FocalPoint II

A 300nS, 240 Gb/s switch/router

Agenda

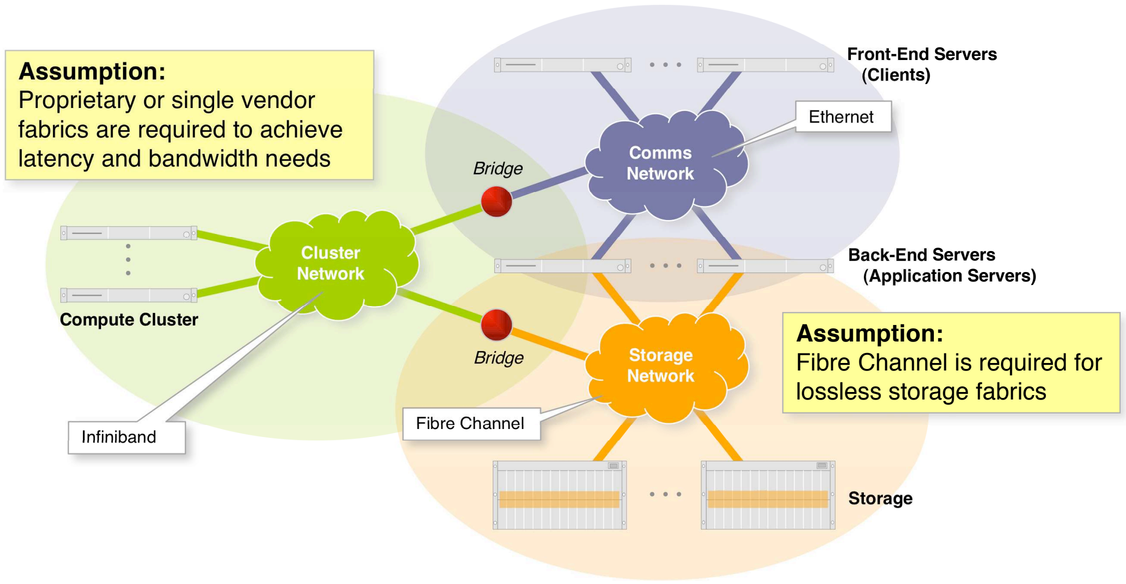
Datacenter Interconnect Requirements

FocalPoint I Status Update

FocalPoint II (Bali) Overview

Anatomy of the Multi-Fabric Data Center

Inefficient islands add complexity and cost; limit scale-out

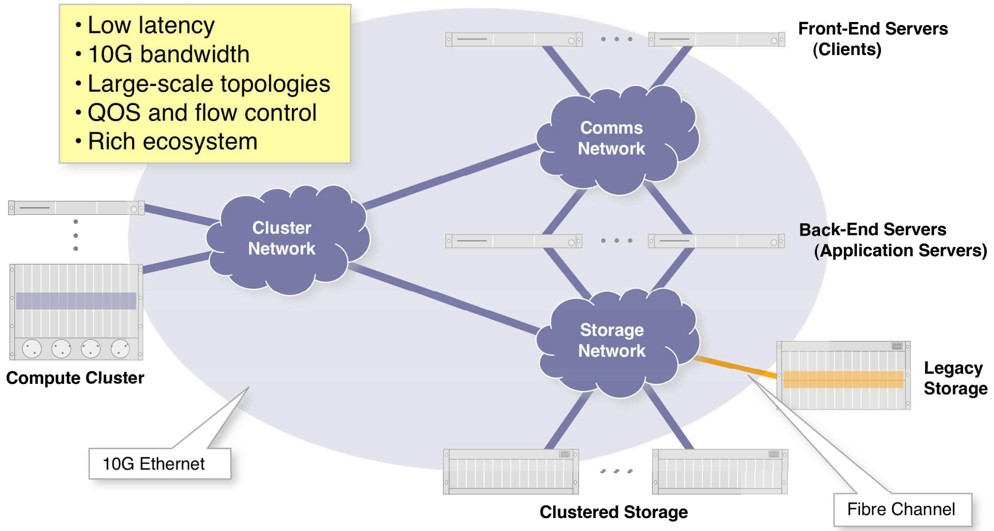


Assumption:
Proprietary or single vendor fabrics are required to achieve latency and bandwidth needs

Assumption:
Fibre Channel is required for lossless storage fabrics

10GE: Unifying Datacenter Interconnect

Datacenter Ethernet enables full cross-sectional bandwidth and a single management domain over all three networks



- Low latency
- 10G bandwidth
- Large-scale topologies
- QOS and flow control
- Rich ecosystem

Step 1: Solve Latency and Port Density

The world's most powerful 10G Ethernet switch chip

- **Highest port density** (24 10GE ports)
- **Highest bandwidth** (240Gbps)
- **Lowest latency** (200ns)
- **Most scalable** (fat trees, 1,000s of ports)
- **Most integrated** (single chip)

FocalPoint Evaluation Platforms

(The world's most integrated 10G Ethernet systems)

Reno
24-port 10GE design

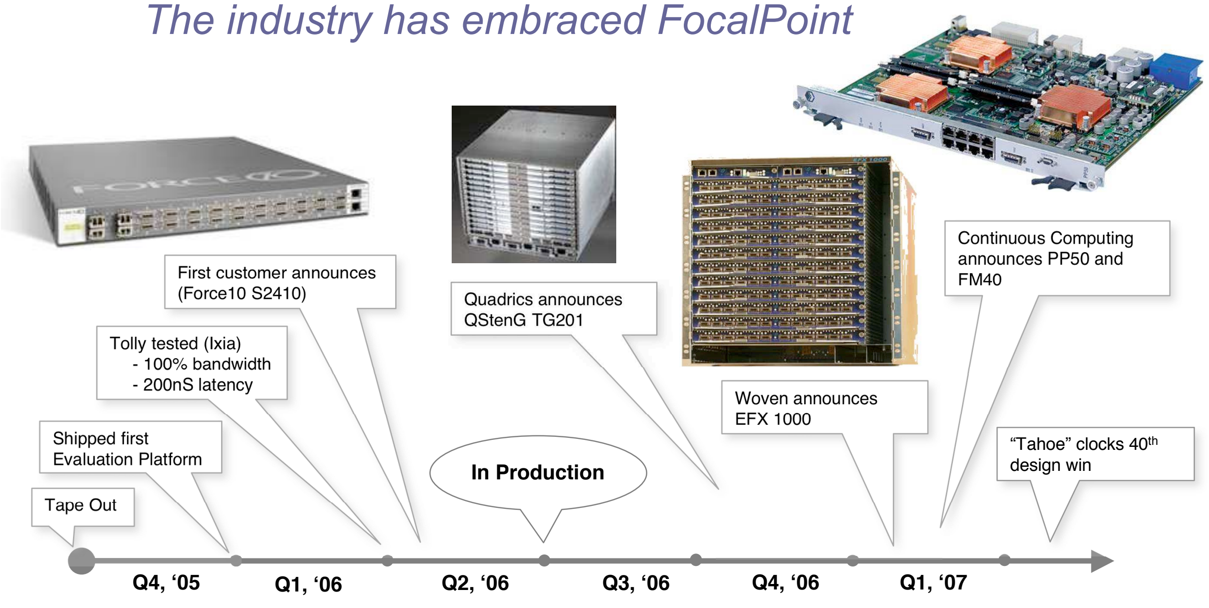
San Marino
Highest 10GBase-T density

Heavenly
IBM BladeCenter-H fabric

Vegas
Non-blocking 1GE platform

FocalPoint Status Report

The industry has embraced FocalPoint



Tape Out

Shipped first Evaluation Platform

Tolly tested (Ixia)
- 100% bandwidth
- 200nS latency

First customer announces (Force10 S2410)

In Production

Quadrics announces QStenG TG201

Woven announces EFX 1000

Continuous Computing announces PP50 and FM40

"Tahoe" clocks 40th design win

Step 2: Routing & Network Performance

FocalPoint II (Bali) project goals

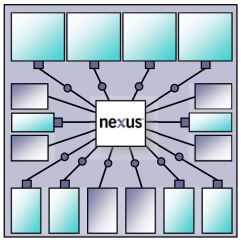
- **Maintains Gen I Performance**
  - **24 10GE ports, 200nS, 360 MPPS**
  - **But increase to 2MB memory**
- **IPv4 & IPv6 unicast & multicast routing**
  - **16k IP addresses**
- **L2-L4+ ACLs with deep inspection**
- **Chip cascades**
  - Virtual switch of Clos, rings & meshes
- **Clos Improvements**
  - As close to full bandwidth as possible

Converged Enhanced Ethernet (CEE)
- **Enable lossless Ethernet fabrics**

Architecture Enabling Circuits

Two key IP blocks demonstrate the virtues of the technology

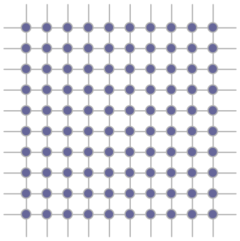
Nexus*
(Terabit Crossbar)



- Gigahertz performance
- Terabit capacity
- Nanosecond latency
- No power penalty

* Licensed to PMC PMC-SIERRA for SoC interconnect

RapidArray
(Asynchronous SRAM)



- 720 MHz SRAM
- 1200 MHz interconnect
- 518 Gbps throughput
- Scalable for any use

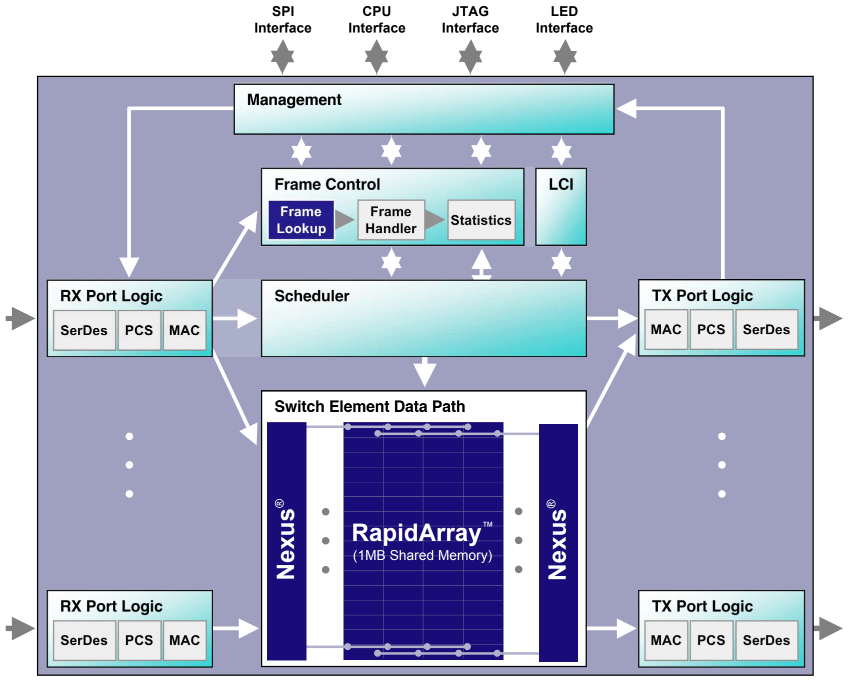
Key Benefits:

- **Easily integrates independent clock domains**
- **Provides 4x overspeed**
- **Reduces overall chip area**

- **2x the speed of vendor cores (same size, density, yield)**
- **Reduces power consumption (based on activity)**
- **CAM circuit is close relative**

FocalPoint I & II Architecture

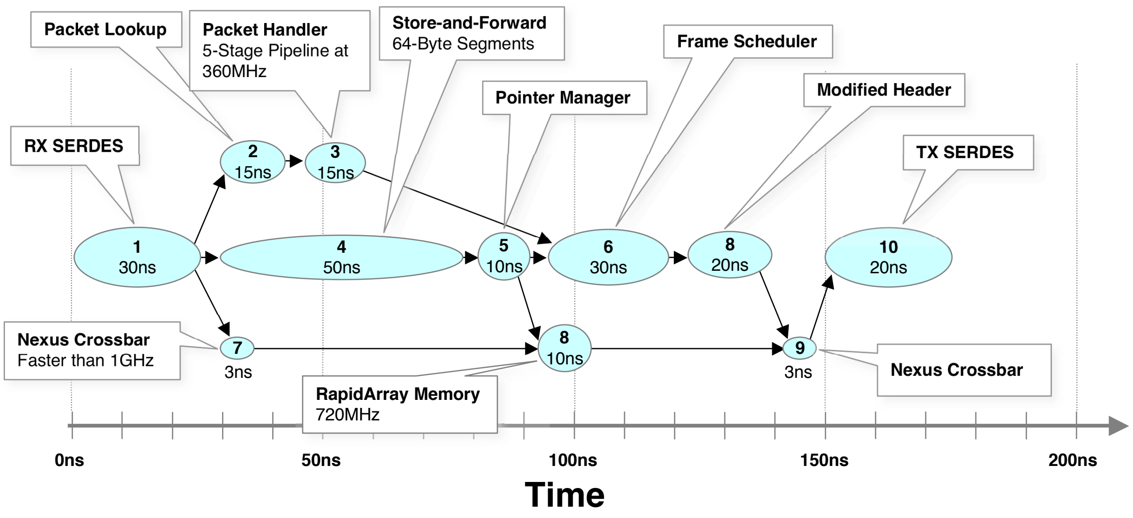
Modular architecture, centralized control



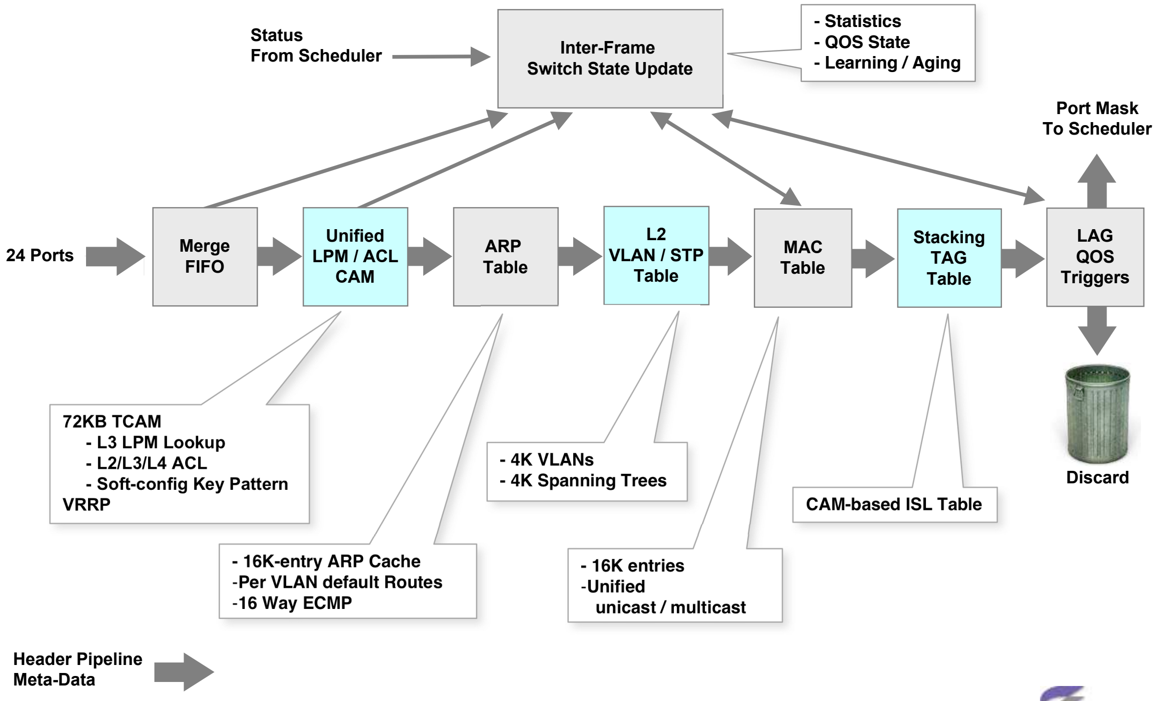
FocalPoint I & II Latency Detail

Requirement: Full L3 latency less than 300nS with 360 MPPS

Budget: 35 clock-cycle pipeline (2.6 nS / cycle) for all L3



FP II L2-L4 Packet Processing Pipeline



Status From Scheduler

Inter-Frame Switch State Update

- Statistics
- QOS State
- Learning / Aging

Port Mask To Scheduler

24 Ports

Merge FIFO

Unified LPM / ACL CAM

ARP Table

L2 VLAN / STP Table

MAC Table

Stacking TAG Table

LAG QOS Triggers

Discard

72KB TCAM
- L3 LPM Lookup
- L2/L3/L4 ACL
- Soft-config Key Pattern

VRRP

- 16K-entry ARP Cache
-Per VLAN default Routes
-16 Way ECMP

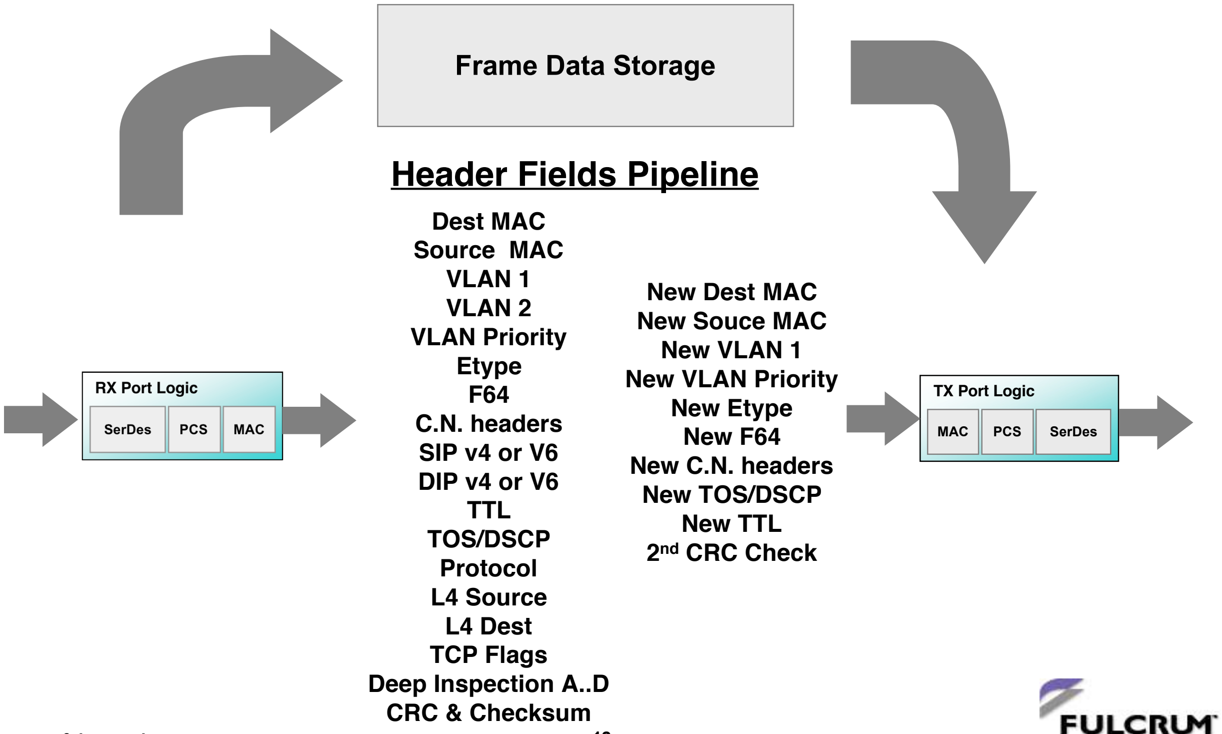
- 4K VLANs
- 4K Spanning Trees

- 16K entries
-Unified
unicast / multicast

CAM-based ISL Table

Header Pipeline Meta-Data

L2-L4 Packet Parsing and Manipulation



Frame Data Storage

Header Fields Pipeline

Dest MAC
Source MAC
VLAN 1
VLAN 2
VLAN Priority
Etype
F64
C.N. headers
SIP v4 or V6
DIP v4 or V6
TTL
TOS/DSCP
Protocol
L4 Source
L4 Dest
TCP Flags
Deep Inspection A..D
CRC & Checksum

New Dest MAC
New Souce MAC
New VLAN 1
New VLAN Priority
New Etype
New F64
New C.N. headers
New TOS/DSCP
New TTL
2nd CRC Check

RX Port Logic

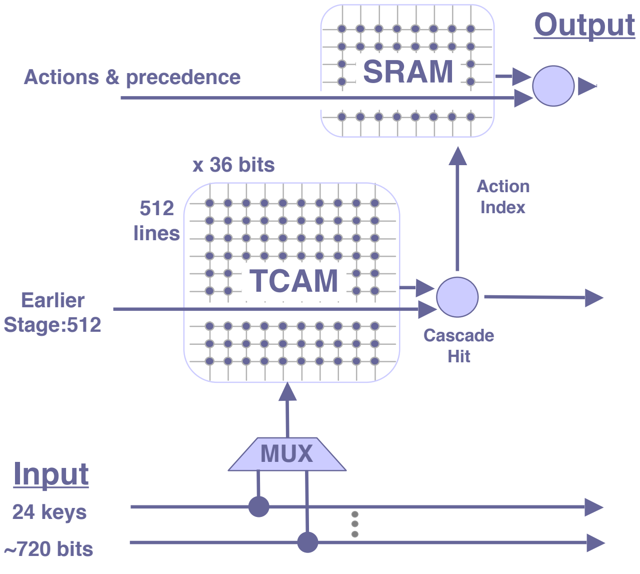
SerDes PCS MAC

TX Port Logic

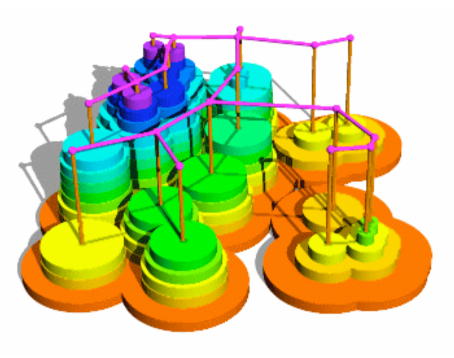
MAC PCS SerDes

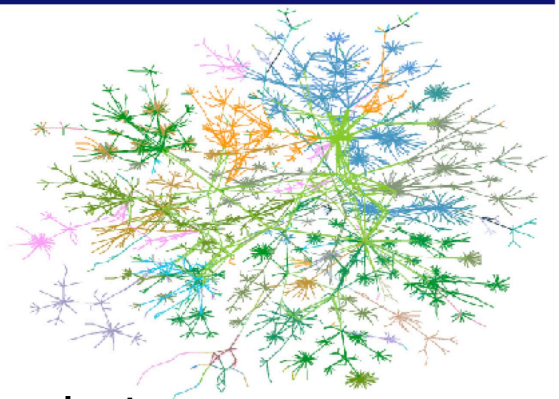
Filter & Forwarding (FFU) Organization

- **Input header keys**
  - Contains input fields
  - Any key available to any and all banks
- **TCAM Organization**
  - 32 banks, 72 KB, 16k min rules
  - Combine up to 32 banks
  - SRAM encodes 1 or multiple actions
  - Precedence in action combine allows multiple levels of non-orthogonal rules
- **Performance**
  - 1 gate delay per stage
  - 6 stages per flop
  - 6 clocks overall: 15nS of Latency
  - 360 MHz



Topology Enhancements and Ethernet

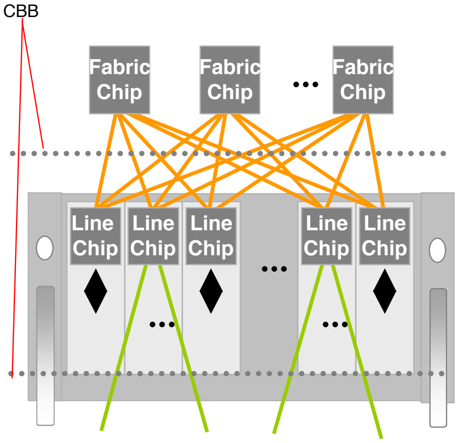




- **An Ethernet switch uses a single spanning tree**
  - Networks should scale in a non-blocking fashion
  - Spanning tree hashing and resilience
- **A Chip is not an Ethernet switch**
  - We may want to use one or many chips in a “switch”
  - We need to link data plane port state between chips
  - Multi-chip and multi-box link aggregation and management

Clos Architecture Theoretical Performance

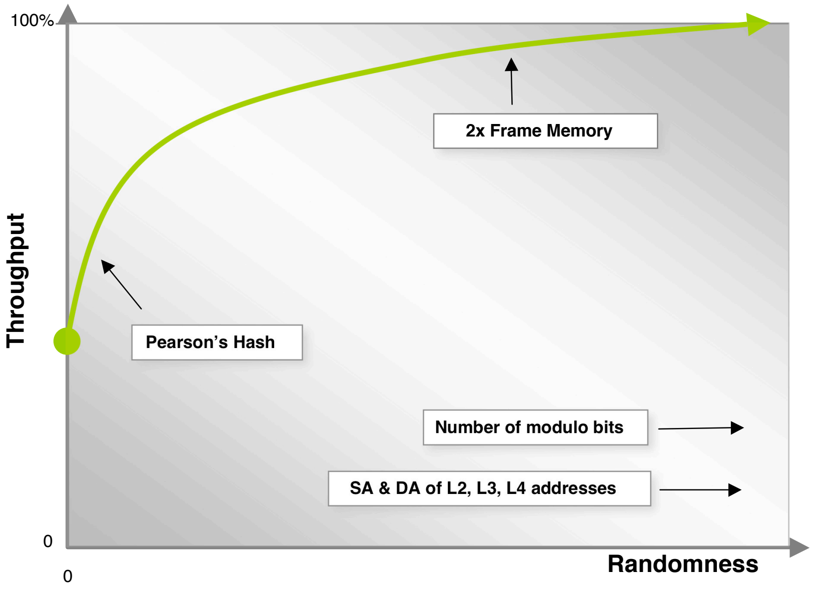
- **Significant Economic advantages are driving the move to Clos**
  - 10x reduction in per port price
  - 10x reduction in latency
  - 2-8x increase in port density / BW
- **All modern Clos architectures are really statistical multi-path, multi-hop networks**
  - Examples are Infiniband (Mellanox), fiberchannel (Brocade), Ethernet (Fulcrum)
- **Clos architectures achieve ideal performance if**
  - The path selection (often hashing) is sufficiently stochastic
  - There is enough over-speed in the system to compensate for any non-ideal path selection
  - There is enough memory per switch to compensate for collisions so that flow control is infrequent



FocalPoint II achieves new performance levels

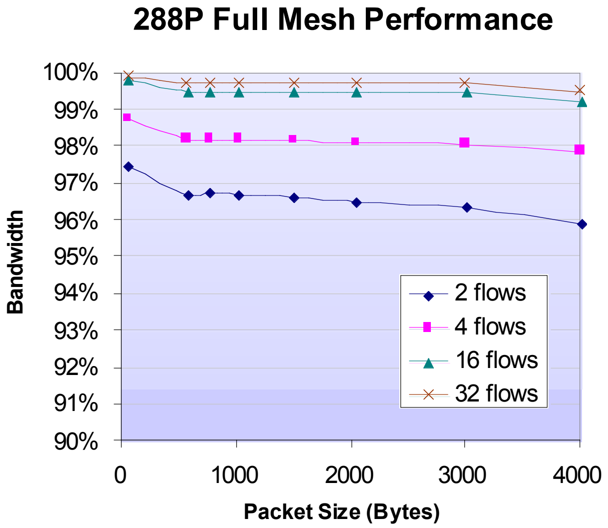
All flow based systems face the same challenges

How improvements in FocalPoint II capture increase throughput



Simulated Clos Performance

- **Standard Full Mesh**
  - Each port sends to every other port in the system every cycle
  - Every packet is randomly assigned to a flow on its port
  - 2-32 flows per port simulated
  - 576-9216 system flows (288P)
  - 288P system is made from 36 24P switch chips in a Clos configuration
- **Amount of over-speed from line card switching (24P line cards)**
  - 3% in 288P Clos
  - 23% in 48P Clos
- **Conclusion**
  - 16 flows per port is nearly perfect in a 288P system (four flow is very good)
  - 1 flow per port is nearly perfect in a 48P system given the 23% overspeed



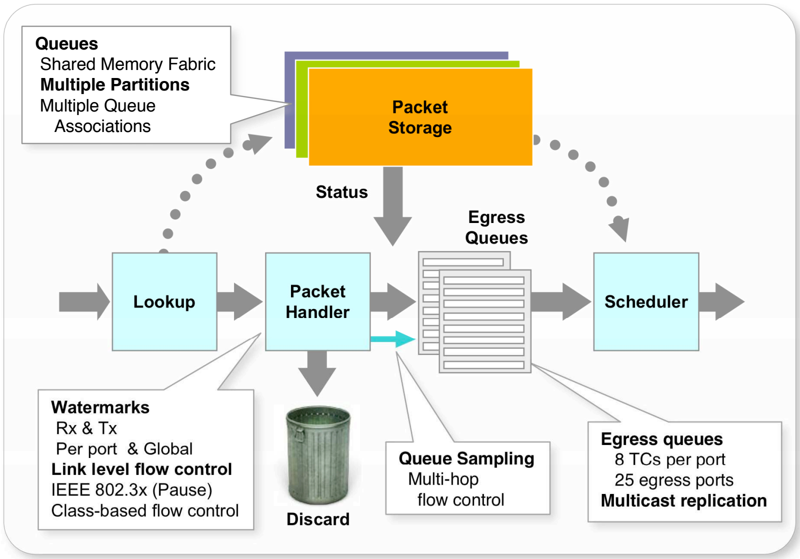
Congestion Control

- **There are three drivers for congestion control enhancements in the enterprise datacenter**
  - Storage & IPC traffic is highly loss and jitter sensitive
  - Multiple traffic types, like storage and LAN, require different best in class congestion control practices
  - Cost oriented fully integrated full bandwidth switch chips are required to use memory very efficiently
- **As a result there is a race to produce lossless, non-HOL blocking, low latency fabrics with optimal bandwidth**
  - Congestion control is being standardized by the IEEE in 802.1au and potential future working groups

Congestion Management and Scheduling Architecture

Traffic separation enables virtual switching

FocalPoint II Chip-Level Architecture



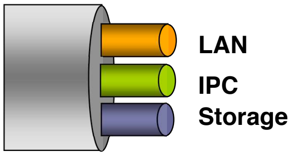
Queues
Shared Memory Fabric
Multiple Partitions
Multiple Queue
Associations

Watermarks
Rx & Tx
Per port & Global
Link level flow control
IEEE 802.3x (Pause)
Class-based flow control

Queue Sampling
Multi-hop
flow control

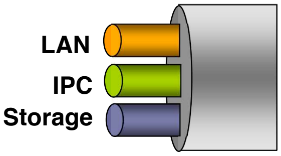
Egress queues
8 TCs per port
25 egress ports
Multicast replication

Egress Features



- Multi-level scheduling
  Bandwidth groups with priority
- Traffic distribution
  Deficit weighted round robin

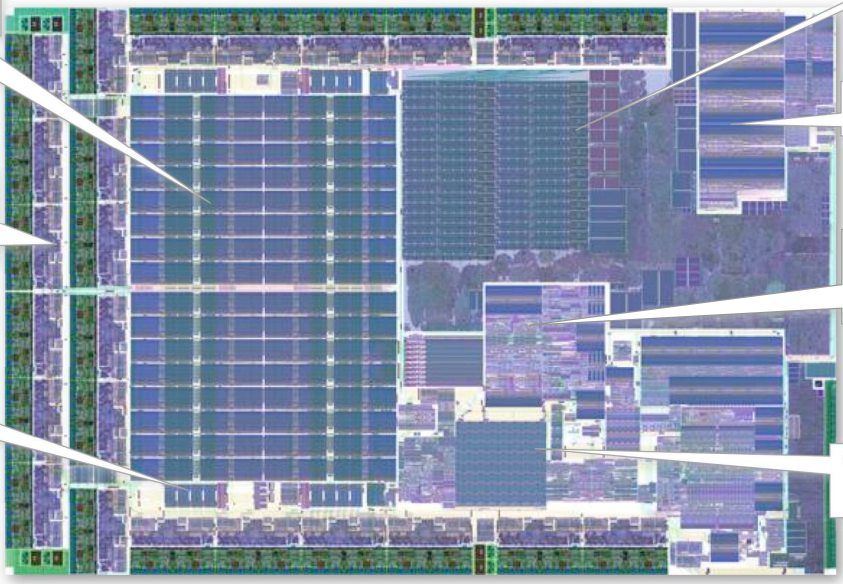
Ingress Features



- Flow Control
  Link & Per class pause
- Static & Dynamic rate-limiting
  Pause pacing & policing

Bali Chip Plot

Fabricated in TSMC 0.13um
250 Million Transistors



RapidArray Memory
2 MB

Ethernet Port Logic
- Phy (SerDes)
- PCS
- MAC

Nexus Crossbar
- Terabit capacity
- 3ns latency

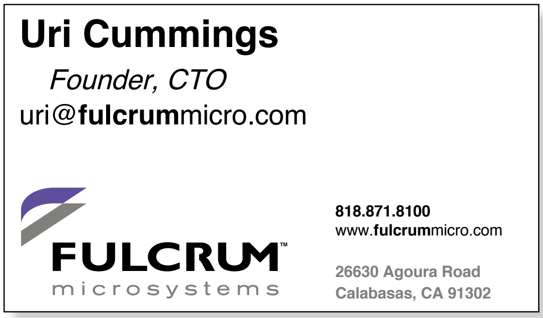
TCAM

MAC Table
- 16K addresses

Frame Control
- Frame handler
- Lookup
- Statistics

Scheduler
- Highly optimized
- High event rate

Thank You!



Uri Cummings

Founder, CTO

uri@**fulcrum**micro.com

FULCRUM™
microsystems

818.871.8100
www.**fulcrum**micro.com

26630 Agoura Road
Calabasas, CA 91302