

FORCE 10™

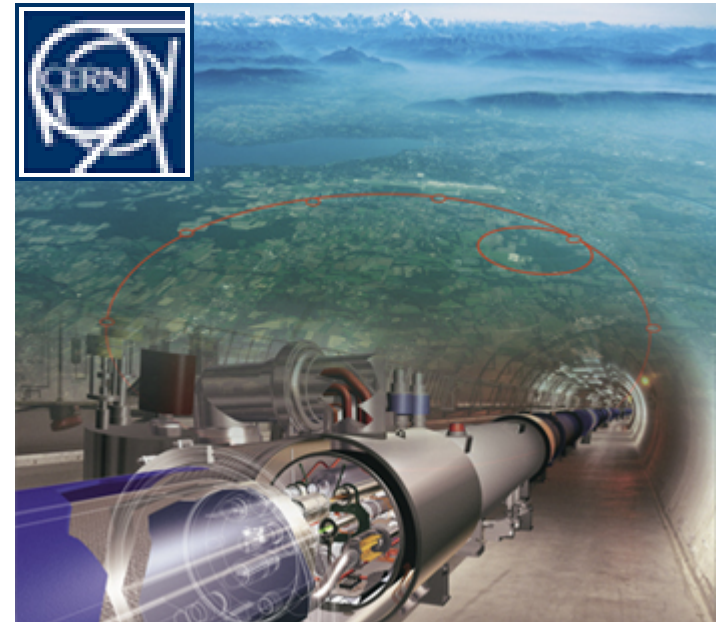
High Performance Ethernet for Grid & Cluster Applications

Adam Filby
Systems Engineer, EMEA



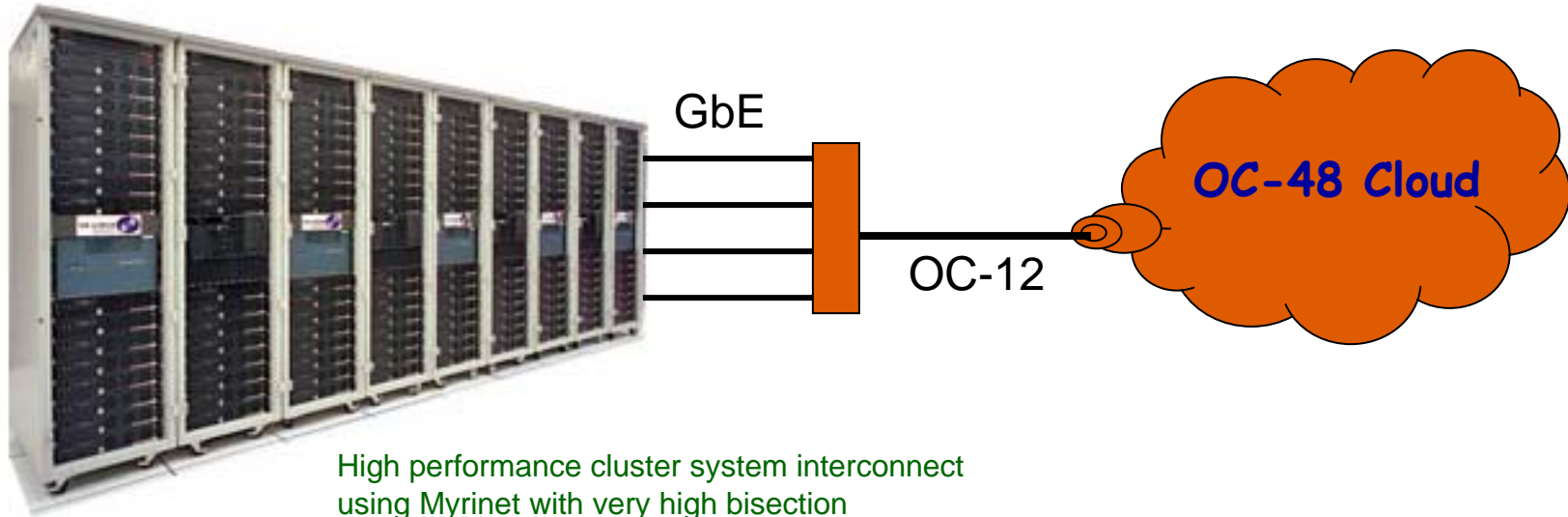
- Drivers & Applications
- The Technology – Ethernet Everywhere
 - Ethernet as a Cluster interconnect
 - Ethernet as the Grid interconnect
- Ethernet in the LAN/WAN
 - LAN PHY
 - WAN PHY
 - XFP Technologies – Distances 300m to 80km
- Requirements for placing all these eggs in one basket
 - Carrier Grade Stability
 - Carrier Grade Resiliency
- Beyond 10GE what's Next 40GE or 100GE

- Cluster Computing – How does the data get to the cluster
- CERN will be generating data at a minimum of 8 Petabytes per year
- 70,000 CPU's are required to analyse this data
- A Serious grid infrastructure is required to support this level of data transfer

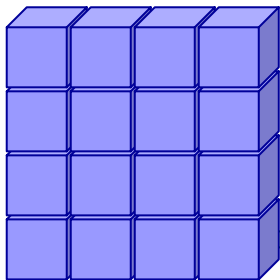


- Cluster connectivity requirements can be
 - MPI
 - I/O
 - Management
- Each has its own performance criteria
 - MPI often requires a low latency interconnect
 - I/O often requires high non blocking bandwidth
 - Management usually only requires 10/100 Ethernet bandwidth and latency is not an issue
- Grid connectivity requires long range and sometimes high bandwidth connectivity

Traditional Cluster Network Access



High performance cluster system interconnect using Myrinet with very high bisection bandwidth (hundreds of GB/s) with external connection of $n \times \text{GbE}$, n is small integer.



(Time to move entire contents of memory)

2000 s (33 min)


13k s (3.6h)

1 TB

0.5 GB/s

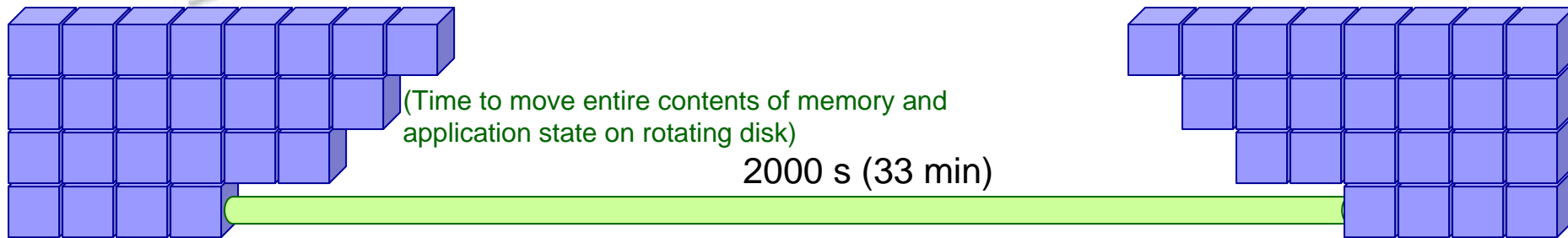
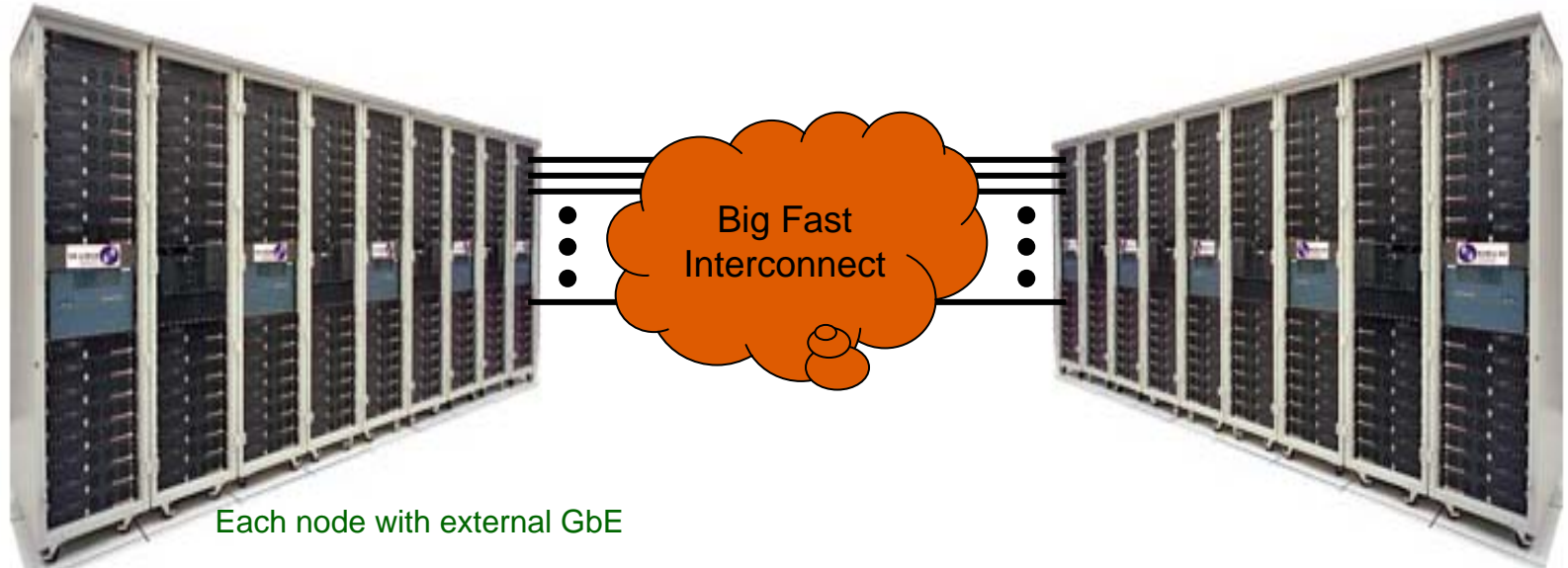
78 MB/s

 64 GB

5  1024 MB

Traditionally, high-performance computers have been islands of capability separated by wide area networks that provide a fraction of a percent of the internal cluster network bandwidth.

To Build a Distributed Cluster




100 TB

5 GB/s

100 TB

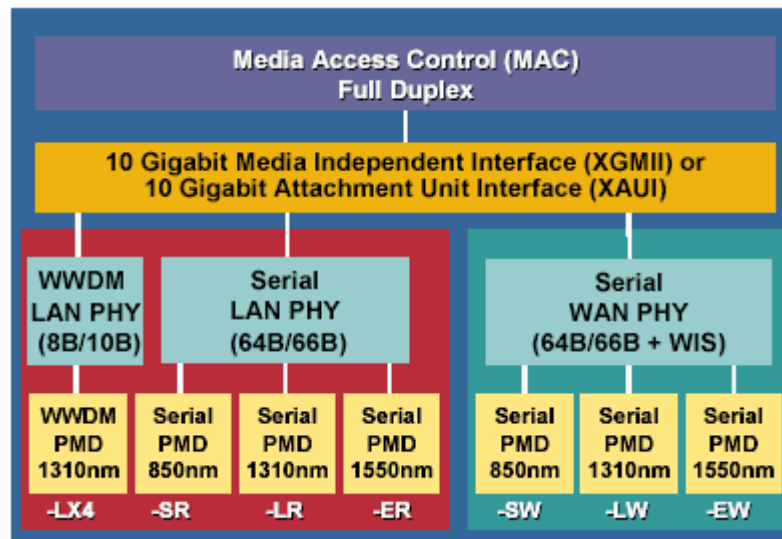
5 GB/s = 200 nodes x 25 MB/s (=20% of GbE per node)

 4096 GB

6  64 GB

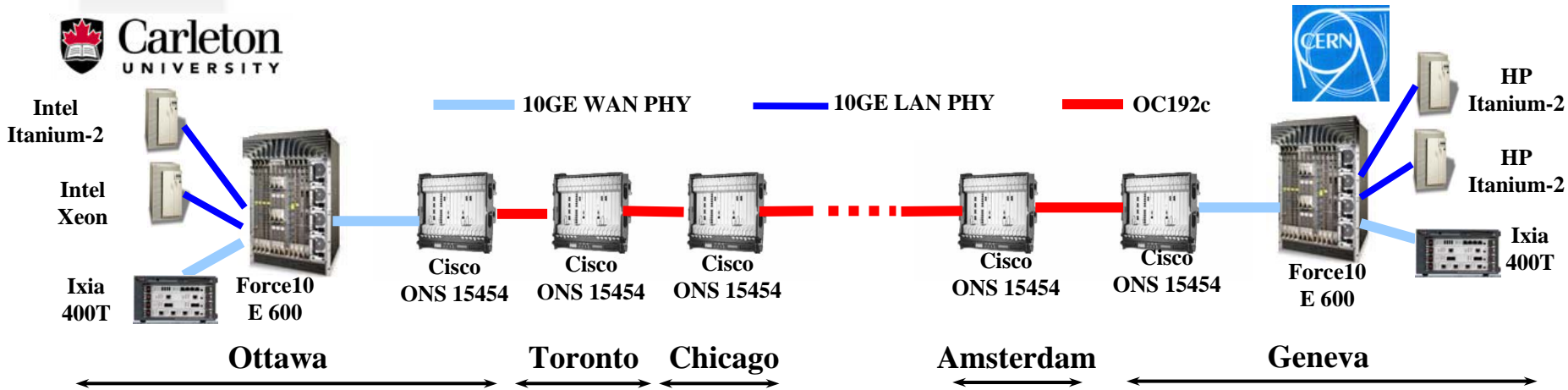
- Existing POS technology relies on expensive underlying SDH / ATM technology to offer service differentiation and resiliency
- Ethernet is the standard for LAN interconnect
- To Move Ethernet into the WAN what is required
 - Service differentiation
 - Resiliency
 - High Speed Interconnect
 - Manageability

- LAN PHY encodes the 10G stream using the 64B/66B PCS (Physical Coding Sublayer)
- WAN PHY uses the same PCS but with rate adaptation to match the SDH payload
- WAN PHY also used the WIS (WAN Interface Sublayer) to provide a simplified SDH framer



- 802.1q and 802.1p provide service separation and differentiation
- 802.3ad provides both resiliency and link aggregation
 - Sub 50ms link failover
 - Force10 E-Series supports up to 16 10GE links in a bundle
 - 802.1w provides sub second topology failover
- 802.3ae provides High Speed Ethernet and a seamless WAN interconnect
- SNMP and RFC 3176 offer a management interface and a way of acquiring flow information

Transatlantic 10 GE



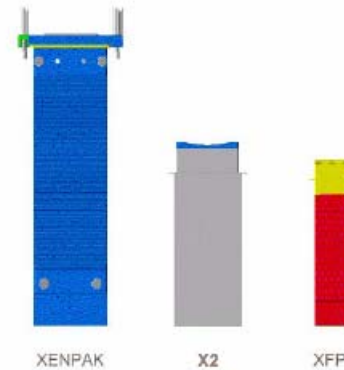
WAN PHY over an OC-192c circuit using lightpaths provided by SURFnet and CANARIE

9.24 Gbps using traffic generators

6 Gbps using UDP on PCs

5.5 Gbps using TCP on PCs

- There are 3 types of pluggable optics
 - XENPAK
 - X2
 - XFP



- XFP Technology allows complete flexibility for a 10GE Solution
 - SR 300m or LR 10km for Campus connectivity
 - ER 40km or ZR 80km for intercampus technology
 - LAN or WAN PHY on a per port basis independent of XFP
- XFP Technology allows us to increase the density of 10GE and reduce the price per port

- In order to use 10GE as a WAN technology the platforms supporting it need to offer carrier grade resiliency and stability
- At a minimum they should offer
 - Passive Copper Backplane
 - Redundant Power and Fans
 - Redundant route processors
 - Redundant switch fabrics
 - Complete separation of control and forwarding
 - No forwarding packets should traverse the control plane (Not even the first packet of an IP Flow)
 - Hitless failover of control and forwarding planes
 - Support for graceful restart protocols

- 14 Slot Chassis
- 56.25Gbps per slot
- 1.6875 Gbps Switching capacity
- 672 ports of line rate GE
- 1260 ports of GE
- 56 ports of line rate 10GE
- Hitless failover of Control and Forwarding plane



TeraScale E-Series Architecture

Resiliency, Scalable Performance & Security



Modular OS with memory protection for stability

FTOS

RPM

ASICs

Switching

Routing

Management

Switch Fabric

Backplane

First Tbps switch/router AND massive ACL security filter scalability (>1M)

3-CPU architecture for resiliency

Resiliency through integrated CPU protection

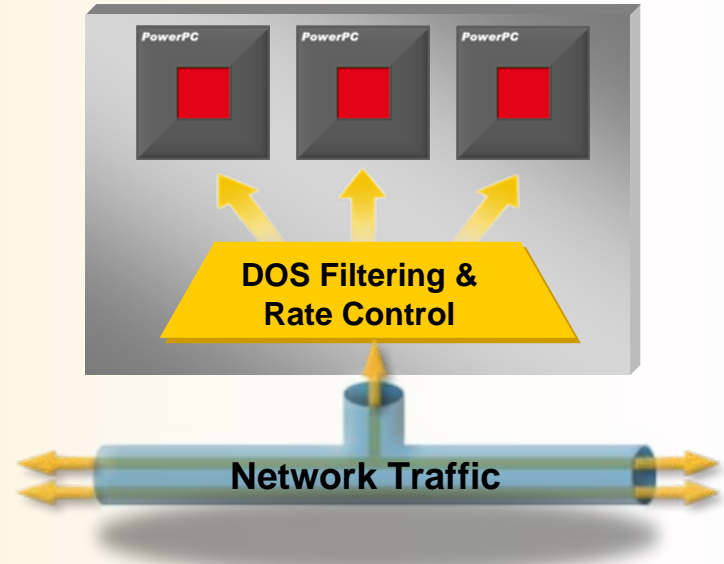
5 Tbps capacity (100 GbE ready) providing long-term investment protection

1.68 Tbps, supporting 672 line-rate GbE ports per chassis

Modular Operating System (FTOS)

3 RPM CPUs, and LC CPUs

Resiliency, Scalability and STABILITY



Control Processor

- CLI, SNMP, Telnet
- Configuration
- Interface Manager
- Event Logger
- Boot Manager

Route Processor 1

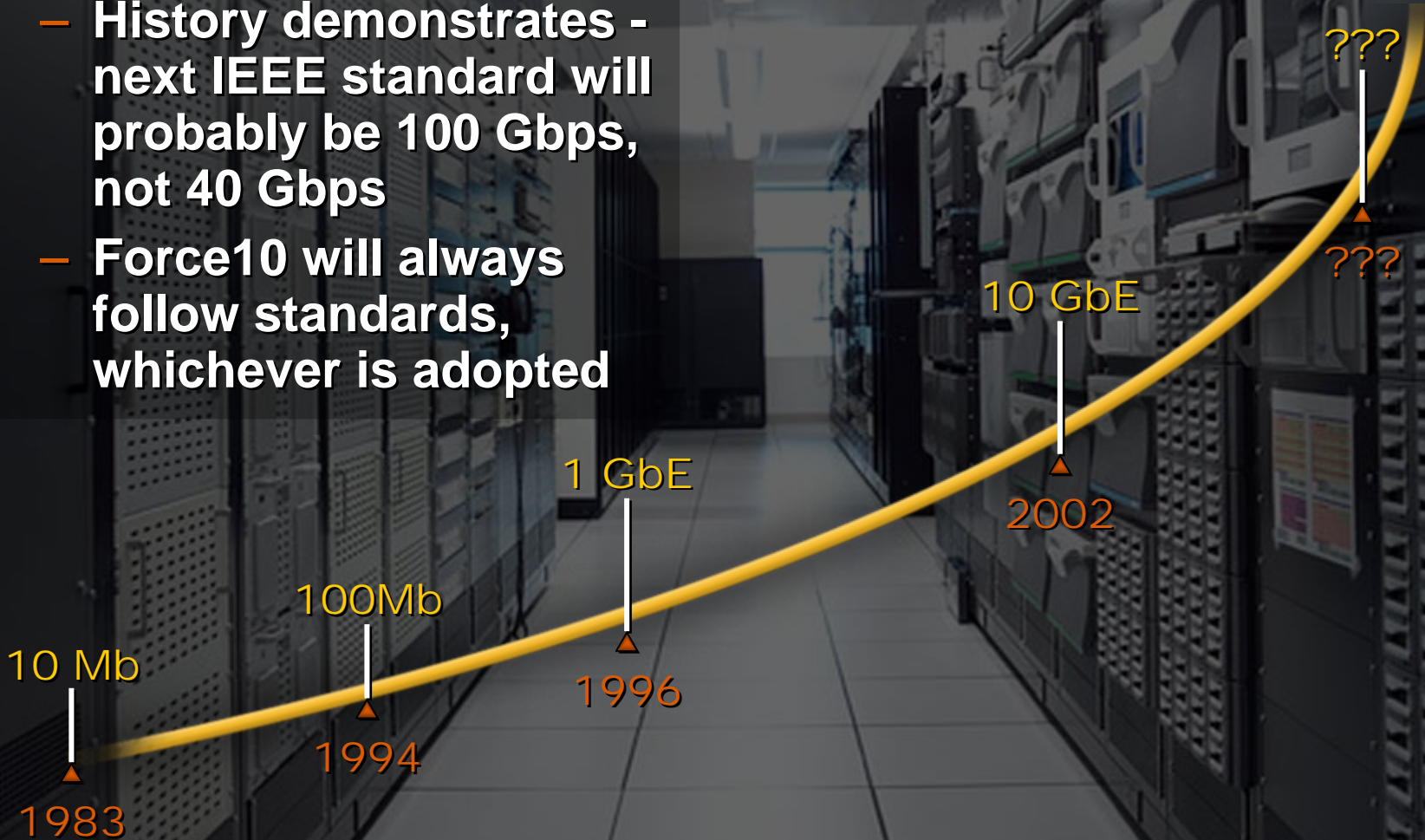
- RIP/OSPF/IS-IS/BGP
- Static Routes
- Route Table Manager
- ACL Manager
- Route Maps
- Redistribution

Route Processor 2

- MAC Manager
- Spanning Tree
- ARP Manager
- Link Aggregation (LAG)
- VRRP, ICMP, PPP

■ **Force10 positions:**

- History demonstrates - next IEEE standard will probably be 100 Gbps, not 40 Gbps
- Force10 will always follow standards, whichever is adopted



■ Pros –

- 40 GbE is the logical next step leveraging SONET / SDH OC-768 / STM-64 PHYs
- 40 GbE can leverage OIF efforts underway to standardize OC768 interfaces

■ Cons

- Cost per port of existing OC-768 technology is too high
 - >\$1M per port
 - Should be 4 x 10 GbE pricing for adoption
- Not a significant enough performance improvement to warrant the cost of development/adoption

- Pros –
 - Next logical step in Ethernet
 - Low cost optics (samples) are becoming available
 - Standards process favors neither 100 or 40 GbE yet since nothing has officially started
 - Development cost of 40GbE vs 100GbE will be about the same, why not go for the higher bandwidth that will scale for future applications
 - Easier to get to ASP goal for technology to get to mass adoption
 - GbE - \$500/port
 - 10GbE - \$2.5K/port
 - 40GbE – 4.5K/port
 - 100GbE - \$10K/port

- Backplane that's supports required bandwidth
 - 334G is required per slot
- Optics Supporting 100G Over a Single Fiber
 - Four lambda at 25 Gb/s per lane
 - Ten lambda at 10 Gb/s per lane
- High Speed Silicon Support from Component Suppliers
 - PHY Silicon has to run at 25Gb/s
- System Design Issues
 - 420W required per slot for 100GE
- Next Generation ASIC Development
 - High speed interconnects to RAM
 - Serdes technology for all interfaces

- Force10 Guides the Industry in key Technologies
 - Joel Goergen, Chair, ad-hoc subcommittee IEEE P802.3ap Task Force – “Backplane Channel Model”
- Working with leading component suppliers
 - 40 Gbps and 100 Gbps Ethernet
 - Next generation Backplane SERDES devices for higher speed



Thank You

FORCE ™