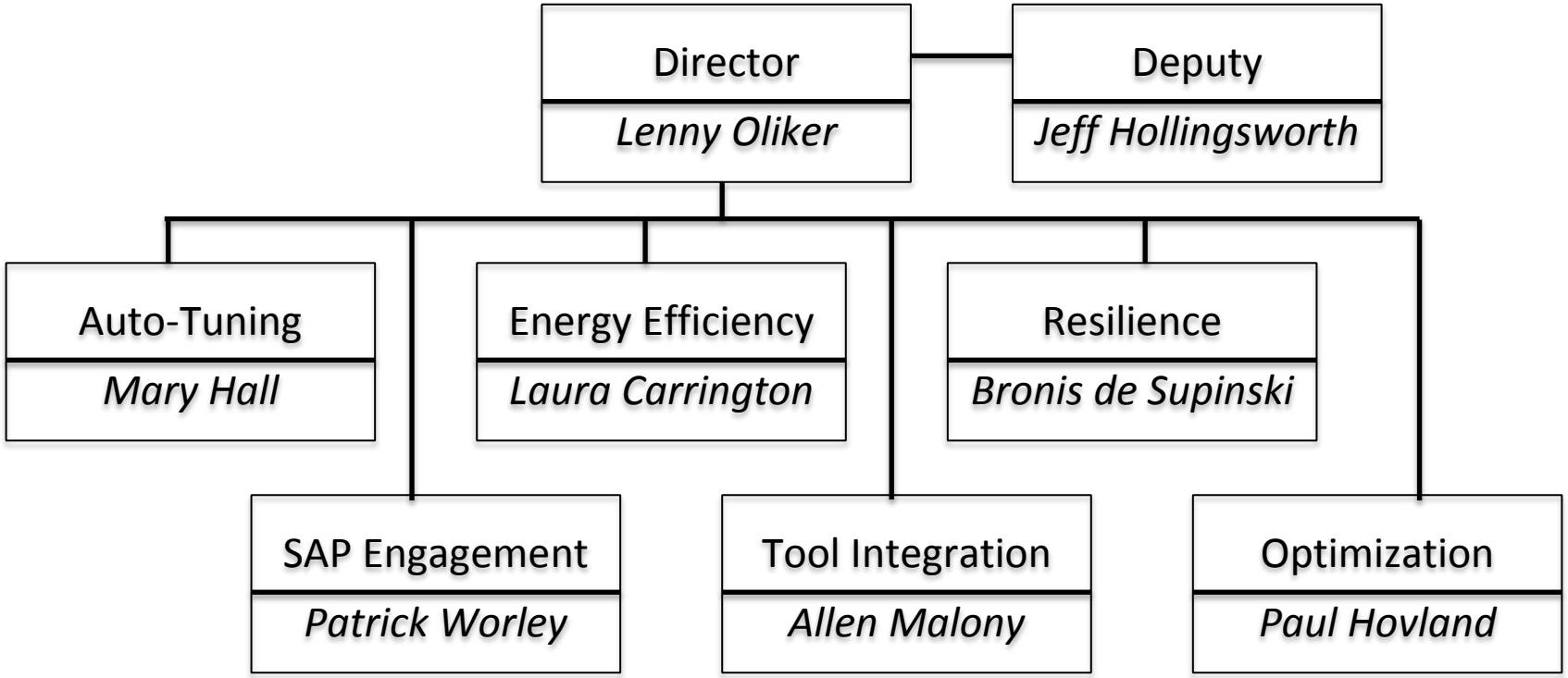
SUPER: Institute for Sustained Performance, Energy, and Resilience

Lenny Oliker
Lawrence Berkeley National Laboratory

Support for this work was provided through the Scientific Discovery through Advanced Computing (SciDAC) program funded by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research

INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

SUPER Organization Change



SUPER is large project and leadership is distributed across the PIs

SUPER's Engagement

- **Application Engagement**
  - SUPER engaging with 14 application partnerships
  - This talk will focus on 5 of those partnerships
  - Highlight MPAS-Ocean – collaborative effort, enable optimization while driving SUPER research
  - GEANT4 supplement – HEP simulation toolkit for passage of particles through matter
- **Architecture Awareness**
  - Application partnerships to exploit CPU/BGQ-based and MIC/GPU-accelerated DOE systems
  - Tools (ROSE, PAPI, TAU, CHiLL, Pbound, ORIO, ActiveHarmony)
  - Roofline to characterize CPU and GPU-based computing systems
  - GreenQueue to maximize energy efficiency
  - Resilience for soft-error detection & mixed solutions
- **Institute Engagement**
  - Roofline collaboration with FastMath
  - NUCLEI SpMV optimizations with FastMath
  - Linking Performance to Scientific Visualization with SDAV
  - Model-based I/O Optimization with SDAV

Application Engagement

SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

Participation in 14 SciDAC-3 Application Partnerships

BER Applying Computationally Efficient Schemes for BioGeochemical Cycles
BER MultiScale Methods for Accurate, Efficient, and Scale-Aware Models of the Earth Sys.
BER Predicting Ice Sheet and Climate Evolution at Extreme Scales

BES Developing Advanced Methods for Excited State Chemistry in the NWChem S/W Suite
BES Optimizing Superconductor Transport Properties through Large-scale Simulation
BES Simulating the Generation, Evolution and Fate of Electronic Excitations in Molecular and Nanoscale Materials with First Principles Methods

FES Advanced Tokamak Modeling
FES Partnership for Edge Plasma Simulation
FES Plasma Surface Interactions

HEP Community Petascale Project for Accelerator Science and Simulation

NP A MultiScale Approach to Nuclear Structure and Reactions
NP Computing Properties of Hadrons, Nuclei and Nuclear Matter from QCD
NP Nuclear Computational Low Energy Initiative

NNSA ParaDIS: Parallel Dislocation Simulator

Supplement (Non-SciDAC) HEP GEANT4 simulation toolkit for passage of particles through matter

Linking Performance MPAS-Ocean Data into Scientific Visualization Tools
SDAV/SUPER Collaboration (BER/ASCR Multiscale PI: Bill Collins)

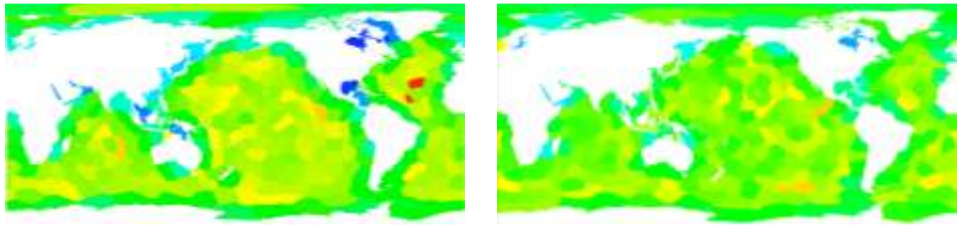
Objectives

- Model for Prediction Across Scale (SCVT)
- Map TAU performance measurements to the MPAS-Ocean spatial domain to assist in optimization of partition strategies

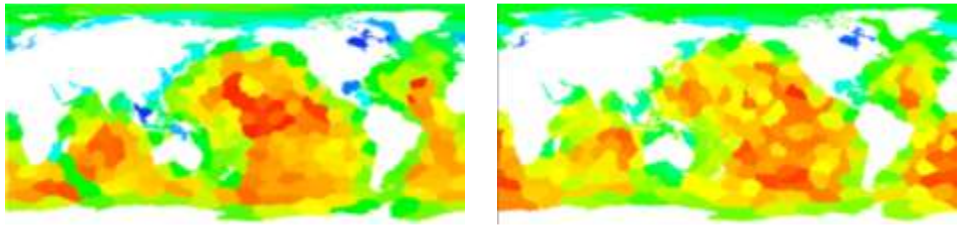
Impact

- Integrated TAU performance measurement data with application scientific data in VisIt
- Reduced execution time up to 15% for 60km model on 256 processes

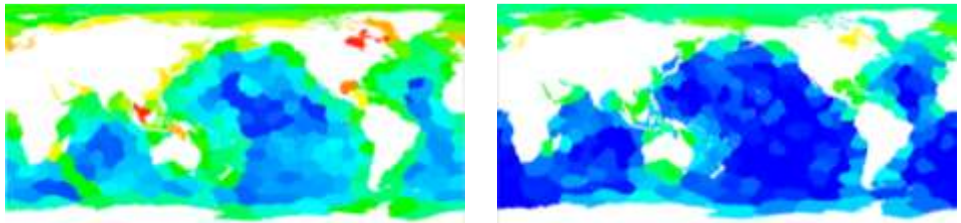
Original Partitions | Refined Partitions



min: **473**, max: **846** ***Total Cells per Block*** min: 535, max: **771**



min: **83**s, max: **250**s ***Computation Time*** min: 98s, max: **240**s



min: **27**s, max: **190**s ***MPI_Wait Time*** min: **9**s, max: **150**s

Progress & Accomplishments

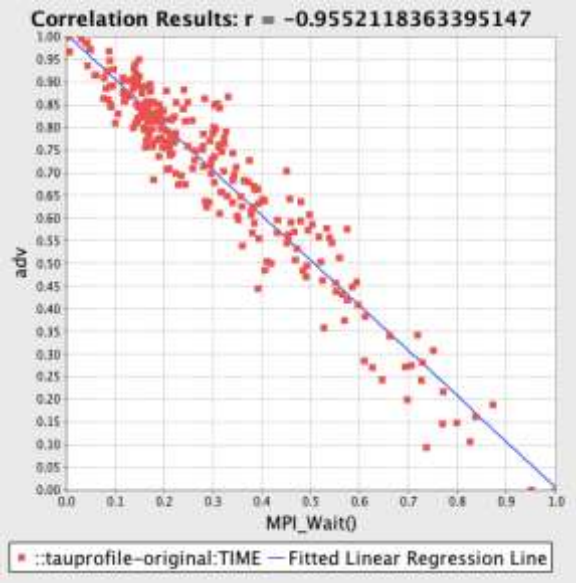
- **Demonstrated that the load imbalance problem is correlated with variability among partition block size due to relatively large halo regions**
- **Visualizations also show that vertical depth, coastlines and number of neighbors affect computation, communication times**
- **Hindsight partition refinement using block+halo weights reduced mean MPI_Wait times by 40%, and overall execution time up to 15%**
- **Huck et al., "*Linking Performance Data into Scientific Visualization Tools*", Visual Performance Analytics at SC'14**

Results collected on Edison@NERSC using 60km data and 240 processes

TAU Measurement of MPAS Load Imbalance

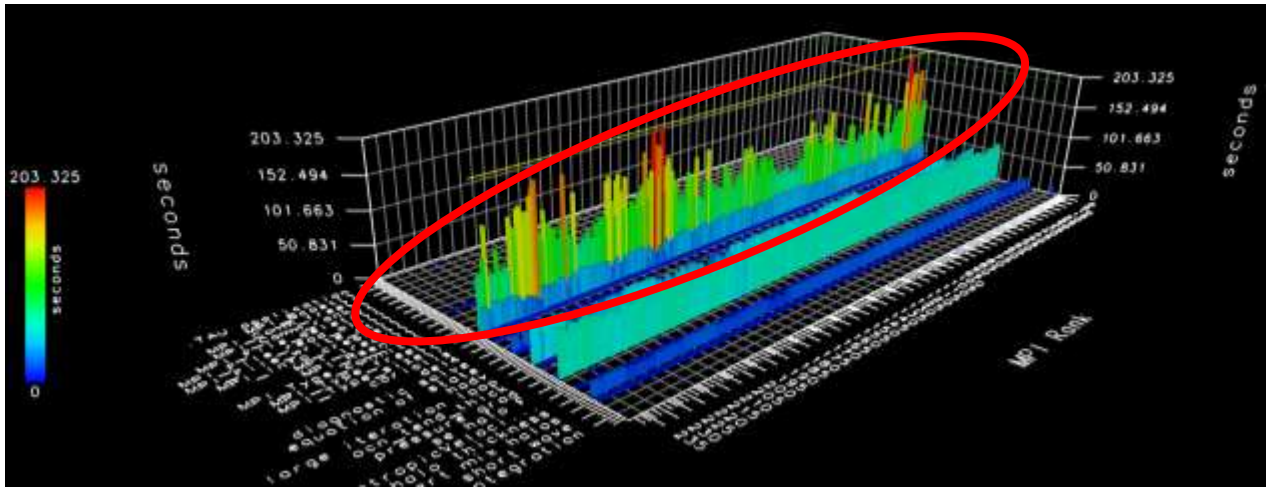
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

- Unusual stencil problem: 3 layers of halo/ghost cells lead to imbalance
- *Performance measurement in isolation is not enough*
  - Need to combine:
  - Performance measurements
  - Application metadata
  - Correlations between measurements and metadata
  - Visualization of performance data in application domain



TAU Measurement / Analysis Approach:

- MPAS application instrumentation modified to use TAU_start/stop, link with TAU library
- Use applicable MPI wrappers, OpenMP introspection and PAPI support
- Data properties instrumented / captured as TAU metadata at runtime
- Results archived in TAUdb database
- PerfExplorer script to analyze and extract data from TAU profiles
- ParaProf, PerfExplorer, VisIt visualizations



Performance Optimizations for MPAS-Ocean

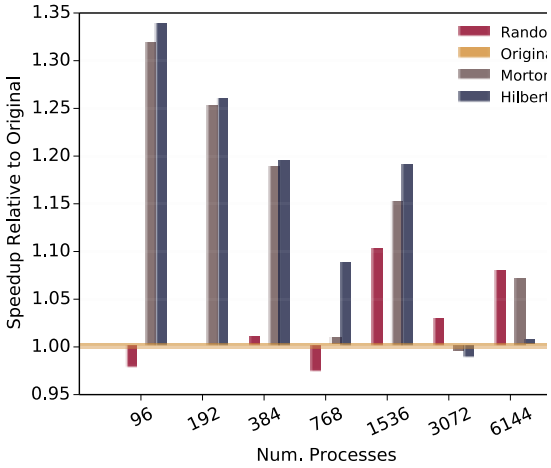
Objectives:

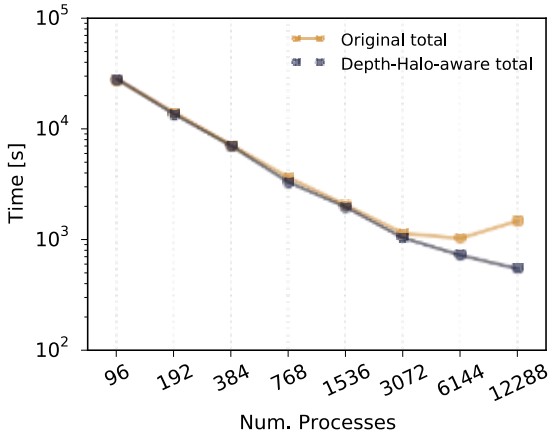
- Accelerate MPAS-Ocean code performance on state-of-the-art supercomputers
- Prepare MPAS-Ocean for transition to next generation highly parallel architectures
- Work collaboratively with SUPER Institute to leverage performance optimization tools and expertise

Impact:

- Demonstrated that space-filling curve based ordering essential for intra-node communication reduction
- Developed partitioning optimization approaches broadly applicable to numerous of unstructured-mesh based computations
- Allow higher Simulated-Years-Per-Day throughput for ocean modeling simulations with MPAS-Ocean.

(Left) Low concurrency speedup via mesh reordering on Edison
(Right) High concurrency speedup via hypergraph partitioning





Progress and Accomplishments:

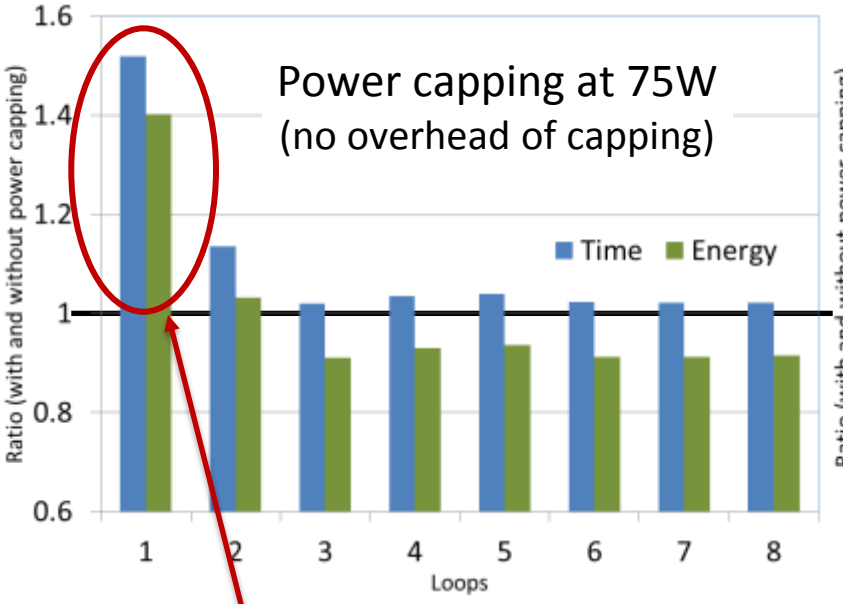
- **SFC-based mesh data reordering: average 1.25x performance gain**
- **New halo-aware hypergraph partitioning algorithm improves scalability at high concurrency by over 2x.**
- **Combined with SUPER optimizations, including pointer reduction at 12K cores: 3-4x MPAS overall speedup**
- **A. Sarje et al. MSESM/ICCS 2015 nominated best paper**

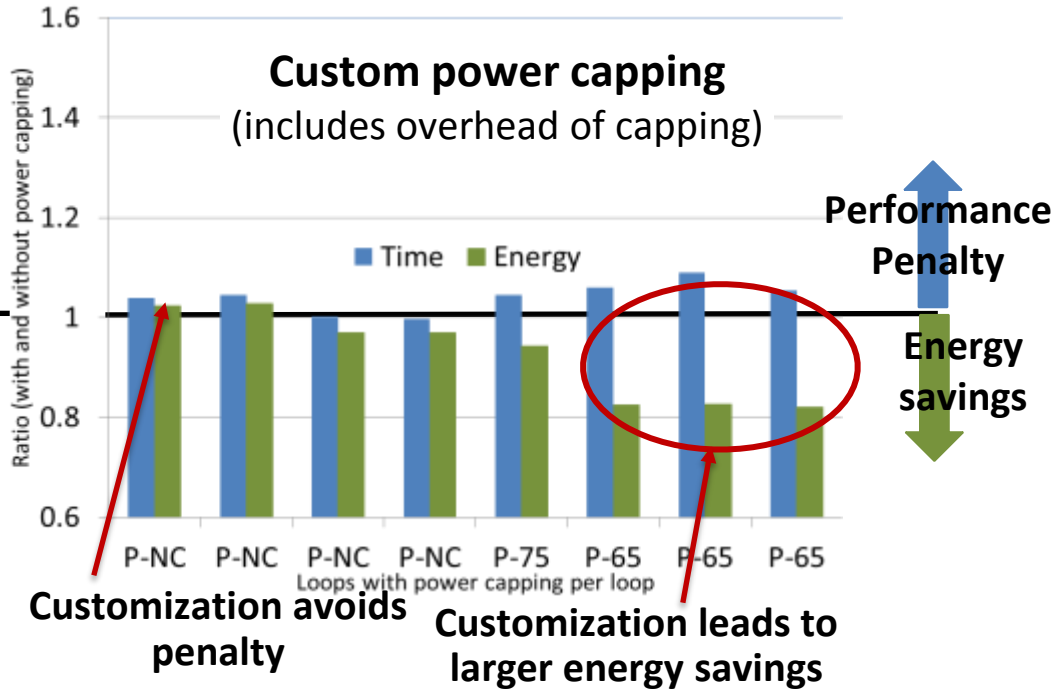
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

Green Queue: Customize Loop-level Power Capping in MPAS-O for Energy Efficiency

Tools & models to automate custome loop-level power capping for better energy efficiency.

- Automated identification of hotspot loops in large-scale application
- Identify energy-optimal power capping level vis modeling the impact of power changes
- Automate power-capping per loop/function throughout large-scale application



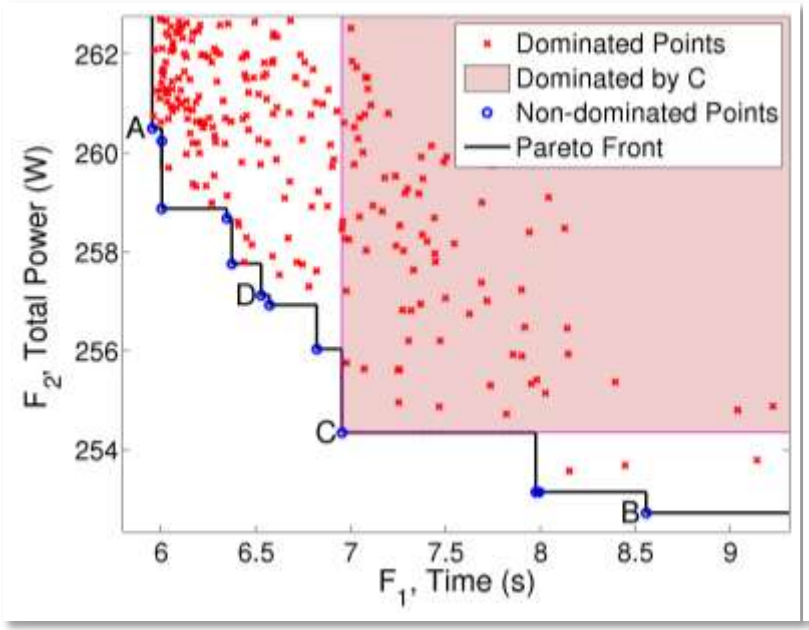


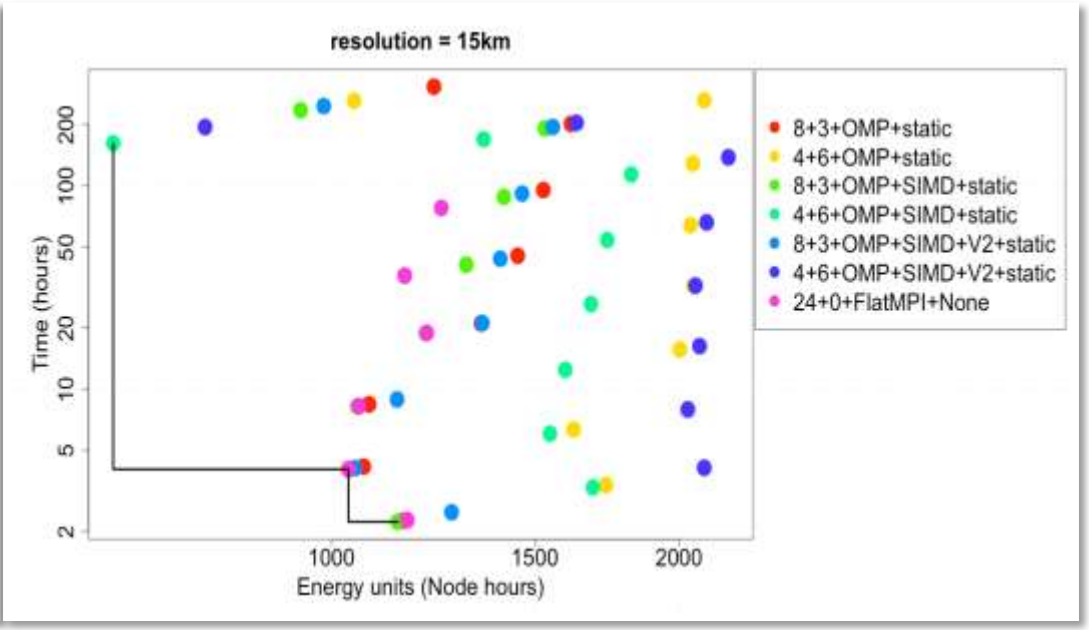
Customize loop-level power capping for MPAS

MPAS: Ocean modeling application 20-day simulation on 512 cores

Customized loop-level power capping avoids performance penalties while reducing power & energy costs.

MPAS-Ocean Multi-Objective Optimization





- Multiple objectives: execution time, energy consumption, resilience to errors, power demands, and memory footprint, etc.
- Multi-objective optimization problem with relative weights/constraints are not known at search time
- Pareto optimality: non dominating points ABCD (left fig.)
- MPAS-Ocean - Node hours as a proxy for energy (and utilization) on **Edison**
- For 15km, **Pareto points** from **3 different configurations**, due to better-than-ideal scaling

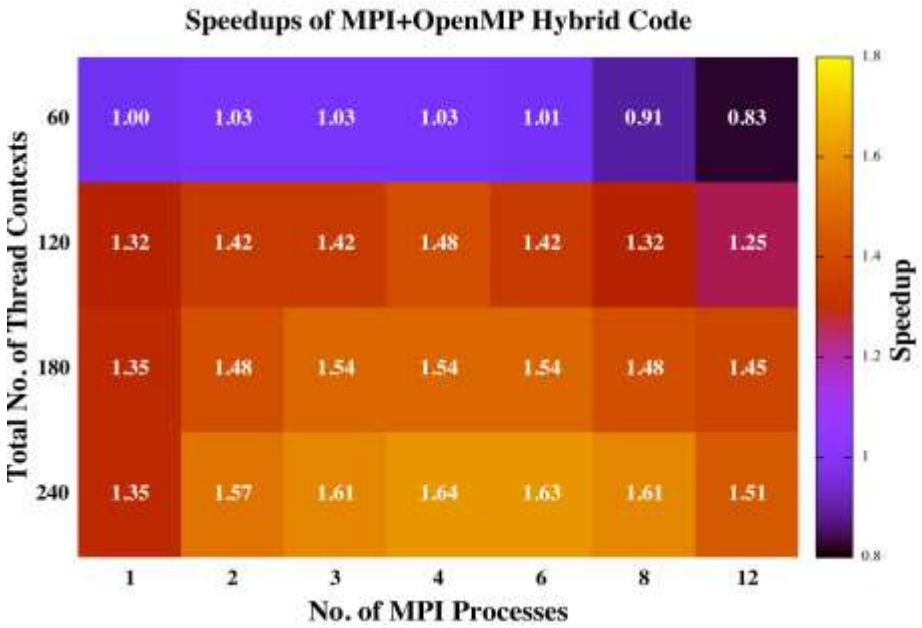
65x and 1.6x Speedup for CCSD and Fock matrix for NWChem
PI: Chris Cramer BES/ASCR Collaboration

Objectives

- Accelerate NWChem performance by implementing thread-level parallelism on the Intel Phi many-core architecture.

Impact

- Faster time-to-solution enabling larger and more accurate excited-state simulations with NWChem.



Performance improvement of hybrid MPI+OpenMP over flat MPI code for Fock matrix construction on Intel MIC architecture. The hybrid code exploits all 240 hardware thread contexts on card

Progress & Accomplishments

SUPER Institute collaboration implemented OpenMP parallelism for two NWChem modules

- **Native mode optimization to prepare for next-generation NERSC8 Cori**
- **Threading is essential to exploit full capability of MIC architecture**

Performance of triples part of CCSD(T) improved 65x over original flat MPI implementation

- **Flat MPI constrained to single process because of memory limitation**

Performance of Fock matrix construction improved 1.64x over original flat MPI

- **Flat MPI constrained to 60 MPI processes**
- **Presented at SC15 workshop & NERSC Tutorial**

Up to 40% Performance Improvement from New Load Balancing Scheme

PI: C.S. Chang, Fusion SciDAC Center for Edge Physics Simulation (EPSi)

Objectives

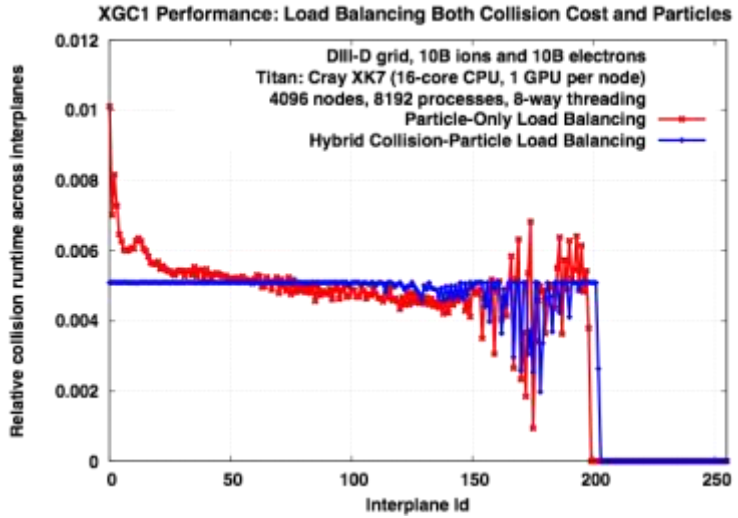
ASCR–FES SciDAC Highlight: Two-way collaboration among EPSi and SUPER

- XCG1: Gyrokinetic particle-in-cell code designed for simulating edge plasmas in tokamaks
- Address performance degradation due to load imbalance in nonlinear collision calculation for XGC1 on DOE Leadership Computing Systems.

Impact

- Low overhead automatic adjustment of parallel decomposition improves computational performance robustly and with minimal user input.

Accomplishment highlight



Example load imbalance in collision operator cost, comparing load balancing only particle distribution with also load balancing collision cost. Cost is summed over rows of virtual 2D processor grid. Full model performance improvement is 30% for this example.

Progress and Accomplishments:

Challenge

- **Existing particle load balancing algorithm does not also balance distribution of collision cost in parallel decomposition.**
- **Both particle count and collision cost per grid cell distributions evolve with the simulation.**

Solution

- **Two level optimization strategy: (a) balance collision cost subject to constraint on particle load imbalance, (b) optimize XGC1 performance by varying constraint periodically, converging to the optimum if distributions are static and adapting to the changing distributions otherwise.**

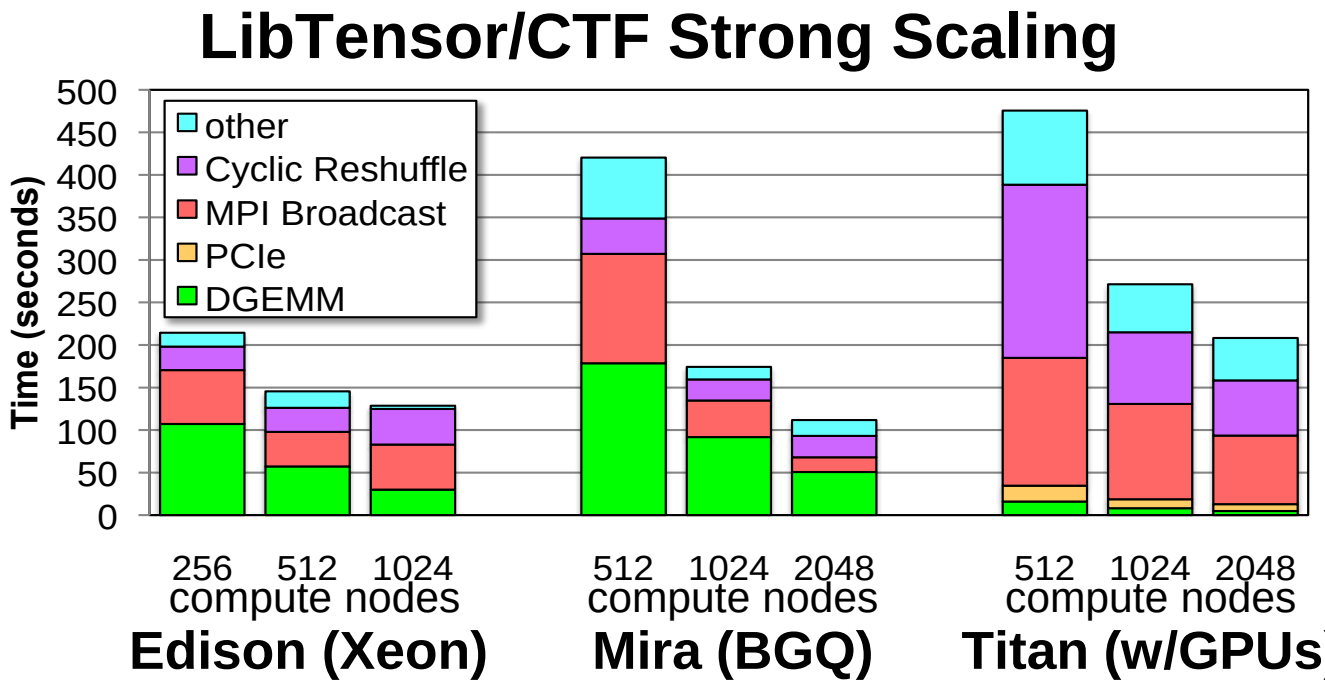
Result

- **<u>10%-40% improvement for production runs</u>.**

INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

LibTensor / CTF Optimization

- Electronic Excitations in Molecular and Nanoscale Materials BES **(PI: Martin Head-Gordon)**
- Originally restricted to running on large SMPs with spinning disks (bad match DOE HPC)
- Our optimized version leverages **Cyclops Tensor Framework (CTF)** developed at UCB
- Cyclic distribution of tensor via MPI All-to-All (stresses network) & SUMMA (Bcast & BLAS)
- GPU-acceleration **reduced DGEMM to 3% of the run time**.
- Mira had slowest DGEMM, but fastest Broadcast & Reshuffle: best runtime & scalability
- **New version: 2K Mira nodes is 150x faster vs original large-scale SMP**



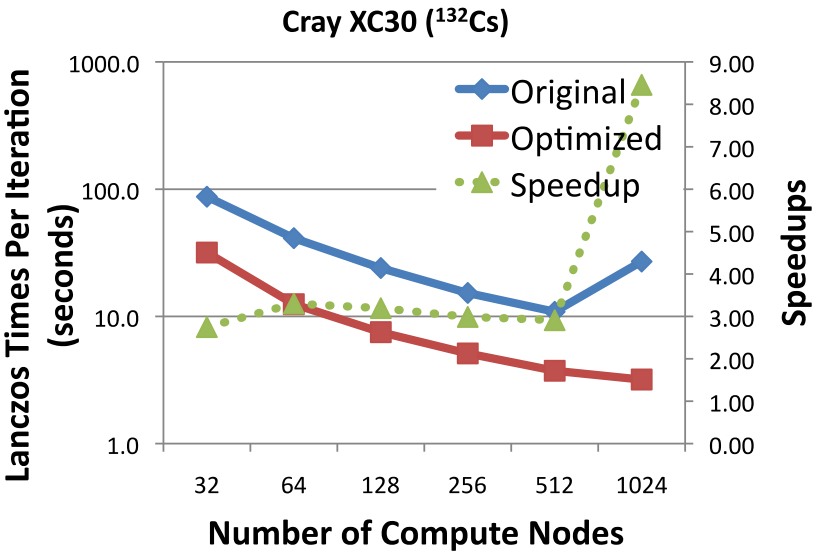
Methylated uracil water dimer test problem with 302 basis functions and C_s symmetry

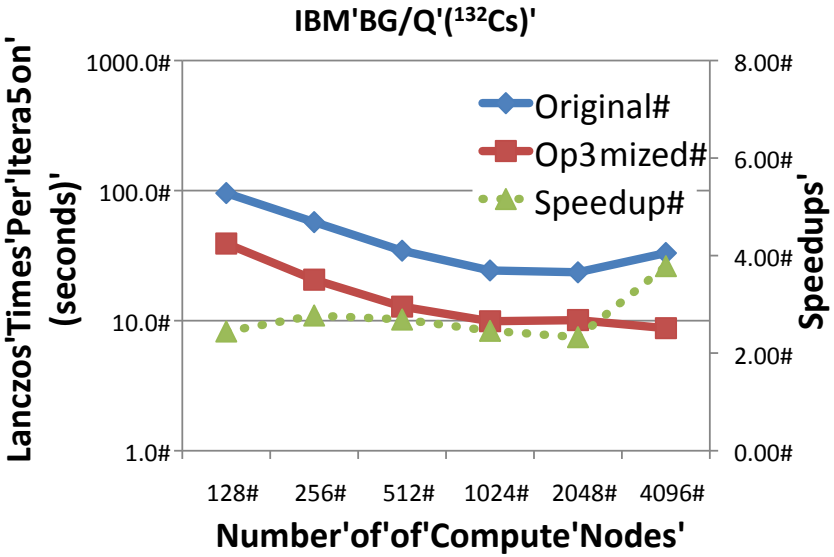
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

BIGSTICK Optimization

- CalLat NP SciDAC Partnership (PI: Wick Haxton)
- Configuration-interaction (CI) popular technique for solving quantum many-body systems
  - BIGSTICK: scalable, memory-efficient CI code, large eigenpair problem via iterative methods
  - Series of tables used to **compute non-zeros on the fly**, **reduces memory from 100TB to 0.5TB**
  - Challenges: variable non-zero performance, load imbalance due to complex decomposition
- Applied both intra- and inter-node optimizations
  (1) Empirical load balance for OpBundles nonzeros (2) Tuning MPI+OpenMP for each system
  (3) Fusing collectives (4) empirically-tuned reduced concurrency for Lanczos reorthogonalization

Overall, improved scalability and performance by 1.3x – 8x (Accepted SC15)





^{132}Xe using a ^{100}Sn frozen core with 4 valence protons and 8 valence neutrons

INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

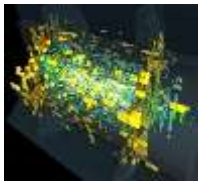
GEANT 4 Enhancements

GEANT4 library used to build simulators for the interactions between physics particles and materials.

- Medical imaging/therapy, design of radiation shielding, Monte Carlo simulation of high energy physics detectors at CERN.
- Validate/calibrate/compensate detectors
- Well-designed, 20 year old, C++ code.
- One estimate LHC costs: $100M/year (1% improvement worth $1 M/year.)

SUPER activities.

- Performance analysis of GEANT4 toolkit and LHC applications.
- Reorganize loops for vectorization
- Prototype and study automated GPU transformations
- Identify particle-material (XC) cross section calculations as a significant "warm region" (10 to 22% of time)



Cross section (XC) improvements.

- Cache one recent result for each (particle, material, process, energy) → ~2% speedup
- XC of complex material is the weighted sum of XCs for every isotope in the material.
  - 50 (particle, material, process) triples account for >90% of XC costs.

Fast Path for Hadronic Process XCs

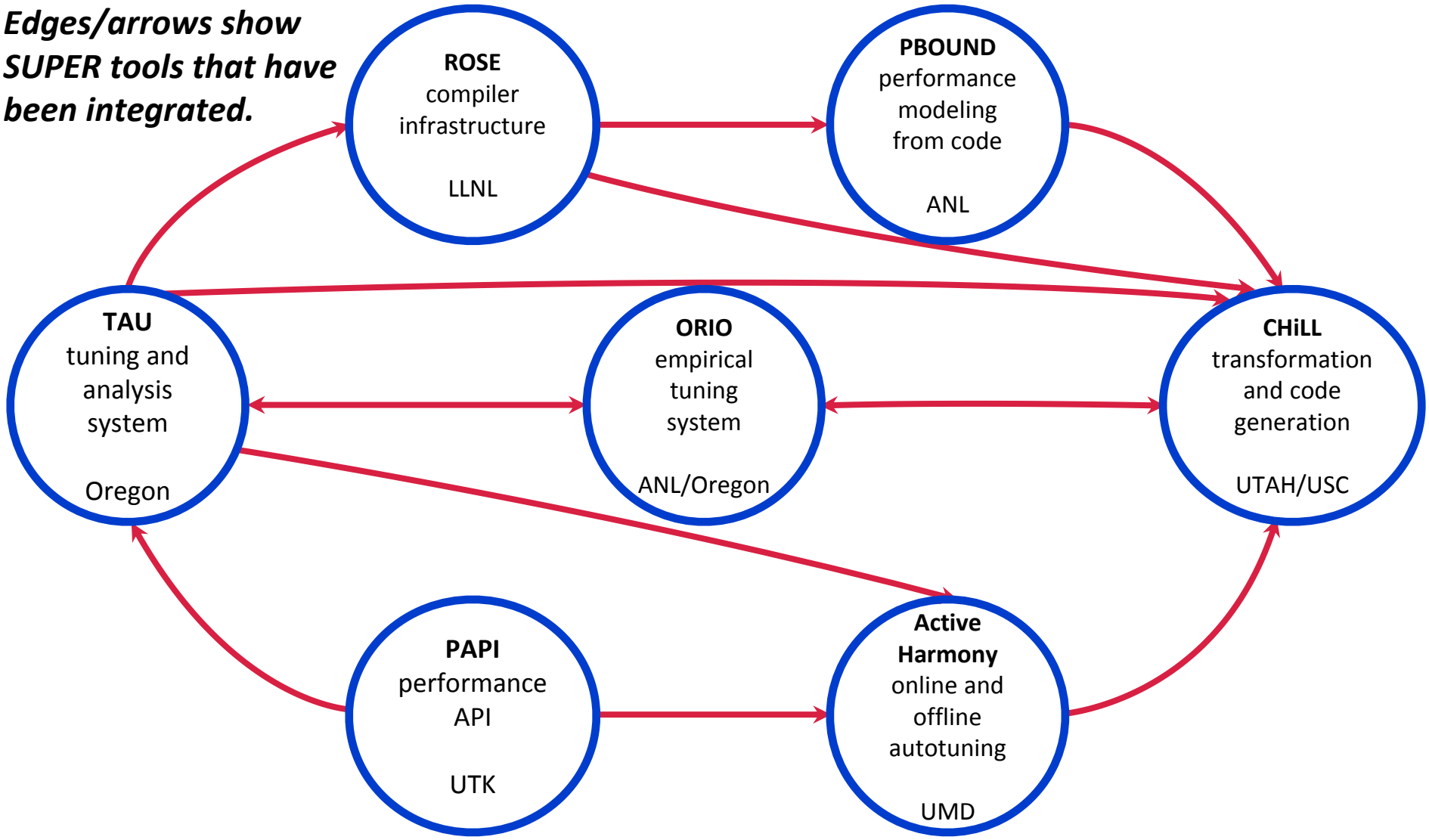
- Over-sample curves for common triples.
- Create piecewise linear approx within *tol* of originals and preserves area(s)
- Fast path to determine interactions
- Average XC comp reduced by 5.8X
- Additional 3% speedup on a test input for CMS.

Architecture Awareness

SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

SUPER Tool Integration

Edges/arrows show SUPER tools that have been integrated.



Commercial tools are conservative, while SUPER can afford to be aggressive (autotune etc)

SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

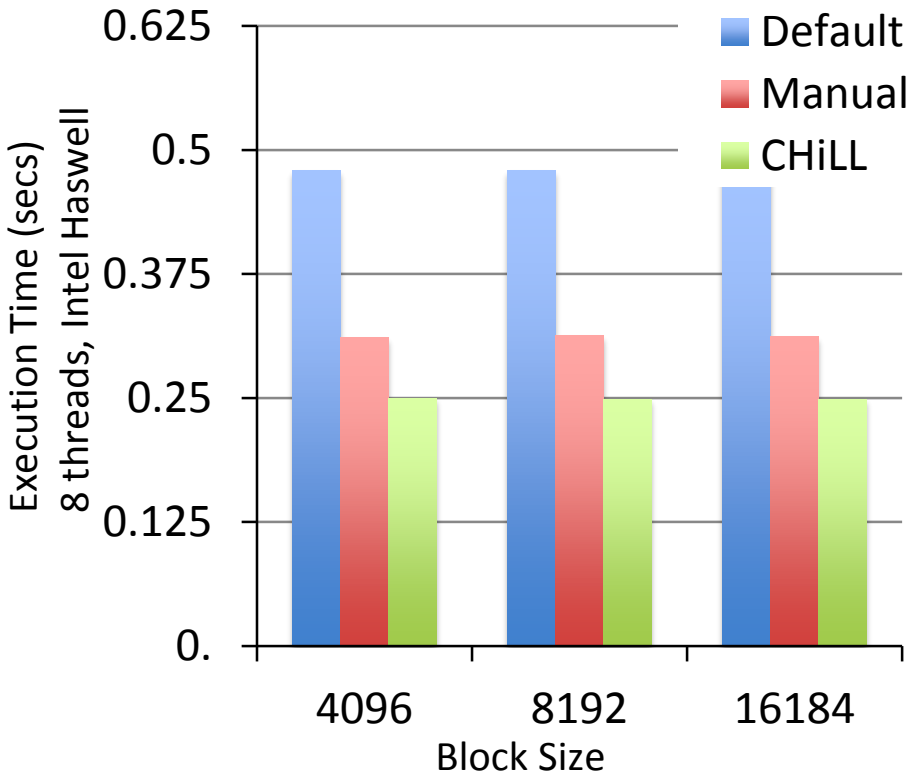
State-of-the-Art Tools Applied to NUCLEI SciDAC App

CHiLL autotuning compiler

- Transformations support sparse matrix computations and matrix format conversion
- Automatically-generated inspector, converts matrix format
- Composes with other compiler transformations for optimized code

NUCLEI Computation

- SpMV for very large sparse symmetric matrix
- Only store triangular portion and generate remaining matrix
- Compressed sparse block (CSB) representation exploits symmetry
- FastMath collaboration



Compiler-generated code has less indirection, outperforms fast, manually-tuned version

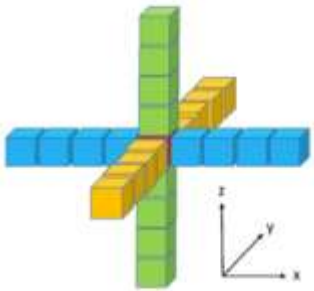
SciDAC unique capability enabled research collaboration from application to computational science (FastMath) to automation via computer science (SUPER)

Enhancing Resilience:

Error Detector Synthesis to Capture Silent Data Corruption

- Developed Customized Solutions for Soft Error Detection in Stencil Computations
- Example: Given a 25-point RTM Stencil Function (used in Geo-Science Applications)
  - Synthesize approximations of this function using fewer spatial / temporal points
  - Synthesis accomplished through Linear Regression (Machine Learning)
  - Method is general for PDE Stencil Codes
- Learned Functions Serve as Error Detectors (outlier)
  - Release of SORREL : Platform for Machine-Learning based Error Detector Synthesis
  - Inject bit-flips using our LLVM-level fault injector
- Error Detectors Customizable for Overhead / Detection Ratio
  - Can Achieve 88% error detection for 32% overhead (more probes -> increased overhead)
  - Future Work : Better methods for overhead reduction and amortization

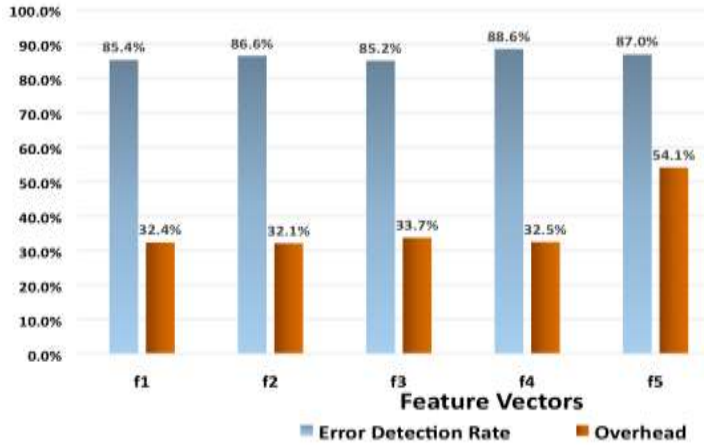
RTM Stencil



Partial-Differential Equation (PDE) used by RTM Algorithm

$$\frac{\partial^2 P}{\partial n^2} = v^2 \left(\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} + \frac{\partial^2 P}{\partial z^2} \right)$$

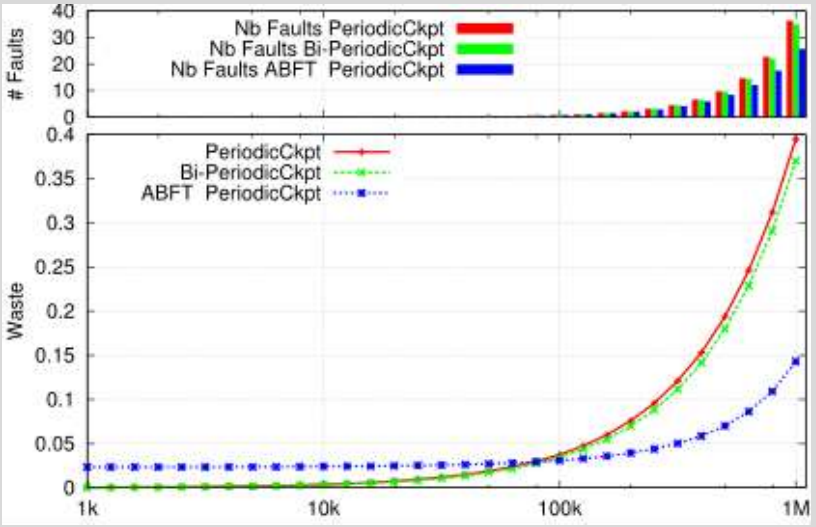
RTM Kernel ; Finite Difference Discretization
(2nd Order Accurate in Time, 8th Order Accurate in Space)



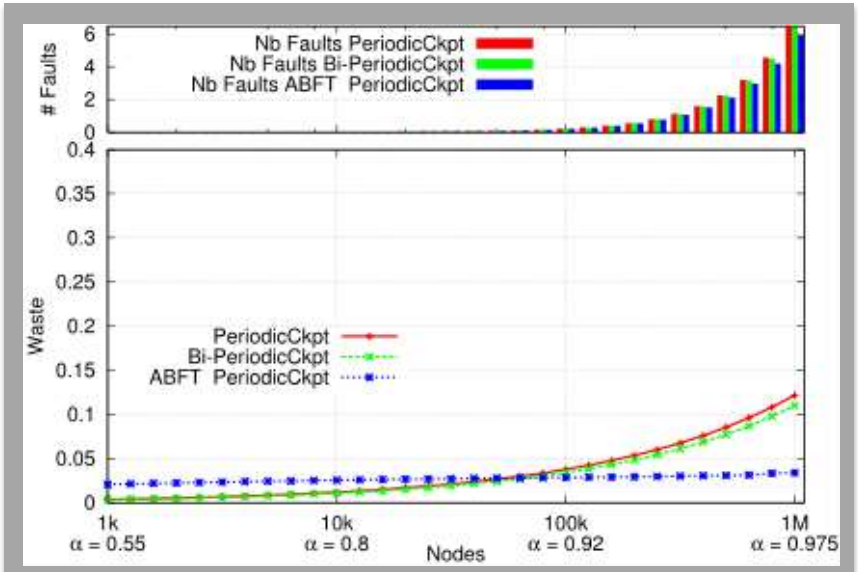
SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

Balancing Resilience Alternatives to Address Reliability Wall

- Comparing fault management schemes for iterative application using specialized numerical libraries (ScaLAPACK, PETSC, etc) and external non-library components
- Uses Algorithm Based Fault Tolerance (ABFT): adapting the algorithm so that the application dataset can be recomputed without costly checkpoint/restart (C/R)
- Method used composite protocol to combine traditional C/R together with ABFT



C (=R) is in O(n)



C (=R) independent of n (O(1))

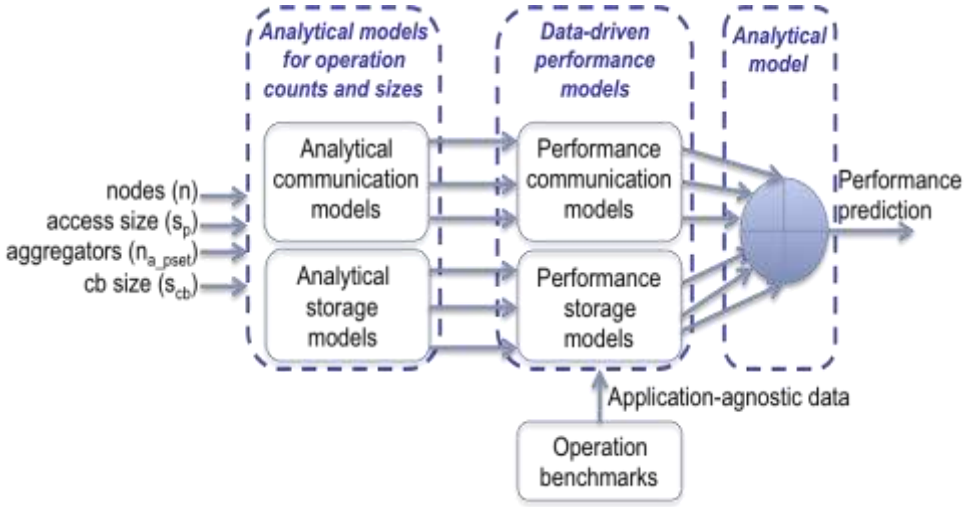
Evolutionary I/O system: up to 40% overhead using traditional C/R, only 15% overhead using our C/R + ABFT approach

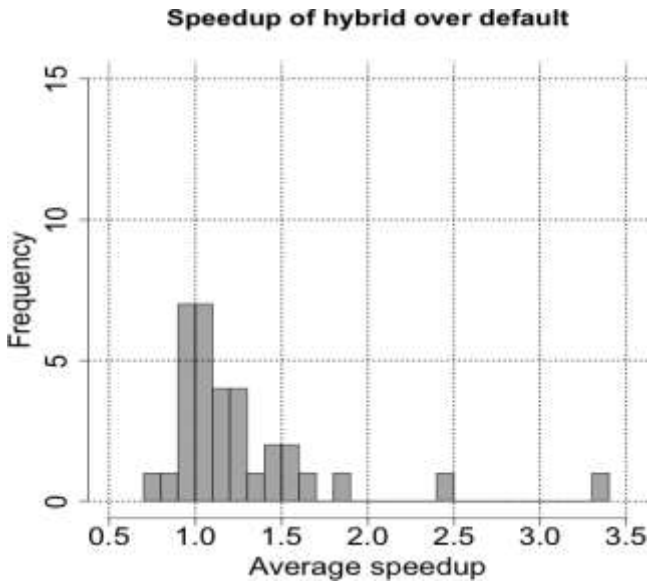
Revolutionary NVRAM I/O: C/R remains constant, limited to 15%. Our C/R + ABFT is only 3% overhead.

Institute Engagement

SUPER/SDAV Collaboration: Improving Computational Science Throughput via Model-Based I/O Optimization

INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE





- Leverage ***SUPER multi-objective optimization capabilities with SDAV I/O*** expertise and models
- Model subsystems for performance, power, energy via hybrid approach: ***analytical and machine learning***
- Design ***model-based optimization*** & evaluate w/ exemplar SciDAC workloads, includes: architecture, system software and noise

- Quadratic models effective at predicting parallel I/O write times *in the absence of noise*
- Analytical models for communication + Empirical model for storage
- On IOR benchmark from ROMIO (used in C/R), model-based optimization obtains ***speedup from 1.5x to 2.5x over default*** [Cluster 2015a, Cluster 2015b]

The Roofline Toolkit
SUPER/FASTMath Collaboration

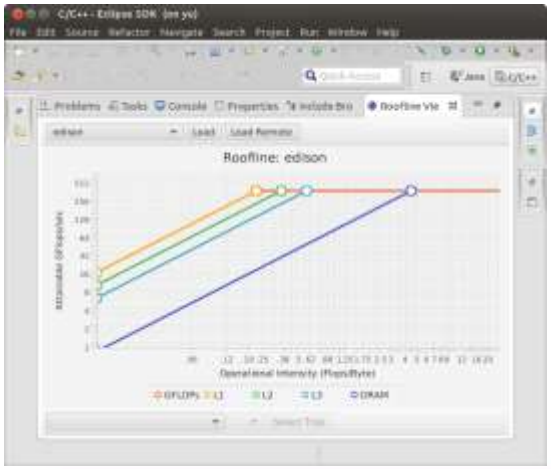
Objectives:

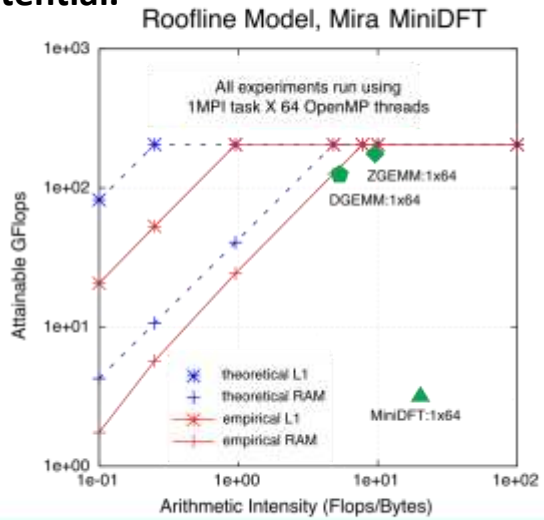
- Performance modeling is critical for identifying and ameliorating bottlenecks on emerging HPC systems
- Roofline performance model is a highly recognized approach for quantifying system and program behavior, but requires expert knowledge
- Goal: produce set of software tools allowing non-experts to automatically leverage Roofline modeling capabilities

Impact:

- Automated roofline code could be used to diagnose performance problems for DOE & SciDAC codes
- **The community can focus on addressing appropriate performance impediments via optimization, algorithm design, or hardware selection**
- Parameterized roofline models can be used to **predict behavior of future systems, to help drive forward-looking algorithms and architectures**

(Left) Automated Roofline Visualizer using ERT Edison data. (Right) Mira Roofline bounds relative to MiniDFT compact app highlighting performance potential.





Progress and Accomplishments:

- **Detailed roofline analysis of 4 leading HPCs: Edison, Mira, Babbage, and Titan**
- **Quantified benefits of emerging GPU software managed cache technologies *Y. Lo et al. PMBS2014***
- **Insights resulted in 03/2015 *public release* V1.0: Empirical Roofline Tool (LBNL) & Roofline Visualizer (U Oregon)**
- Roofline Toolkit is a community tool for automatic hardware introspection & analysis

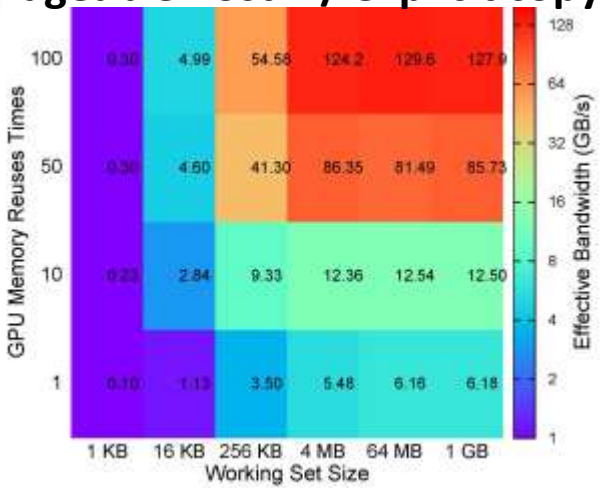
U.S. DEPARTMENT OF ENERGY | Office of Science

SUPER INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

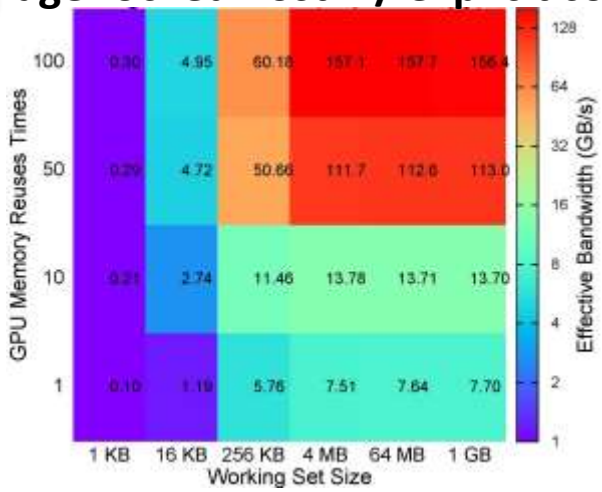
SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

Roofline for GPU Memory Performance

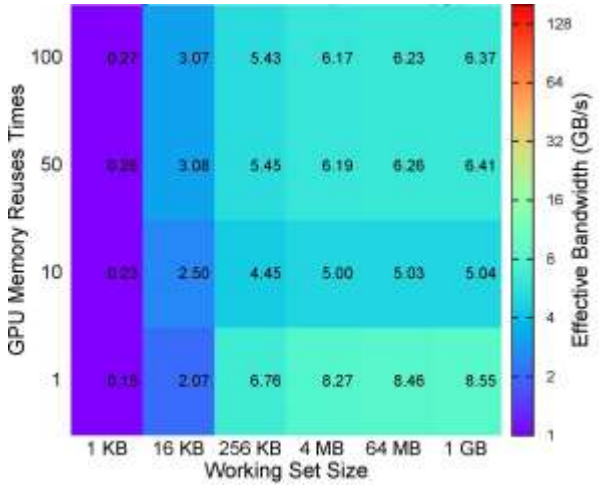
1 Pageable host w/ explicit copy



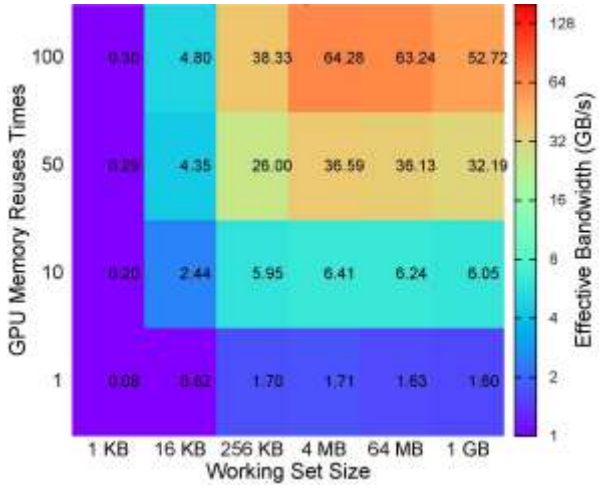
2 Page-locked host w/ explicit copy



3 Page-locked host w/ zero copy



4 Unified Memory



Bandwidth dramatically impacted by working set, reuse and memory access scheme.

Empirical study completed, GPU functionality expected in new Roofline release in 2015.

Summary

- **Performance engineering of scientific software**
  - Significant optimization progress for numerous SciDAC applications
  - Speedups: XGC1 10-40% on Titan, MPAS-O 3-4x on Edison, BIGSTICK 2-9x on BG/Q, LibTensor 150x on Edison, NWChem CCSD 65x & 1.6x Fock Matrix on MIC
- **Tool integration**
  - Significant progress in measurement/optimization tool integration & co-design
- **Energy minimization**
  - Fine grained power capping for energy savings while maintaining performance
- **Multi-objective optimization**
  - Coarse grained auto-tuning of energy and performance tradeoffs
- **Automatic performance tuning**
  - Chill automated NUCLEI tuning outperforms optimized hand-tuned version
- **Resilient computing**
  - Customized solutions for error detection & modeling mixed resilient solutions
- **Inter-institute collaboration**
  - Perform visualization, I/O optimization, Empirical Roofline Toolkit, NUCLEI/SpMV tuning

EXTRA SLIDES

SUPER
INSTITUTE FOR SUSTAINED PERFORMANCE, ENERGY, AND RESILIENCE

SUPER Performance Tools

Commercial Tools

- Target general-purpose workloads (FP↓, CF↑, RE↓, SY↑)
- Programmer productivity is primary driver
- Performance measurement focuses on execution time
- Optimization is conservative (static, architecture independent)
- Proven technology

SUPER Tools

- Target scientific simulation (FP↑, CF↓, RE↑, SY↓)
- Performance is primary driver
- Performance measurement extensive, pinpoints opportunities for improvement
- Optimization is aggressive (dynamic, autotuning, architecture specific)
- State-of-the-art technology