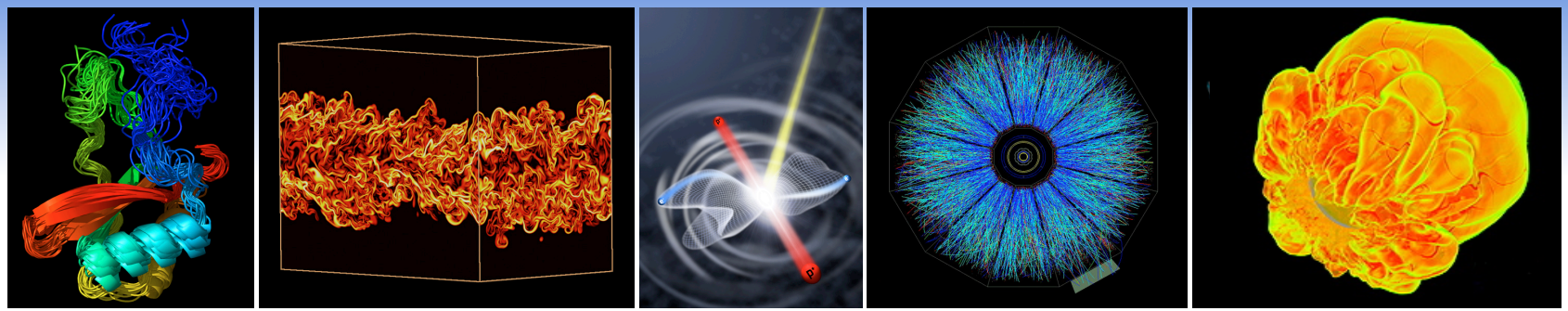


Workshop Goals & Process

Large Scale Computing and Storage Requirements for High Energy Physics Research

Joint HEP/ ASCR / NERSC Workshop

Harvey Wasserman
NERSC User Services

November 12-13, 2009

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC

BERKELEY LAB

Logistics: Schedule

- Agenda on workshop web page
  - http://www.nersc.gov/projects/science_requirements/HEP/agenda.php
- Mid-morning / afternoon break, lunch
- Self-organization for dinner
- 5 “science areas,” one workshop
  - Science-focused but cross-science discussion
  - Explore areas of common need (within HEP)
- Breakout sessions Friday AM in one room

Why is NERSC Collecting Computational Requirements?

- Help ASCR and NERSC make informed decisions for technology and services.
- Input is used to guide procurements, staffing, and to improve the effectiveness of NERSC services.
  - Includes hardware, software, support, data, storage, analysis, work flow
  - Time scale: 5 years
- Result: NERSC can better provide what you need for your work.

Logistics: Case Studies

- One co-lead (for each science area)
  - help roll up discussions into major case studies
- Case Studies:
  - Narrative describing science & NERSC reqmts
  - Audience is NERSC, DOE program managers
  - Initial set suggested by Amber
    - Minimum set to capture HEP mission and unique NERSC requirements
    - Actual number may vary
  - Encourage participation by all; roundtable

Logistics: Templates

- Web templates: web "Reference Material"
  - Based on NERSC info
  - Summary of projects as we know them
  - Good point of departure
    - A framework for discussion
    - But not necessarily the entire story

Logistics: Final Report Content

- Format similar to ESnet
  - But NERSC requirement space much broader than Esnet
  - See “Reference Material” on web site
  - Contents
    - Executive summary,
    - ~2-page case study reports,
    - NERSC synthesis of all results

Logistics: Final Report Schedule

- Revised case studies due to NERSC .. Nov 29
- NERSC draft report …….. Dec 23
- Participants review period ……Jan 11, 2010
- NERSC Near final ..... Feb 7
- BER AD approval .....
- NERSC Revisions .....
- Final Report posted on Workshop Webpage

  .....

Examples of Information Sought

- Type of simulation, #, reason for #, algorithms, solver
- Parallelism: method, weak or strong scaling, implementation, concurrency, limits
- Key physical parameters and their limits:
  - spatial resolution, # of atoms/energy levels, integration range, …
- Representative code
- Key science result metrics and goals

Examples of Information Sought

- Typical science process (workflow)
- Data: amount stored / transferred for input, results, and fault mitigation
- Special needs for data intensive projects
  - Grids, gateways, workflows, provenance, `
- Special query regarding multicore/manycore

- How all of this is
  - Driven by the science
  - Likely to change and why

Lattice QCD

- Doug Toussaint (University of Arizona), Lead
  - QCD with three flavors of dynamical quarks
- Paul McKenzie (Fermilab)
  - Chair of USQCD
- Don Sinclair (ANL)
  - Lattice Gauge Theory Simulations
- Bernd Berg (FSU)
  - Deconfined Phase in small Volumes with Cold Boundary Conditions
- Junko Shigemitsu (OSU)
  - Heavy-Light Physics with NRQCD Heavy and Improved Staggered Light Quarks

Astrophysics: Modeling

- Stan Woosley (UC SC), Lead
  - Computational Astrophysics Consortium
- John Bell (LBNL)
  - Low Mach Number Astrophysics, Compressible Astrophysics, Nuclear Flames
- Mike Norman (SDSC)
  - The Cosmic Frontier
- Primack, Joel (UC SC)
  - Galaxy Formation
- Edward Baron
  - Synthetic Spectra of Astrophysical Objects

Accelerator Science

- Panagiotis Spentzouris (Fermilab), Lead
  - COMPASS
- Warren Mori, Frank Tsung (UCLA), Cameron Geddes (LBNL), Phillip Sprangle (NRL), David Bruhwiler (T-X)
  - Laser Wakefield, plasma accelerators
- Lie-Quan Lee, Kwok Ko (SLAC NAL)
  - Advanced Modeling for Particle Accelerators
- Ji Qiang (LBNL)
  - Beam Delivery Optimization for X-Ray FEL

Astrophysics: Data Analysis

- Julian Borrill (LBNL), Alex Szalay (JsHU), Co-Leads
  - CMB
  - Sloan Digital Sky Survey
- Peter Nugent (LBNL)
  - Palomar Transient Factory, La Silla Supernova Search, DeepSky Gateway, Baryon Oscillation Spectroscopic Survey
- Greg Aldering (LBNL)
  - The Nearby Supernova Factory
- George Smoot (LBNL)
- Dan Werthimer (UCB)
  - Berkeley High Resolution Neutral Hydrogen Sky Survey

Detector Simulation and Data Analysis

- Craig Tull (LBNL), Lead
- PDSF
- Big community, pre-conceived structure, workflow, grid based
- Daya Bay Neutrino Experiment
- ATLAS (A Toroidal LHC ApparatuS)
- AstroGFS (Smoot: *Large Astrophysical Data Sets: Data analysis and simulation of astro-physical neutrinos, dark matter and dark energy.)*
- Nearby Supernova Factory

Final Thoughts

- LBNL will try to record – could use help
- Requirements characterization process is not complicated.
- Mutually beneficial.

Scaling Science

Inspired by **P. Kent**, "*Computational Challenges in Nanoscience: an* ab initio *Perspective*", Peta08 workshop, Hawaii (2008) and **Jonathan Carter** (NERSC).

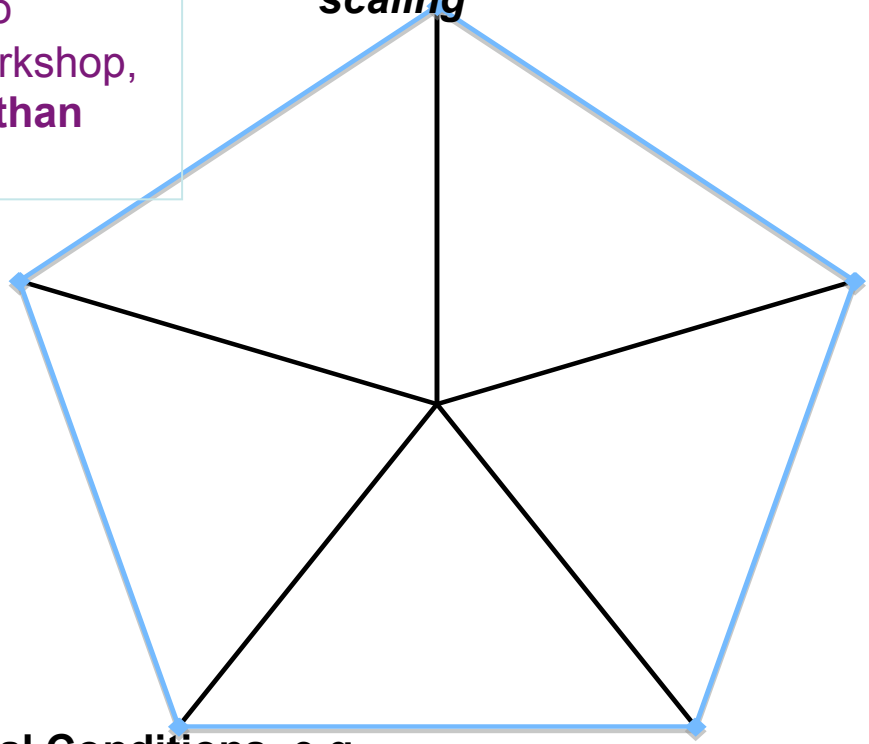
Length, Spatial extent, #Atoms, *Weak scaling*

Time scale Optimizations, *Strong scaling*

Simulation method, e.g. DFT, QMC or HF/ SCF; LES or DNS

Initial Conditions, e.g. molecule, boundaries, *Ensembles*

Convergence, systematic errors due to cutoffs, etc.



NERSC NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

BACKUP SLIDES

Workload Analysis

- Ongoing activity within NERSC SDSA*
- Effort to drill deeper than this workshop
  - Study representative codes in detail
- See how the code stresses the machine
  - Help evaluate architectural trade-offs

***Science Driven System Architecture Team, http://www.nersc.gov/projects/SDSA/**

Workload-Driven Characteristics

- Memory requirements as *f*(algorithm, inputs)
- Memory-to-floating-point operation ratio
- Memory access pattern
- Interprocessor communication pattern, size, frequency
- Parallelism type, granularity, scaling characteristics, load balance
- I/O volume, frequency, pattern, method, desired percent of total runtime
- How science drives workload scaling: problem size, data set size, memory size

How Science Drives Architecture

Science areas \ Algorithm	Dense linear algebra	Sparse linear algebra	Spectral Methods (FFTs)	Particle Methods	Structured Grids	Unstructured or AMR Grids	Data Intensive
Accelerator Science		X	X	X	X	X	
Astrophysics	X	X	X	X	X	X	X
Chemistry	X	X	X	X			X
Climate			X		X	X	X
Combustion					X	X	X
Fusion	X	X		X	X	X	X
Lattice Gauge		X	X	X	X		
Material Science	X		X	X	X		
BioScience			X	X			X

Machine Requirements

Science areas \ Algorithm	Dense linear algebra	Sparse linear algebra	Spectral Methods (FFT)s	Particle Methods	Structured Grids	Unstructured or AMR Grids	Data Intensive
Accelerator							
Astrophysics							
Chemistry							
Climate							
Combustion							
Fusion							
Lattice Gauge							
MatSci							
BioScience							

High Flop/s rate

High performance memory system

High bisection bandwidth

High performance memory system

High flop/s rate

Low latency, efficient gather/scatter

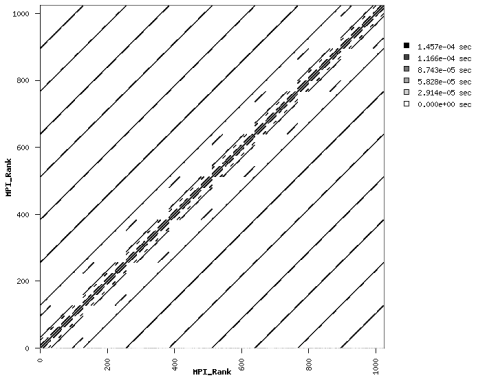
Storage, Network Infrastructure

Workload-Driven Characteristics

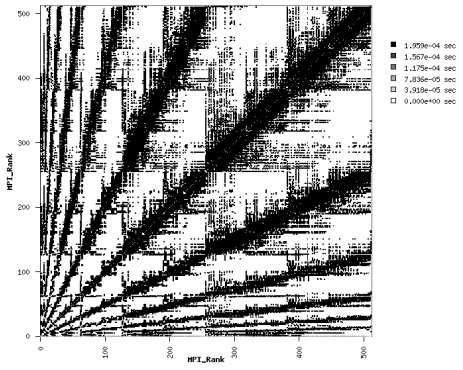
- What follows are data and descriptions of three benchmark codes used by NERSC recently that represent portions of the NERSC HEP workload.
- The full report concerning these data is available as LBNL Technical Report LBNL-1014E, available from the NERSC web site.

http://www.nersc.gov

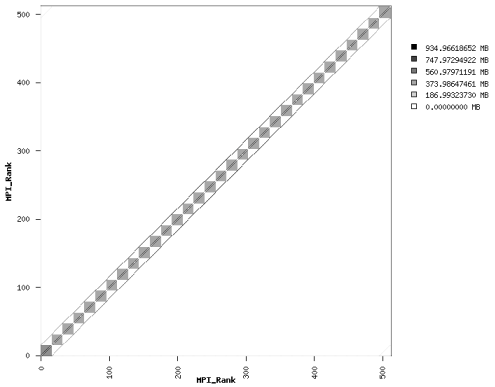
Communication Topology



MILC (QCD)



MAESTRO (Low Mach Number Flow)



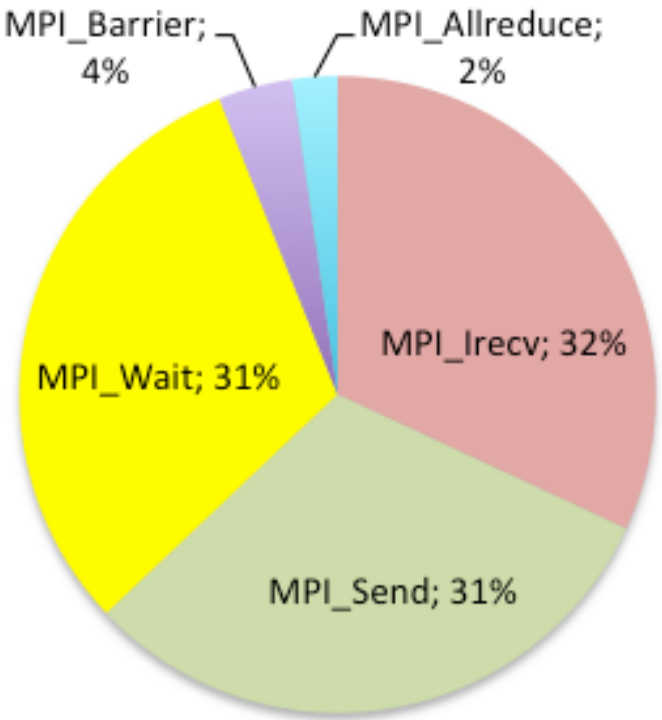
IMPACT-T (Accelerator Physics PIC)

IMPACT-T: Accelerator Science

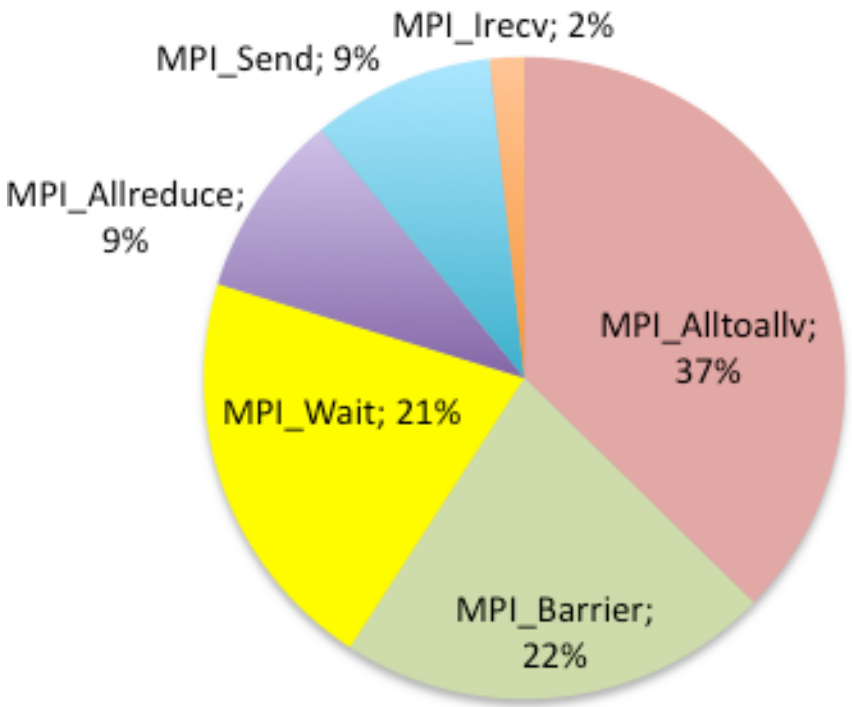
- Author: J. Qiang, et al., LBNL Accelerator & Fusion Research Div.
- Relation to NERSC Workload
  - DOE High Energy Physics (HEP) and Nuclear Physics (NP) programs, plus SciDAC COMmunity Petascale Project for Accelerator Science and Simulation.
  - Part of a suite of codes, IMPACT-Z, Theta, Fix2d/3d, others.
  - Wide variety of science drivers/approaches/codes: Accelerator design, electromagnetics, electron cooling, advanced acceleration
- Description: 3-D PIC, quasi-static, integrated Green Function, moving beam frame; FFT Poisson solver.
- Coding: 33,000 lines of object-oriented Fortran90.
- Parallelism: 2-D decomposition, MPI; frequent load-rebalance based on domain.
- NERSC-6 tests: photoelectron beam transported through a photoinjector similar to one at SLAC; strong scaling on 256 and 1024 cores; 50 particles per cell

IMPACT-T Characteristics

MPI Calls by Count



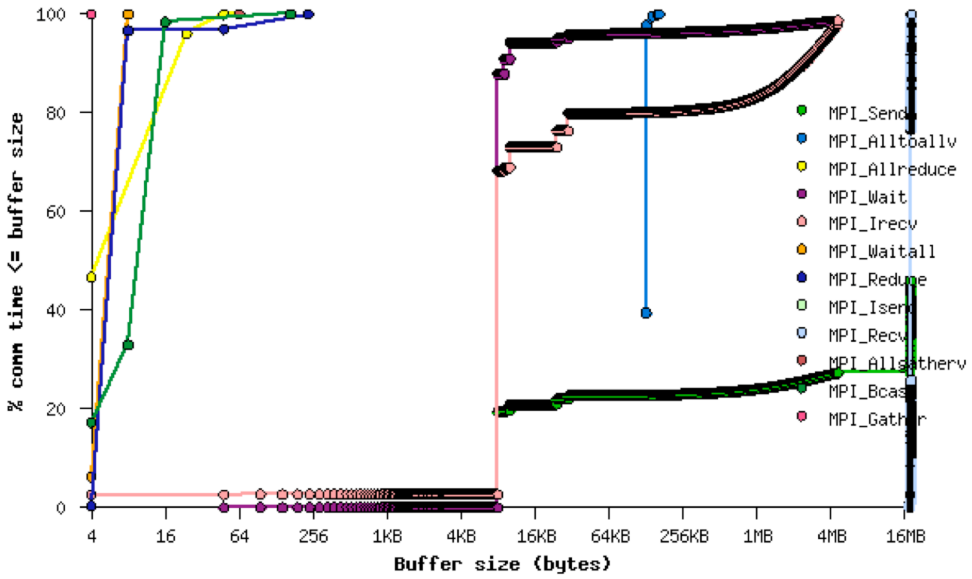
MPI Calls by Time



Data from IPM using IMPACT-T on 1024 cores of Franklin.

IMPACT-T Characteristics

MPI Event	Msg Buffer Size (Bytes)	Percent of Total Wall Clock Time
`MPI_Altoallv`	132096	9%
`MPI_Send`	8192	3%



MPI message buffer size distribution based on time for IMPACT-T from IPM on Franklin

IMPACT-T: Performance

P	Itanium HLRB-II		Opteron Ranger		Power5 Bassi		IBM BG/P		Opteron Jaguar		Opteron Franklin	
	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.
256	116	7%	94	8%	143	7%	34	4%	111	5%	130	10%
1024	309	5%	436	9%	n/a		174	5%	513	6%	638	12%

- Differentiation from GTC:
  - Lower computational intensity, percentage of peak;
  - Bigger communication component, different ops;
  - Different performance ratios relative to Franklin.

What IMPACT-T Adds to NERSC-6

- FFT Poisson solver stresses collective communications with small to moderate message sizes;
- Fixed global problem size causes smaller message sizes and increasing importance of MPI latency at higher concurrencies.
- Different from other PIC codes due to external fields, open boundary conditions, multiple beams;
- Relatively moderate computational intensity;
- Object-oriented Fortran90 coding style.

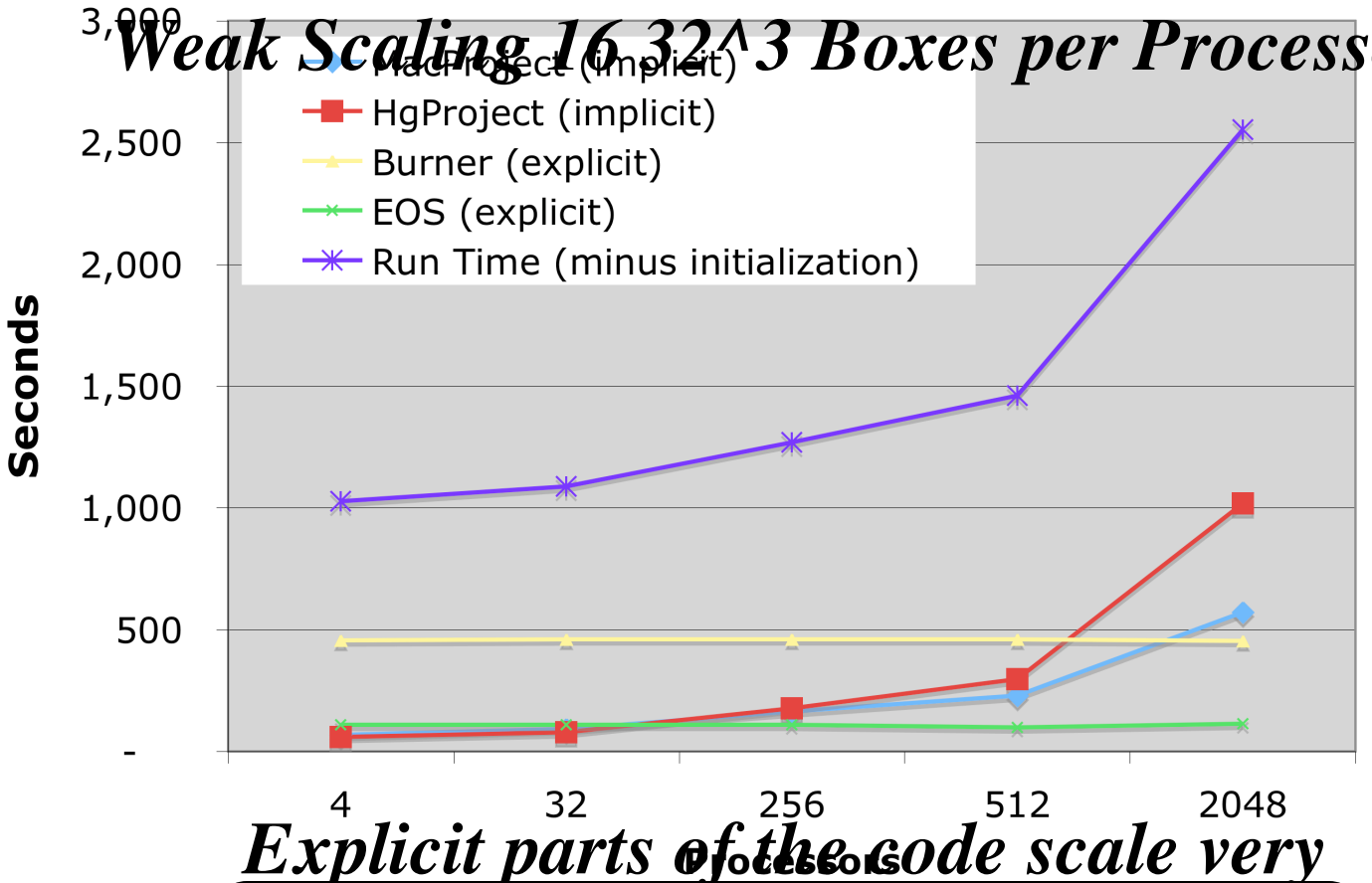
MAESTRO: Low Mach Number Flow

- Authors: LBNL Computing Research Division; SciDAC07
- Relation to NERSC Workload:
  - Model convection leading up to Type 1a supernova explosion;
  - Method also applicable to 3-D turbulent combustion studies.
- Description: Structured rectangular grid plus patch-based AMR (although NERSC-6 code does not adapt);
  - hydro model has implicit & explicit components;
- Coding: ~ 100,000 lines Fortran 90/77.
- Parallelism: 3-D processor non-overlapping decomposition, MPI.
  - Knapsack algorithm for load distribution; move boxes close in physical space to same/close processor.
    - More communication than necessary but has AMR communication characteristics.
- NERSC-6 tests: weak scaling on 512 and 2048 cores; 16 boxes (32^3 cells each) per processor.

MAESTRO Scaling

MAESTRO White Dwarf Convection

Weak Scaling 16 32^3 Boxes per Processor



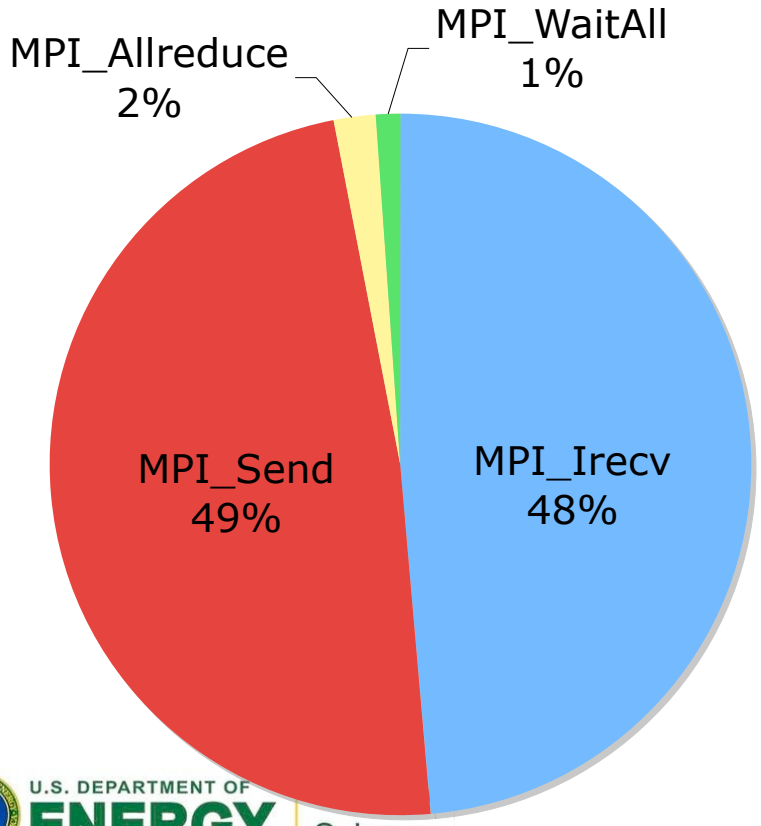
Explicit parts of the code scale very well but implicit parts of code pose more challenges to systems due to global communications

Maestro Communication Patterns

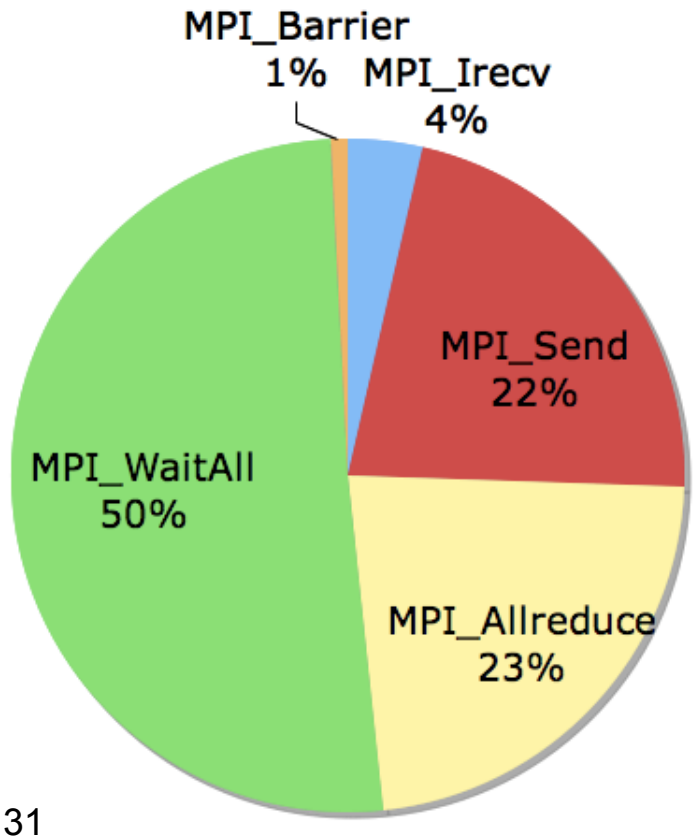
MAESTRO White Dwarf Convection

512 Processors 512x512X1024 Grid from Cray_Pat on Franklin

MPI Calls by Count



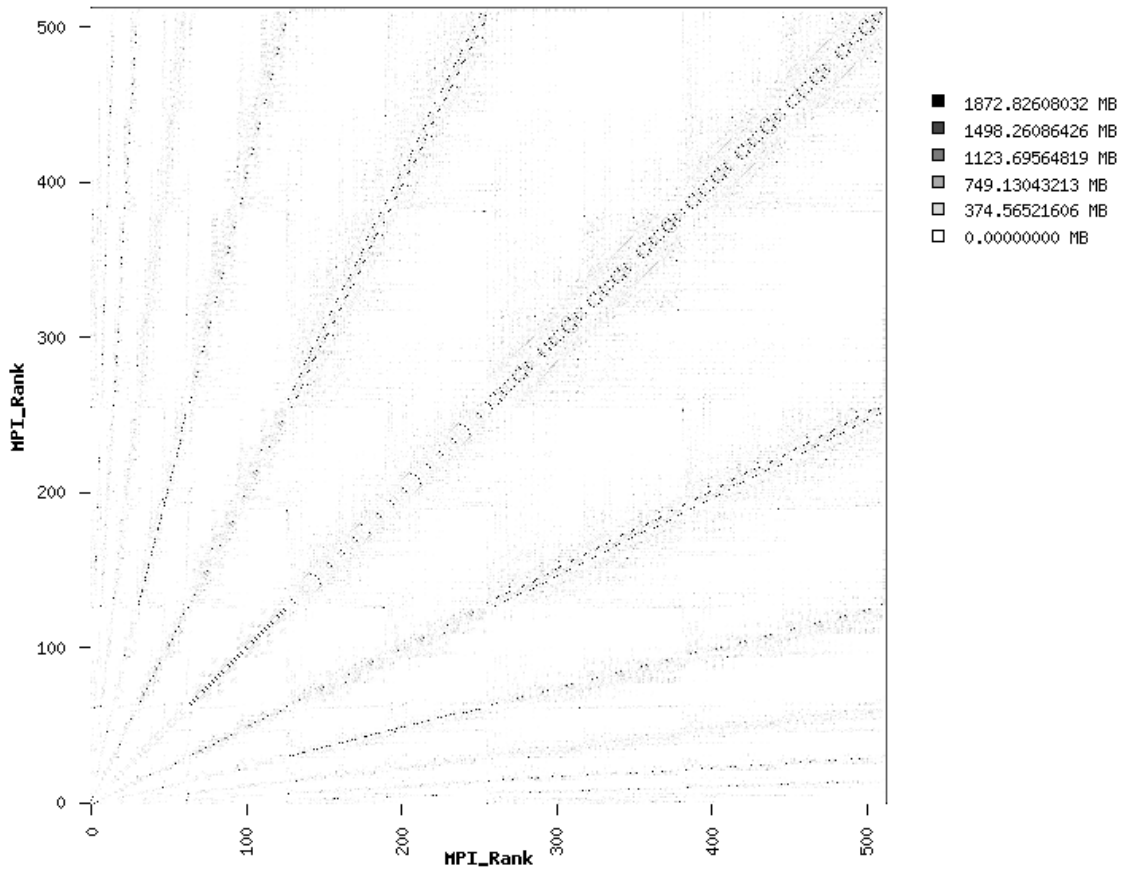
MPI Calls by Time



Maestro Communication Topology

512 procs, 16 32^32 boxes per processor - grid size

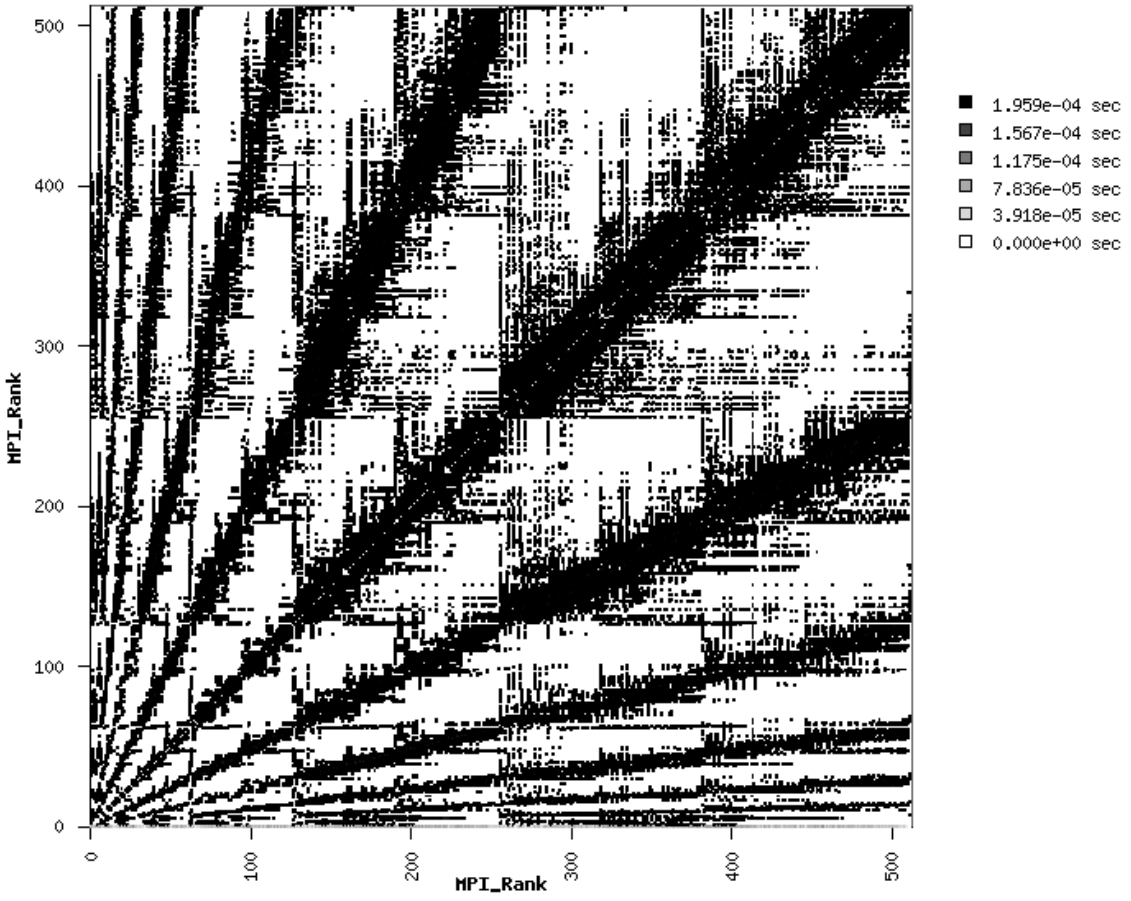
sent



- Communication pattern based on Boxlib grid
- Boxlib works for both adaptive and uniform meshes
- Boxes distributed to be load balanced across processors
- Next, box location optimized for locality
- Result is a clumping effect

Maestro Communication Topology

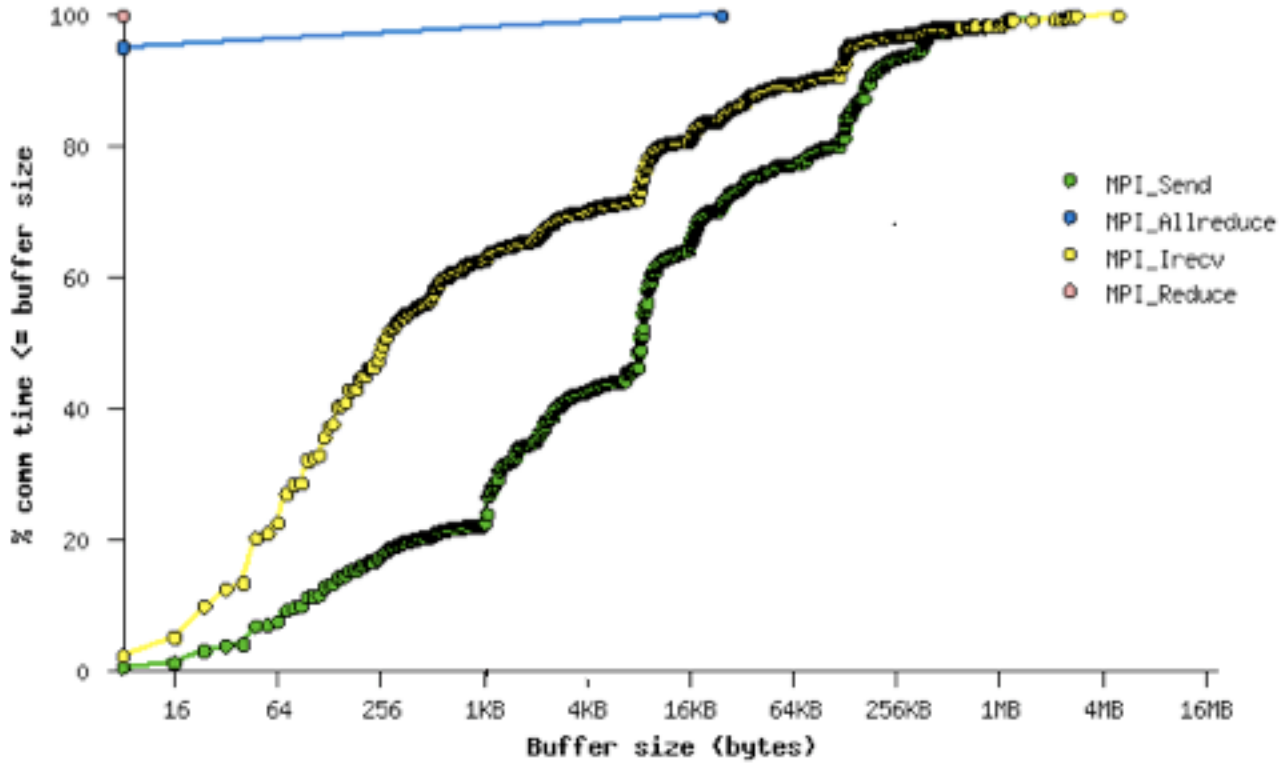
512 procs, 16 32^32 boxes per processor - grid size



- Examining communication topology by time shows global communications more clearly

Maestro Message Sizes

512 procs, 16 32^32 boxes per processor - grid size 512x512x1024



Message Buffer Size Distribution by Time

Maestro: Performance

P	Power5 Bassi		IBM BG/P		Opteron Jaguar		Opteron Franklin	
	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.
512	178	5%	52*	3%*	230	5%	245	9%
2048	n/a				406	2%	437	4%

- All architectures at low percentage of peak for this memory- and communications-intensive benchmark.

*Estimated

What MAESTRO Adds to NERSC-6

- MAESTRO: Unusual communication topology should challenge simple topology interconnects and represent characteristics associated with irregular or refined grids.
- Very low computational intensity stresses memory performance.
- Implicit solver technology stresses global communications;
- Wide range of message sizes from short to relatively moderate.

MILC: MIMD Lattice Gauge QCD

- Authors: MILC collaboration, especially S. Gottlieb
- Relation to NERSC Workload
  - Funded through High Energy Physics Theory
  - Understand results of particle and nuclear physics experiments in terms of Quantum Chromodynamics
- Description: Physics on a 4D lattice, CG algorithm, sparse 3x3 complex matrix multiplies - highly memory bandwidth intensive.
- Coding:
  - V7; ~ 60,000 lines of C; POWER and x86 assembler (Cray redid for Opteron DC & QC); wants gcc.
  - Extensive hard-coded prefetch;
  - CG algorithm with MPI_Allreduce
- Parallelism: 4-D domain decomposition, MPI.

(Continued…)

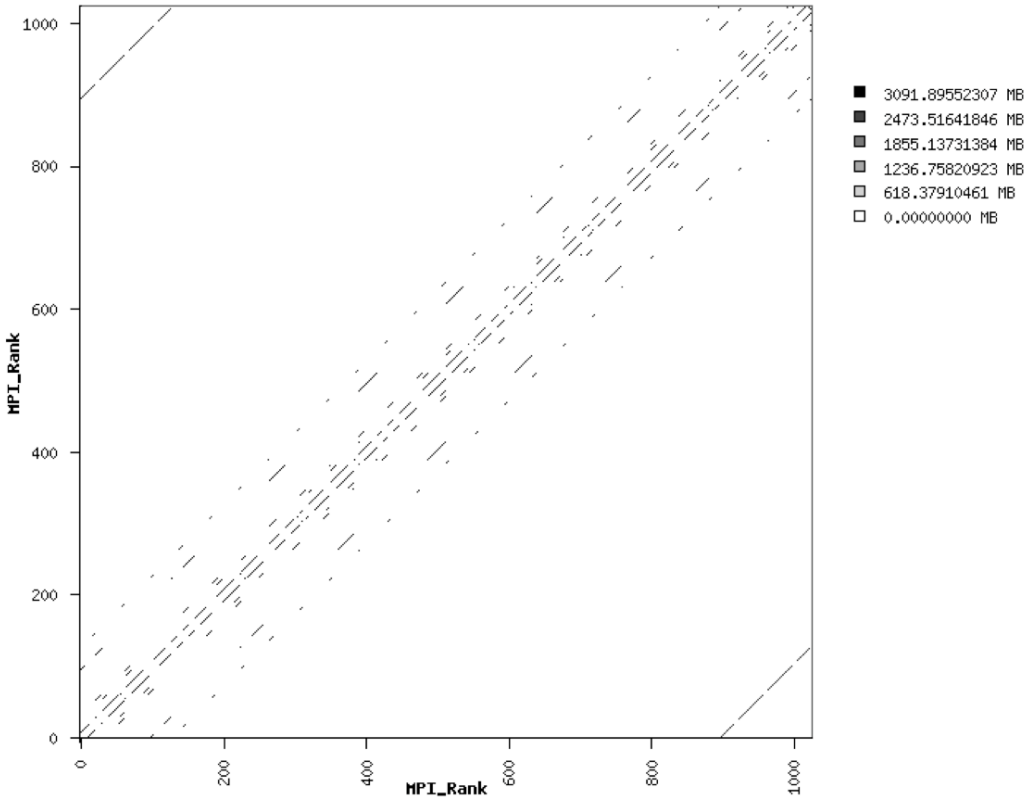
MILC: MIMD Lattice Gauge QCD

- NERSC-6 tests: weak scaling

Concurrency	Global Lattice	Local Lattice
256	32 x 32 x 32 x 36	8 x 8 x 8 x 9
1024	64 x 64 x 32 x 72	8 x 8 x 8 x 9
8192	64 x 64 x 64 x 144	8 x 8 x 8 x 9

  - Much smaller subgrid than NERSC-5.
  - Each test does two runs, one to “prime” the solver, the other to do the measurements.
  - Results in greater emphasis on the interconnect, which tends to dominate performance of some actual QCD runs (due to CG solver).
  - Extra-Large problem same size as Toussain production runs on Franklin in early 2008.
- Profile: Franklin %comm ranges from: 24 - 41%, mostly MPI_Allreduce & MPI_Wait.

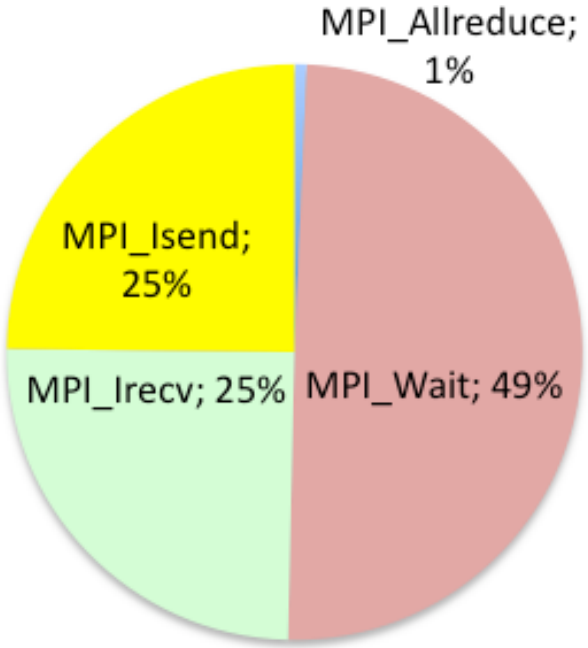
MILC Characteristics



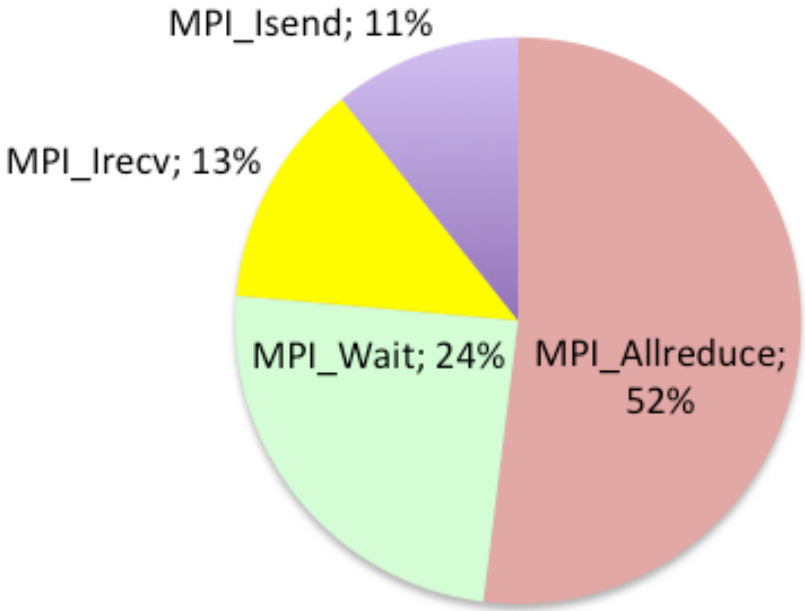
Communication topology for MILC from IPM on Franklin.

MILC Characteristics

MPI Calls by Count

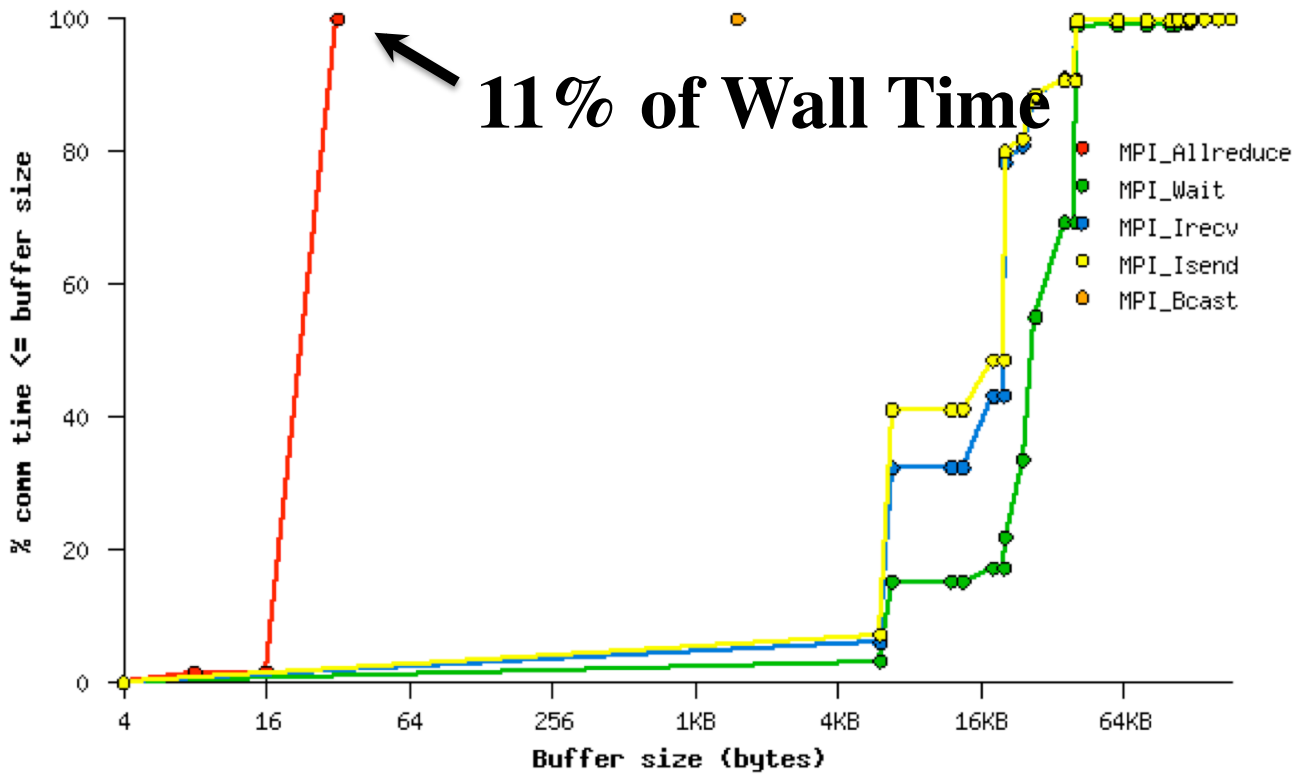


MPI Calls by Time



IPM Data for MILC on 1024 cores of Franklin.

MILC Characteristics



MPI message buffer size distribution based on time for MILC on Franklin from IPM.

MILC: Performance

P	Power5 Bassi		IBM BG/P		Opteron Jaguar		Opteron Franklin	
	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.	GFlops	Effic.
256	488	25%	113*	13*%	203	9%	291	22%
1024	n/a		456*	13*%	513	6%	1101	21%
8192	n/a		n/a		3179	5%	5783	14%

- "Multicore effect" largest for all NERSC benchmarks.
- Does not use QC SSE on Jaguar

*Estimated

What MILC Adds to NERSC-6

- CG solver with small subgrid sizes used stresses interconnect with small messages for both point-to-point and collective operations;
- Extremely dependent on memory bandwidth and prefetching;
  - Large dual-core->quad-core performance reduction.
- High computational intensity;
- Used in NSF Trac-I benchmarking.