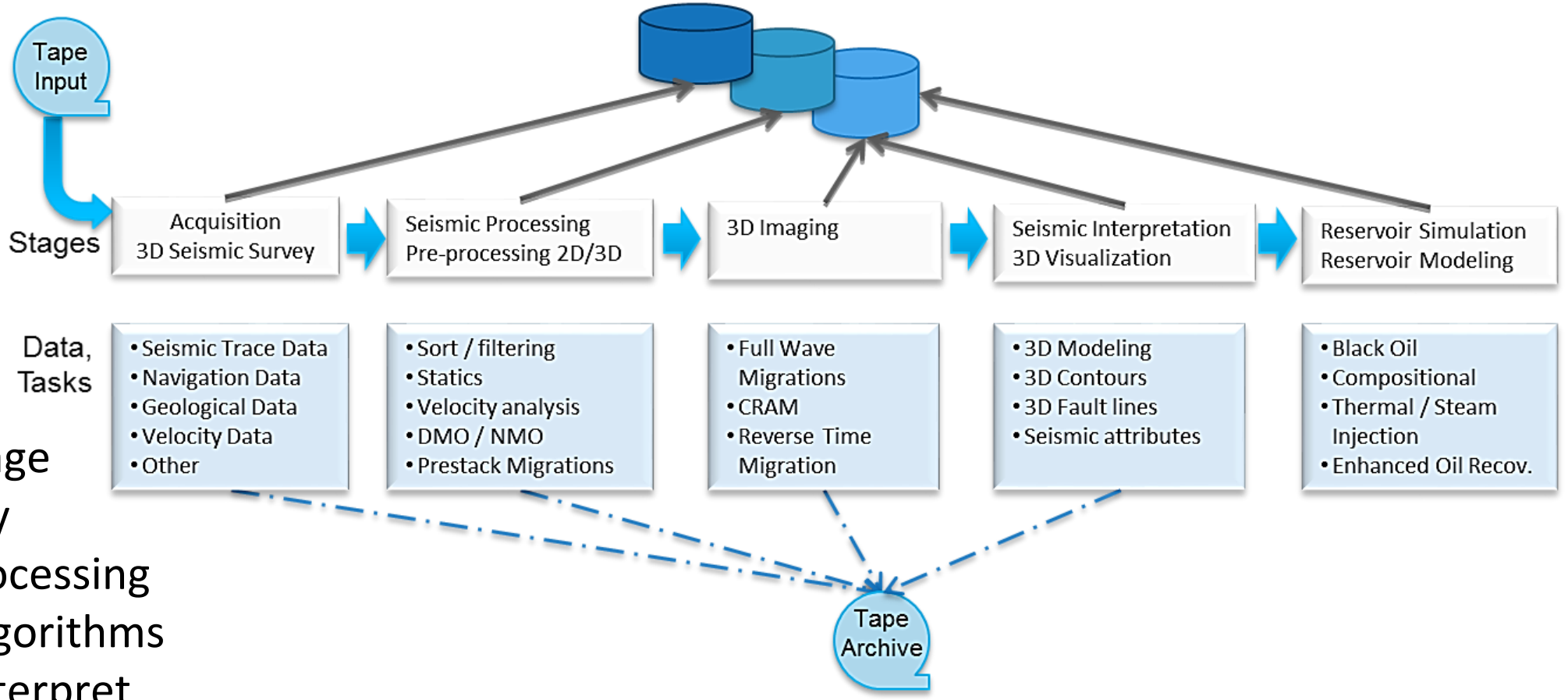
HPC Strategy & OpenPOWER

Our view of high-performance computing has evolved

- **The "Old View" of HPC**
  - The Value of an HPC System is Measured by FLOPS and TOP500 rank
  - The Objective is to Make an Algorithm Run Faster
  - HPC is a Special Category of Computing
  - HPC Looks Only at the Cluster/Server
  - Storage is an Afterthought

- **The IBM View of HPC**
  - Value is Measured by Application Performance
  - The Objective is to make a workflow optimized
  - HPC is another form of Analytics
  - Influx of Large Data demands consideration of Data Management and Storage in HPC: We Must Look Beyond the Server. Performance and data availability are imperative

Workflows Define HPC: Oil and Gas Example



- Read from storage
- Load to memory
- Perform pre-processing
- Execute RTM algorithms
- Visualize and interpret
- Simulate and model

The capability of any single piece of hardware is not what drives workflow.

Portfolio of HPC Solutions

Processors & Systems

- High Performance Processors & Systems
- Accelerator, networking, storage integration via NVLink & CAPI
- Highest memory throughput

High Speed Interconnect

- High speed interconnect / network fabric from Mellanox Technologies
- MPI acceleration in the IB fabric, reducing CPU overhead
- Support for GPUDirect, NVMe over fabric

High Performance File System & Storage

- Highest Performance HPC Storage: Elastic Storage Server
- High Performance Spectrum Scale (GPFS) Parallel File System
- Data centric design

HPC Software

- Deployment tools, integrated management
- Compilers: gcc, IBM XLC, LLVM OpenMP4, PGI Fortran/C/C++, Java, OpenACC, OpenMP
- Debuggers, Profilers, Math libraries, MPI & HPC apps

OpenPOWER: Open Architecture for HPC & Analytics

Processor IP Licensing	Licensing processor core to enable semiconductor partners like Suzhou Powercore to build POWER chips
Open Interfaces	Tight integration using CAPI & NVLink with Accelerators (NVIDIA, Xilinx), Networking (Mellanox), Storage (CAPI Flash)
Systems & Software	Enabling System Partners to build POWER-based servers and Open Sourcing Software including Firmware & Hypervisor

Introducing the OpenPOWER Foundation…
5 Founding members in 2013

OpenPOWER

2016: 250+ Members

Membership Options

Anyone may participate in OpenPOWER. Membership levels are designed for those that are investing to grow and enhance the OpenPOWER community and its proliferation within the industry.

Membership Level	Annual Fee $ USD	FTEs	Technical Steering Committee	Board / Voting position
Platinum	$100k	10	One seat per member not otherwise represented	Includes board position Includes TSC position
Gold	$60k	3	May be on TSC if Work group lead	Gold members may elect one board representative per three gold members
Silver	$20k $5k if <300 employees	0	May be on TSC if Work group lead	Sliver members may elect one board representative for all silver members
New Silver ISV	$0 if ISV is <300 employees	0	May be on TSC if Work group lead	Sliver members may elect one board representative for all silver members
Associate & Academic	$0	0	May be on TSC if Work group lead	May be elected to one community observer, non-voting Board seat

www.openpowerfoundation.org

2300+ Linux Applications on POWER

HPC

CHARMM
GROMACS
NAMD
AMBER
RTM
GAMESS
WRF
HYCOM
HOMME
LES
MiniGhost
AMG2013
OpenFOAM
miniDFT
CTH
BLAST
Bowtie
BWA
FASTA
HMMER
GATK
SOAP3
STAC-A2
SHOC
Graph500
Ilog

Cloud

openstack
TOMCAT
Apache
Chef
STORIX
REFLEXIS
DIASOFT
OneView Commerce
POWERSENSE
IFS
KANA.
FIS
YUCHENG Technologies Limited
sky solutions
betasystems
EXIGEN
ProjectWare
ALPHINAT

Big Data & Machine Learning

hadoop
PostgreSQL
MariaDB
redislabs
SUGARCRM
gpudb
Spark
Datameer
neo4j
EDB ENTERPRISEDB
AsiaINFO
CROSSVIEW
INTERSYSTEMS
ZAO
POLARIS
Valogix

Mobile Enterprise

zend
亿阳信通 BOCO Inter-Telecom
php
CSC
RabbitMQ Open Source Enterprise Messaging
Kingdee 金蝶
SYBASE An SAP Company
FIS
NARI 南瑞集团公司 NARI GROUP CORPORATION
FINEOS
ALPHINAT
comptel
Corent
TEMENOS
KUTIR

Major Linux Distros

ubuntu
SuSE
redhat
FreeBSD
Linux
Power
PowerVM

IBM Power Systems LC Line

OpenPOWER servers for cloud and cluster deployments that are different by design

Big Data

High Performance Computing

Compute Intensive

S812LC

1 socket, 2U

Storage rich for big data applications

Memory Intensive workloads

TYAN

S822LC For Big Data

2 socket, 2U

Storage-centric and high through-put workloads

Big data acceleration with work CAPI and GPUs

SUPERMICRO

S822LC For High Performance Computing

2 socket, 2U

POWER8 with NVIDIA NVLink

Up to 4 integrated NVIDIA “Pascal” P100 GPUs

wistron

S821LC

2 sockets, 1U

Dense computing

SUPERMICRO

S822LC

2 socket, 2U

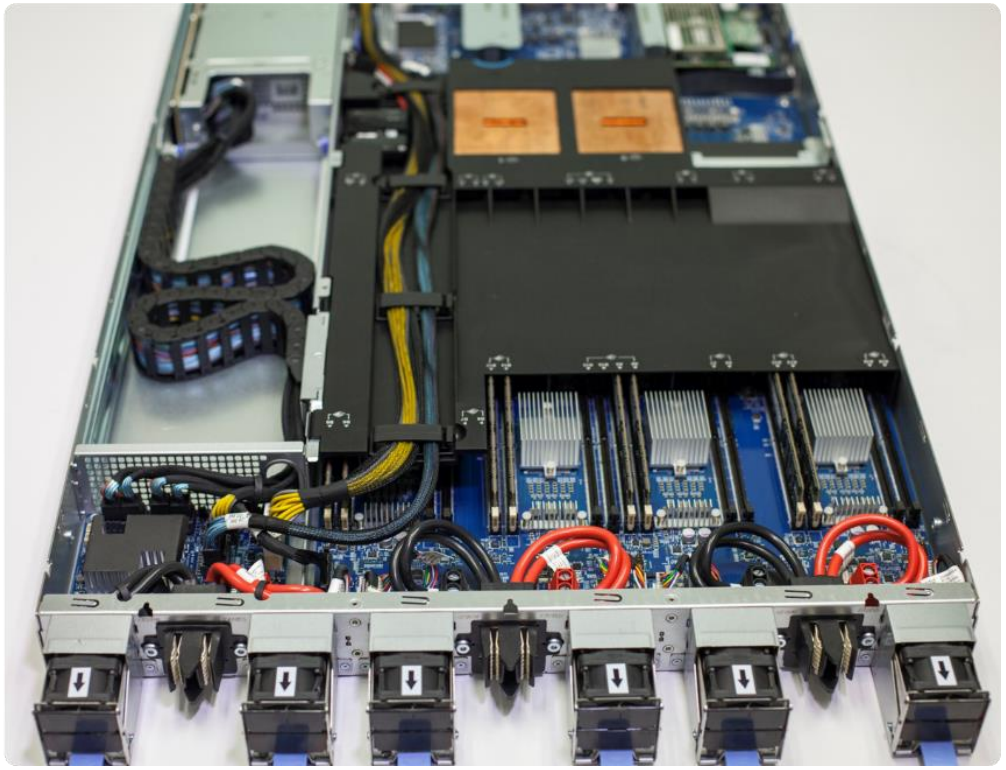
Memory Intensive workloads

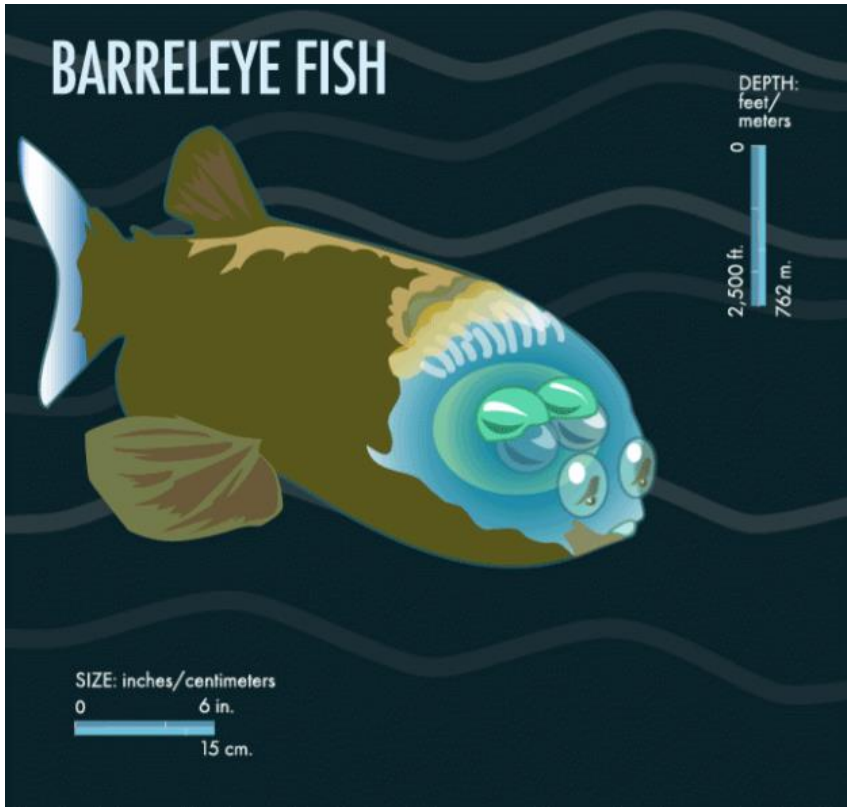
wistron

rackspace HOSTING

Available now: Barreleye

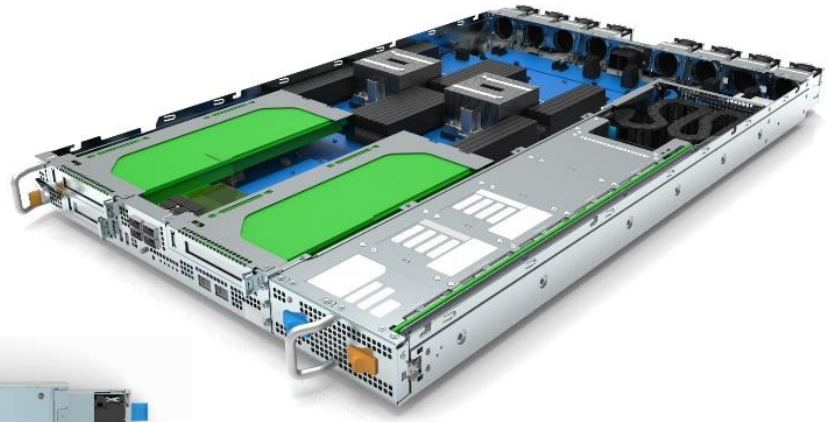
In partnership with Avago, IBM, Mellanox, PMC & Samsung

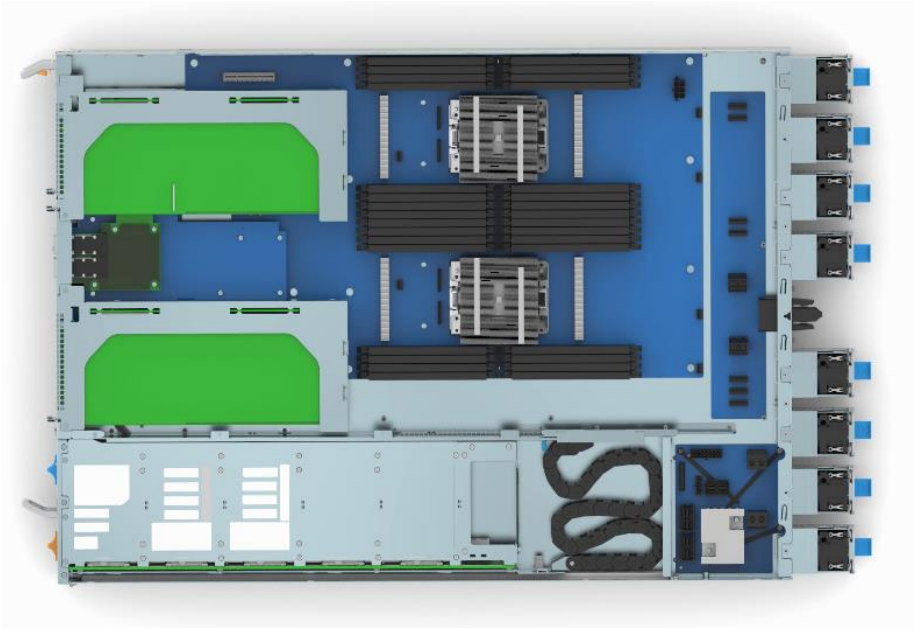




OPEN Compute Project

rackspace HOSTING Google





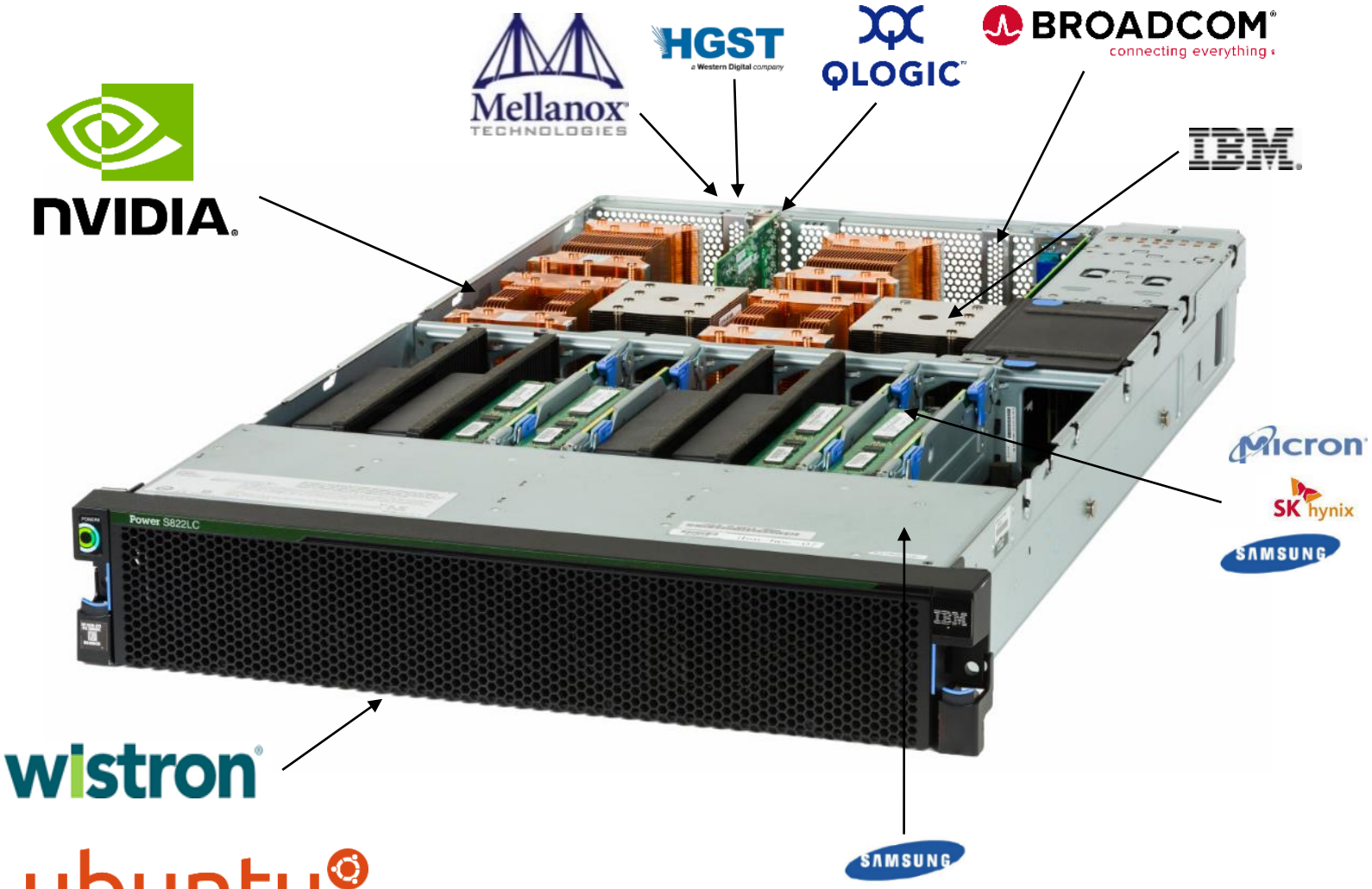
OPEN Compute Project

Zaius 1.25 OU

- 2 POWER9 CPUS
- 32 DDR4 DIMM SLOTS
- 2X G4 PCIE X16 FHFL SLOTS
- 1X G4 X16 HHHL SLOT
- 1X G4 X16 OCP MEZ
- 1X M.2 SATA PORT
- 1X SATA PORT
- 15X 2.5” SAS/SATA/NVME SLOTS
- BMC W/GBE LOM
- “DISKLESS” OPTION

OpenPOWER Innovation in the Design

Power Systems S822LC for High Performance Computing (aka Minsky)



NVIDIA:
Tesla P100 GPU Accelerator with NVLink (GPU↔GPU & GPU↔CPU)

Ubuntu by Canonical:
Launch OS supporting NVLink and Page Migration Engine

Wistron: Platform co-design

Mellanox: InfiniBand/Ethernet Connectivity in and out of server

HGST: Optional NVMe Adapters

Broadcom: Optional PCIe Adapters

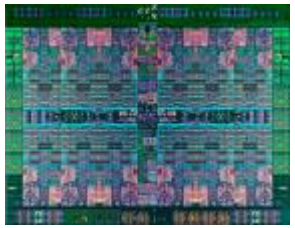
QLogic: Optional Fiber Channel PCIe

Samsung: 2.5” SSDs

Hynix, Samsung, Micron: DDR4

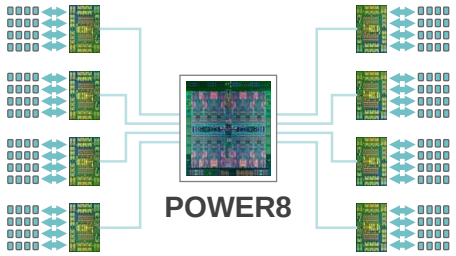
IBM: POWER8 CPU with NVLink

IBM Strategy for HPC Systems



High Performance Cores

Faster Cores than x86



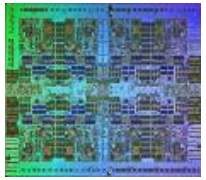
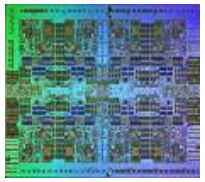
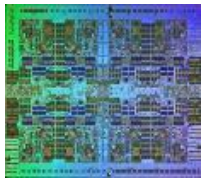
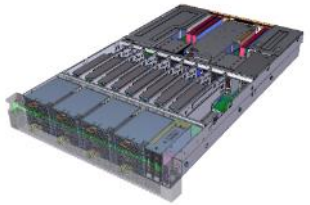
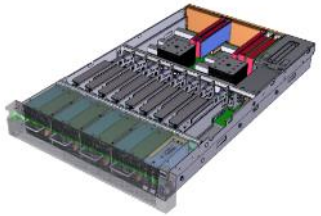
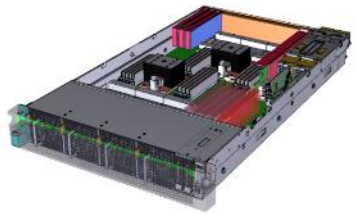
Fast & Large Memory System

Larger Caches Per Core than x86

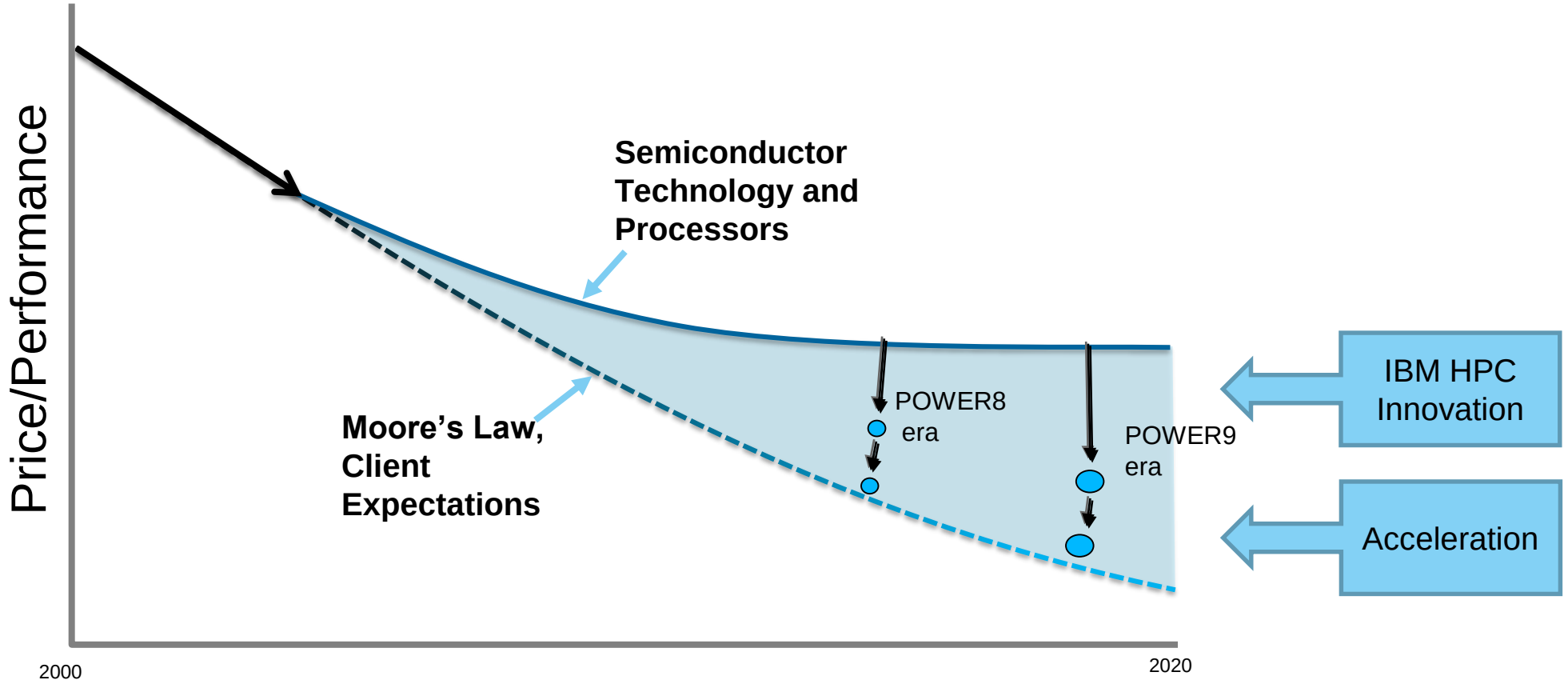
Fast PowerAccel Interconnects for Accelerators

5x Faster Data Communication between POWER8 & GPUs

Roadmap for HPC / HPDA

	2015	2016	2017
Mellanox Interconnect Technology	Connect-IB FDR Infiniband PCIe Gen3	ConnectX-4 EDR Infiniband CAPI over PCIe Gen3	ConnectX-5 Next-Gen Infiniband Enhanced CAPI over PCIe Gen4
NVIDIA GPUs	Kepler PCIe Gen3	Pascal NVLink	Volta NVLink Next Gen
IBM CPUs	POWER8  OpenPower CAPI Interface	POWER8 with NVLink  PowerAccel Interfaces: NVLink, CAPI, PCIe Gen3	POWER9  PowerAccel: Enhanced CAPI, NVLink Next Gen, PCIe Gen4
IBM Nodes			

Why Accelerators and GPUs?



Shift back towards the Moore's Law prediction through:

1. **IBM HPC Innovation** (processor architecture enhancement, scalable filesystems, workflow management
2. **Acceleration** through partner ecosystem (e.g. NVIDIA GPUs deliver 2X perf/watt)

POWER8: Designed Memory Bandwidth

IBM 22nm Technology

- Silicon-on-Insulator, 15 metal layers,
- ~4.2 billion Transistors
- Deep trench eDRAM

Compute

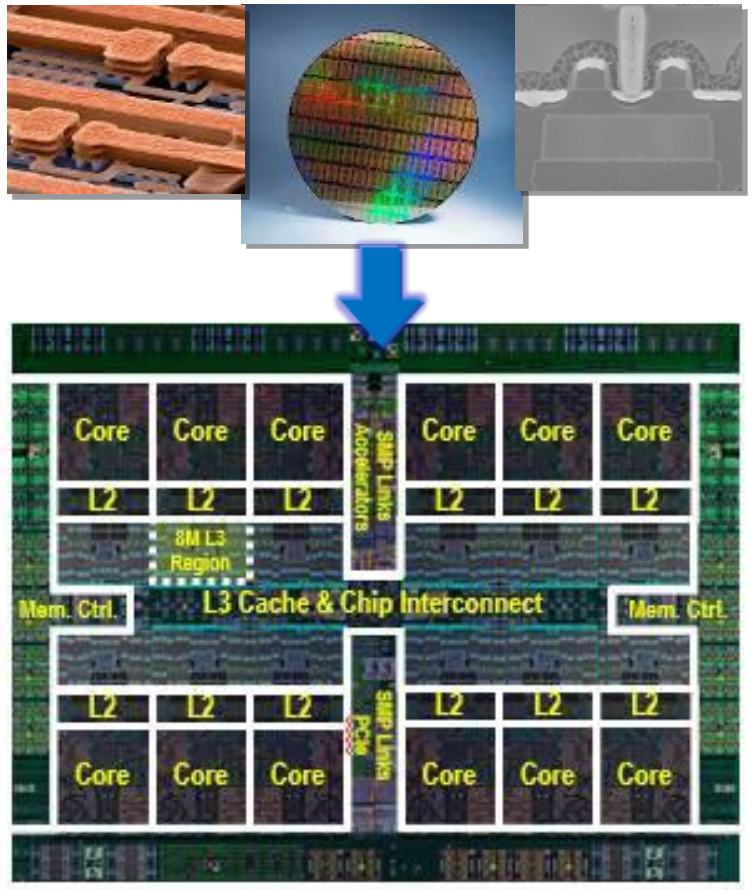
- 6/12 cores, ST/SMT2/SMT4/SMT8
- Enhanced, Auto balancing threads
- 8 dispatch/16 execution pipes/224 instructions in flight
- Transactional Memory/ Crypto & Crc instructions

Cache

- 64KB L1 + 512KB L2 / core
- 96MB L3 + up to 128MB L4 / socket

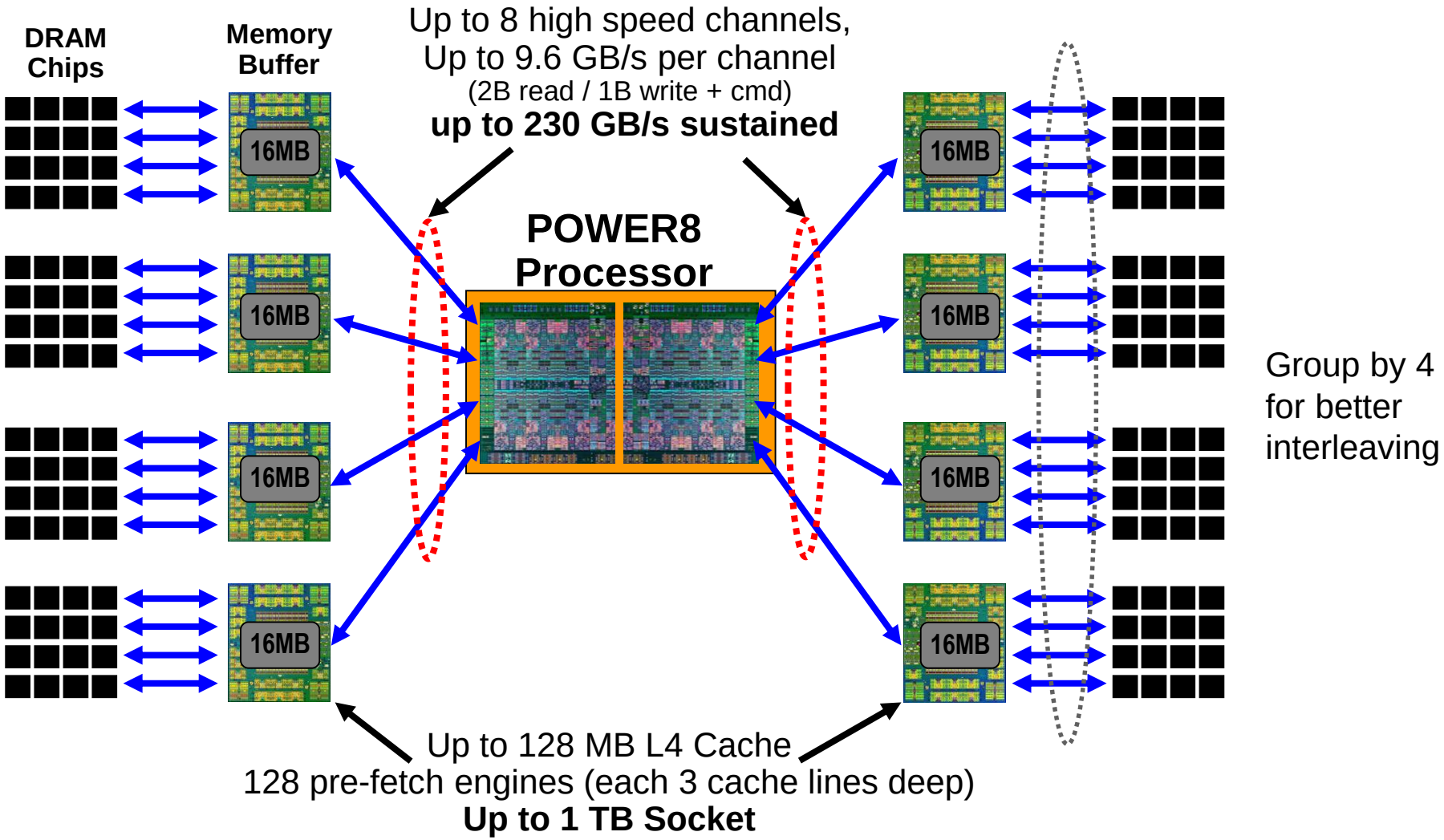
System Interfaces

- 230 GB/s memory bandwidth / socket
- Up to 48x Integrated PCI gen 3 / socket
- CAPI (over PCI gen 3)
- Robust, Large SMP Interconnect
- On chip Energy Mgmt, VRM / core



POWER8

POWER8 Memory Organization (Max Config)



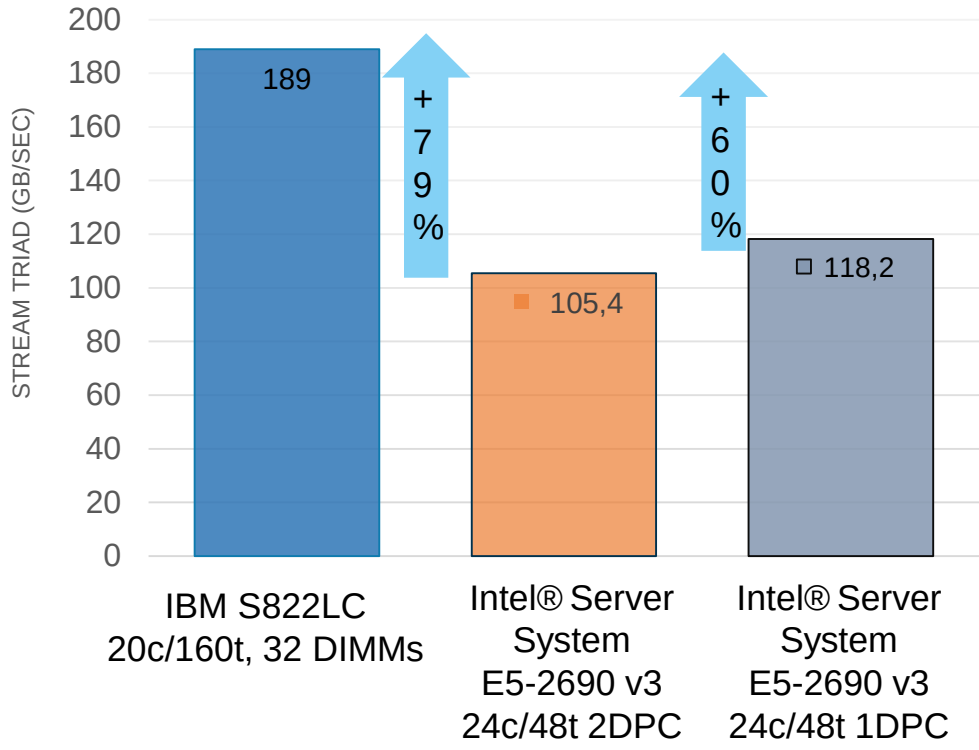
Memory bandwidth vs. most Xeon E5-2600v3 Configurations

Based on STREAM Triad memory bandwidth

Deliver 79% greater memory bandwidth compared to Xeon E5-2600 v3 configurations with 2DPC

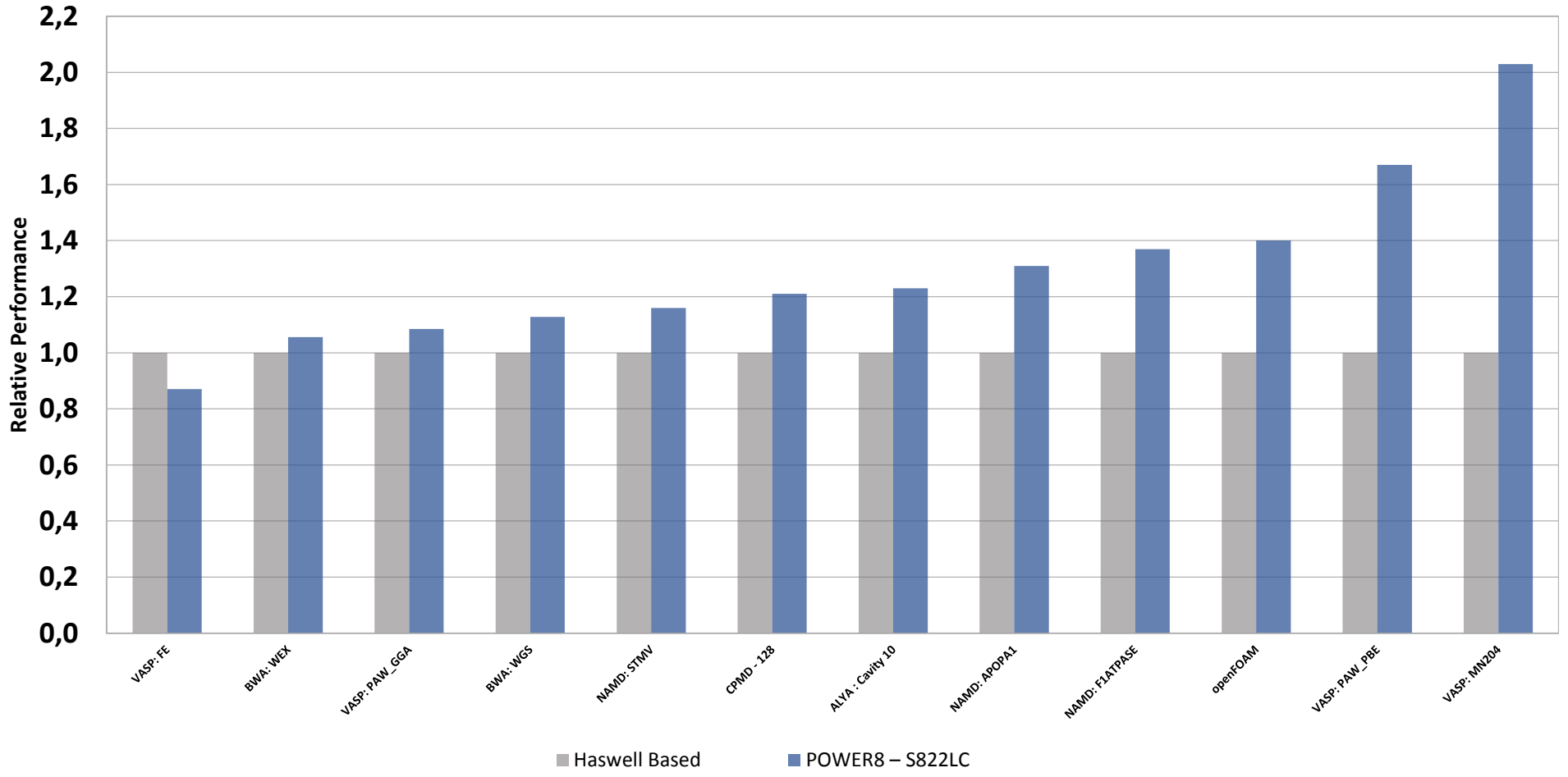
Deliver 60% greater memory bandwidth compared to Xeon E5-2600 v3 configurations with 1DPC

Only minor change vs Xeon E5-2600v4



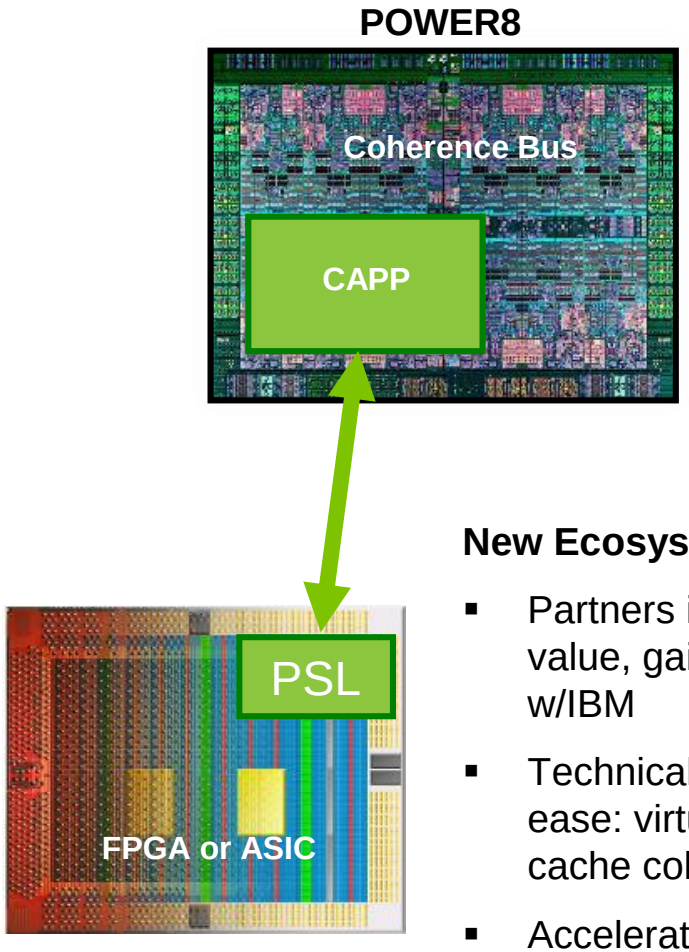
- IBM Power System S822LC results are based on IBM internal measurements of STREAM Triad; 20 cores / 20 of 160 threads active, POWER8; 3.5GHz, up to 1TB memory,
- Intel Xeon data is based on published data of Intel® Server System R2208WTTYS running STREAM Triad; 24 cores / 24 of 48 threads active, E5-2690 v3; 2.3GHz. For more details see http://www.intel.com/content/www/us/en/benchmarks/server/xeon-e5-2600-v3/xeon-e5-2600-v3-stream.html

What Does it Mean? Excellent CPU-Only Application Performance



Differentiated Acceleration - CAPI and NVLink

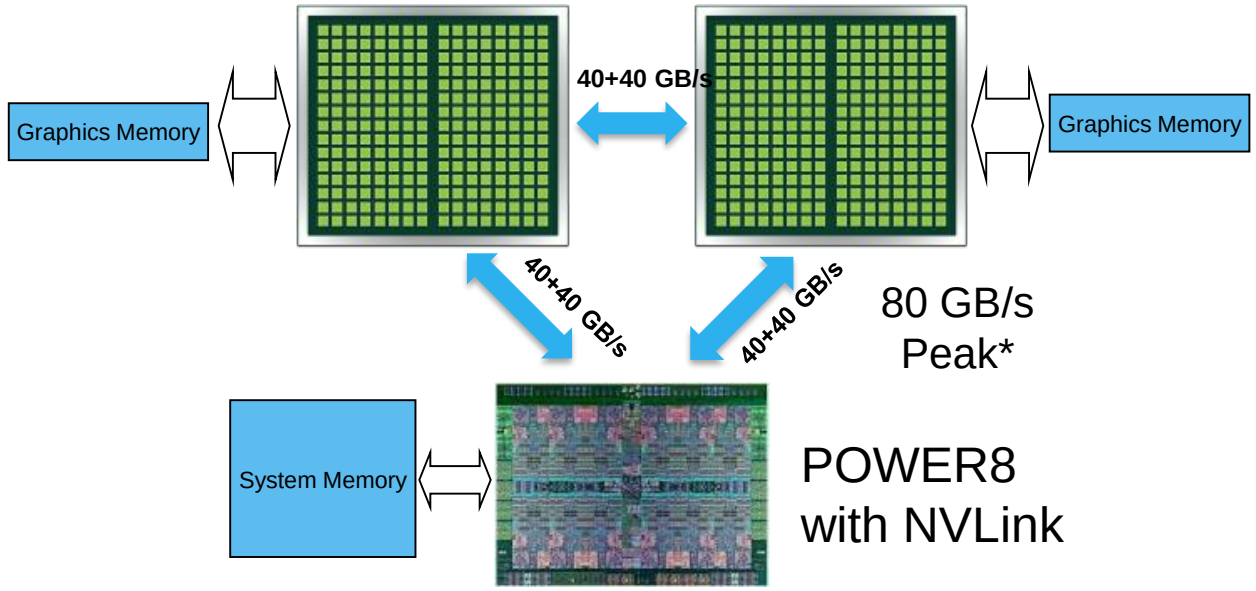
CAPI-attached Accelerators



New Ecosystems with CAPI

- Partners innovate, add value, gain revenue together w/IBM
- Technical and programming ease: virtual addressing, cache coherence
- Accelerator is hardware peer

NVIDIA Tesla GPU with NVLink



Future, Innovative Systems with NVLink

- Faster GPU-GPU communication
- Breaks down barriers between CPU-GPU
- New system architectures

Power 8 CAPI – Coherent Accelerator Processor Interface

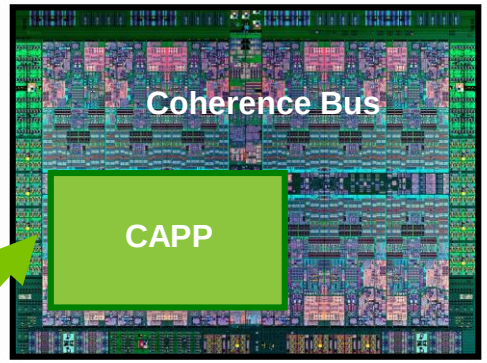
Virtual Addressing

- Accelerator can work with same memory addresses that the processors use

Hardware Managed Cache Coherence

- Enables the accelerator to participate in “Locks” as a normal thread Lowers Latency over IO communication model

POWER8



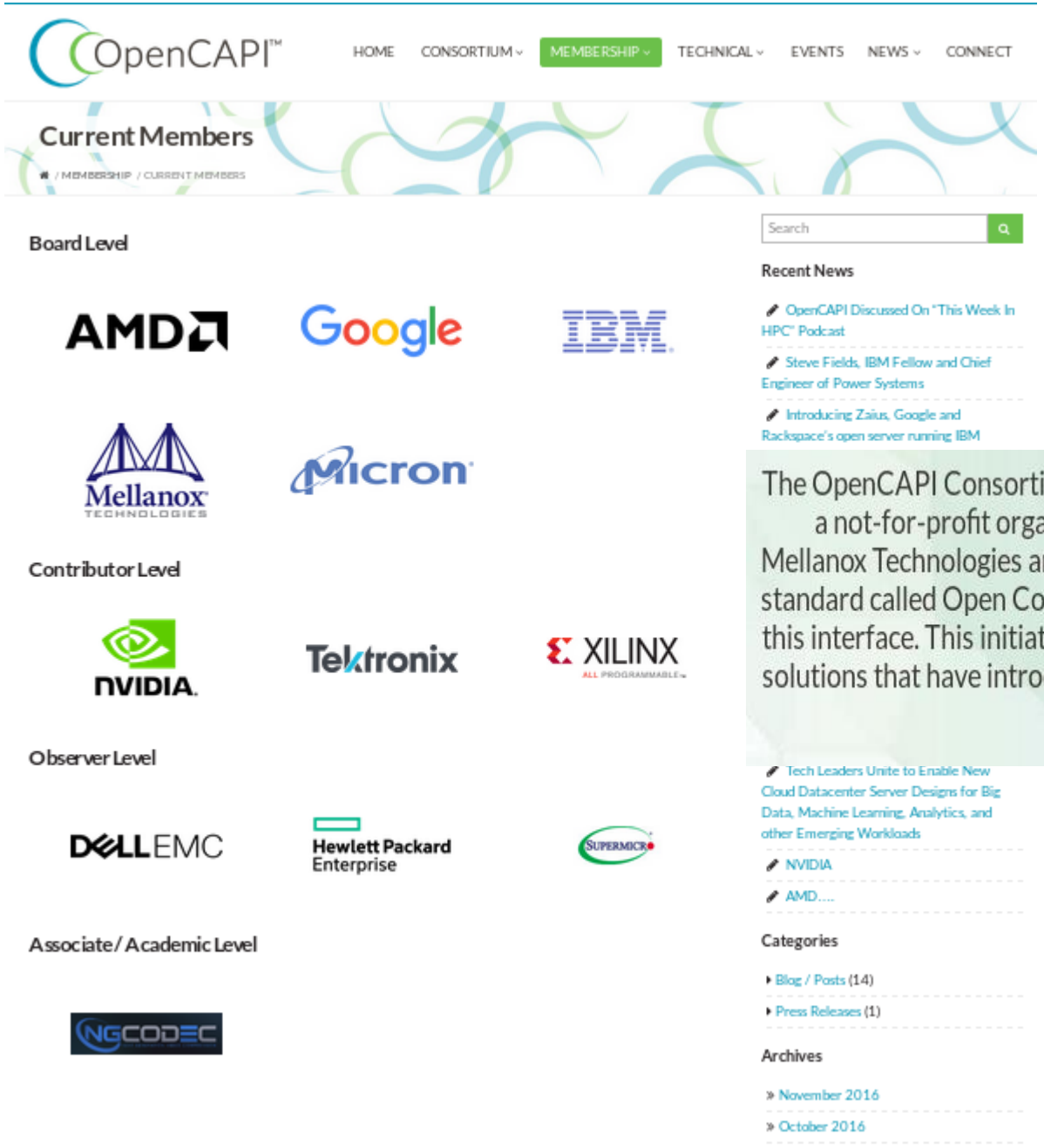
PCIe Gen 3
Transport for encapsulated messages

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



OpenCAPI.org

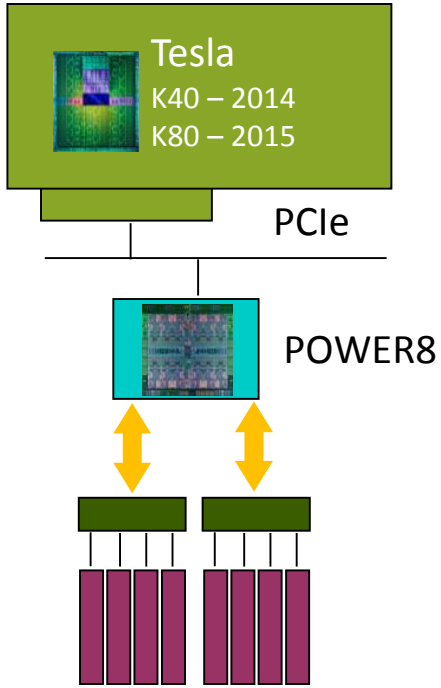
The OpenCAPI Consortium is an open forum to manage the OpenCAPI specification and ecosystem. OpenCAPI is a not-for-profit organization formed in October 2016 by OpenCAPI Board Members AMD, Google, IBM, Mellanox Technologies and Micron to create an open coherent high performance bus interface based on a new bus standard called Open Coherent Accelerator Processor Interface (OpenCAPI) and grow the ecosystem that utilizes this interface. This initiative is being driven by the emerging accelerated computing and advanced memory/storage solutions that have introduced significant system bottlenecks in today's current open bus protocols and requires a technical solution that is openly available

NVIDIA GPU Roadmap

Kepler

CUDA 5.5 – 7.0
Unified Memory

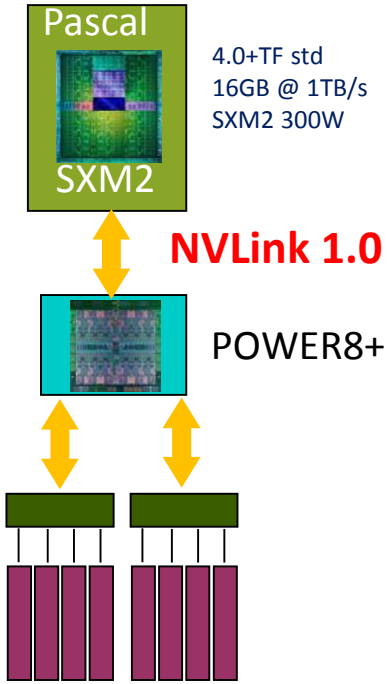
1 -2 GPU per board
235W – 300W versions
1.5TF – 1.9TF std; 2.7TF (boost)
12GB @ 288GB/s
Or 24GB @ 480GB/s



Buffered Memory

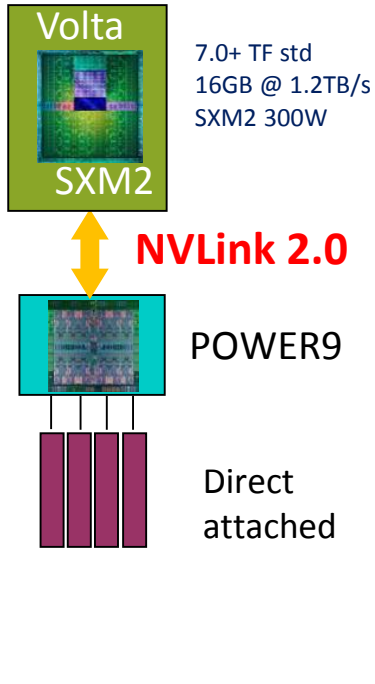
Pascal

CUDA 8
Full GPU Paging



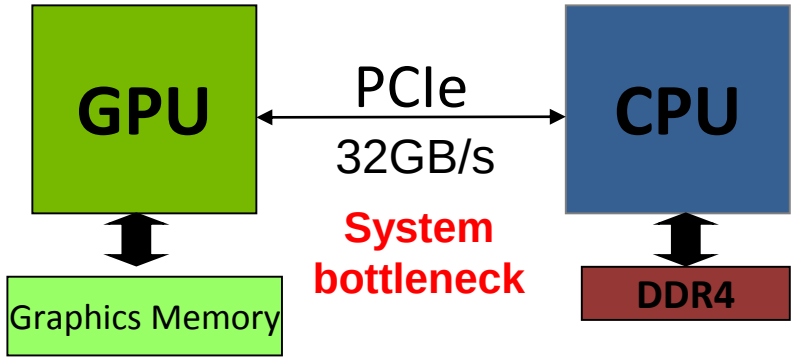
Volta

CUDA 9
Cache Coherent

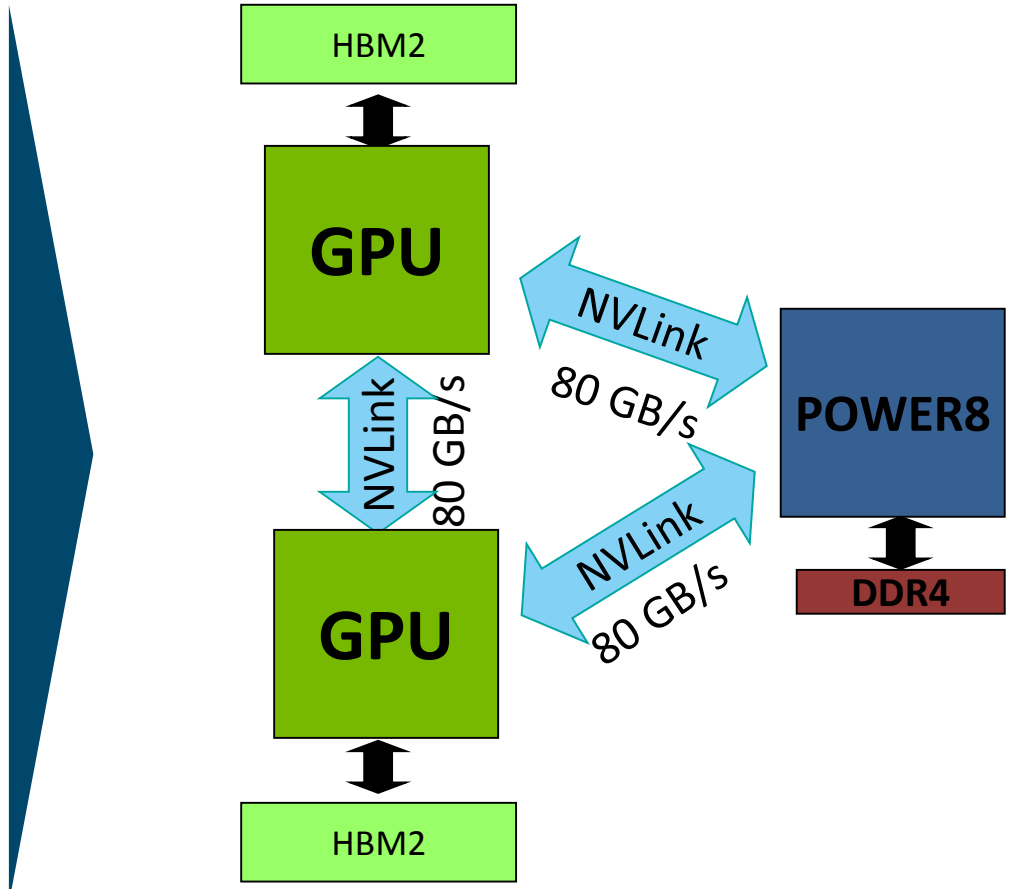


2014-2015 | **2016** | **2017**

POWER8 with NVLink: 2.5x Faster CPU-GPU Connection

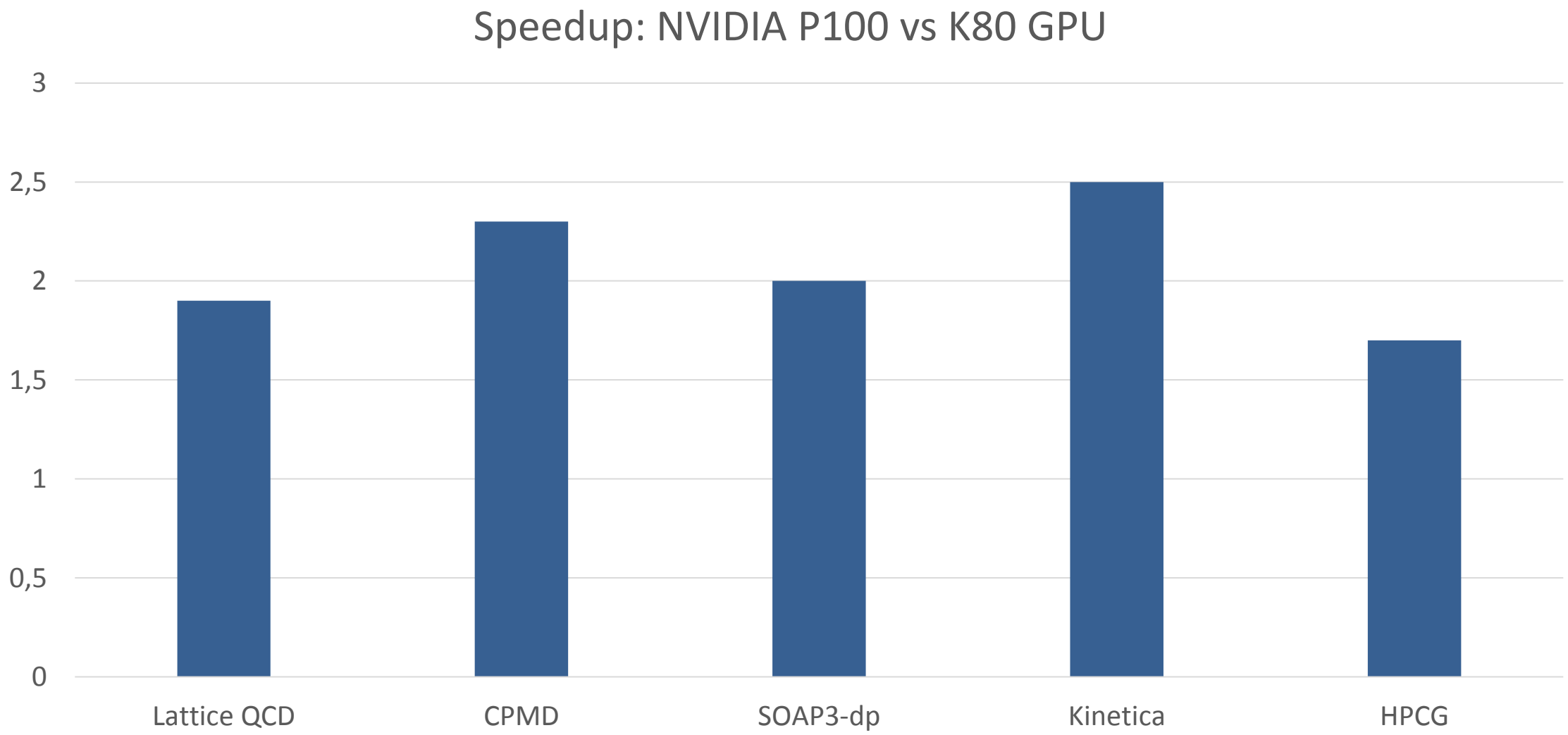


GPUs Limited by PCIe Bandwidth
From CPU-System Memory



NVLink Enables Fast Unified Memory Access
between CPU & GPU Memories

Early Performance Results on Minsky



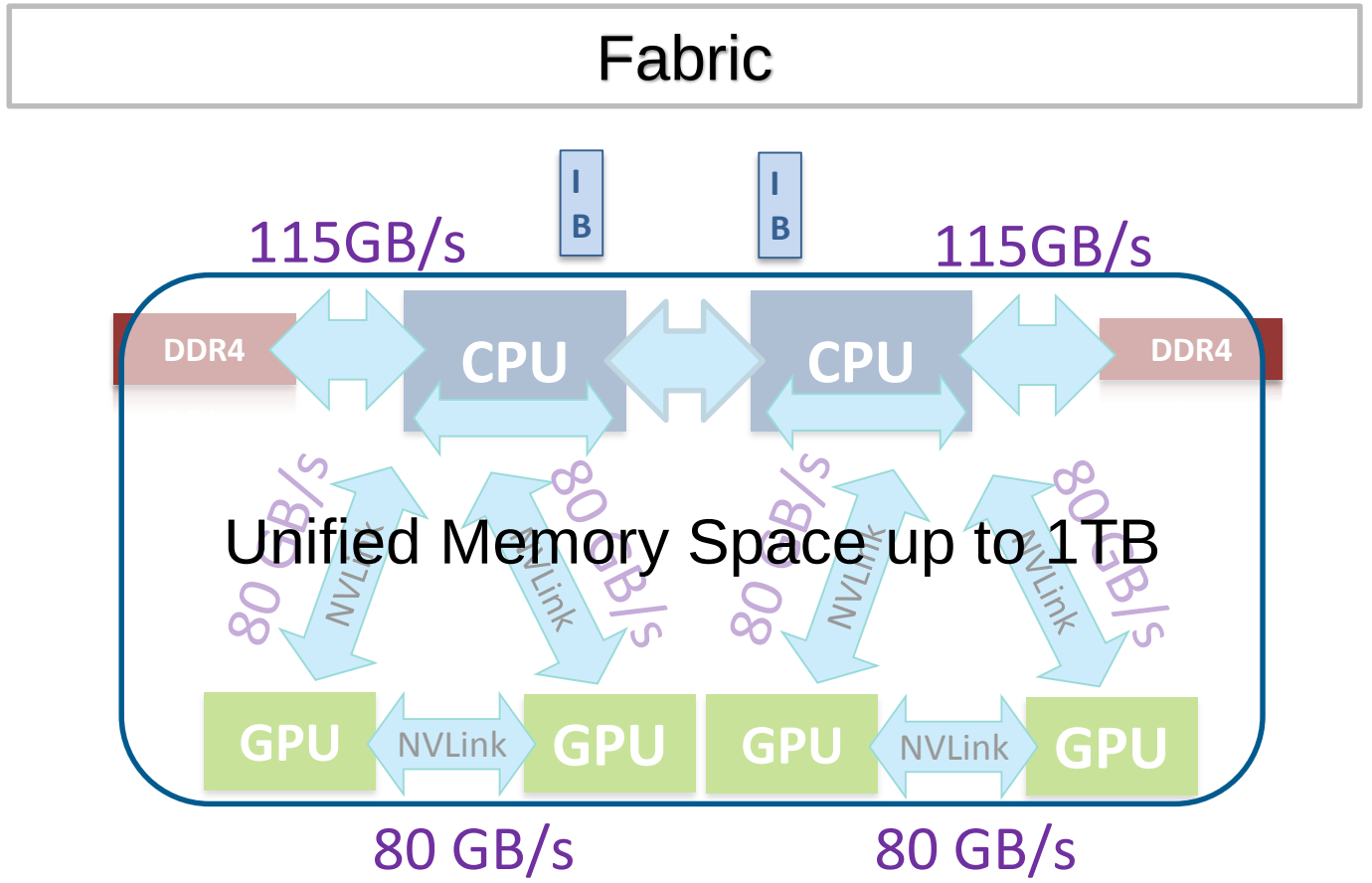
Better Design: Flat and Fat

Minksy is engineered both **flat** and **fat**

- Data flows freely across system
- Nearly as broad from CPU: GPU as System Memory: CPU
- Big pipes between GPUs on the same socket

Addresses PCI-E Bottleneck for numerous usage models

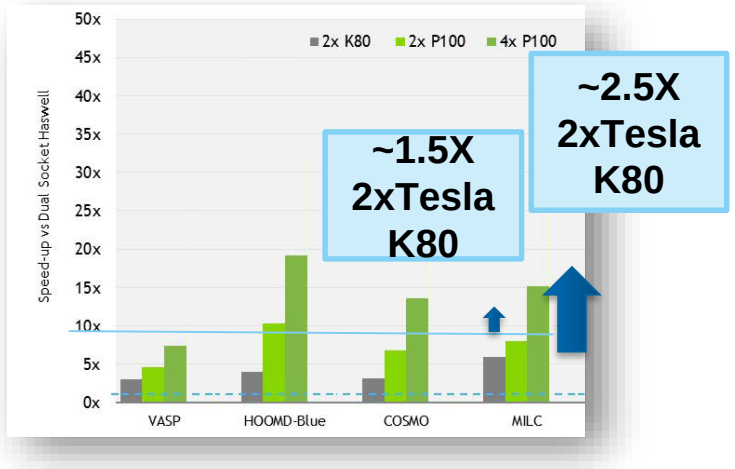
- Burst at startup/teardown
- Stream data constantly Host-Device
- Constant Transfers between 2 GPUs
- Hidden Bus Transfers from Host-Device
  (due to insufficient BW)



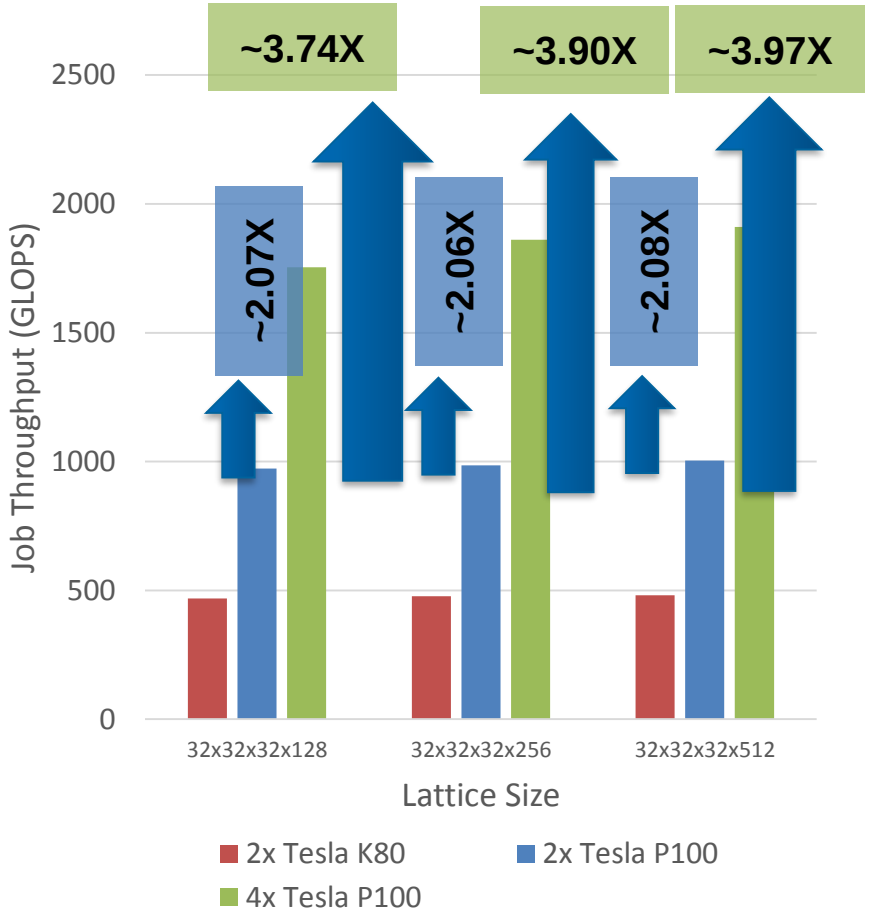
Performance improvement with Power Architecture

POWER8 with NVLink Platforms: up to a *4X performance uplift on Lattice QCD codes compared to their predecessors*

x86 Alternatives: typically delivering 1.5-2.5X performance differentials on the same types of code



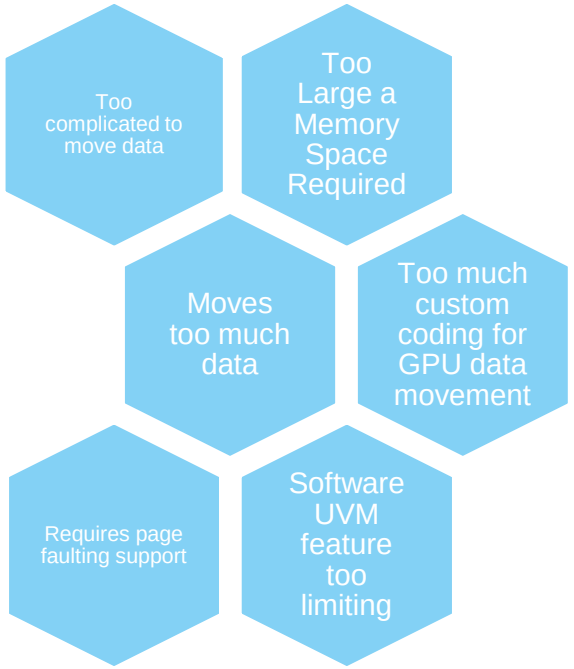
x86 Platform Speedup, vs CPU, 2x Tesla K80



Minksy Performance Increase vs 2x Tesla K80 System: MILC/LQCD

Page Migration Engine & POWER8 with NVLink

Barriers to Entry Removed



Far easier to create new applications on Tesla P100 + Minsky

- **NVIDIA Page Migration Engine** ensures unified memory space
  - **Unified memory:** address space spans CPU and GPU, 1TB+
  - **Hardware managed transfers:** eliminates explicit data transfers
  - Testing program implementing these advantages
- **POWER8 with NVLink** ensures speedy data throughput
  - 1TB memory space requires faster CPU:GPU data movement
  - Bus masks transfer times
- Close code-base to parallel CPU code

CPU Code

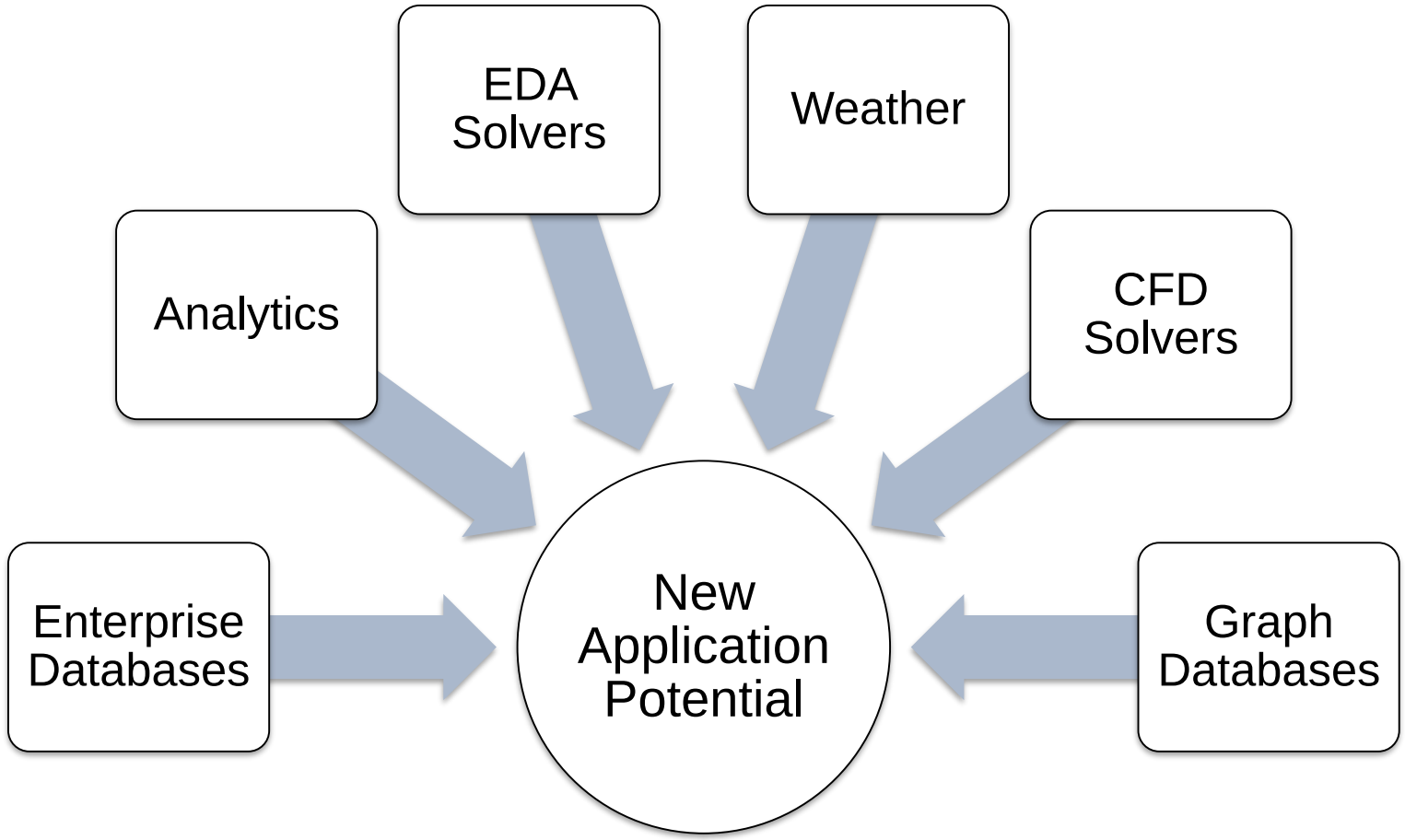
```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort(data, N, 1, compare);

  use_data(data);

  free(data);
}
```

Pascal Unified Memory*

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  free(data);
}
```

*With operating system support

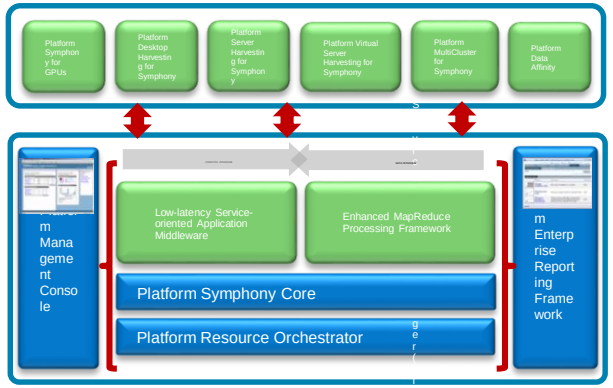
Application Potential Unlocked By Page Migration Engine and NVLink



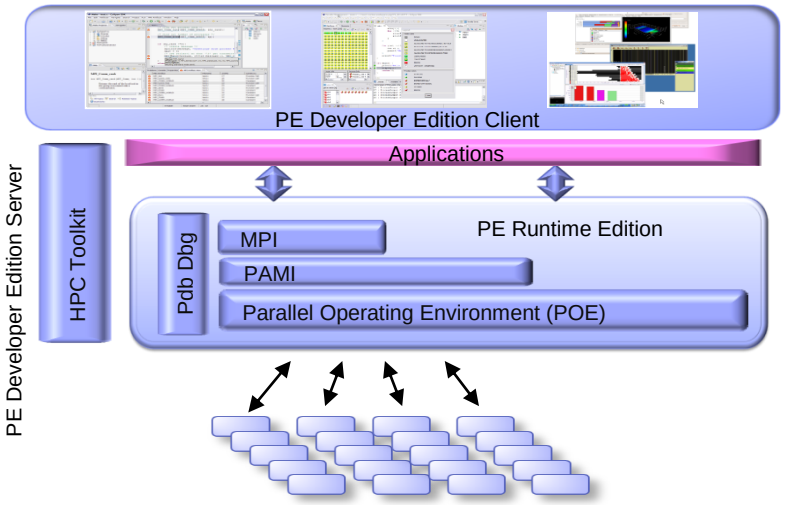
IBM Technical Computing Software Portfolio

Spectrum LSF
Workload Management

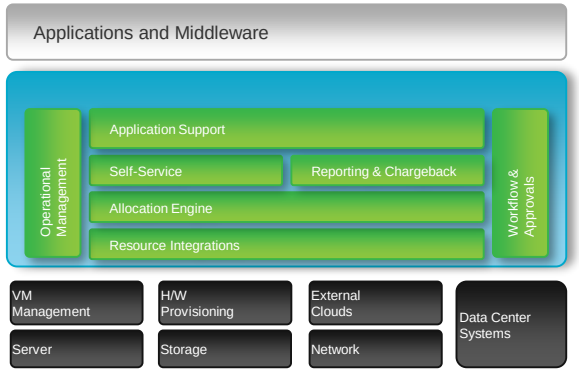
Spectrum Symphony
HPC Grid Services Management



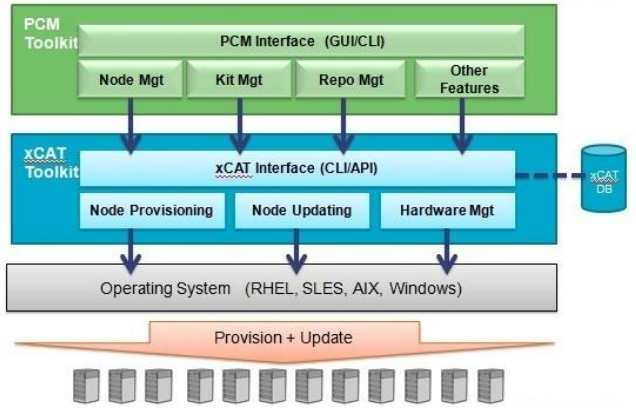
Parallel Environment (PE)



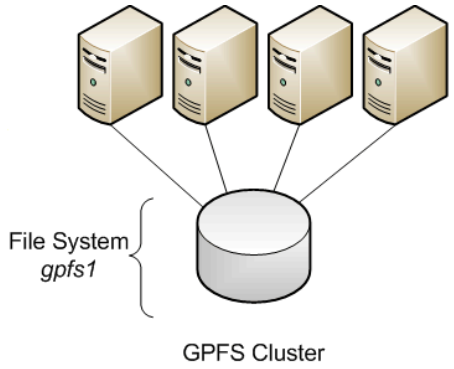
Spectrum Cluster Foundation
Dynamic HPC Infrastructure Management



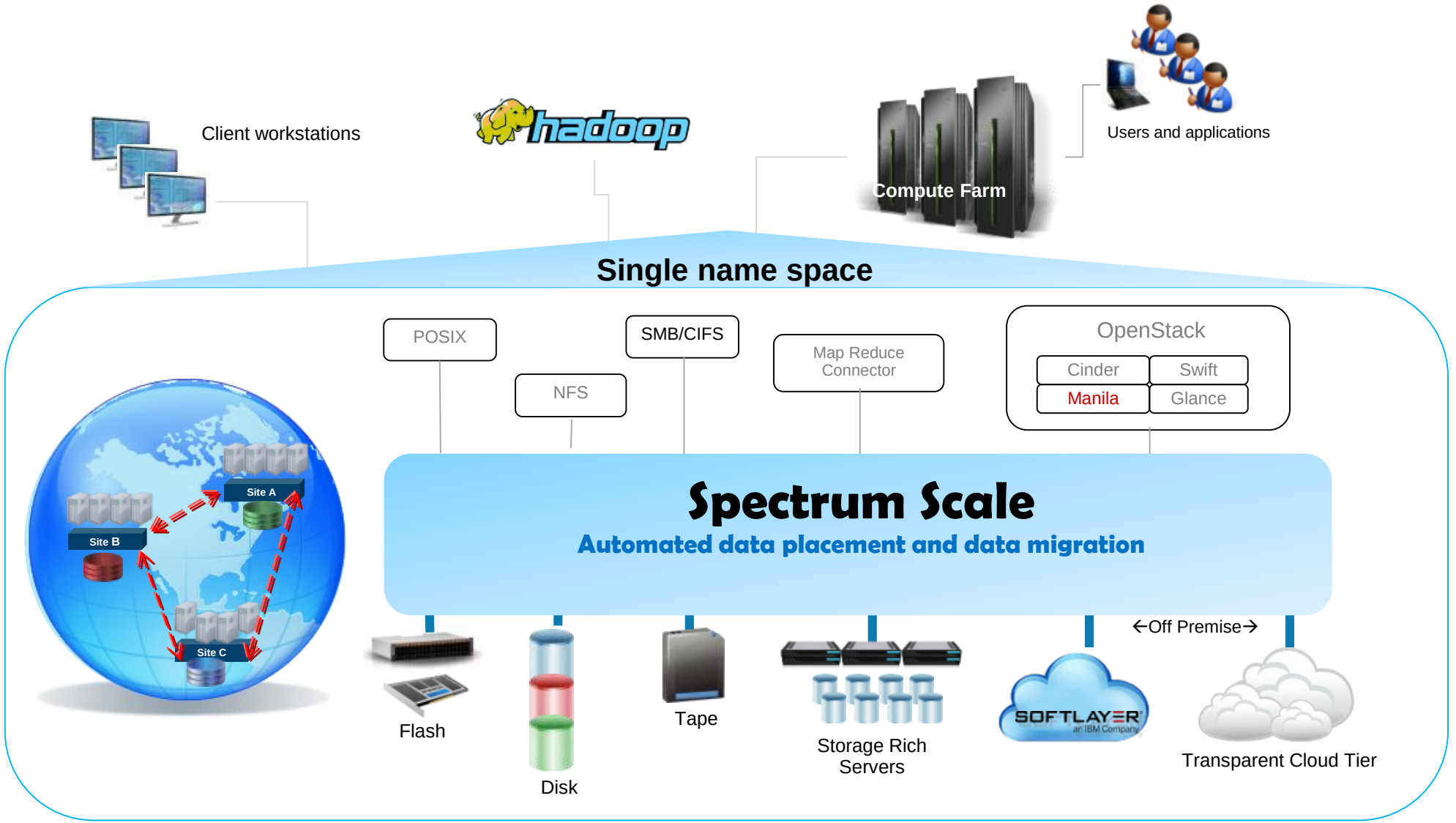
Spectrum Cluster Foundation
Systems Management and Provisioning



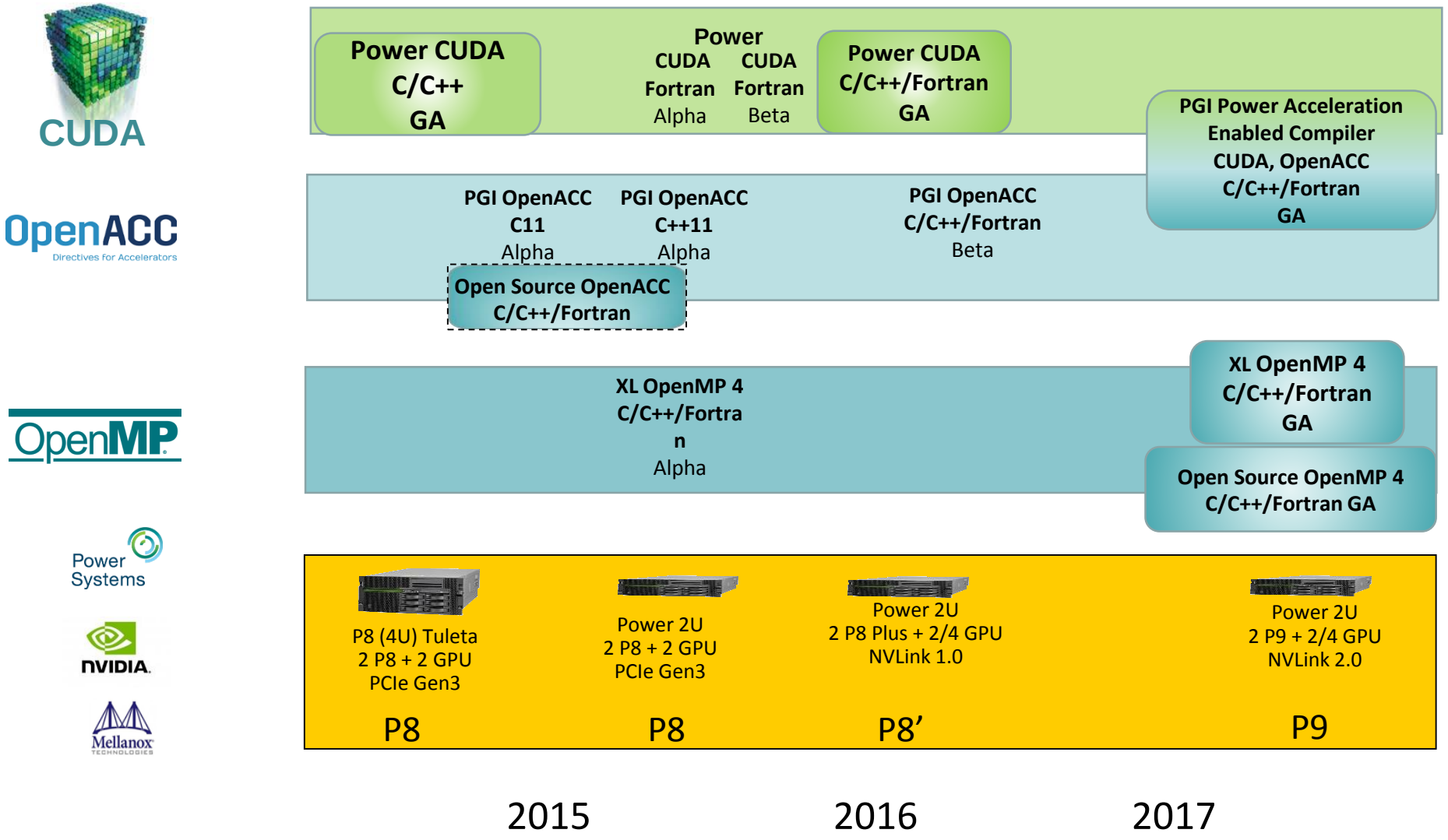
Spectrum Scale (GPFS)
File System Management



Spectrum Scale – An High Performance Parallel File System



Power GPU Acceleration for HPC Compiler Roadmap



Summarizing our strategy

- IBM remains committed to HPC
- We have a long term HPC roadmap already committed to multiple customers
- OpenPower is a broad play for entire HPC market, not just high end, and offers an alternative to the x86 monoculture
- Power outperforms x86 on key HPC apps
- We are actively attracting developers and ISVs to our platform
- We have differentiated solutions with accelerators and networking with CAPI and NVLink
- We have excellent storage solution for HPC (ESS)
- IBM Research is paving the way to exascale through innovation and collaboration