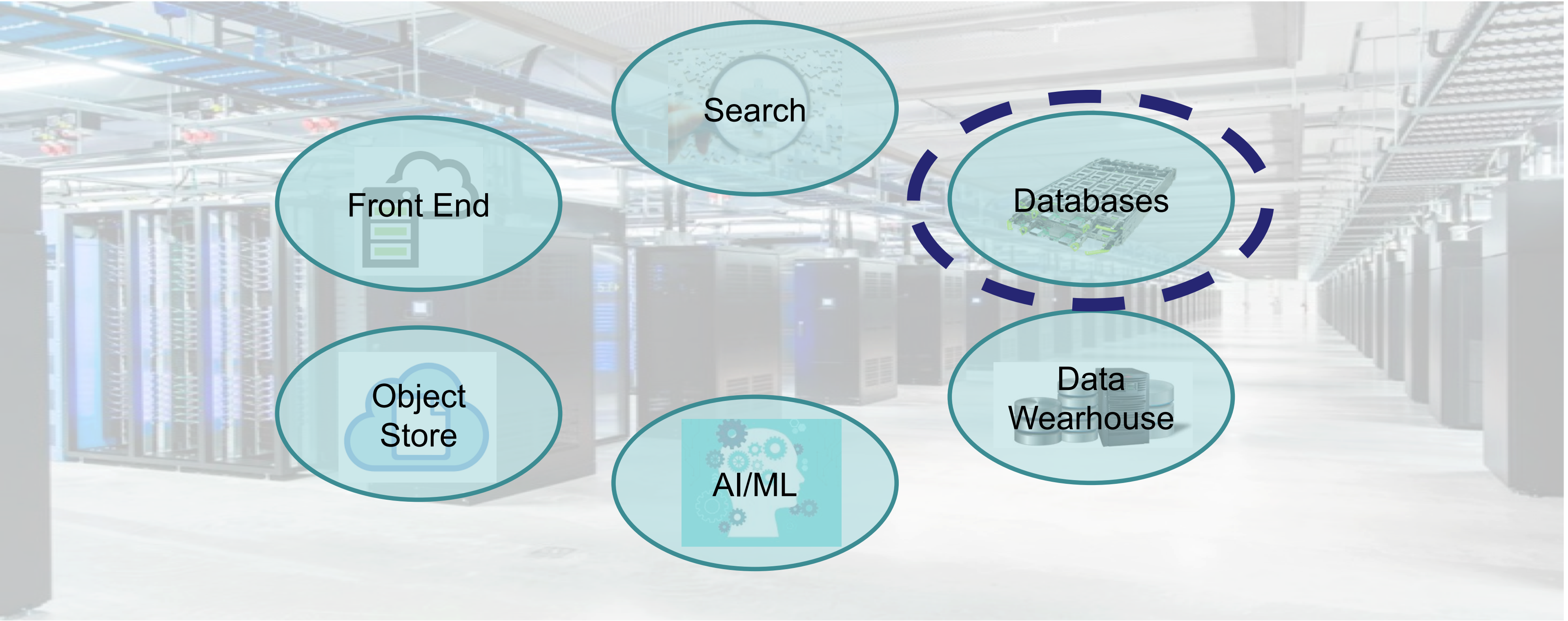
Flash Memory Summit

INVT-202A-1: Handling the Network Requirements of High-Speed NVMe SSDs

Manoj Wadekar, Storage Engineer, Facebook
Rob Davis, VP Storage Technology, Mellanox

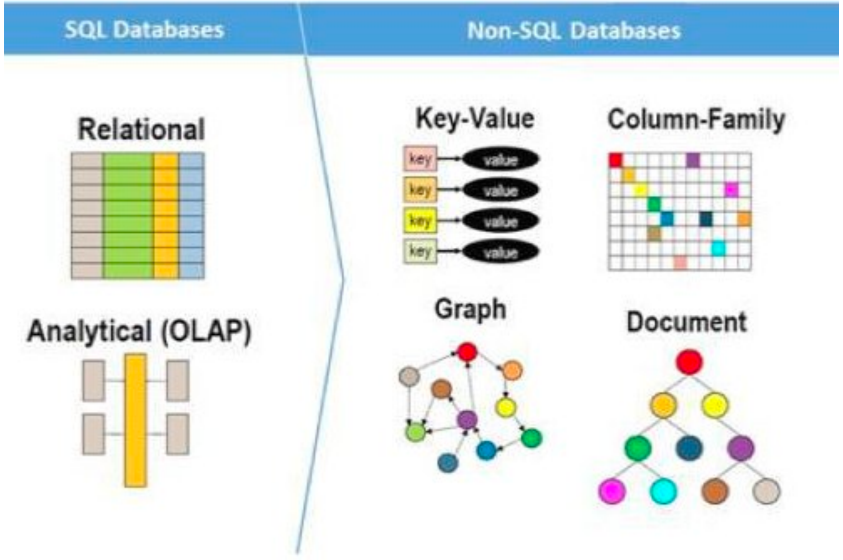
Typical Hyperscale Server Infrastructure

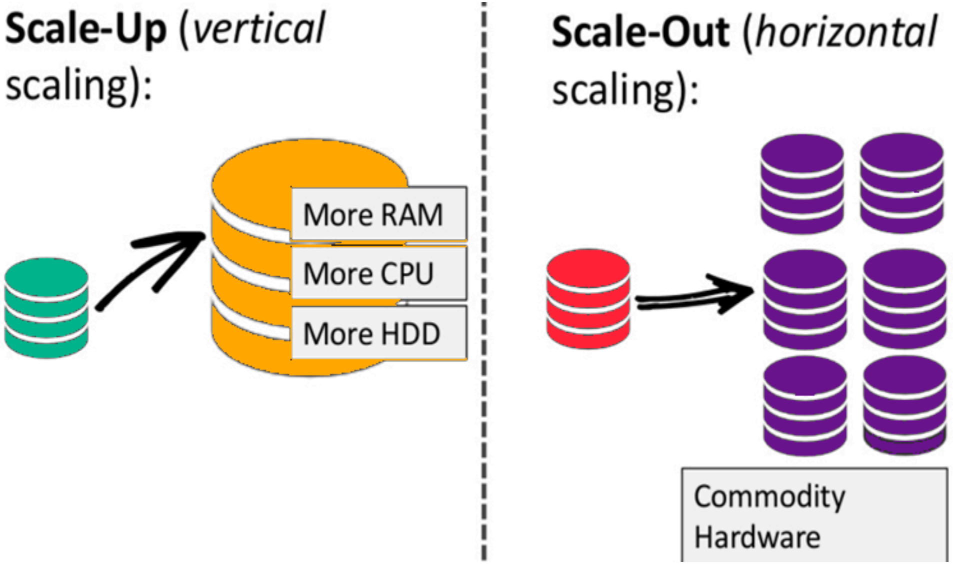
Flash Memory Summit



Flash Memory Summit

Databases for Hyperscale



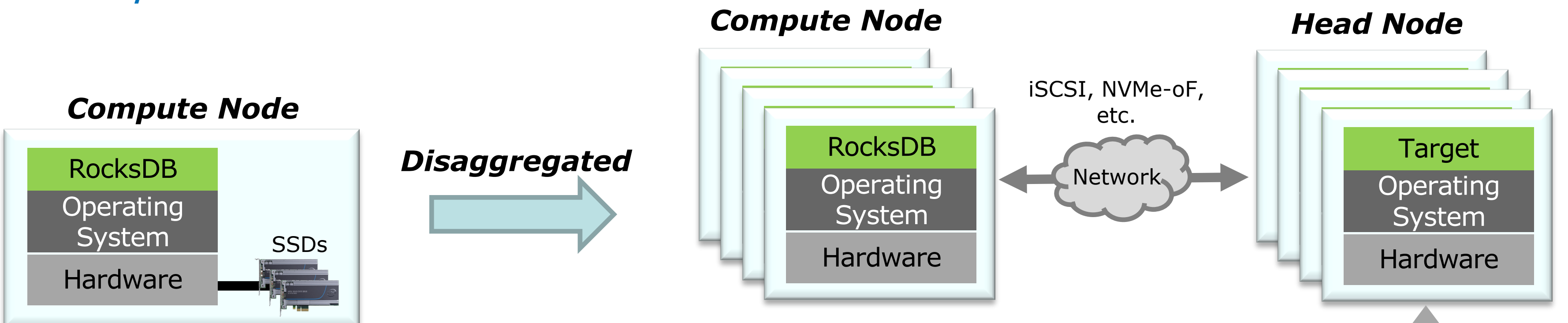


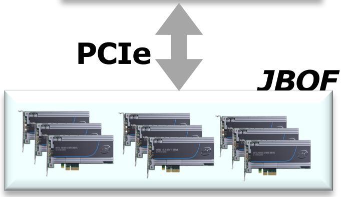
Open Source Software

Open Hardware: E.g. OCP

Ref: https://github.com/UWCoffeeNCode/resources/wiki/SQL-and-NoSQL-Databases

Disaggregated Storage Architecture



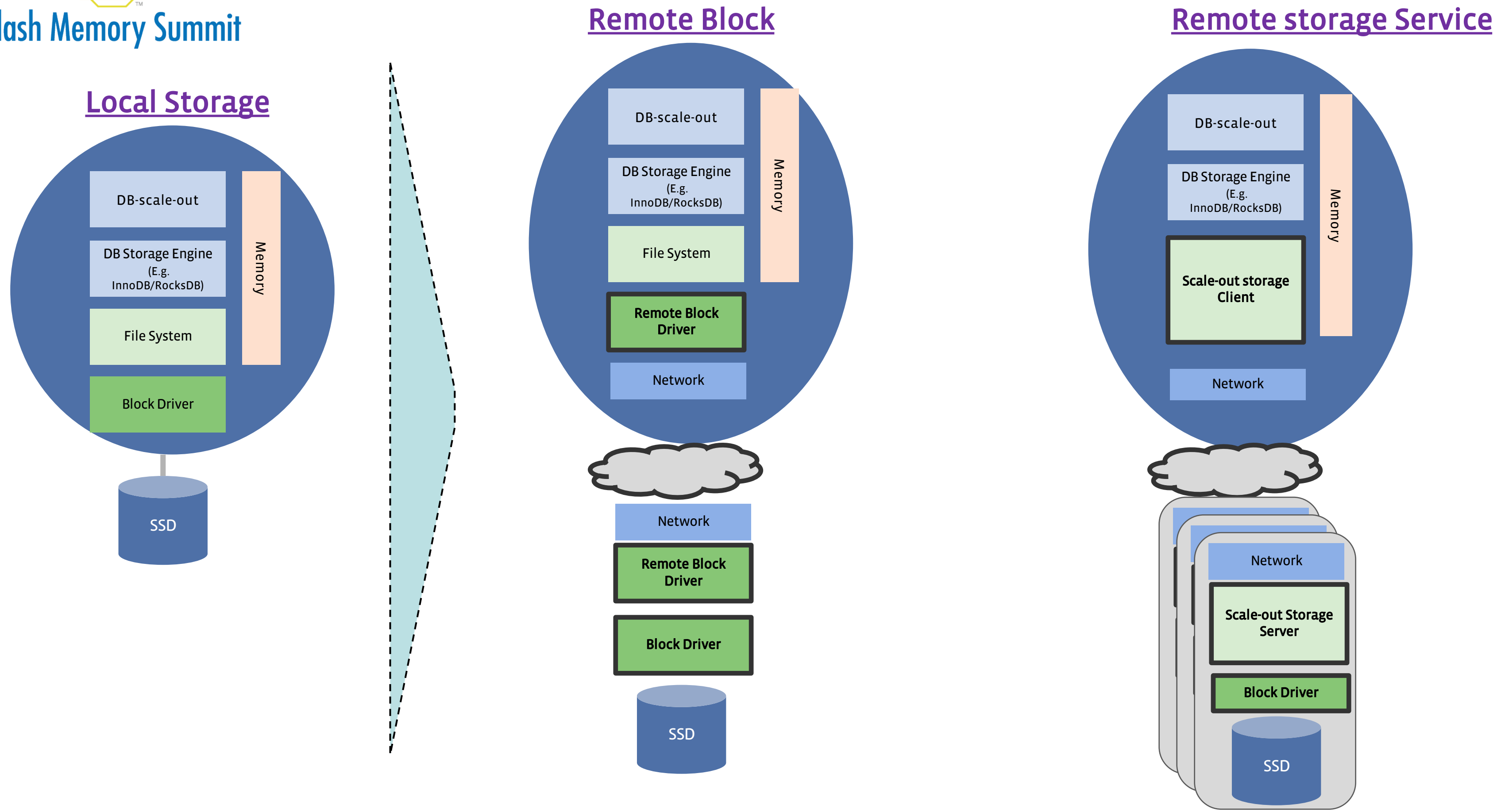


- **Local attached storage**
- **Static binding**
- **Stranded capacity**, IOPS
- Inefficient, increased TCO

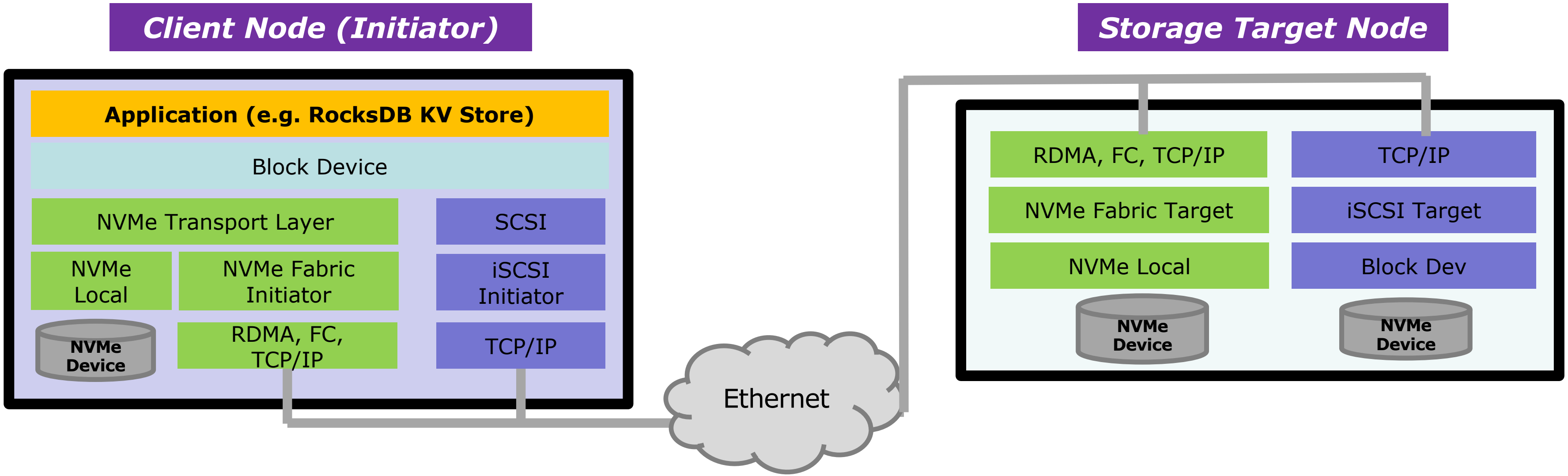
- **Logical disaggregation**
- Consumes physical or logical block devices
- **Dynamic binding** based on workload requirements
- Efficient, improved TCO

- **Physical disaggregation**
- **Static binding**
- **Shared resources**
- Target can expose physical or logical devices

Storage Disaggregation

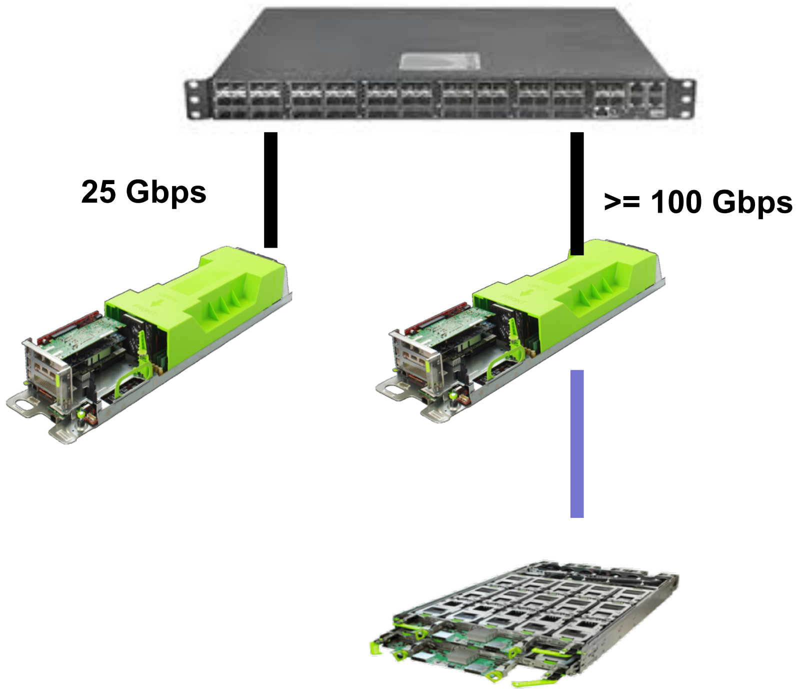


Remote Block Storage



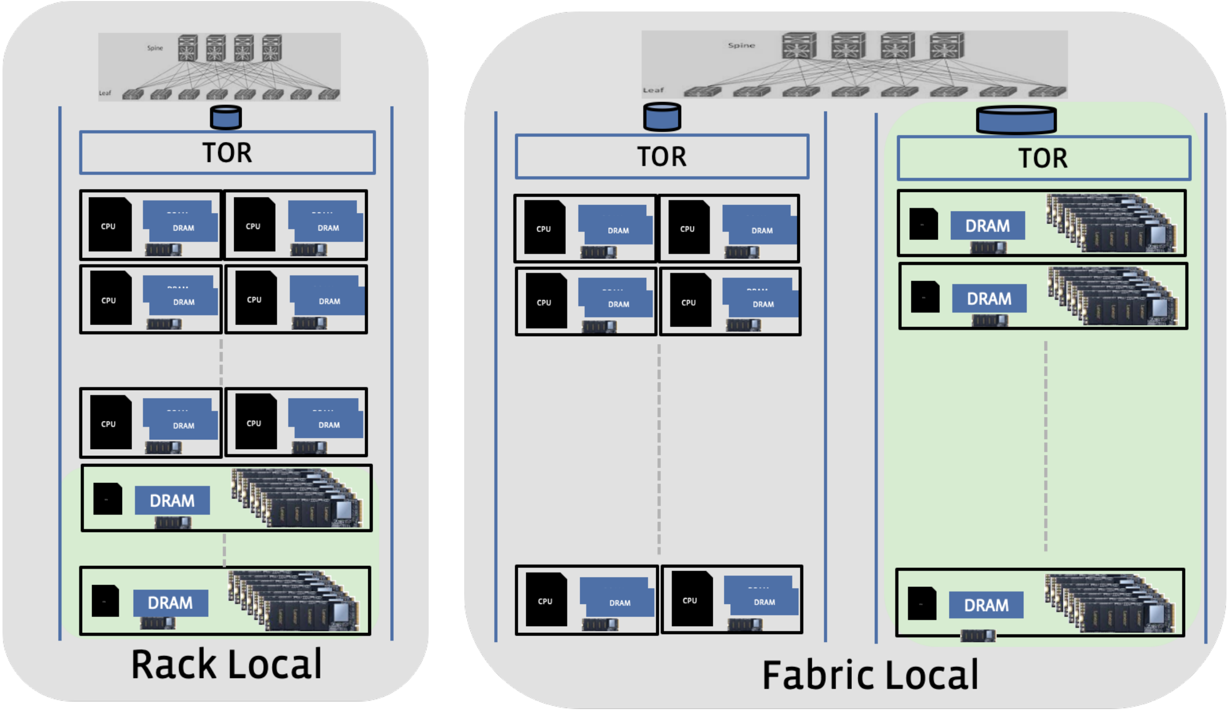
- **Enables sharing** of NVMe flash storage over network
- Can use **traditional** block protocols (e.g. iSCSI) or **NVMe optimized** protocols (e.g., NVMe/TCP)
- **NVMe over Fabrics** – supports multiple transports, extends **NVMe efficiency over network**
  - Poll and interrupt mode architecture
  - Kernel and user mode implementations

Remote Block Storage: OCP HW



- FB Lightning supports 30 M.2 NVMe SSDs
- Storage can be accessed over Ethernet using Tioga Pass server

Network: Speed



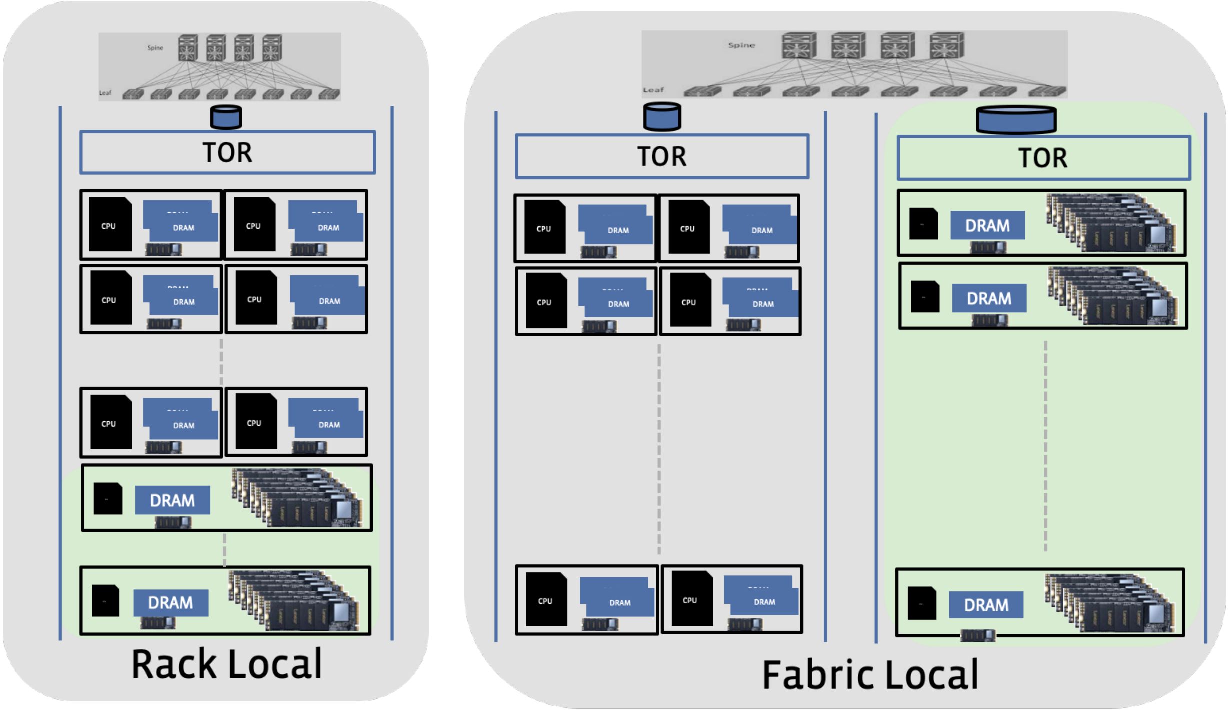
100s Gbps

Tbps

- Disaggregation
- Multi-tenant access
- High bandwidth applications
- ..

Flash Memory Summit

Network: Latency



100s Gbps

Tbps

- Link Latency
- Stack Latency
- Number of hops
- Queuing
- Congestion
- ..

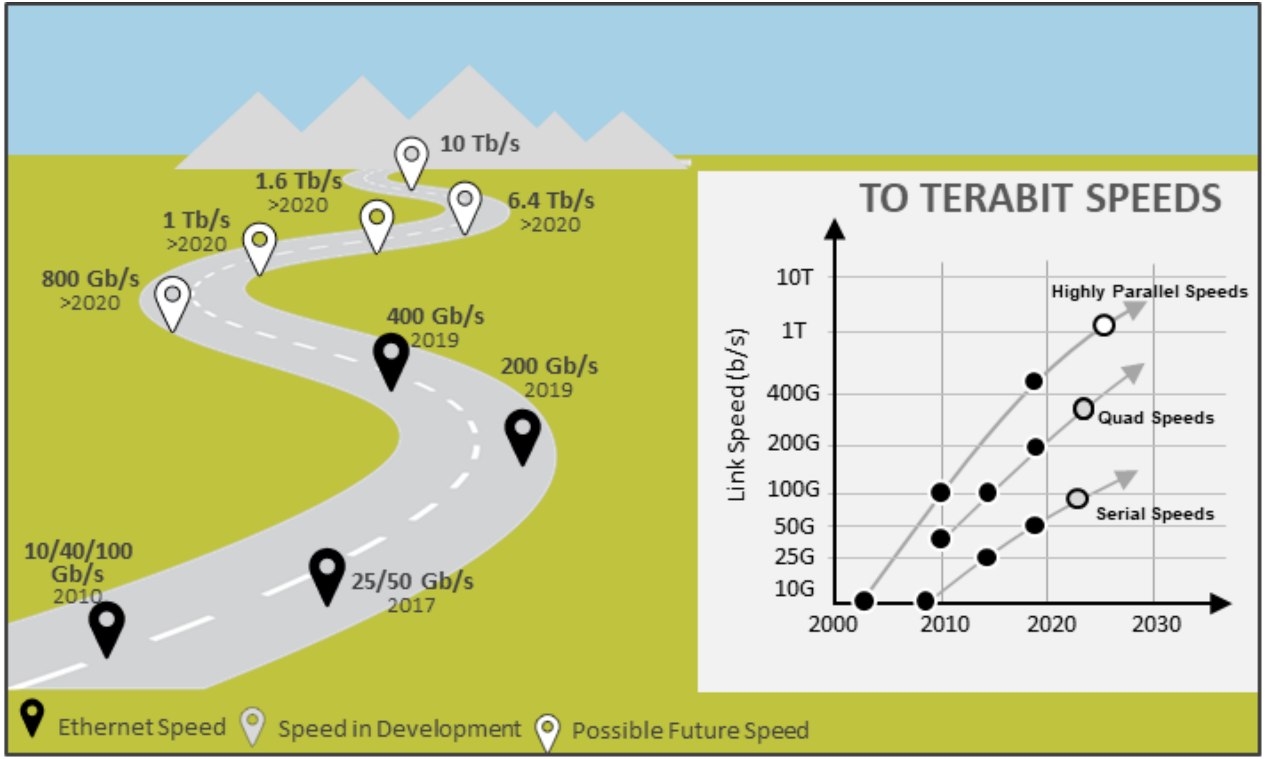
Summary – hyperscale storage

Flash Memory Summit

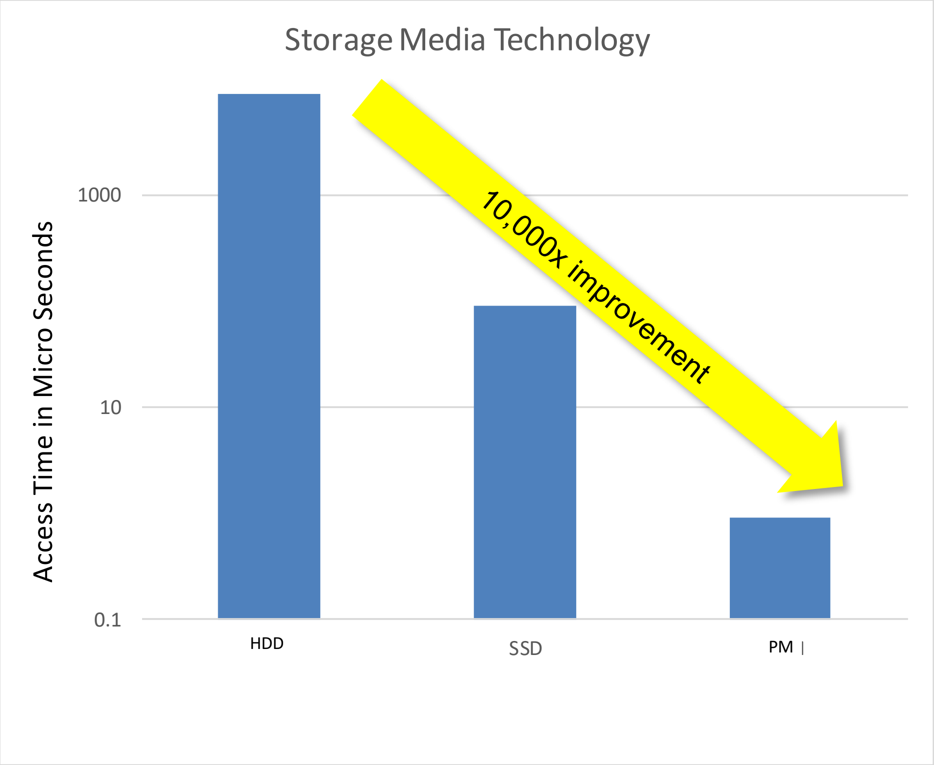
- Storage Disaggregation important for hyperscale efficiency
- It depends on high performance network
- Is Ethernet Ready?

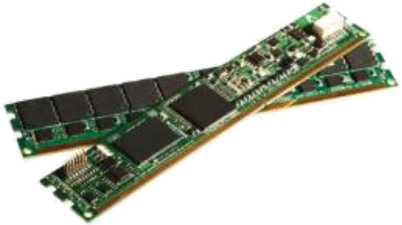
Ethernet is Ready!

Ethernet Technology Roadmap

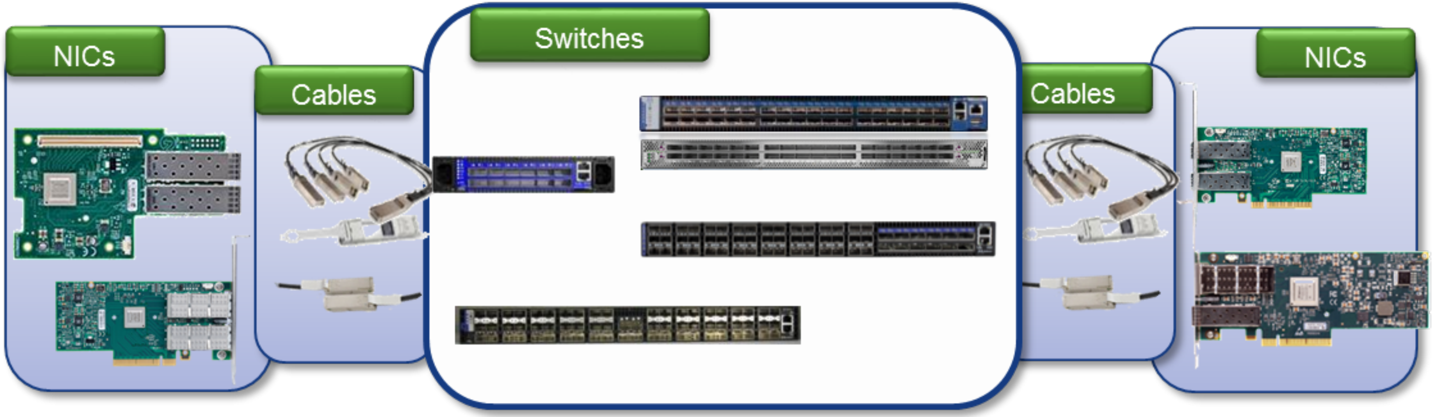


We Need Network Speed for Flash!





High Performance Networking is Here Today

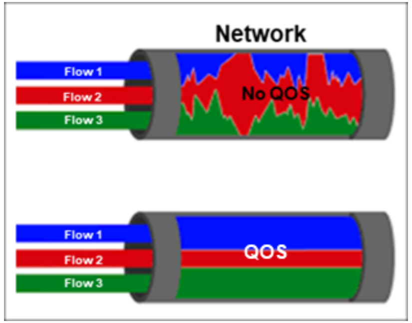


End-to-End 25, 40, 50, 100, 200GbE and soon 400Gb

But High Bandwidth is Only Part of the Solution

- We also Need Low Latency

- Effective Performant Network Congestion Control



- QOS
- Security

Low Latency

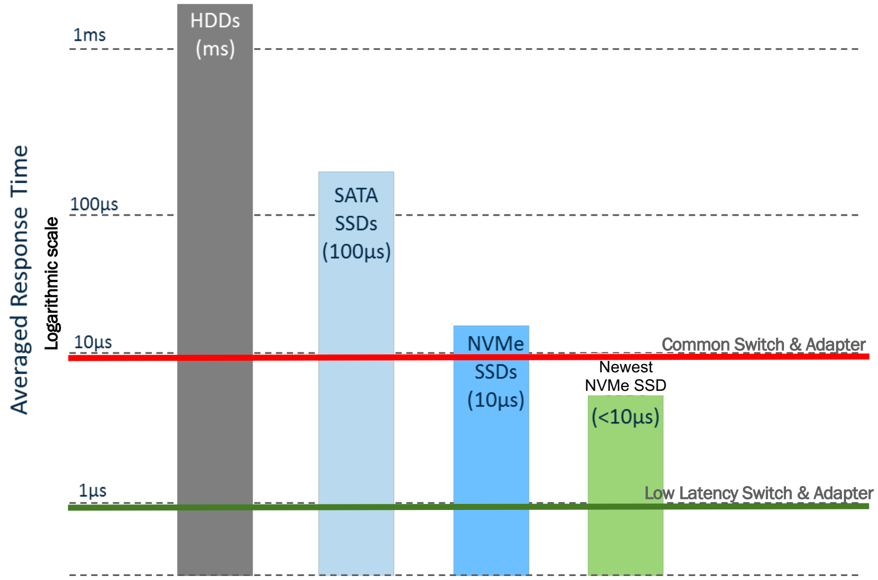
Network Congestion Control

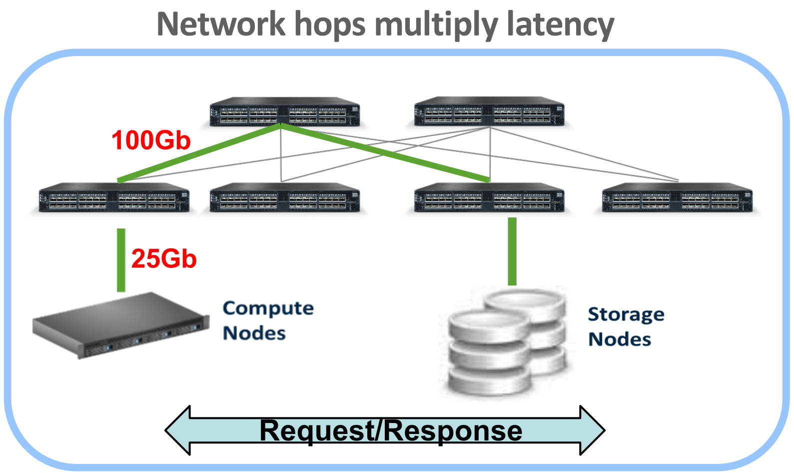
QOS

Security

At Scale

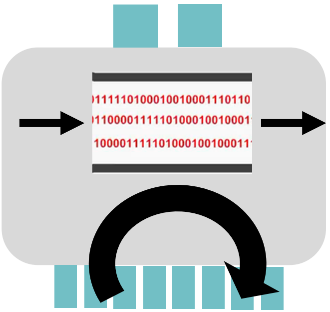
Importance of Latency with Flash Storage





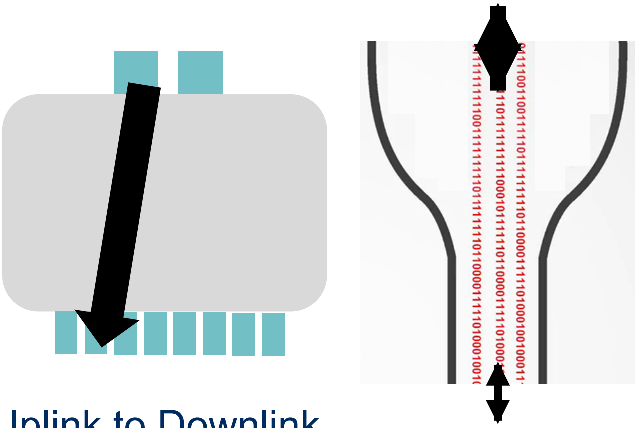
Flash Memory Summit

Intelligent Cut-Through Reduces or Eliminates Store & Forward Latency



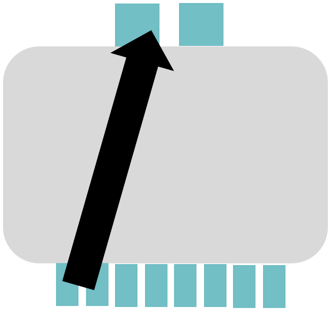
Downlink to Downlink

Full Cut-Through



Uplink to Downlink

Full Cut-Through

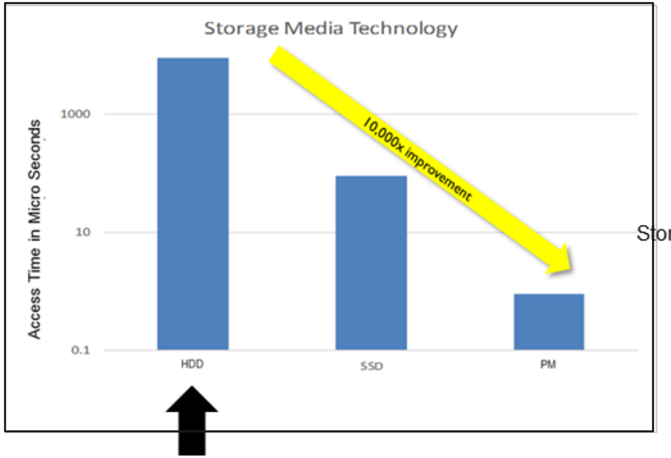


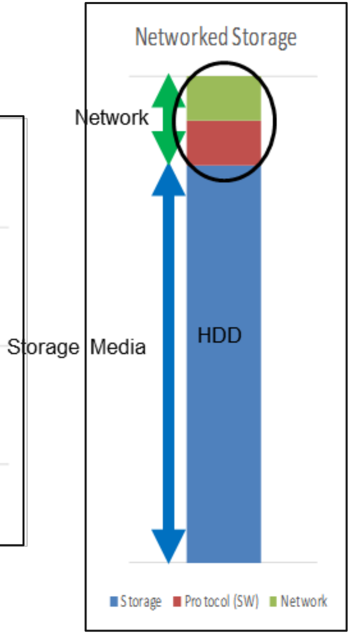
Downlink to Uplink

Smart Store and Forward

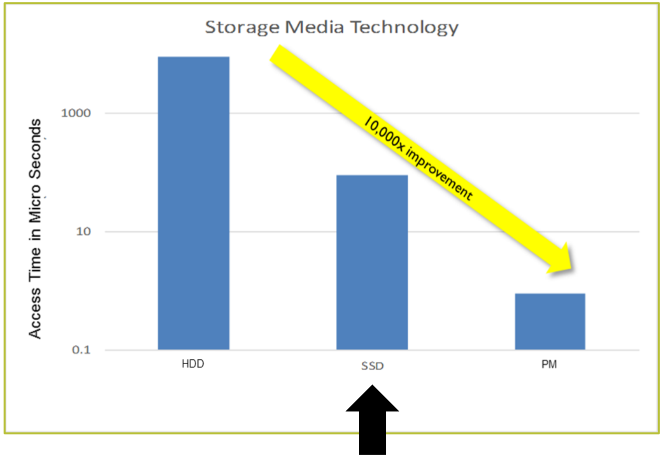
Intelligent cut-through also reduces network congestion

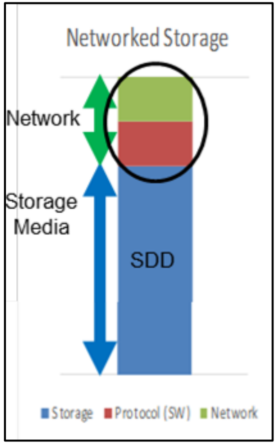
Latency is Not Only About the Hardware



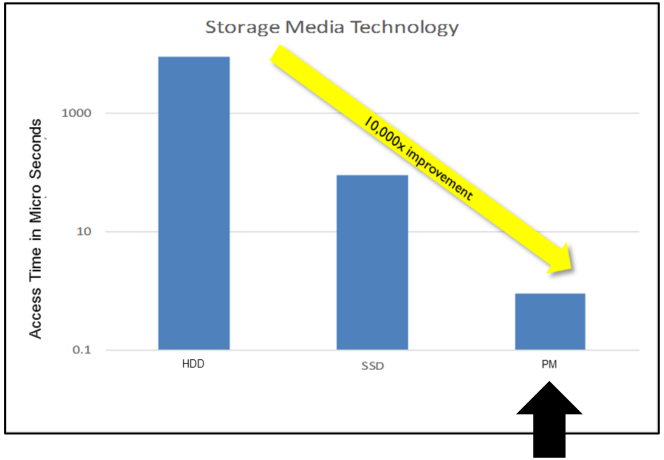


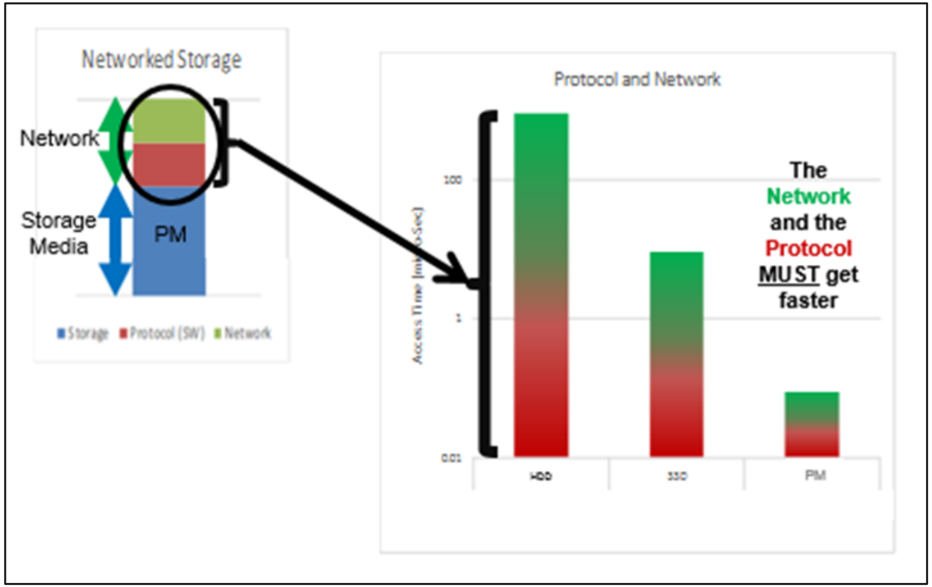
Latency is Not Only About the Hardware



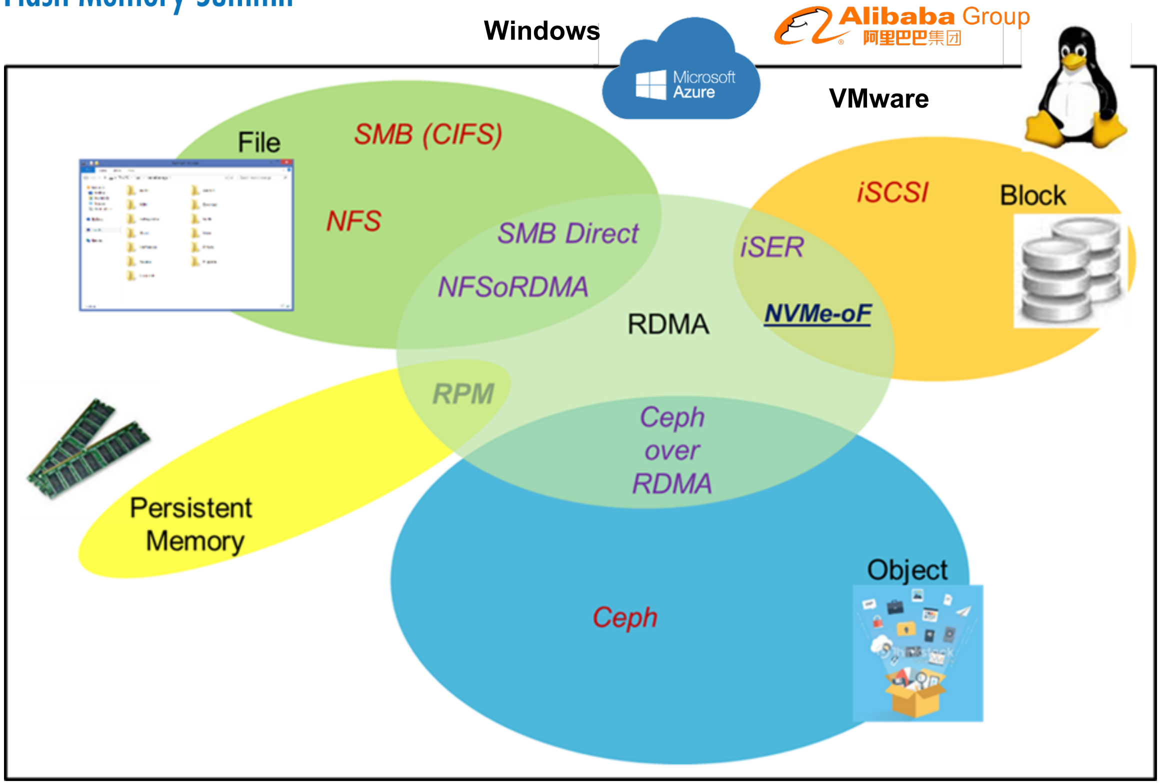


Latency is Not Only About the Hardware

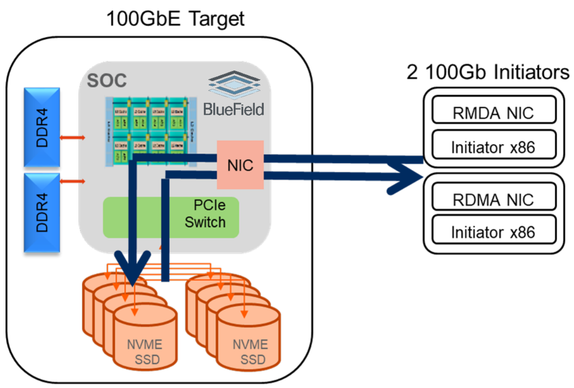




RDMA protocol



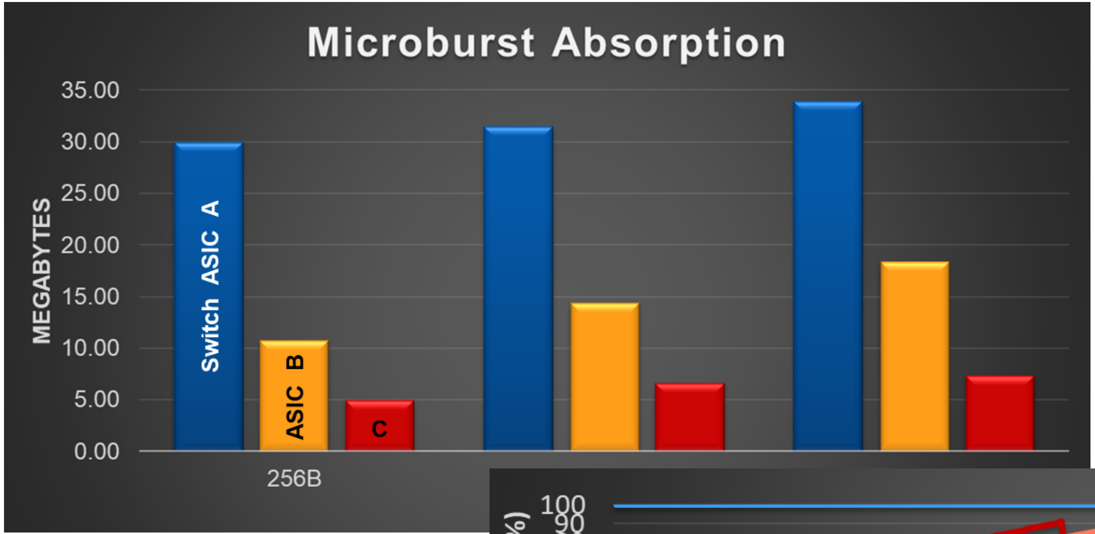
NVMe-oF over RoCE

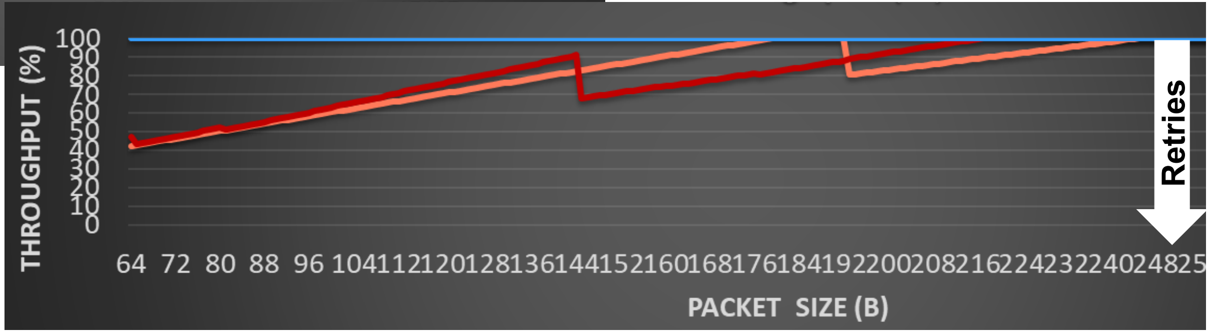


- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- 0.01% CPU utilization
- ~3usec latency

Effective Performant Network Congestion Control

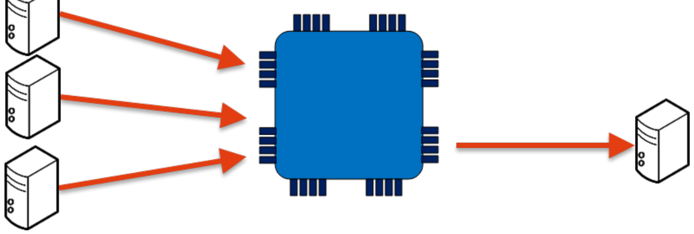
Switch Buffer Size and Congestion



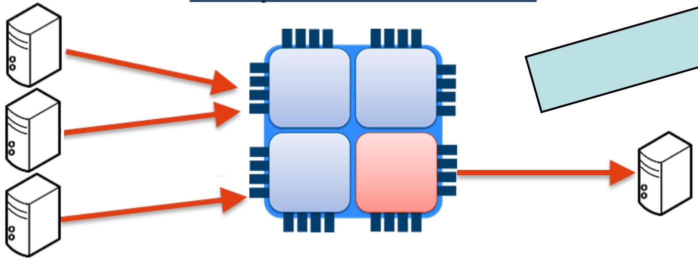


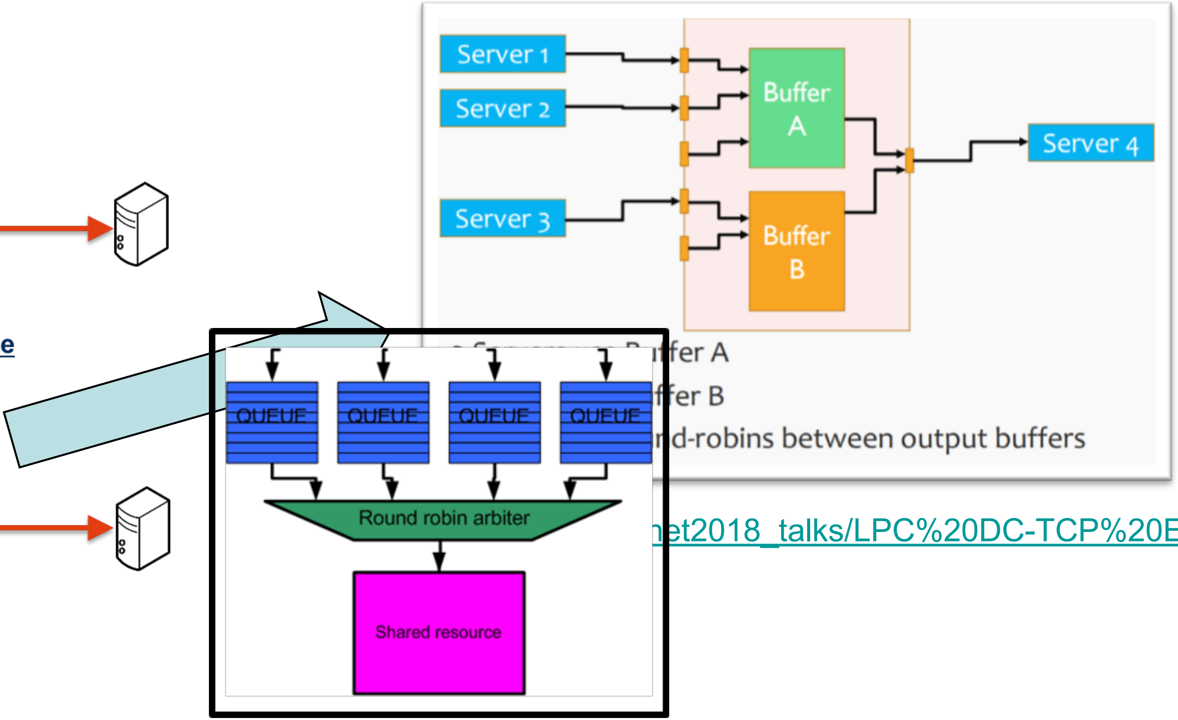
Fairness in Switch Architecture

Fully shared packet buffer



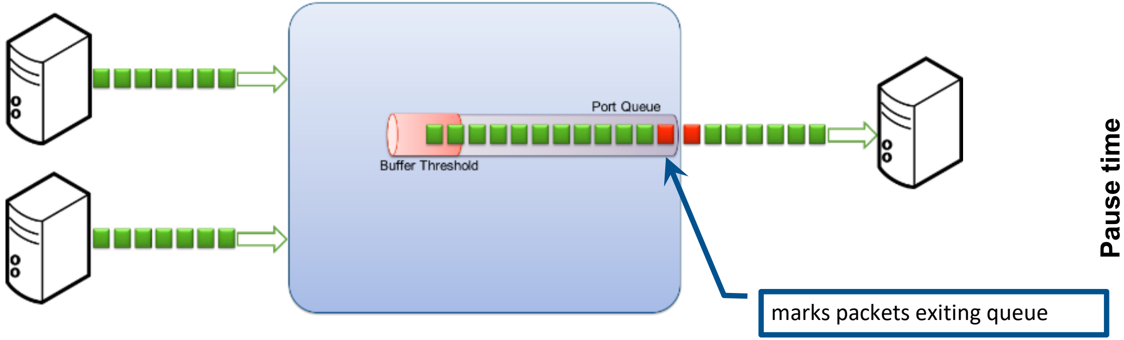
Multi packet buffer scheme



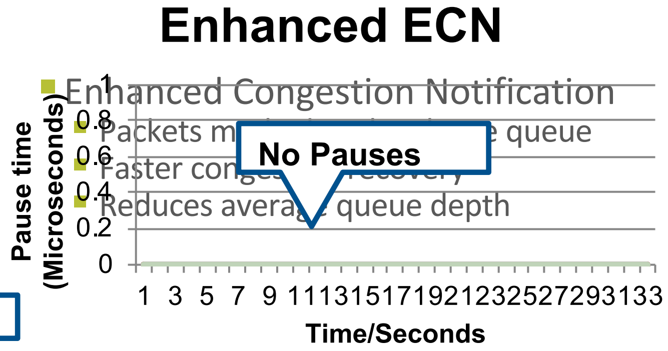


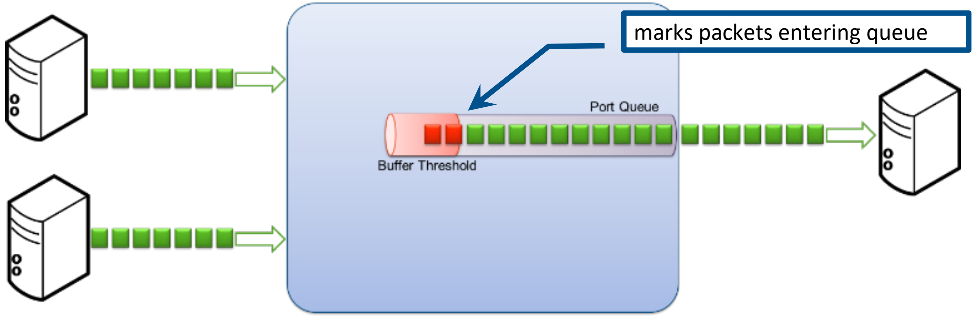
et2018_talks/LPC%20DC-TCP%20Eval.pdf

All ECN is Not Equal

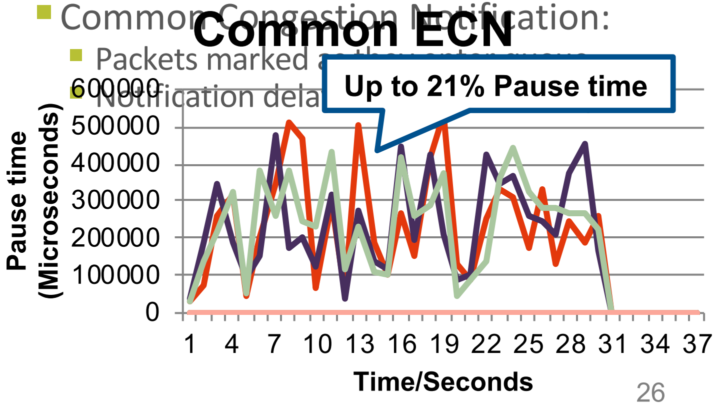


- Enhanced Congestion Notification
  - Packets m… e queue
  - Faster cong… recovery
  - Reduces average queue depth



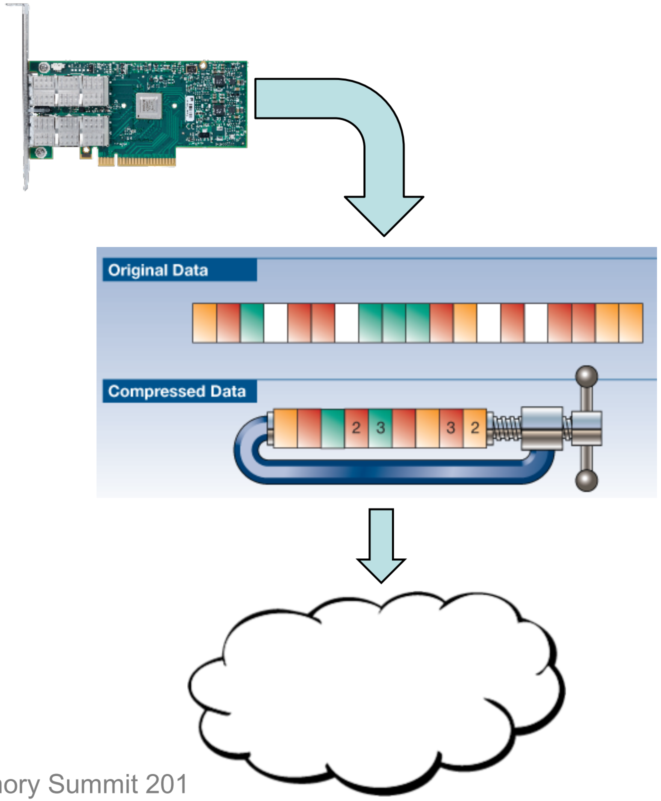


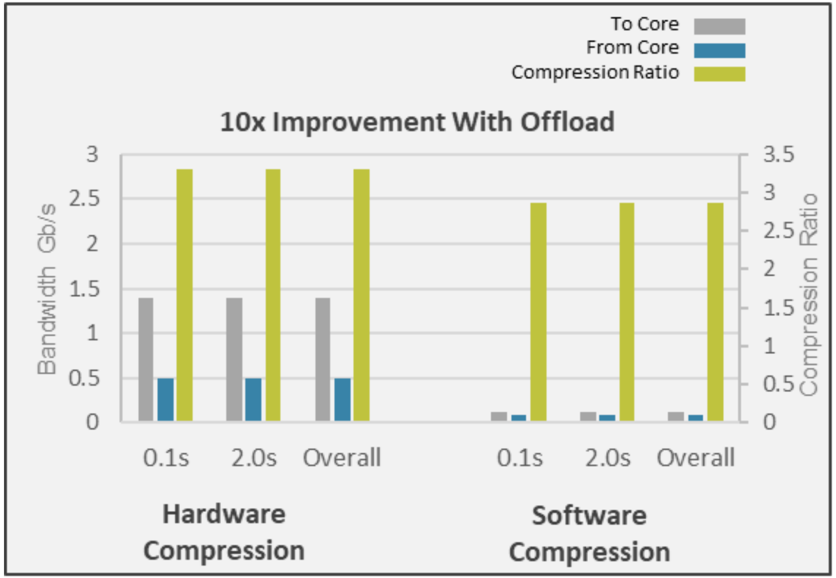
- Common Congestion Notification:
  - Packets marked a… enter queue
  - Notification dela…



Flash Memory Summit

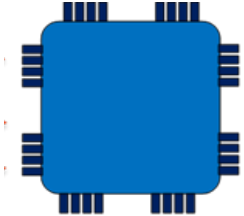
Reduce the Data Before Sending



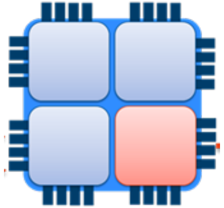


Quality of Service (QOS)

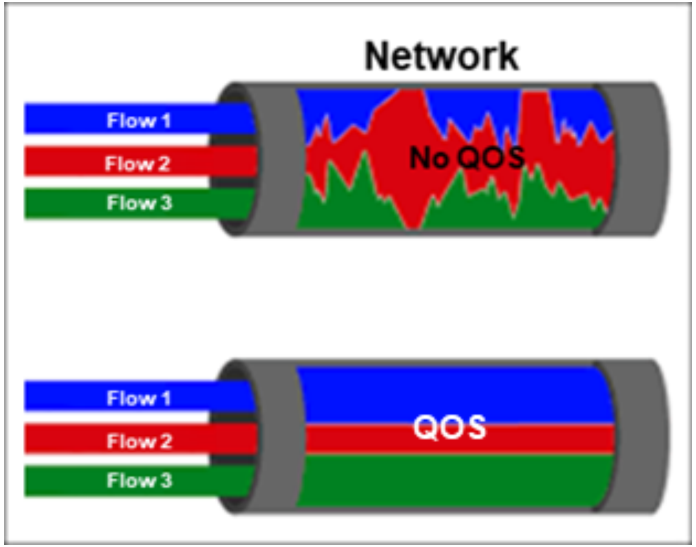
- A fully shared buffer architecture is best for implementing effective QOS
- The QOS algorithm must adapt the bandwidth allocation to the incoming priorities at wire speed



Fully Shared buffer

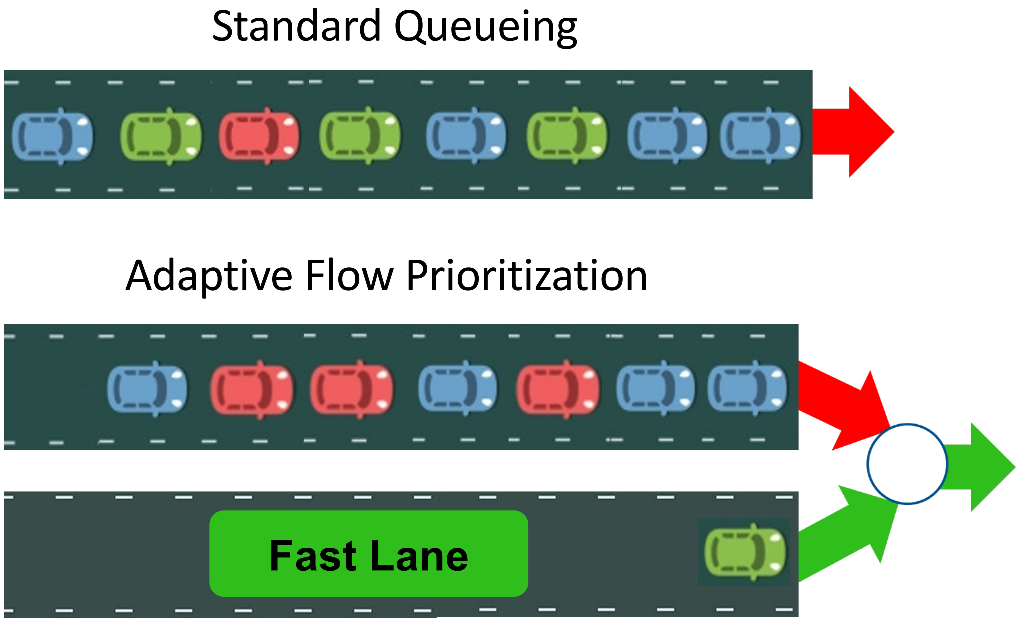


Multi-buffer



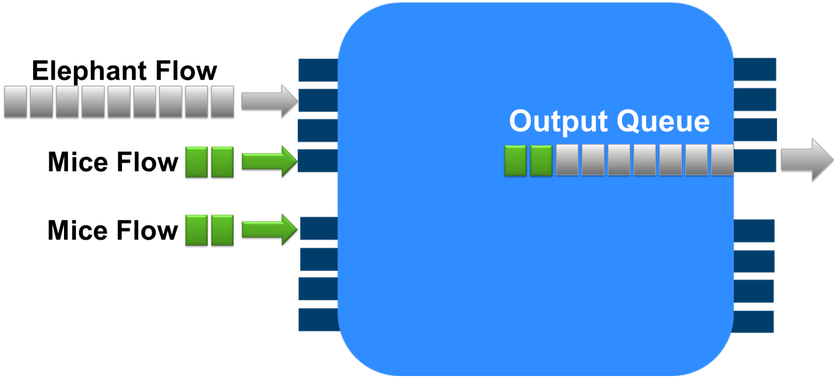
Adaptive Flow Prioritization

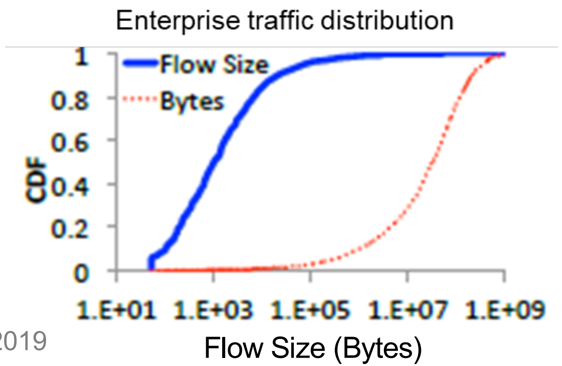
- Egress flow prioritization
- Short flows get benefits
- Reduce flow completion time



Elephants and Mice

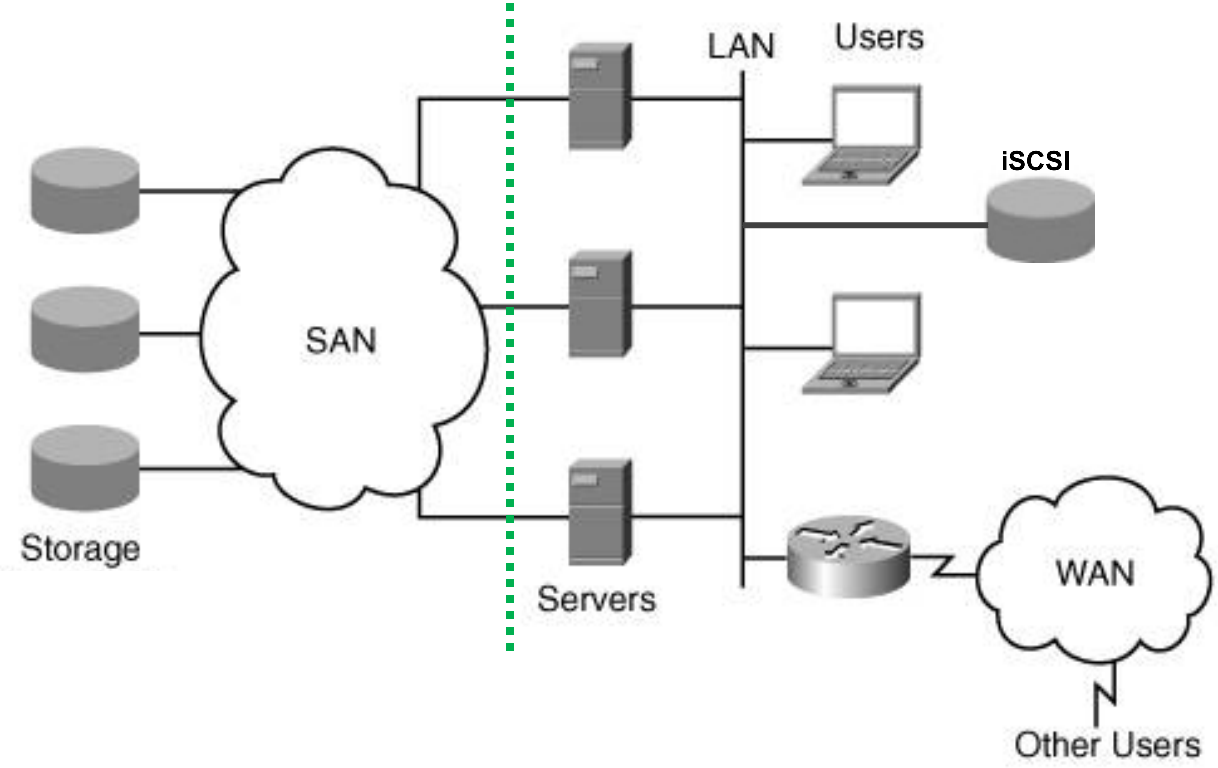
- Majority of flows in the datacenter are small – Mice Flows
- Majority of packets belong to a few large flows – Elephant Flows



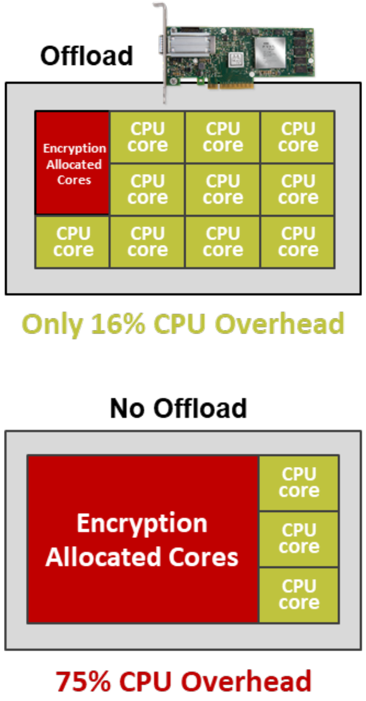


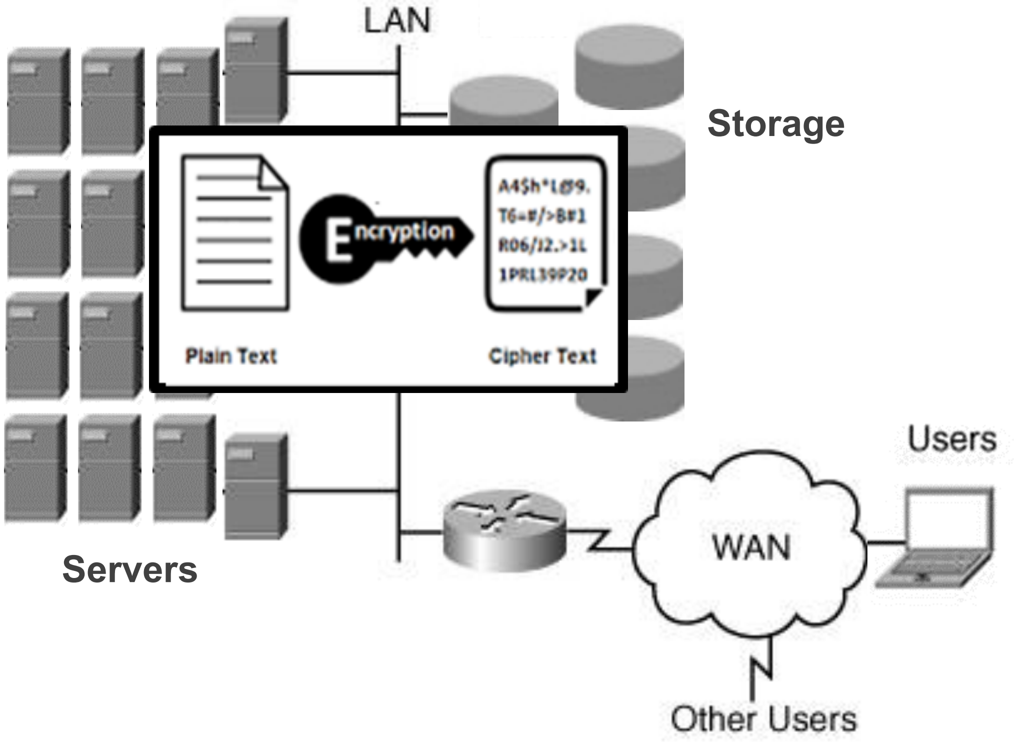
Security

Isolated Fibre Channel SAN

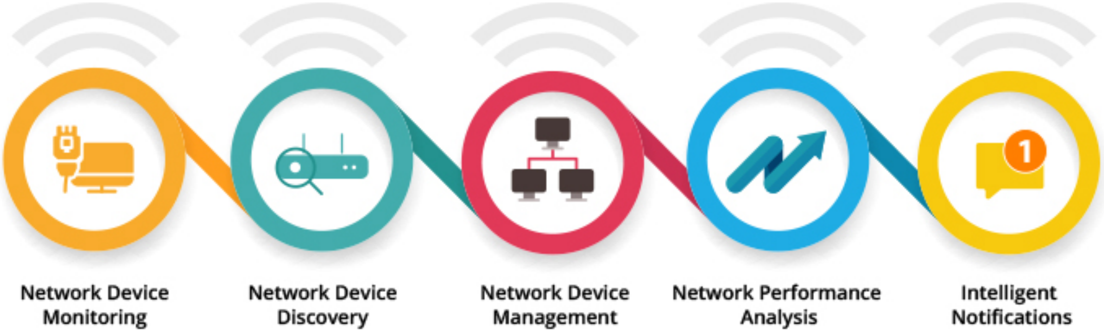


No Longer Isolated Storage

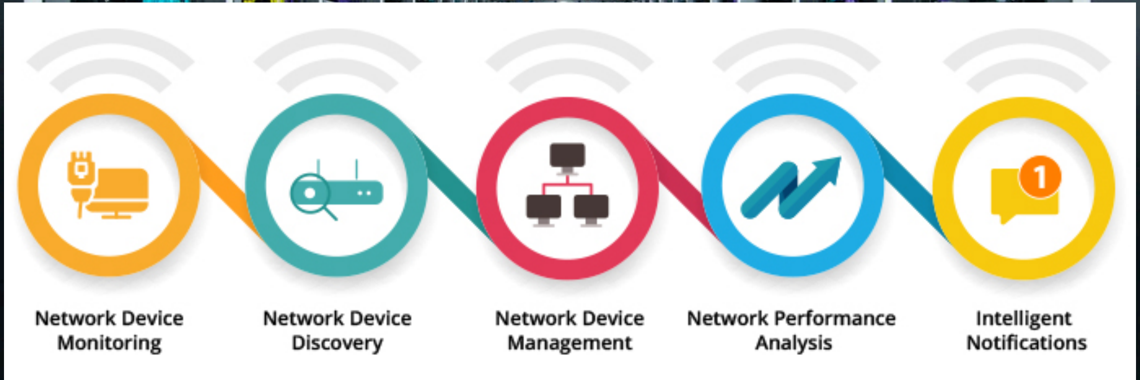




Management

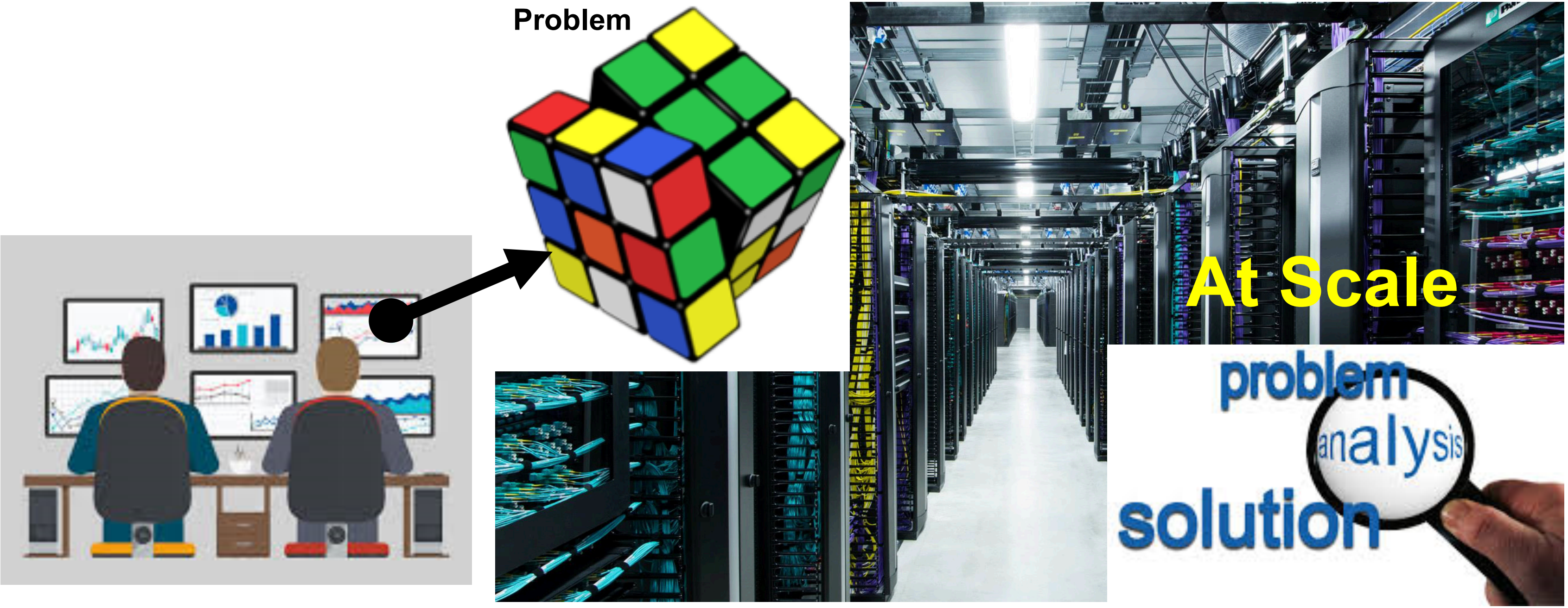

Network Device
Monitoring
Network Device
Discovery
Network Device
Management
Network Performance
Analysis
Intelligent
Notifications

Management



At Scale

Normal Operation is Not the Issue



The Answer is Automated Telemetry Capture and Analysis

Successful analysis of the telemetry data…

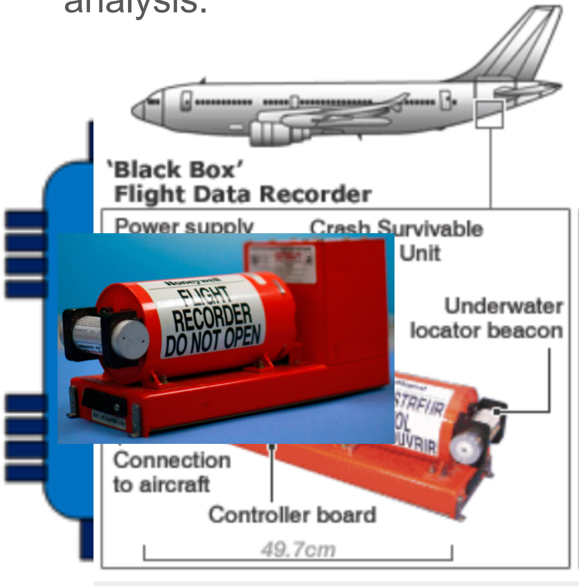
…leads to root cause

What is Telemetry?

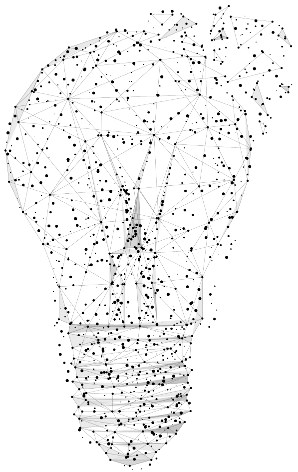
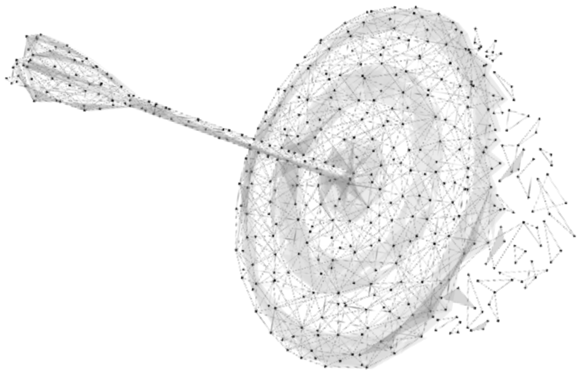
The Important Questions

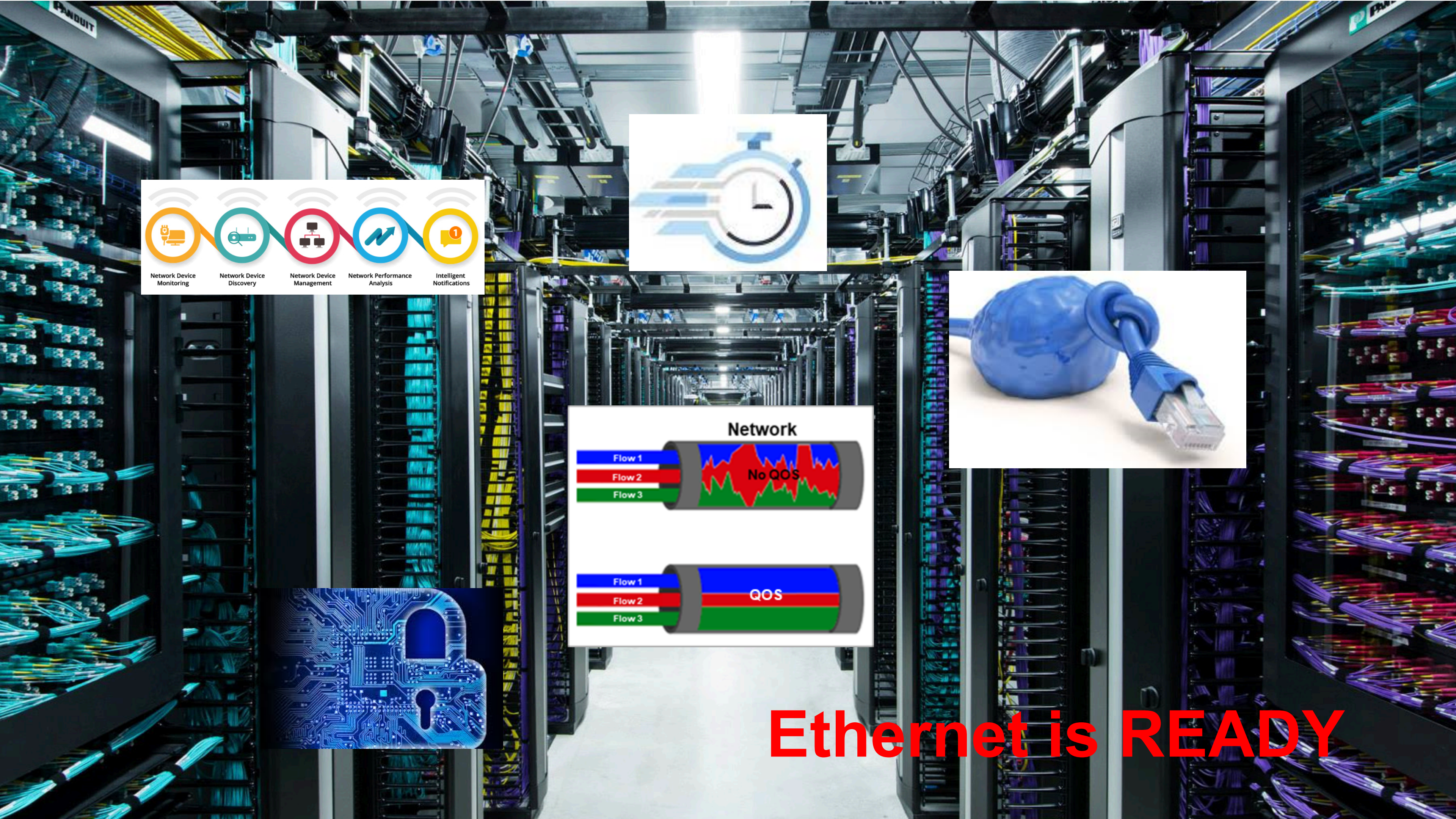
- ✓WHO is being impacted
- ✓WHEN it happened
- ✓WHAT is causing the problem
- ✓WHERE is the problem
- ✓WHY it is happening

Telemetry is an automated communications process by which measurements and other data are collected at remote or inaccessible points and transmitted to receiving equipment for monitoring and analysis.



Telemetry Tell “What Just Happen?”

Faster Time to Innocence	Faster Time To Resolution	Get more out of the Network
		


Network Device Monitoring
Network Device Discovery
Network Device Management
Network Performance Analysis
Intelligent Notifications
Network
Flow 1
Flow 2
Flow 3
No QOS
Flow 1
Flow 2
Flow 3
QOS
Ethernet is READY

Flash Memory Summit

Questions?

Thank You!