

# FM4000

## A Scalable, Low-latency 10 GigE Switch for High-performance Data Centers

Uri Cummings  
Dan Daly

Rebecca Collins

Virat Agarwal  
Fabrizio Petrini  
Michael Perrone  
Davide Pasetto

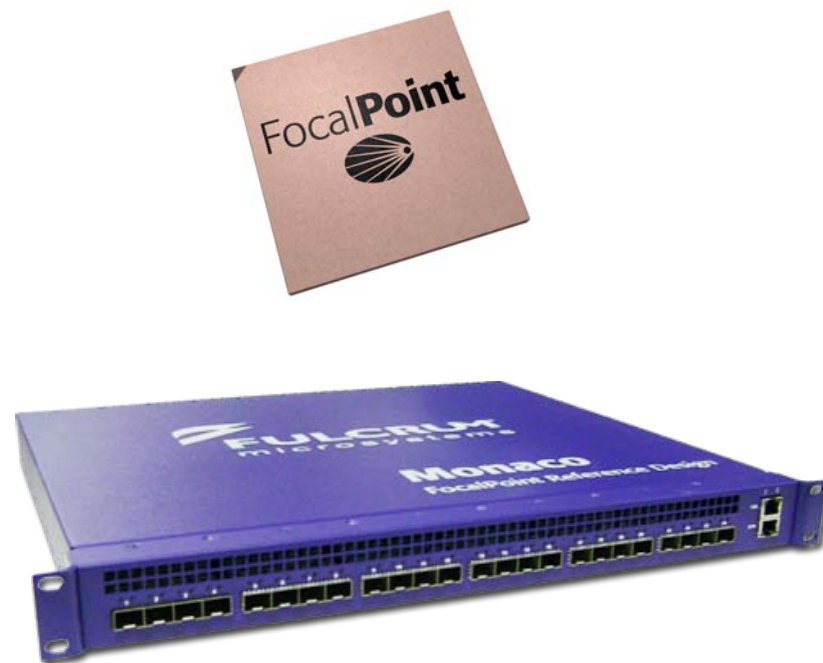


Columbia University  
IBM TJ Watson Research Center, US  
IBM Computational Science Center, Ireland

# What is FM4000?

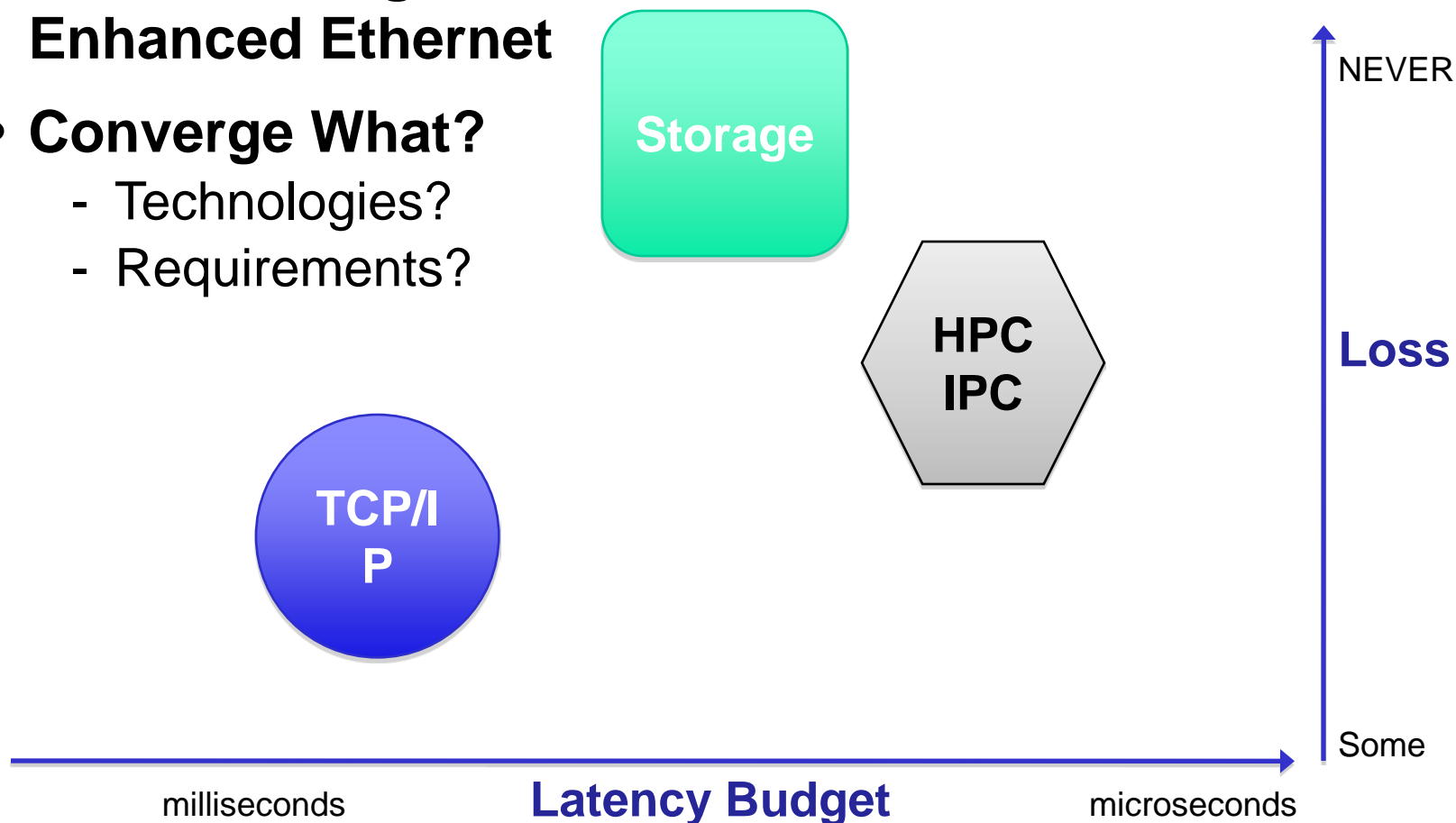
FocalPoint  
FM4000

- 10 GigE Silicon
- L2/L3 Switch/Router
- 300 ns Latency
- Parallel Multicast
- CEE Feature Set



# Convergence

- **CEE: Converged Enhanced Ethernet**
- **Converge What?**
  - Technologies?
  - Requirements?



# Latency Becomes Key Metric

- **All Non-IP Flows Are Latency Sensitive**
- **Old Foes Incur New Costs:**
- **Loss** → **Latency**
  - Retransmits become too expensive
- **Queuing** → **Latency**
  - Big buffers have big side effects

Network Speed	1B Queue Delay	9K Queue Delay
1 Gigabit / sec	0.008 microseconds	72 microseconds
10 Gigabit / sec	0.0008 microseconds	7.2 microseconds

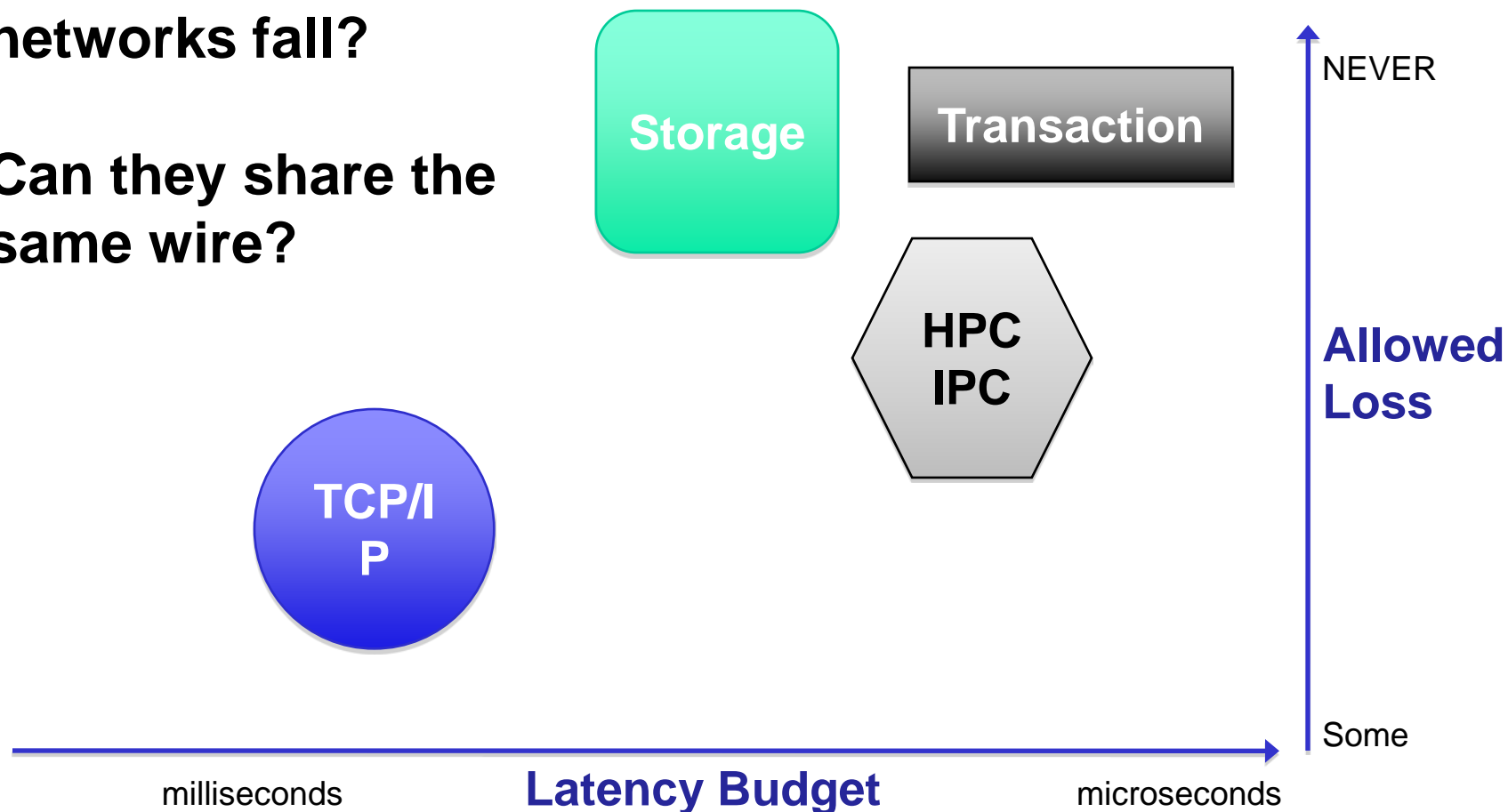
- **Cut-through Avoids Per-hop Queuing**

- **Latency Becomes System Level Metric**
- **Guaranteed Lossless**
  - Non-Blocking Architectures
    - Unicast & Multicast
  - Flow Control
    - Throttle senders before loss occurs
    - Link-Level (PFC) and End-to-End (QCN)
    - Fast response time critical
- **Transmission Selection**
  - Transmit latency sensitive traffic first
  - IP traffic can queue when needed to

# Convergence

Where do transaction networks fall?

Can they share the same wire?



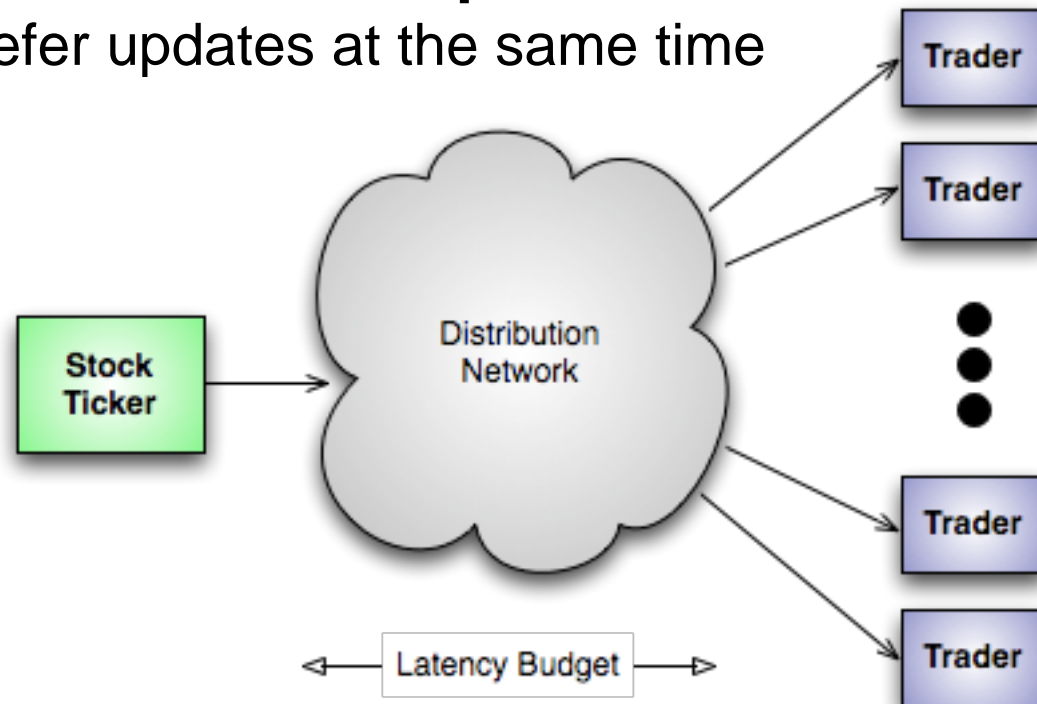
# Converged Network Requirements

## Step 1: Distribute Market Information

Must be multicast, low latency and low jitter

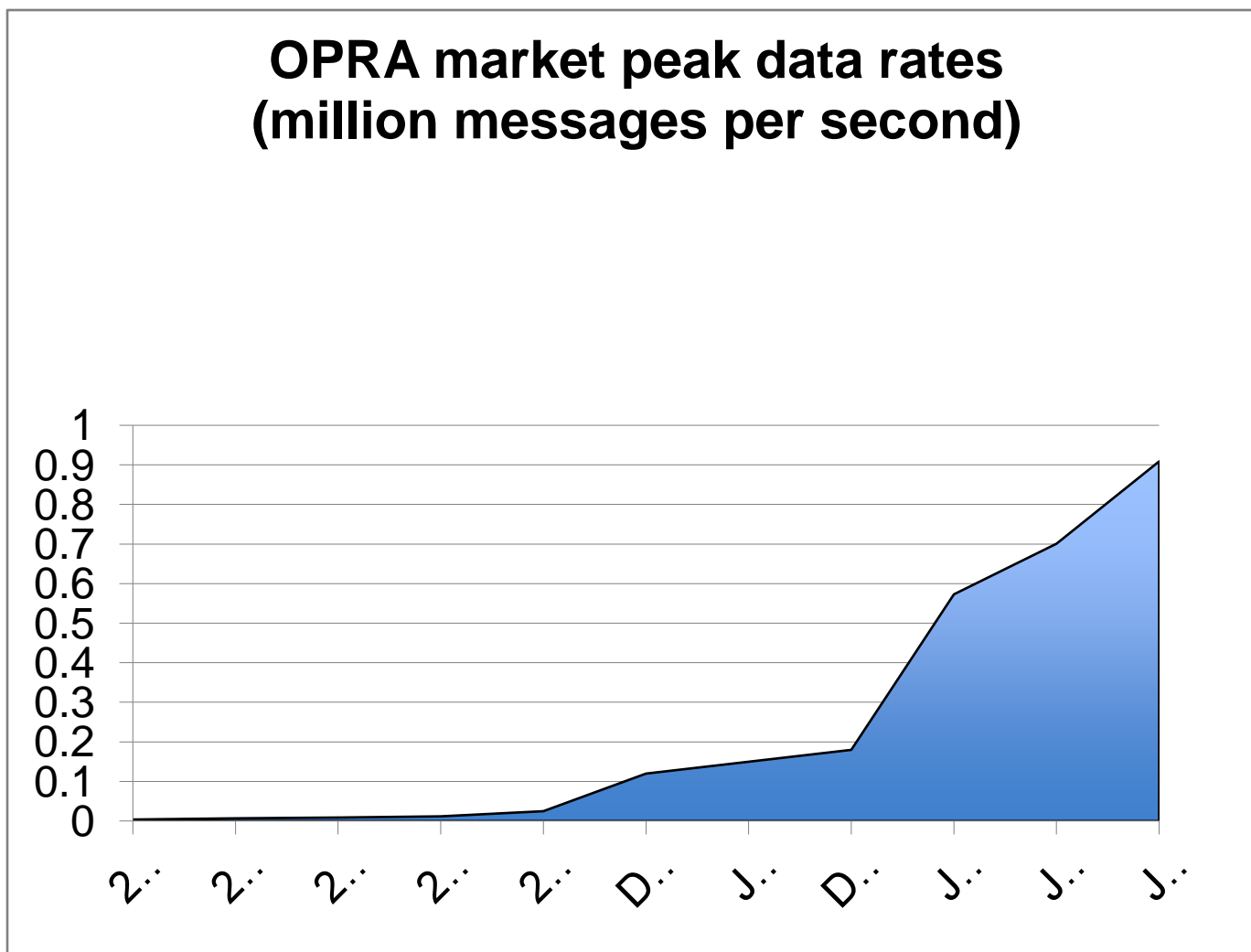
## Step 2: End Users Process Updates

End users prefer updates at the same time



## Step 3: Profit

# OPRA Feeds: Exponential Growth



# Transaction Network Requirements

- **Latency is THE System Level Metric**
- **Provisioned Lossless**
- **Transmission Selection**

1. **Can CEE switches be used to implement a transaction network?**

***Yes – using low latency Ethernet***

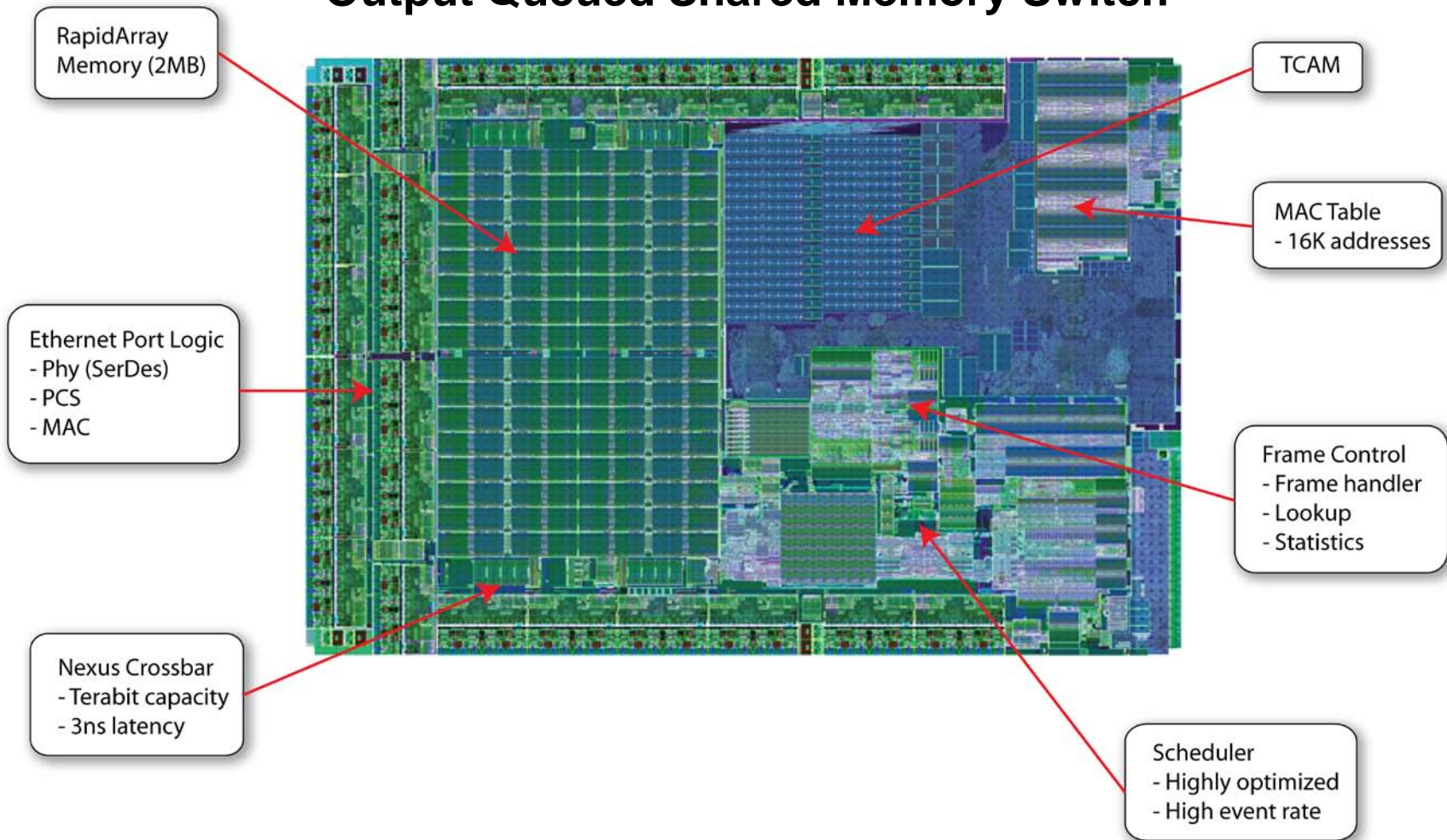
2. **Can I use this same network to check my email?**

***Yes – using CEE enabled switches***

3. **Can this functionality be demonstrated without building out an entire network?**

***Yes – by pushing the envelope***

## Output Queued Shared Memory Switch

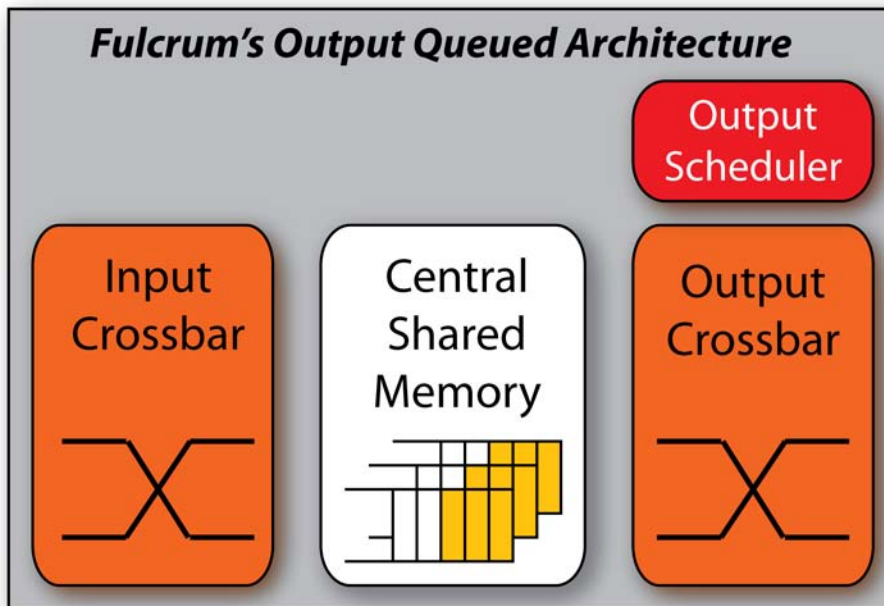


**N Ports \* (Min-frame-size framerate) = Packet Rate**

**24 Ports \* 14.88 M packets/sec = 357 Mpps**

**N Ports \* 10GigE = Data Rate (input and output)**

**24 Ports \* 10GigE = 240 Gigabits / second**

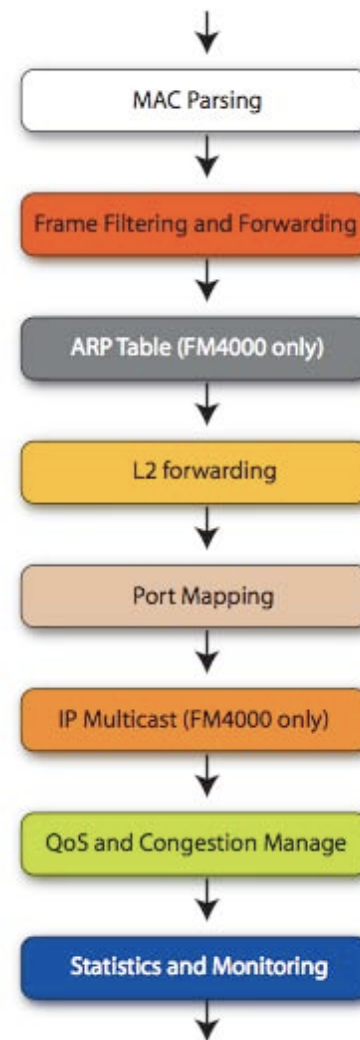


**Output Scheduler:  
Runs at Double  
Packet Rate**

**Shared Memory:  
Loads and Stores at  
Data Rate**

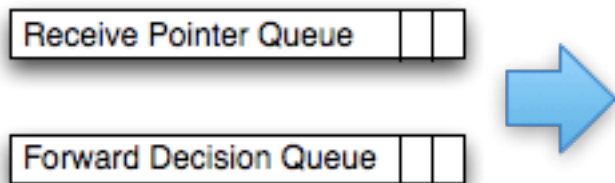
# Forwarding Pipeline

- **Ethernet Specific Tables**
  - MAC, ARP & IP Tables
- **Fully Flexible TCAM**
  - ACLs, flow tracking/forwarding, etc
- **Port Mapping**
  - Global addressing for thousands of ports in a fabric
- **QoS and Congestion Management**
- **Runs at frame rate ( > 360 Mpps )**
- **Sits in forwarding latency path**

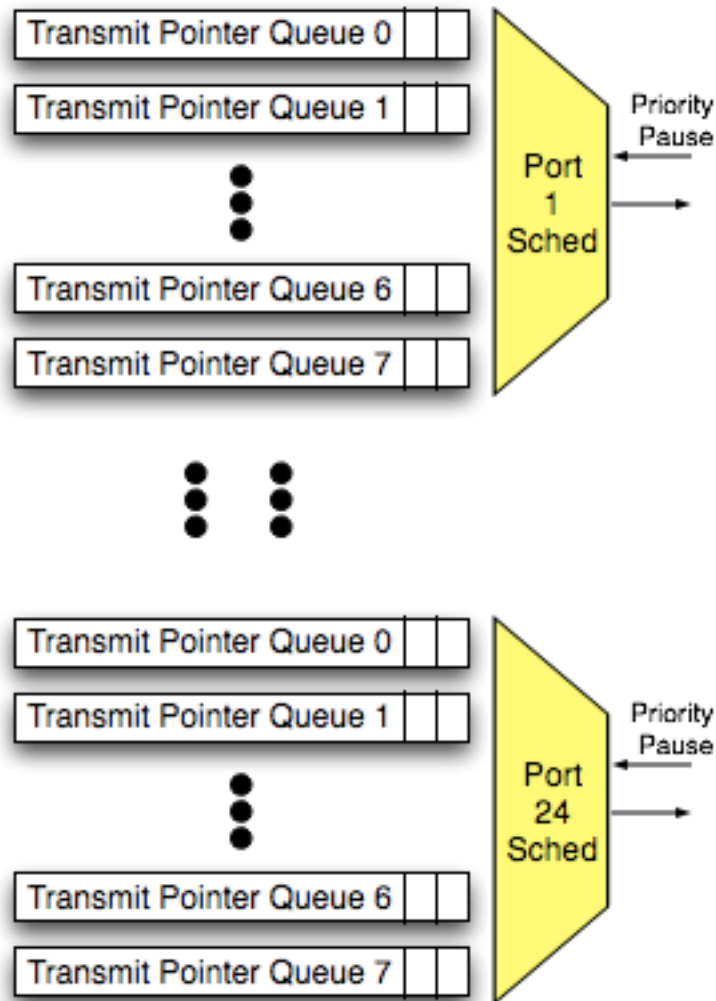


# Multicast Replication

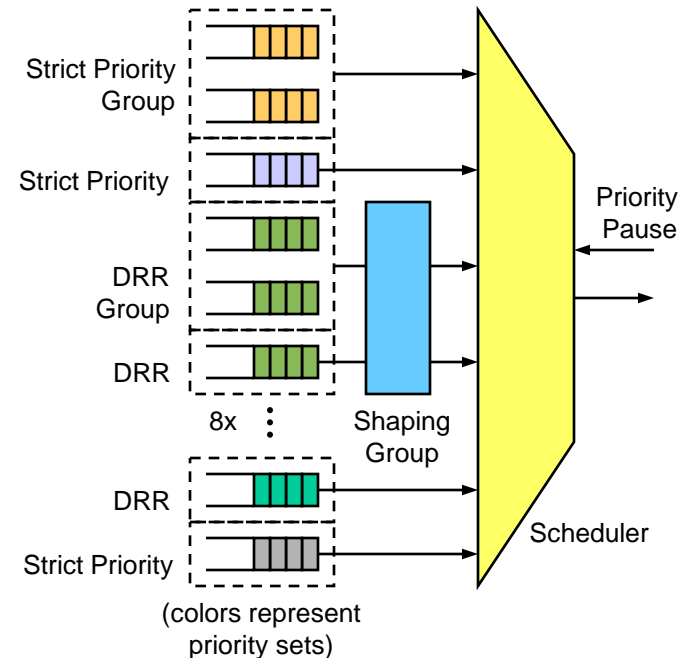
**Replication on Egress**  
**Pointer tracking**  
**Parallel draining at port**  
**67.2 nsec loop**



**Multicast copies**  
**transmit synchronized**  
**within 67.2 nsec**



- **8 queues per port**
- **Two Level Scheduler**
  - Scheduler first by group
  - Schedule second by priority
- **Min Bandwidth Guarantees**
  - Deficit round-robin
- **Max Bandwidth Guarantees**
  - Bandwidth shaper groups



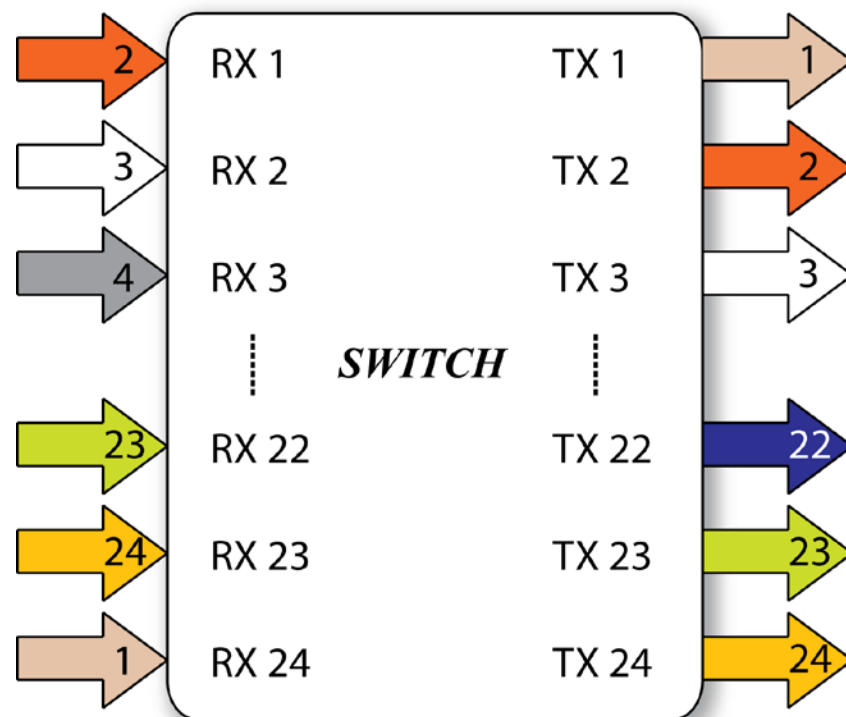
Example Configuration

- **Existing RFC Benchmarks**
- **Most RFC performance tests unlikely to occur in real network**
- **Defines performance envelop with extreme corner case traffic flows**

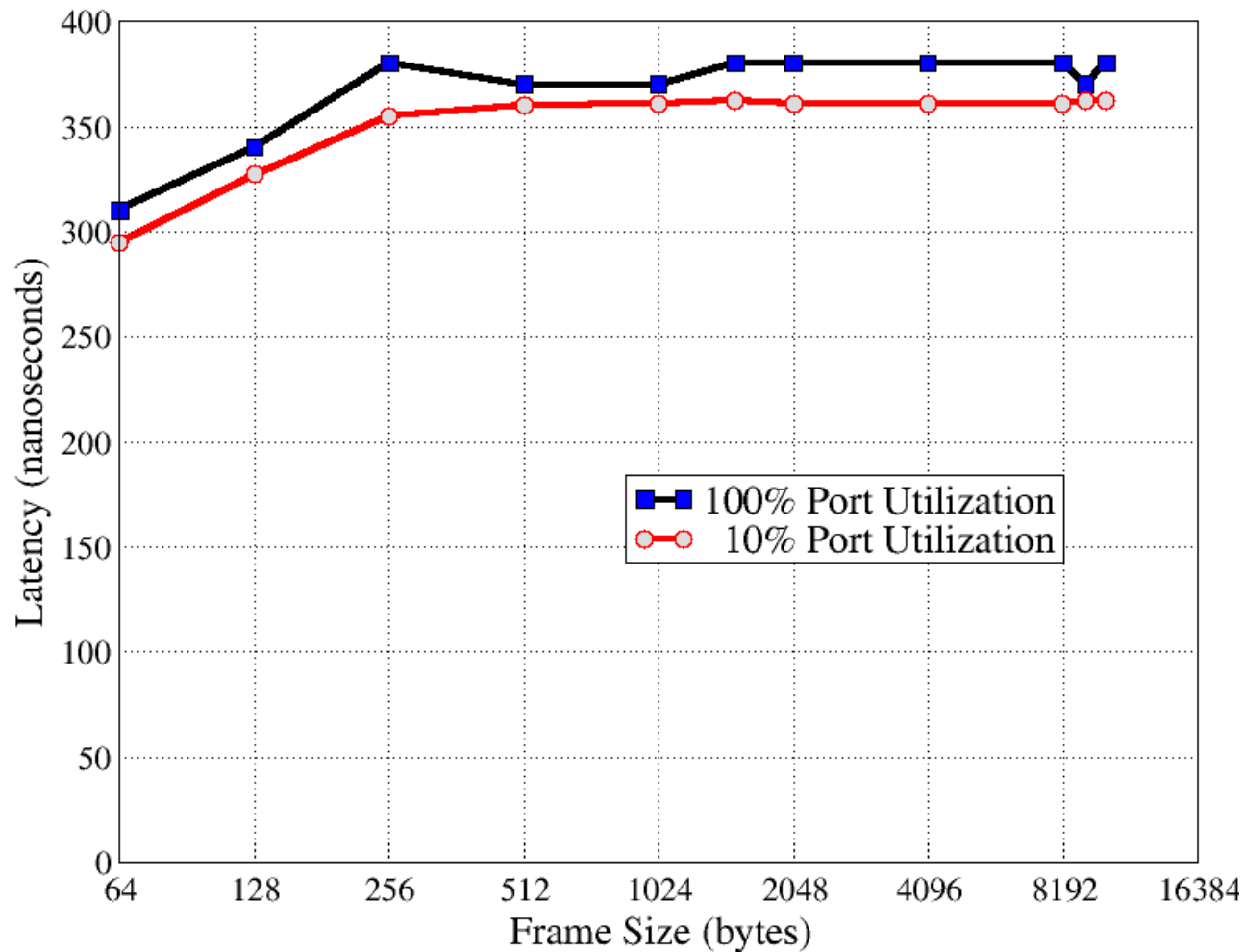
**Testing Goal: Start with RFCs, then expand to converged multicast traffic**

# Port to Port Benchmark

- **RFC 2544**
- **No inherent congestion**
  - expect zero queuing
- **Inability to forward a rate incurs:**
  - Loss
  - Queuing latencies



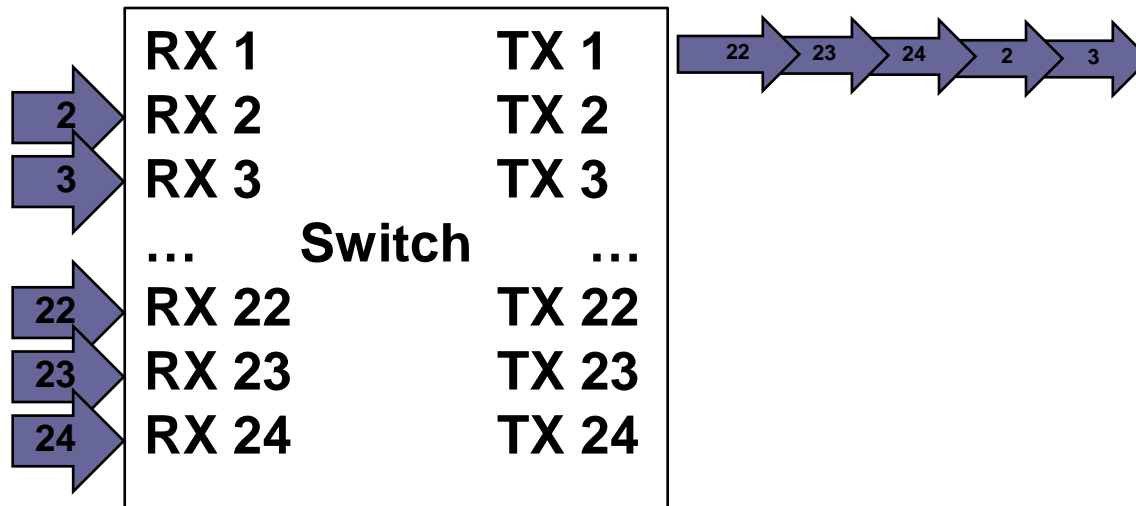
# Port to Port Benchmark Results



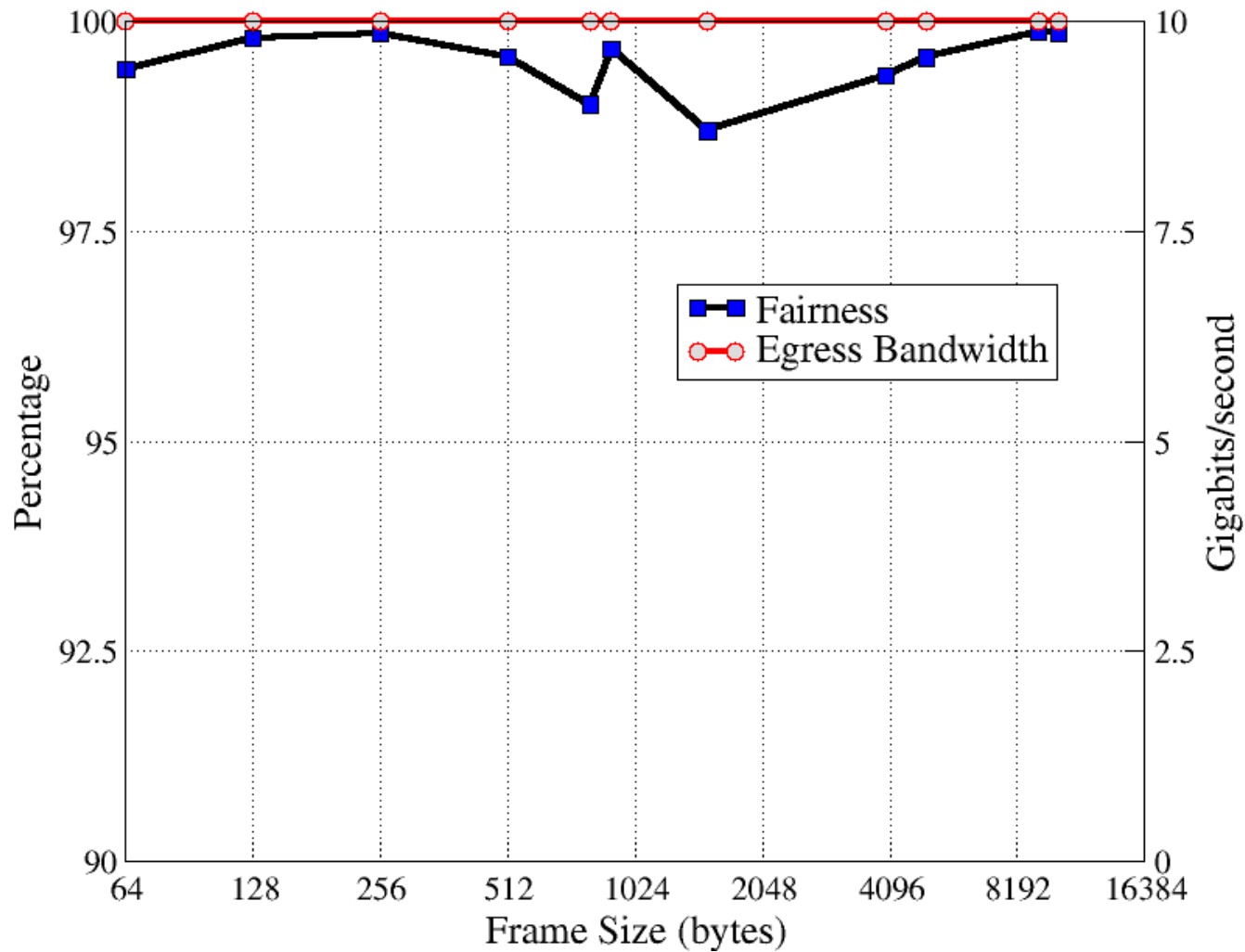
Zero Loss, nanosecond latency across frame sizes

# 23 to 1 Flow Control

- **Single Hot Spot**
- **Simultaneous Congesting Flows**
- **Flow Controlled to  $1/23^{\text{rd}}$  of Line Rate**
- **Must Be Fair**



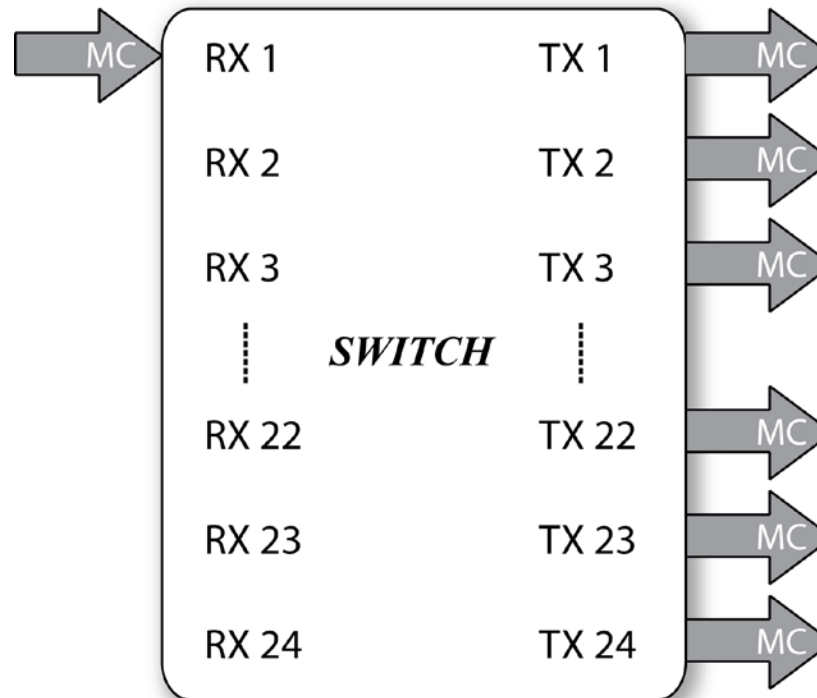
# 23 to 1 Benchmark Results



Zero Loss, ports throttled evenly (within 2%)

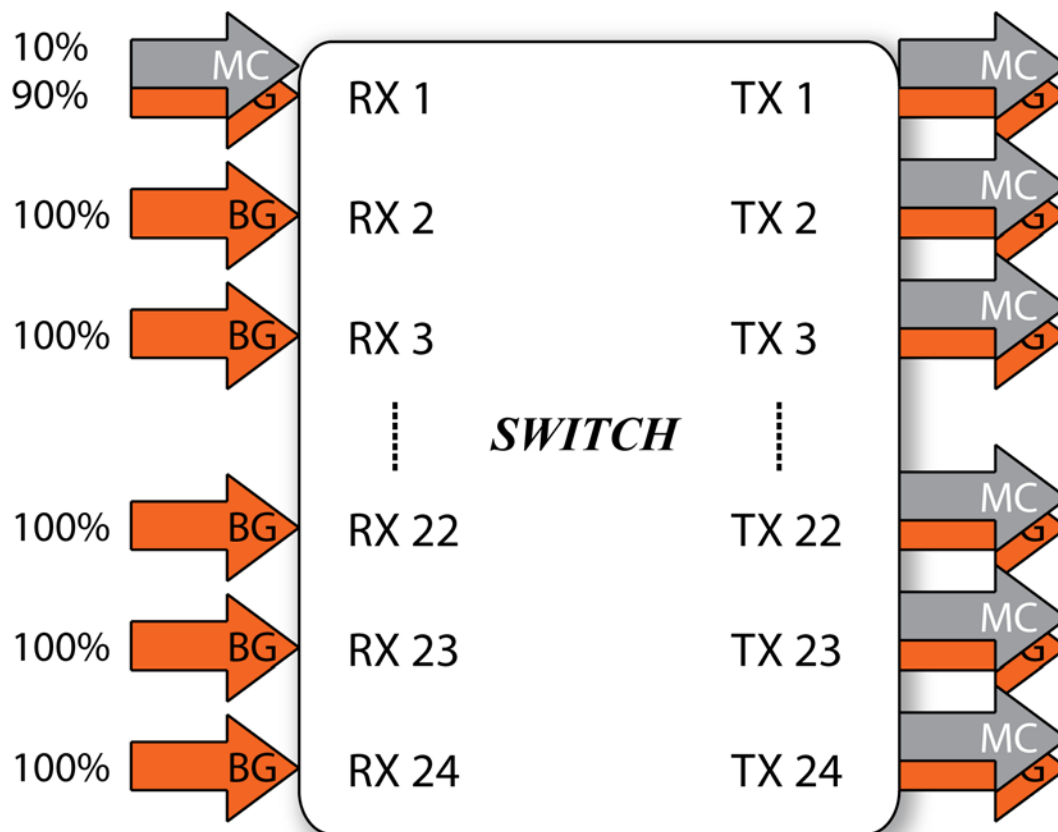
# 1 to 24 Multicast Test

- **Single Multicast Transmitter**
- **Loopback Suppression Disabled**
  - Traffic hairpin turns back to sender
- **Non-blocking Multicast Forwarding**

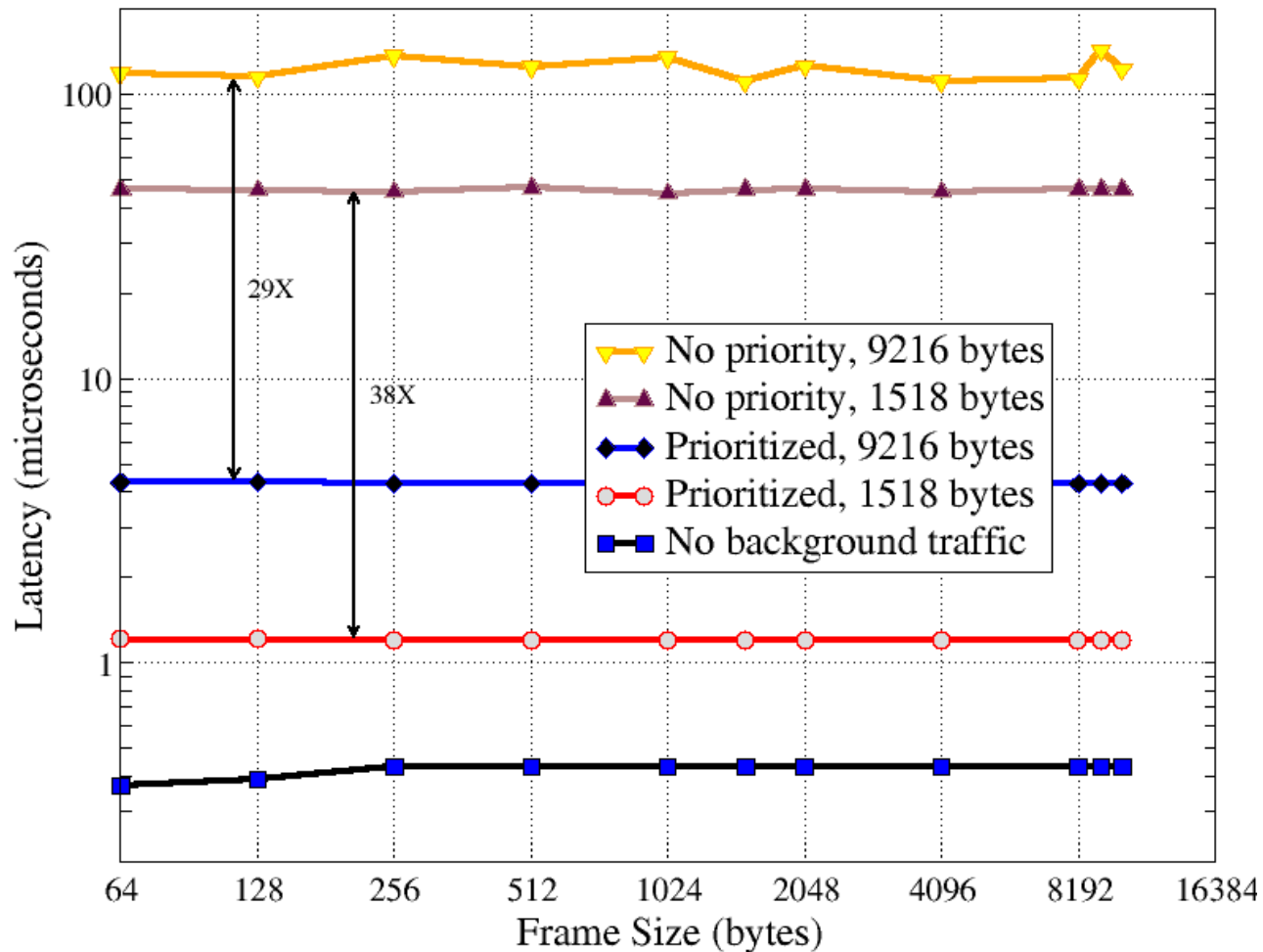


# Multicast Background Traffic Test

- Add Broadcast Traffic On All Ports
- Switch Filled With 100% Broadcast Traffic



# Multicast Background Traffic Results



Prioritization Guarantees BW and Latency

- **Without background traffic:**
  - 300 nanosecond latency, 1 to 24 multicast
  - Nanosecond jitter between multicast copies
- **With low priority background traffic:**
  - Adds up to 1 low priority frame's worth of queuing

- **Tests Demonstrate:**
  - Non-blocking architecture
  - Cut-through latency
  - Predictable delay under stress
  - Lossless flow control on all ports
  - Multicast transmission selection of high vs low priority
- **Data Determines Latency Budget**
  - Same wire for transactions and IP incurs latency on transactions
  - Separate wires, same switch incurs no penalty

Contributor	Organization
Rebecca Collins	IBM TJ Watson Research Center, US and Columbia University
Virat Agarwal	IBM TJ Watson Research Center, US
Fabrizio Petrini	IBM TJ Watson Research Center, US
Michael Perrone	IBM TJ Watson Research Center, US
Davide Pasetto	IBM Computational Science Center, Ireland
Uri Cummings	Fulcrum Microsystems
Dan Daly	Fulcrum Microsystems

**Many thanks to the team for putting together this paper!  
Questions?**