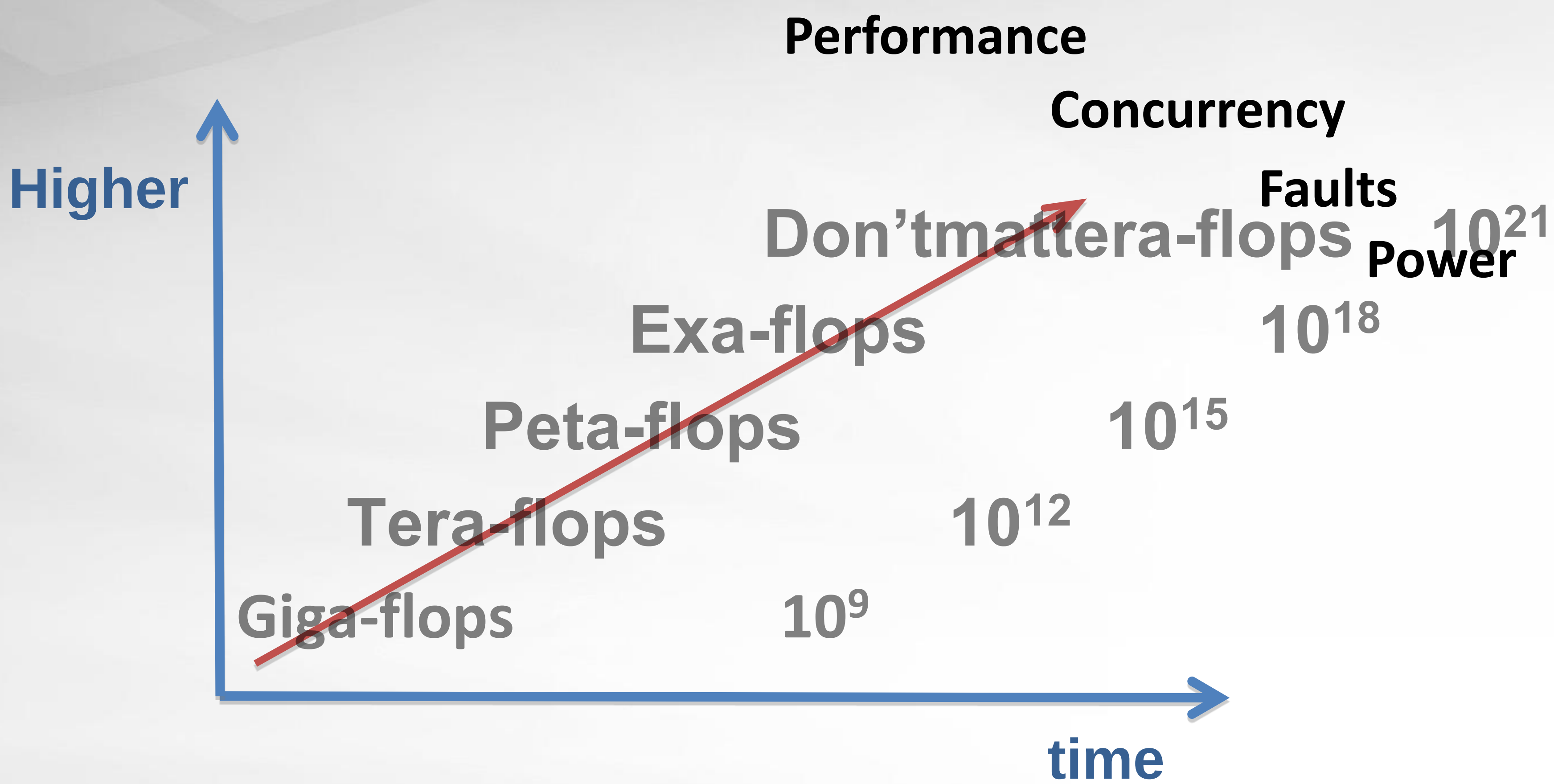
Preparing Applications for current and future Systems: Experiences at PNNL

Darren J. Kerbyson
Laboratory Fellow, HPC Group Lead

August 27th 2014

High Performance Computing



Mega-watt years not Million CPU hours ?
How many DM/s not GF/s ?

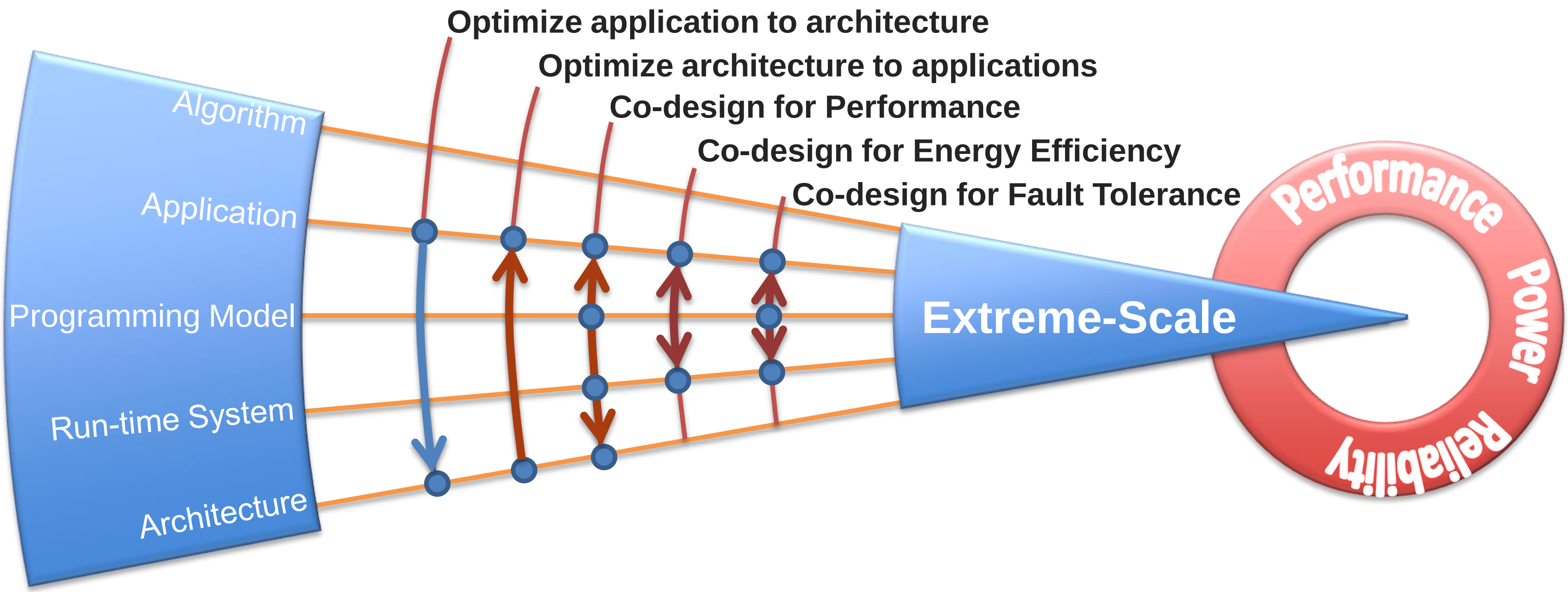
Pacific Northwest NATIONAL LABORATORY

HPC @ Pacific Northwest National Lab

- **A member of the E7 grouping dealing with Exascale**
  - Assisting DOE ASCR in advancing technology development
  - Lead of Performance Execution *Nexus*, a focus for performance related activities
- **HPC at PNNL has long standing research capabilities for exploring software and hardware solutions for advanced systems. From programming models to application development to fault tolerance & highly energy efficient embedded and supercomputing systems.**
- **Collaborative research, partners in other national labs, industry, academia. Many visiting researchers, interns, and post-docs**
- **Main clients include: DOE ASCR, DoD, DARPA**
- **Examples of our research**
  - Modeling and simulation
  - Programming models and runtimes for future systems
  - Energy and Power Optimization
  - Advanced Testbeds

Need for a Holistic approach: Throughout the software stack & down to the hardware

- Combination of interests required for future systems
- In addition multiple metrics of interest



- Underpinning technology is common across computing domains
  - Embedded -> HPC

"Co-design for Exascale systems: Performance, Power, Reliability", Kerbyson, Vishnu, Barker, Hoisie, IEEE Computer, Nov. 2011

MVAPICH2 @ PNNL

" 2:1 Ratio of MVAPICH use to other"

" MVAPICH best performing"

"MVAPICH is MPI of choice"

Systems	towards establishing North West Supercomputing Center
Applications	computational chemistry, subsurface, climate, HEP
Research	techniques into practice

- **Aimed to nurture a culture of computational science and have impact on PNNL mission areas**
  - Capacity cluster funded in part by PNNL, and in part by project “buy-ins”
- **Main Cluster (olympus)**
  - 650 dual-socket AMD Interlagos nodes (22,000 cores)
  - 64 GB per node
  - QDR InfiniBand network (2:1 oversubscription)
  - “Buy-ins” increased node count by 220
  - > 90% utilization
  - Typically smaller jobs (50% < 32 nodes)
- **4 PB Lustre File-system**
- **Additional small-scale production systems:**
  - 18 node Hadoop cluster, 16 nodes Intel Phi, 32 nodes Nvidia, 32 nodes windows HPC
- **Upgrade to Intel Haswell & FDR, September 2014**
- **Stepping stone to larger capability systems within the DOE**

Cascade: 3.4 Pf/s peak production system

- **EMSL – Environmental and Molecular Sciences Lab**
  - DOE Biological and Environmental Research (BER)
- **Cascade: 4th in a series of capability HPC systems @ PNNL**
  - Intel Xeon + Phi processors
  - 1440 Compute Nodes
  - 2-sockets 8-core Ivybridge &
  - 2x 68-core Phi co-processors
  - 128 GB per node (8 GB per Xeon core)
  - FDR InfiniBand Network
  - 2.7 Petabyte shared parallel filesystem (60 GB/s read/write)
  - 3.4 Pf/s peak (2.5Pf/s Linpack)

- **Note: High memory/node enables processing of certain problems in biology, climate research, chemistry and materials science.**
- **Second Cascade system expected CY2015**
- **Working with OSU to make MVAPICH2-MIC available to users**

Advanced Testbeds

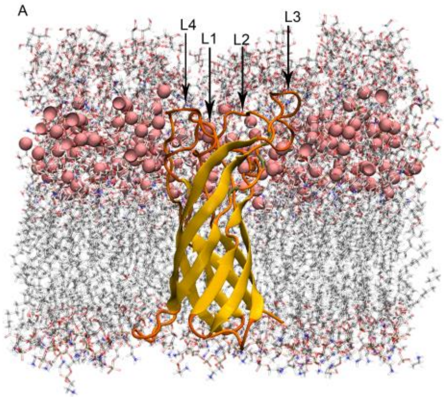
- **State-of-the-art computational resources**
- **Project based and often made available nationally**
- **Part of the Embedded HPC lab (PNNL investment)**

- **DARPA PERFECT**
  - Power Efficiency Revolution For Embedded Computing Technologies
  - Goal to achieve 75 Gflops/W
  - Phase 1: diverse set of technology research across 16 performer teams

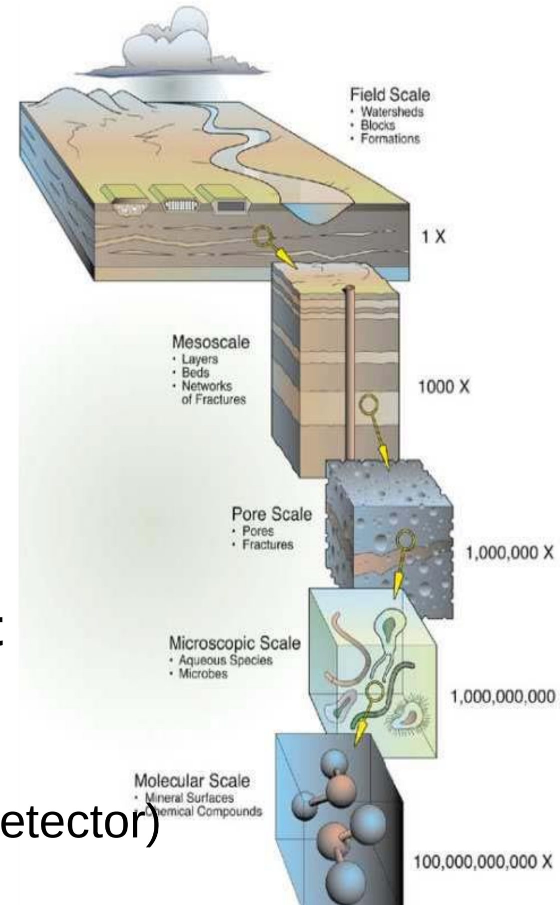
- **DARPA SEAK**
  - Suite of Embedded Applications and Kernels
  - Develop and evaluate rankings for applications of interest to DoD

- **Possible DOE Test, Evaluation and Design**
  - Analyze state-of-the-art technologies for applications of interest to DOE
  - Empirical @ small-scale, predictive @ large-scale

Some examples of our Applications



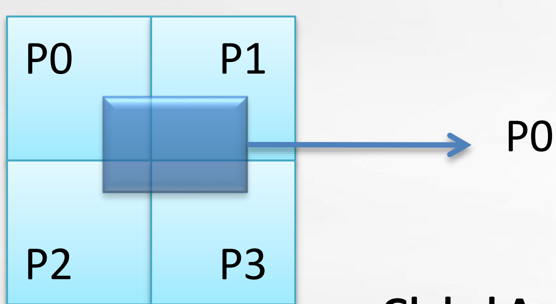
- **Computational Chemistry – NWChem**
  - PNNL lead framework for computational chemistry
  - Examples: Coupled Cluster, Molecular dynamics, Plane-wave DFT, …



- **Subsurface Modeling – eSTOMP, PFloTran**
  - Modeling of subsuface contaminant fate and transport, carbon sequestration
- **Climate – Atmospheric modeling (CAM, WRF)**
  - Integrated Multi-scale modeling, Community Atmospheric Model, Weather Research & Forcasting
- **Physics – Examining why the universe does not contain anti-matter (Belle2)**
  - 6KHz event rate (1.8GB/s), real-time transfer from Tsukuba to PNNL ($400M upgrade to accelerator & detector)
  - Processing to be distributed world-wide

Long standing research in programming models including Global Arrays: Supports #1 app @ ORNL

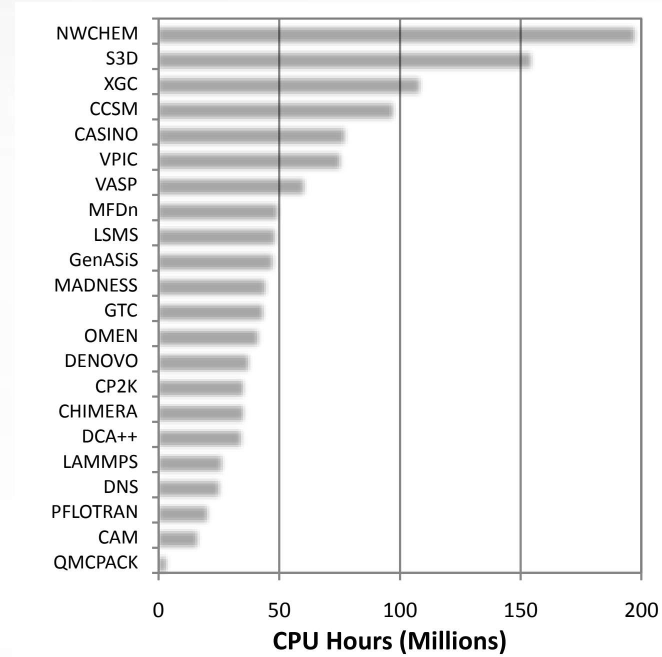
- **Global Arrays: Programming Model that Provides Easy Access to Distributed Data**
  - Simplicity of data access while retaining performance
  - Traditionally suited irregular access to dense arrays
  - Applications include Chemistry, Bio-informatics, sub-surface modeling
  - Use of one-sided communication
  - Library based, inter-operable with MPI



Global Arrays:

```
if (me = P0)
        NGA_Get(g_a, lo, hi, buffer, ld);
```

"Of the 2.3 billion core-hours tracked over the 23 month reporting period, NWCHEM is the top user with 197 million core-hours of 7.5% of the total"



"An Analysis of Computational Workloads for the ORNL Jaguar System", W. Joubert, S.Q. Su, in Proc. ACM Int. Conference on Supercomputing (ICS), Venice, Italy, June 2012, pp. 247-256.

Computational Chemistry NWChem

- **Widely used framework for computational chemistry**
  - Open source
  - Developed by a consortium of developers and maintained at PNNL
- **Use of appropriate programming model**
  - Extensive use of Global Arrays
  - MPI/MVAPICH for many packages
- **NWChem can handle**
  - Biomolecules, nanostructures, and solid-state
  - From quantum to classical, and all combinations
  - Ground and excited-states
  - Gaussian basis functions or plane-waves
  - Properties and relativistic effects
- **Some on-going work**
  - Density functional theory (DFT), time-dependent DFT (TD-DFT)
  - Plane-Wave Density Functional Theory (DFT), Ab Initio Molecular Dynamics
  - High-level Coupled-Cluster methods

M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, W.A. de Jong, "NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations" Comput. Phys. Commun. 181, 1477 (2010)

CC methods on Intel Xeon Phi Coprocessor

Efficient Implementation of Many-body Quantum Chemical Methods on the Intel® Xeon Phi™ Coprocessor

Edoardo Aprà
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
edoardo.apra@pnnl.gov

Michael Klemm
Software and Services Group
Intel Corporation
michael.klemm@intel.com

Karol Kowalski
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
karol.kowalski@pnnl.gov

SC14 New Orleans, LA hpc matters.

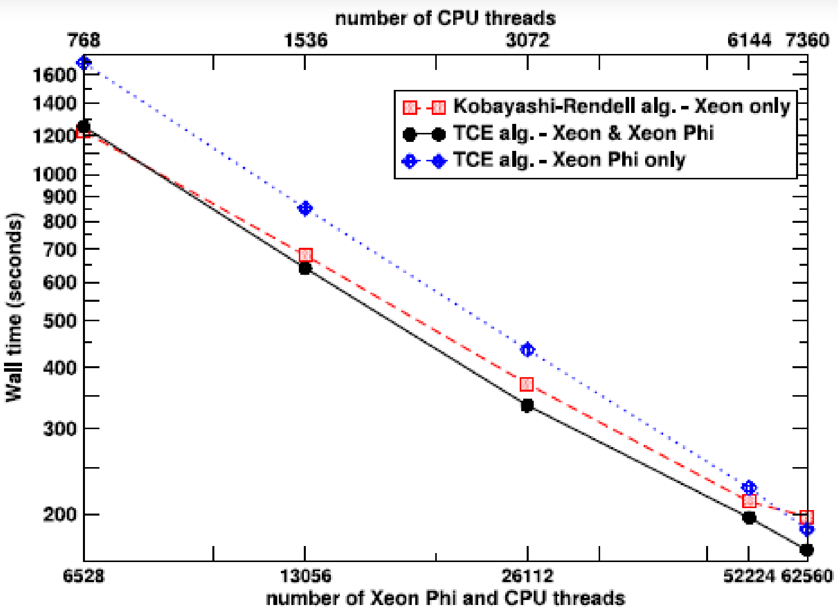


Fig. 8. Wall time to solution (in seconds) for the perturbative triples correction to the CCSD(T) correlation energy of the pentacene molecule ($C_{22}H_{14}$). A logarithmic scale is used on all the axis.

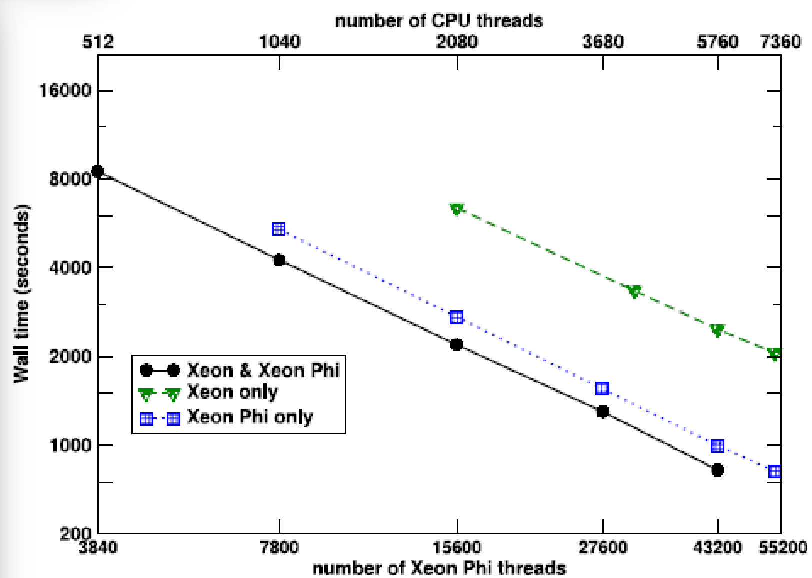
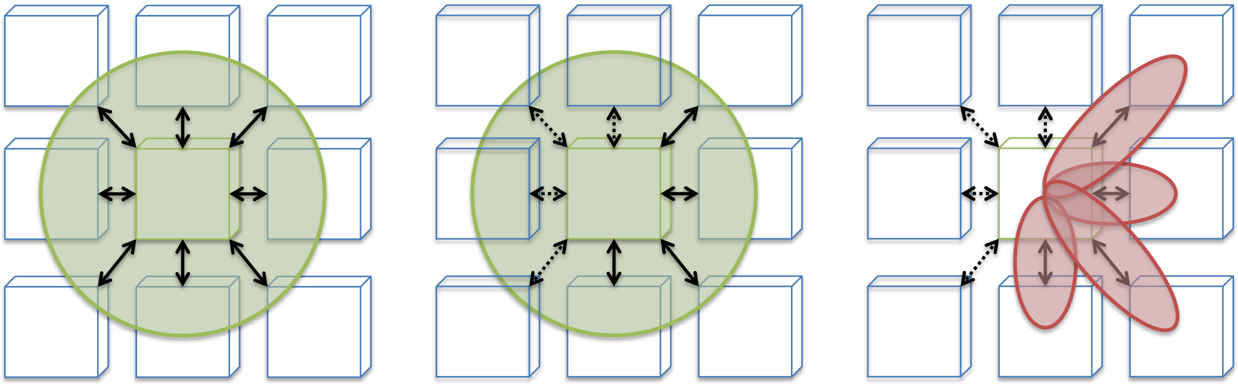


Fig. 9. Wall time to solution (in seconds) for the perturbative triples correction to the CCSD(T) correlation energy of the 1,3,4,5-tetrasilylimidazol-2-ylidene molecule (formula $Si_4C_3N_2H_{12}$) in its triplet state. A logarithmic scale is used on all the axis.

Example – Molecular Dynamics (ARGOS)

- **Removal of explicit synchronization from the basic MD time steps**
- **Order of magnitude increase in scalability through data distribution by cell pairs**



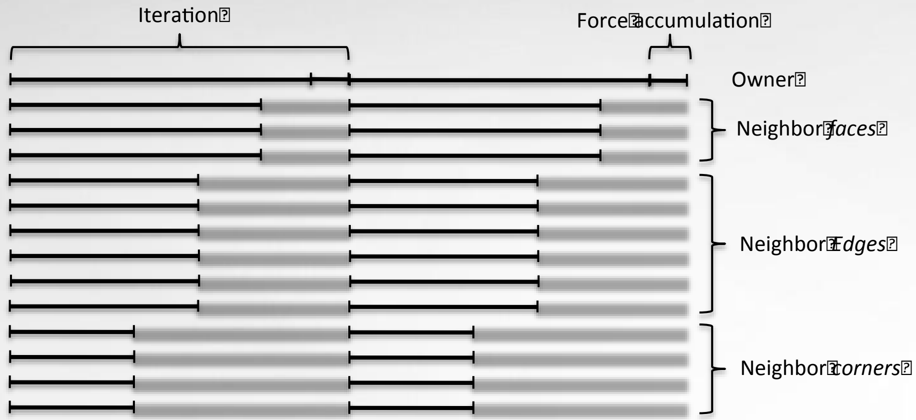
Spatial Decomposition: Particle Interactions are calculated within a cutoff distance	**Symmetry: Only half of all interactions must be computed**	**ARGOS Partitioning: Cell-cell interactions are partitioned among available processors**

- **This partitioning method leads to load imbalance**
  - Load imbalance is inherent in the algorithm and does not arise as a result of poor partitioning
  - The degree of load imbalance will evolve over time as the simulation evolves

Energy Optimization using Energy Templates: exploiting dynamic concurrency in applications

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- **Figure shows load variation for ARGOS across 14 cores (3D partitioning)**



- **Application provides information describing per-core volume of computation for the near-term future**
  - Information pushed to the Energy Template may be periodically updated to reflect current simulation conditions
- **Energy Template "dilates" active computation (black) to reduce idle time (green)**
  - Goal is to ensure all processor cores reach the iteration boundaries simultaneously
  - Dynamic Frequency Scaling is the mechanism used to affect processing rate
  - Care must be taken to avoid performance impact on "force accumulation" phase, which lies in the critical processing path

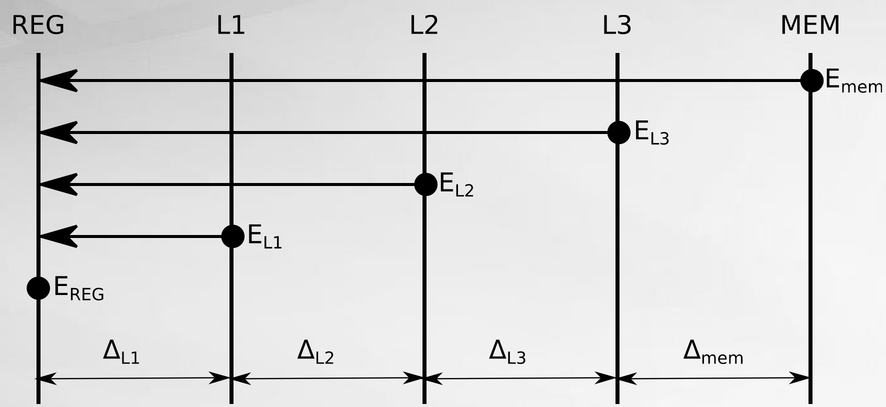
Energy Templates are effective tools for energy optimization

- **Energy Templates are the interface between the application and runtime layers**
  - ETs allow applications to describe computational behavior that could not be determined by lower software layers
  - ETs separate the *policy* describing when to apply energy-saving techniques from the *mechanisms* used to implement these techniques
- **We have applied Energy Templates in several scenarios**
  - To applications with dramatically different computational patterns (wavefront pipelined processing and more traditional BSP)
  - To systems with different mechanisms for reducing power consumption (e.g., idling cores, DVFS, interrupt vs. polling message delivery)
- **Results demonstrate low overhead, significant power/energy savings, efficiency across scales, and applicability to wide variety of applications and systems**

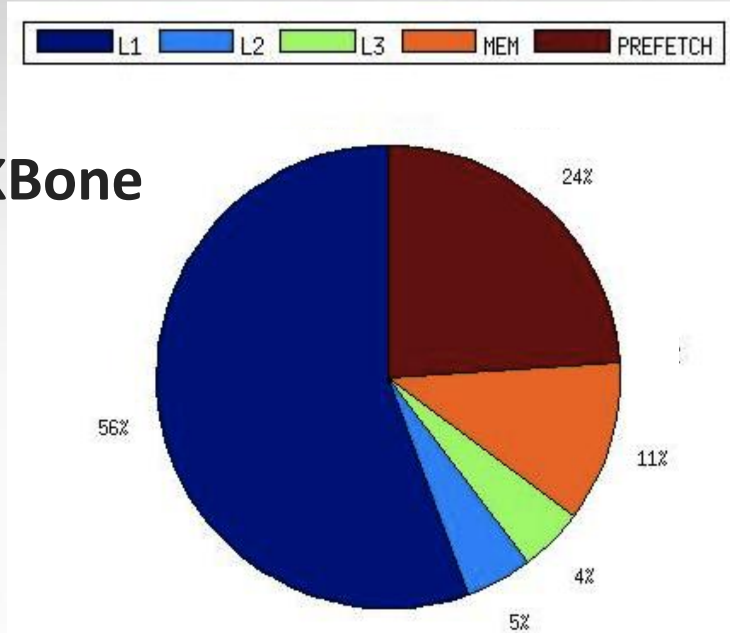
Quantify Energy costs of Data Movement in the memory hierarchy

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Energy costs measured for individual operations



NEKBone



Basis of model

$$E_{DM} = \sum_{i} E_i * N_i$$

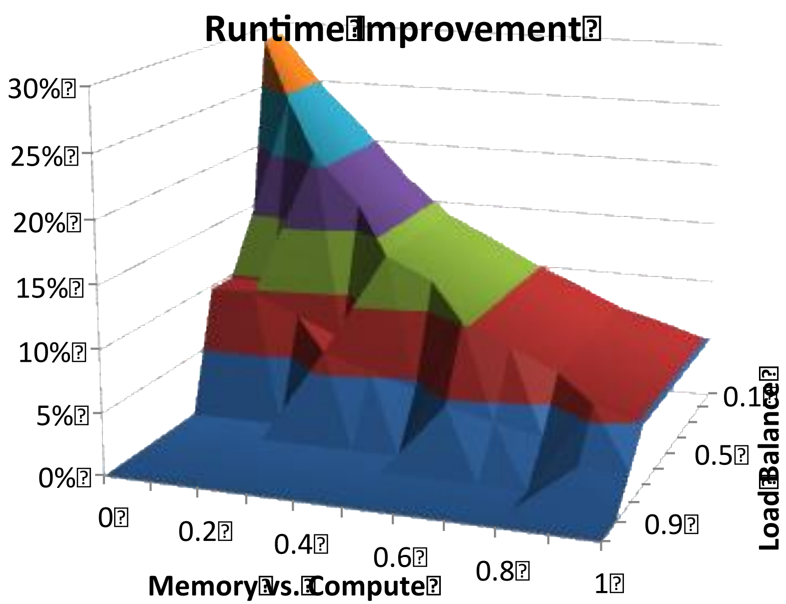
Isolation of operations is complex and required carefully designed benchmarks

- Moving data from the L1 to the processor's registers is dominant
- Memory prefetcher can have a large energy consumption in data movement
- Memory prefetcher may waste energy prefetching unused data (CG solver)
- < 50% of enegy cost currently in data movement

"Quantifying the Energy Cost of Data Movement in Scientific Applications", Kestor, Gioiosa, Kerbyson, Hoisie, IEEE Int. Symp. on Workload Characterization, Portland, Sept. 2013, Best Paper

Power Steering – Steering power to where the data is

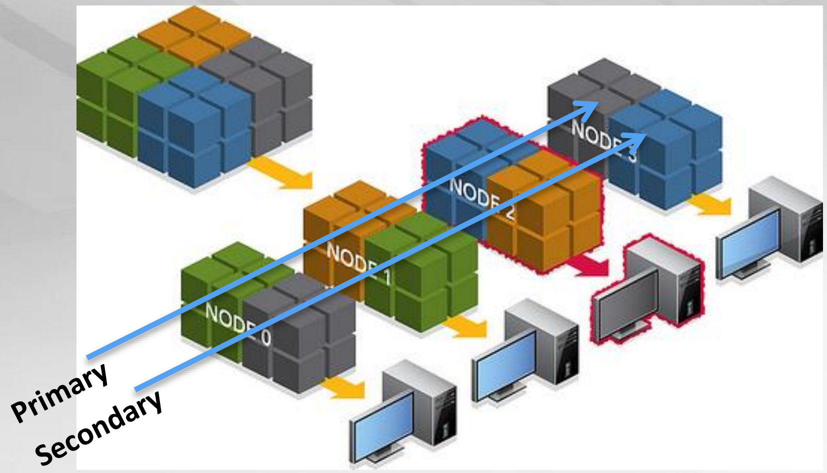
- **Power Steering – Direct system power to cores with work to do (Power-balance) rather than load-balance work across a system**
- **Mimic on current system using mid P-state**
- **Explored using a *synthetic workload* for a variety of characteristics**
    1. Compute Intensity: Workloads that demonstrate higher compute intensity are more sensitive to processor core *p-state*.
    2. Load Balance: Imbalanced workloads exhibit *slack* that can be exploited for improvements in energy efficiency



Runtime improvement when dynamically allocating power to overloaded cores within a fixed global power budget.

Maximum when load-imbalanced & compute bound

Fault-Tolerance:
Application + Programming Model + Run-Time



Example data mapped across 4 nodes:
Node 2 dies, data recovery from secondary copy (node 3)

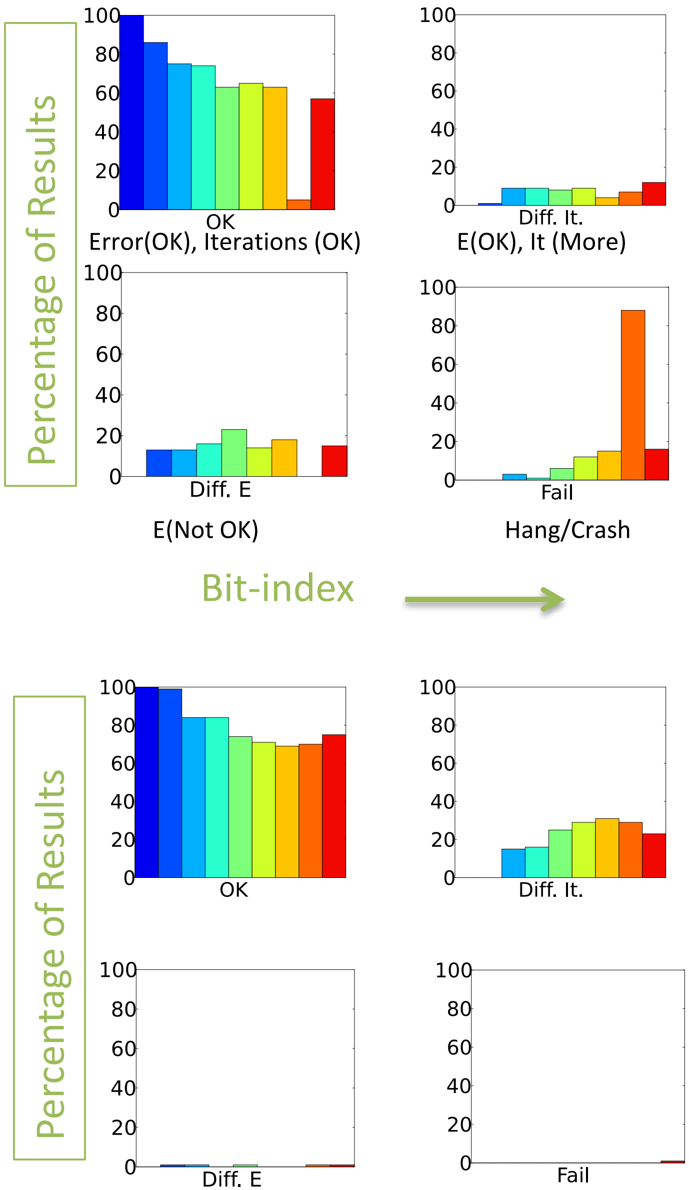
Time-space Trade-off
- **Time: k*t for k replicas (2-3% typical)**
- **Space: k*M (5%) typical**
- **Increased Job-Level MTBF ($P^{-1/k}$)**

- Increased system size & inherent increases in technology faults will lead to a greater reliability issue. 85% of failures at node level
- Combined application / run-time approach
  - Selective Replication
  - identify node failures &
  - ensure continued execution
- Use an extension to the Global Arrays (GA) programming model
  - Task based programming in the application
- Application specifies which data is critical (for replication) & how to re-compute a task should a node fail

Resiliency Co-design: Soft Errors

- Near-threshold Voltage execution is a likely cause for soft errors
  - Undetected multi-bit flip
- Potential for silent data corruption

- At PNNL, we have performed an in-depth analysis of impact of soft errors
- Algorithms for scalable detection and correction
  - Practically no undetected soft errors



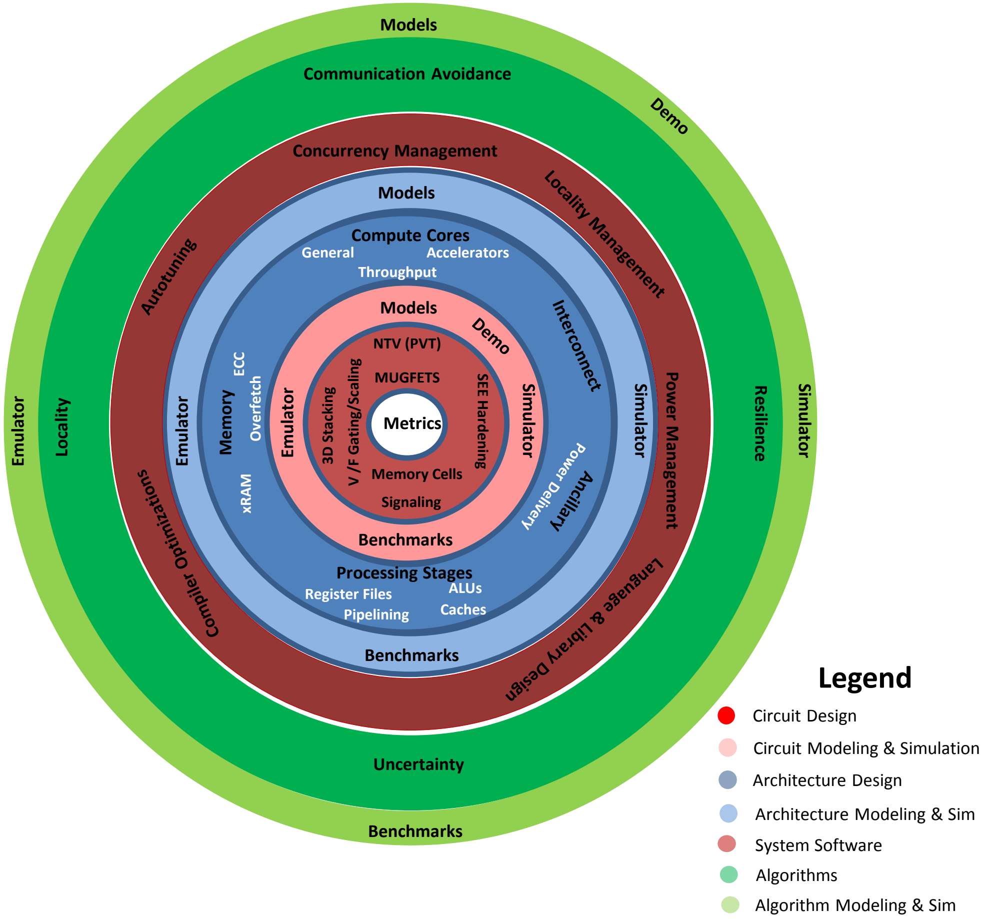
Van Dam et al., A Case for Soft Errors Detection and Correction in Computational Chemistry, Journal of Chemical Theory and Computation (In Review) 2013,

Diversity in DARPA PERFECT

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **PERFECT Challenge**
  - Goal: Embedded system delivering "75 GFLOPS/W" by 2018
  - Performers contribute only part of a system (architecture to algorithms)
  - TAV must assess Performer's contribution w.r.t. entire system
- **Three pillars to the assessment strategy**
  - Baseline Architectures
  - PERFECT Suite
  - Proxy Architecture
- **Support and evaluate 16 Performer teams**

- PERFECT Landscape



DARPA PERFECT : Baseline Architectures

- Architectures that reflect state-of-the-art systems
- Real world data points for power/performance
- Goals:
  - Performance/power profiles for the PERFECT Suite
  - To calibrate modeling and simulation environments
  - Testbeds accessible by16 PERFECT performers
- EHPC lab @ PNNL
- Power Instrumentation:
  - Tier 1: Watts-up power meters,
  - Tier 2: Internal (RAPL, Amrester)
  - Tier 3: DAQ

Platform	# Cores (Threads)	Peak Perf (GFLOPS)	Clock (GHz)	Peak Power (Watts)	GFLOPS per Watt	Mem (GB)
nCore BD-Y TI Keystone II	16+96 (ARM+DSP)	614.4 (SP)	1.2+ 1.4	36-56	17.1 - 11 (SP)	56
NVIDIA Kayla	4+2(384) (ARM+SMX)	270 (SP)	1.2 / 1.05	22		2/1
IBM POWER7	8(32)	264.96	4.2	240*	1.104	16
Intel x86 Haswell	4(8)	294.4 (SP)	2.3	45	6.54	16

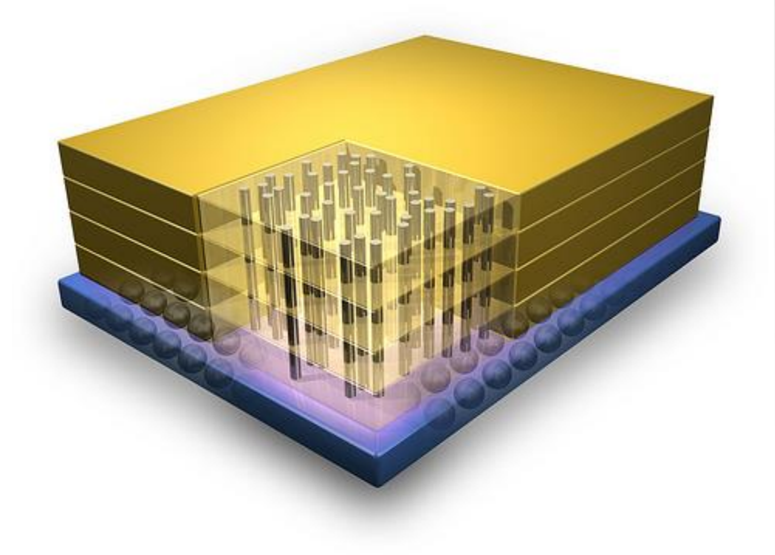
Kayla

P7

nCore

Haswell

Hybrid Memory Cube (HMC) Test Board

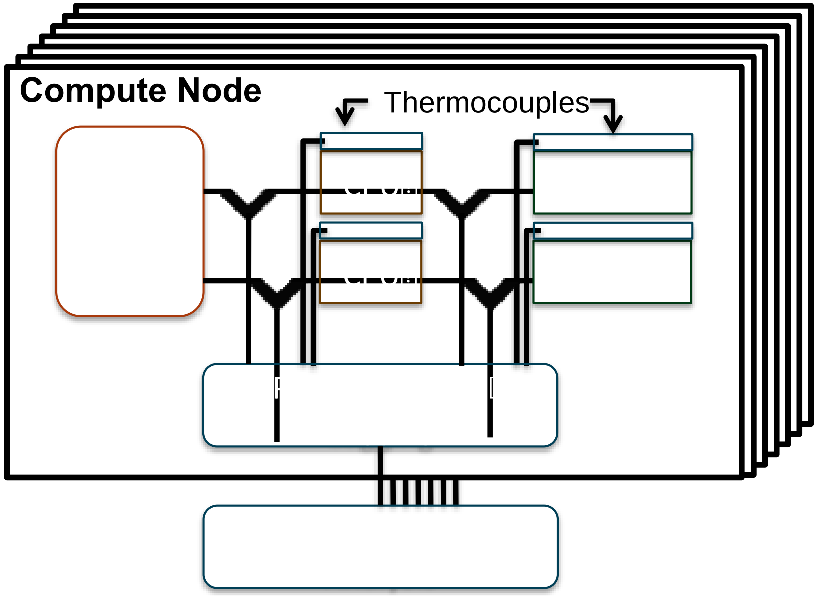


- One of 8 within the DOE lab community
- HMC Gen2
- Novel ultra high bandwidth (BW) DRAM memory organization
  - 4x socketed DDR3L-1600 SODIMMs (2GB, x 72, Dual Rank, ECC)
  - 4x channels with peak 20GB/s per channel
- Altera FPGAs to exercise the HMC device
  - input vectors exercise HMC

- We are exploring
  - Impact on future systems of current design point (& other points)
  - Effective BW for various access pattern of interest
  - Possible impacts on resiliency
  - Power & Temperature profiles under various loads

SEAPEARL: Power and thermal analysis @ reasonable scale

- Critical need
  - Ability to study power consumption and thermal effects *at scale*
  - Correlation of measurements to workload features (not steady-state)
  - Platform for development of modeling and optimization capabilities



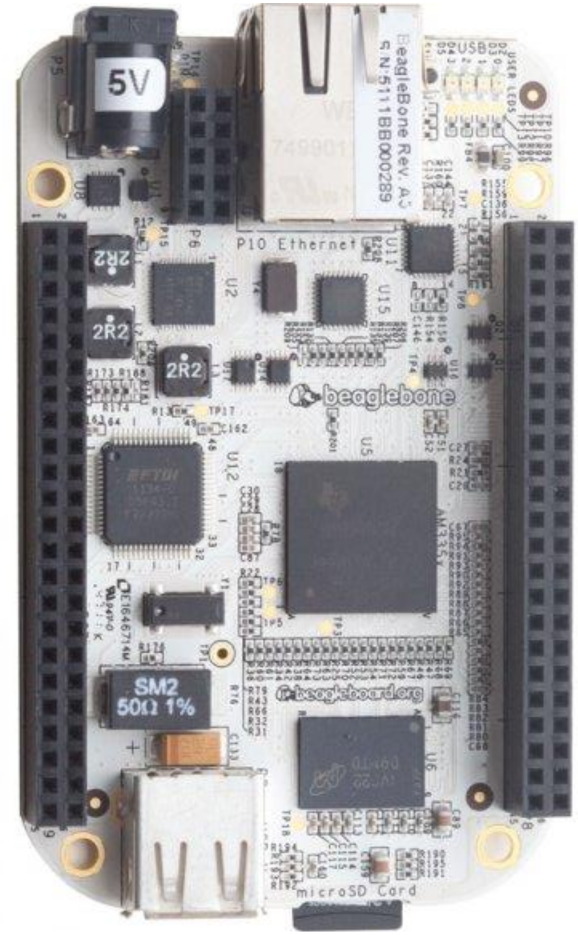
- SEAPEARL: a Unique Resource
  - High fidelity power measurement
    - Spatial: separate CPU from memory
    - Temporal: low sampling period of 1ms
  - Coupled thermal information
  - Advanced architectures: x86 multi-core and AMD Fusion (integrates CPU and GPU)
- Off-line analysis + potential for on-line (dynamic) optimization

Cluster Power and Thermal Instrumentation

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

PowerInsight (Penguin Computing)

- Designed to instrument commodity HW
- Data logging system in each compute node
  - ARM Cortex A8 processor
  - Connectivity via 10/100 Ethernet and USB
  - Full Linux SW stack
- ~1KHz sampling rate per data channel
- 21 available data channels
  - Power
    - In-line Hall effect current sensors and voltage divider ensure low impact on power rail
    - Separate data channels for CPU sockets and associated memory
  - Thermal
    - Pico Lab TC-08 Thermocouples
    - Data range from -270° to +1820°F
    - Measure at component level and node inflow/outflow



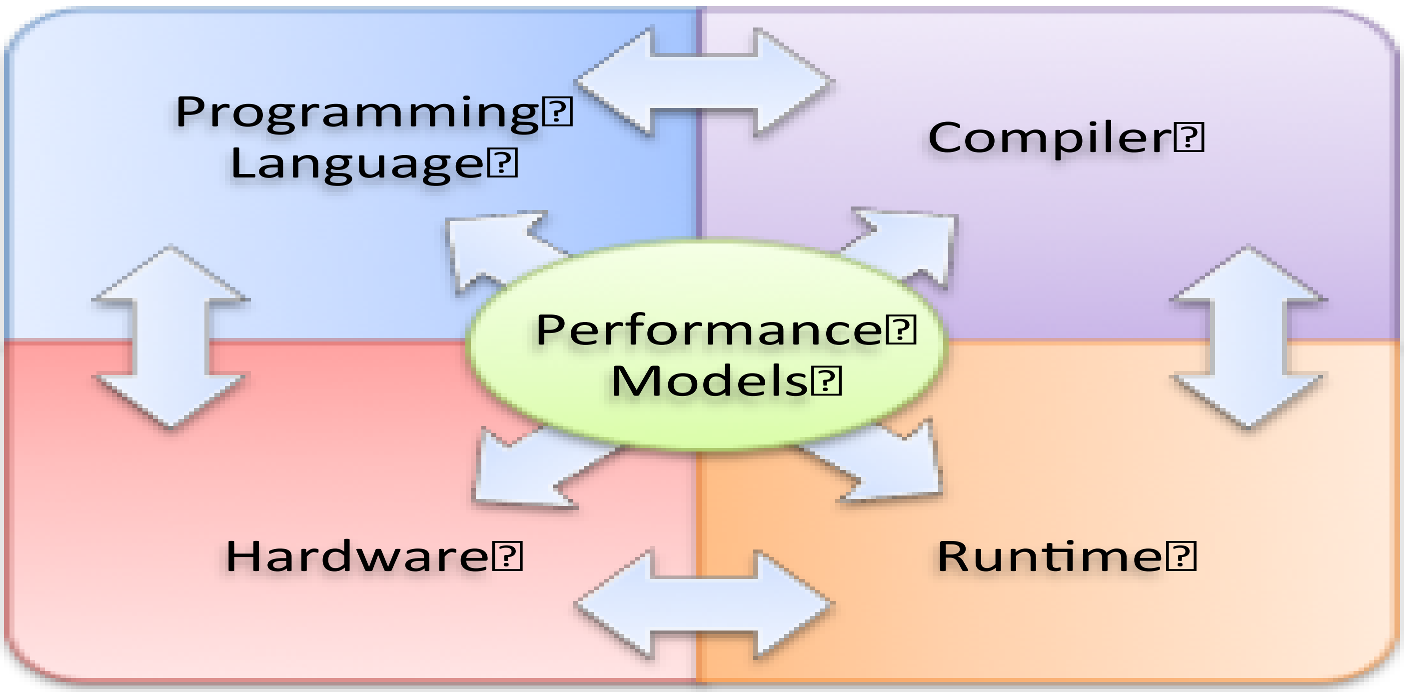
BeagleBone hardware in each node logs power and thermal measurements and connects with external data logging system

Further testbeds

- **Power8 + GPUs**
  - OpenPower consortium
- **ARM64**
  - Initially tried to procure in September 2013
- **Empirical analysis and optimization at small-scale coupled with performance modeling for prediction of possible large-scale systems**
- **Testbeds also allow for optimization of system software (including MVAPICH)**
- **Further modeling work includes**
  - DARPA funded study on the possible advantages of Silicon Photonics
  - DOE ASCR on modeling of extreme-scale scientific workflows

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965



Summary

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- **MVAPICH is widely used at PNNL**
  - Users experience a turn-key operation
  - Looking forward to MVAPICH2-MIC for Cascade
- **Application base is diverse**
  - Programming model matched to application
  - Global Arrays interoperable with MPI
- **Systems**
  - Large-scale production systems, often provide a stepping-stone to LCF's
  - Advanced Testbeds allow analysis at small-scale
- **Research in exploration of future system performance / power / Reliability & thermal issues**
  - Programming models, power optimization, fault tolerance
  - Application centric, of interest to various clients
  - Modeling

explore-in-advance & optimize at run-time

Acknowledgements

- Researchers at PNNL:
  - Kevin J. Barker, Daniel Chavarria, Roberto Gioiosa, Adolfy Hoisie, Gokcen Kestor, Sriram Krishnamoorthy, Joseph Manzano, Andres Marquez, Leon Song, Nathan Tallent
- Advanced Scientific Computing Research (ASCR), DOE
  - Center for Exascale Simulation of Advanced Reactors (CESAR) Co-design center lead by Argonne National Laboratory
  - Beyond the Standard Model (BSM)
  - Performance Health Monitoring (PHM)
  - Modeling Execution Models (MEMS)
  - Integrated End-to-End Performance Prediction and Diagnosis of Extreme Scale scientific Workflows (IPPD)
- EMSL – Environmental and Molecular Sciences Lab
- Biological and Environmental Research (BER), DOE
- Advanced Computer Systems Research Program (ACS)
- DARPA
  - Power Efficiency Revolution for Embedded Computing Technologies (PERFECT)
  - Suite of Applications and Kernels (SEAK)