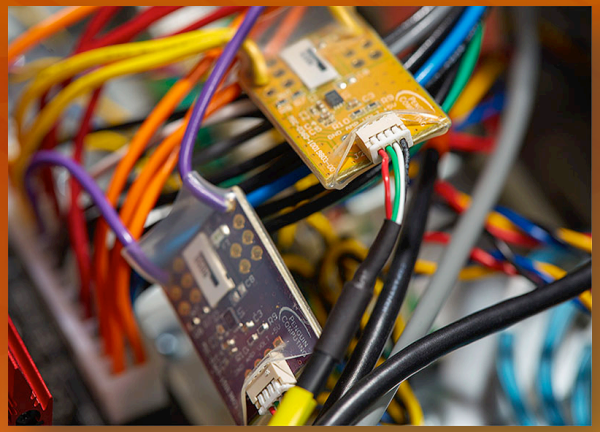


Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Evaluating Novel Networks: Combining Empirical and Predictive Test-beds in CENATE

Performance and Architecture Lab (PAL)
Pacific Northwest National Laboratory

Darren Kerbyson
Associate Division Director and Lab Fellow, High Performance Computing

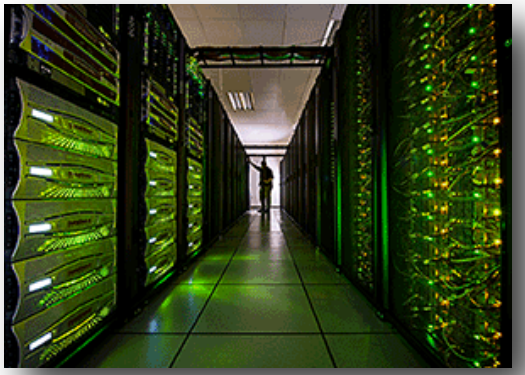
Work with: Kevin J. Barker, Ryan Friese, Roberto Gioiosa, Nitin Gawande, Adolfy Hoisie, Gokcen Kestor, Andres Marquez, Matthew Macduff, Shuaiwen Leon Song, Nathan Tallent, Antonino Tumeo

MUG, August 16th-17th, 2016

U.S. DEPARTMENT OF ENERGY

Overview of Systems @ PNNL

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



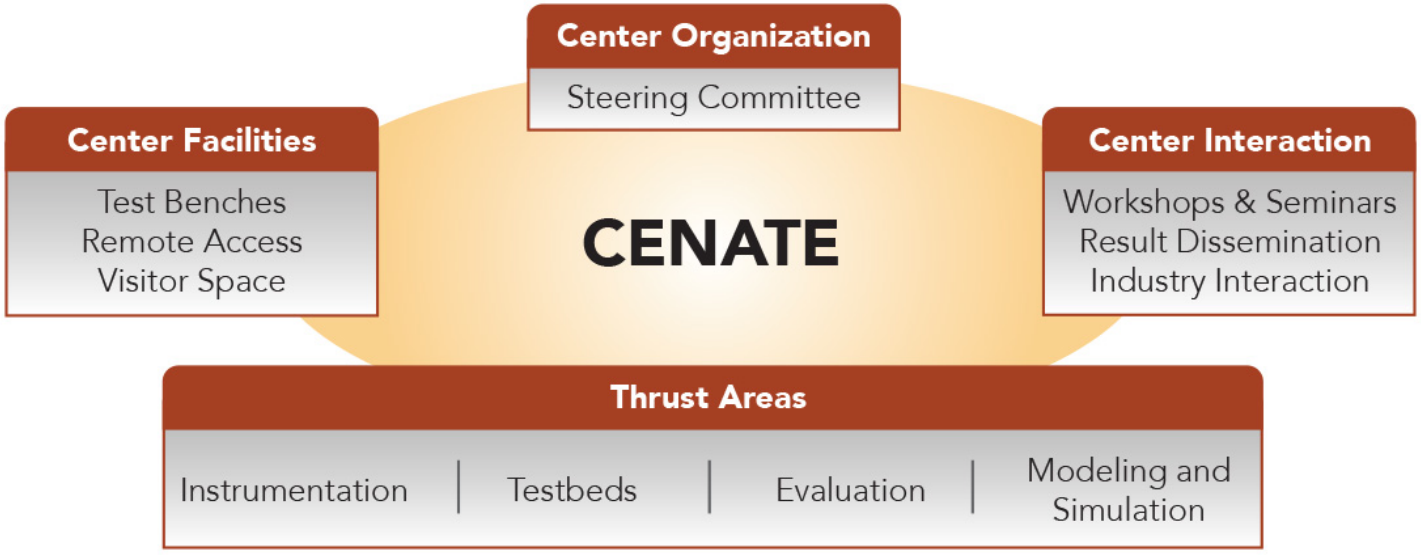
- Cascade (3.4PF): User facility for DOE SC BER
  - 1440 2-socket 8-core Ivybridge + 2-socket Intel Phi + 128GB/node, InfiniBand
- Constance (0.26PF): Institutional Computing
  - 452 2-socket 12-core Haswell + 64GB/node, InfiniBand
- "On-ramp systems"
  - Nvidia K80's, Cloud Testbed, Hadoop cluster, KNL
- Testbeds (CENATE)
  - Seapearl (instrumented cluster), Data Vortex, DGX-1, HMC, Contutto, network testbed …
- MVAPICH available on most systems
  - Library of choice
  - Higher performing, easier to use, easier to install
  - Leading edge optimizations, binding processes support, SLURM integration,
  - Interest in monitoring tools: INAM, OEMT

Center of Advanced Technology Evaluation (CENATE)

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

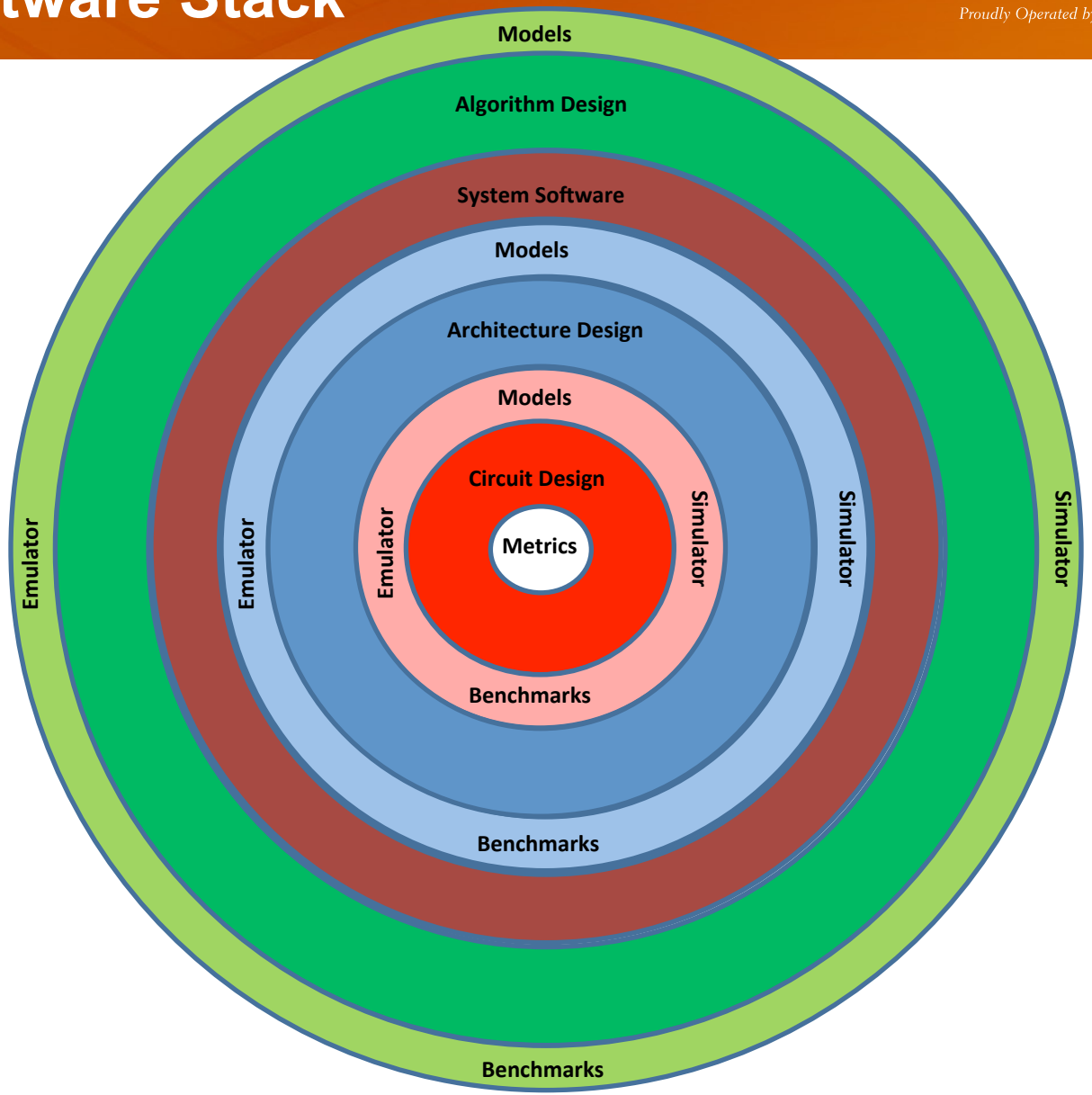
U.S. DEPARTMENT OF ENERGY | Office of Science

- Advanced technology evaluations
- Instrumentation for power and performance
- Testbed infrastructure for high-throughput evaluation
- Predictive exploration: integration of results from empirical evaluations with modeling and simulation
  - Impact at scale; “what-ifs”; for DOE apps



Technology Fragmentation Across the Hardware/Software Stack

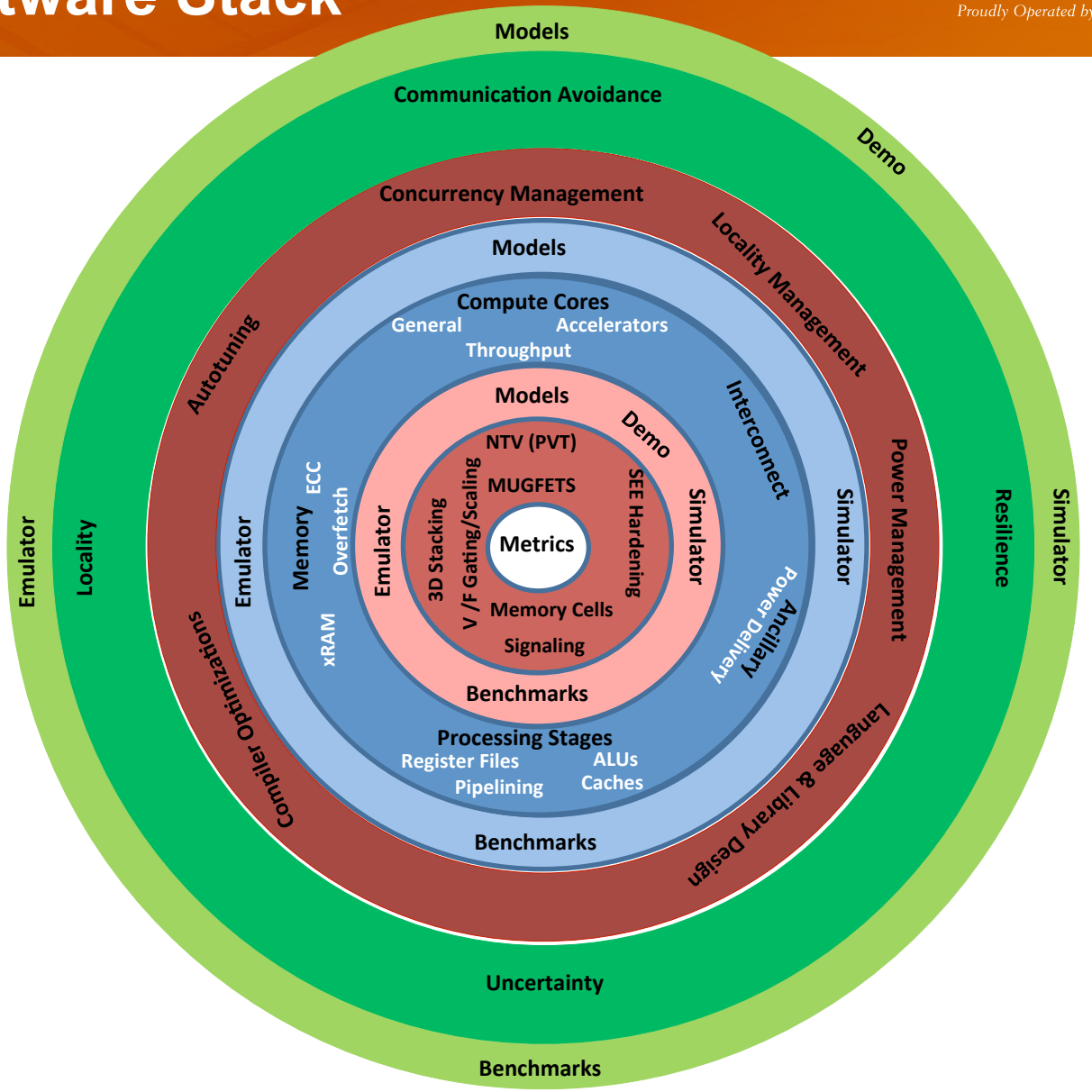
- Circuit Design
- Circuit Modeling & Simulation
- Architecture Design
- Architecture Modeling & Simulation
- System Software
- Algorithm Design
- Algorithm Modeling & Simulation



Technology Fragmentation Across the Hardware/Software Stack

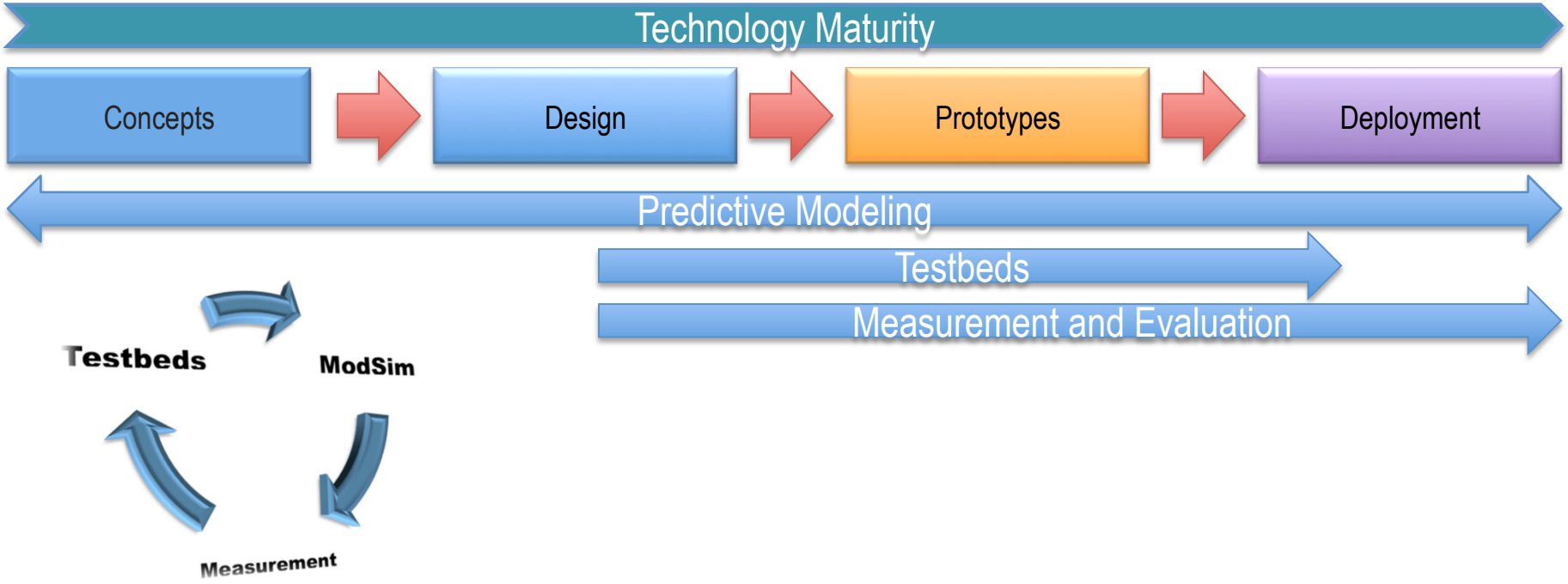
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Circuit Design
- Circuit Modeling & Simulation
- Architecture Design
- Architecture Modeling & Simulation
- System Software
- Algorithms
- Algorithm Modeling & Simulation



CENATE Covers a Multidimensional Technology Space

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



Sub-systems	Processing	Memory	Network	Storage and I/O
Emerging Paradigms	Approximate	Quantum	Neuromorphic	Superconductive
Workloads	Numeric	Machine Learning	Data Analytics	Graph Analytics

Advanced Measurement Laboratory (AML)

- AML provides infrastructure to measure
  - Early Engineering Boards
  - Subsystem Prototypes (e.g., HMC)
  - Small Systems
- AML measures
  - Performance
    - Time-to-solution
    - Performance counters
  - Power
    - System Wall Power
    - Internal Shunts and Hall-sensors
  - Temperature
    - Thermo-Couples
    - Thermal cameras
- Building up FPGA capabilities
  - Xilinx and Altera Toolkits
  - Mentor Graphic's Modelsim

SEAPEARL: Integrated Power, Performance, and Thermal Measurement

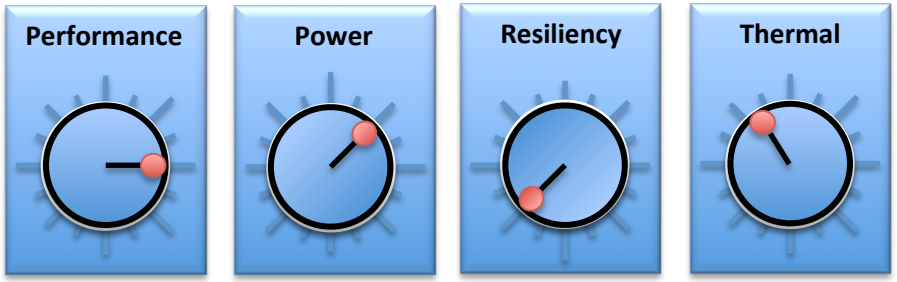
- Critical needs
  - Ability to study power consumption and thermal effects ***at scale***
  - Correlation of measurements to workload features (not steady state)
  - Platform for development of modeling and optimization capabilities
- SEAPEARL: A Unique Resource
  - High-fidelity power measurement
    - Spatial: separate CPU from memory
    - Temporal: low sampling period of 1 ms
  - Coupled thermal information
  - Advanced architectures: x86 multicore and AMD Fusion (integrates CPU and GPU)
- Offline analysis and potential for online (dynamic) optimization

CENATE: Establishing Best Practices for Measurements

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

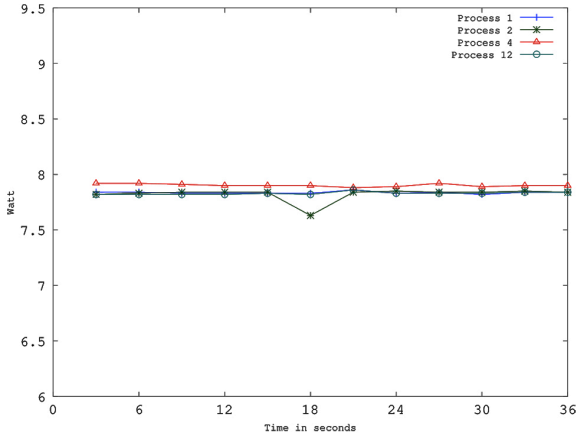
- Integrating state-of-the-art measurements into Idiosyncratic systems
  - Best practices: Measurement is a science and a craft
- Multi-tier approach:
  - Tier 1: external, low-resolution, available to all system
  - Tier 2: internal, system specific, vendor based (e.g., RAPL, Amester, Data Vortex …)
  - Tier 3: external, high-resolution, invasive, need vendor support (e.g., Power Insight)
- Measurements:
  - In-band: synchronous with apps (e.g., performance counters)
  - Out-of-band: asynchronous with apps (e.g., power meters)
- Interaction with vendors is essential:
  - Understand position of sensors
  - Access to “special” systems and proprietary software

- Multiple Metrics of interest:

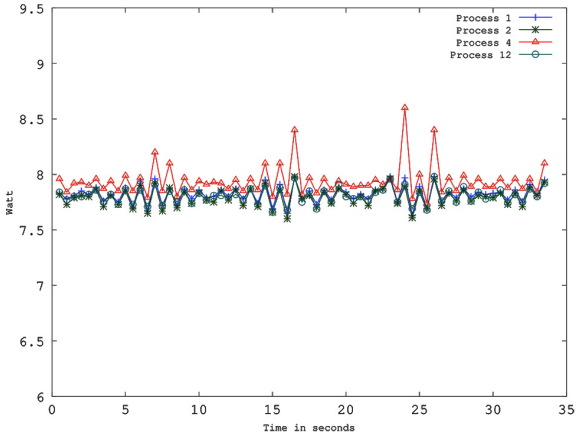


Instrumentation Granularity Affects Insight

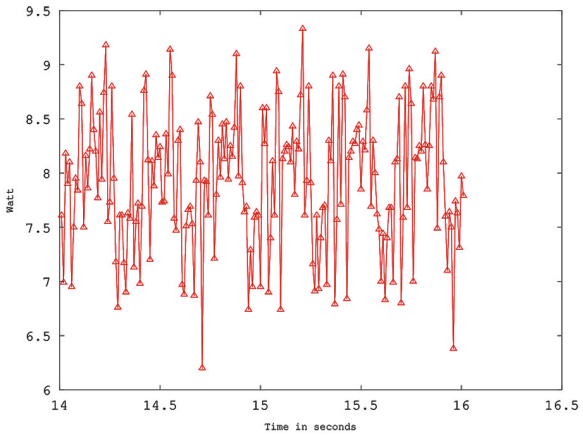
- Coarse *spatial* and *temporal* instrumentation may hide important information
  - e.g., peak power/temperature consumption
- Example for scalar pentadiagonal solver with 32 parallel threads
  - Peak power measured with 0.1 second granularity is much higher (9.7 W/core) than the one measured with 1 second granularity (7.8 W/core)



1 second



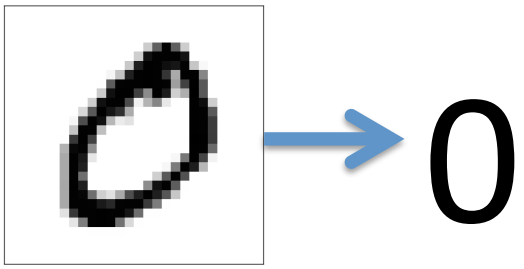
0.5 second

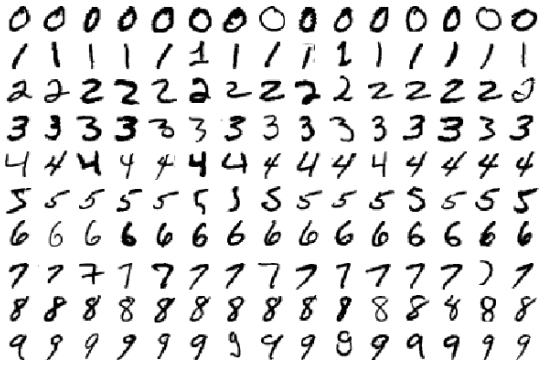


0.1 second

Evaluation Strategy: Need to Overcome Architectural Diversity

- Solution-oriented:
  - What is the best hardware/software solution for my problem?
  - "Future proof": Same problems, different solutions
    - Leverage experiences gained from DARPA SEAK
    - Functional descriptions (input/output)
    - Input generators, output correctness checkers
  - Problem interface is fixed; algorithm and architecture are variables
  - End-user trade-off evaluation: performance versus power versus accuracy
  - Encourage creative solutions, e.g., co-design
- Text Image Classification: Recognize handwritten digits

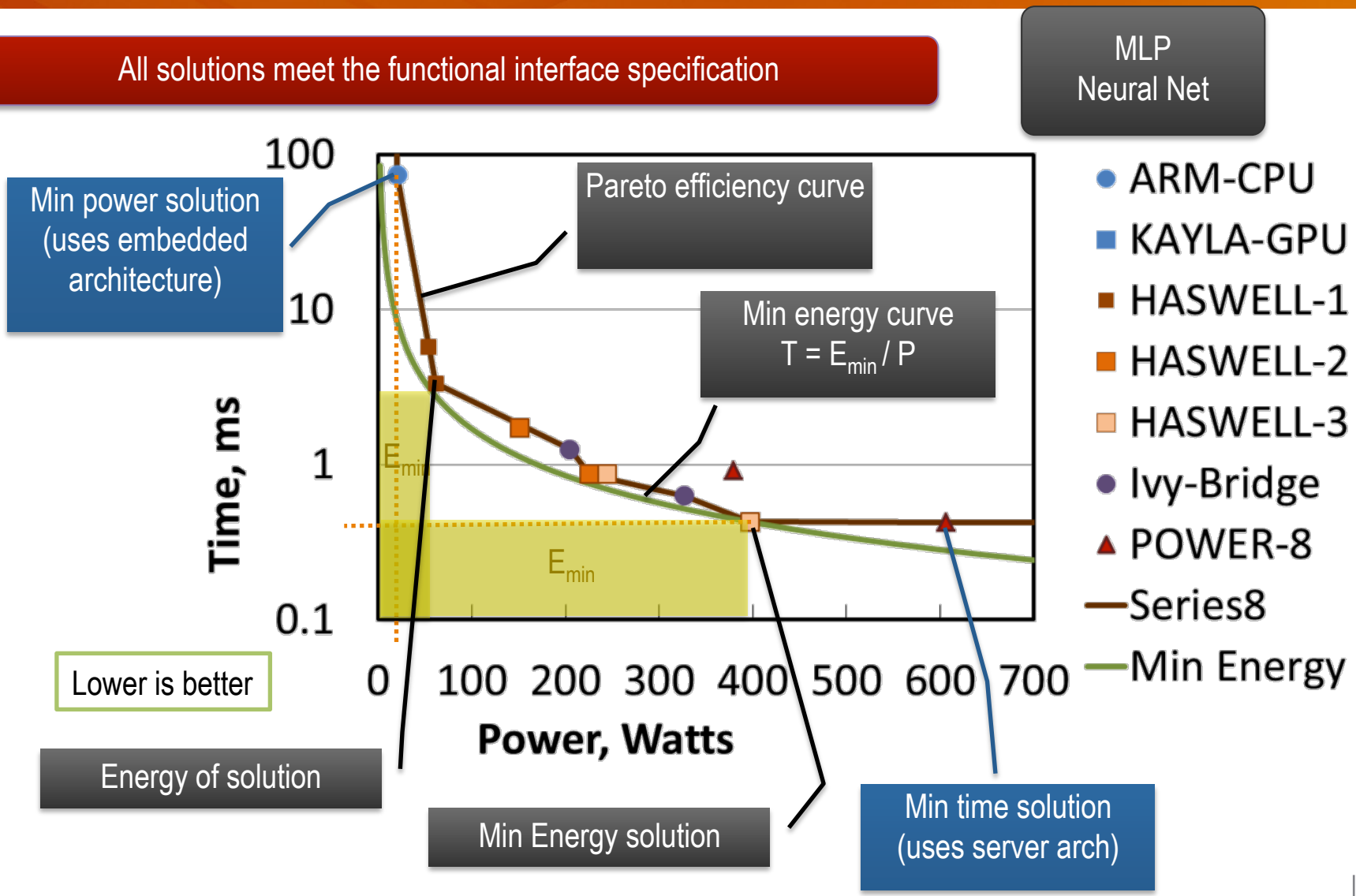




Output:
label (0 to 9),
error rate

Trade-off Evaluation

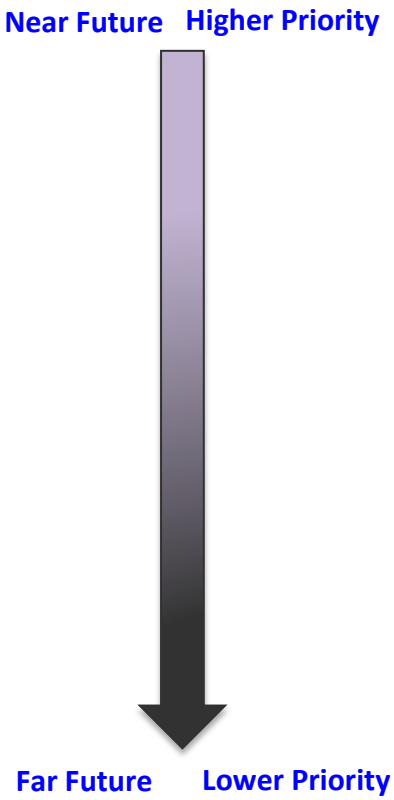
All solutions meet the functional interface specification



Sample Workloads

- Numeric:
  - nekbone, coMD, miniAMR, miniFE, SNAP, LULESH, MCB, AMG2013,
  - UMT2013, RSBench, XSBench, HPCG,
  - FFT, Space-Time Adaptive Processing, Synthetic Aperture Radar, Text Image Classification, Wide Area Motion Imaging, Image Fusion
- Machine Learning/Data Analytics:
  - Support Vector Machine, K-Nearest Neighbor, K-Means, Spectral Clustering, FP-Growth, MLP Neural Networks
- Graph Analytics:
  - Graph clustering, Vertex matching, page rank, Breadth-first-search

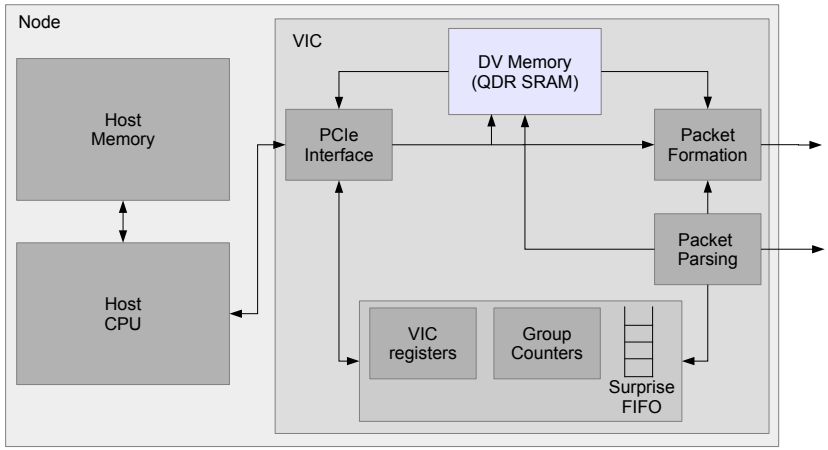
Testbed Classification



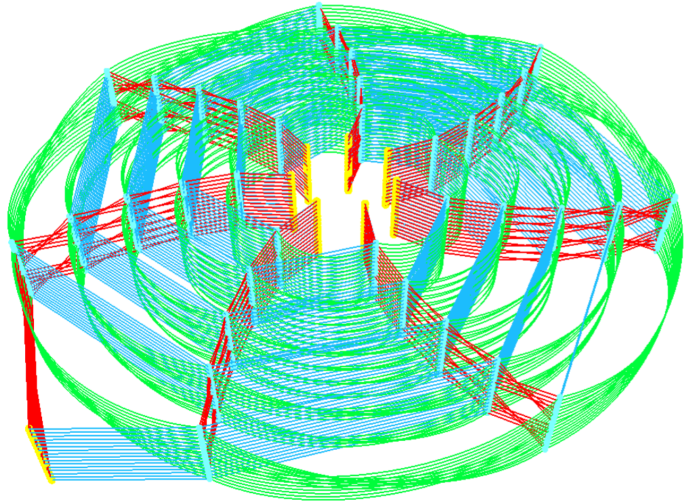
- Testbed Classes
  - System Technologies with future high impact factor to HPC
    - Computing (Throughput or Latency Optimized Computing , Processing Near Memory)
    - Memory (Phase-Change, Spin Transfer Memory, Memristor, 3D DRAM, 3D Flash)
    - Networks (*Data Vortex*, *Mellanox*, Software Defined Networks)
    - Storage (high bandwidth/low latency buffers, fast container support *netCDF*/ *HDF5*, Software Defined Storage)
  - Novel Processing Paradigms beyond traditional computing paradigms
    - Neuromorphic Processing
    - Approximate Computing
    - Quantum Computing
    - Silicon Photonics
  - Emerging Technologies (Materials -> Devices) beyond Moore's Law transitioning towards rudimentary logic/ memory cell and circuit basic blocks
    - Superconductive Processing (Single Flux Quantum (SFQ), Adiabatic Quantum Computing (AQC))

Networks: Data Vortex

- Novel network technology based on topology fundamentals
  - Ensures high probability of contention-free transport
  - Enables very small packet (64-bit) transport
  - Analyzing network performance/power for applications of interest: high-dimensional PDE, FFTs, finely partitioned data-intensive applications
- Vortex Interface Controller (VIC)
  - FPGA + 32 MB static RAM
  - Preparation and packet buffering



The Data Vortex Switch Topology



"Exploring Data Vortex Network Architectures" R. Gioiosa, T. Warfel, J. Yin, A. Tumeo and D. Haglin, Hot Interconnects, Santa Clara, August 2016.

Networks and Storage: Scalability Cluster

- Advanced HPC scale out solutions with emphasis on network and storage exploration
- Aim to provide flexible test-bed to explore issues at scale
- Emulate large network within smaller node footprint
  - Enable large-scale issues to be explored (collectives, contention, topology)
  - Validation on larger network for modeling, and prediction at scale

Mellanox TECHNOLOGIES

- Exploring a flexible network reconfiguration (OCS) option
  - Dynamic topology reconfiguration (system book, job submission, intra-job also possible)

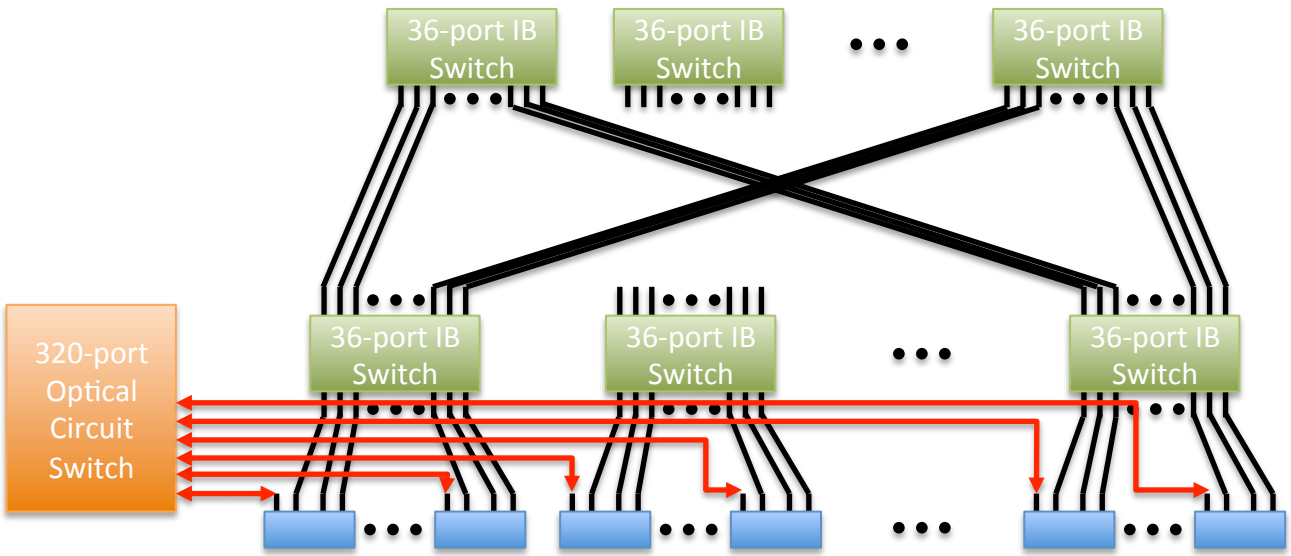
Fixed topology

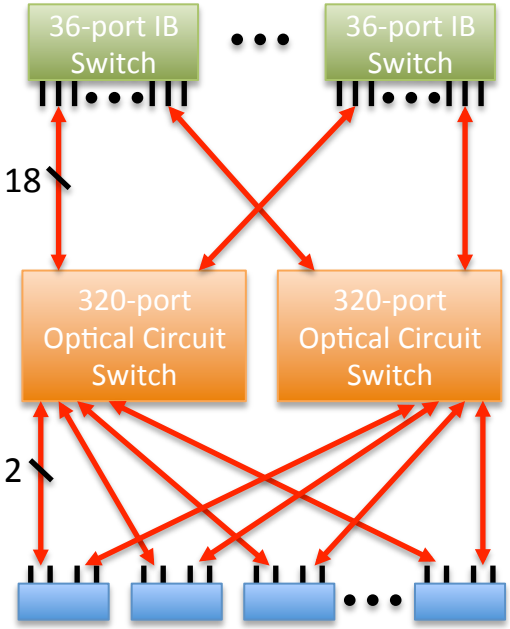
Flexible topology

- First 128 network end-points scalability cluster system expected fall 2016 (15-month Technology Refresh Cycle)

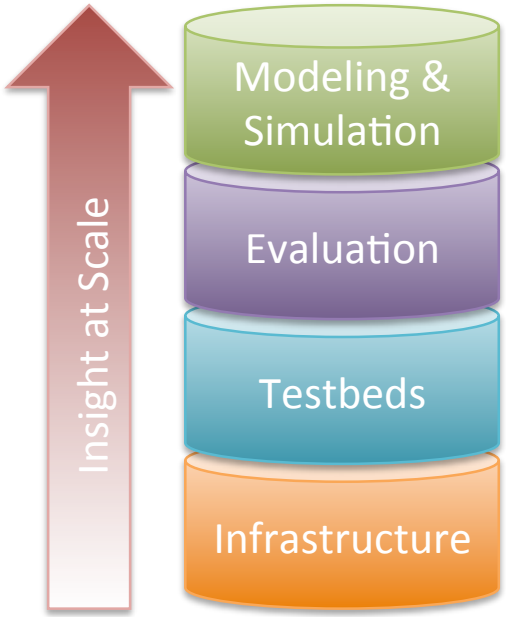
Infiniband + OCS

- Re-cabling of system in 50ms
- Possible configurations:
  - A) OCS to side of IB
  - B) OCS between nodes and IB





ModSim within CENATE

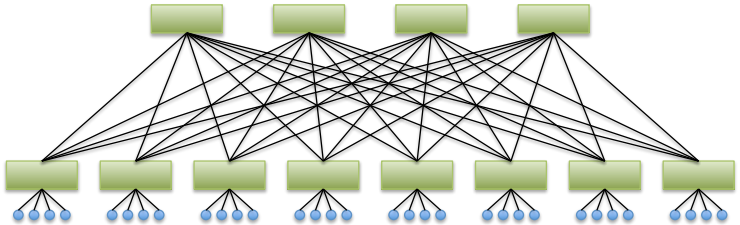


- Modeling and Simulation will be used to explore:
  - System scales that cannot be directly measured
  - Systems integrating disparate technologies
  - Multiple alternative system configurations
- Quantify trade-offs between multiple metrics of interest:
  - Performance
  - Power and energy consumption
  - Impact of thermal variation, faults, and fault mitigation
- Modeling builds on the CENATE foundation:
  - Application-centric models are derived from workload applications
  - Models are parameterized using measurements taken on instrumented testbeds (micro-benchmarks isolate “atomic” performance characteristics)
  - Models are validated at small-scale
- Key contribution of modeling is insight:
  - Rapid turnaround from system specification to performance quantification
  - Issues in performance can be traced to root causes
  - Quantify interplay between application characteristics and system

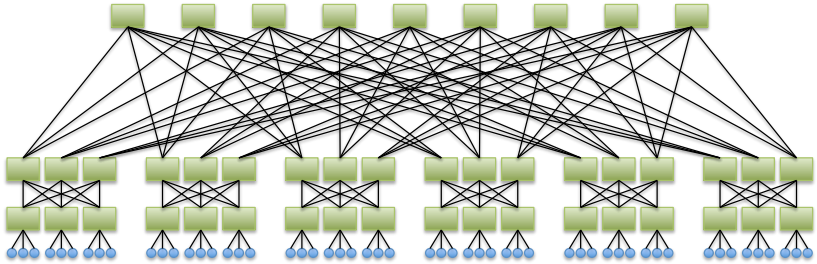
CENATE Modeling @ Scale: Network Analysis

- Model impact of network topology on communication performance and energy
- Explore mainstream networks considered in DOE *Forward programs

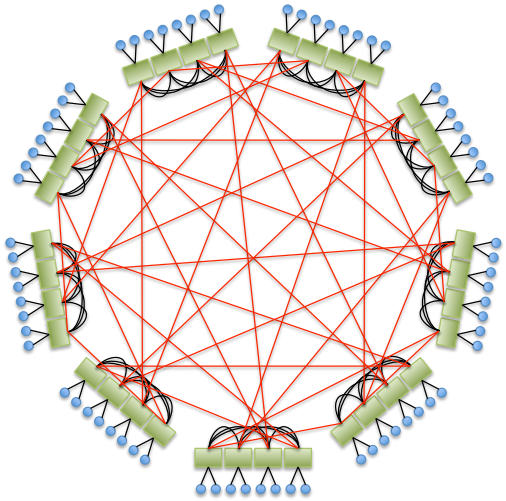
2-Level Fat-tree R=8



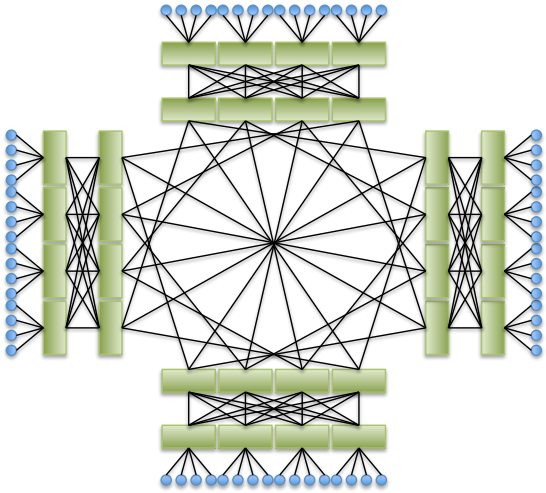
3-Level Fat-tree R=6



Dragonfly R=8



Dragonfly+ R=8

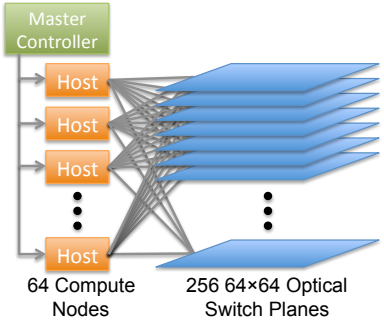


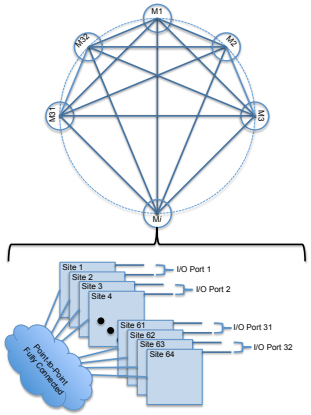
**** Note: Examples shown with small-radix switches.**

Modeling Possible Future Silicon Photonics Networks

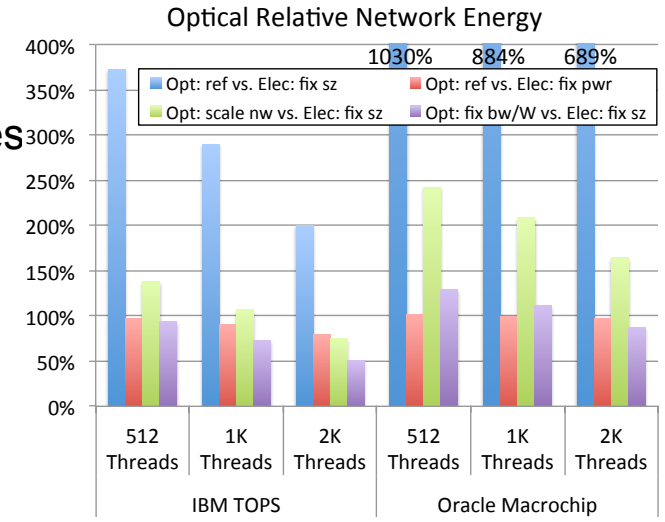
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Disparate technologies from IBM (internode) and Oracle (intranode)
- Modeling enabled:
  - Possible "marriage" options to be explored overcoming separation barriers
  - Quantified advantages over expected future electrical networks
  - Analyzed in the context of key graph analytic applications





- Oracle Macrochip intra-node network
  - 64 compute/memory sites fully connected
  - 2 GB/s per site pair (128 GB total)
  - 32 ports I/O per macrochip (for internode)
- IBM TOPS inter-node network
  - 64 node system
  - 256x 64x64 optical switch planes
    - 16 wavelengths per fiber
    - 20 GB/s BW per wavelength
- Improvement due to:
  - Improved link bandwidth
  - Greater link *concurrency*
  - Varied topological routing



CENATE: Center for Advanced Technology Evaluation

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Empirical evaluation of current “interesting” technologies
  - From device to system level
- Predictive evaluation of possible systems
  - Scalability: from empirical evaluations at small scale
  - Future technologies
  - allow for the “virtual” integration of technologies (e.g. Silicon Photonics, DV +HMC)
- Application centric
- Network testbeds:
  - Data Vortex: One-sided MPI over DV ??
  - IB + OCS Cluster: handling dynamic topology changes (inter- or intra- job) ??
  - Silicon Photonics: Exploring potential (performance / Energy)