NVMe-oF SAN Telemetry with Programmable Switches

Surendra Anubolu, Maggie Gao and Mohan Kalkunte

Broadcom Inc

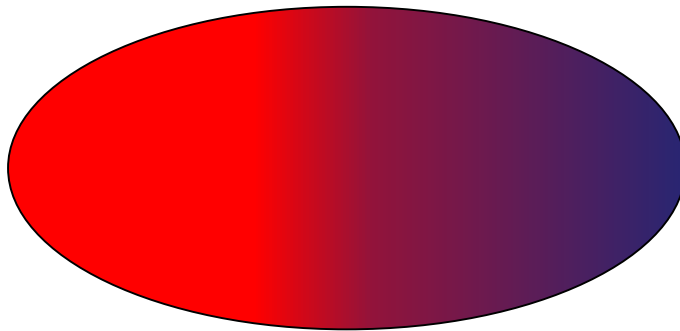# Fabric Solution Key Parameters

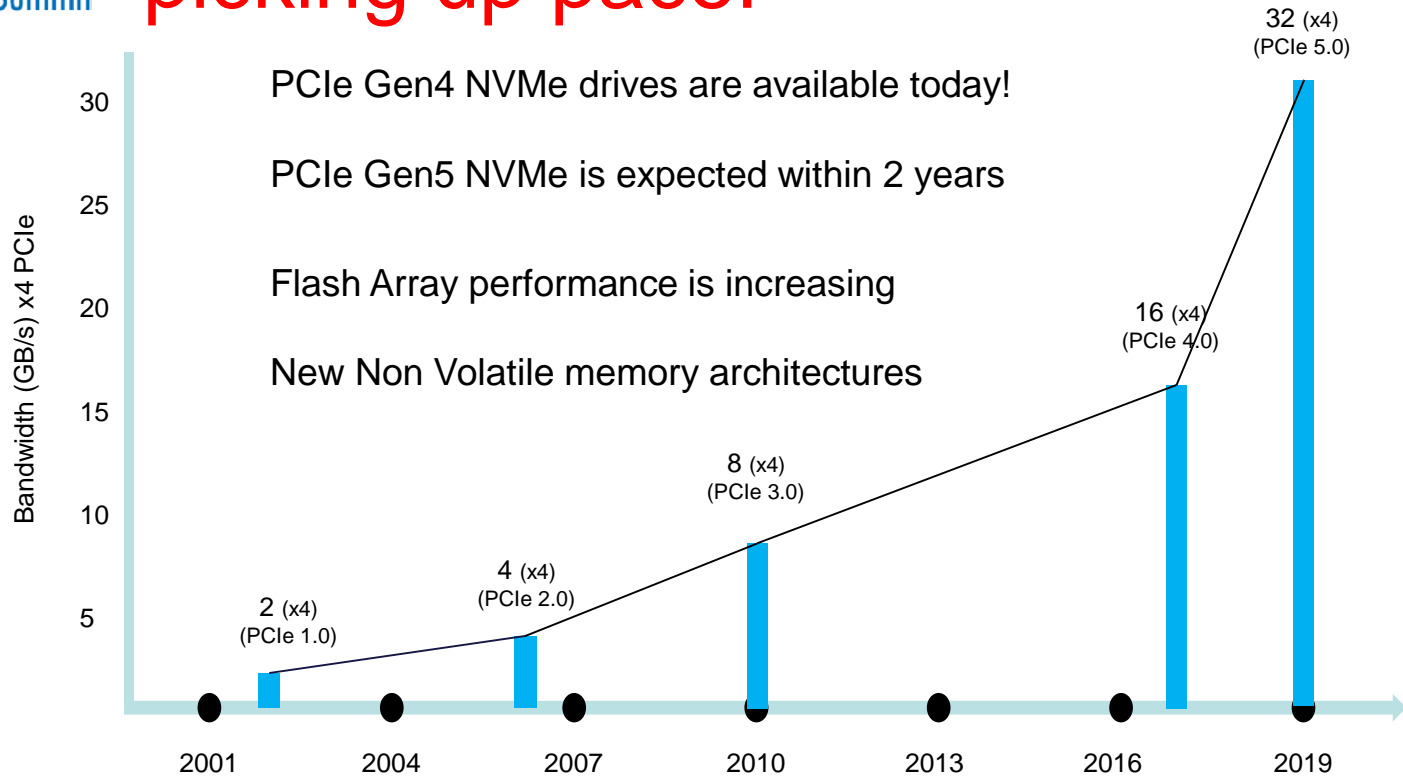Port Speed

Scale

Cost

Operations

SLAs

Monitoring

NVMe and PCIe bandwidth is picking up pace!

PCIe Gen4 NVMe drives are available today!

PCIe Gen5 NVMe is expected within 2 years

Flash Array performance is increasing

New Non Volatile memory architectures

Bandwidth (GB/s) x4 PCIe

- 2 (x4) (PCIe 1.0)
- 4 (x4) (PCIe 2.0)
- 8 (x4) (PCIe 3.0)
- 16 (x4) (PCIe 4.0)
- 32 (x4) (PCIe 5.0)

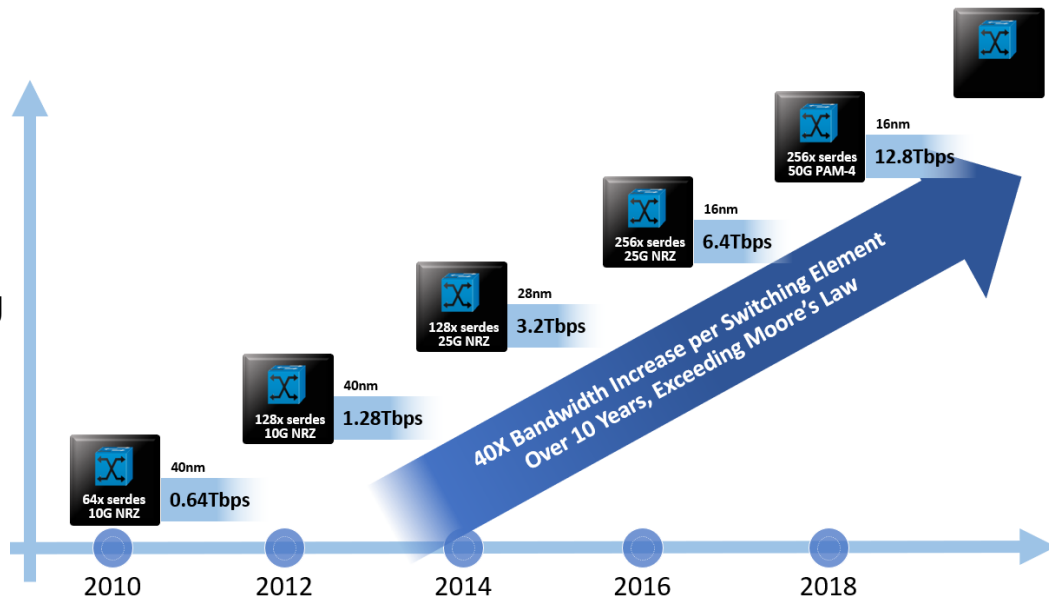|						
2001	2004	2007	2010	2013	2016	2019

Ethernet as Fabric for Storage

Ethernet continues to grow as an unified Fabric

- 40X increase over 10 years
  - Faster than Moore's Law!

- Large Ecosystem

- Economies of Scale
  - Cost per Terabit bandwidth is falling

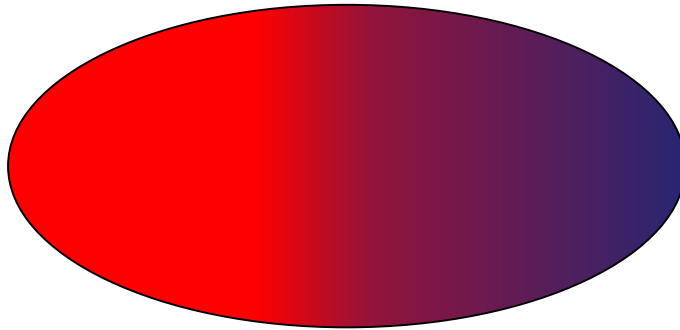- Management tools
  - TCP/IP

- Unification

64x serdes 10G NRZ — 40nm — 0.64Tbps
128x serdes 10G NRZ — 40nm — 1.28Tbps
128x serdes 25G NRZ — 28nm — 3.2Tbps
256x serdes 25G NRZ — 16nm — 6.4Tbps
256x serdes 50G PAM-4 — 16nm — 12.8Tbps

40X Bandwidth Increase per Switching Element Over 10 Years, Exceeding Moore's Law

2010   2012   2014   2016   2018

Fabric Solution Key Parameters
- Ethernet As Fabric

- ✓ Port Speed
- ✓ Scale
- ✓ Cost

? Operations

? SLAs

? Monitoring

Network Operations

SLA	Deploy	Monitor
- Uptime	- Plan	- SLA perf
- IOPs	- Provision	- Errors
- Latency	- Access	- Hotspots
- …		

Failures occur
 - Need tools to isolate the problem and resolve them

Is it a Server issue? App Issue? Network? Storage?

Storage Centric Analytics

Discovery · **IO Type (read/write)** · **Completion Times** · **IOPS** · **Latency Tracking**

- Where are my storage devices ?
- Can I track IOPs by Connection? Host? Storage Server?

- Do I have hotspots?
- Are SLAs met?

Event Driven Streaming Telemetry

Switch tracks
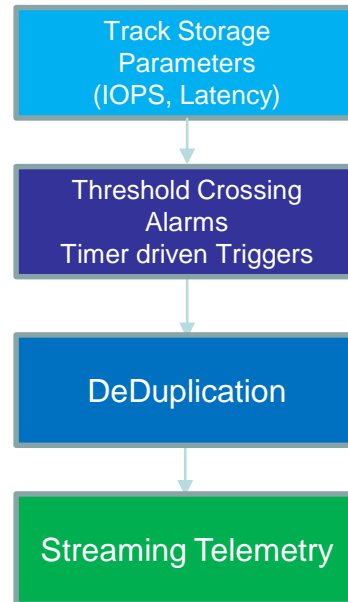- internal state (queues, buffers)
- Flows

Triggers
- Can set Threshold Crossing Alarms
  - Latency > N microseconds
- Timer driven periodic report

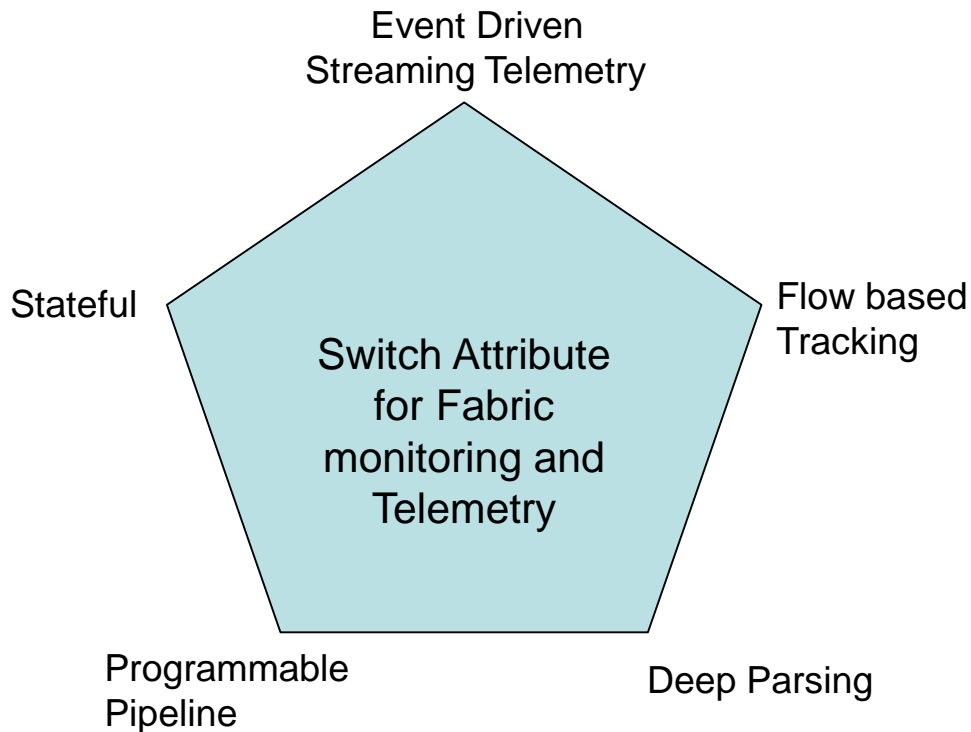Deduplication of reports
- Reduce load on collector

Steaming Telemetry
- Protobuf encoding

Track Storage Parameters (IOPS, Latency)

Threshold Crossing Alarms Timer driven Triggers
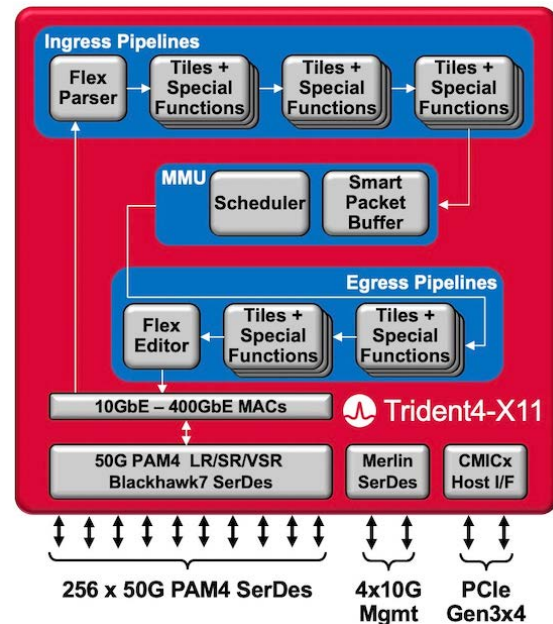
DeDuplication

Streaming Telemetry

Programmable Switches for NVMe-oF
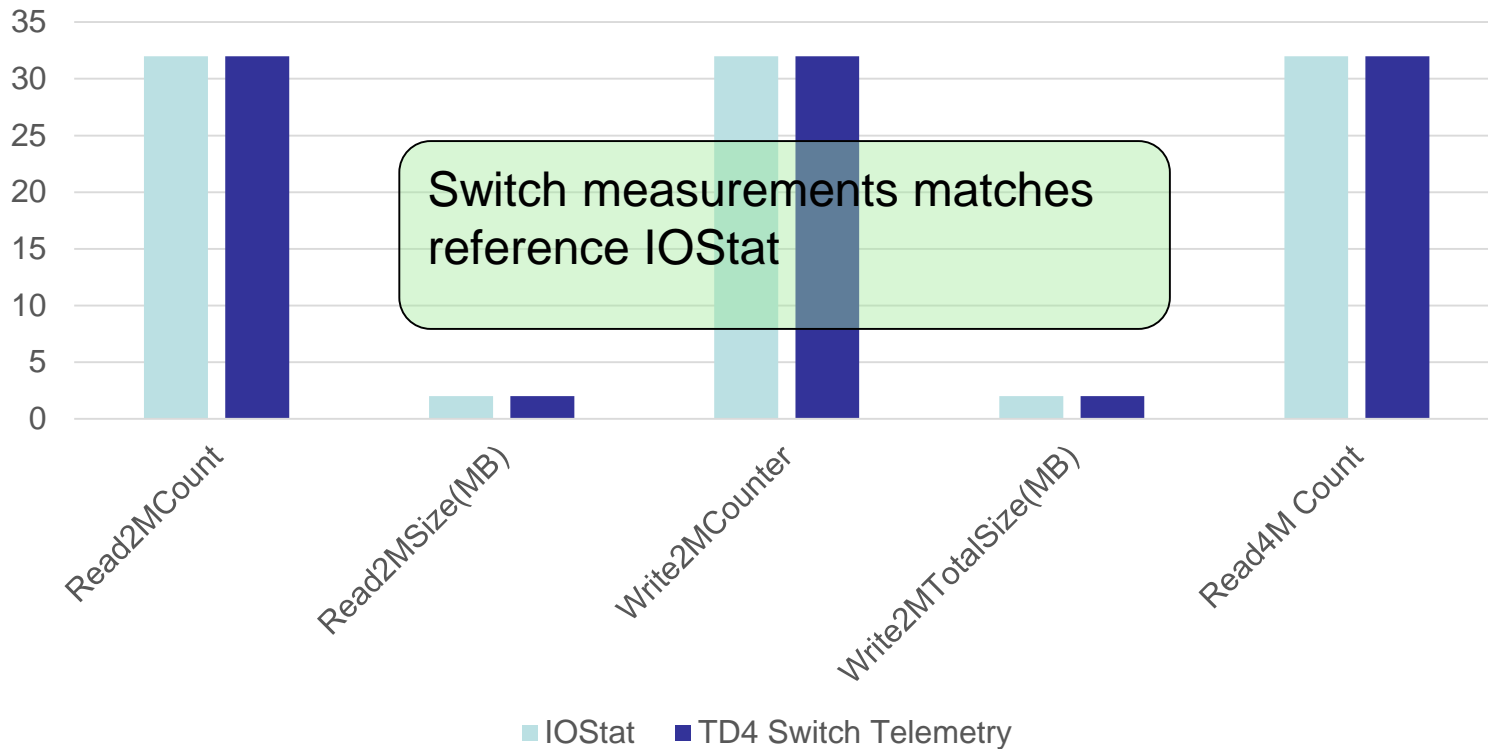
Trident 4 - Programmable Switch

- User programmable Instrumentation Capabilities

- 12.8Tbps capacity per instance

- Flexible Tile architecture
  - Compiler driven dataplane

- Stateful tracking
  - Tracking at full packet rate

- Streaming Telemetry

Storage Telemetry Demonstration

Test Results

IOStat(/proc/diskstats) vs Trident 4 Switch Telemetry

Switch measurements matches reference IOStat

Categories: Read2MCount, Read2MSize(MB), Write2MCounter, Write2MTotalSize(MB), Read4M Count

Legend: IOStat, TD4 Switch Telemetry

Summary

- Ethernet as Fabric presents a matching growth in connection speed and Switch Scale for Storage

- Trident 4 can Parse and Track NVMe-oF packets at high speeds

- Trident 4 stateful capabilities enable native storage metric reports

- Trident 4 offers native support for Rich Streaming Telemetry

Thank You!