The Interconnect Dilemma: Next-Gen Fabrics

Flexibility and Performance with GigaIO™ FabreX™

Niraj Mathur – VP of Product
Scott Taylor – Head of S/W
Market Driven Transformation

AI/DL/ML
Big Data Analytics
= Heterogenous Compute – GPU, FPGA, ASICs
Storage – NVMe, SCM

…. but how to efficiently:

• Share resources without impacting performance
• Build systems of scale
Rack Scale Interconnect Today

- Emergence of NVMe & NVMe-oF has enabled
  - Efficient protocol for SSDs
  - Disaggregation and sharing of higher cost NVMe SSDs
  - Capacity for server’s with PCIe enumeration limitations
  - Reasonable performance trending to ~10µs transport latency

- Still needs multiple transport protocols

- Need capability beyond storage
  - GPUs, FPGAs, NICs etc. still in servers
Future Performance Demands

- SSD media is getting faster
  - Driving to 1µs latencies

- Transport / Network needs to offer proportional gains
  - Currently 40% of total latency
  - Diminishing returns
  - Will get worse with new media

- Requires a Next-Gen network!

Source: NVMe Developer Days 2018: Ultra Low Latency NVMe-oF Controller Design, Samsung
Extending PCIe/NVMe Transport

- No protocol conversions
- Extreme performance: PCIe latencies and bandwidth across rack / cluster
- Disaggregation and S/W defined composition without sacrificing performance
- Native support for storage, compute and networking
- Rich Roadmap: Gen4 now, Gen 5 standard ratified, Gen 6 in progress, CXL adds cache coherency
- ... but need is beyond simple PCIe switching
  
  Server-to-server communication
  Native NVMe-oF, GDR
GigaIO FabreX: Next-Gen PCIe/NVMe Network

- P2P Networking via non-transparent bridges
- Excellent Performance - latency AND bandwidth
- All types of storage and compute nodes supported
- Scales without limits
- Support for NVMe-oF and GDR
- Open-program platform via Redfish APIs

FabreX OS & FabreX Non-Blocking Switch
FabreX Network Adaptor
Open-source server S/W
Cables
FabreX: Scale Up and Scale Out

Application Servers

- PCIe
- NVMe-oF

GDR

GPU Direct

PCIe

NVMe

Storage Server NVMe-oF Target

JBOG

JBOX

JBOF / Computational Storage

NVMe-oF

SCM

Memory Semantics

GPU Server
Easy Integration in S/W Workflows
Legacy NVMe-oF Implementation

- NVMe Host Resident Memory
  - NVMe-oF SQ
  - NVMe-oF CQ
  - Data Buffers

- NVMe Target Resident Memory
  - Data Buffers
  - NVMe-oF CQ
  - NVMe-oF SQ

- RDMA / TCP / FC Transport

SSD
FabreX NVMe-oF Implementation

NVMe Host Resident Memory
- NVMe-oF SQ
- NVMe-oF CQ
- Data Buffers

Block Layer

NVMe Target Resident Memory
- NVMe-oF CQ
- NVMe-oF SQ
- Data Buffers

RDMA/TCP/FC Transport

SSD

Flash Memory Summit 2019
Santa Clara, CA
Simultaneous operation of:
- DAS SSDs
- Composed NVMe SSDs
- NVMe-oF SSDs
- ML Suite on disaggregated FPGAs

- PCIe/NVMe transport at native PCIe bandwidth and latencies

- Live Demo at FMS19 (Booth 1045)
FabreX Latency Performance

- FabreX delivers 75% latency reduction over legacy transports today
  - ~2.5µs for reads and writes
- Future releases will optimize latency further to <1.5µs
IOPS and Bandwidth Performance

DAS = 5GB/s, 160K IOPS
Composed = 5GB/s, 160K IOPS
NVMe-oF = 4.8GB/s, 150K IOPS
Key Takeaways

- Higher Performance & Disaggregation are required for emerging data center workloads and lower TCO

- Next-Gen storage and accelerators will require a Next-Gen network to enable performance

- GigaIO’s FabreX extends PCIe/NVMe to the network and delivers significant bandwidth and latency benefits TODAY