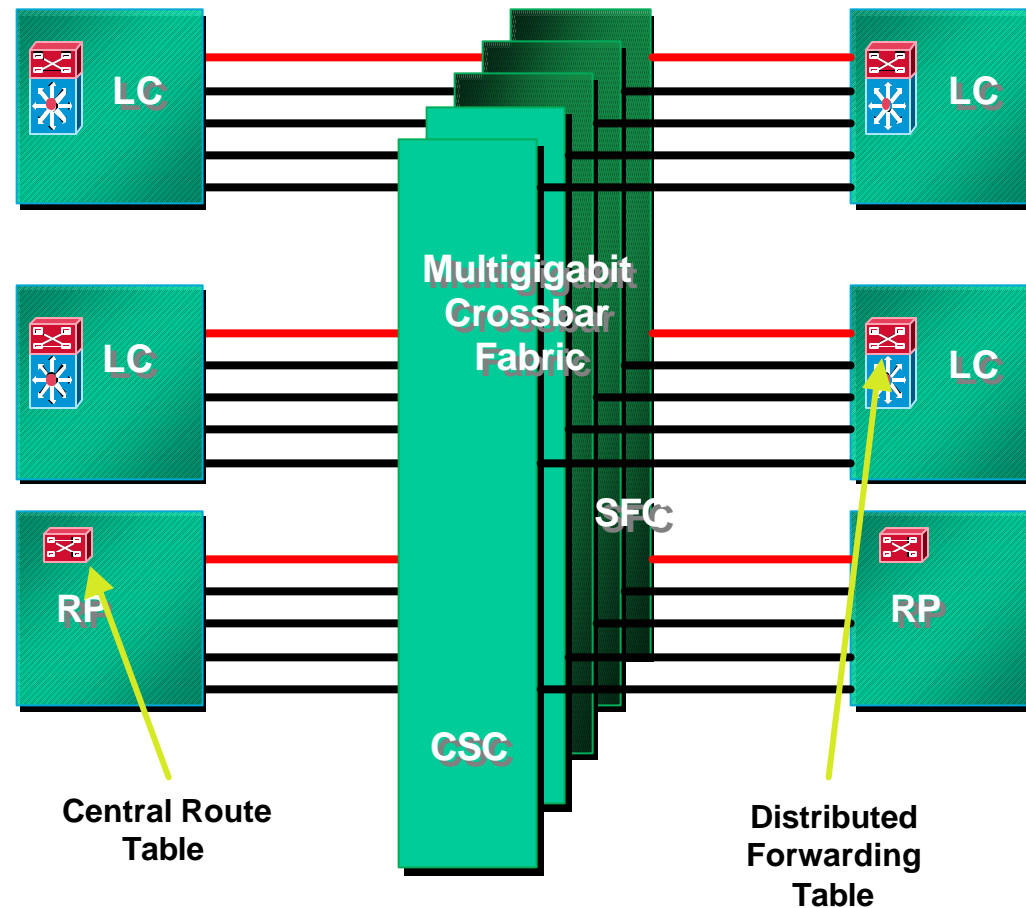


Notes on C12000 Queuing Architecture
for
MDRR testing

12000 GSR Architecture (Diagram: from Cisco)

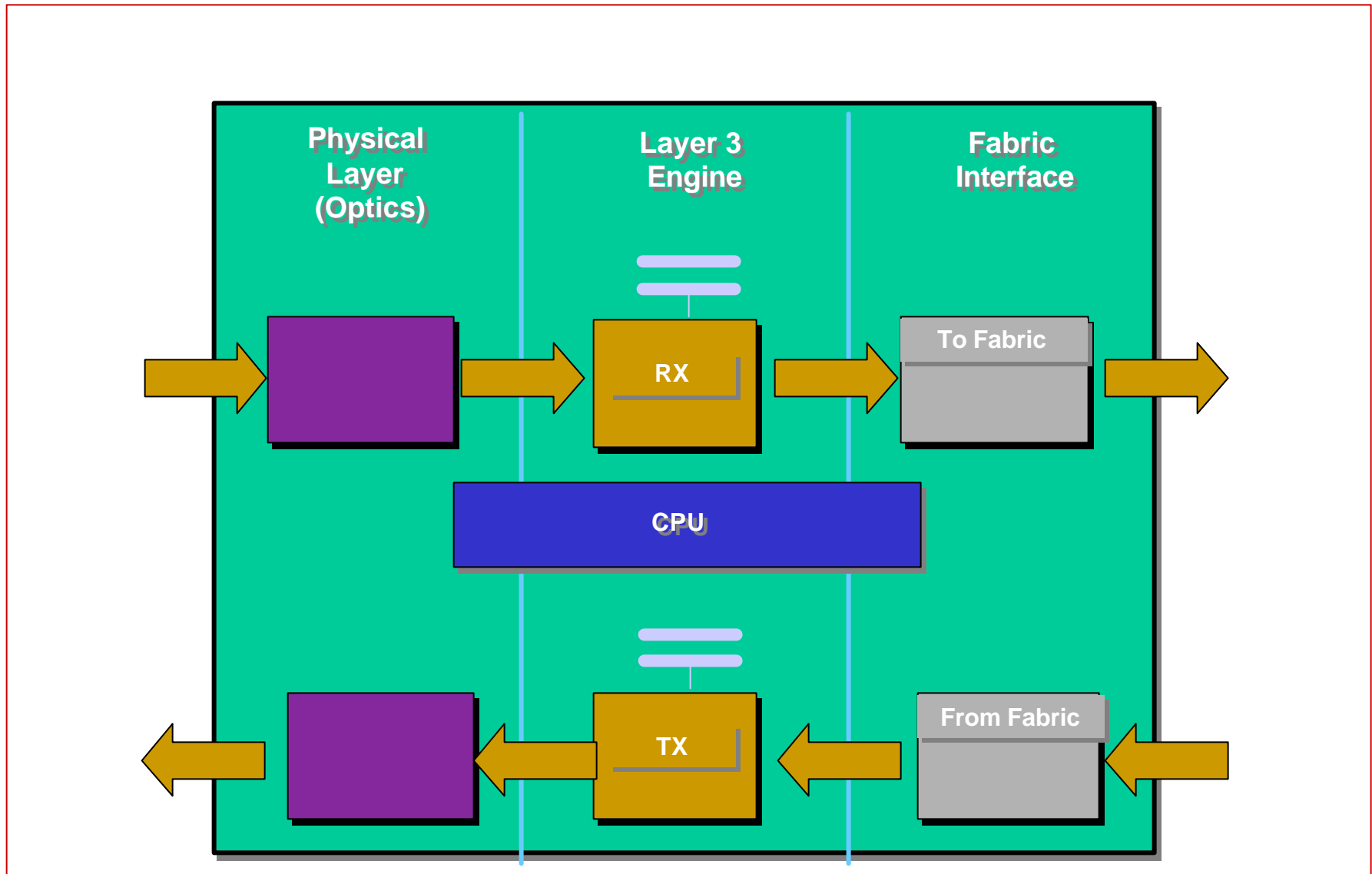


- SFC: Switch Fabric Card // 3 SCS + 1 CSC for full bandwidth
- CSC: Clock Scheduler Card // clocking + SCS
- With max configuration for SFCs and CSCs, 5 Gbps per line card (8B/10B)

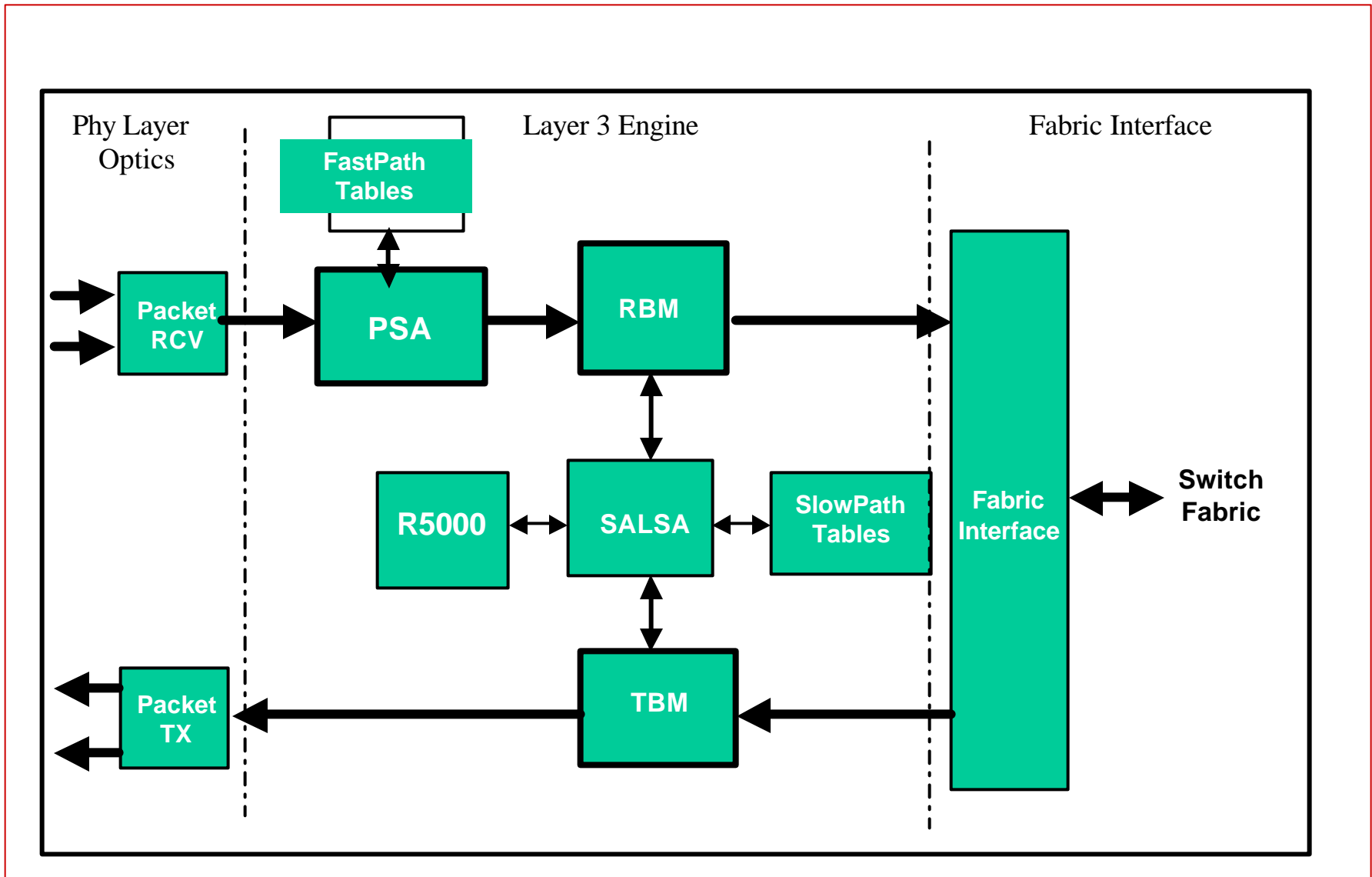
Advantages

- Crossbar switch:
 - Communication between LCs through **point-to-point links**
 - Crossbar configuration defined by a **centralized scheduler**, which ensures that each input is connected to at most one output and vice versa, in a fair way
 - **Synchronization** of transmission on each bus by dividing time into fixed length **time slots**
- Advantages:
 - **Internally non blocking**: it allows all inputs and outputs to transfer packets simultaneously (i.e. parallel transmissions on different point-to-point links through the closing of several crossbar crosspoints)
 - **Minimization of delay** introduced by the contention present in shared buses

12000 GSR Line Card Architecture (Diagram: from Cisco RST-301)



12000 - LC - engine 2 – cont (Diagram: from Cisco RST-301)



12000 - LC - engine 2

- Components:

PSA: Packet Switch ASIC

- (logic + memory) for IP packet analysis
- IPv4 protocol, known MAC encapsulation format, no IP options, correct IP checksum, and the Multi-Protocol Label Switching (MPLS) label
- decision on fast-path function routing (in hardware), if not, packet is stored in RBM and sent to local processor

Fast Path Route Tables: produced by GRP and downloaded to line cards

Slow-path Route table:

- to perform the routing operation on any packets not processed by the packet switch manager.
- stored in the line card *processor memory* (configured with 256 MB of parity-protected memory)
- line card *processor memory* also contains the local *program image* and *data structures* for the line card processor.

SALSA: helps CPU with label lookup

12000 - LC - engine 2 (cont)

- Buffering:

RBM/TBM: rx/tx buffer manager. logic and memory that supports buffer management functions.

RBM:

- each packet received by line card written to a buffer in the receive buffer memory
- Each packet received by the line card written to a buffer *consecutively*. The buffer is released only when all **16K** have been completely allocated and then another contiguous buffer can be allocated.

TBM: logic and memory that supports transmit buffer manager functions

- Each cell received from the switch fabric is written to a buffer in the SDRAM transmit buffer memory
- 256 MB Packet buffering (ECC SDRAM)
- reassembling of incoming cells before transmission
- Fixed size buffers of **16K** each

12000 - LC – engine 2

Line card processor: using the slow-path route table, performs the routing function on any packets unable to be directly processed in hardware. Functions:

- Reading and validating the layer 3 (L3) packet header.
- Modifying the appropriate fields of the L3 header.
- Determining the **appropriate output interface** for this packet (slot and port numbers).- Linking the completed packet onto an appropriate output queue in the receive buffer manager, if it is to be transmitted, or discarding the packet if necessary
- **64-byte** cells that contain header, **CRC error detection/correction, and data**

Switch Fabric Interface:

- accepts data from the receive buffer manager and **transmits it to the appropriate** destination line cards
- **makes requests** to the switch fabric scheduler for access to line cards for which it has transmit data
- **transmits** the granted data to the switch fabric.

Maintenance Bus Interface:

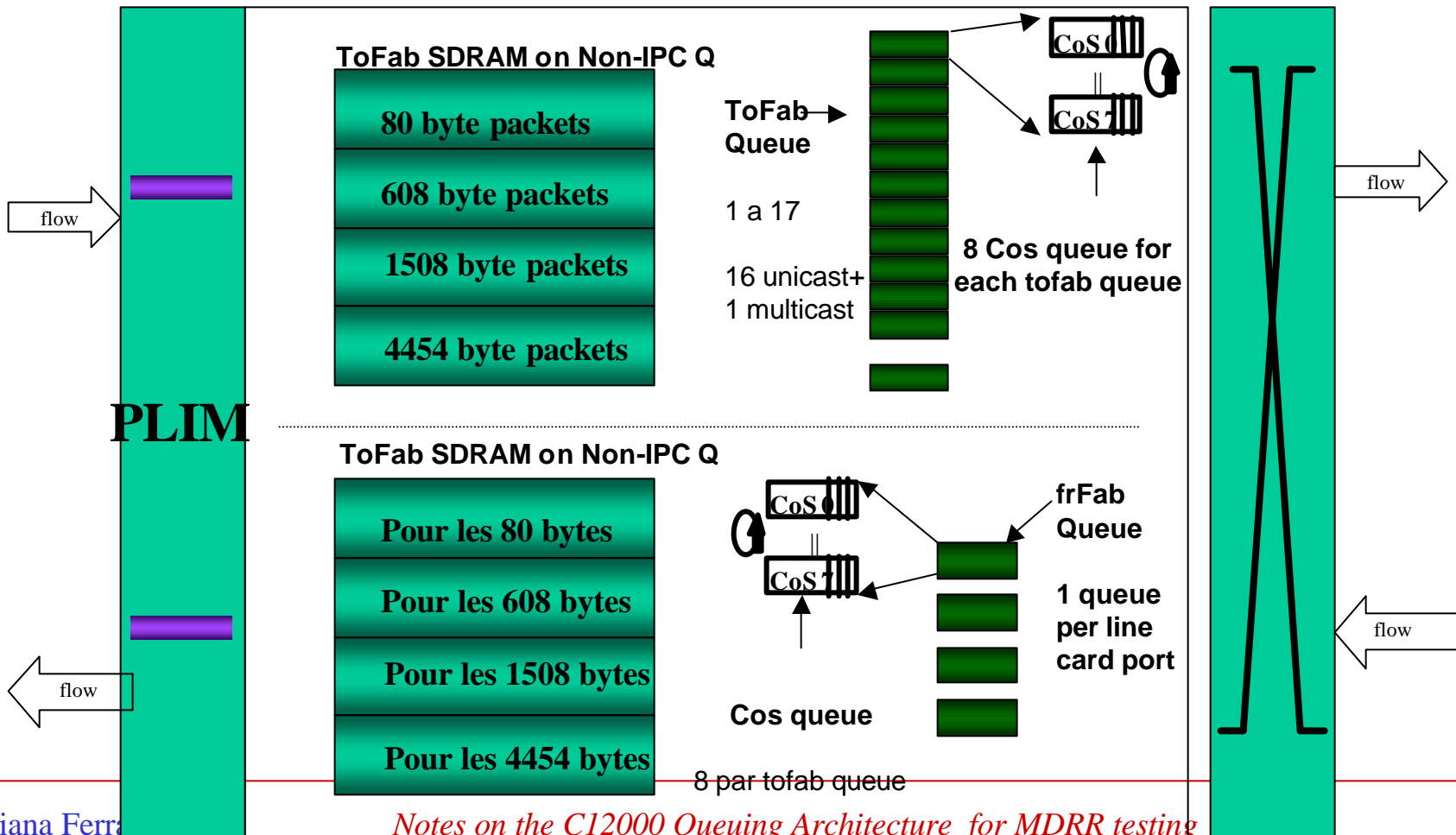
- line card communicates with the GRP through the MBus interface
- MBus interface provides a **redundant MBus controller** area network (CAN)
- communications interface for access to the GRP by means of a simple message-passing mailbox protocol.

Phases

- 1. Packet comes into the PLIM
- 2. Buffer allocation in SRAM for each input packet
- 3. CEF makes switching decision and determines next hop interface/slot from adjacency database
- 4. packet is enqueued in one of the 17 (16 unicast + 1 multicast) ToFab queues (local output queues) – WRED
- Queue manager (component of Buffer Manager) determines *CoS queue* for queuing packet

Buffer Management (Diagram: from Cisco)

- PLIM: Physical Layer Interface Module
- IPC: Inter Processor Communication
- SRAM kept separated from tx/rx buffers



Advantages

- **head of line blocking**: prevented through *virtual output queuing*, i.e. By grouping packets into different queues according to the output port they are destined to
- **Input and output blocking**: only one cell at a time can be transferred/received at a time by a crossbar bus.
 - No impact on crossbar throughput
 - Impact on packet delay ->
 - **Input blocking**: multiple packets to the same *output port*
 - Solution: packet prioritization (input)
 - **Output blocking**: different input interfaces “talking to the same output interface”
 - Solution: *speedup*, i.e. crossbar run faster than external line rate (tradeoff: speedup factor of 4 deemed sufficient, 2 is the minimum value if priority in use)

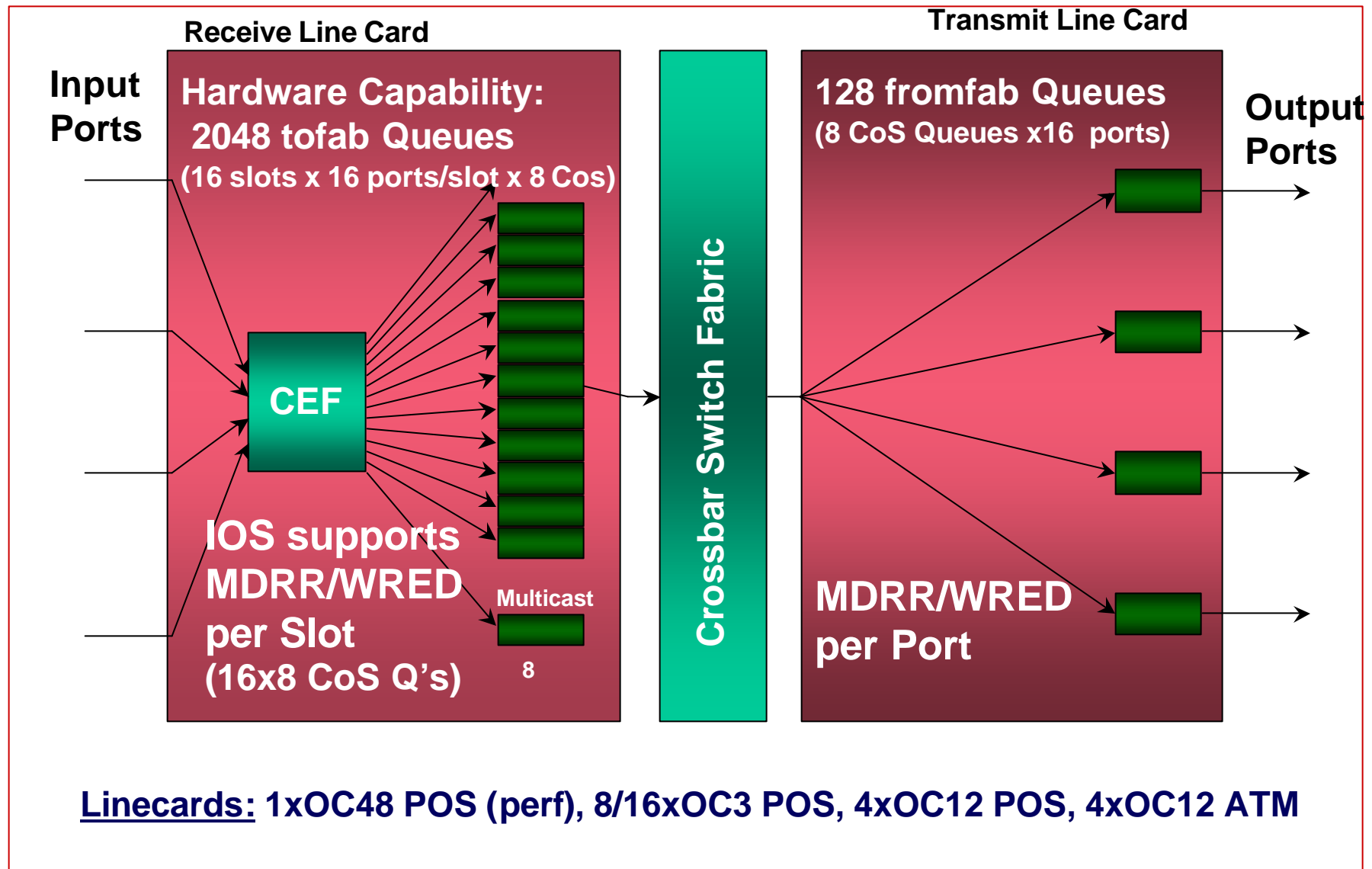
Crossbar scheduling algorithm

- High throughput
- Starvation free
- Fast (algorithm cannot be a performance bottleneck)
- Simple to implement (in special-purpose hw)
- **iSLIP: rotating priority arbitration**
 - Step 1: **Request**. Each active input sends a request to *every* output for which it has a queued cell
 - Step 2: **Grant**. If an output receives a request, it chooses the one that appears next in a fixed round-robin schedule starting from the highest priority element, i.e. the port which should be serviced next (fairness) Input is notified.
 - Step 3: **Accept**. If an input receives a grant it accepts the one which appears next in a round robin order, starting from the highest priority element.
 - pointer to highest priority element is increased by each matched input and output
 - Iterations continue until all active inputs/outputs are matched

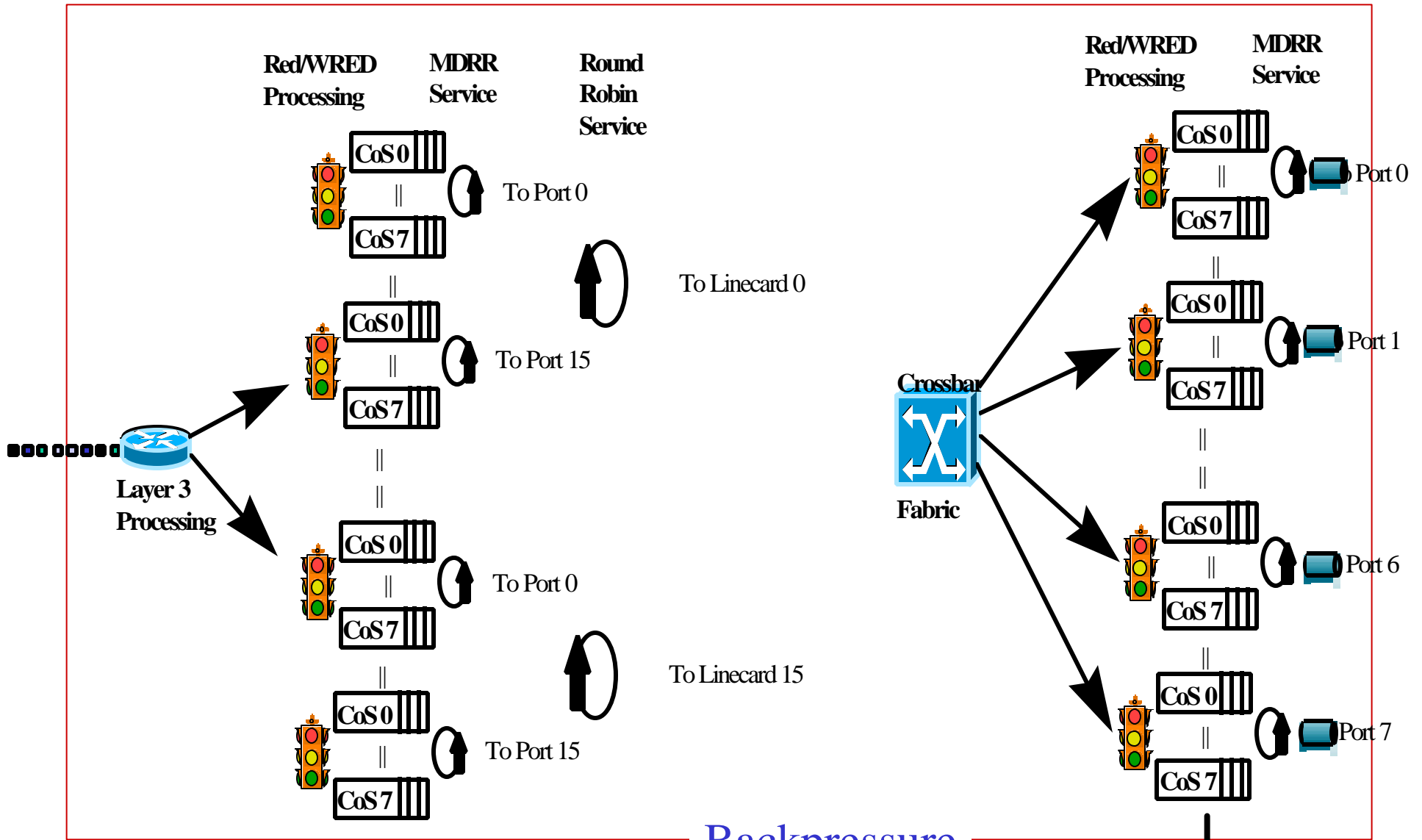
Crossbar scheduling algorithm - cont

- iSLIP Extension needed to support multicast transmission and priorities:
 - A separate set of pointers for multicast traffic is maintained.
 - Multicast pointers are shared by all input ports and all output ports separately
 - Unicast traffic cannot starve multicast traffic and vice versa
- Result: ESLIP

Buffer Management – cont (Diagram: from Cisco)

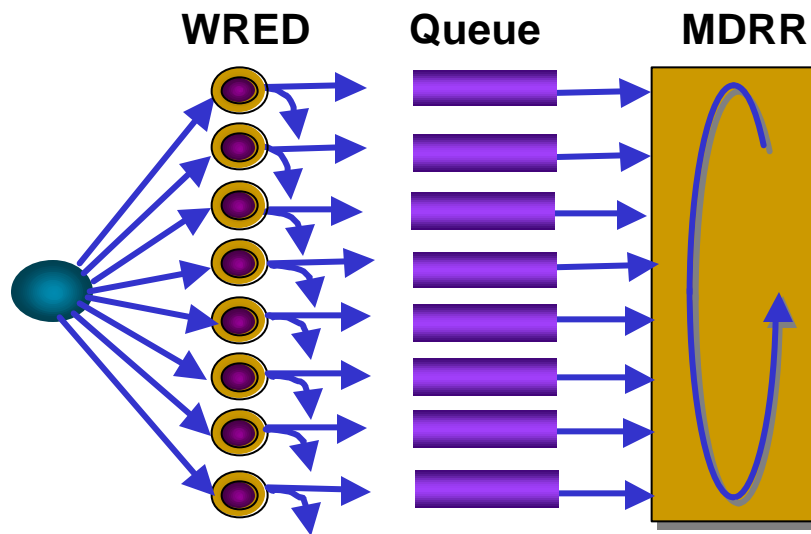


Overall Queue Scheduling (Diagram from CISCO)



Class of Service Processing in a VOQ (Diagram from CISCO)

WRED -> Scheduling



MDRR and WRED configuration

- **MDRR queue quantum** (byte):
 - $\text{MTU (in byte)} + (\text{queue_weight} - 1) * 512$
- **Bus share**(q_i) = $\text{quantum}(j) / \sum_i \text{quantum}(i)$
- **WRED min threshold**:
 - $\text{Min} = 1/10 * \text{pipe_size}$ // pipe size: num of packets which can be transmitted in 1 RTT
 - $\text{Pipe_size} = \text{RTT} * \text{bw} / (\text{MTU} * 8)$
- **WRED max threshold**:
 - Adjusted to be a power of 2

References

- Cisco RST-301: Router Architecture and Cisco IOS Internals
- Cisco 12000 Series: IP QoS Features
- GSR Product Training: GSR Class of Service Features (CoS)
- Fast Switched Backplane for a Gigabit Switched Router, N. McKeown
- Quad OC-48c/STM-16c Packet-Over-SONET Line Card Installation and Configuration Note