Who Says Data Center Storage Has to be Inefficient?

Larry Chisvin
PLX Technology, Inc.
Agenda

- Who is PLX (and why does it matter)?
- Current usage of PCIe in storage
- Improving on data center fabrics
  - How a PCIe fabric works
  - Why it is better
- Usage models for PCIe-based storage
- New data center architectures
- Development Tools
- Call to action
PLX at a Glance

Public Company
NASDAQ: PLXT, $100M+ Rev., ~58% GM, Profitable

Global Enterprise
~155 Employees, Headquartered in USA
Design Center in India

The Market Leader
Over 70% Market Share and Growing

Market-Leading Customers
Widespread use by 1000’s of key accounts

Fabless Model
TSMC Partner with Designs and Production at 40nm
Customers Serving Cloud/Data Center

Servers
First Market that Developed for Switches

All Market Leaders Using PLX PCIe Products

Storage
Largest Market Growth & Attach Rate

Switches & Routers
Control Plane Standard
I/O Fan-out, Redundancy & Backplane

Large Switches with PLX added features used throughout Enterprise Storage Systems. Represents most of the storage revenue today in the pie chart below.

- Total Storage Doubling Every Two Years
- Standard Bus for Market Leaders
- Storage Systems & HBAs/CNAs

Storage
Largest Market Growth & Attach Rate
Basic Fanout of Storage Systems

Rest of System

PCIe Switch

Storage Subsystem

Storage Subsystem

Storage Subsystem

SAS/SATA Controller

Storage

PCle Switch

CPU
SSD-Based Aggregation

Rest of System

PCIe Switch

PCIe SSD Controller

Flash Memory

PCIe SSD Controller

Flash Memory

PCIe SSD Controller

Flash Memory

PCIe SSD Controller

Flash Memory
Making the Data Center Better
The Rack in the Cloud
Current Solutions

- Multiple fabrics
- “Bridging” Devices on each module
- More
  - Parts
  - Power
  - Cost
  - Latency
  - Complexity
Problem…or Opportunity?

Since one major problem here is the need for all of those “bridging” devices…

Then what is the answer?
Is There Something Common?
PCI Express Fabric

- Single Converged Fabric
- Simple, Inexpensive Redriver
- High performance, Low cost, Low latency, Low power Interconnect
PCIe-Based Fabric Solution

Standard capabilities of fabric provided by PCIe

Host-Host
Using standard DMA & RDMA Infrastructure

Shared I/O
Using existing Hardware and Drivers

Management CPU
To configure system and manage events
Mixing Data Types on PCIe

Enterprise SSDs are rapidly standardizing on PCIe

Hybrid systems can be easily implemented
PCI Express Initiatives

- SR-IOV
  - Allows sharing of end points

- SRIS (Separate Refclk Independent SSC)
  - Allows each side of link to have its own clock
  - Enables use of general purpose cables

- New PCIe Cable spec (OcuLink)
  - Targeted for low cost copper and optical

- Signal integrity
  - Better defined repeater & retimer functions
DPC/eDPC

- Downstream Port Containment (DPC)
  - Enables the prevention of host time-out or “blue screen”
  - When triggered by error from an endpoint
    - Take down link to endpoint
    - Send error message/interrupt to the host
    - Reply to read request within time-out window
- Completion synthesis for root ports
- Host software to handle switch responses
- Important for storage systems
Quantifying the Advantages

- Replacing existing Ethernet-based racks with PCI Express can offer tangible savings
  - PCI Express inside the rack
  - Ethernet to rest of data center
- Top-of-rack switch similar cost to PCIe
  - Greater reduction if converged data
- Remove NIC/LOM
  - Each slot costs $200-$300
  - Each slot dissipates 5-7w
  - Rack has 24-64 slots
- PCIe-base solution uses simple redriver
  - Each redriver can be $2-$5
  - Power is less than 1w
Cost and Power Advantages of PCIe

**Cost/Gbps ($) 10GE**

- ExpressFabric: $1.00
- 10GE SFP+: $1.40
- 10GE LOM: $1.20

**Power/Gbps (w) 10GE**

- ExpressFabric: 0.00 w
- 10GE SFP+: 0.80 w
- 10GE LOM: 1.40 w
Storage Usage Models for a PCIe Fabric
SSD Adapter Card

- Two SSD Controllers
  - NVMe Form-Factor – 2.5”
  - Only ASIC1 (controller) exposed to Host
  - ASIC1 servicing interrupts from ASIC2

- Two or more SSD controllers aggregated with a switch
  - Adapter Card Form-Factor
  - Host isolation capability
  - One ASIC exposed to Host
    - Services interrupts from other ASICs
Server Motherboards

- Capacity expansion in server enclosures
  - Aggregation through PCIe switch
Hybrid Storage System

- Mixing storage mediums for cost/performance
  - Move data between these mediums
Storage Expansion Boxes

- Creating storage expansion box with PCIe
  - Connect to server through PCIe Cu/Op cable
  - No protocol translation overhead

![Diagram of storage expansion box with PCIe connection](image-url)
Flash Appliance

- Share an array of SSD modules in a chassis
  - Associate a segment of SSDs to specific server
Server Redundancy

➢ Build Server Redundancy
SR-IOV Sharing of SSDs

- Share SR-IOV SSD modules in expansion chassis
  - Shared by multiple servers/hosts or blade servers
Can Form Factors Help?
From Rack to Blade

- Rack servers are self-contained systems
  - Traditional architecture in stackable form
  - CPU, memory, I/O, power, cooling, etc.
  - Standard form factors

- Blade servers consolidate common subsystems
  - Common power, cooling, external connection
  - Blades offer processing, storage, I/O
  - Systems are vendor-specific

- Improvement on rack
  - Allows some savings
  - But non-standard subsystems
  - And doesn’t really address issue completely
Next Step - MicroServers

- Traditional (brute force) approach to computing limits compute density due to power
- Open source OS's and application software allow portability of CPUs and platforms
- Large scale applications can be effective with smaller compute engines
  - Web servers
  - Content delivery
  - Analytics
- Bottleneck in such applications are memory and I/O bandwidth rather than raw compute power
What is a MicroServer?

- Array of low power, low cost processors
  - Less than 45w
  - Moving to less than 10w
- Each processor has modest performance
  - Aggregate performance can be powerful
- Can execute applications where processing can be easily distributed
- CPU, memory, & I/O on compact module
- ARM-based systems are most common
  - X86 platforms are expected to be popular
Problems with Current Solutions

- Cannot adequately handle compute-intensive applications easily
  - Many applications cannot distribute processing
- Cannot easily trade off processing, storage & I/O, since the subsystems are self-contained
- Most systems are proprietary
  - Need to buy into entire system
- Similar to blade servers vs rack servers
  - Definite advantages, but…
  - Limited ability to mix best-in-class subsystems
PCIe-Based MicroServer

µServer

Network of SOCs

PCIe Aggregation

PCI Express Fabric

Network of SOCs

µServer

PCIe Aggregation
Powerful Storage Appliance

Network of SOCs

μServer

PCIe Aggregation

PCI Express Fabric

Flash

SSD Controller

μServer

PCIe Aggregation

Flash

SSD Controller

Network of SOCs

PCIe Aggregation

Flash

SSD Controller
Consolidate Complete System

Network of SOCs

µServer

PCIe Aggregation

PCI Express Fabric

Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Flash
Fla
Building a Complete Fabric

- Development platforms for customer products
- ExpressNIC
  - PCIe bus extender
  - Based on PEX 8725
  - 2 QSFP+ ports, x4 PCIe each
- Argo
  - Rack-mounted PCIe switch fabric
  - 1U form
  - 32 QSFP+ ports, x4 PCIe
Better Storage Through PCIe

- PCIe-based solutions offer substantial benefits to storage systems
  - Direct connection to vast number of devices
  - Low latency, high bandwidth solution
  - Flexible sharing with existing infrastructure
  - Expand as specification improves

- Especially attractive to SSD-based systems
Thank You!
For More Info, please visit

www.plxttech.com