IBM PowerPRSTM

A Scalable Switch Fabric to Multi-Terabit: Architecture and Challenges July, 2002

Speaker: François Le Maut





Outline

- -Introduction
- -General Architecture
- -Flow Control
- -Multicast
- -Redundancy
- -Physical Constraints
 - I/O Constraints
 - Integrated SERDES





Today's Switching Challenge

A Versatile Switch Fabric to Meet Bandwidth Demand and Requirements at Nodes of Communications Networks Handling Multi-Applications on a Single Infrastructure:

-Real-Time Applications such as Voice and Video-conferencing

- •i.e., Applications Requiring a QoS (Quality of Service)
- •VoIP as well as Carrier-Class Voice Transport
- -Best Effort Service
 - •Data File Transfer, e.g.: E-mail

To Allow Equipment Manufacturers to Build Platforms aimed at Switching all Protocols in Common Use:

- IP, ATM, Ethernet, legacy TDM ...



An Output Queuing Shared-Memory Switch

High Performance Architecture at a Reasonable Cost:

- -OQ Switch Known to be Best for Performance
 - •Full Outbound Throughput
 - •No Internal Blocking
- –Best Buffer Memory Utilization Through Sharing of Up to 16k Packets (Q-128G)

Requires High Thruput Buffers:

 Four-ported High Speed (2 Ns Cycle) Memories implemented in IBM CMOS Cu11 Technology (Leff=0.11)





VOQ's in Queue Manager (CSI)

Common Switch Interface or CSI, is Queue Manager Companion Chip to PowerPRS (1 per IN/OUT Port):

- -Holds VOQ's in Ingress CSI
 - Prevents HOL blocking
 - •Allows a Per Port & Per Priority Flow-Control
- -C192 is Companion Chip to PowerPRS Q-64G





Holding Traffic in Ingress CSI VOQ's

To Prevent OQ's from Overflowing, Packets are No Longer Admitted in Switch Core (for That Port) when an Output Port Congestion is Detected

–Lower Priority Packets are Held First According to a Series of Thresholds Associated to the Set of OQ's

-Priorities are Intrinsically Fully Preemptive





Distributed Hop By Hop Flow Control

If Egress Buffer Starts to Build Up, Packets May Be Denied Permission to Leave Switch Core (Sent Grant is Removed)

-On a Per Priority Basis

There is NO Centralized Scheduler

Decisions are Made
Independently in Each
Component (Switch Core,
Ingress & Egress Egress
CSI) on the Basis of the
Broadcast of Flow Control
Information (In Each
Packet Header)



Handling of Multicast Traffic

Is 'Built In' in PowerPRS Architecture (Output Queuing Shared-Memory Switch)

- There is a Single Copy of Packets Transmitted in Shared Memory from Ingress CSI's
- -One Reference in each OQ Concerned by MC
- –Packet Released when Last Copy Forwarded
 - •NOT a Requirement that Copies Have to Leave Switch Core in the Same Packet Cycle (to be compared with Crossbar)
- -MC Traffic is Independently Flow Controlled thru a Simple Metric





Credit Tables & Exhaustive Scheduling

Two Mechanisms to be Used with CoS (Class of Services i.e., Priorities) to Implement QoS requirements

-Credit Table

•to Provide a Minimum BW to a CoS to Avoid Starvation in Presence of Higher Priority Traffic

–Exhaustive Scheduling

- •to Force a Higher Priority Traffic to be Processed Exhaustively Irrespective of the Credit Table Allocations
- •to Process Real-Time Traffic, i.e.: Voice







1+1 Redundancy

CSI, e.g.: C192, Can be Connected to Two Sets of PowerPRS i.e., Two Switching Planes

–Lossless Scheduled Switch-Over

- -50 ms Automatic Switch-Over in Case of Failure
- –Enables Load Sharing Capability

IBM Serial Link @ 2.5 Gbps



PowerPRS Scalability

PowerPRS Architecture (i.e.: Output Queuing + Shared-Memory) Is Scalable Mainly Through the Combination of Two Mechanisms

-Shared Memory is 'Naturally' Scalable

- Memory Can be Grown because of Technological Innovations and Lithography Progresses
- •Memory Can be Shared as Speed and Number of RAM Ports Increase

Each Switch Module Handles only Packet Slices:
 a Master Module Drives Slaves Modules through
 a SPEX (SPeed EXpansion) Bus

 Packets Can be Sliced Down to Packet Header Size

Up to 1 Tbps Now !



Master

Slave

Slave

Slave

SP

Counting Signal I/O's

To Move One Tbps of Data IN and OUT e.g., of a 64-port OC192 (10 Gbps) Switch:

-@ 16 Gbps (1.6 Speedup Factor) per Port (IN & OUT)

•Requires 8 Serial Link @ 2 Gbps effective (without 8B/10B Overhead)

•64 x (8 + 8) = 1024

i.e.: 1024 x 2.5 Gbps Serial Links per Tbps of Data to Switch (One Differential Pair per Link)

Denser Connectors Can Handle 70 Pairs per Inch:

–About 15 Inches (37 cm) of Board Connector Required pe Switched Tbps of Data

Faster Serial Links (10 Gbps) Required for Multi Tbps Switches



512GBPS PowerPRS Reference Switch Board



Fully Self Testable (100% BW) from SWICC Interface



