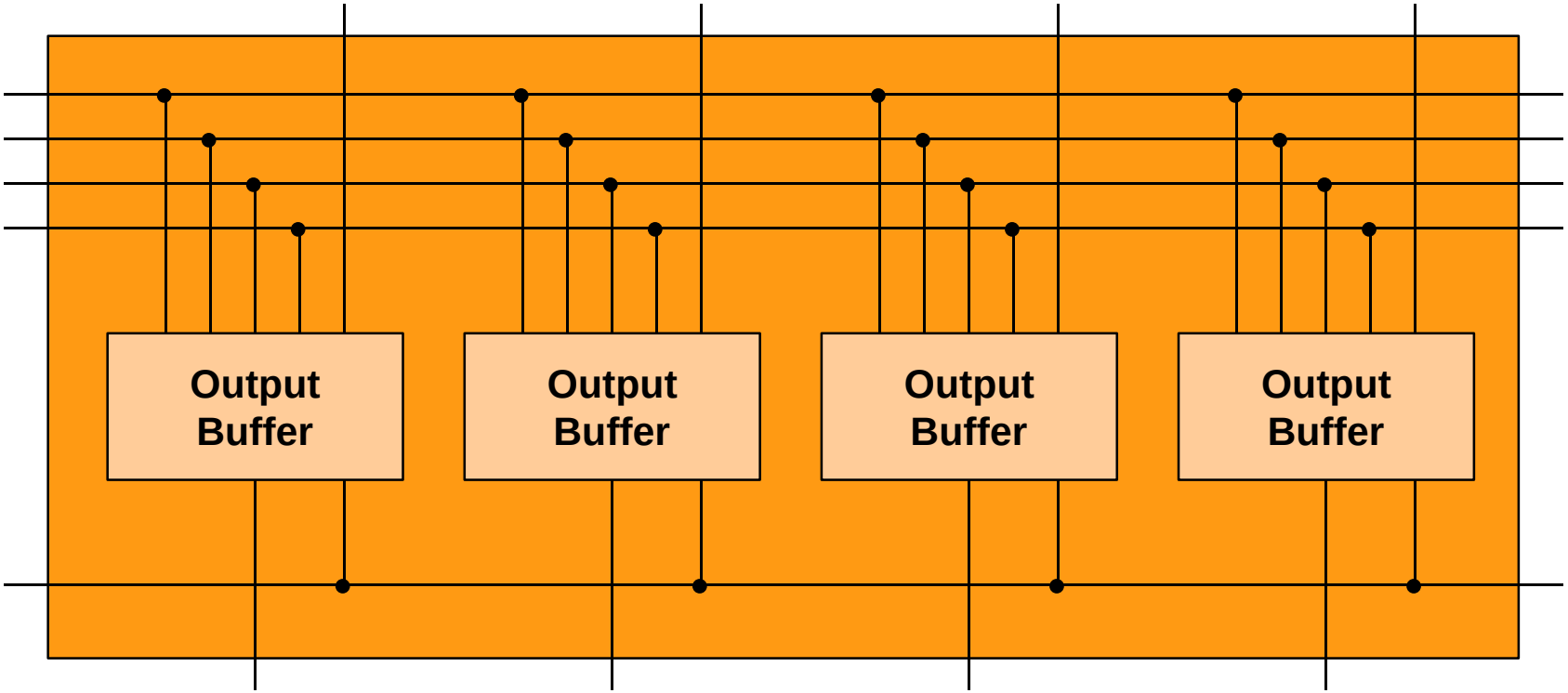


Fujitsu Chipset



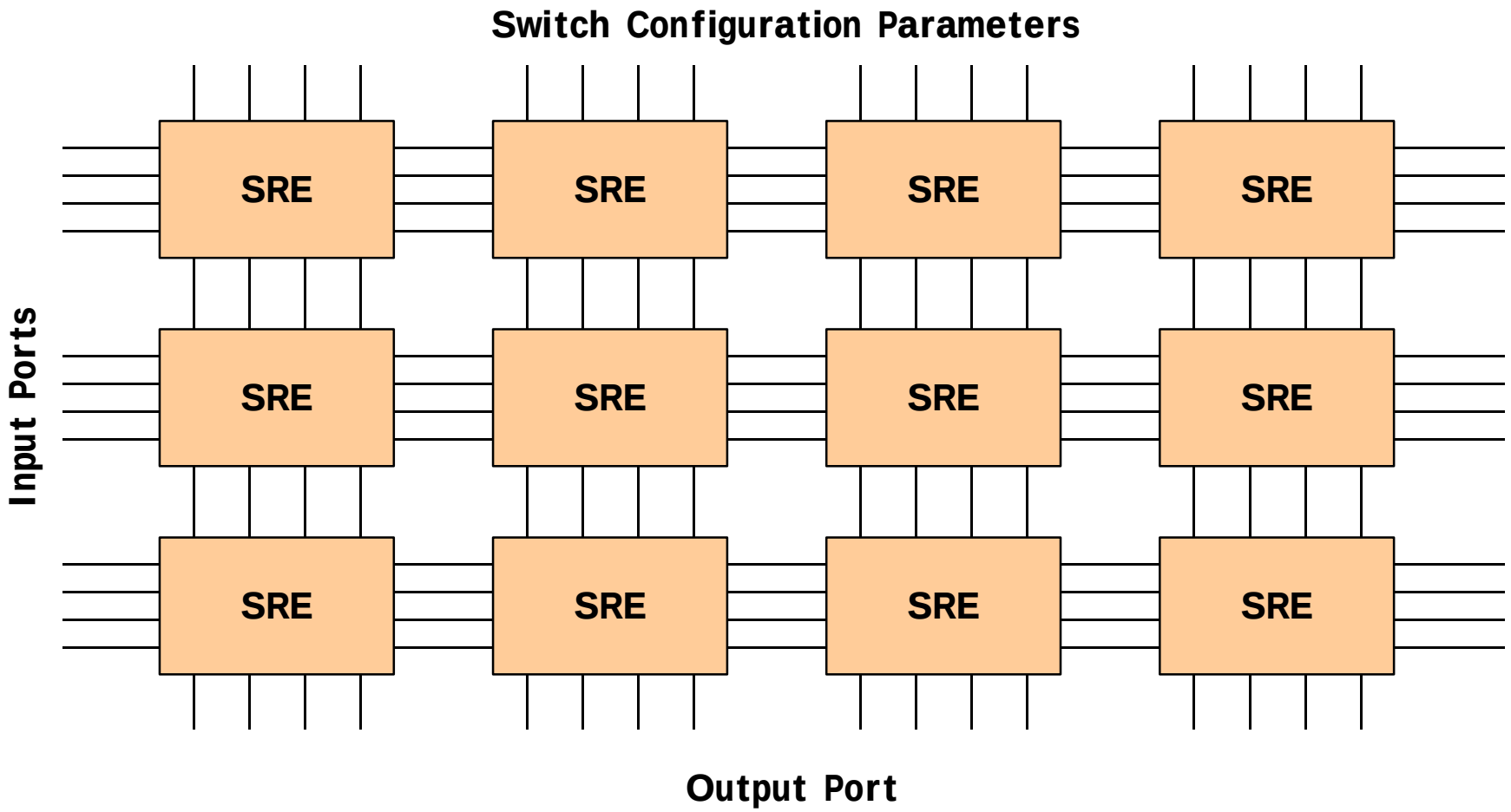
SRE	Self-Routing Switch Element
NTC	Network Termination Controller
ALC	Adaptation Layer Controller
ATC	Address Translation Controller

SRE Block Diagram



- **ATM Switch Element SRE (Self-Routing switch Element)**
  - 4x4 155 Mbps cell switch building block
  - Selectable high and low priority queues
  - Output queued for nonblocking operation
  - Multicast support

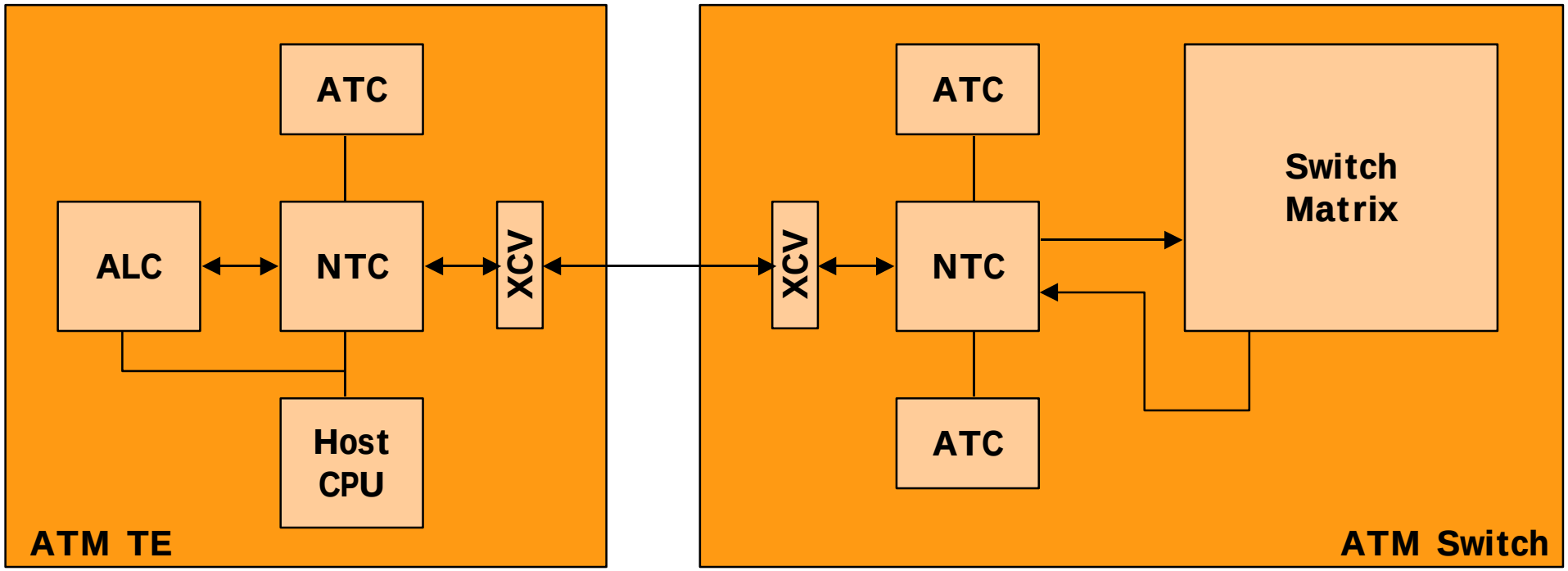
SRE Switch Matrix



*Delta Configuration allowed

ITRI
CCL

Network Termination Controller

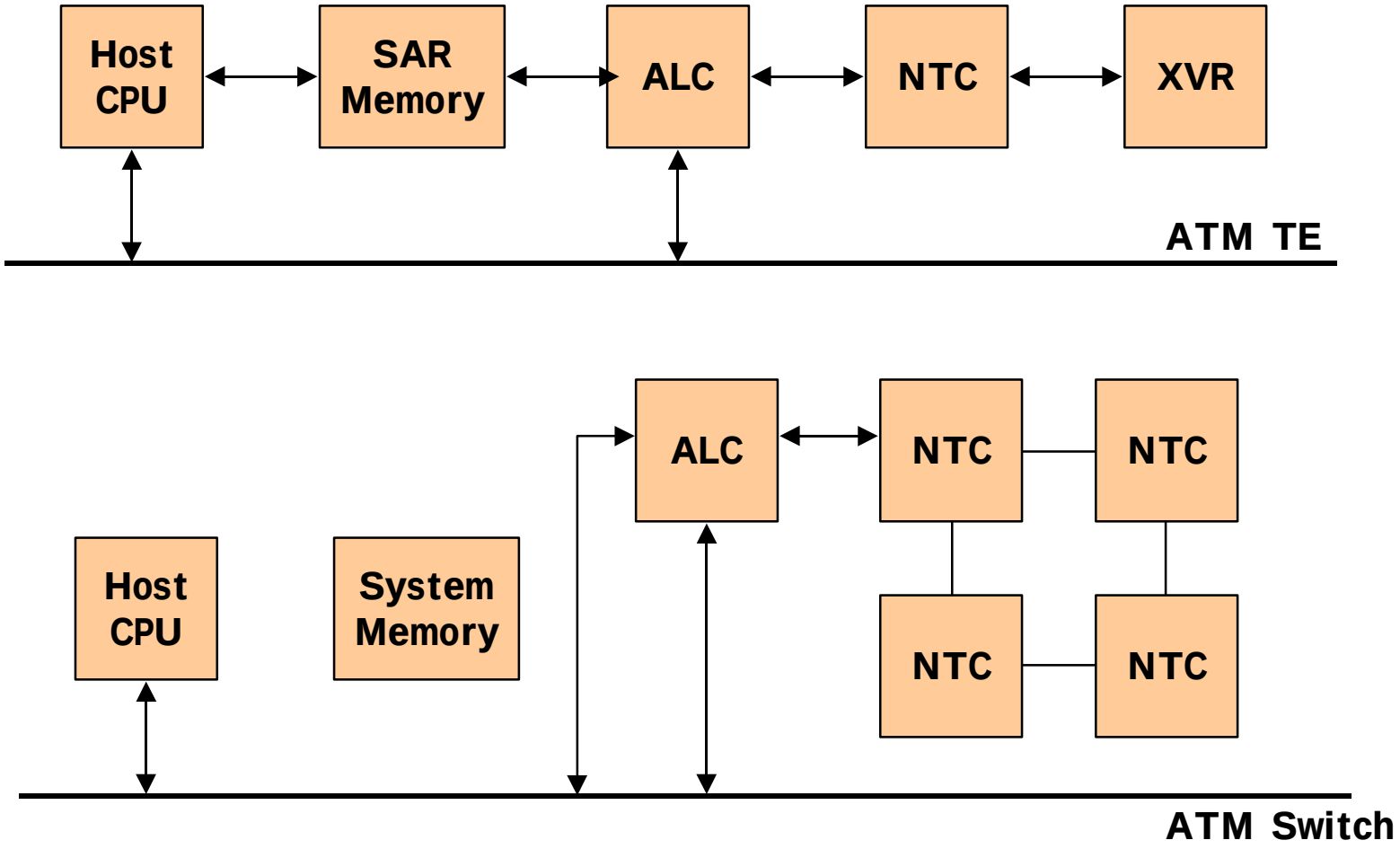


- Network Termination Controller (NTC)
  - » On-chip DMA controller for high-speed transfer of statistics to system memory.
  - » Interface to Address controller to perform real time cell header translation.
  - » Cell rate decoupling

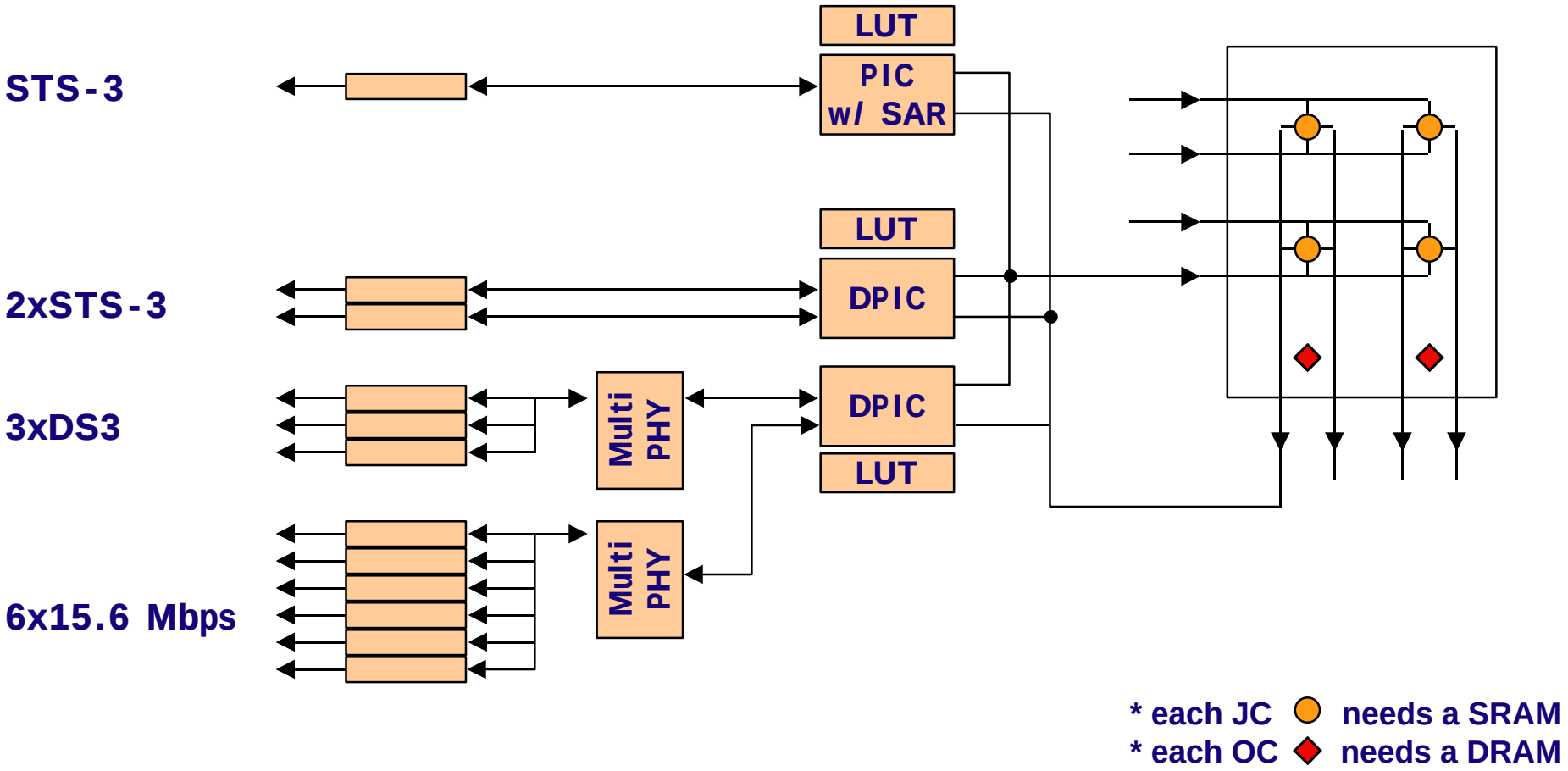
Address Translation Controller (ATC)

- Address Translation Controller (ATC)
  - » Real time translation of ATM header information up to 155 Mbps
  - » Supply 3 byte switch-internal routing tag.
  - » 1024 entry CAM
  - » Full 28 bit comparison for each entry, with optional bit-masking
  - » Supports multiple matches for multicast operation.
  - » Multiple ATCs can be cascaded to support larger addressing range
  - » Supports CLP and congestion indication/removal for each entry.

Adaptation Layer Controller



Scorpio Chipset



ITRI
CCL

Scorpio Chipset

- ATM Junction Controller (JC)
  - » Provides x-point switching and buffer mgt. for 4 junctions of a switch fabric.
- ATM Output Controller (OC)
  - » Provides cell buffer mgt. and scheduling for up to 16 junctions.
- ATM Dual Port Interface Controller (Dual PIC)
  - » Interfaces a 640 Mbps fabric channel to 2 UTOPIA interfaces.
- ATM Port Interface Controller with SAR (PIC w/ SAR)
  - » Interfaces a single 640 to a single UTOPIA interface and includes SAR functionality to allow fabric access by a host processor.
- ATM Multiple Physical Support Chip (MultiPHY)
  - » Multiplexes up to 6 UTOPIA interfaces, with an aggregate bandwidth of less than or equal to 155 Mbps to a single UTOPIA interface.

Scorpio Chipset - Cont'd

- Switch Features
  - » NxN switch fabric (640 Mbps/channel; N < 8; < 5.12 Gbps)
  - » Non-blocking architecture
  - » Distributed output cell buffering
    (combination of central and output buffering)
    - – Buffer sharing to allow for statistical variance in port buffer usage
    - – Average buffer sizes up to 64k cells per output channel
    - – Guaranteed minimum buffer space per physical port (between 2 to 64 cells).
    - – Configurable per-port limits on buffer usage to maintain fairness.
    - – Configurable per-port thresholds at which cells with CLP=1 are discarded.
  - » Full support for multicast

Scorpio Chipset - Cont'd

- Port Interface Module
  - » Support for up to 14 ports (max. aggregate capacity of 640 Mbps)
  - » ATM Layer mgt. - VP/VC label swapping
  - » SAR function integration
  - » Back pressure indication for flow control between physical ports and the SF.
  - » Multiplexes input port data onto a single Fabric channel
  - » Demultiplexes Fabric channel data to physical ports.

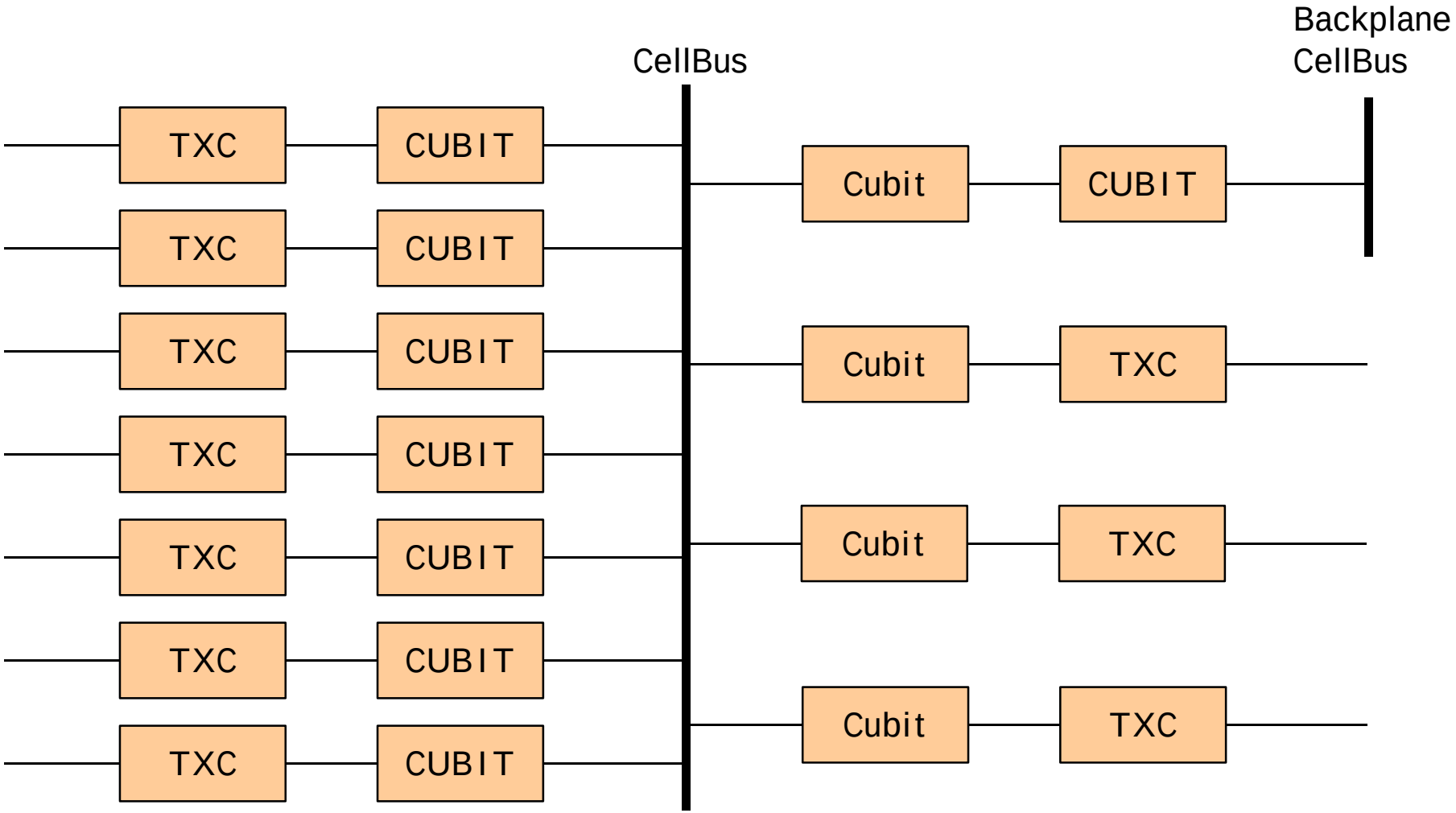
Scorpio Chipset - Cont'd

- Junction Controller
  - » Total capacity of 1.28 Gbps
  - » Ability to manage 124 individual queues (112 unicast; 12 multicast)
  - » Supports 2 x 14 physical port per output bus.
  - » Dynamic allocation of 1k to 32k of ext. cell memory among all managed queues
- Output Controller
  - » Coordinates queue management for up to 4 JC.
  - » Provides multicast functionality
  - » Provides cell transmission scheduling for all queues based on a combination of weighted round robin and strict priority algorithms.

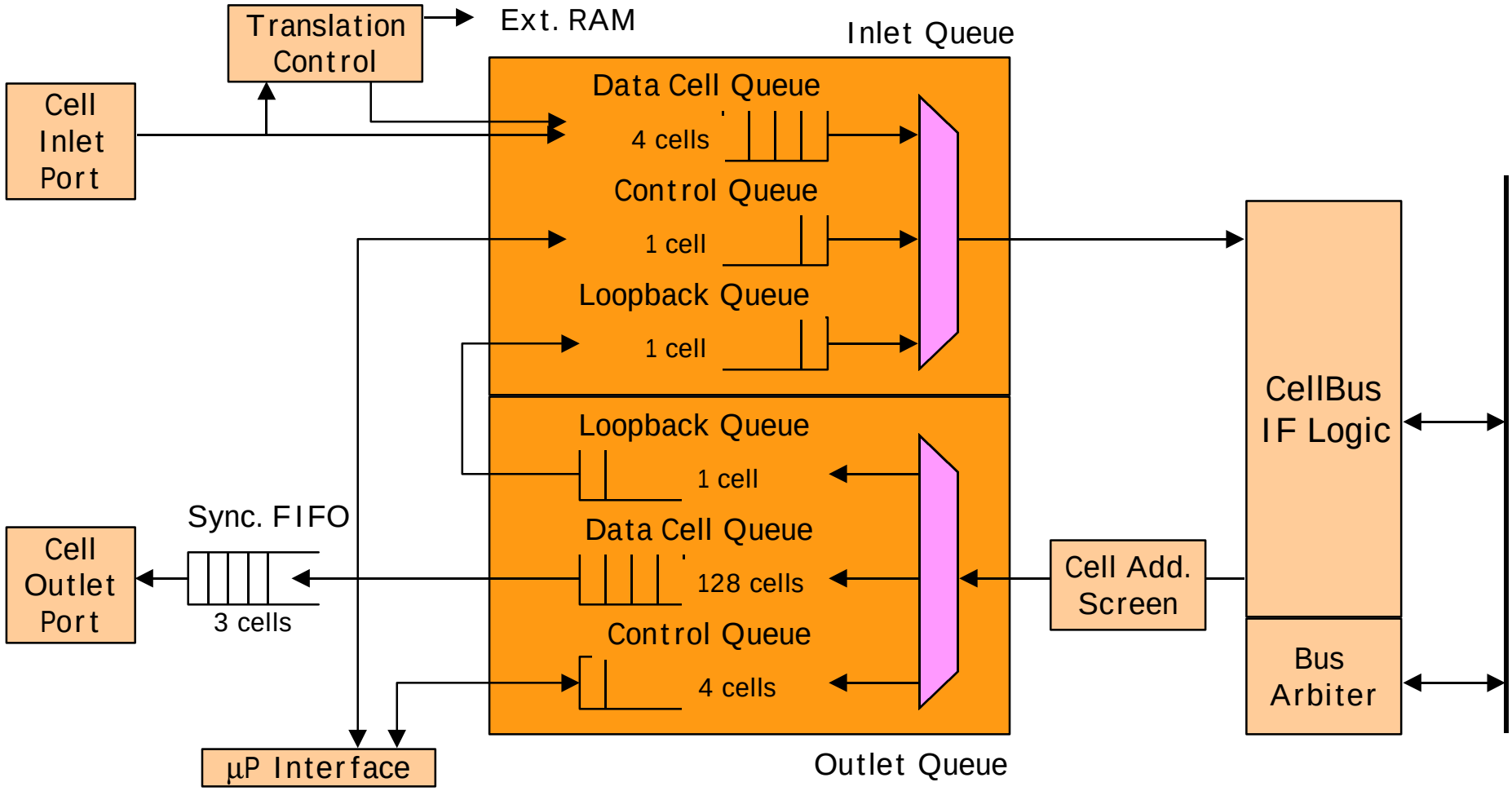
Scorpio Chipset - Cont'd

- Port Interface Controller
  - » Interfaces a 640 Mbps fabric channel to two (for dual) UTOPIA devices
  - » Provides VPI/VCI mapping
  - » Provides cell drop accounting
  - » SAR functionality (for PIC w/ SAR)
- MultiPHY
  - » Multiplexes/demultiplexes up to six UTOPIA devices
  - » Provides back pressure mechanism to the switch fabric by monitoring FIFOs within the physical layer devices

Transwitch - CellBus Switch



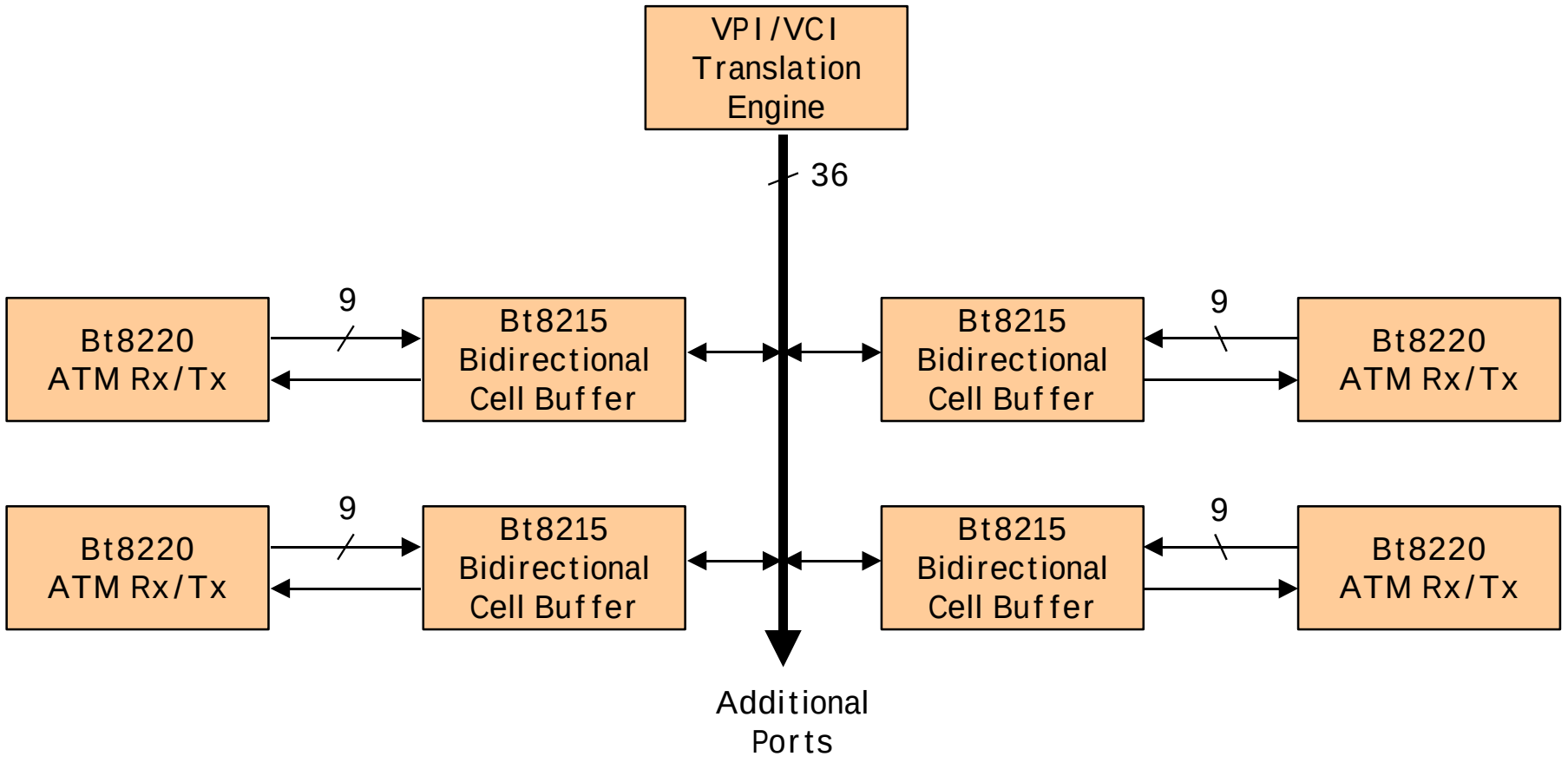
Transwitch Chipset - CUBIT



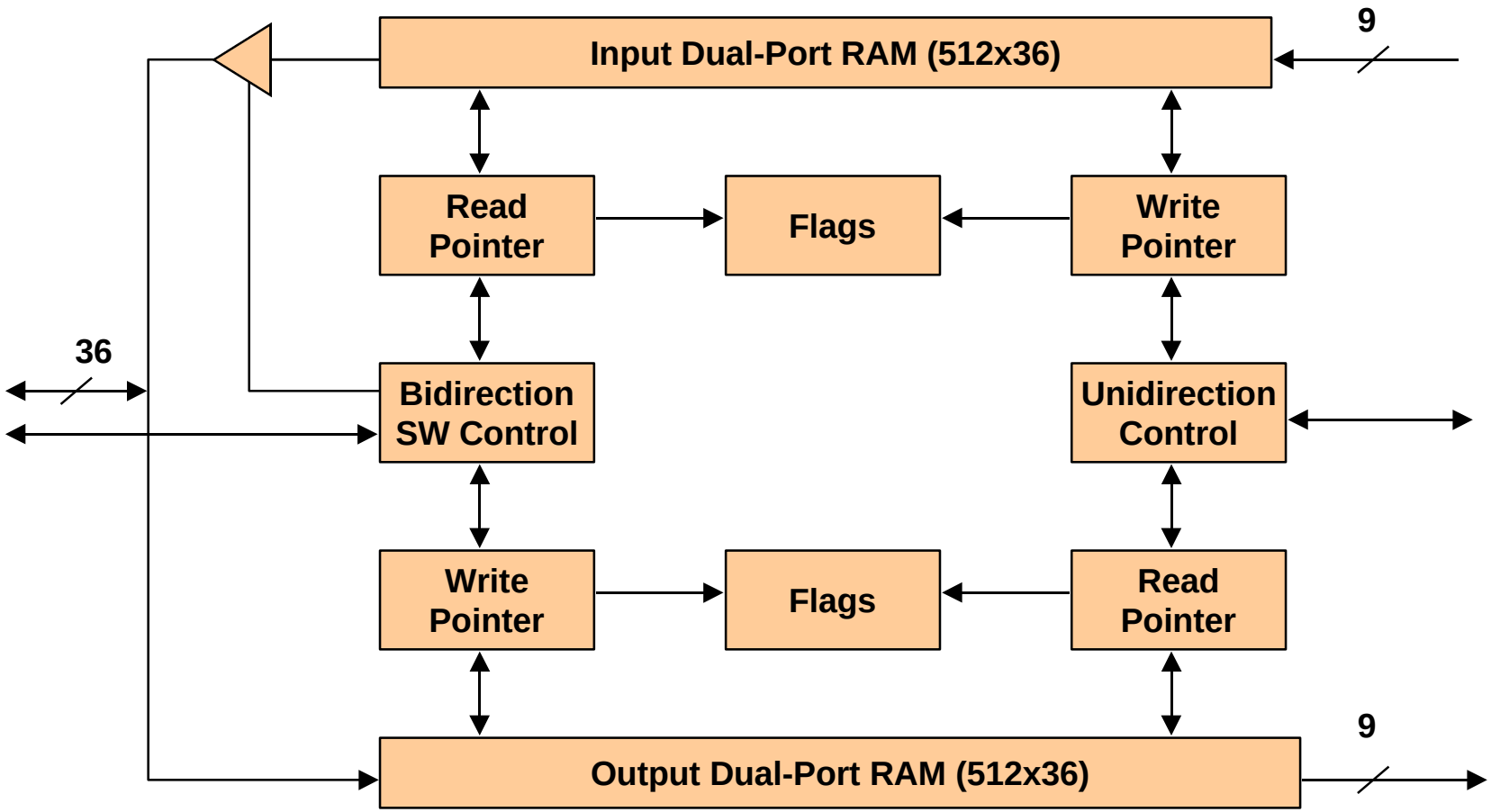
Transwitch Chipset

- CUBIT Device: Cellbus Switch TXC-05801
  - » Inlet-side address translation and routing header insertion (ext. SRAM)
  - » *Cellbus* access request, grant reception and bus transmission
  - » *Cellbus* cell reception and address recognition
  - » Outlet cell queueing; various modes
  - » Master bus arbiter included in each CUBIT
  - » Interface port to translation table SRAM
- CUBIT
  - » 37 line common bus @ 38 Mhz clock rate = 1 Gbps bandwidth.
  - » max 256 multicast sessions.
  - » 4 different service queues at output.

Brooktree Chipset



Brooktree Chipset

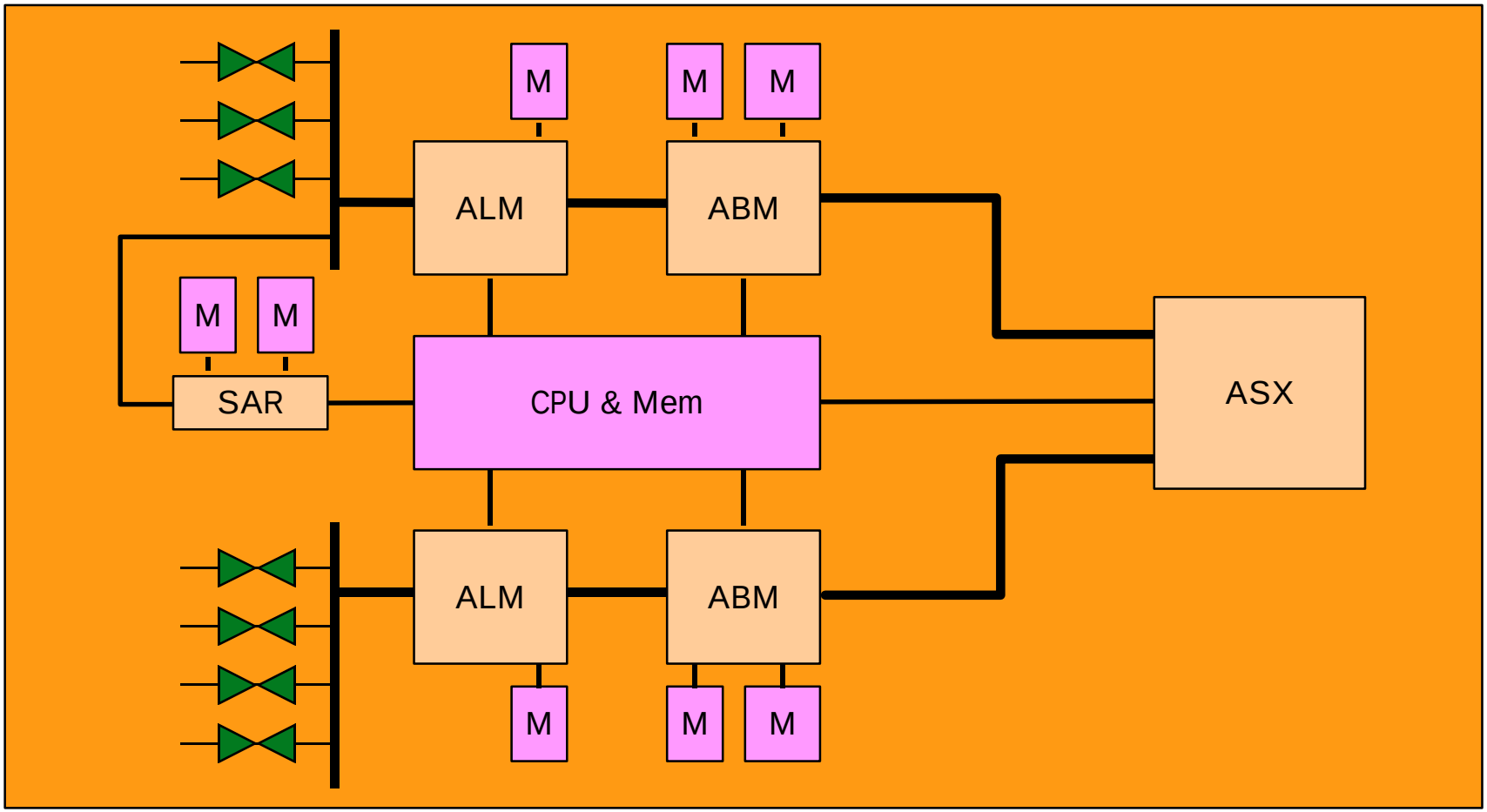


BrookTree Chipset - Cont'd

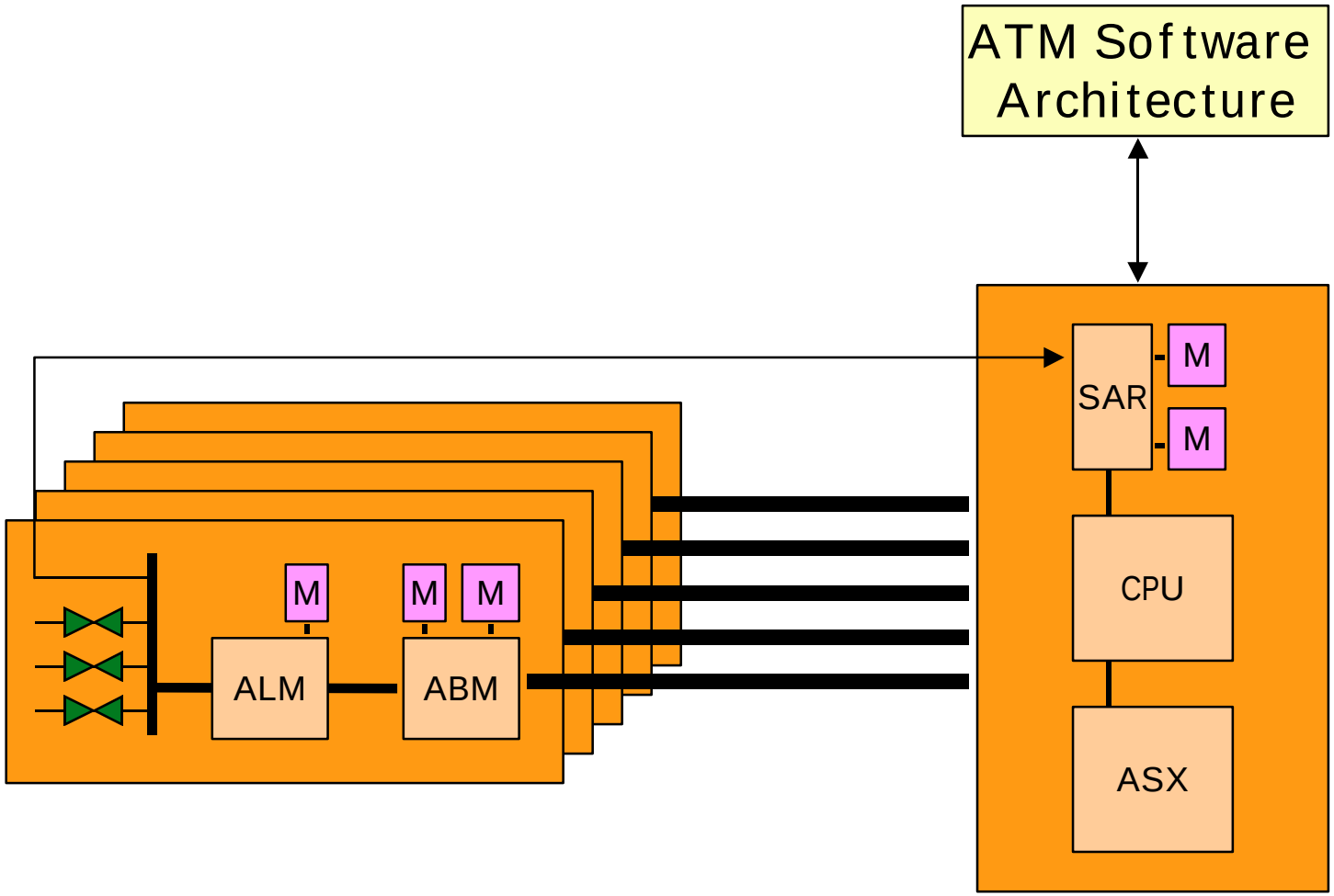
- Bt8215 Bidirectional Cell Buffer
  - » Simplifies interface between the processors and the peripherals by integrating memory and control logic.
  - » Replaces eight separate FIFO memories and associated control logic for the byte-to-word format conversion; 8 bit to 32 bit data alignment
  - » Cascade with off-the-shelf FIFOs for greater depth
  - » Supports fixed-length cell switching.
  - » 33 Mhz operation for 36 bit port; 20 Mhz operation of 9 bit port
- Additional Features
  - » 2 Bt8215s may be cascaded to form a 64-bit interface.
  - » Full, empty, almost-empty, almost-full, and half-full flags provide for buffer control.
  - » Bidirectional 36-bit port with integral parity check
  - » Separate unidirectional 9 bit ports with integral parity check
  - » 512 x 36 bit buffer memory in each direction
  - » Synchronous or asynchronous interfaces on all ports

ITRI
CCL

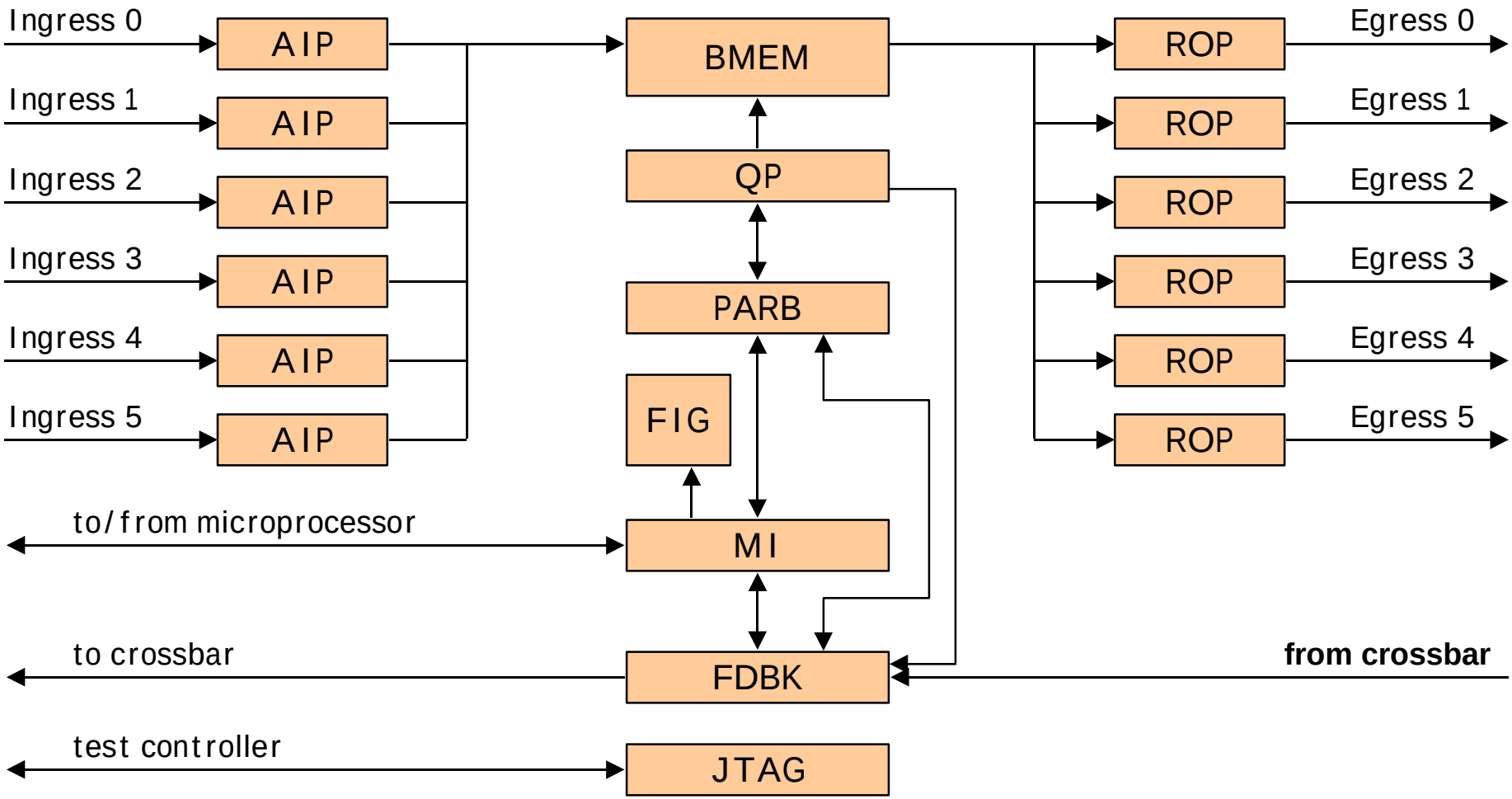
AT&T System Architecture (1)



AT&T System Architecture (2)



ATLANTA - ASX


Ingress 0
Ingress 1
Ingress 2
Ingress 3
Ingress 4
Ingress 5
AIP
AIP
AIP
AIP
AIP
AIP
BMEM
QP
PARB
FIG
MI
FDBK
JTAG
ROP
ROP
ROP
ROP
ROP
ROP
Egress 0
Egress 1
Egress 2
Egress 3
Egress 4
Egress 5
to/from microprocessor
to crossbar
from crossbar
test controller

ITRI
CCL

ATLANTA Chipset - ASX

- Switching capacity ~ 6x6 622 Mbps (3.7 Gbps total bandwidth)
- Integrated internal buffer (shared) ~ 512 cells (No external buffering necessary)
- Scalable
  - » 18x18 622 Mbps (11 Gbps total bandwidth) on a single fabric card
  - » 36x36 622 Mbps (24 Gbps total bandwidth) on a dual fabric card
  - » Support up to 1080x1080 UNI / NNI ports
  - » Redundant fabric possible
- Functionality
  - » 4 WRR (16 programmable weights) delay priorities/port, each with 2 loss priorities and configurable dynamic thresholding
  - » Smart backpressure specific to congested fabric port and priority level
  - » HOL blocking avoided via flow control
  - » Multicasting (only one cell copy)

ITRI
CCL

ATLANTA Chipset - ABM

- Scalable buffer size up to 32k cells/fabric port
- 4 separately configurable WRR (16 programmable weights) delay priorities/port, each with 2 loss priorities and configurable dynamic thresholding
  - » ensures higher priority sub-queues served more frequently
  - » prevents starvation of lower priority levels (i.e. guaranteed some BW sharing)
  - » If empty or blocked, serves highest, non-empty, non-backpressured delay priority
- HOL blocking avoided via flow control
- "Smart" backpressure from fabric to ingress and (w/ override) from egress to fabric
- Rate scheduling for all subports on output (1.5 to 622 Mbps)
- Provides multicasting to all subports

ITRI
CCL

ATLANTA Chipset - ALM

- ATM Layer functions
  - » Up to 30 physical UNI/NNI supported via MultiPhy Utopia II interface
  - » Up to 32k VCs (in/out-bound) per fabric port
  - » VPI/VCI translations, optional HEC check & header correction.
- Traffic Management
  - » per-VC configurable dual-leaky bucket UPC
    (scalable granularities: 64 kbps to 622 Mbps, with ~ 0.1% steps)
  - » Cell monitoring: CLP0, CLP1, CLP0+1
  - » Parameter monitoring: PCR, CVT, SCR, BT.
  - » Extract/Insert OAM cells with optional routing to/from local µP interface.
- per-VC statistics collection (6 counters)

ITRI
CCL

ATLANTA - Memory requirements

- VC Tables: ~ 2200 VCs per 1 Mbit SRAM
  - » VC connection parameters, Traffic statistics, Policing parameter
  - » w/o policing ~ 4000 VCs per 1 Mbit SRAM
  - » w/o policing & statistics ~ 8000 VCs per 1 Mbit SRAM
- Cell Buffer: ~ 2k cells per 1 Mbit SRAM
  - » Minimum 4k cells needed w/ 2 units of 32k x 32 SRAM
  - » Does not include associated pointer space
- Example: 4 STS--3c ports on a line card
  - » ~ 8k VCs (no policing): 2 Mbits
  - » 4k cell buffer: 2 Mbits
  - » Pointer space: 1 Mbits