Reconfigurable Accelerators for Big Data and Cloud
RAW 2016

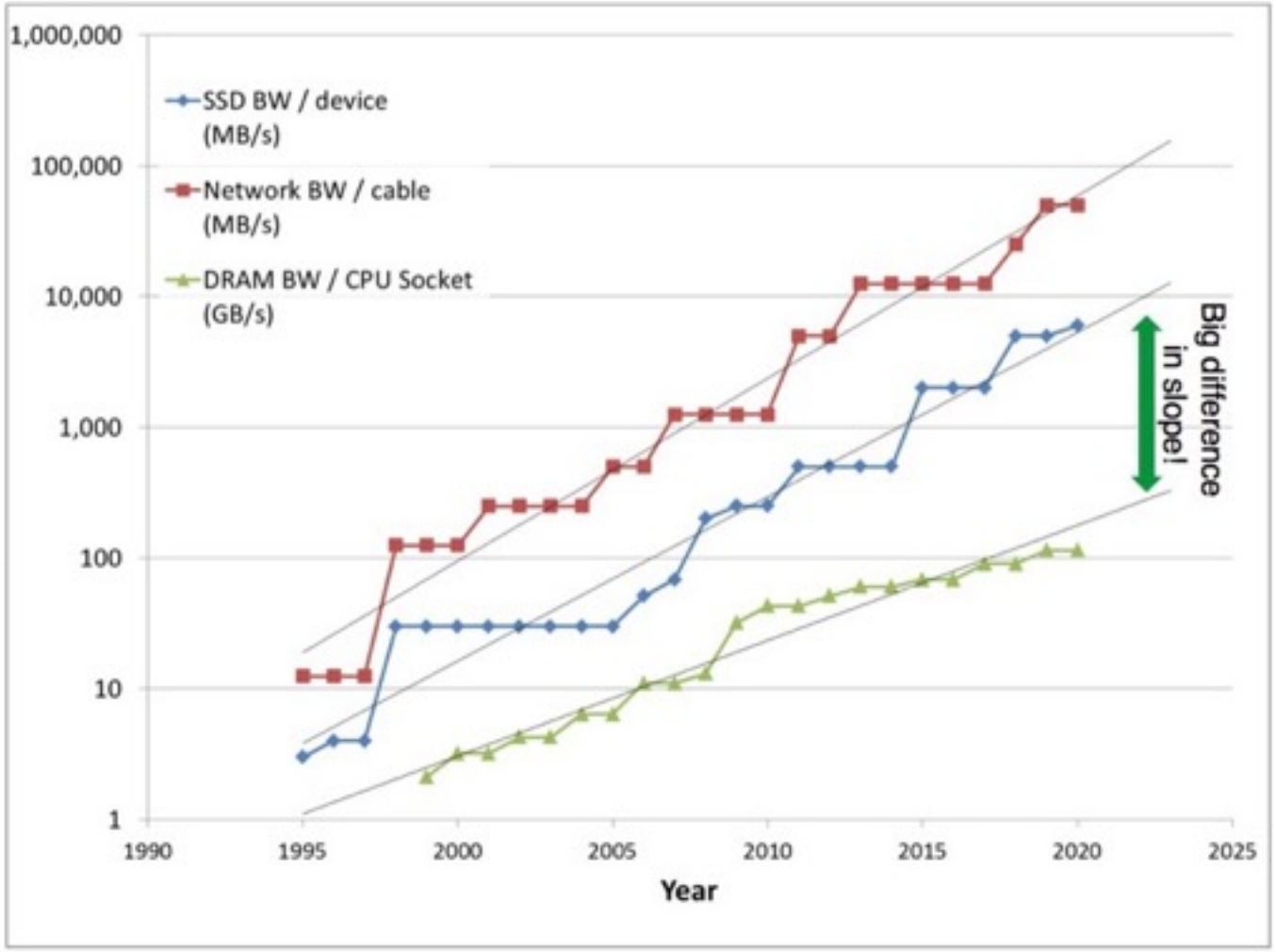
(23rd Reconfigurable Architectures Workshop @ IPDPS)

H. Peter Hofstee, Ph.D.
Distinguished Research Staff Member, IBM Austin Research
Professor, Big Data Systems, TU Delft
May 23, 2016

Network, Storage, & DRAM trends

Log scale

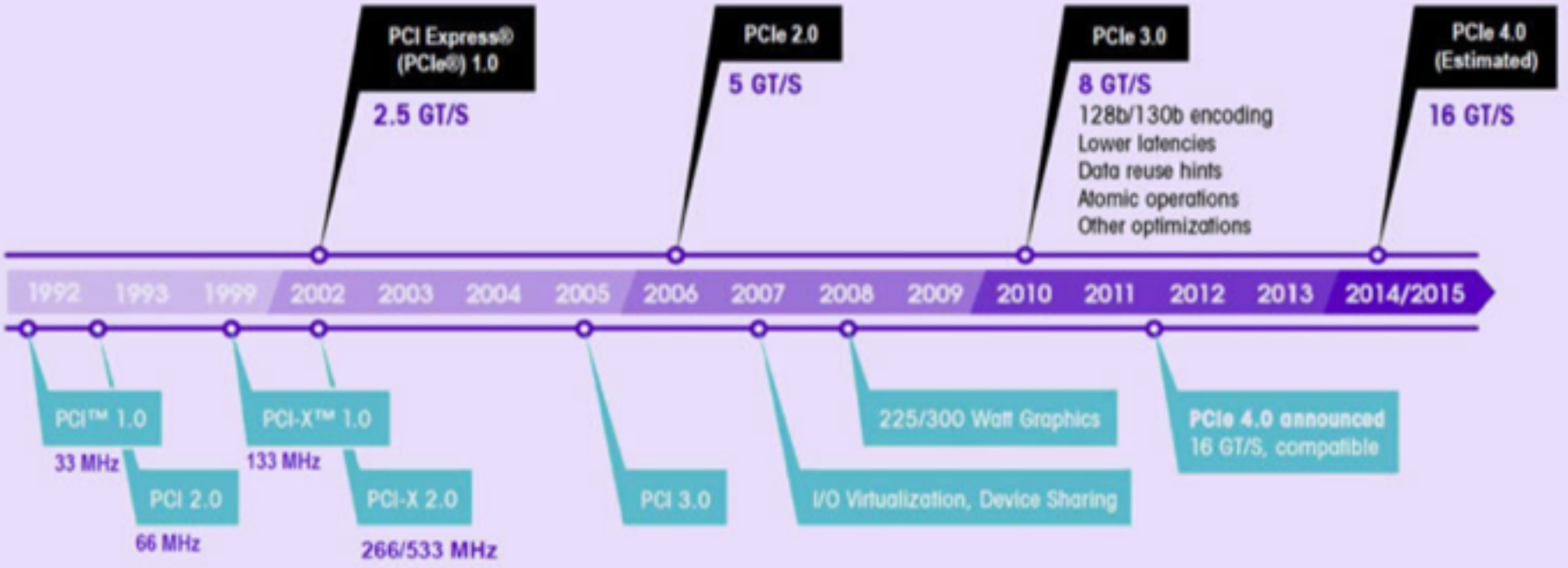
- Use DRAM Bandwidth as a proxy for CPU throughput
- Reasonable approximation for DMA and poor cache performance workloads (e.g. Storage)



Source: Sandisk IT Blog

PCI & PCIe Technology Timeline

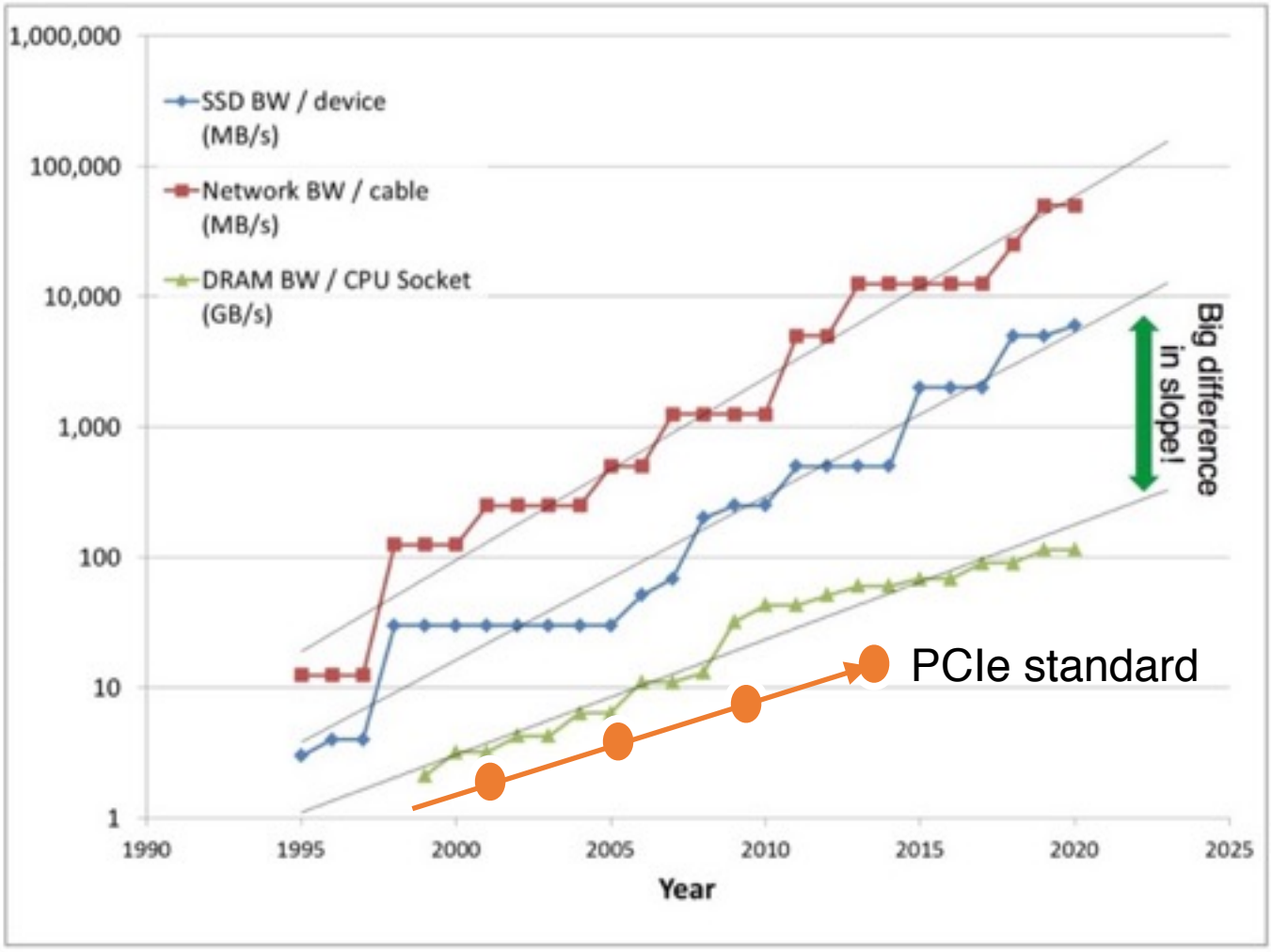
- PCI-SIG is the leader in I/O standard development and continues to evolve to meet existing and emerging usage models across the computing industry



Network, Storage, & DRAM trends

Log scale

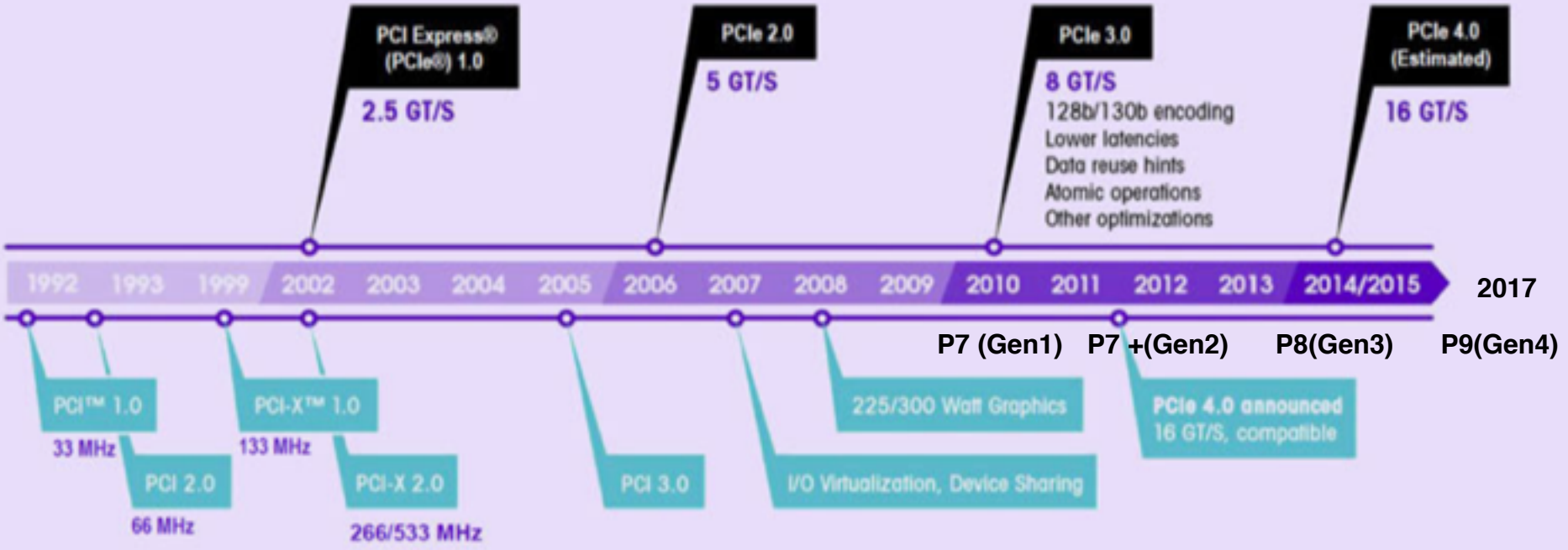
- Use DRAM Bandwidth as a proxy for CPU throughput
- Reasonable approximation for DMA and poor cache performance workloads (e.g. Storage)



Source: Sandisk IT Blog

PCI & PCIe Technology Timeline

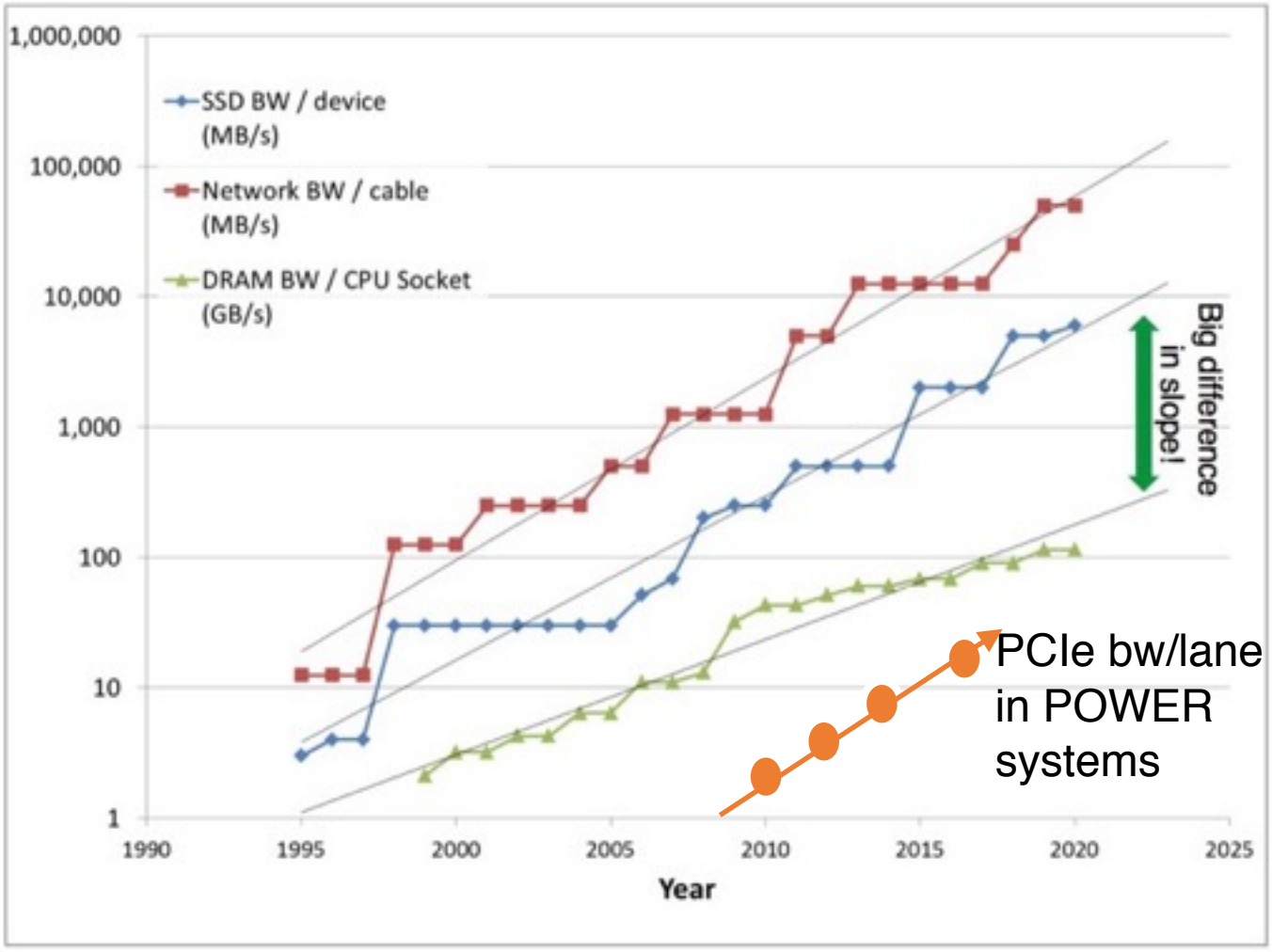
- PCI-SIG is the leader in I/O standard development and continues to evolve to meet existing and emerging usage models across the computing industry



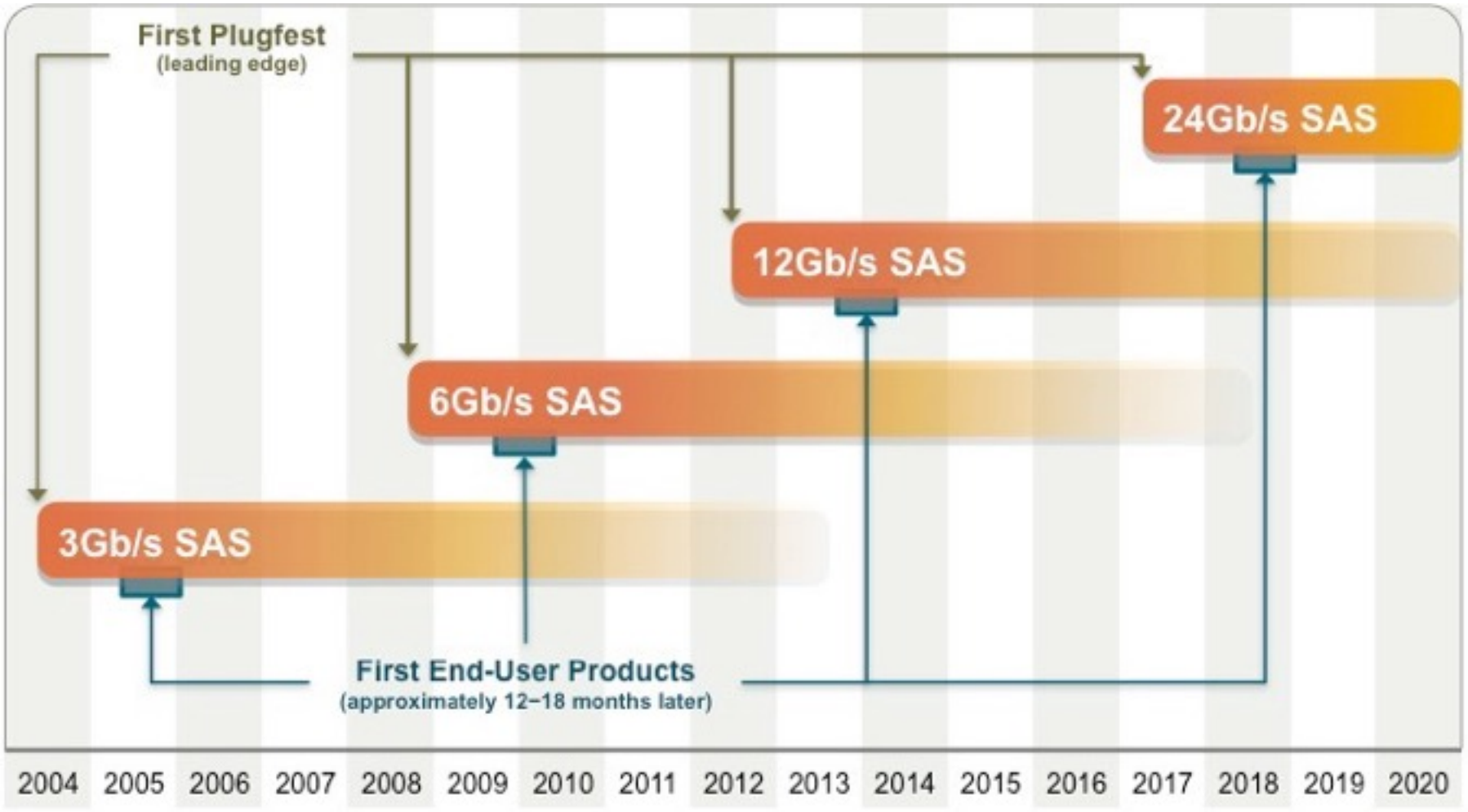
Network, Storage, & DRAM trends

Log scale

- Use DRAM Bandwidth as a proxy for CPU throughput
- Reasonable approximation for DMA and poor cache performance workloads (e.g. Storage)



Source: Sandisk IT Blog



SCSI Trade Association – Aug 2015

Summary Performance Data – HDD, SSHD, SSD

Class	Type	FOB IOPS	IOPS (higher is better)			Throughput (larger is better)		Response Time (faster is better)	
Storage Device	Form Factor, Capacity, Cache	RND 4KiB 100% W	RND 4KiB 100% W	RND 4KiB 65:35 RW	RND 4KiB 100% R	SEQ 1024KiB 100% W	SEQ 1024KiB 100% R	RND 4KiB 100% W AVE	RND 4KiB 100% W MAX
HDD & SSHD									
7,200 RPM SATA Hybrid R30-4	2.5" SATA 500 GB WCD	125	147	150	135	97 MB/s	99 MB	15.55 msec	44.84 msec
15,000 RPM SAS HDD IN-1117	2.5" SAS 80 GB WCD	350	340	398	401	84 MB/s	90 MB/s	5.39 msec	97.28 msec
Client SSDs									
mSATA SSD R32-336	mSATA 32 GB WCD	18,000	838	1,318	52,793	79 MB/s	529 MB/s	1.39 msec	75.57 msec
SATA3 SSD IN8-1025	SATA3 256GB WCD	56,986	3,147	3,779	29,876	240 MB/s	400 MB/s	0.51 msec	1,218.45 msec
SATA3 SSD R30-5148	SATA3 256GB WCE	60,090	60,302	41,045	40,686	249 MB/s	386 MB/s	0.35 msec	17.83 msec
Enterprise SSDs									
Enterprise SAS SSD R1-2288	SAS 400GB WCD	61,929	24,848	29,863	53,942	393 MB/s	496 MB/s	0.05 msec	19.60 msec
Server PCIe SSD IN1-1727	PCIe 320GB WCD	133,560	73,008	53,797	54,327	663 MB/s	772 MB/s	0.05 msec	12.60 msec
Server PCIe SSD IN24-1349	PCIe 700GB WCD	417,469	202,929	411,39	684,284	1,343 MB/s	2,053 MB/s	0.0 msec	0.58 msec

Source: Calypso Testers, 2013

Today

- PCIe NVMe commoditized
  - Better BW/$ than HDD !
- Typical enterprise NVMe
  - 3.2 GB/s seq. read
  - 3.2 TB
  - 800K IOP (4K)
  - 25W
  - 50-100usec
- 2U with 24 SFF NVMe
  - 76.8 GB/s (96 lanes of PCIe Gen 3)
    - all available lanes in a dual-socket POWER8
  - 76.8 TB
  - 19.2 Million IOPs
  - 600W(max) / 24 NVMe

SuperMicro

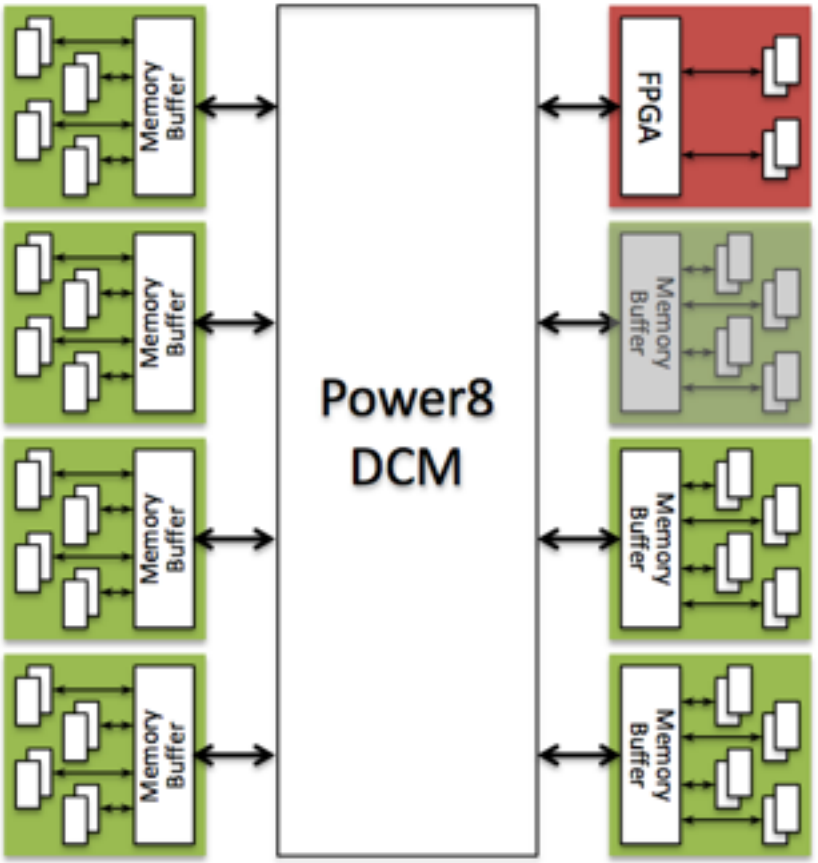
Memory-class bandwidth even before storage-class memory!

Three Ways to Attach

- As memory
  - Real address space
  - Any accelerator is highest privilege level only
  - Not attractive if latencies are high
- As I/O
  - I/O space
  - Offload (including DMA) to pinned memory only
  - Typically involves extra copy operation and coordination w.CPU
- On the SMP bus
  - Full access to all system resources
  - Best support for offload and near-memory compute
  - Most efficient synchronization mechanisms
  - Adapters can interact with other adapters w/o any CPU cycles

P8 Memory Sub-System with ConTutto

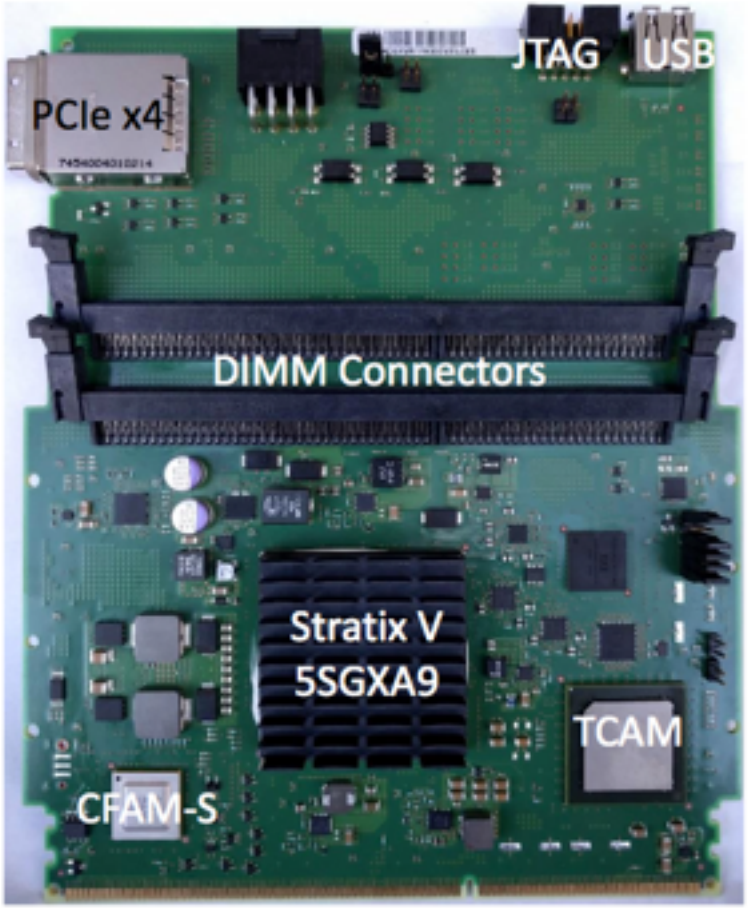
- Built an FPGA-based card that plugs into the DMI slot
- Enables regular system operation with any mix of CDIMMs and ConTutto cards populated
- Full compatibility with DMI protocol
- Memory controllers implemented in fabric logic and independent of DMI protocol logic
- Flexible system architecture enables easy implementation of additional features



OpenPOWER™

ConTutto Card

- Intended to be an experimentation and proto-typing vehicle
- Card characteristics
  - 10 signal layers
  - 10 power/ground layers
- Plug compatible with CDIMM, but 2.5" higher-- and DIMMs add width
- Large Altera FPGA with capacity for additional function incorporated
- CFAM-S (connection to service processor) enables system integration

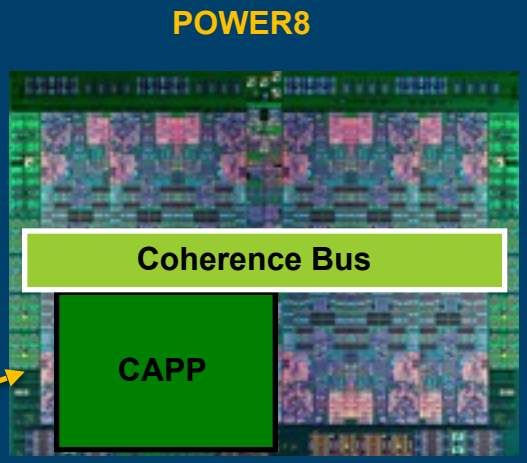


Attach on the SMP bus: CAPI

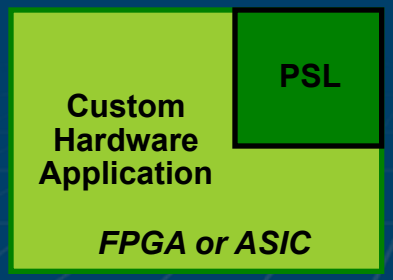
POWER8 CAPI Coherent Accelerator Processor Interface

Virtual Addressing
- Accelerator can work with same memory addresses that the processors use
- Pointers de-referenced same as the host application
- Removes OS & device driver overhead

Hardware Managed Cache Coherence
- Enables the accelerator to participate in "Locks" as a normal thread Lowers Latency over IO communication model



PCIe Gen 3
Transport for encapsulated messages



Processor Service Layer (PSL)
- Present robust, durable interfaces to applications
- Offload complexity / content from CAPP

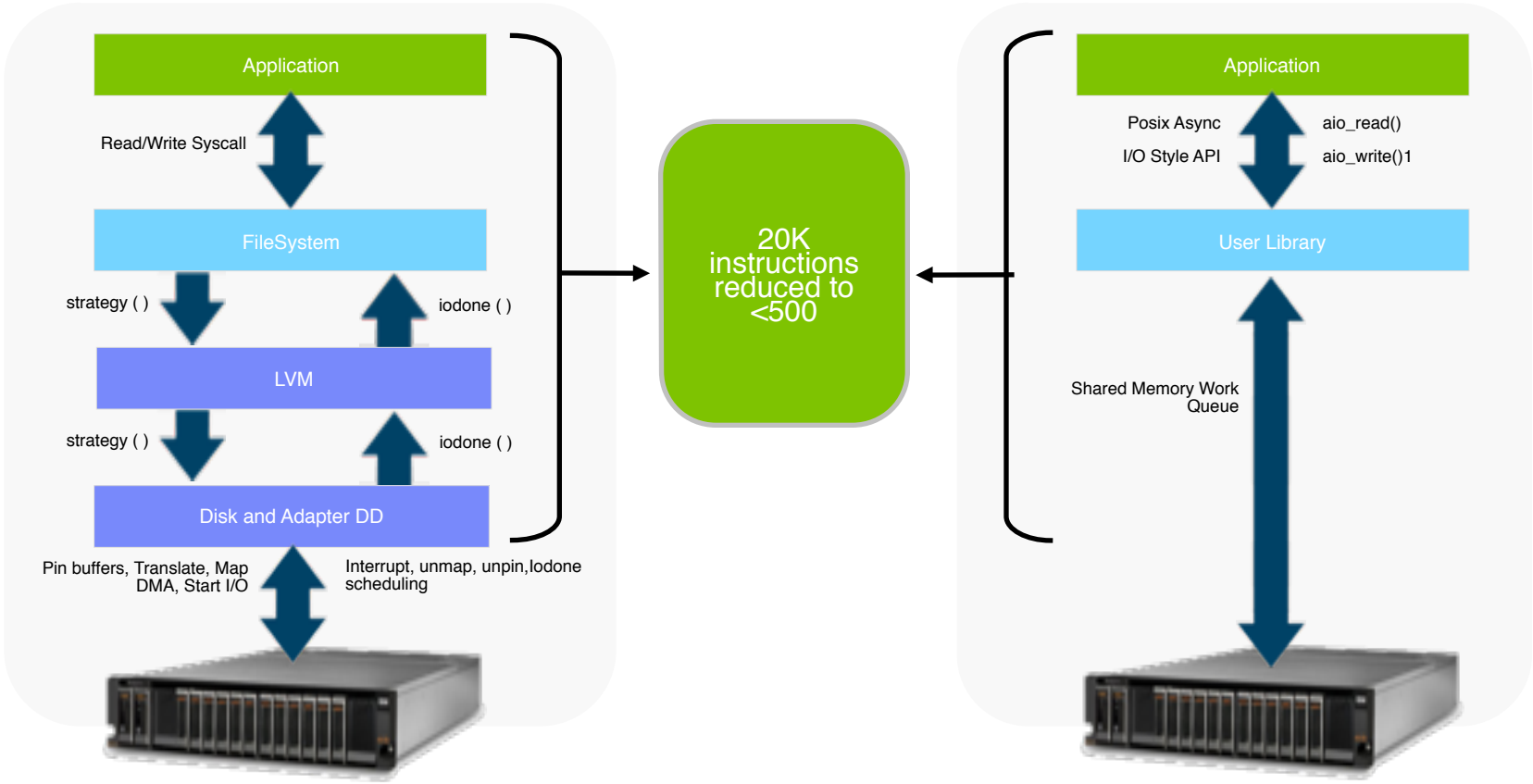
Customizable Hardware Application Accelerator
- Specific system SW, middleware, or user application
- Written to durable interface provided by PSL

Attach on the SMP bus: CAPI example

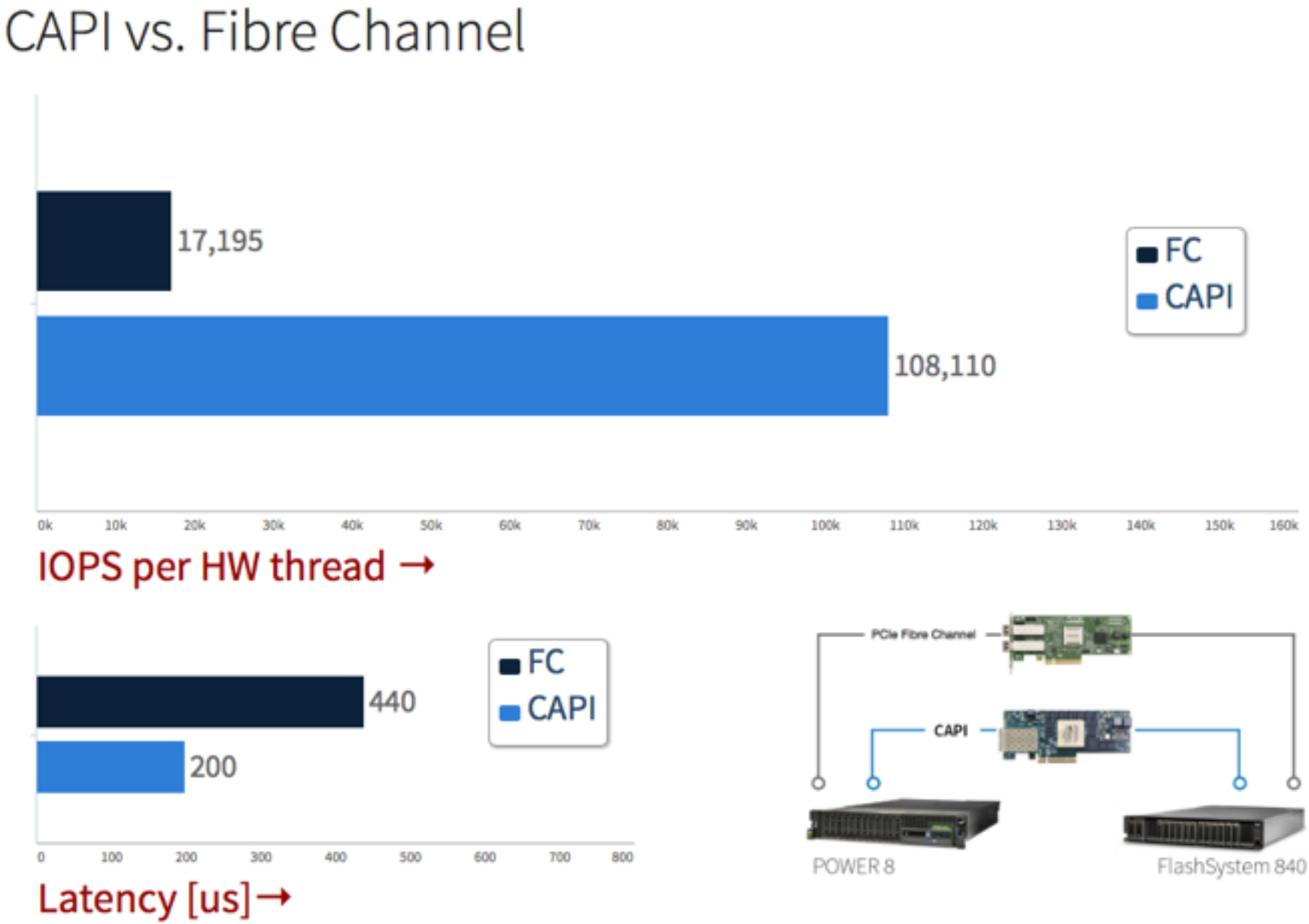
CAPI Attached Flash Optimization

IBM

- Attach TMS Flash to POWER8 via CAPI coherent Attach
- Issues Read/Write Commands from applications to eliminate 97% of code pathlength
- Saves 20-30 cores per 1M IOPs

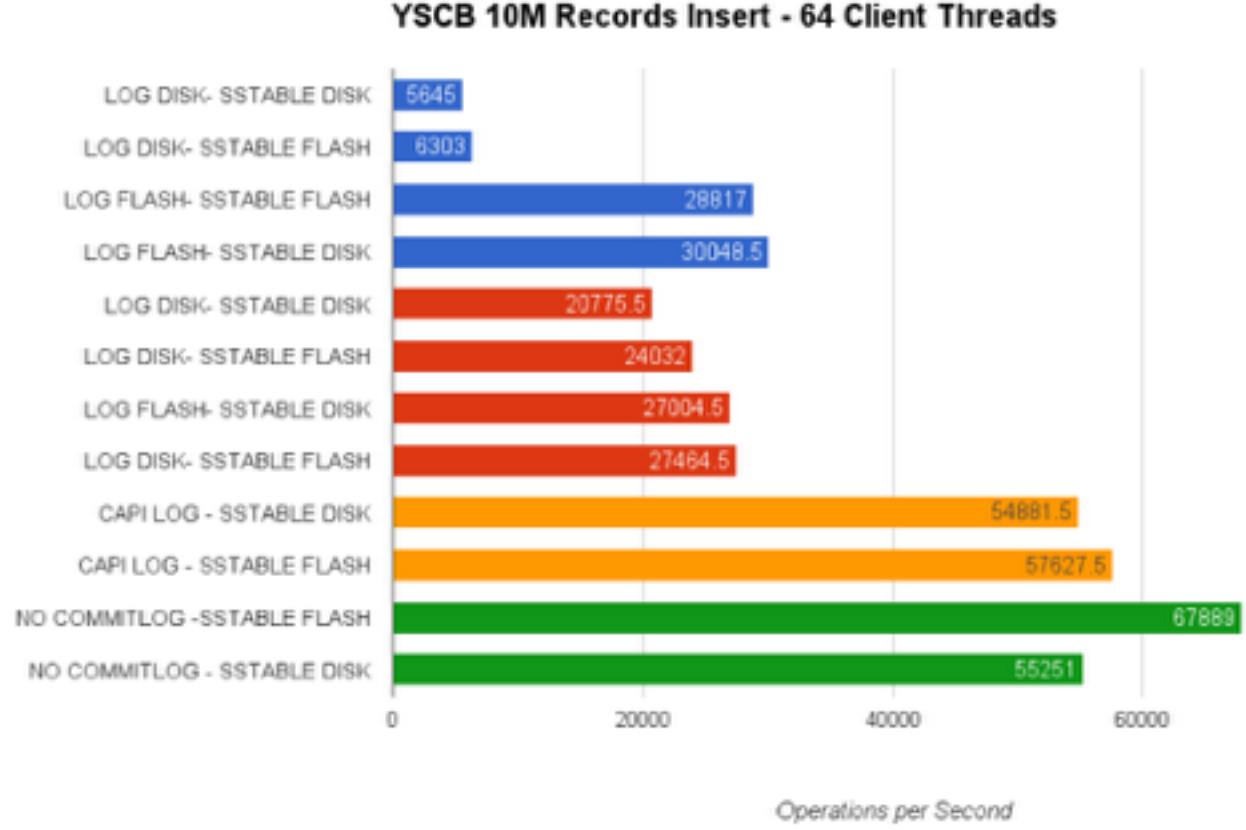


Attach on the SMP bus: CAPI example



Attach on the SMP bus: CAPI example

CASSANDRA WRITES
w. Bedri Sendir



- **Periodic Commitlog:** Syncs every **commitlog_sync_period_in_ms**. *Writes are not durable.* **commitlog_sync_period_in_ms** : 10000ms
- **Batch Commitlog:** Collects requests in batches **commitlog_sync_batch_window_in_ms** and sync's & acknowledge each add. *Durable writes.* **commitlog_sync_batch_window_in_ms** : 2ms
- **CAPI-Flash Commitlog:** Writes each mutation to the flash as soon as it is received on Cassandra server. It will guarantee that it writes on flash before acknowledging writes. *Fast* ***Durable writes.***
- **Commitlog Disabled:** Writes go directly to the in memory data structure.

Attach on the SMP bus: CAPI example

datanami

• BIG DATA • BIG ANALYTICS • BIG INSIGHTS •

Home About Whitepapers Events Subscribe

HOME FEATURES ▾ SECTORS ▾ APPLICAT

Detecting fraud hidden in huge data sets should be easier with Neo4j running on Power8.

HPC wire

ENTERPRISETECH

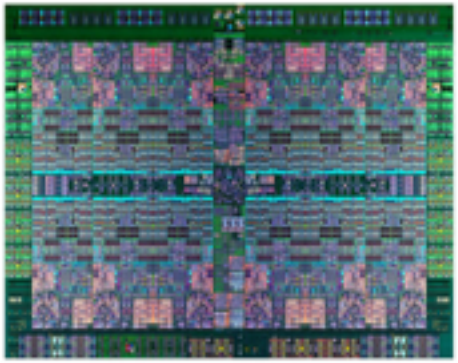
HPC wire JAPAN

Top News from Leading Solution Providers

October 21, 2015

Neo4j Touts 10x Performance Boost of Graphs on IBM Power FPGAs

Alex Woodie



Neo Technology today announced that it's working with IBM to support its graph database on IBM Power8-based servers equipped with field programmable gate arrays (FPGAs). This will enable customers to run graph databases with hundreds of billions to trillions of edges, thereby tackling a new class of intractable big data problems in bioinformatics, fraud detection, and IoT analytics. In other news, Neo also open sourced a key piece of software.

Attach on the SMP bus: CAPI example

Acceleration Use Case Example:

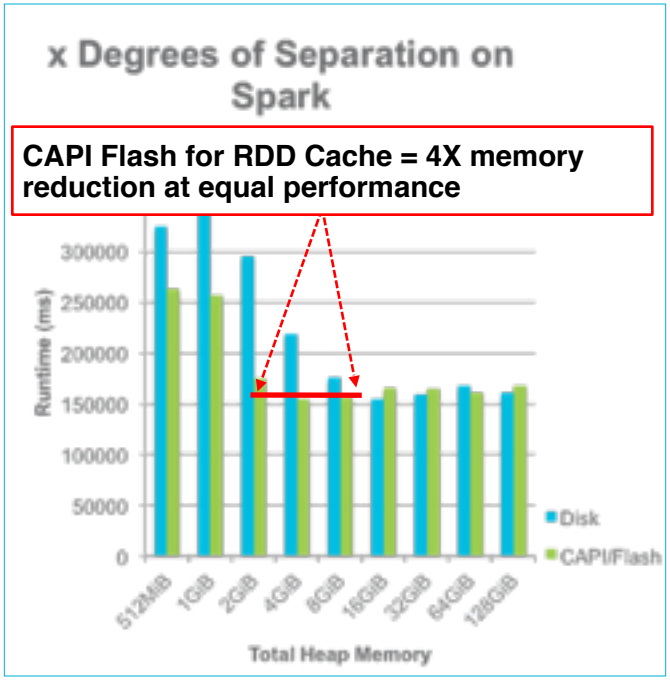
Power Systems

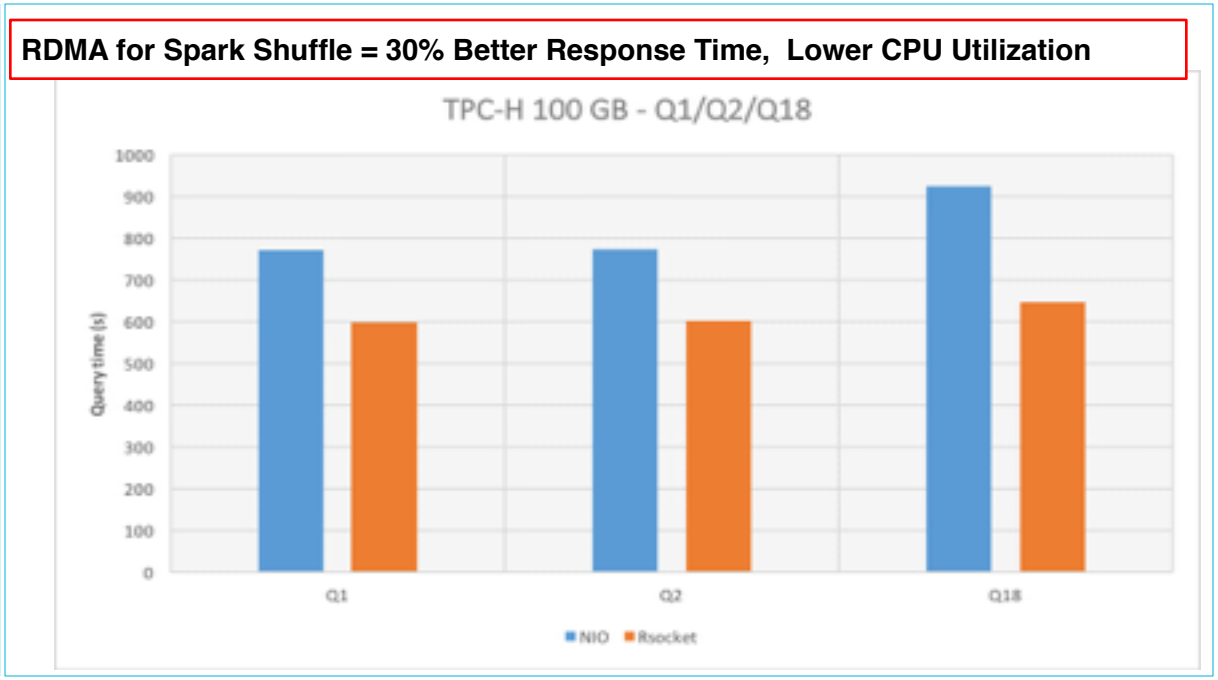
Spark

Fast and general engine for large-scale data processing

IBM

- CAPI Flash and RDMA Leveraged Transparently to Spark Applications
- Coming…. HDFS CAPI FPGA Erasure Code Acceleration, CAPI FPGA Compression Acceleration, ….





Attach on the SMP bus: CAPI example

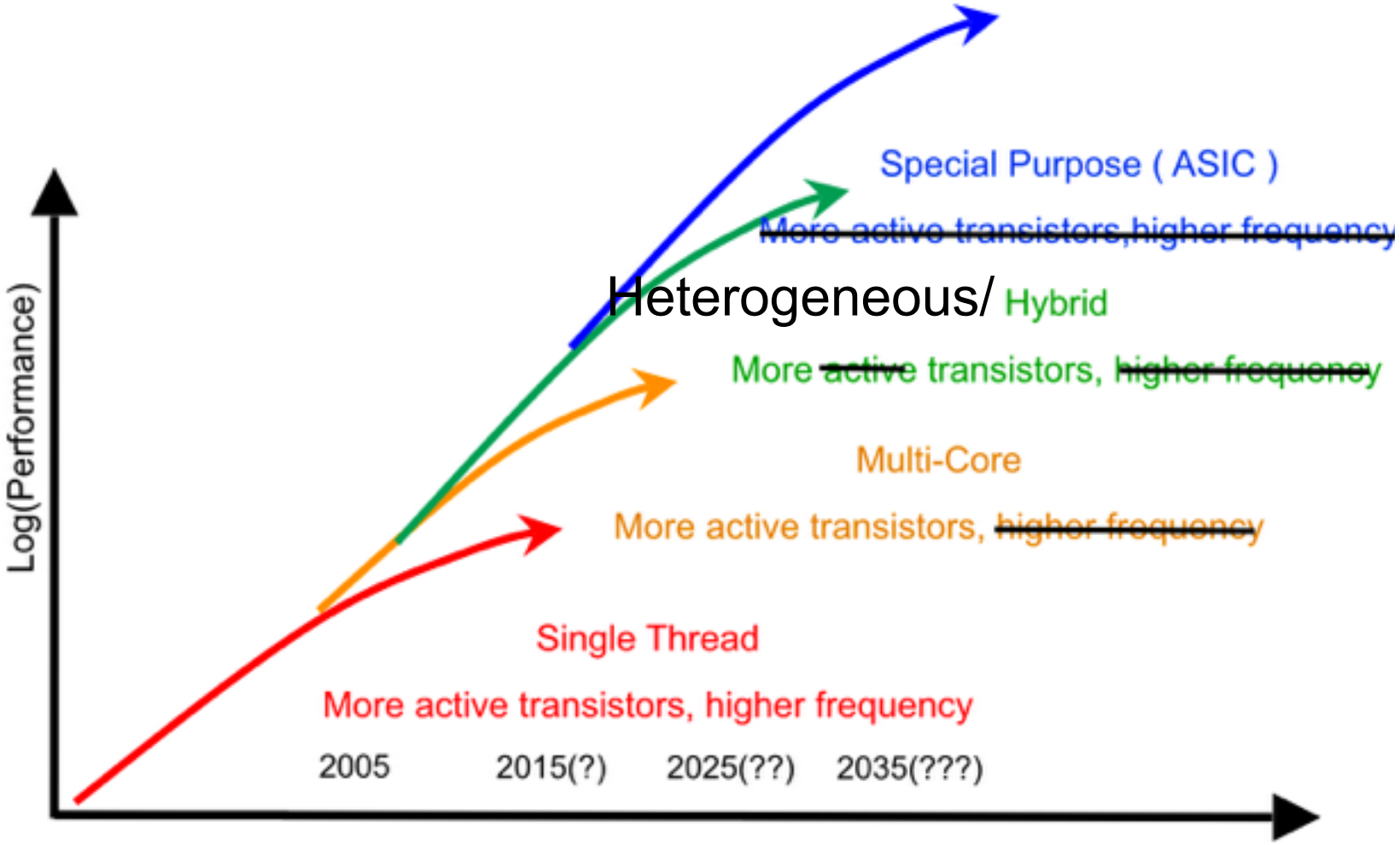
In-the-box version of CAPI-attached Flash
(Uses standard M.2 formfactor flash)

Nallatech CAPI Flash Accelerator based on a Xilinx Kintex UltraScale KU060 FPGA

Similar Story for Networking

- Fast transition from 1Gb/s to 10Gb/s to 100Gb/s
  - 25Gb, 50Gb really subsets of 100Gb
  - 40Gb looks less attractive (to me)
- Partly driven by in-memory paradigms like Spark
  - No longer just trying to match HDD speed
- Transition to optical (especially across racks)
  - Opens the door for much more bandwidth in future
- Coherent attach allows:
  - Most offload to network adapter
  - Network adapter to reach all resources (including storage) without touching the CPU

Usual Story for Compute



Reminder: Multi-Core Drives Memory Pressure

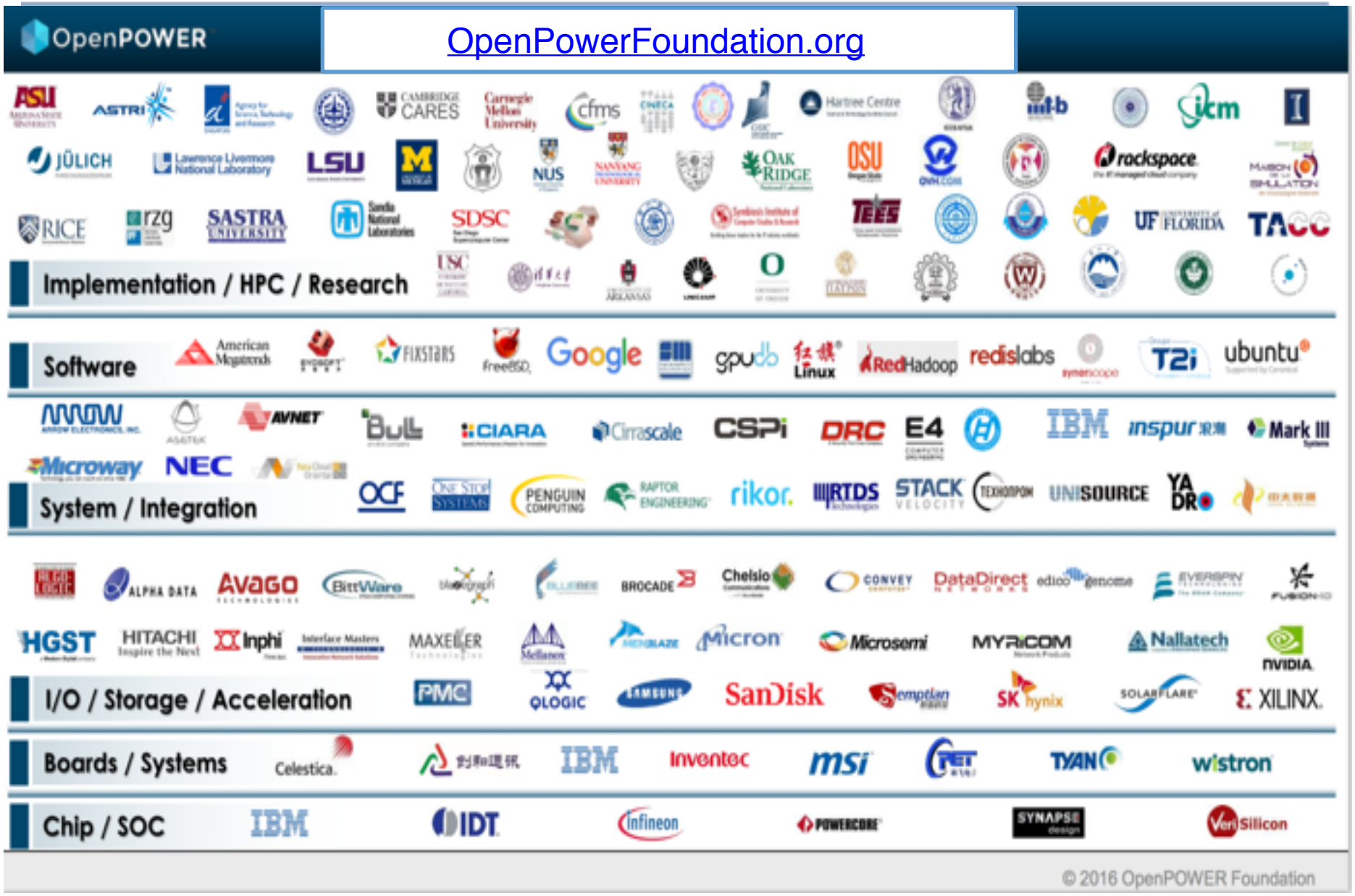
Conclusions so far …

A LOT of pressure on bandwidth

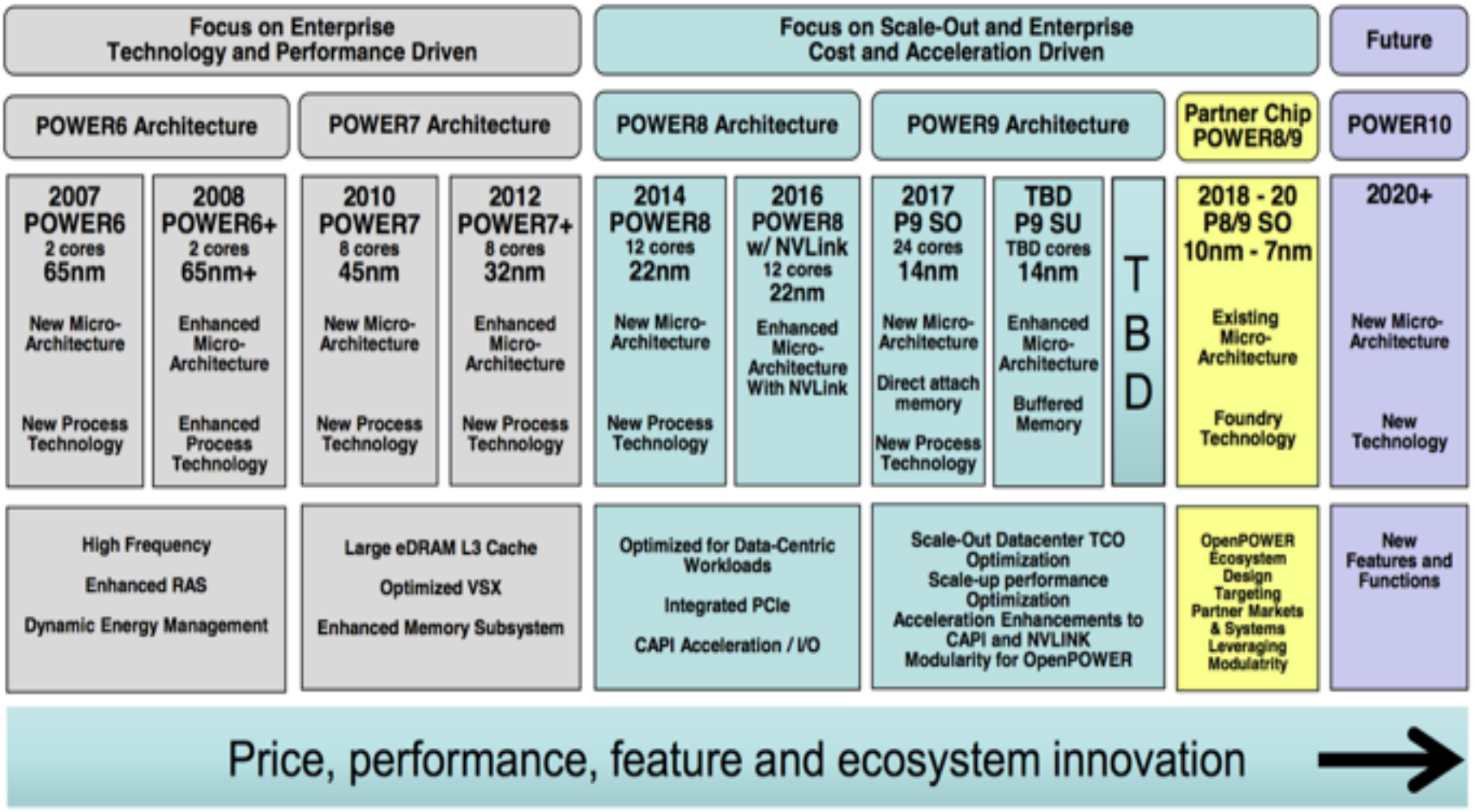
Strong desire to be SMP-attached

So what are we doing about it?

1: This is a big challenge: don't go it alone …



2: Publish a roadmap …

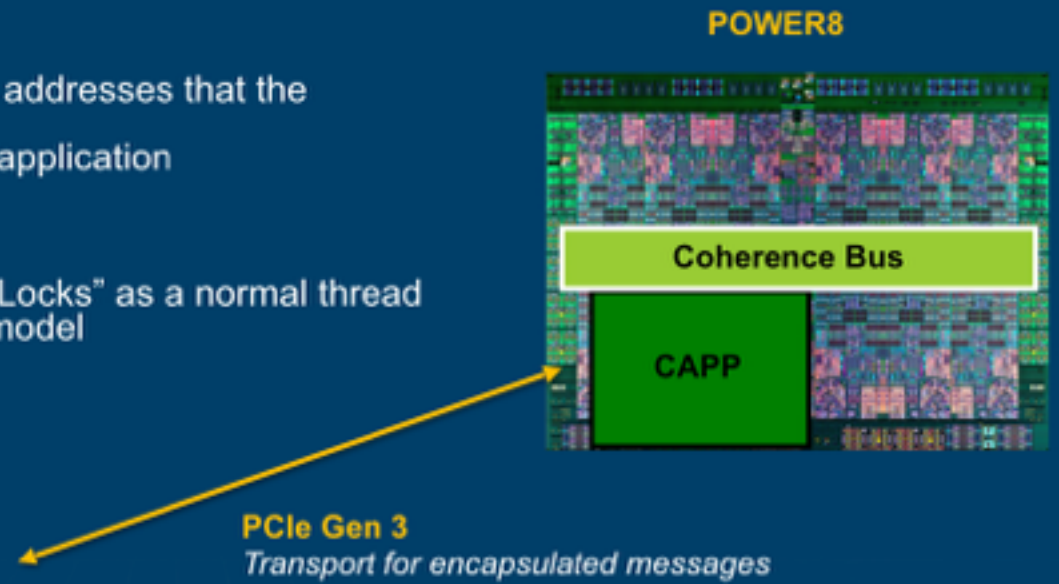


source: B. McCredie 2016 OpenPOWER summit

POWER8 CAPI Coherent Accelerator Processor Interface

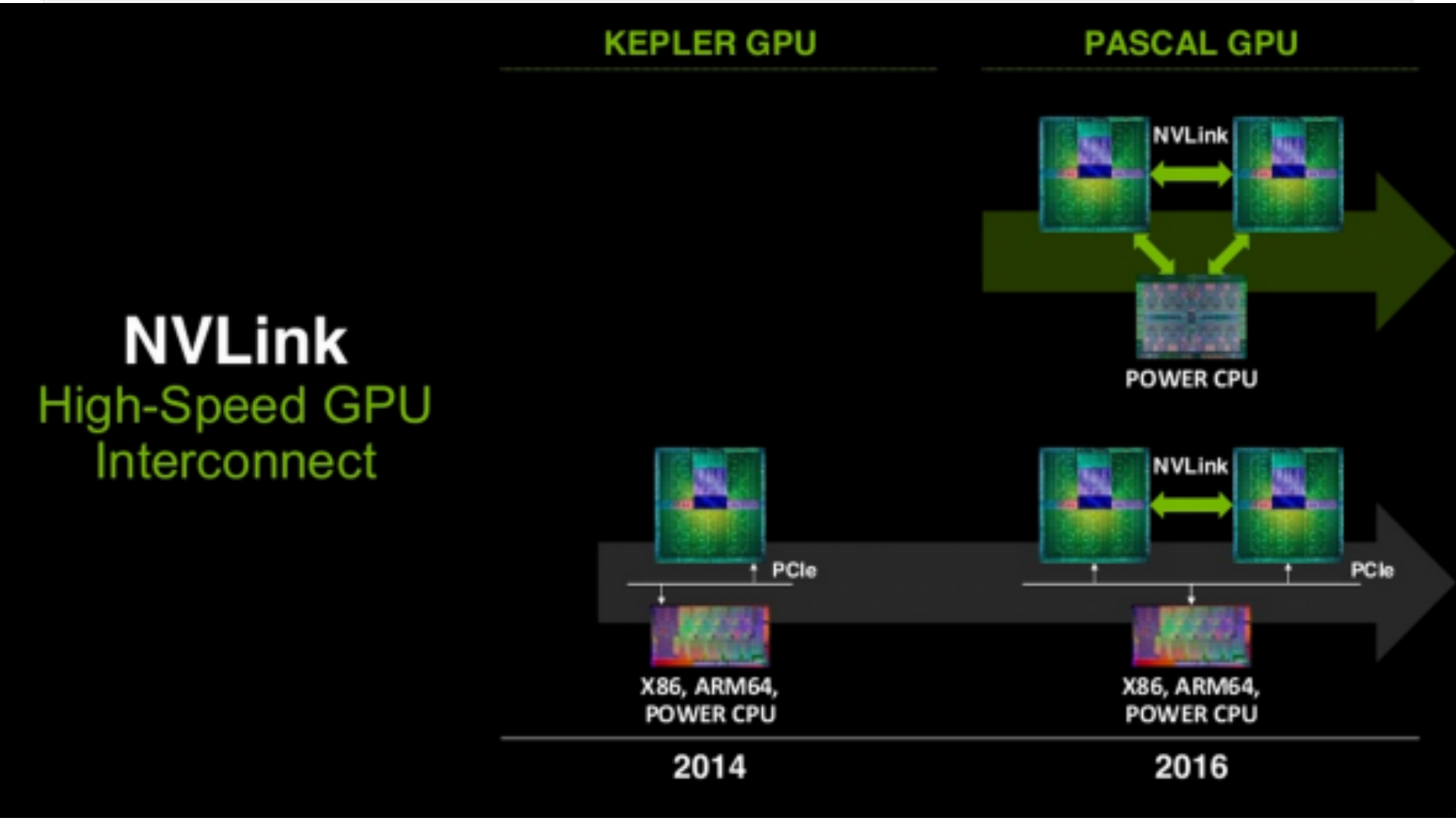
Virtual Addressing
- Accelerator can work with same memory addresses that the processors use
- Pointers de-referenced same as the host application
- Removes OS & device driver overhead

Hardware Managed Cache Coherence
- Enables the accelerator to participate in "Locks" as a normal thread
  Lowers Latency over IO communication model



Processor Service Layer (PSL)
- Present robust, durable interfaces to applications
- Offload complexity / content from CAPP

Customizable Hardware Application Accelerator
- Specific system SW, middleware, or user application
- Written to durable interface provided by PSL



5-12x PCIe Gen3 x16

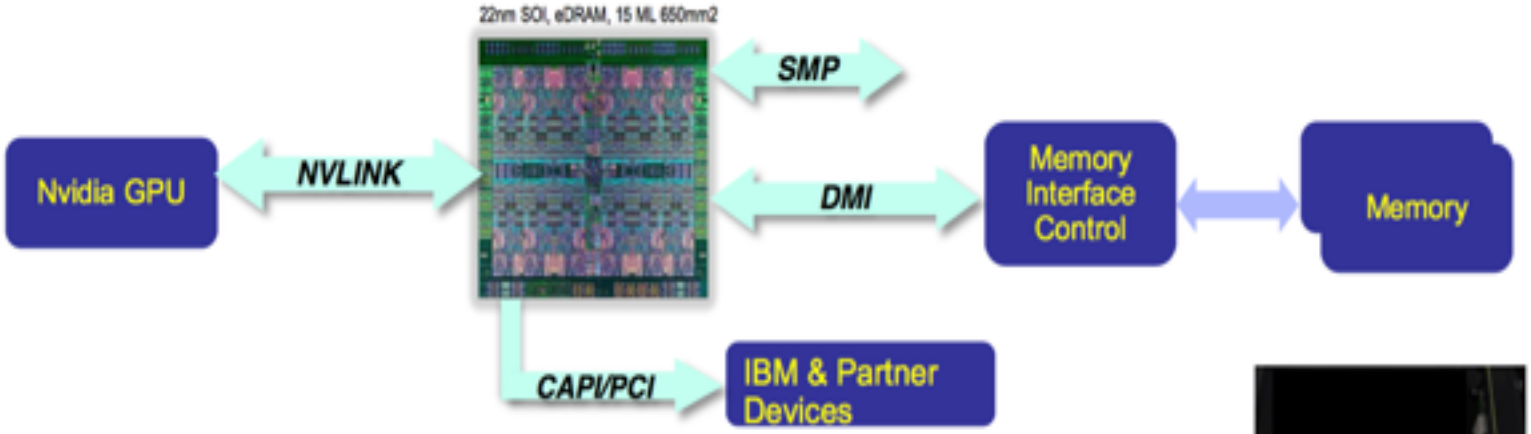
Source: NVIDIA

POWER8 + NVLink

Power Systems

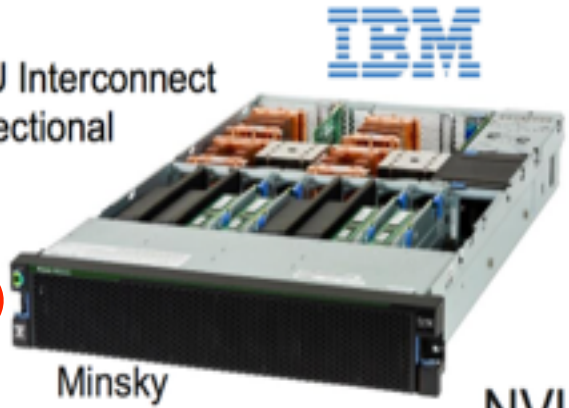
POWER8 with NVLink

IBM

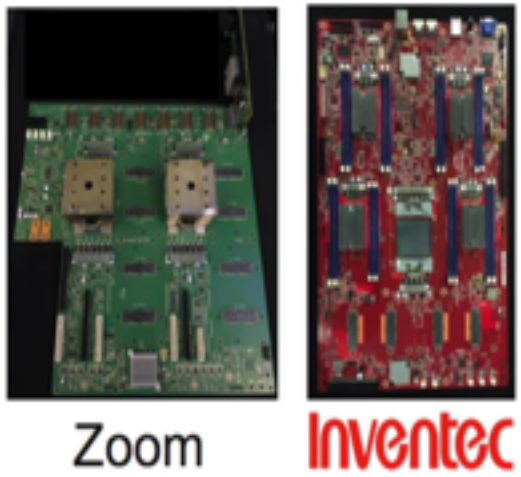


- NVLink High Speed CPU <-> GPU Interconnect
- 160+ GigaBytes per second bi-directional
- 5-12x faster than PCIe Gen3 x16

Nvlink Accelerator Lab
accellab@us.ibm.com



Minsky



Zoom Inventec

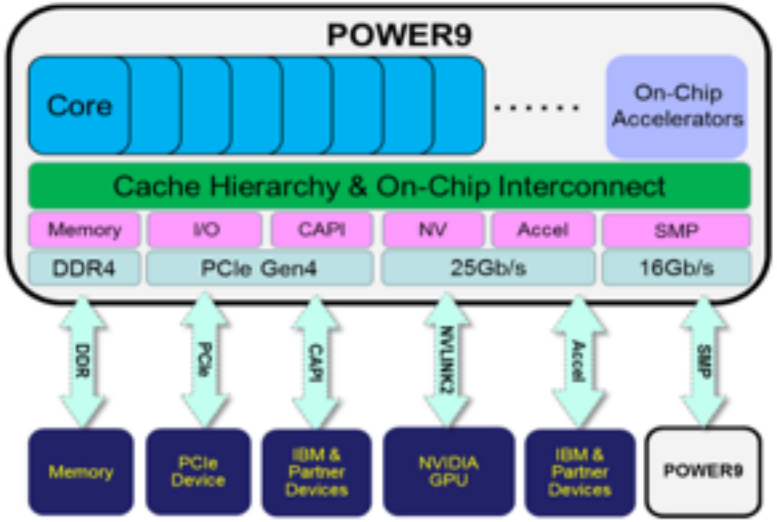
NVLink POWER Systems

source: B. McCredie 2016 OpenPOWER summit

5: New high-bandwidth standard for coherent attach: OpenCAPI (CAPI 2.0)

Power Systems

POWER9 SO with Advanced Accelerator Attach



24 newly designed POWER9 cores

- Leveraging execution slices for improved performance on cognitive, analytic, and big-data applications

Large, low-latency, eDRAM cache for big datasets

Global Foundries 14HP finFET technology with eDRAM

Cloud-focused innovation in Energy Efficiency, Security, and Quality of Service

State-of-the-art IO subsystem using PCIe Gen4

Leadership platforms for hardware acceleration

- High bandwidth, GPU interconnect (NV link2.0)
- Next-generation CAPI2.0 interface for coherent accelerator and storage attach
- On-chip compression & cryptography accelerators
- New 25Gb/s advanced accelerator attach bus

1st chip in POWER9 family

- Optimized for 2 socket scale out servers & hyperscale datacenters
- DDR4 direct attach memory channels

Full POWER9 family will address a broad range of scale out & enterprise servers

7

source: B. McCredie 2016 OpenPOWER summit

Get in the game: Free resources to try this out & develop your own …

FPGA Acceleration Service

3 use cases to experience the FPGA acceleration services.

- **Try FPGA acceleration service with a demo**
- **How to upload a private accelerator**
- **How to create a VM with accelerator**

Try FPGA acceleration service with a demo

Play a demo with a pre-installed application and attached with an FPGA accelerator

- Visit SuperVessel Cloud (http://www.ptopenlab.com), click "Application Acceleration"
- Clone a dedicated virtual machine to run the Informix® demo. Get your host IP address.

- Connect VPN, go to Informix® dashboard and play the demo

How to upload private accelerator

To work with user's own accelerator, upload it to the cloud.

- Login to the FPGA Maker Zone

- Fill in accelerator information and upload.

① Upload CAPI accelerator or PCIe accelerator
② Choose a picture in local disk
③ Choose the accelerator package in local disk

- Check accelerator deployment status. All procedures should be green.

How to create VM with accelerator

Work in Linux environment with FPGA accelerator attached.

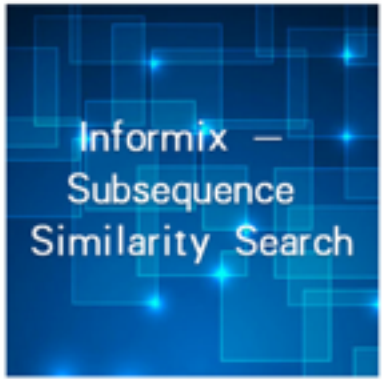
- Visit SuperVessel Cloud and click Apply VM->Launch instance

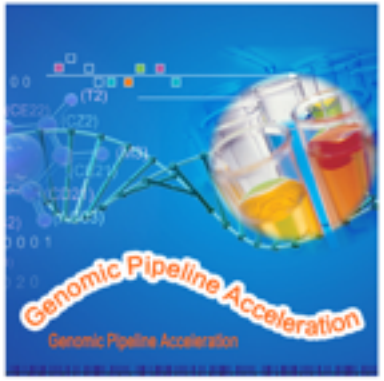
① For PCIe, use KVM image. for CAPI, use docker
② Choose public or private accelerator
③ Launch

- Login the host with web console or SSH tool, start your development

- **Apply SuperVessel services from here:**
  http://www.ptopenlab.com
- **Facebook:**
  http://www.facebook.com/supervesselcloud
- WeChat "SuperVessel"
- SuerVessel QQ group **344373069**

Get in the game: SuperVessel examples

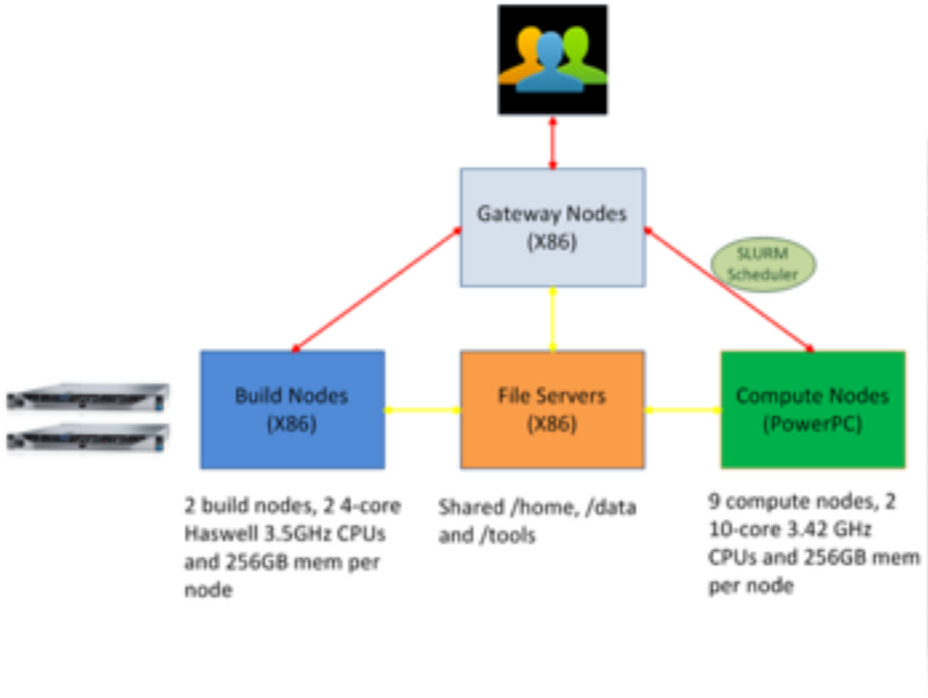




FAbRIC: FPGA Accelerator Research Infrastructure Cloud

IBM POWER8+CAPI cluster

The IBM POWER8+CAPI Cluster is a cluster of several x86 servers and nine POWER8 servers. Each POWER8 node is a heterogeneous platform capable of running GPGPU- and/or FPGA- accelerated applications.



Our current setup supports three accelerating devices:

- Nallatech 385 A7 Stratix V Altera-based FPGA adapter
- Alpha-data 7V3 Virtex7 Xilinx-based FPGA adapter
- NVIDIA Tesla K40m GPGPU card

https://wikis.utexas.edu/display/fabric/Home

OpenPOWER
DEVELOPER CHALLENGE

Two tracks to challenge and win:

1. The Open Road Test
   - Port and optimize for OpenPOWER
   - Go faster with accelerators *(optional)*
2. The Spark Rally
   - Train an accelerated DNN and recognize objects with greater accuracy
   - Show you can scale with Spark

SuperVessel Cloud

OpenPOWER™

Key Dates

Register today
openpower.devpost.com

Register Now
Contest:
June 1 - Sep 1, 2016

Grand prizes include a trip to **Supercomputing 2016**
Other prizes include **iPads, Apple Watches**

Join the conversation at #OpenPOWERSummit

Thank You

Follow @OpenPOWERorg and use #OpenPOWER to share your thoughts and join the conversation!