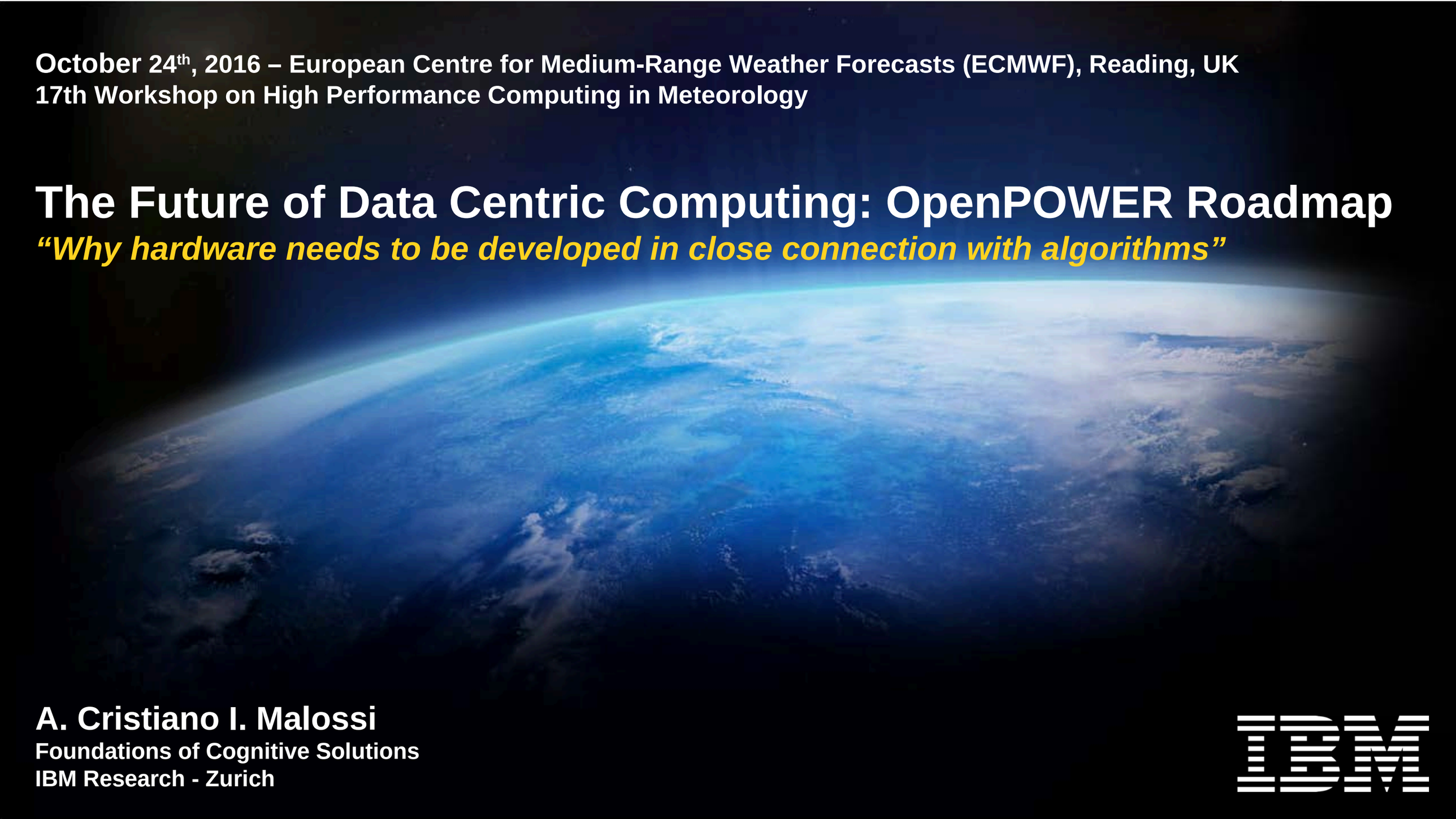
October 24th, 2016 – European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
17th Workshop on High Performance Computing in Meteorology

The Future of Data Centric Computing: OpenPOWER Roadmap

"Why hardware needs to be developed in close connection with algorithms"



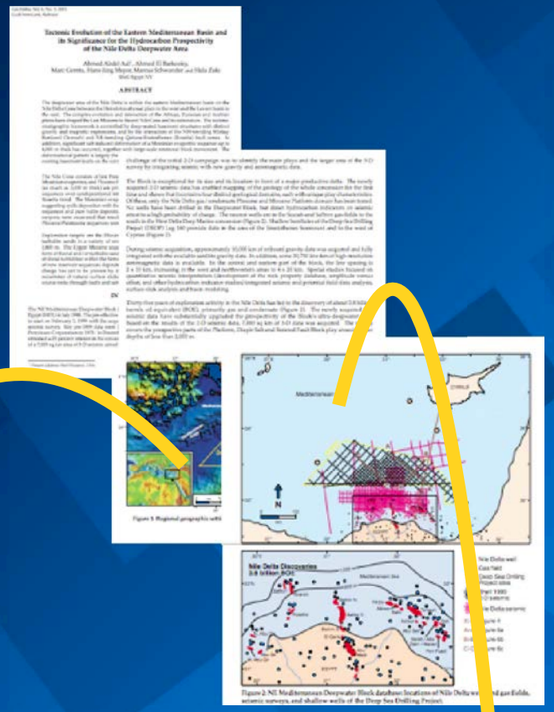
A. Cristiano I. Malossi
Foundations of Cognitive Solutions
IBM Research - Zurich

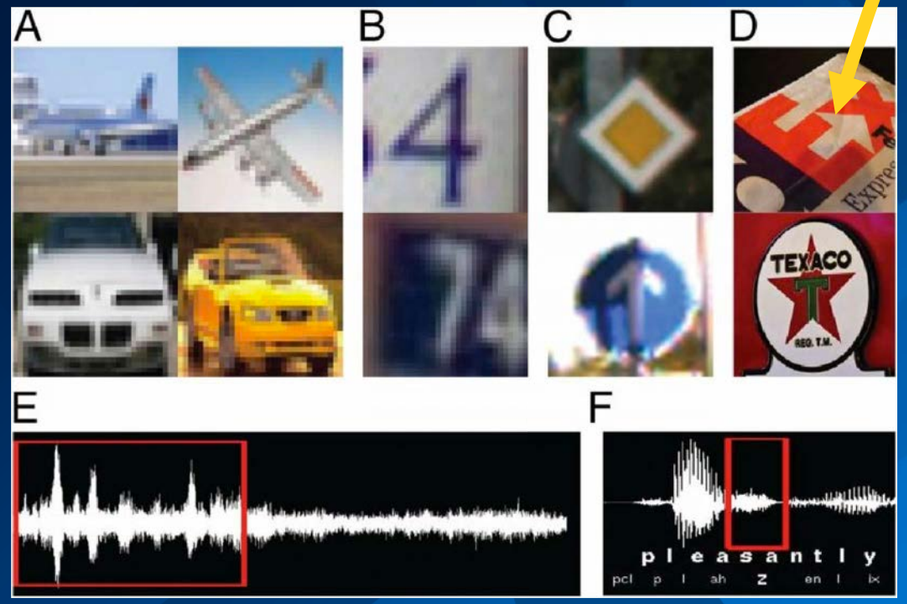
IBM

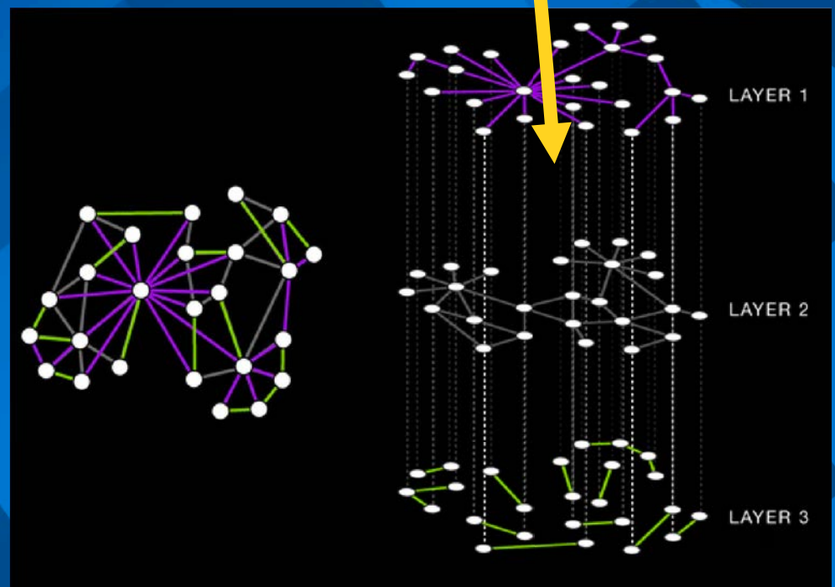
Today's research focus on “Big Data” and “Cognitive”

Plenty of new algorithms in **Graph Analytics**, **Deep Learning**, and more, for:

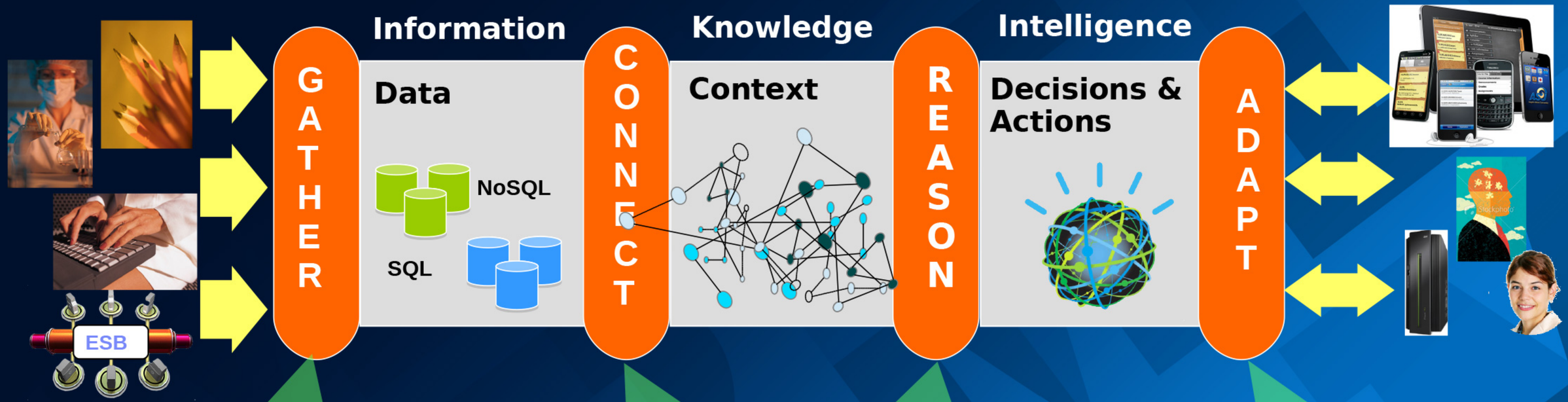
- Knowledge extraction / discovery / mining
- Image recognition (face, objects, captions generation, ...)
- NLP, speech and text (language identification, automatic translation, ...)
- Sentiment analysis (emotions, views, impact of thoughts, …)
- On demand simulations to reduce uncertainty (parameter optimization)
- … and many more (potentially unlimited!)







Data driven knowledge discovery pipeline



Gather

Collect all relevant data from a variety of sources: **Publications, RSS, APIs, DBs, Patents, etc.**

Connect

Extract features and build context using multiple **diverse data sources,** new sources added at run-time: **User Defined**

Reason

Analyze data in context to uncover hidden information and find new relationships.

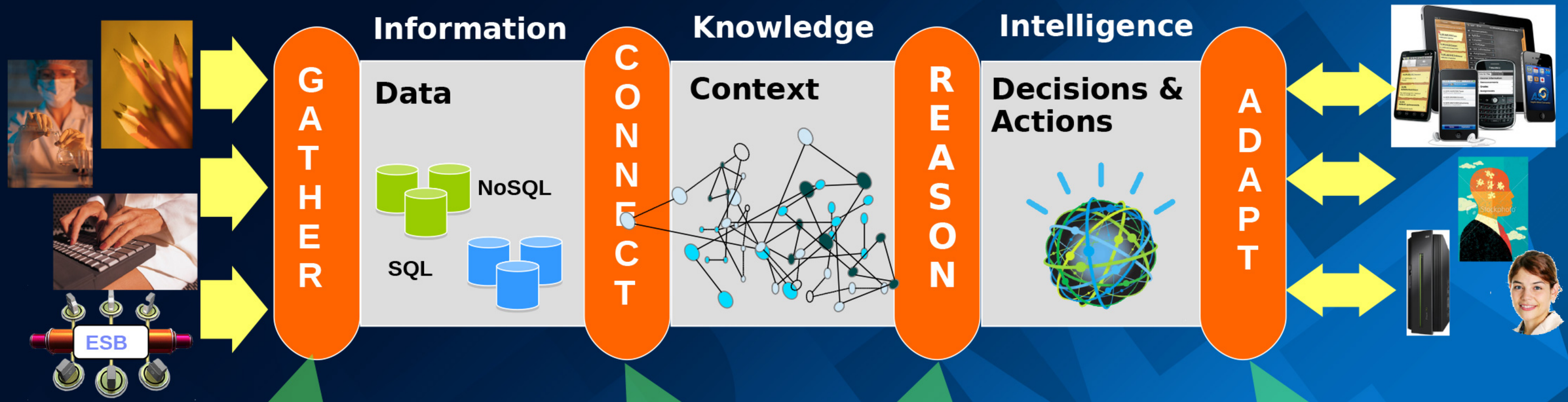
Analytics both add to context **via metadata extraction,** and use context **to broader information exploited**

Adapt

Compose recommended interactions, use context to deliver to point of action.

E.g.: Suggest material properties; Suggest simulations

Data driven knowledge discovery pipeline



Gather

Collect all relevant data from a variety of sources: **Publications, RSS, APIs, DBs, Patents, etc.**

Hardware tasks: intense Storage, Network Traffic

Connect

Extract features and build context using multiple **diverse data sources,** new sources added at run-time: **User Defined**

Hardware tasks: intense I/O, memory use/capacity

Reason

Analyze data in context to uncover hidden information and find new relationships.

Analytics both add to context **via metadata extraction,** and use context **to broader information exploited**

Hardware tasks: intense computations, flops, mem. bandwidth use, latency

Adapt

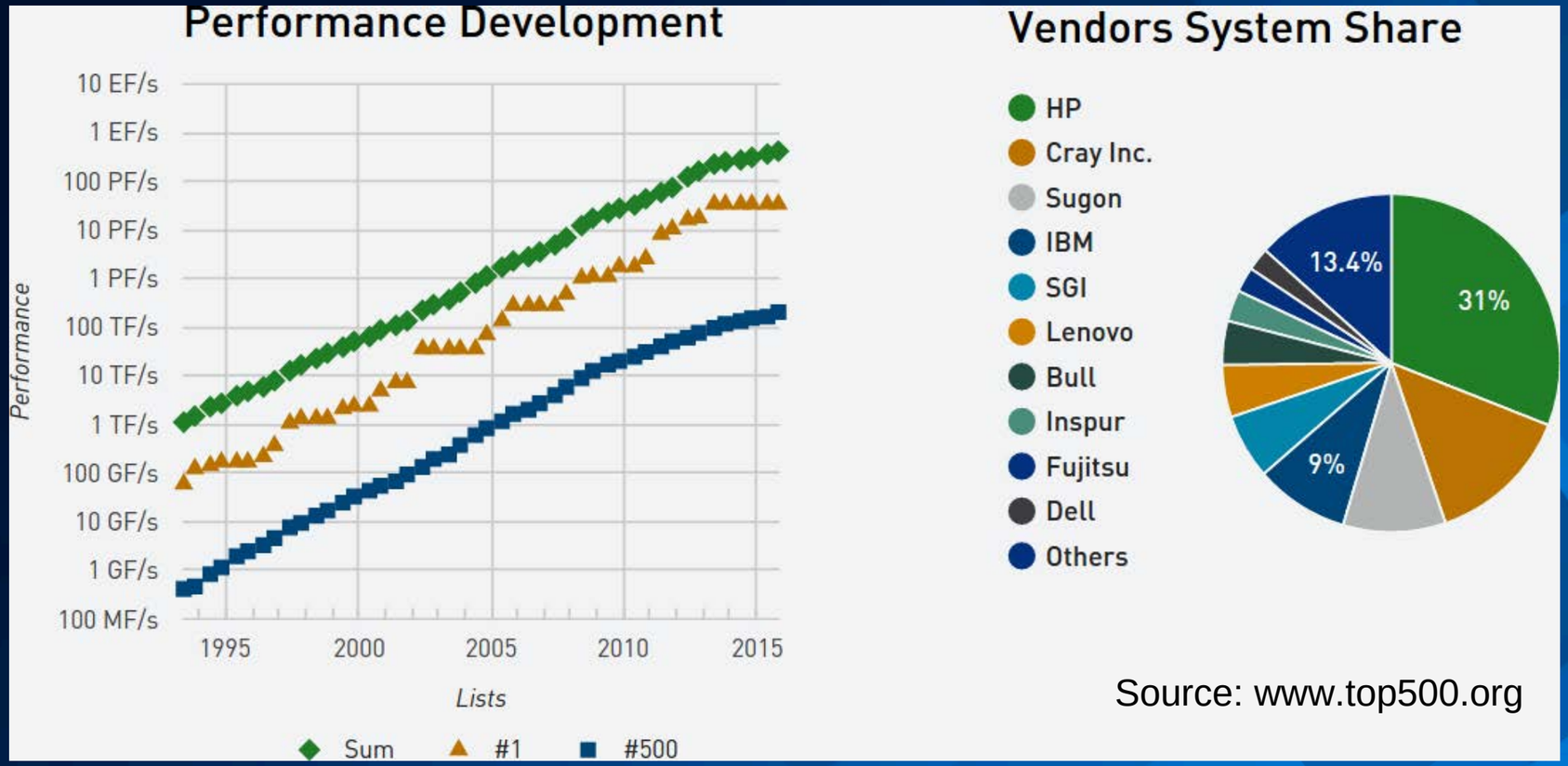
Compose recommended interactions, use context to deliver to point of action.

E.g.: Suggest material properties; Suggest simulations

Hardware tasks: intense UI, rendering

How HPC systems have been measured so far: TOP500

- Initially designed in ~1979 to provide information on execution time to solve a dense linear system
- From the '90 considered *de-facto* as the main metric to rank supercomputers



- ✔ Provide a reasonable indication of speed vs. problem size
- ✖ Emphasizes FLOPs and peak performance (do not account for network, bandwidth, energy, etc.)

How HPC systems have been measured so far: HPCG

- Introduced in 2013 to better represent today's real scientific applications

HPCG Rank	Computer	HPCG [PFlops]	Rmax [PFlops]	HPCG/HPL [%]	HPL Rank
1	Tianhe-2	0.58	33.863	1.7	2 (+1)
2	K computer	0.4608	10.51	5.3	5 (+3)
3	Sunway TaihuLight	0.37	93.02	0.4	1 (-2)
4	Sequoia	0.33	17.17	1.9	4(=)
5	Titan	0.3223	17.59	1.8	3 (-2)
6	Trinity	0.1826	8.1009	2.3	7 (+1)
7	Mira	0.167	8.587	1.9	6 (-1)
8	Pangea	0.1627	5.28	3.1	11(+3)
9	Pleiades	0.155	4.089	3.8	15(+6)
10	Hazel Hen	0.138	5.64	2.4	9 (-1)

Source: http://www.hpcg-benchmark.org
June 2016

- ✔ Emphasis not only on Flops, but also on memory bandwidth and interconnects
- ✔ Ranking provides a different picture than HPL (memory bandwidth impact)
- ■ Specific code optimizations, and problem domain (Nx-Ny-Nz) impact too much performances
- ✖ Based on one single problem, one algorithm, and one type of matrix

How HPC systems have been measured so far: GREEN500

- Introduced in 2007 to complement TOP500 and rank top supercomputers by energy efficiency

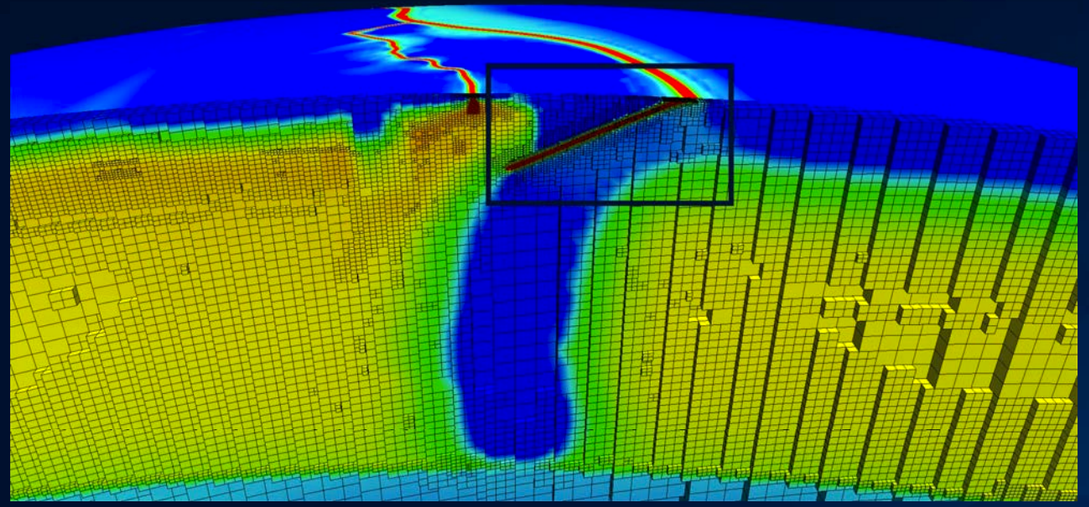
Green500 Rank	MFLOPS/W	Site	System	Total Power(kW)
1	6673.8	Advanced Center for Computing and Communication, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	150.0
2	6195.2	Computational Astrophysics Laboratory, RIKEN	ZettaScaler-1.6, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband FDR, PEZY-SCnp	46.9
3	6051.3	National Supercomputing Center in Wuxi	Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway	15371
4	5272.1	GSI Helmholtz Center	ASUS ESC4000 FDR/G2S, Intel Xeon E5-2690v2 10C 3GHz, Infiniband FDR, AMD FirePro S9150	57.2
5	4778.5	Institute of Modern Physics (IMP), Chinese Academy of Sciences	Sugon Cluster W780I, Xeon E5-2640v3 8C 2.6GHz, Infiniband QDR, NVIDIA Tesla K80	65
6	4112.1	Stanford Research Computing Center	Cray CS-Storm, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, Nvidia K80	190
7	3775.5	Internet Service (B)	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	110
8	3775.5	Internet Service (B)	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	110
9	3775.5	Internet Service (B)	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	110
10	3775.5	Internet Service (B)	Inspur TS10000 HPC Server, Intel Xeon E5-2620v2 6C 2.1GHz, 10G Ethernet, NVIDIA Tesla K40	110

Only one large system in first 10 positions!

Source: www.green500.org (June 2016)

- ✔ Increased energy awareness
- ✖ Do not promote large supercomputers (do not account for problem size and scalability)
- ✖ Measure energy per Flop and not energy-to-solution

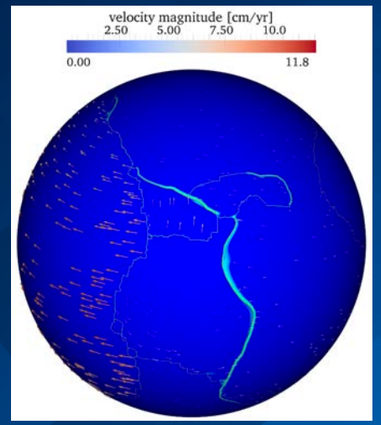
How HPC system should be measured: scalability in real problems

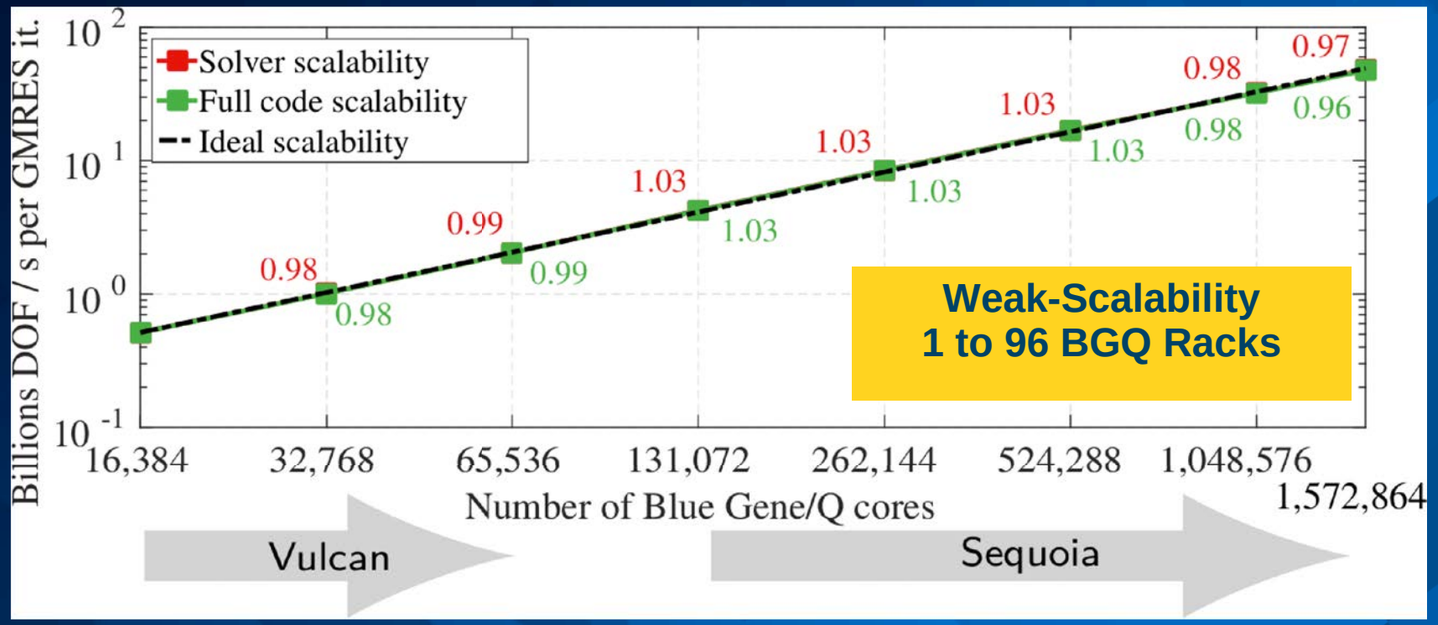


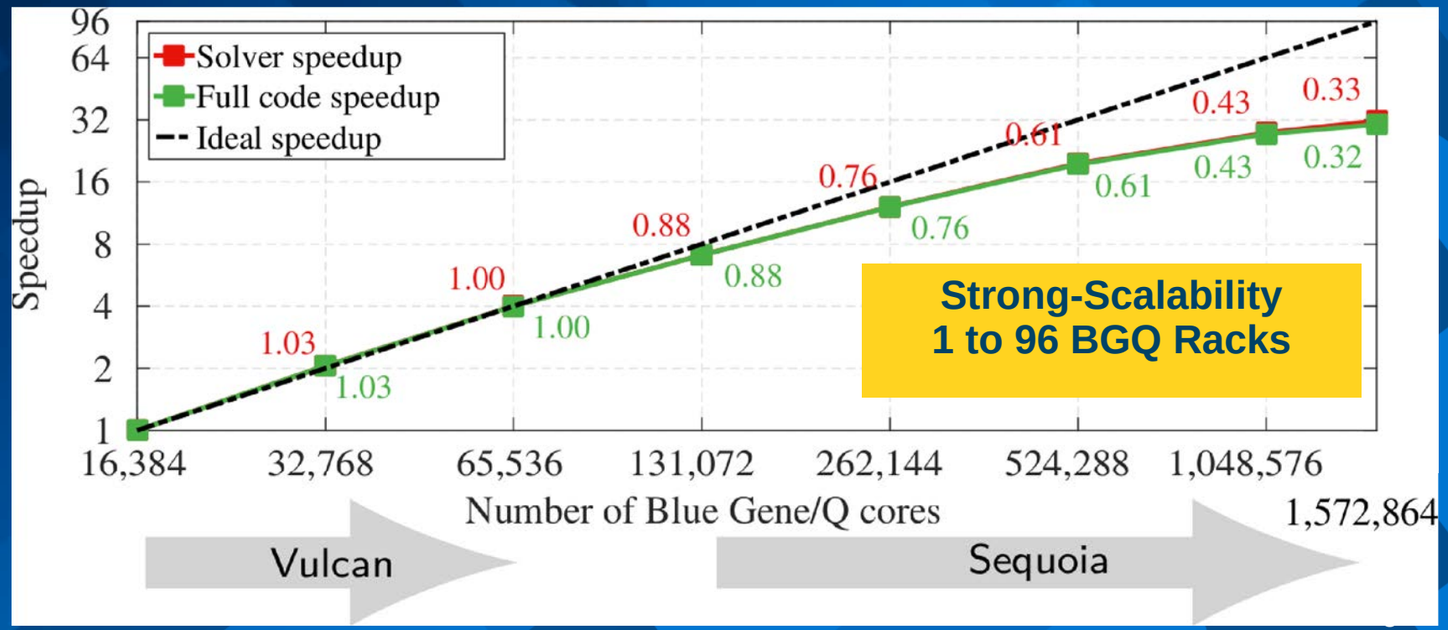
An Extreme-scale Implicit Solver for Complex PDEs: Highly Heterogeneous Flow in Earth's Mantle

J. Rudi[1], A.C.I. Malossi[2], T. Isaac[1], G. Stadler[3], M. Gurnis[4], P.W.J. Staar[2], Y. Ineichen[2], C. Bekas[2], A. Curioni[2], O. Ghattas[1]

1: The University of Texas at Austin
2: IBM Research – Zurich
3: New York University
4: California Institute of Technology



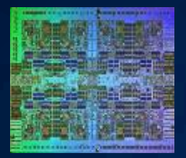
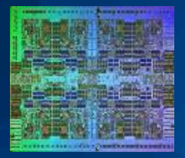
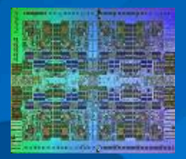
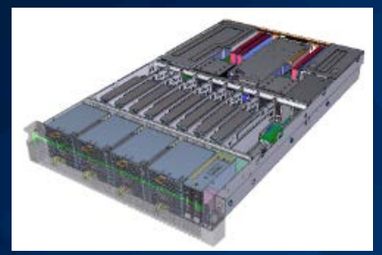
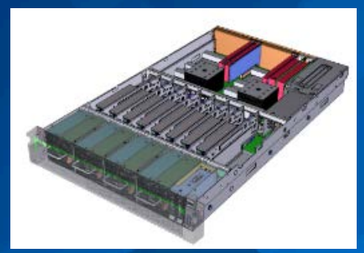
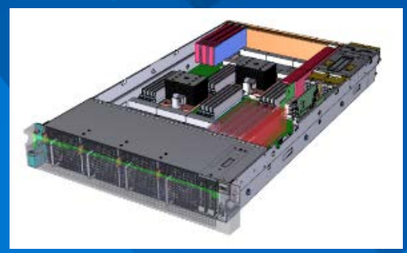




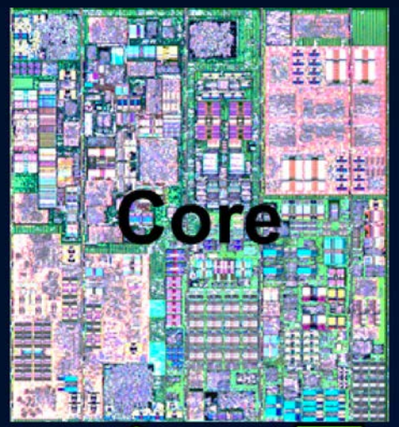
OpenPOWER Foundations: 200+ members and growing

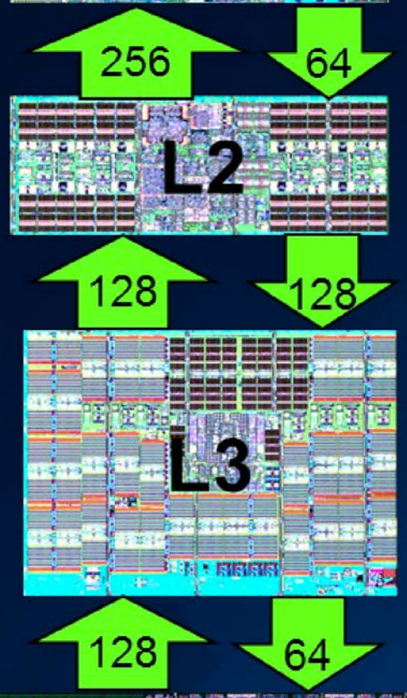
- Members can **customize POWER CPU processors and platforms** for their business needs
- Innovations include **custom systems** for data centers, **workload acceleration** through GPU, FPGA or advanced I/O, **platform optimization** for SW appliances, or **advanced hardware technology** exploitation

OpenPOWER Roadmap

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 (Firestone)**  **OpenPower CAPI Interface**	**POWER8 (Minsky)**  **Acceleration: NVLink 1.0, CAPI 1.0, PCIe Gen3**	**POWER9 (Witherspoon)**  **Acceleration: CAPI 2.0, NVLink 2.0, BlueLink, PCIe Gen4**
IBM Nodes	**2015** 	**2016** 	**2017** 

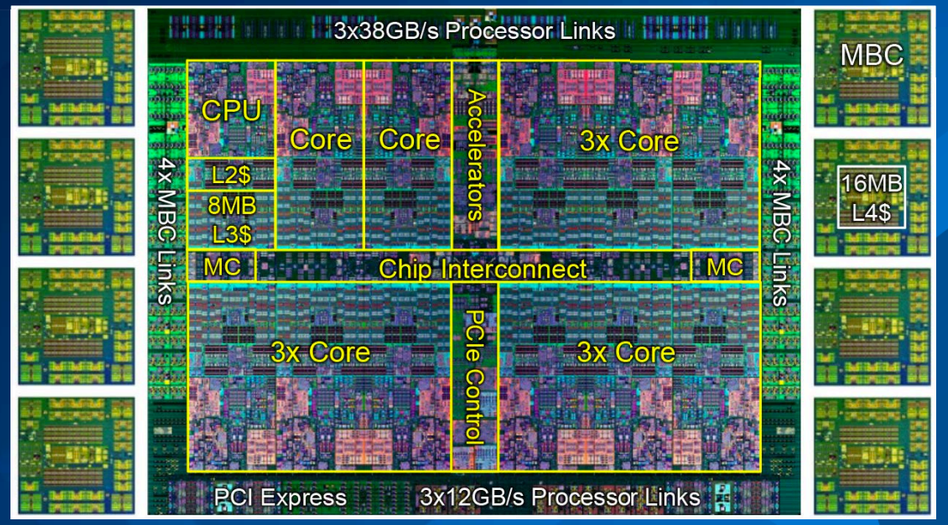
Minsky: CPU and cache/memory overview





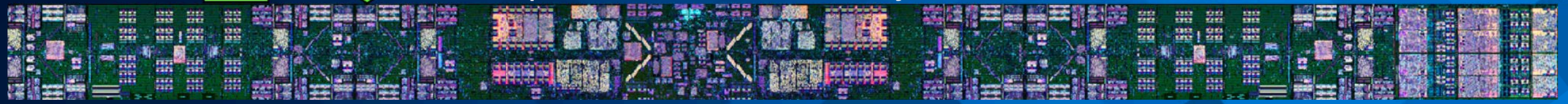
CPU

- 22nm SOI Technology
- Up to 4GHz
- Up to 12 cores (SMT8)
- 8 dispatch, 10 issue, 16 exec pipe
- 2 FXU, 2 LSU, 2 LU, 4 FPU, 2 VMX, 1 Crypto, 1 DFU, 1 CR, 1 BR
- 64K data cache, 32K instruction cache

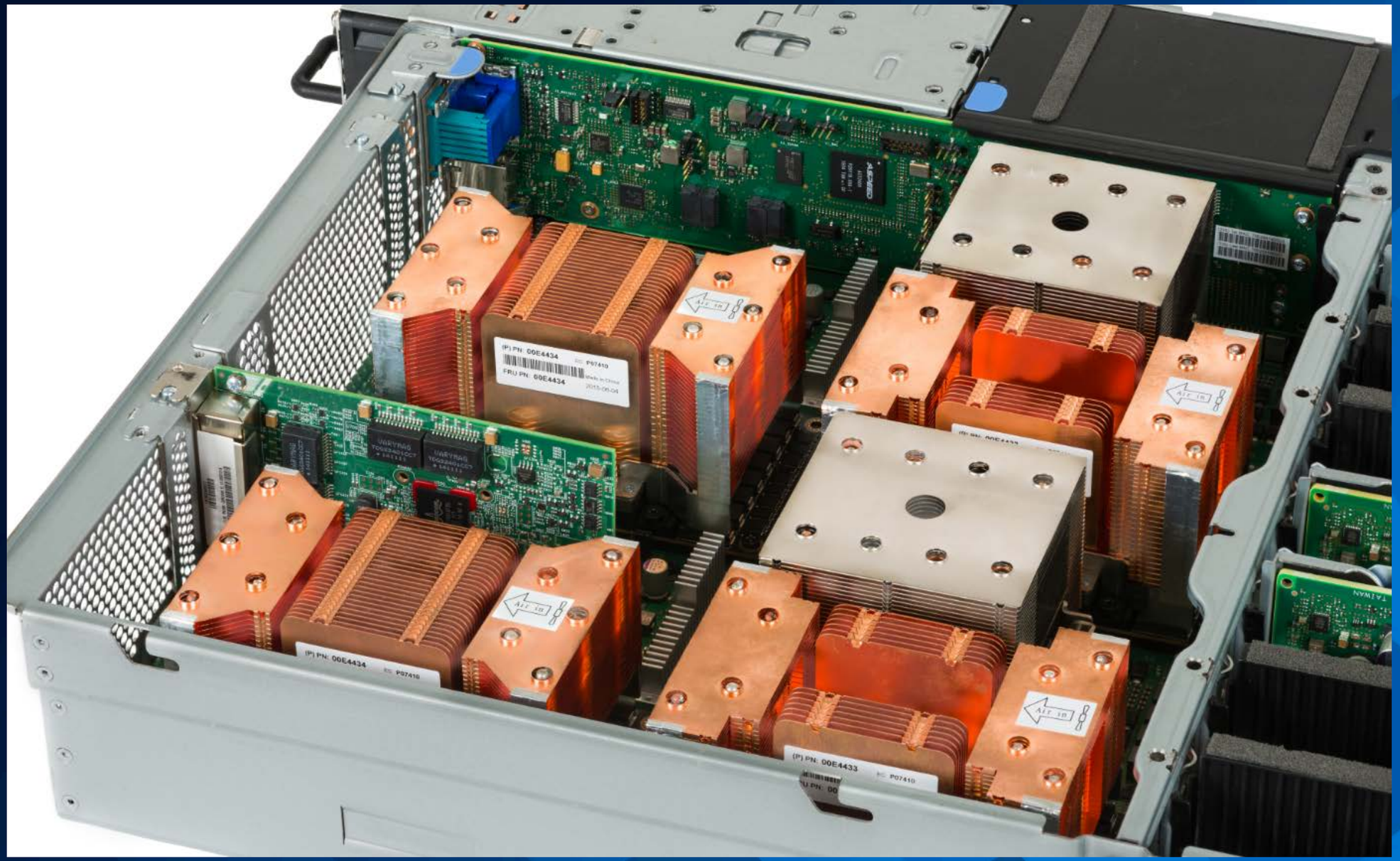


Cache hierarchy

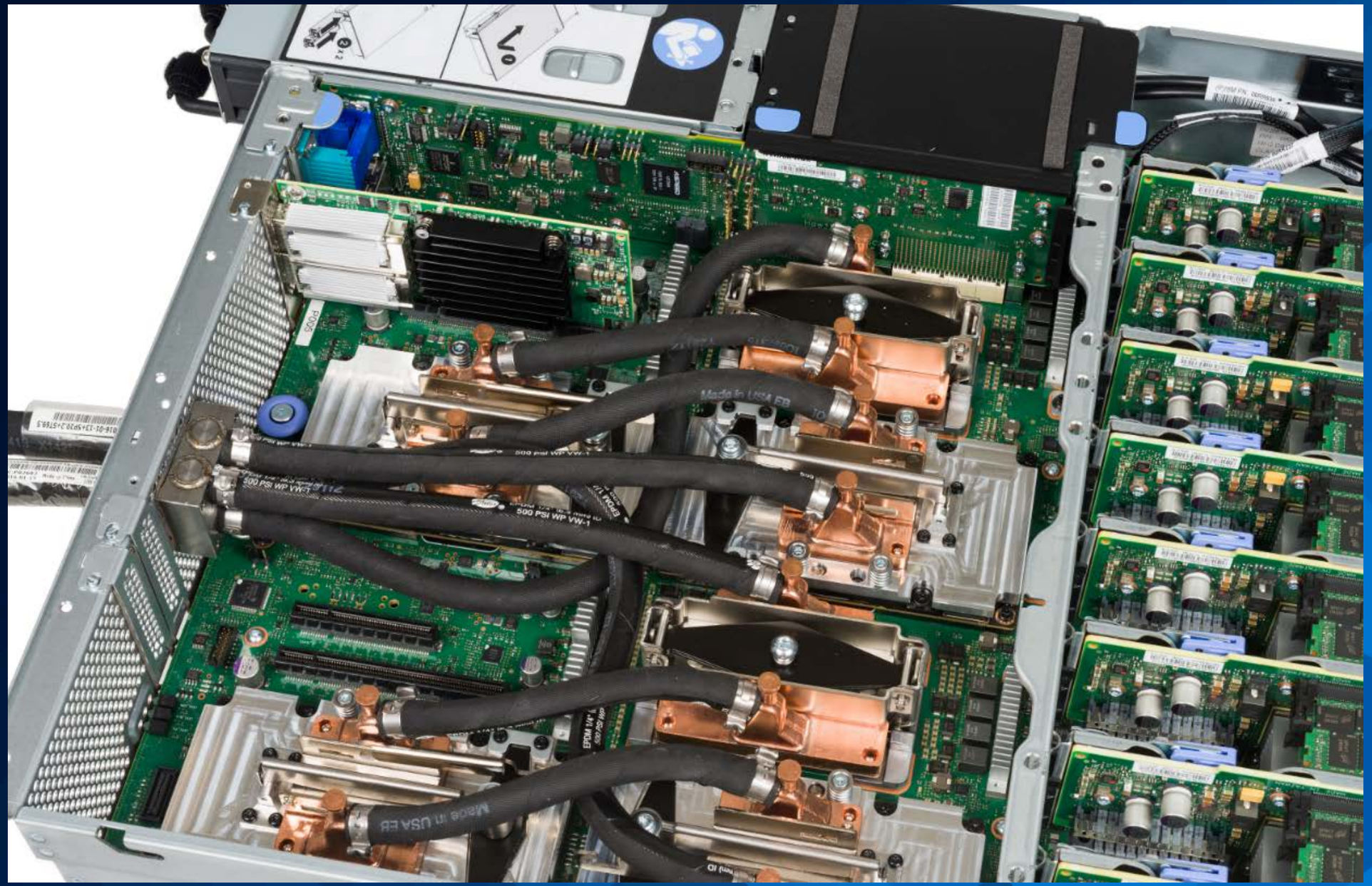
- L2: 512 kb SRAM 8 way per core
- L3: 96 MB (12 x 8 MB eDRAM 8 way Bank)
- L4: 128 MB eDRAM (on Centaur)
- “NUCA” Cache policy (Non-Uniform)
- Cache bandwidth:
  - Up to 4 TB/sec L2 Bandwidth
  - Up to 3 TB/sec L3 Bandwidth
  - Up to 230 GB/sec Memory Bandwidth



Minsky: a view on the four Pascal GPUs (air-cooled version)

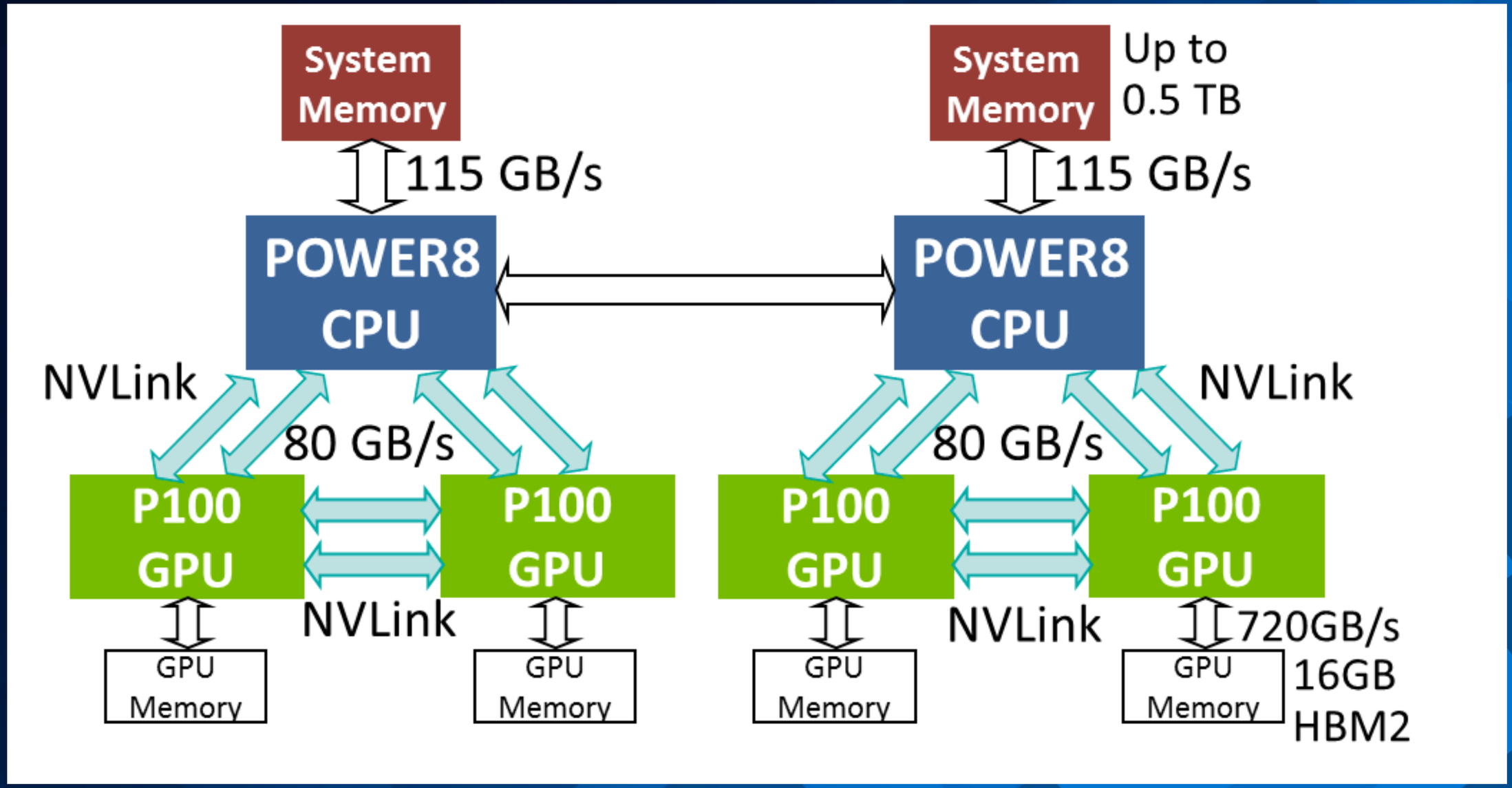


Minsky: a view on the four Pascal GPUs (water-cooled version)



Minsky: NVLink bandwidth 5x faster than PCIe

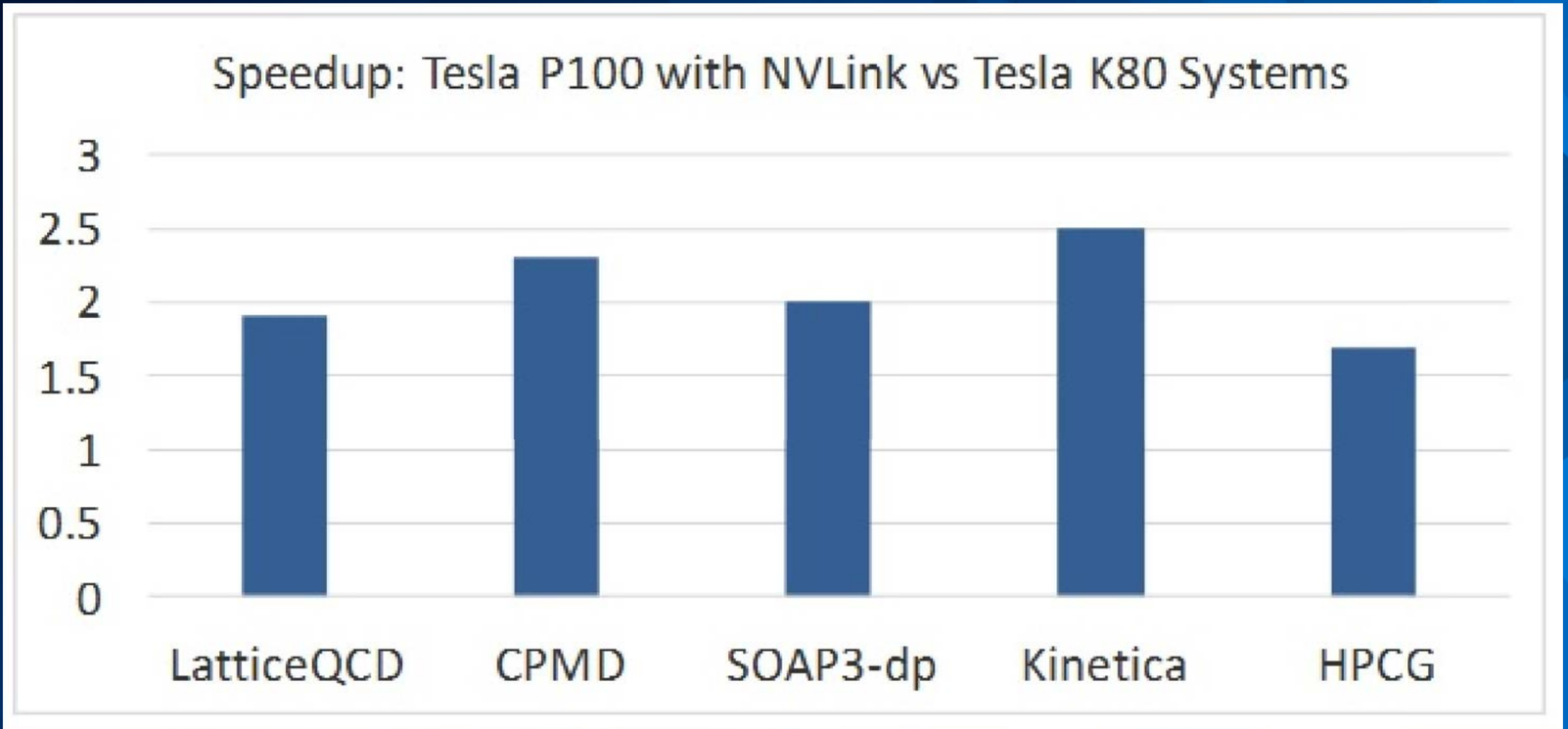
- The NVLink ports delivers 80 GB/s (40 GB/s bidirectional) point-to-point between GPU-CPU and GPU-GPU



Minsky: early performance benchmarks

Speedup with the new S822LC (Minsky – P100 + NVLink 1.0) vs. Firestone (K80 – PCIe/No NVLink):

- 2x increase for LatticeQCD, a quantum chromodynamics application for computational physics
- 2.25x increase for CPMD, a computational chemistry application
- 2x increase for SOAP3-dp, a bio-informatics (genomics) application
- 2.4x increase for Kinetica, an in-memory, relational database
- 1.75x increase for HPCG, a high performance computing benchmark



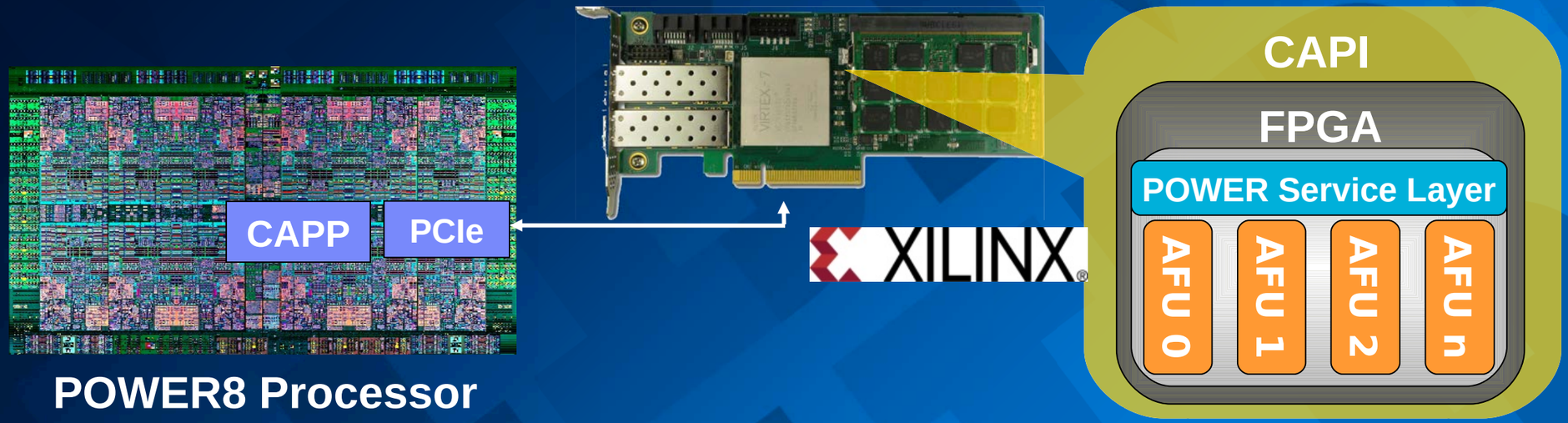
Minsky: coherent accelerator processor interface (CAPI)

Standard I/O Model Flow ~ 15 µs for data preparation

DD Call → Copy/Pin → MMIO Notify → Accelerate → Poll / Int → Copy/Unpin → Return DD

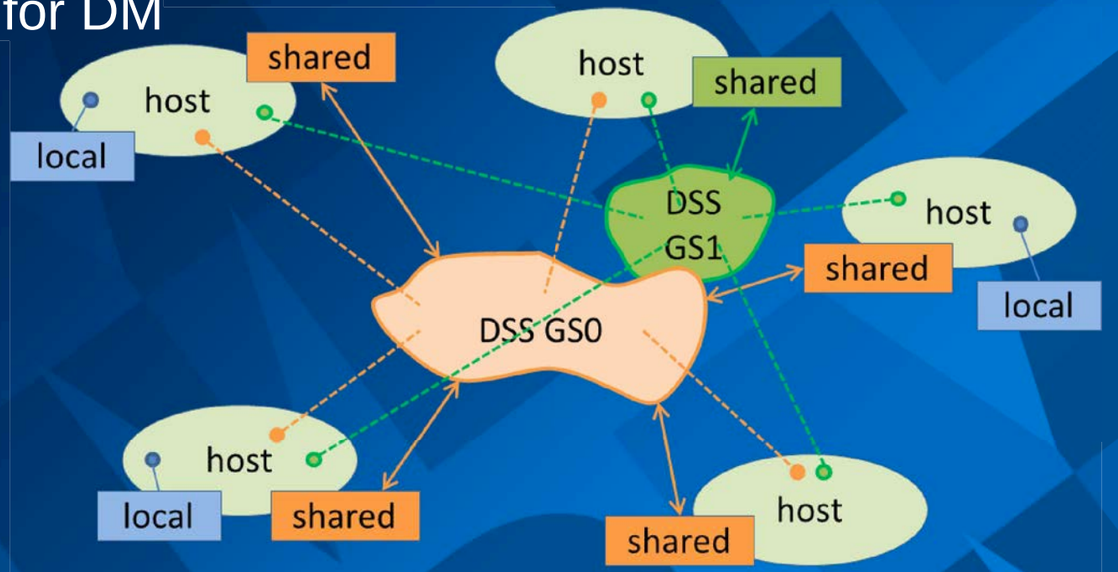
Flow with a Coherent Model ~ 0.5 µs for data preparation

Shared Mem. Notify Accelerator → Accelerate → Shared Memory Completion



Minsky: dense storage memory (research work)

- Prototype of **Dense Memory Integration Software Stack**
  - byte addressable, distributed globally accessable DM resource
  - exports industry standard asynchronous RDMA API for DM read and write access
- Software components
  - kernel module for local NVMe access
  - kernel module for storage abstraction and global sharing
  - kernel module for OpenFabrics RDMA stack integration
  - user library for OpenFabrics compatible user interface
  - user level demon process for global/inter-node resource management
- Implements efficient local and remote DM access
  - zero copy local access via direct DMA device - application buffer
  - zero copy remote access via IB RDMA remote host - application buffer
- First performance results
  - **local DM access** at NVMe devices performance limits (**3.5 GB/s read, 1.8 GB/s write** of 4k buffers)
  - **remote DM access** at network (100Gb/s InfiniBand) and device limits: **12.5 GB/s distributed DM random read** with 4 storage nodes, all equipped with one NVMe SSD each
  - close to 900k IOPs for single device short sequential red/write operations



Looking ahead: POWER9

New Core Microarchitecture

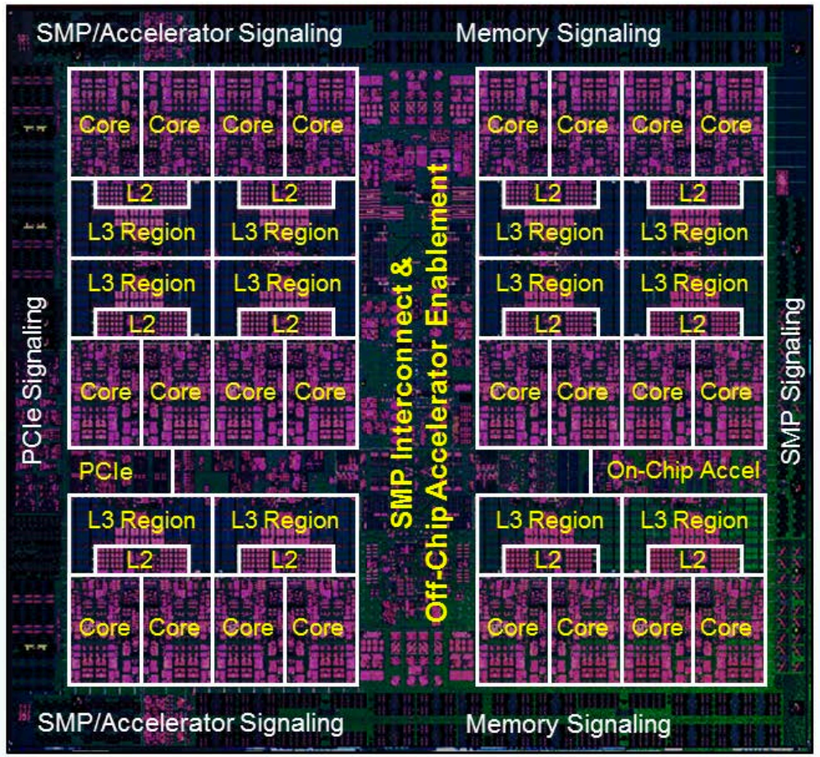
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy

- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation

- Quality of service assists
- New interrupt architecture
- Workload optimized frequency
- Hardware enforced trusted execution



14nm finFET Semiconductor Process

- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors

Leadership Hardware Acceleration Platform

- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth, advanced new features
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- New CAPI: Improved latency and bandwidth, open interface

State of the Art I/O Subsystem

- PCIe Gen4 – 48 lanes

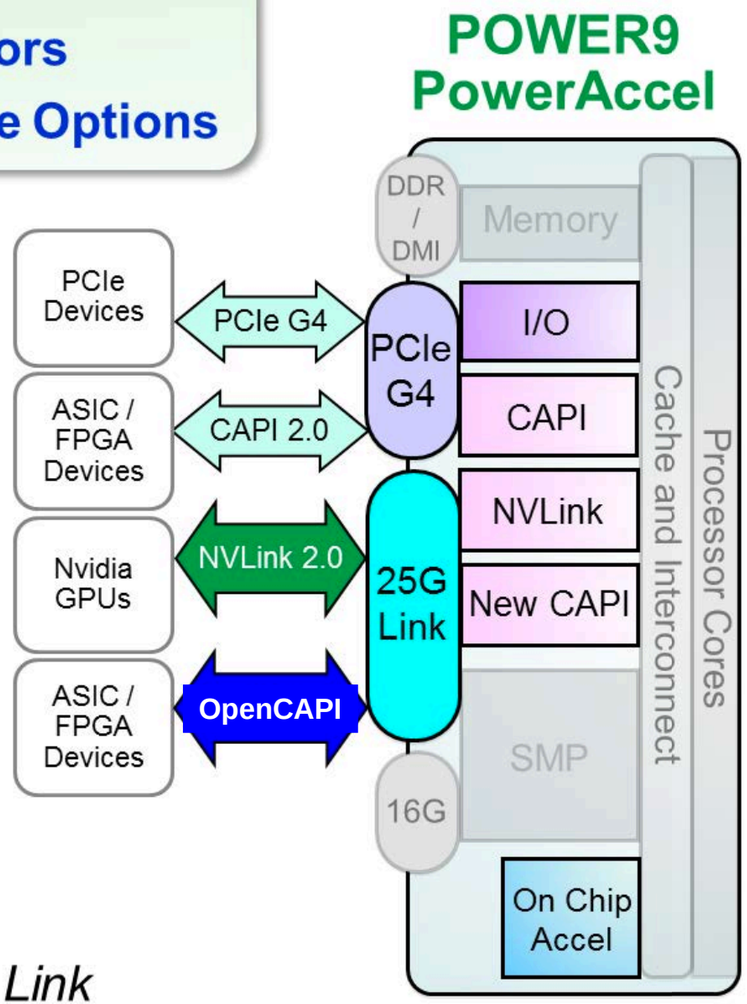
High Bandwidth Signaling Technology

- 16 Gb/s interface
  - Local SMP
- 25 Gb/s interface – 25G Link
  - Accelerator, remote SMP

POWER9 accelerator interfaces

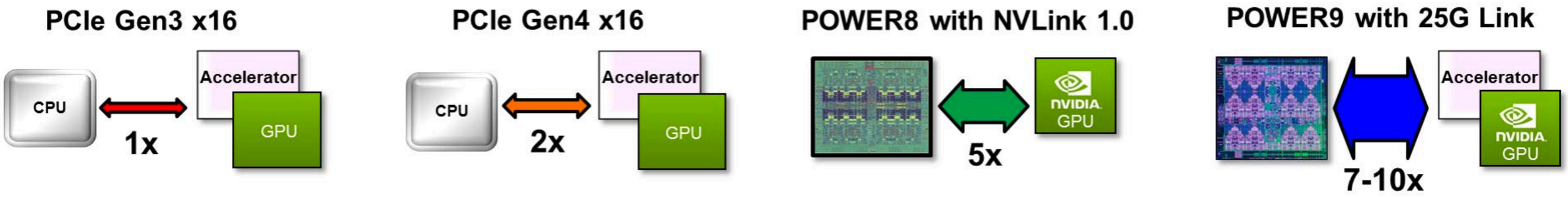
- **Extreme Processor / Accelerator Bandwidth and Reduced Latency**
- **Coherent Memory and Virtual Addressing Capability for all Accelerators**
- **OpenPOWER Community Enablement – Robust Accelerated Compute Options**

- **State of the Art I/O and Acceleration Attachment Signaling**
  - **PCIe Gen 4** x 48 lanes – 192 GB/s duplex bandwidth
  - **25G Link** x 48 lanes – 300 GB/s duplex bandwidth
- **Robust Accelerated Compute Options with OPEN standards**
  - **On-Chip Acceleration** – Gzip x1, 842 Compression x2, AES/SHA x2
  - **CAPI 2.0** – 4x bandwidth of POWER8 using *PCIe Gen 4*
  - **NVLink 2.0** – Next generation of GPU/CPU bandwidth and integration
  - **OpenCAPI** – High bandwidth, low latency and open interface using *25G Link*



POWER9 accelerator speedup

Extreme CPU/Accelerator Bandwidth



Increased Performance / Features / Acceleration Opportunity

Seamless CPU/Accelerator Interaction

- Coherent memory sharing
- Enhanced virtual address translation
- Data interaction with reduced SW & HW overhead

Broader Application of Heterogeneous Compute

- Designed for efficient programming models
- Accelerate complex analytic / cognitive applications

References

- J. Rudi, A. C. I. Malossi, T. Isaac, G. Stadler, M. Gurnis, P. W. J. Staar, Y. Ineichen, C. Bekas, A. Curioni, O. Ghattas. An Extreme-scale Implicit Solver for Complex PDEs: Highly Heterogeneous Flow in Earth's Mantle. SC'15, ACM, 2015. – **ACM Gordon Bell Prize Winner 2015**
- A. C. I. Malossi, Y. Ineichen, C. Bekas, A. Curioni. Fast Exponential Computation on SIMD Architectures. HiPEAC 2015 - 1st Workshop On Approximate Computing (WAPCO), 2015
- P. Klavik, A. C. I. Malossi, C. Bekas, A. Curioni. Changing computing paradigms towards power efficiency. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 372(2018), The Royal Society, 2014
- P. W. J. Staar, P. Barkoutsos, R. Istrate, A. C. I. Malossi, I. Tavernelli, N. Moll, H. Giefers, C. Hagleitner, C. Bekas, A. Curioni. Stochastic Matrix-Function Estimators Scalable Big-Data Kernels with High Performance, IPDPS, 2016. – **IPDPS Best Papers Session 2016**
- F. E. Faisal, Y. Ineichen, A. C. I. Malossi, P. Staar, C. Bekas and A. Curioni. Massively Parallel and Near Linear Time Graph Analytics, SC'14, Poster, 2014

THANK YOU!

POWER9: more details on OpenCAPI