Best Practices: Accelerated System Configuration

John E. Stone

Theoretical and Computational Biophysics Group

Beckman Institute for Advanced Science and Technology

University of Illinois at Urbana-Champaign

http://www.ks.uiuc.edu/Research/gpu/

Workshop on Accelerated High-Performance Computing in Computational Sciences (SMR 2760),

International Centre for Theoretical Physics (ICTP),

Trieste, Italy, June 2, 2015

Overview

- Memory coherency, NUMA, futures, die stacked memories, deeper hierarchies
- Node configuration, PCIe speed, topology, Hwloc, ‘lstopo’, BIOS flash updates
- Memtest86, cpuburn, GPU diag tools
- Use of nvidia-smi, persistent mode, ECC, …
- Power supplies, GeForce vs. Tesla
- Infiniband+MPI that support RDMA w/ GPUs, pinned memory, …

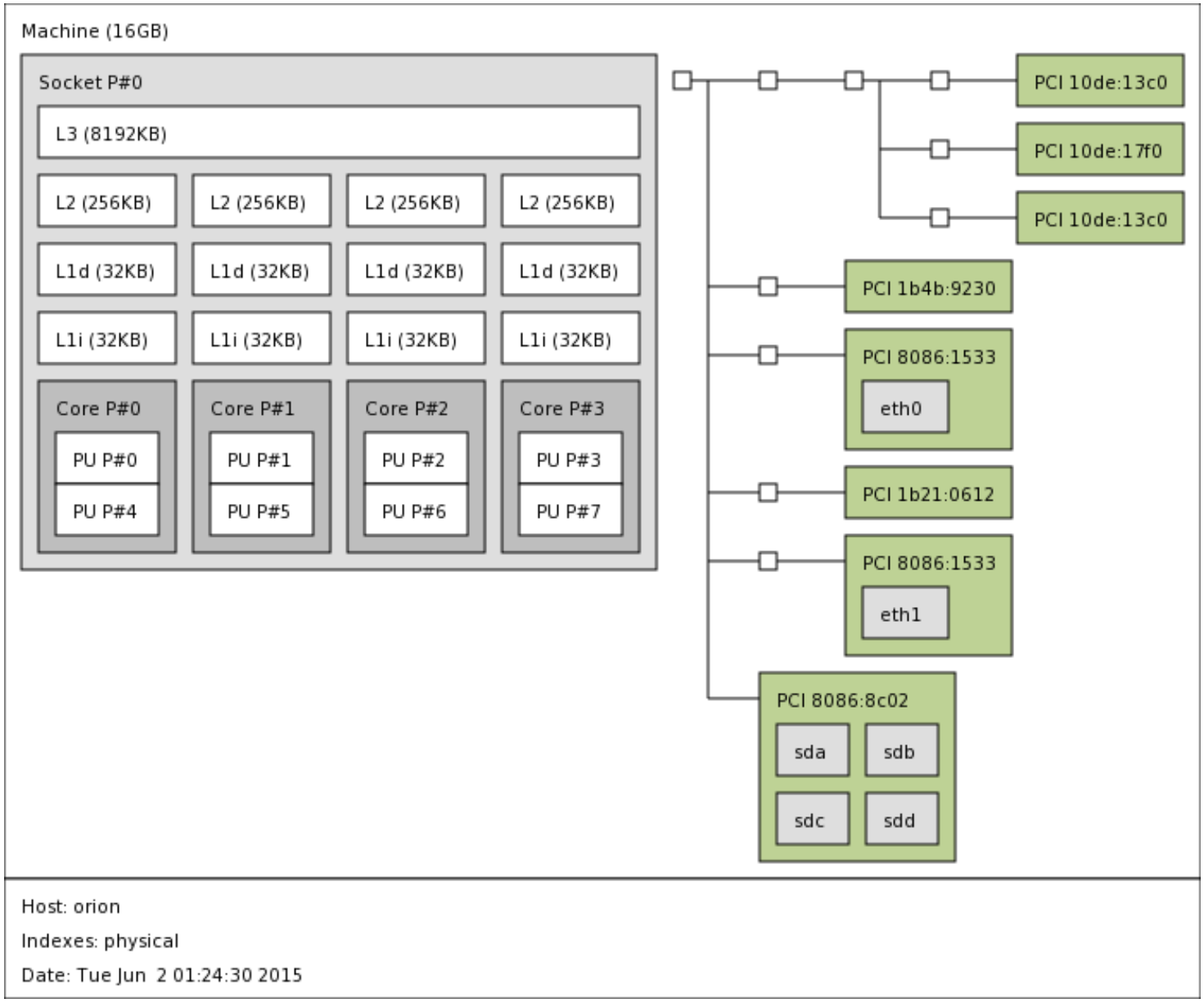
Memory Coherency (Oversimplified)

- Memory writes by CPU cores that share the same L1/L2 caches are visible to peers
- Memory writes by one CPU socket are (eventually) visible to all other peer CPUs in a multi-socket system
- Cache hardware permits multiple cores to read without penalties
- Write conflicts are solved by mutual exclusion locks, memory barrier instructions, atomic CAS and other ops

Issues With Memory Coherency

- Hardware that supports coherency is MUCH more complex than hardware that does not
- Coherency is difficult to achieve between hardware devices that are not intimately tied together in the same memory system
- Compromise solutions (e.g. GPUs) gain speed, energy efficiency, lower cost by giving up some coherency
- Future CPU/GPU hardware will likely make new compromises so that performance and efficiency keep increasing, albeit with more complexity for programmers
- Future memory systems will have more complicated and deeper memory hierarchies, and bigger performance costs when not used correctly…

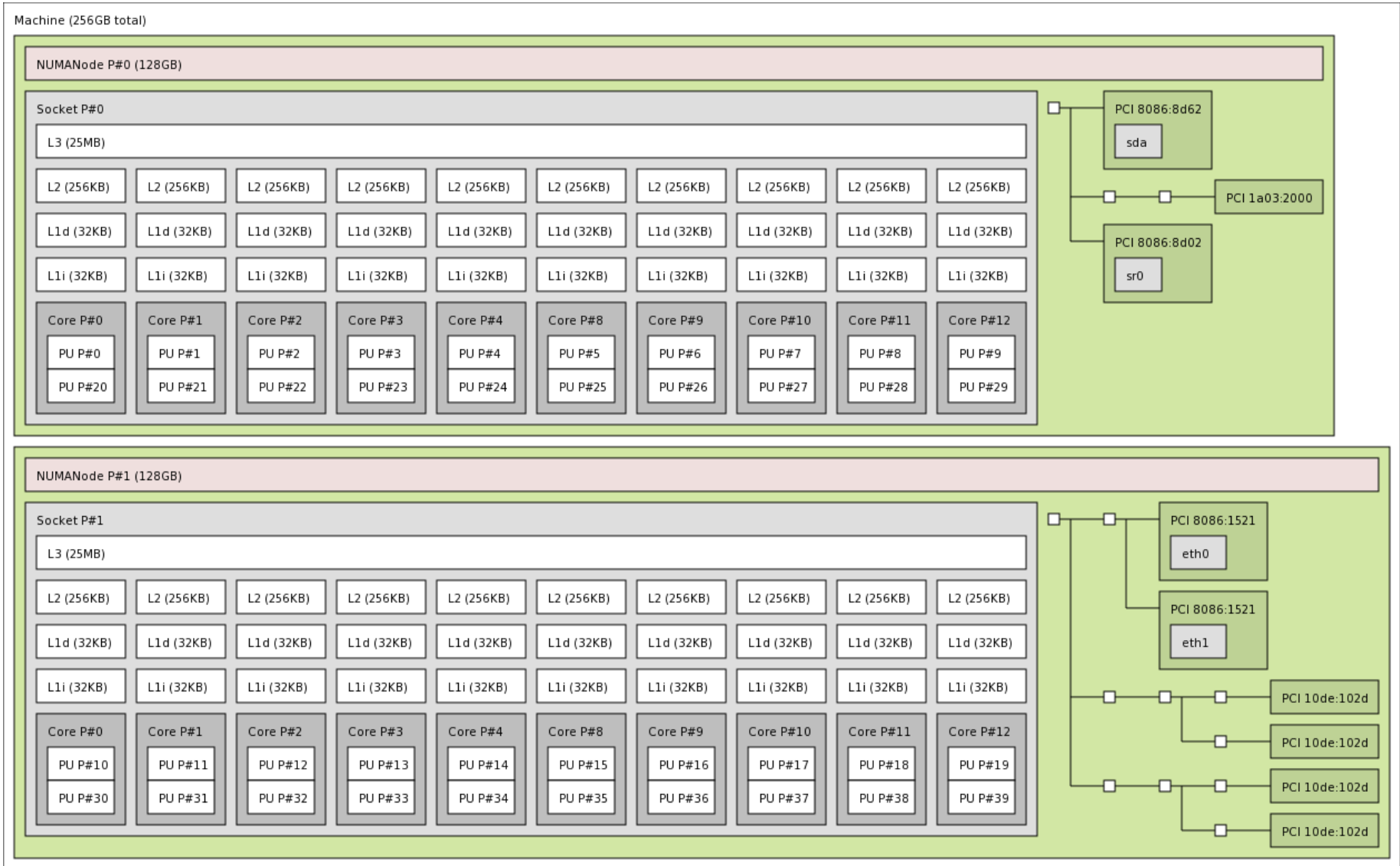
Hwloc ‘lstopo’: Intel Haswell 4771, 3 GPUs, No NUMA



NUMA

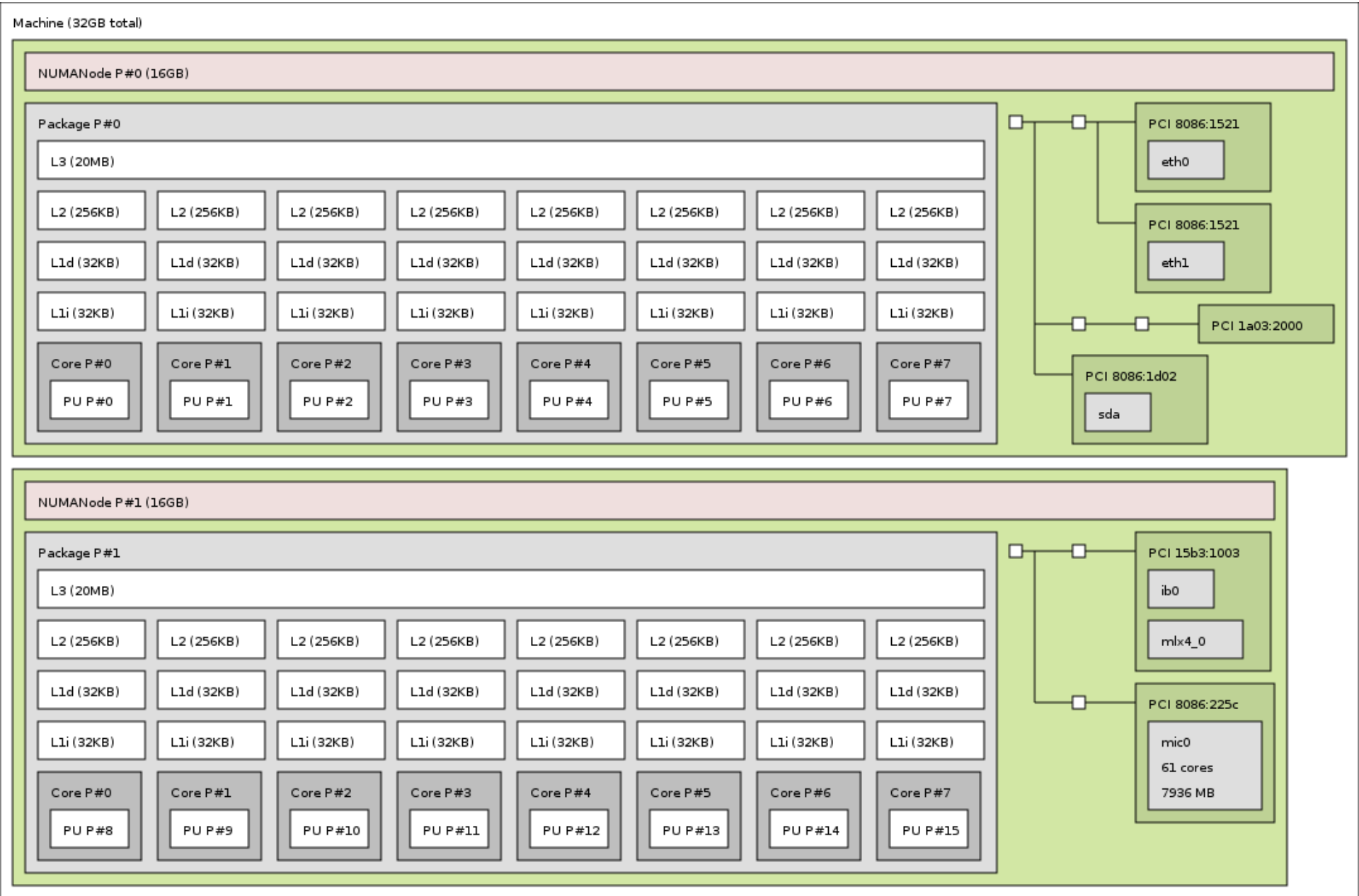
- NUMA == Non-Uniform Memory Access
- CPUs directly incorporate memory controllers
- Memory system is not flat, it is instead implemented as a (small) network
- Each CPU socket is associated with its own memory
- Access to memory on a peer CPU socket has much higher latency and lower bandwidth than socket-local memory, must traverse CPU bus topology (QPI or HT)

Hwloc ‘lstopo’: 2x Xeon E5, 2x Tesla K80



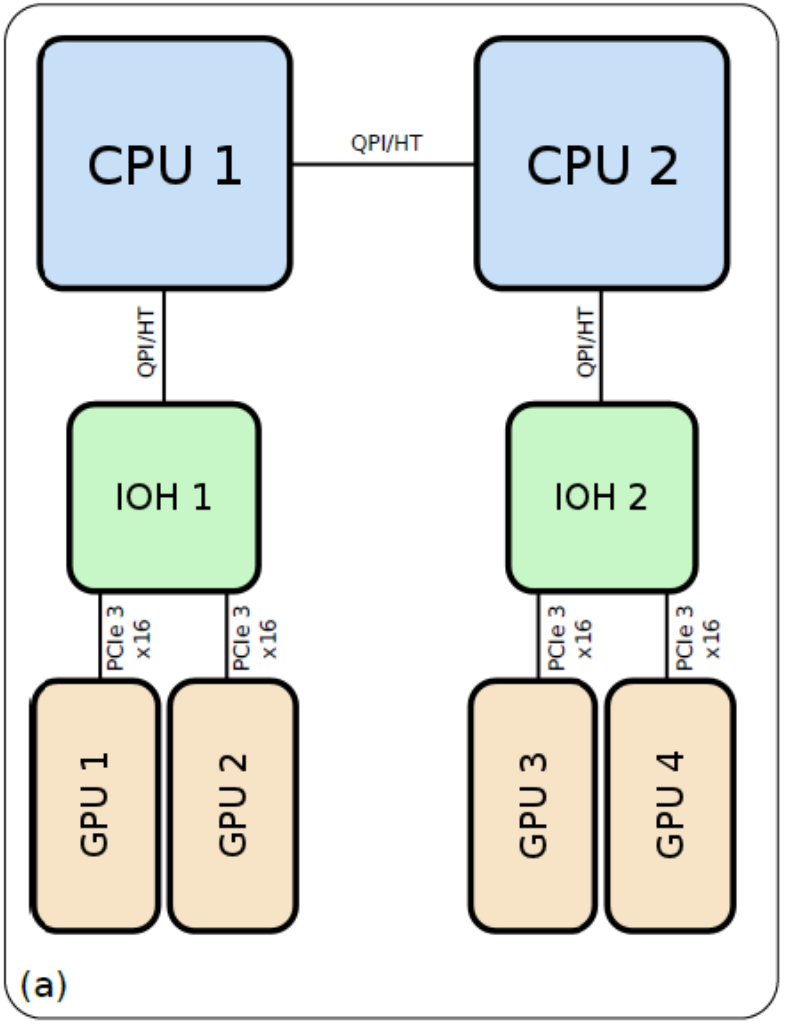
Hwloc ‘lstopo’: 2x Xeon E5, 2x Xeon Phi, 2 GPUs

Hwloc ‘lstopo’: 2x Xeon E5 + IB + Xeon Phi



Multi-GPU NUMA Architectures:

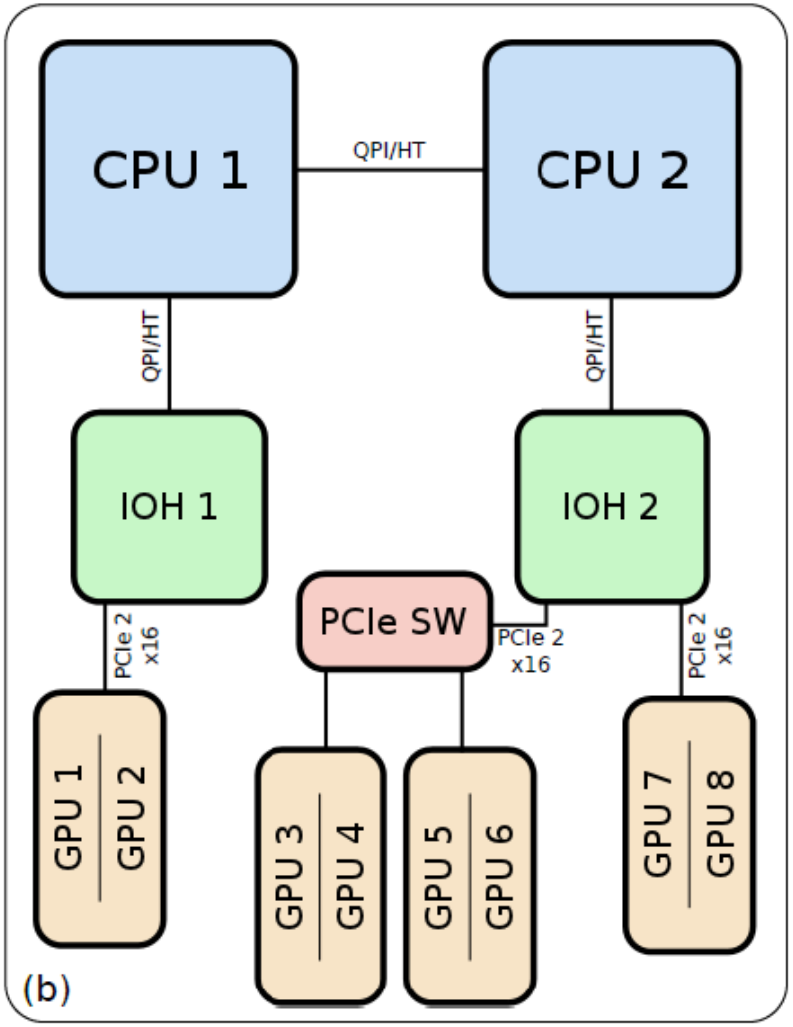
- Example of a “balanced” PCIe topology
- NUMA: Host threads should be pinned to the CPU that is “closest” to their target GPU
- GPUs on the same PCIe I/O Hub (IOH) can use CUDA peer-to-peer transfer APIs
- Intel: GPUs on different IOHs can’t use peer-to-peer



Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations
Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten.
Journal of Parallel Computing, 40:86-99, 2014.
http://dx.doi.org/10.1016/j.parco.2014.03.009

Multi-GPU NUMA Architectures:

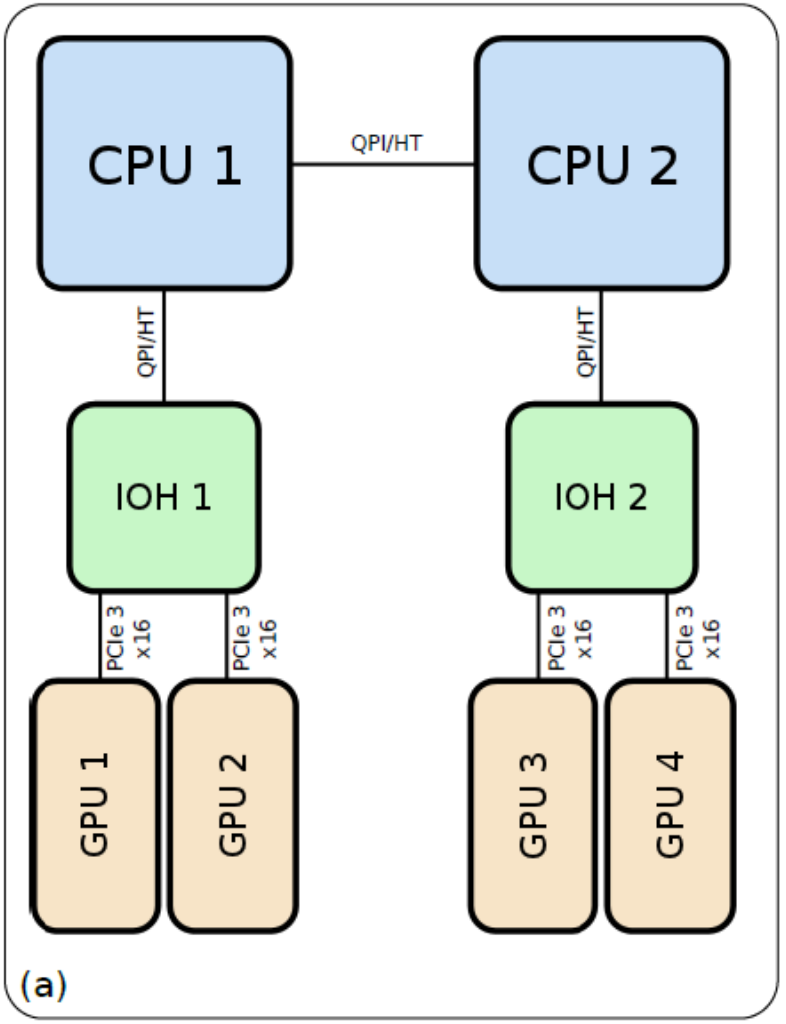
- Example of a very "unbalanced" PCIe topology
- CPU 2 will overhelm its QP/HT link with host-GPU DMAs
- Poor scalability as compared to a balanced PCIe topology



Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations
Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten.
Journal of Parallel Computing, 40:86-99, 2014.
http://dx.doi.org/10.1016/j.parco.2014.03.009

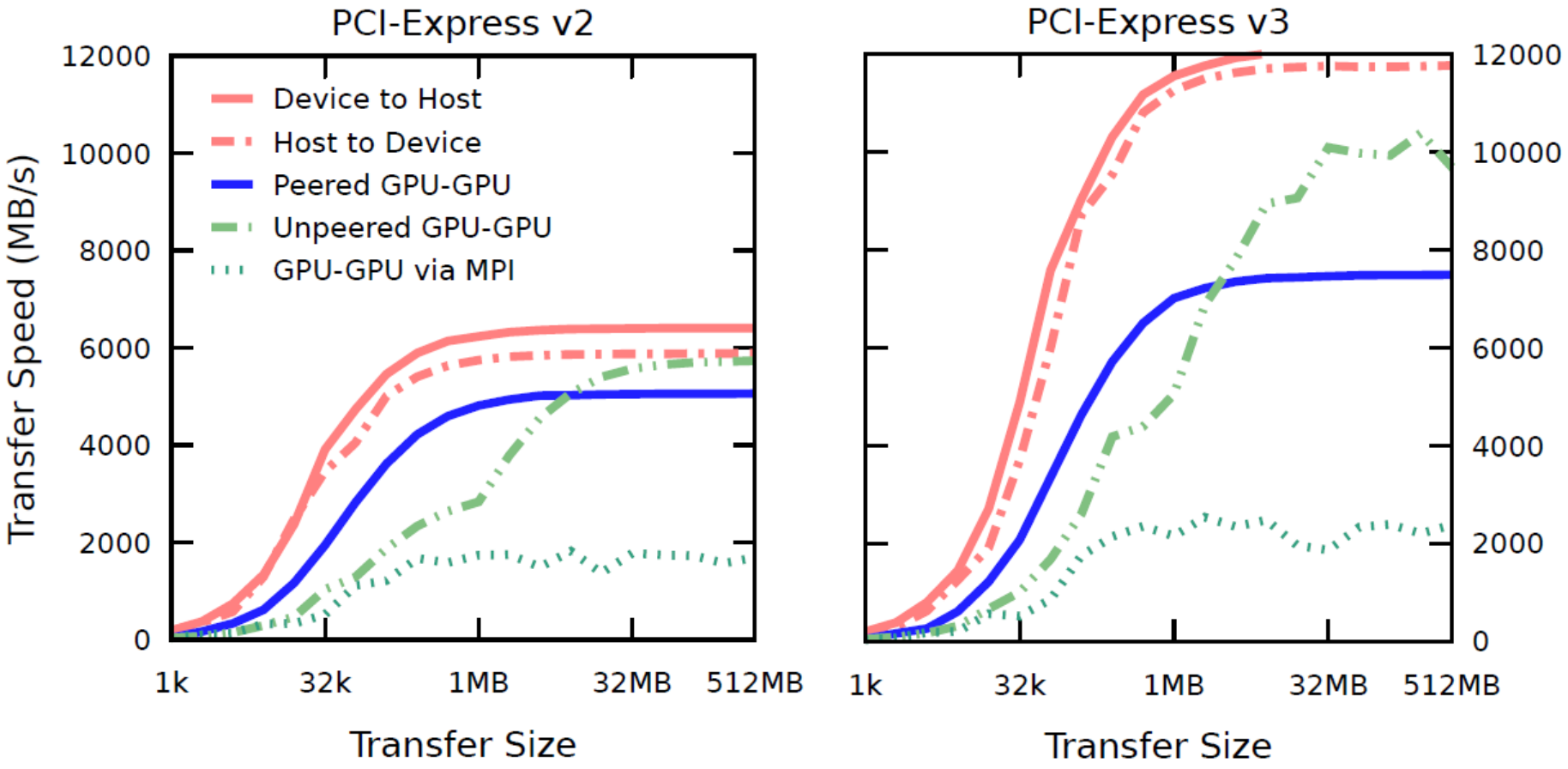
Multi-GPU NUMA Architectures:

- GPU-to-GPU peer DMA operations are much more performant than other approaches, particularly for moderate sized transfers
- Likely to perform even better in future multi-GPU cards with direct GPU links, e.g. announced “NVLink”



Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations
Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten.
Journal of Parallel Computing, 40:86-99, 2014.
http://dx.doi.org/10.1016/j.parco.2014.03.009

GPU PCI-Express DMA



Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations
Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten.
Journal of Parallel Computing, 40:86-99, 2014.
http://dx.doi.org/10.1016/j.parco.2014.03.009

Acknowledgements

- Theoretical and Computational Biophysics Group, University of Illinois at Urbana-Champaign
- NVIDIA CUDA Center of Excellence, University of Illinois at Urbana-Champaign
- NVIDIA CUDA team
- NCSA Blue Waters Team
- Funding:
  - NSF OCI 07-25070
  - NSF PRAC “The Computational Microscope”
  - NIH support: 9P41GM104601, 5R01GM098243-02


NIH BTRC for Macromolecular Modeling and Bioinformatics
1990-2017
Beckman Institute
University of Illinois at
Urbana-Champaign

GPU Computing Publications

http://www.ks.uiuc.edu/Research/gpu/

- **Runtime and Architecture Support for Efficient Data Exchange in Multi-Accelerator Applications** Javier Cabezas, Isaac Gelado, John E. Stone, Nacho Navarro, David B. Kirk, and Wen-mei Hwu. IEEE Transactions on Parallel and Distributed Systems, 26(5):1405-1418, 2015.
- **Unlocking the Full Potential of the Cray XK7 Accelerator** Mark Klein and John E. Stone. Cray Users Group, Lugano Switzerland, May 2014.
- **Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations** Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten. Journal of Parallel Computing, 40:86-99 2014.
- **GPU-Accelerated Analysis and Visualization of Large Structures Solved by Molecular Dynamics Flexible Fitting** John E. Stone, Ryan McGreevy, Barry Isralewitz, and Klaus Schulten. Faraday Discussions, 169:265-283, 2014.
- **GPU-Accelerated Molecular Visualization on Petascale Supercomputing Platforms.** J. Stone, K. L. Vandivort, and K. Schulten. UltraVis'13: Proceedings of the 8th International Workshop on Ultrascale Visualization, pp. 6:1-6:8, 2013.
- **Early Experiences Scaling VMD Molecular Visualization and Analysis Jobs on Blue Waters.** J. E. Stone, B. Isralewitz, and K. Schulten. Extreme Scaling Workshop (XSW), pp. 43-50, 2013.
- **Lattice Microbes: High-performance stochastic simulation method for the reaction-diffusion master equation.** E. Roberts, J. E. Stone, and Z. Luthey-Schulten. J. Computational Chemistry 34 (3), 245-255, 2013**.**

GPU Computing Publications

http://www.ks.uiuc.edu/Research/gpu/

- **Fast Visualization of Gaussian Density Surfaces for Molecular Dynamics and Particle System Trajectories.** M. Krone, J. E. Stone, T. Ertl, and K. Schulten. *EuroVis Short Papers*, pp. 67-71, 2012.
- **Fast Analysis of Molecular Dynamics Trajectories with Graphics Processing Units – Radial Distribution Functions.** B. Levine, J. Stone, and A. Kohlmeyer. *J. Comp. Physics*, 230(9):3556-3569, 2011.
- **Immersive Out-of-Core Visualization of Large-Size and Long-Timescale Molecular Dynamics Trajectories.** J. Stone, K. Vandivort, and K. Schulten. G. Bebis et al. (Eds.): *7th International Symposium on Visual Computing (ISVC 2011)*, LNCS 6939, pp. 1-12, 2011.
- **Quantifying the Impact of GPUs on Performance and Energy Efficiency in HPC Clusters.** J. Enos, C. Steffen, J. Fullop, M. Showerman, G. Shi, K. Esler, V. Kindratenko, J. Stone, J Phillips. *International Conference on Green Computing*, pp. 317-324, 2010.
- **GPU-accelerated molecular modeling coming of age. J. Stone, D. Hardy, I. Ufimtsev, K. Schulten.** *J. Molecular Graphics and Modeling*, 29:116-125, 2010.
- **OpenCL: A Parallel Programming Standard for Heterogeneous Computing. J. Stone, D. Gohara, G. Shi.** *Computing in Science and Engineering*, 12(3):66-73, 2010.

GPU Computing Publications
http://www.ks.uiuc.edu/Research/gpu/

- **An Asymmetric Distributed Shared Memory Model for Heterogeneous Computing Systems**. I. Gelado, J. Stone, J. Cabezas, S. Patel, N. Navarro, W. Hwu. *ASPLOS '10: Proceedings of the 15th International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 347-358, 2010.
- **GPU Clusters for High Performance Computing**. V. Kindratenko, J. Enos, G. Shi, M. Showerman, G. Arnold, J. Stone, J. Phillips, W. Hwu. *Workshop on Parallel Programming on Accelerator Clusters (PPAC)*, In Proceedings IEEE Cluster 2009, pp. 1-8, Aug. 2009.
- **Long time-scale simulations of in vivo diffusion using GPU hardware**. E. Roberts, J. Stone, L. Sepulveda, W. Hwu, Z. Luthey-Schulten. In *IPDPS'09: Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Computing*, pp. 1-8, 2009.
- **High Performance Computation and Interactive Display of Molecular Orbitals on GPUs and Multi-core CPUs**. J. Stone, J. Saam, D. Hardy, K. Vandivort, W. Hwu, K. Schulten, *2nd Workshop on General-Purpose Computation on Graphics Pricessing Units (GPGPU-2), ACM International Conference Proceeding Series*, volume 383, pp. 9-18, 2009.
- **Probing Biomolecular Machines with Graphics Processors**. J. Phillips, J. Stone. *Communications of the ACM*, 52(10):34-41, 2009.
- **Multilevel summation of electrostatic potentials using graphics processing units**. D. Hardy, J. Stone, K. Schulten. *J. Parallel Computing*, 35:164-177, 2009.

GPU Computing Publications

http://www.ks.uiuc.edu/Research/gpu/

- **Adapting a message-driven parallel application to GPU-accelerated clusters**. J. Phillips, J. Stone, K. Schulten. *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, IEEE Press, 2008.
- **GPU acceleration of cutoff pair potentials for molecular modeling applications**. C. Rodrigues, D. Hardy, J. Stone, K. Schulten, and W. Hwu. *Proceedings of the 2008 Conference On Computing Frontiers*, pp. 273-282, 2008.
- **GPU computing**. J. Owens, M. Houston, D. Luebke, S. Green, J. Stone, J. Phillips. *Proceedings of the IEEE*, 96:879-899, 2008.
- **Accelerating molecular modeling applications with graphics processors**. J. Stone, J. Phillips, P. Freddolino, D. Hardy, L. Trabuco, K. Schulten. *J. Comp. Chem.*, 28:2618-2640, 2007.
- **Continuous fluorescence microphotolysis and correlation spectroscopy**. A. Arkhipov, J. Hüve, M. Kahms, R. Peters, K. Schulten. *Biophysical Journal*, 93:4006-4017, 2007.