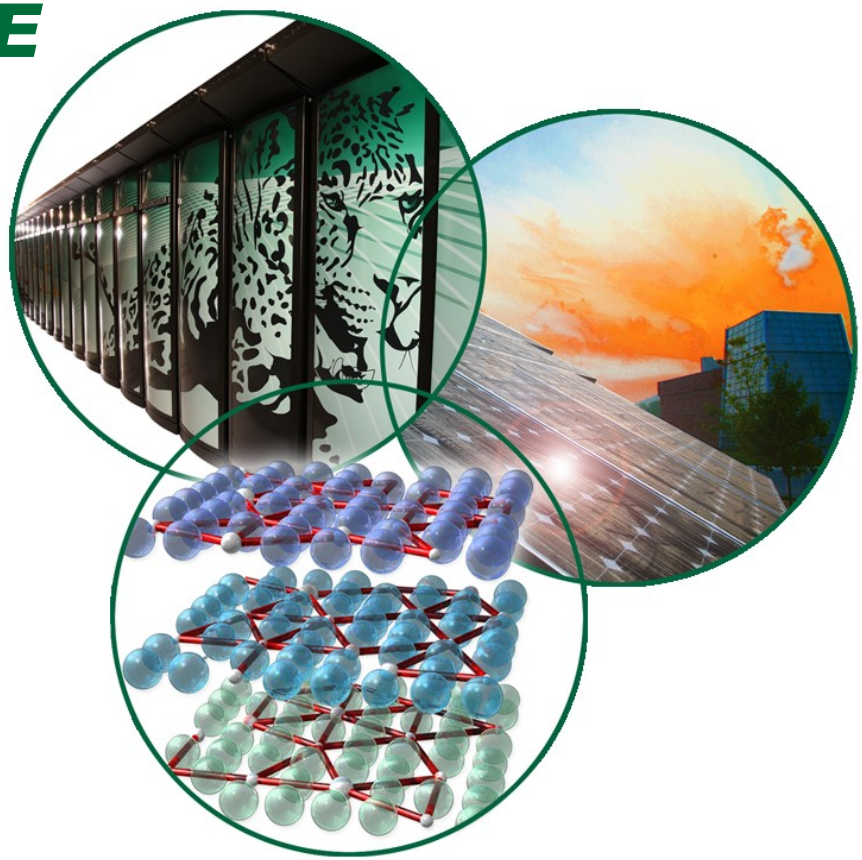
EXTREME-SCALE ALGORITHMS AND SOFTWARE INSTITUTE



PI: Al Geist

Oak Ridge National Laboratory

SC10 Meeting

New Orleans
November 17, 2010

U.S. DEPARTMENT OF ENERGY

Research supported by DOE ASCR

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Joint Math/CS Institute
Extreme-scale Algorithms & Software Institute

Architecture-aware Algorithms for Scalable Performance and Resilience on Heterogeneous Architectures

It's EASI !

We have a Strong Team

Al Geist (ORNL)

Michael Heroux and Ron Brightwell (SNL)
George Fann (ORNL)
Bill Gropp (U ILL)
Jack Dongarra (UTK)
Jim Demmel (UC Berkeley)

Petascale Roadmap

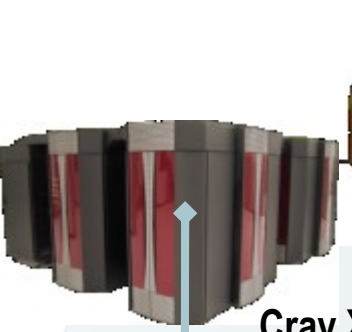
Oak Ridge increased computational capability by almost 1000X in half a decade.

ORNL Leadership Computing Facility successfully executed its petascale roadmap plan on schedule and budget.

Mission: Delivering resources for science breakthroughs. Multiple science applications now running at over a sustained petaflop

Growth was driven by multi-core sockets and increase in the number of cores per node

Growth of the Supercomputers for Energy Research



Cray X1
3 TF

FY2005

Cray XT3
Single-core
26 TF

FY2006

Cray XT3
Dual-core
54 TF

FY2007

Cray XT4
4-core,
dual socket
119 TF

FY2008

Cray XT4
8-core,
dual socket
263 TF

FY2009

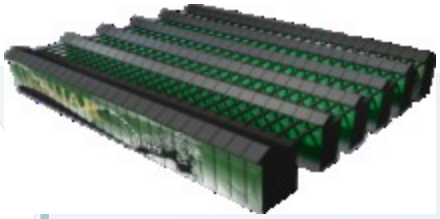
Cray "Baker"
12-core,
dual-socket SMP
2300 TF

Exascale Roadmap

Delivering the next 1000x capability in a decade

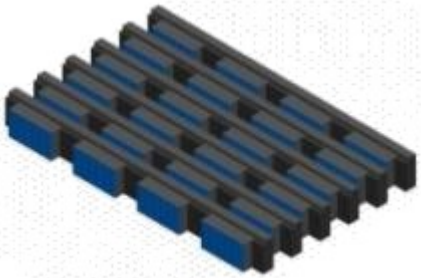
Mission need: Provide the computational resources required to tackle critical national problems	Must also provide the expertise and tools to enable science teams to productively utilize exascale systems

Expectation is that systems will be heterogeneous with nodes composed of many-core GPUs and CPUs



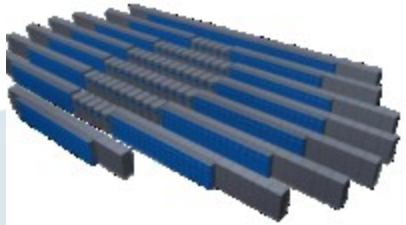
Jaguar: 2 PF Leadership-class system for science

FY2009



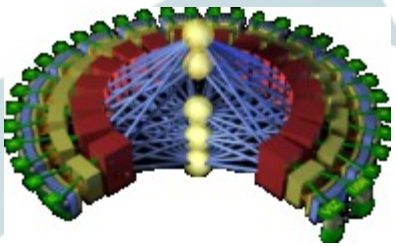
20 PF Leadership-class system

FY2011



100-250 PF

FY2015



Future system: 1 EF

FY2018

EASI Project Overview
Addressing Heterogeneity and Resilience

Applications
HOMME
MADNESS
Charon

Algorithms
Krylov
Poisson
Helmholtz
Dense NLA
BLAS

Runtime
MPI
Shared-memory
Processor affinity
Memory affinity

Architecture **Heterogeneous, multi-core, extreme-scale**

Research Areas in Institute

Heterogeneous programming API
Robust multi-precision algorithms
Communication optimal algorithms
Hybrid programming
Resilient algorithms
New parallelization methods
Auto-tuned BLAS (API)

Task placement and scheduling
Memory management
Architecture-aware MPI

Community Outreach

Deliver codes to community through:
ScaLAPACK
Trilinos
Open MPI
MPICH2
MADNESS
HOMME

Workshops
Training
Publications

EASI uses co-design
to provide both near and long-term Impact:

Integrated team of math, CS, and application experts working together to create new:

Architecture-aware algorithms and associated runtime to enable many science applications to better exploit the architectural features of DOE's petascale systems.

Applications team members immediately incorporate new algorithms providing **Near-term high impact on science**

Numerical libraries used to disseminate the new algorithms to the wider community providing **broader and longer-term impact.**

EASI Technical Research Areas

Heterogeneous programming API

Provide a consistent library interface that remains the same for library developers independent of processor heterogeneity

Robust multi-precision algorithms

Use a mix of single precision and double precision in a way that is transparent to the user.

Communication optimal algorithms

Develop NLA algorithms that reduce communication to a minimum.

Develop Auto-tuned BLAS (API) for heterogeneous processors

Hybrid programming

"MPI only" application with MPI+heterogeneous node support in libraries

Resilient algorithms runtime support

Allows an application to declare certain data as highly reliable, and to Assert that certain computations have to be completed correctly

Runtime support for task and data placement on multi-core nodes, since algorithm performance is extremely sensitive to placement.

EASI
Accomplishments and Highlights to date

highlights **EASI Research Areas**

- ✓ Heterogeneous programming API
- ✓ Robust multi-precision algorithms
- ✓ Communication minimizing algorithms
- ✓ Hybrid programming
- ✓ Resilient algorithms

EASI: Developed heterogeneous programming API – Geist ORNL

ASCR- Math/CS Institute Highlight

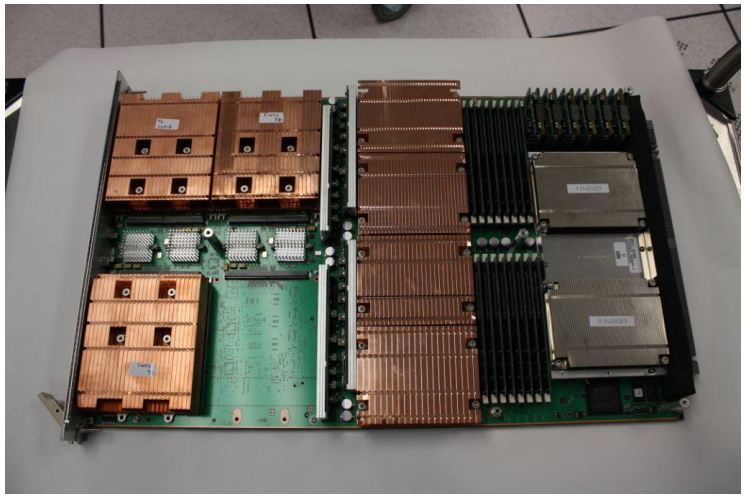
Objectives

- **Develop a programming API for heterogeneous nodes that contain multicore CPUs and GPUs**
- **Make the API portable across GPU vendors**
- **Make the API extensible to other programming models as needed**

Impact

- **Allows writing portable parallel linear algebra software that can use pthreads, OpenMP, CUDA, or Intel TBB (even more than one within the same executable)**

Heterogeneous Supercomputer Node



Progress (and/or Accomplishments w/FY)

- **Completed a portable API for multicore CPUs and GPUs. (2010)**
- **Using the API, we demonstrated compiling and running the same software kernel using pthread, Intel Threading Building Blocks and CUDA. (2011)**
- **This API is now supported in Trilinos. (2011)**
- **The API is documented in http://www.cs.sandia.gov/~maherou/docs/TrilinosNodeAPI.pdf**

Contact: Mike Heroux <maherou@sandia.gov>

EASI: Multi-precision dense matrix algorithms on GPUs – Geist ORNL

ASCR- Math/CS Institute Highlight

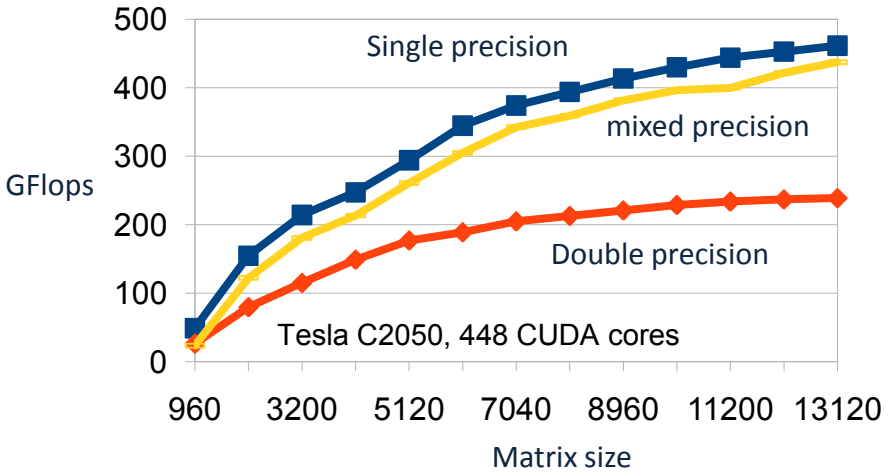
Objectives

- **Develop robust multi-precision algorithms that get answers to double precision accuracy but exploit advantages of single precision as much as possible**
- **Make these algorithms portable across CPU and GPU systems**

Impact

- **Single precision is twice as fast as double precision (2 ops/cycle instead of 4 ops/cycle)**
- **Reduce data motion in half (32 bit data instead of 64 bit data)**
- **Higher locality in cache (Cache is effectively twice as large)**

Performance of new multi-precision algorithm is close to the maximum single precision performance



Progress (and/or Accomplishments w/FY)

- **Completed a portable multi-precision algorithm to solve linear system of equations for multicore CPUs and GPUs. (2010)**
- **Demonstrated performance is nearly as good as single precision while robustly getting double precision accuracy. (2011)**
- **This algorithm now distributed in the MAGMA math library. (2011)**

Contact: Jack Dongarra <dongarra@cs.utk.edu>

EASI: Multi-precision Sparse Matrix Algorithms – Geist ORNL

ASCR- Math/CS Institute Highlight

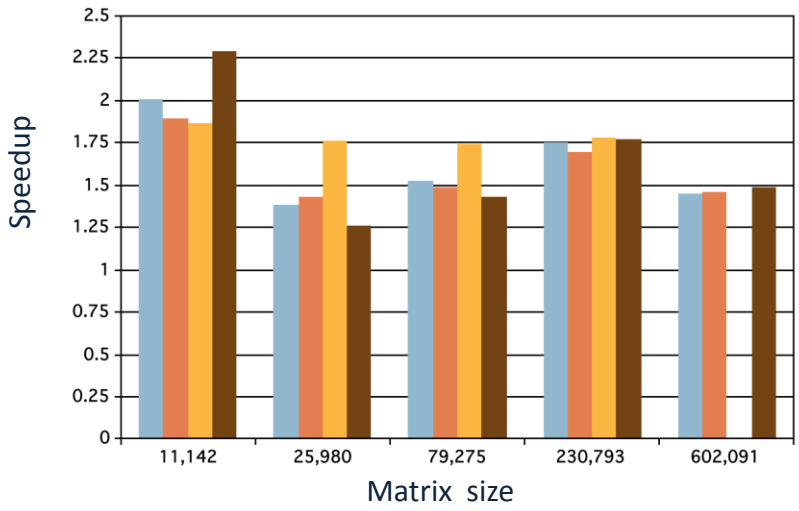
Objectives

- **Develop robust multi-precision sparse algorithms that get answers to double precision accuracy but exploit advantages of single precision as much as possible**
- **Make these algorithms portable and easy to use**

Impact

- **Single precision is twice as fast as double precision (2 ops/cycle instead of 4 ops/cycle)**
- **Reduce data motion in half (32 bit data instead of 64 bit data)**
- **Higher locality in cache (Cache is effectively twice as large)**

Performance of new sparse multi-precision algorithm is up to 2X faster that previous double precision algorithm



Progress (and/or Accomplishments w/FY)

- **Trilinos Library incorporates Multi-Precision Algorithms for Sparse Matrix Solvers (2010)**
- **Trilinos is an object-oriented software framework for the solution of large-scale, complex multi-physics engineering and scientific problems. The latest release utilizes C++ templates to allow users to easily mix precisions in their solvers (2011)**
- **Speedups for mixed precision up to 2X using Inner loop SP and Outer loop DP (SP/DP) *vs* DP/DP (2011)**

Contact: Mike Heroux <maherou@sandia.gov>

EASI: Proving communication lower bounds – Geist ORNL

ASCR- Math/CS Institute Highlight

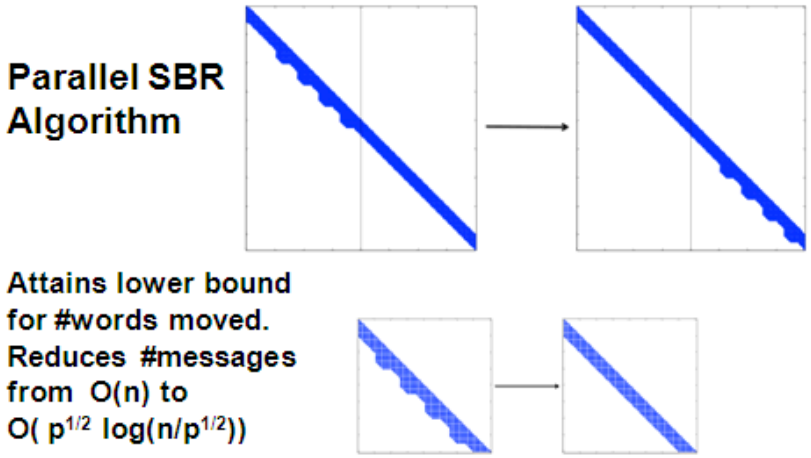
Objectives

- **Prove communication lower bounds for *matrix multiplication, LU, QR, SVD*, and Krylov subspace methods for Ax=b, Ax=λx**
- **Develop algorithms that attain the minimum data movement, in some cases by using more memory**

Impact

- **Cost of communication (moving data between layers of memory or between processors) greatly exceeds cost of arithmetic by 100X**
- **Order of magnitude speedups have been demonstrated**

Performance of new communication minimizing Successive band reduction algorithm is up to 30 times faster than existing implementations



Progress (and/or Accomplishments w/FY)

- Best Paper Prize at SPAA'11 "Graph expansion and communication costs of fast matrix multiplication" for communication lower bounds for Strassen-like algorithms
- Distinguished Paper Award at Euro-Par'11 for "Communication-optimal parallel 2.5D matrix multiplication and LU factorization algorithms "
- New communication-avoiding successive band reduction (SBR) up to 30X speedup over ACML and 17X over MKL.

Contact: Jim Demmel <demmel@cs.berkeley.edu>

EASI: Communication minimizing sparse algorithms – Geist ORNL

ASCR- Math/CS Institute Highlight

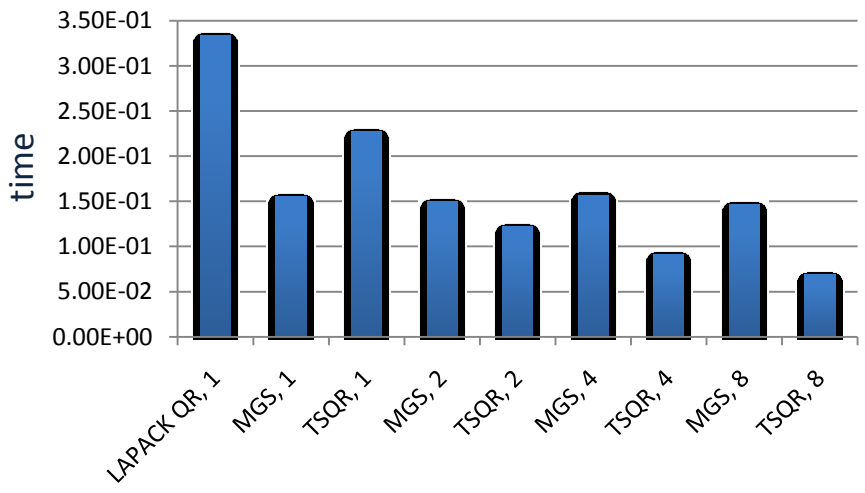
Objectives

- **Develop algorithms that attain the minimum data movement for Krylov subspace methods for Ax=b, Ax=λx**
- **Distribute to users through Trilinos math lib**

Impact

- **Critical for exascale solvers.**
- **Tall Skinny QR factorization (TSQR) communicates less & more accurate**
- **Order of magnitude speedups have been demonstrated**

Performance of new communication minimizing TSQR algorithm is up to 10 times faster than LAPACK



Progress (and/or Accomplishments w/FY)

TSQR capability:

- **Part of the Trilinos scalable multicore capabilities. (2010)**
- **Helps all iterative solvers in Trilinos (available to external libraries, too).**
- **Part of Trilinos 10.6 release (FY2011).**

Contact: Mike Heroux <maherou@sandia.gov>

EASI: High performance Hybrid Algorithms – Geist ORNL

ASCR- Math/CS Institute Highlight

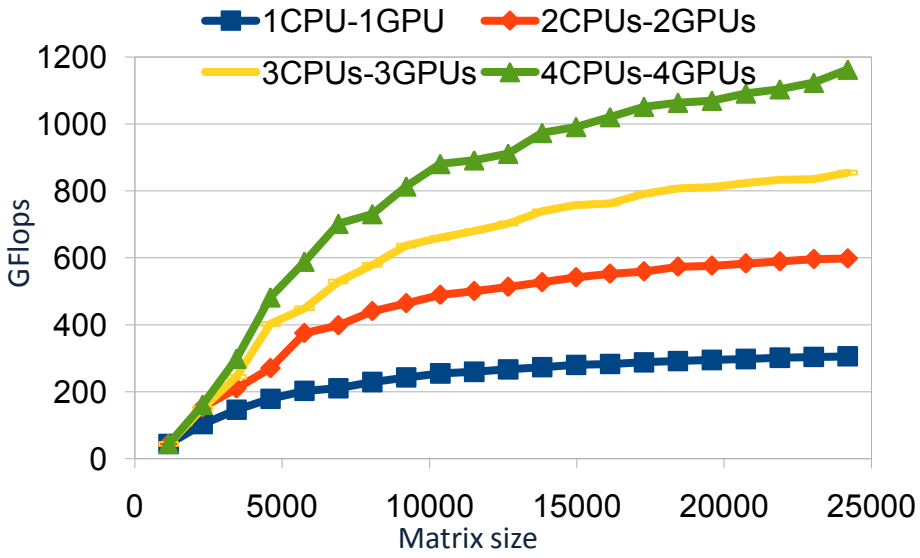
Objectives

- **Develop linear algebra algorithms that run efficiently on hybrid nodes with CPUs and GPUs**
- **Make these algorithms portable and available through MAGMA math library**

Impact

- **Enable applications to exploit much of the power of hybrid manycore and GPUs systems by simply linking in a GPU optimized math library.**

New hybrid Cholesky achieves nearly perfect scale up



Progress (and/or Accomplishments w/FY)

- **"Towards Dense Linear Algebra for Hybrid GPU Accelerated Manycore Systems",** Stanimire Tomov, Jack Dongarra, and Marc Baboulin, Parallel Computing, Volume 36, Issues 5-6, pp 232-240, 2010.
- **"Hybrid Multicore Cholesky Factorization with Multiple GPU Accelerators",** H. Ltaief, S. Tomov, R. Nath, and J. Dongarra, Submitted to IEEE Transaction on Parallel and Distributed Computing, 2010.

Contact: Jack Dongarra <dongarra@cs.utk.edu>

EASI: Resilient linear algebra algorithms – Geist ORNL

ASCR- Math/CS Institute Highlight

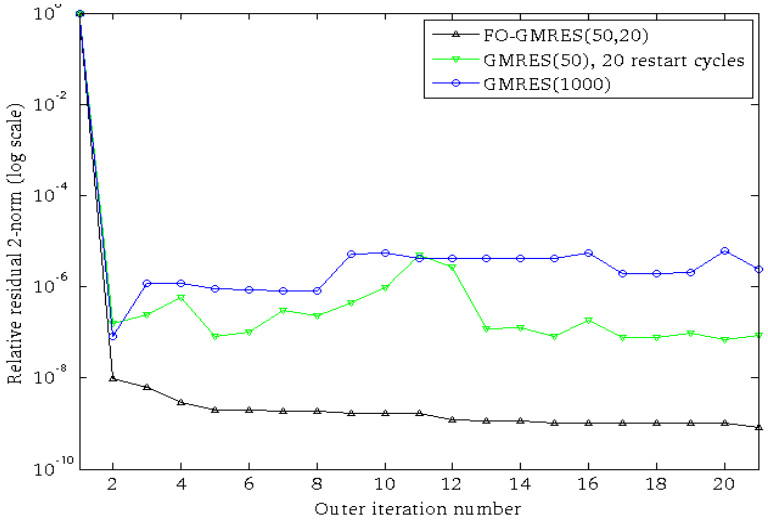
Objectives

- **Exascale systems will be less reliable Including incorrect data and computations**
- **Develop linear algebra algorithms that can tolerate undetected errors and still get the accurate results**

Impact

- **Faults may cause incorrect results *undetectably.* New GMRES algorithm is fault tolerant even to undetected errors.**
- **Opens up a new way to produce resilient algorithms through specifying selective regions of reliable data and computation**

Resilience of FT-GMRES algorithm allows it to converge despite silent errors while existing solution methods do not



Progress (and/or Accomplishments w/FY)

Develop FT-GMRES algorithm (2011)

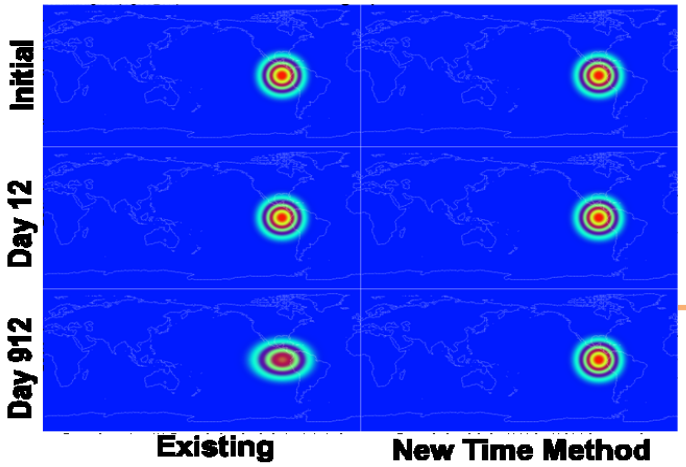
- **Inner solver "preconditions" outer solver**
- **Inner solver runs unreliably, outer solver reliably**
- **Reuse any existing solver stack as "inner solver"**
- **Most time spent in cheap unreliable mode**
- **Faults only delay convergence; don't prevent it**
- **Standard approach of restarting GMRES is not sufficient – doesn't converge**

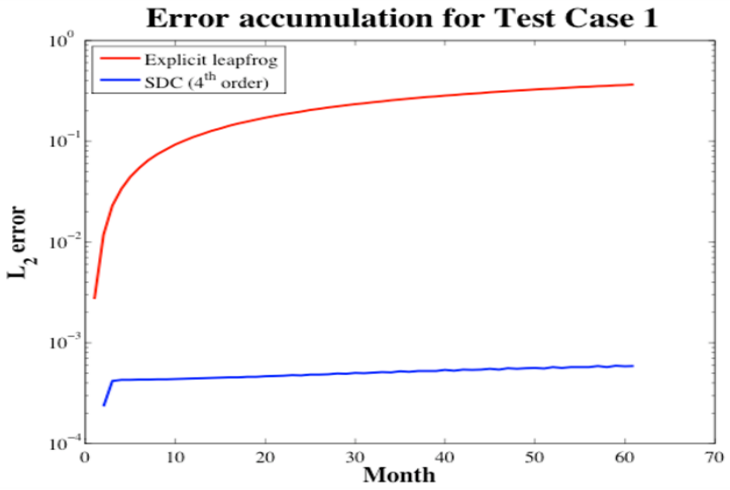
Contact: Mike Heroux <maherou@sandia.gov>

Accurate time stepping for modeling climate dynamics
G. Fann, J. Jia, J. Hill, K. Evans (ORNL), M. Taylor (SNL)

Objectives

- **Scalable and accurate long time simulations of time-dependent physical systems**





Accomplishments

- **Develop and demonstrated new mathematical method and algorithms for high accuracy simulations in time up from 2nd-8th order**
  - **Exceeds the Fully Implicit Jacobian-Free Newton-Krylov method (using it as a preconditioner)**
  - **Developed scalable variants of deferred correction methods**
- **Passed all test cases of the climate dynamics core spectral element code, HOMME, for the shallow water equation**

Accurate time-dependent simulations computes

- *Accurate flow fields integration*
- *Conserves mass and energy over time*
- *Topological and geometric features preserved*

For long numerical simulations, the improved temporal accuracy of the new variant of SDC (right) method is evident by the preservation of the Gaussian shape compared to the leapfrog method (left).

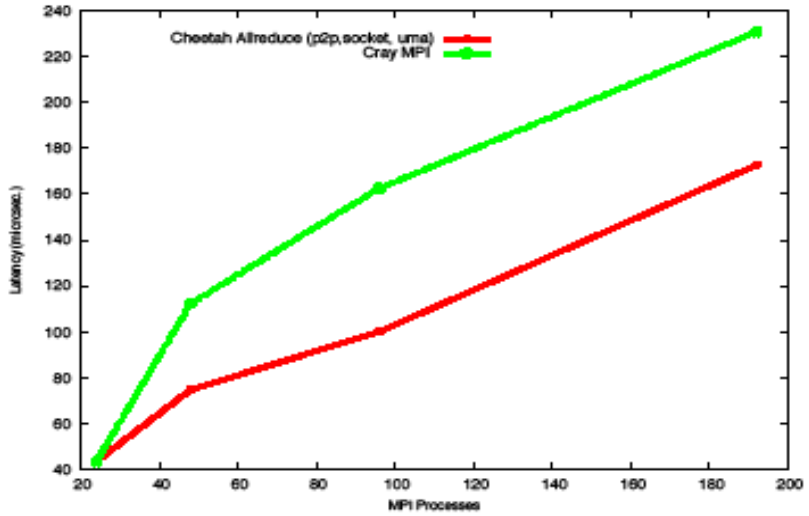
EASI: Faster All Reduce Implementation via Cheetah
R. Graham, M. G. Venkata, J. Ladd and P. Shamis (ORNL)

Objectives

- **Lower communication cost for all-reduce, broacast and global collective operations**
  - **Many HPC apps, solvers and benchmarks (e.g. Fluent, OpenFOAM, Paratec) spend > 50% of collective time in all-reduce**
  - **Improve scalability of applications and iterative solvers (e.g., global dot product in Krylov solvers)**

Accomplishments

- **Faster allreduce and collective via Cheetah**
  - **Outperform Cray MPI allreduce by 33% using 192 MPI processes**
- **Heirarchical collective groups**
  - **Customized collective primitives for each unique communication path/network**
  - **Concurrent progress of optimized comm primitives for asynchronous progress**
  - **Topological aware collective and network decomposition**
  - **K-tree for data distribution and reduction for scaling**



The latency of Cheetah allreduce/global sum compared with Cray MPI allreduce for 1KB message size. Cheetah outperforms Cray MPI allreduce by 33% using 192 MPI processes.

P. Shamis, R. L. Graham, M. G. Venkata, J. S. Ladd , "Design and Implementation of Broadcast Algorithms for Extreme-Scale Systems", at IEEE Cluster 2011.

J. Ladd, M. G. Venkata, R. Graham, P. Shamis, "Analyzing the Effects of Multicore Architectures and On-host Communication Characteristics on Collective Communications", accepted the SRMPDS workshop in conjunction with ICPP.

Contact: Rich Graham <rlgraham@ornl.gov>

EASI: Team Member Awards – Geist ORNL

ASCR- Math/CS Institute Highlight

Objectives

- **Recognition of the world-changing research being done by members of the EASI Team**

Impact

- **Publicity over the awards makes the research done under the EASI project more visible around the world**

Jack Dongarra

Jim Demmel

Progress (and/or Accomplishments w/FY)

- **Jack Dongarra** was awarded the 2011 IEEE Charles Babbage Award in recognition of pioneering working numerical algorithms and linear algebra libraries for parallel computing.
- **Jim Demmel** was elected the National Academy of Sciences in 2011

Contact: Al Geist <gst@ornl.gov>

Broader Impact and Standardization

Goal: Distribute the new algorithms and runtime support through widely used software packages

SCALAPACK	MADDNESS
Trilinos	HOMME
MAGMA	
BLAS	

Runtime: Open MPI and MPICH are the two most widely used MPI libraries.

In the past month we have gotten these MPI extensions officially accepted as a branch of the Open MPI source tree.

Standardization efforts: Begun formal discussions with the **MPI-3** forum about getting these features into the standard

Workshops: