Science Driven System Architectures

(and a few words about benchmarking)

John Shalf

NERSC Users Group Meeting

Princeton Plasma Physics Laboratory

June 2006

U.S. Department of Energy

Office of Science

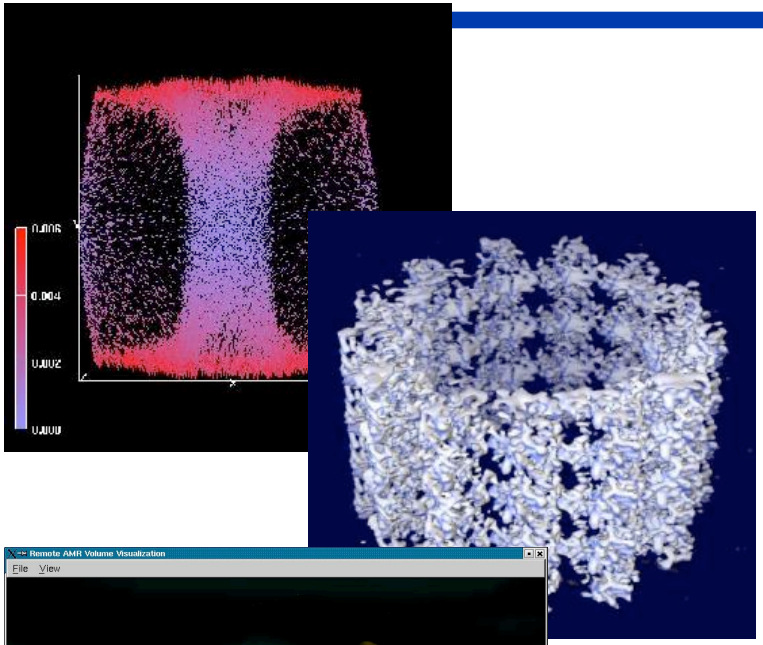
Overview

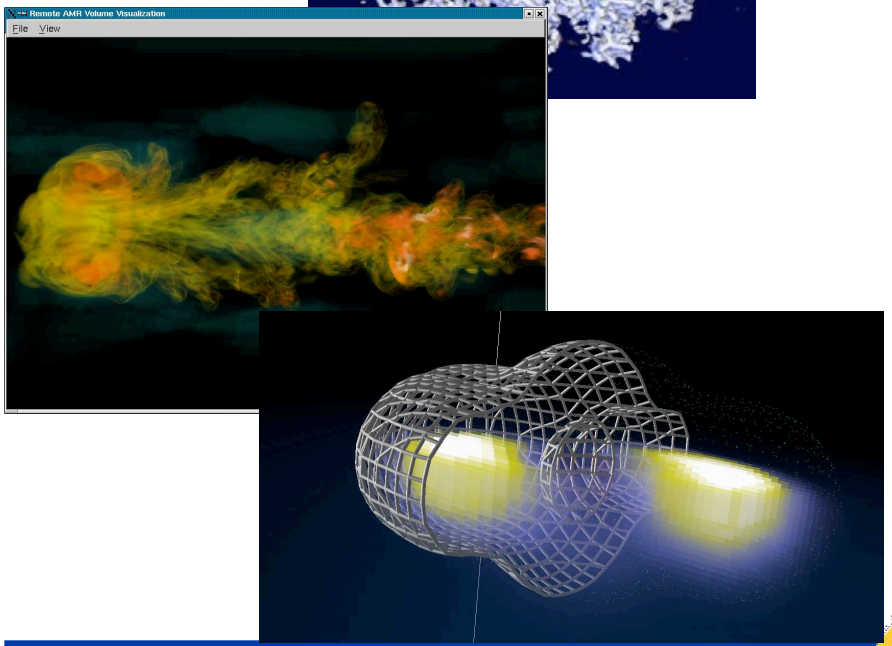
- Overview of the SSP for procurement benchmarks
- Recent Benchmarking Activities
- Workload Characterization
- Wild Ideas

U.S. Department of Energy

NERSC Mission

Berkeley Lab





- Support large scale computational science that cannot be done elsewhere
- Support wide variety of science and computational methods
- Provide a stable production environment to deliver these services

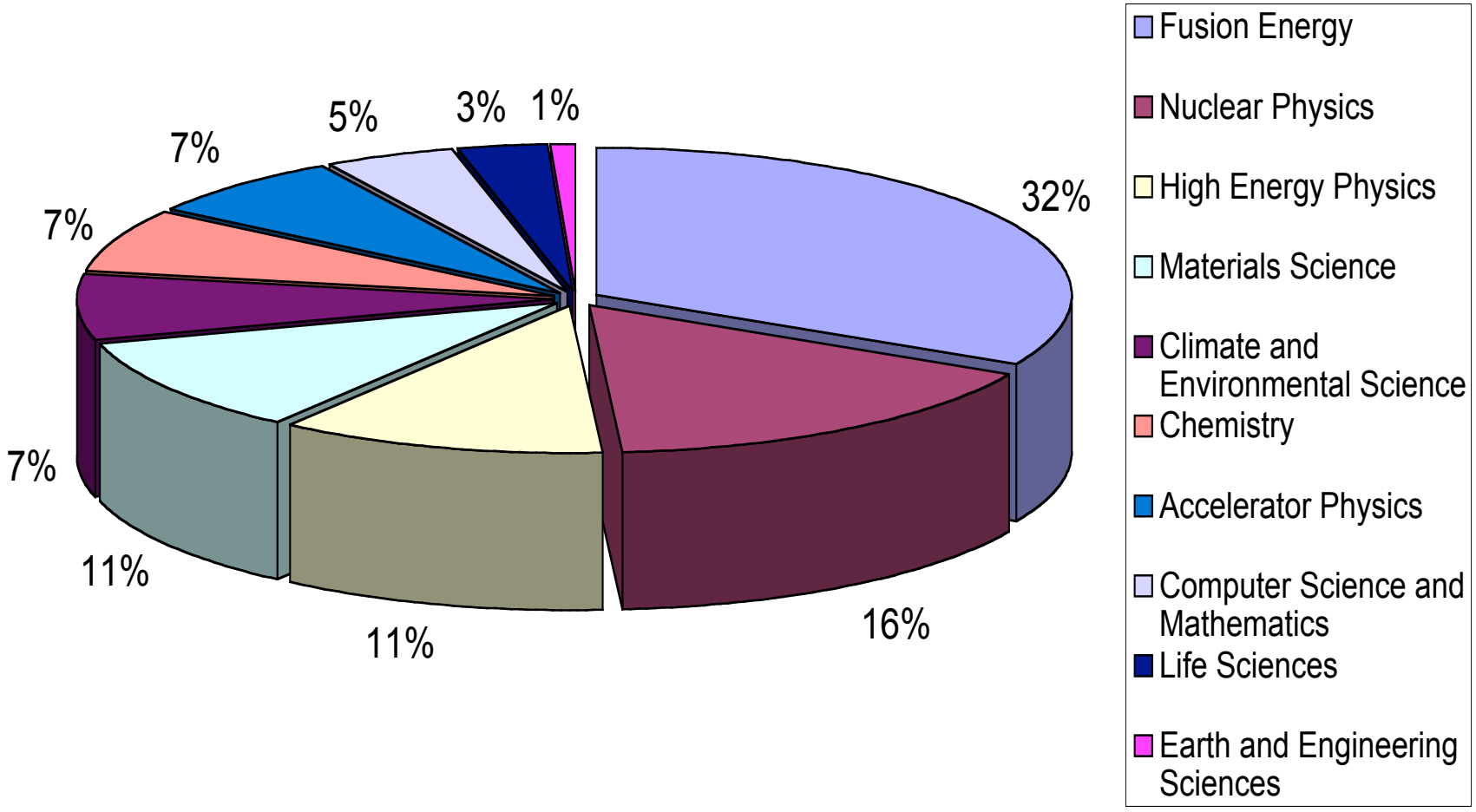
NERSC

U.S. Department of Energy

Office of Science

NERSC's Diverse Workload

Berkeley Lab



NERSC

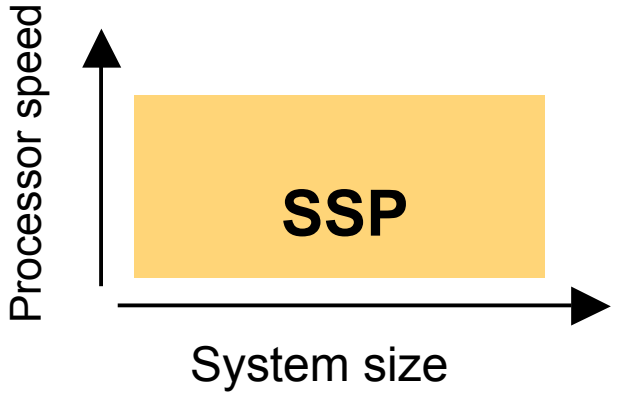
U.S. Department of Energy
Office of Science

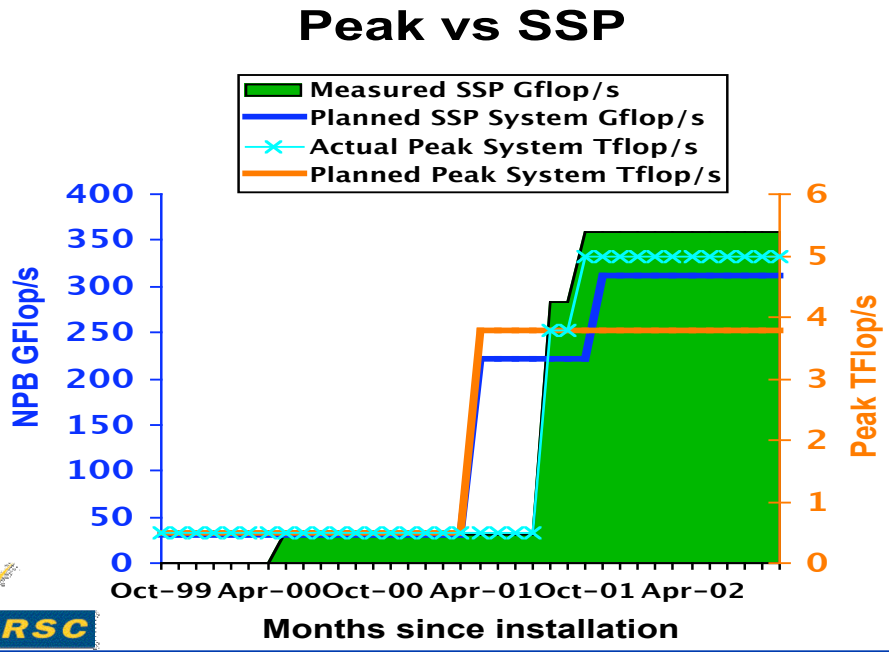
Benchmarking for NERSC System Procurements

- Require a uniform/scientific metric for system "value" over the lifetime of the system that;
  - Assesses effective/delivered system performance
  - Representative of NERSC workload
  - Takes into account system availability and delivery time
  - Focus on the total value of the system to the DOE science community!
- Full Application based benchmark methodology
  - SSP: Sustained System Performance
  - ESP: Effective System Performance
- Same methodology (SSP/ESP) employed for "validation" of the delivered system
  - Factory testing
  - Acceptance testing
  - Continuing testing through the lifetime of the system to assess impact of all system upgrades

SSP Metric

- 7 production applications provide representative subset of NERSC workload
  - Immunity to performance "tweaking"
  - Jobs scaled to match typical/target problem sizes
  - Emphasis on capability jobs
- Uses weighted harmonic mean of job performance
  - *add wallclock times together and divide by total flop count*
- Total "value" is the area under the SSP performance curve!





U.S. Department of Energy
Office of Science

SSP v2 Applications

BERKELEY LAB

Application	Scientific Discipline	Algorithm or Method	MPI Tasks	System Size	Wallclock Timing (sec)
GTC*	**Plasma Physics – (SciDAC)**	**Particle-in-cell**	256	10^7 ions	1682
MADCAP*	**Cosmology (SciDAC)**	**Matrix inversion**	484	40000x 40000	903
Milc*	**Particle Physics (SciDAC)**	**Lattice QCD**	512	32^3x64	1031
NAMD	**Biophysics**	**Molecular dynamics**	1024	92224 atoms	379
NWChem	**Chemistry**	**Density functional**	256	125 atoms	2367
Paratec*	**Material Science**	**Density functional**	128	432 atoms	1386
SEAM*	**Climate (SciDAC)**	**Finite element**	1024	30 days	494

NERSC

ESP

- Throughput of system under normal operating conditions (nontrivial)
  - Batch Scheduler efficiency and validation
  - Job Launcher efficiency
  - Effect of job fragmentation on system performance
    - Issues with < full bisection interconnects
    - Even fat-trees suffer from fragmentation issues
    - Job migration overhead (remediation)

U.S. Department of Energy
Office of Science

Application Selection Issues

Berkeley Lab

- It is difficult to get good coverage
  - Some scientists will not part with “crown jewels”
  - Hopelessly **un**-portable codes
  - Huge time investment for porting and packaging
  - Tuning requirements for novel/unique architectures
  - Difficult for vendors to find test systems of appropriate scale
    - Must test applications at reasonable/native scale
  - Rotation of benchmarks to prevent “performance islands”
  - Rotation of benchmarks to follow workload trends
    - SSP applications will have turn-over as science evolves
  - Vendor “non-compliance” during bidding process
    - Motivates us to simplify benchmarking procedures and methods
- Perhaps we need a persistent effort to manage the SSP?

NERSC

U.S. Department of Energy
Office of Science

Emerging Concerns *(apps not in SSP/ESP)*

Berkeley Lab

- Some good science doesn't scale to thousands of processors
  - AMR
    - Load balancing
    - Locality constraints for prolongation and restriction
    - Pointer chasing (and lots of it!) (Little's Law limitations)
  - Sparse Matrix / SuperLU
    - Domain decomp limits strong scaling efficiency
- Emerging issues with existing applications
  - Implicit Methods
    - Vector inner product required by Krylov subspace algorithms is hampered by latency-bound fast global reductions at massive parallelism
  - Climate Models
    - When science that depends on parameter studies and ensemble runs, capacity and capability are intimately linked! *(capacity vs. capability is a bogus metric)*
- Growth in experimental and sensor data processing requires more attention to I/O performance and global filesystems

NERSC

U.S. Department of Energy
Office of Science

Science Driven System Architecture (SDSA)

Berkeley Lab

- **Stewardship of NERSC SSP Benchmark suite** *(Harvey Wasserman, Lenny Oliker)*
  - Workload characterization (benchmarks only sensible in context of the workload they are intended to model:
  - Selection and packaging of benchmarks + collaboration with other govt. agencies
  - Benchmarking and data collection on available systems (no surprises)
- **Development of New Benchmark Areas** *(Mike Welcome, Hongzhang Shan)*
  - I/O Benchmarking
  - AMR Benchmarks
- **Performance Modeling** *(Erich Strohmaier, Andrew Canning)*
  - Develop microbenchmarks that act as proxies to full application code
  - Develop performance predictive performance models that enable us to predict performance of systems that do not yet exist
  - Use predictive performance models to answer “what-if” architectural questions.
- **Algorithm Tracking and Computer Architecture Evaluation** *(Lin-Wang Wang, Esmond Ng)*
  - What are current resource requirements for current algorithms and how will they affect future computer system architectures?
  - How will future system architecture choices affect the development of future numerical algorithms?
- **Vendor Engagement** *(everyone)*
  - Vendor development cycle 18-24 months!
  - Provide detailed performance analysis & discussion w/vendors to effect changes early in the development cycle *(when it really matters)!*
  - Bring feedback from vendors back to application groups (vendor code tuning assistance)

NERSC

U.S. Department of Energy
Office of Science

Multicore? *(Or not?)*

BERKELEY LAB

Bassi SSP on Power5+ showed only 8.27% performance degradation when run using dual-core mode.

Test Case	SSP					Geometric Mean SSP Time	SSP % Speedup
	FT Time	MG Time	CAM	GTC	PARATEC		
Power5+ "Sparse" 8 procs, 8 cores (one core per processor)	50.450	13.140	2779.224	1365.530	4962.160	416.170	
Power5+ "Dense" 4 procs, 8 cores (2 cores per processor)	56.820	14.520	2841.340	1485.538	5334.200	450.598	-8.27%
Percent Speedup	-12.63%	-10.50%	-2.24%	-8.79%	-7.50%		

Test Case	Memtest Triad BW	MPITest Max Latency	MPITest Min BW	MPITest Nat. Ord. Ring BW	MPITest Rnd. Ord. Ring BW
Power5+ "Sparse" 8 procs, 8 cores (one core per processor)	10222	0.002056	5032.15837	3187.899174	3184.094637
Power5+ "Dense" 4 procs, 8 cores (2 cores per processor)	6693	0.001937	5405.03093	2750.264652	2514.449009
Percent Speedup	-34.52%	5.79%	7.41%	-13.73%	-21.03%

NERSC

U.S. Department of Energy
Office of Science

Multicore? *(Or not?)*

Berkeley Lab

However, AMD Opteron X2 shows >50% degradation for many NAS benchmarks compared to single-core???

Test	Unpacked	Packed	Percent
BT	974	622	63.86%
CG	368	272	73.91%
FT	823	520	63.18%
LU	1072	643	59.98%
MG	1411	960	68.04%
SP	576	374	64.93%
Average			62.46%

	Function	Rate (MB/s)
1task	Copy:	3888.9
	Scale:	3915.4
	Add:	3836.1
	Triad:	3831.8
packed	Copy:	2002.9
	Scale:	2034.9
	Add:	1989.7
	Triad:	1946.0
Difference	Copy:	51.5%
	Scale:	52.0%
	Add:	51.9%
	Triad:	50.8%

- 100% would be perfect speedup
- 50% means each core runs 50% as fast as single core case
- <50% means performance degrades **more** than memory bandwidth alone can explain!!!

NERSC

U.S. Department of Energy
Office of Science

Understanding Interconnects

- CPU clock scaling bonanza has ended
  - Heat density
  - New physics below 90nm *(departure from bulk material properties)*
- Yet, by end of decade mission critical applications expected to have 100X computational demands of current levels *(PITAC Report, Feb 1999)*
- The path forward for high end computing is increasingly reliant on massive parallelism
  - Petascale platforms will likely have hundreds of thousands of processors
  - System costs and performance may soon be dominated by interconnect
- What kind of an interconnect is required for a >100k processor system?
  - What topological requirements? *(fully connected, mesh)*
  - Bandwidth/Latency characteristics?
  - Specialized support for collective communications?

U.S. Department of Energy
Office of Science

Questions

(How do we determine appropriate interconnect requirements?)

Berkeley Lab

- Topology: *will the apps inform us what kind of topology to use?*
  - Crossbars: Not scalable
  - Fat-Trees: Cost scales superlinearly with number of processors
  - Lower Degree Interconnects: *(n-Dim Mesh, Torus, Hypercube, Cayley)*
    - Costs scale linearly with number of processors
    - Problems with application mapping/scheduling fault tolerance
- Bandwidth/Latency/Overhead
  - Which is most important? *(trick question: they are intimately connected)*
  - Requirements for a "balanced" machine? *(eg. performance is not dominated by communication costs)*
- Collectives
  - How important/what type?
  - Do they deserve a dedicated interconnect?
  - Should we put floating point hardware into the NIC?

NERSC

Approach

- Identify candidate set of “Ultrascale Applications” that span scientific disciplines
  - Applications demanding enough to require Ultrascale computing resources
  - Applications that are capable of scaling up to hundreds of thousands of processors
  - *Not every application is “Ultrascale!” (not all good science is Ultrascale)*

- Find communication profiling methodology that is
  - Scalable: N*eed to be able to run for a long time with many processors. Traces are too large*
  - Non-invasive: *Some of these codes are large and can be difficult to instrument even using automated tools*
  - Low-impact on performance: *Full scale apps… not proxies!*

U.S. Department of Energy
Office of Science

IPM *(the "hammer")*

BERKELEY LAB

Developed by David Skinner, NERSC

Integrated Performance Monitoring

- portable, lightweight, scalable profiling
- fast hash method
- profiles MPI topology
- profiles code regions
- open source

```
MPI_Pcontrol(1,"W");
 …code…
MPI_Pcontrol(-1,"W");
```

```
##############################################
# IPMv0.7 :: csnode041 256 tasks  ES/ESOS
# madbench.x (completed) 10/27/04/14:45:56
#
#       <mpi>       <user>       <wall> (sec)
#      171.67       352.16       393.80
#...
##############################################
# W
#       <mpi>       <user>       <wall> (sec)
#       36.40       198.00       198.36
#
# call              [time]       %mpi    %wall
# MPI_Reduce      2.395e+01    65.8      6.1
# MPI_Recv        9.625e+00    26.4      2.4
# MPI_Send        2.708e+00     7.4      0.7
# MPI_Testall     7.310e-02     0.2      0.0
# MPI_Isend       2.597e-02     0.1      0.0
##############################################
...
```

NERSC

U.S. Department of Energy
Office of Science

Application Overview *(the "nails")*

BERKELEY LAB

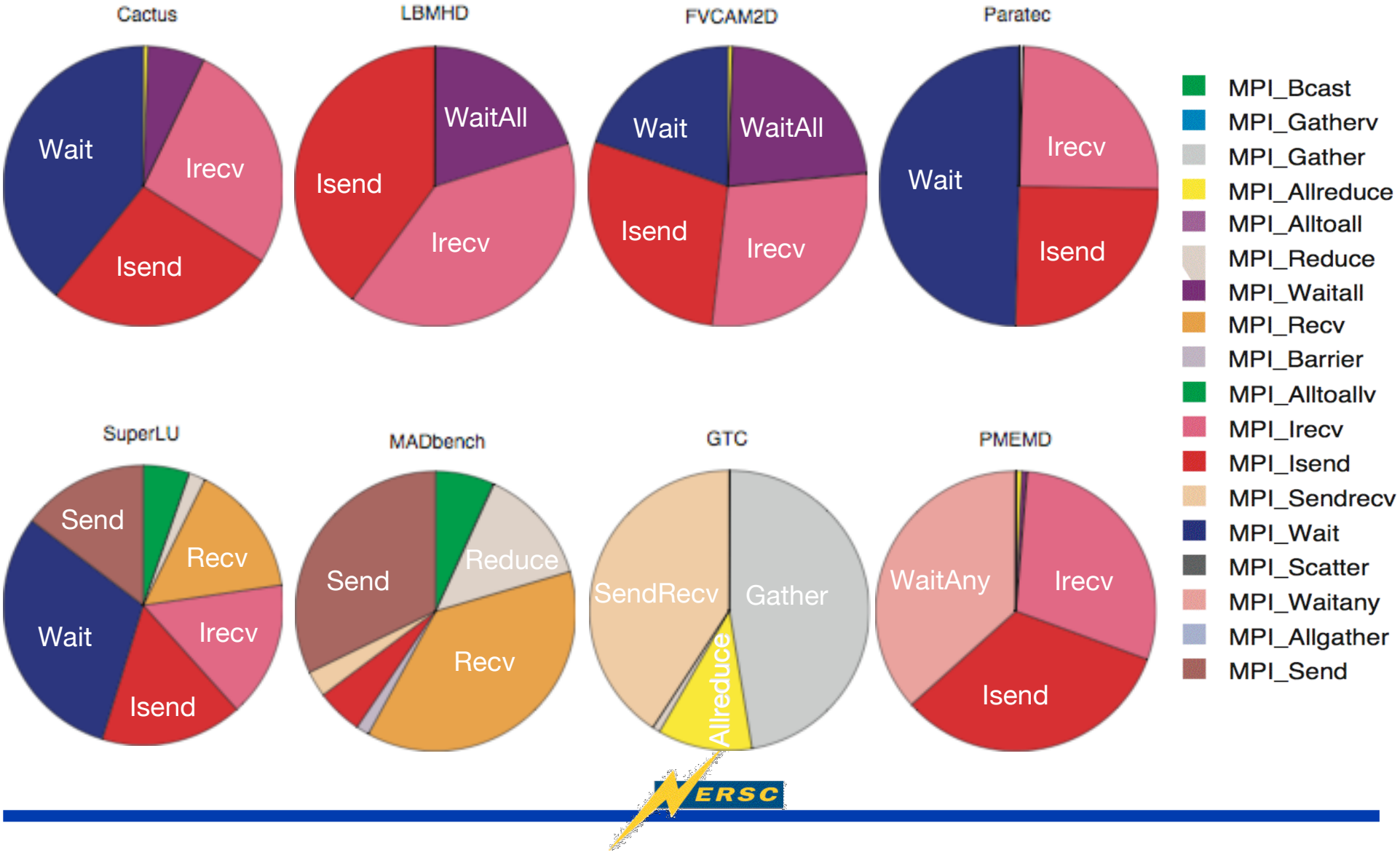
NAME	Discipline	Problem/Method	Structure
MADCAP	Cosmology	CMB Analysis	Dense Matrix
FVCAM	Climate Modeling	AGCM	3D Grid
CACTUS	Astrophysics	General Relativity	3D Grid
LBMHD	Plasma Physics	MHD	2D/3D Lattice
GTC	Magnetic Fusion	Vlasov-Poisson	Particle in Cell
PARATEC	Material Science	DFT	Fourier/Grid
SuperLU	Multi-Discipline	LU Factorization	Sparse Matrix
PMEMD	Life Sciences	Molecular Dynamics	Particle

NERSC

U.S. Department of Energy
Office of Science

Call Counts

Berkeley Lab



NERSC

U.S. Department of Energy

Office of Science

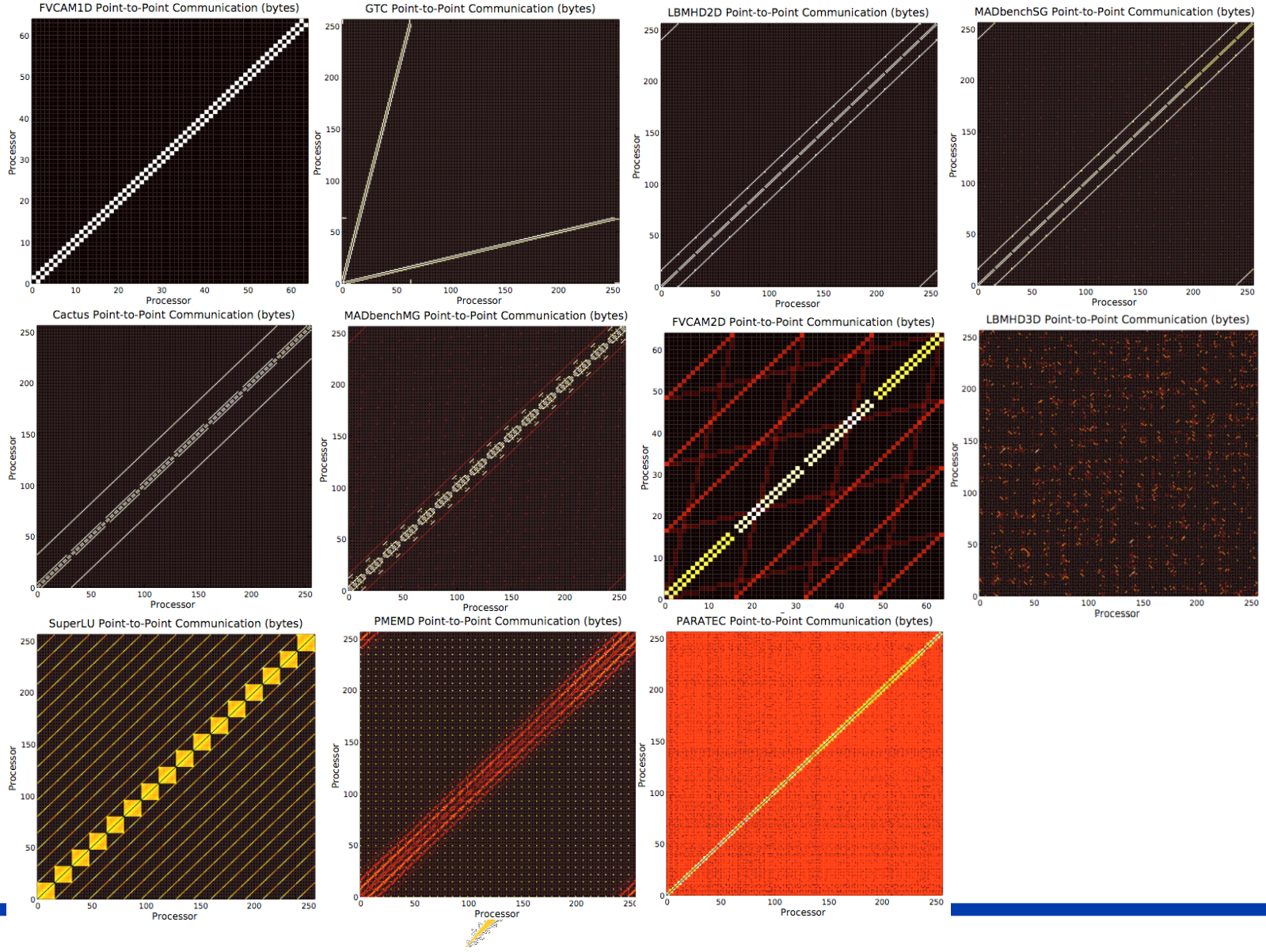
P2P Topology Overview

BERKELEY LAB

Max

Total Message Volume

0

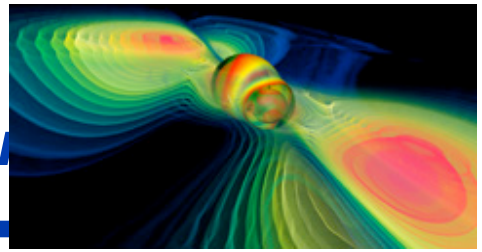


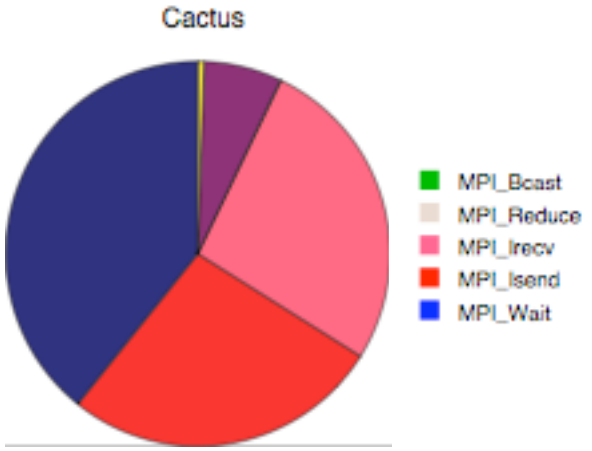
U.S. Department of Energy

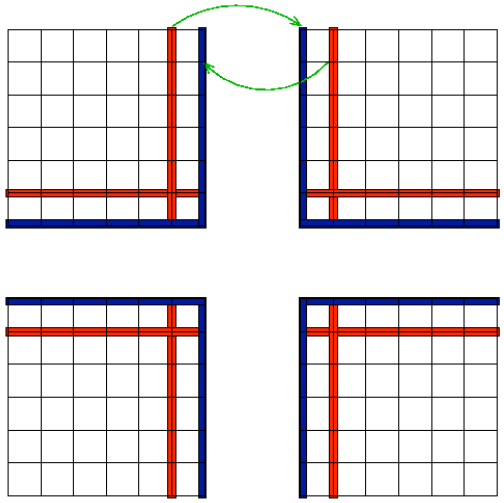
Office of Science

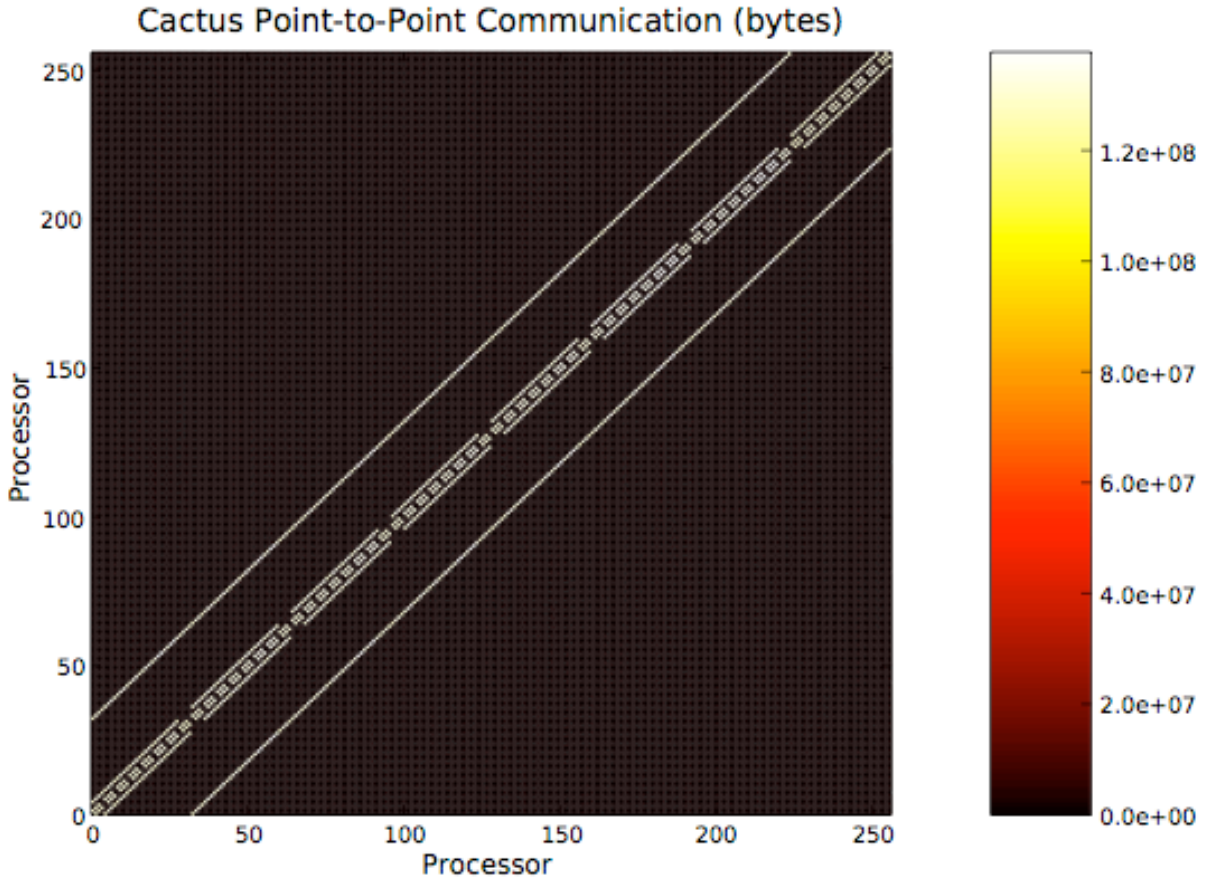
Cactus Communication

PDE Solvers on Block Structured Gr







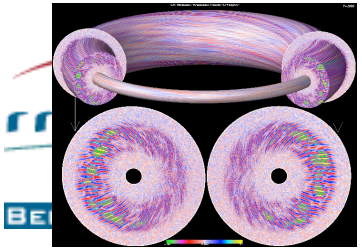


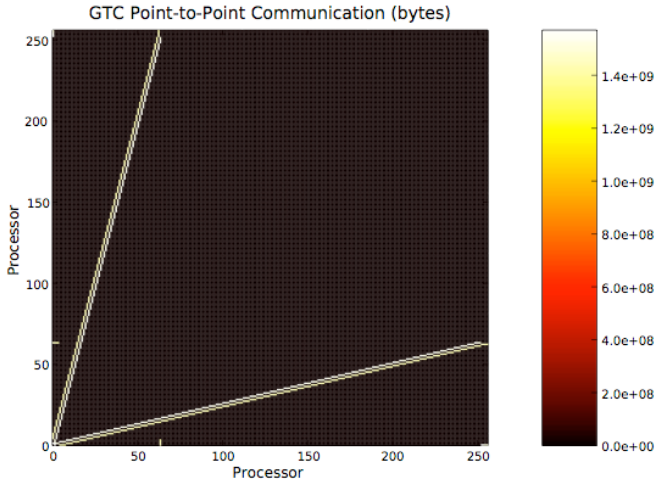
NERSC

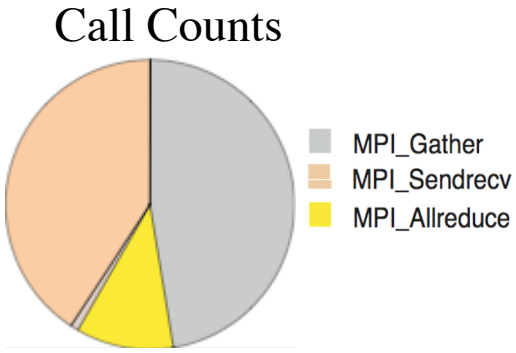
U.S. Department of Energy

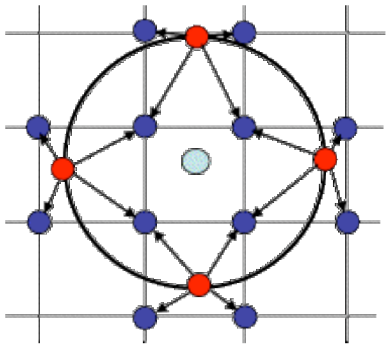
Office of Science

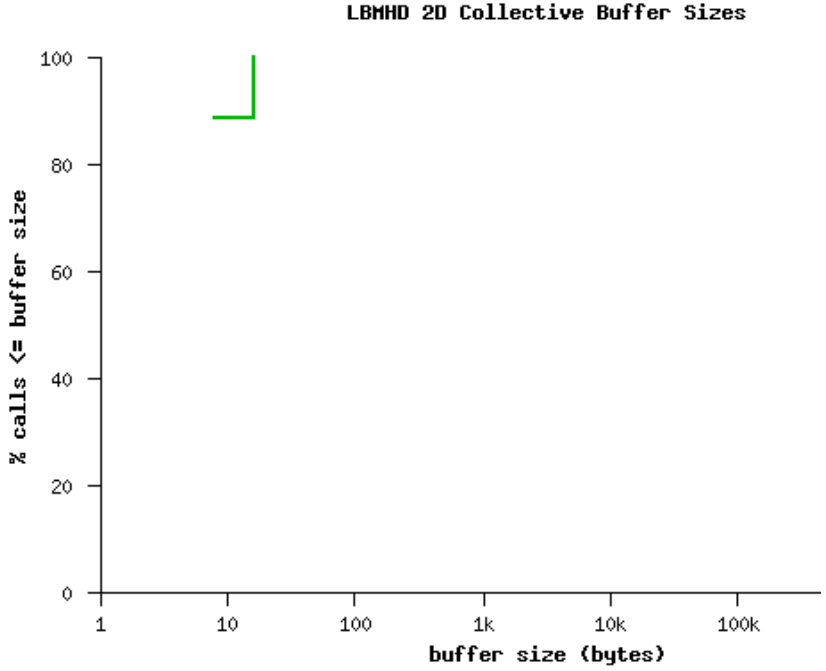
GTC Communication

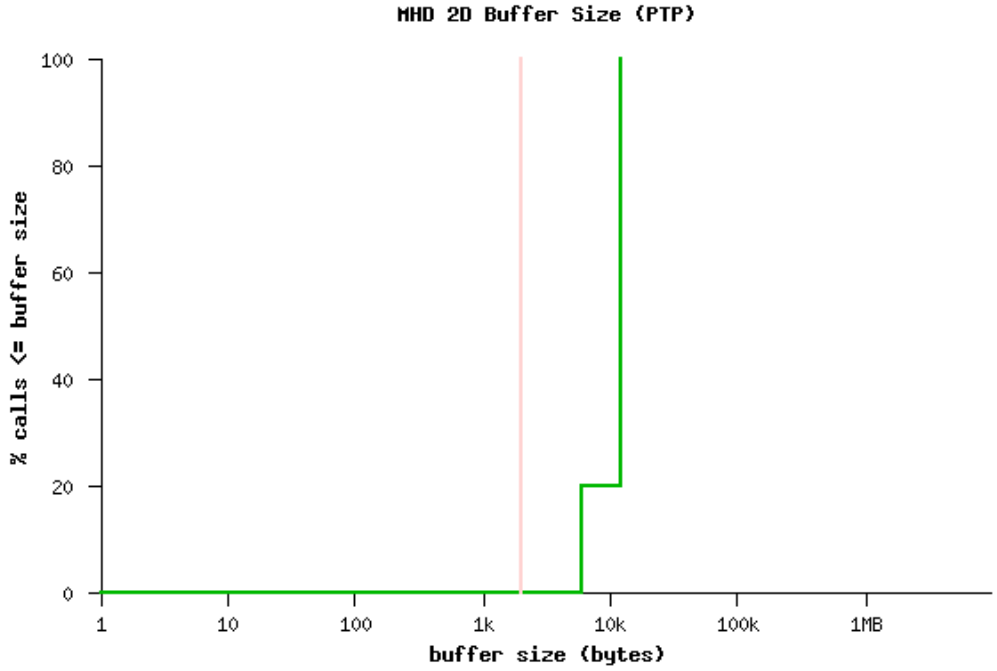








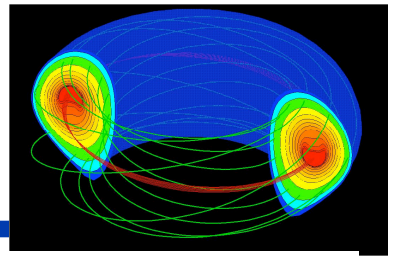


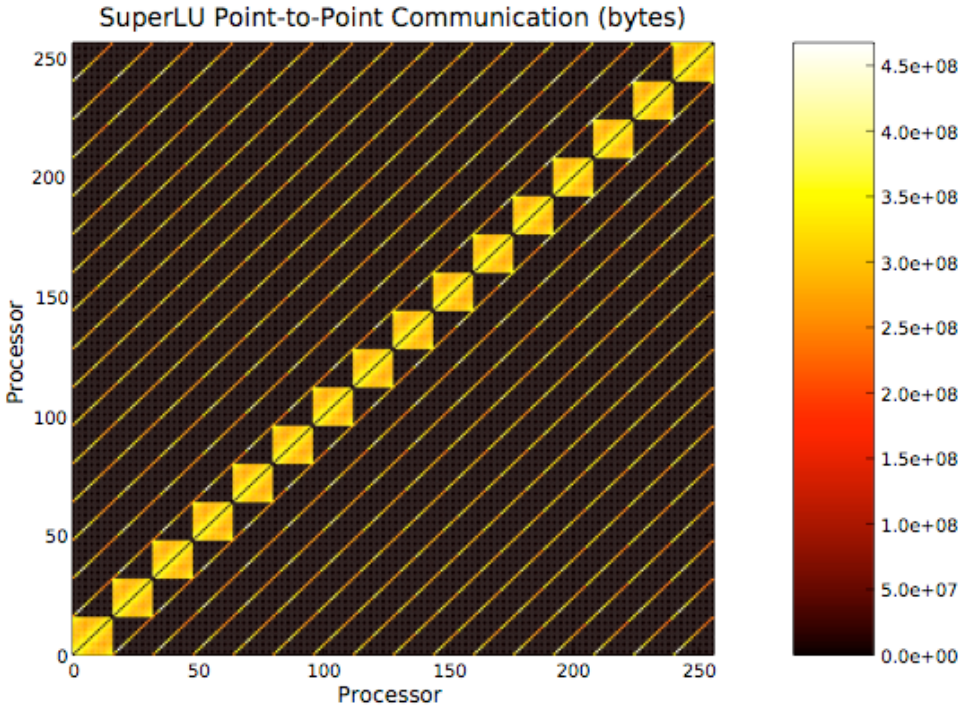


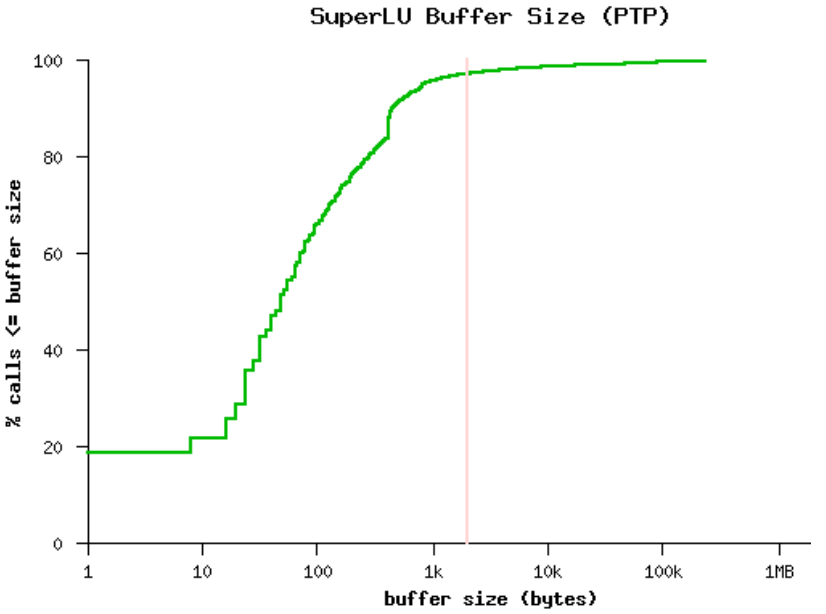
U.S. Department of Energy

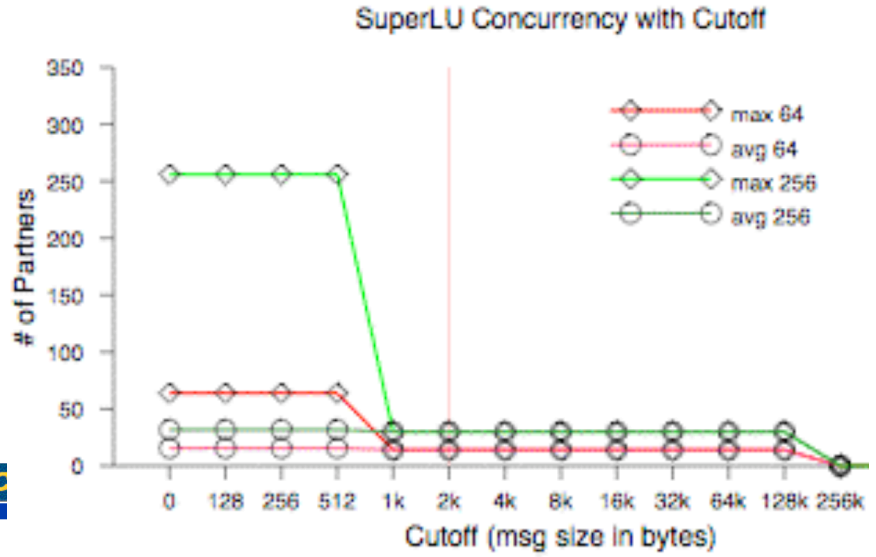
Office of Science

SuperLU Communication







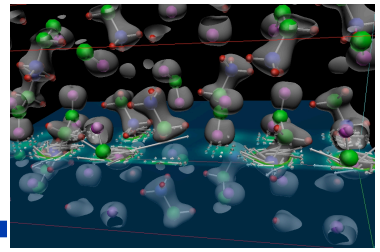


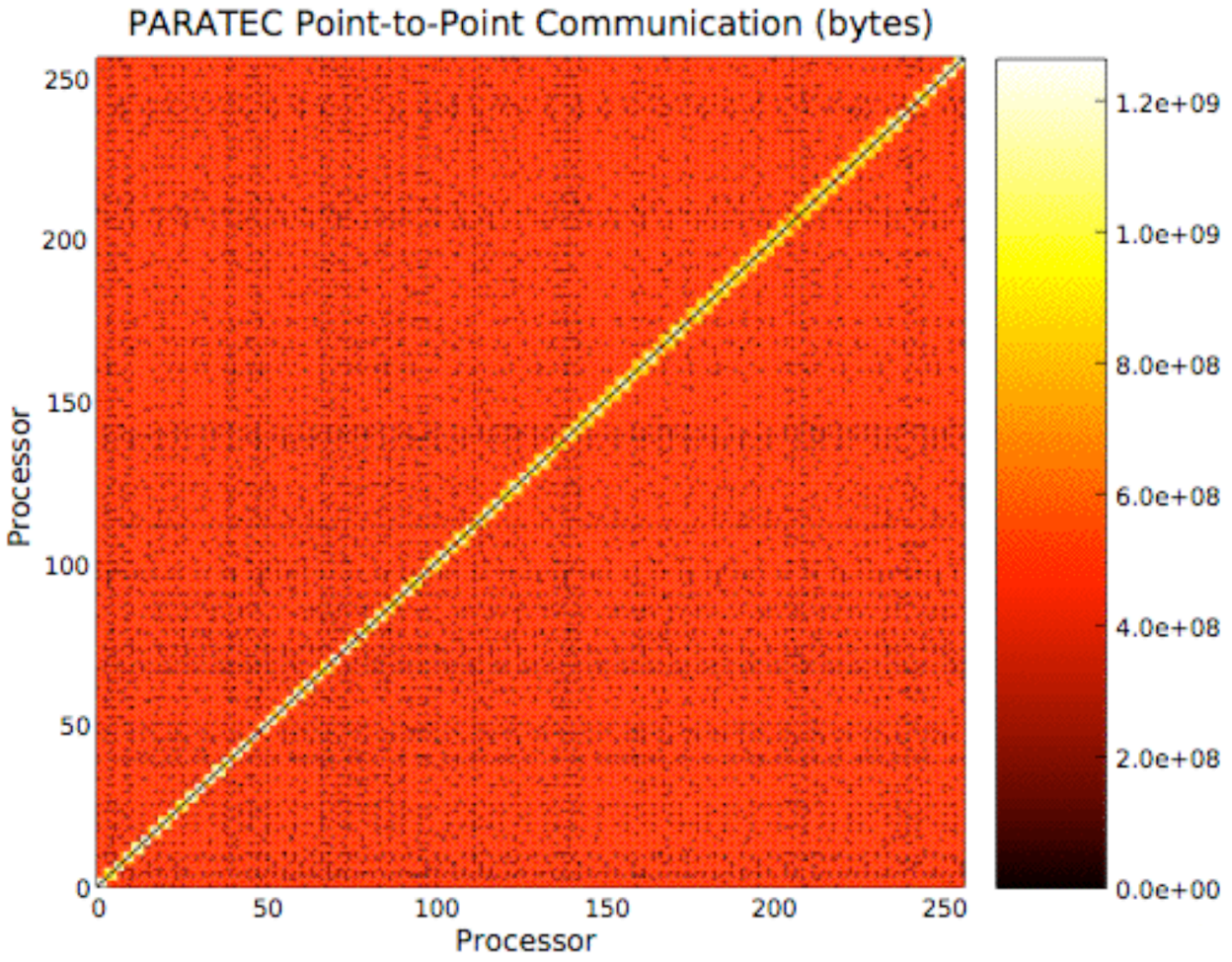
NERSC

U.S. Department of Energy

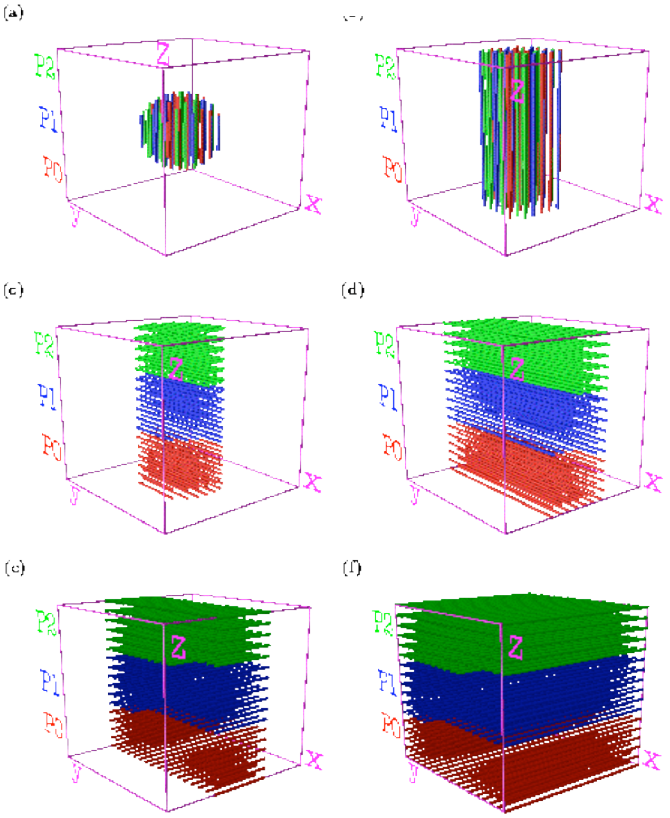
Office of Science

PARATEC Communication





3D FFT

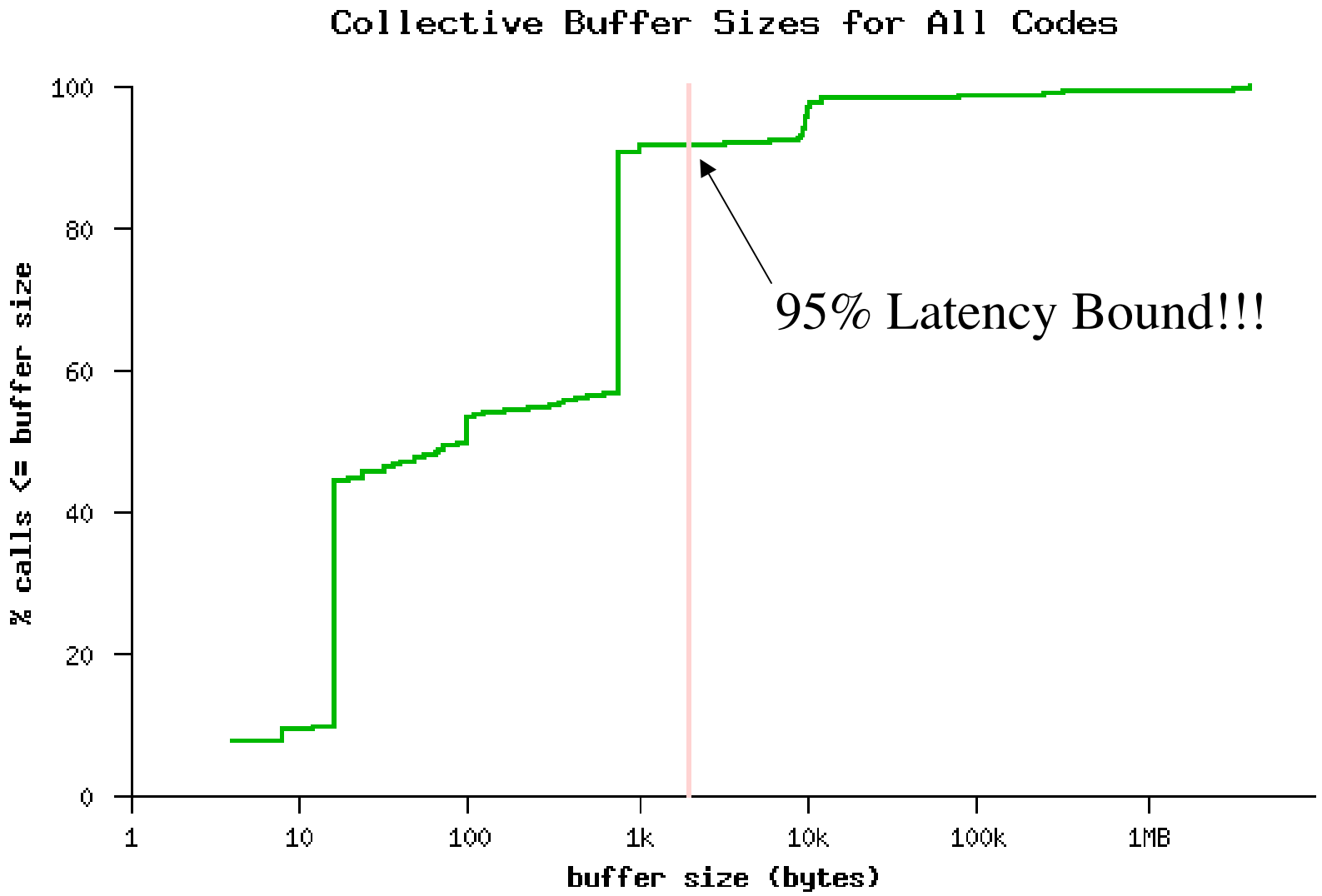


NERSC

U.S. Department of Energy
Office of Science

Collective Buffer Sizes

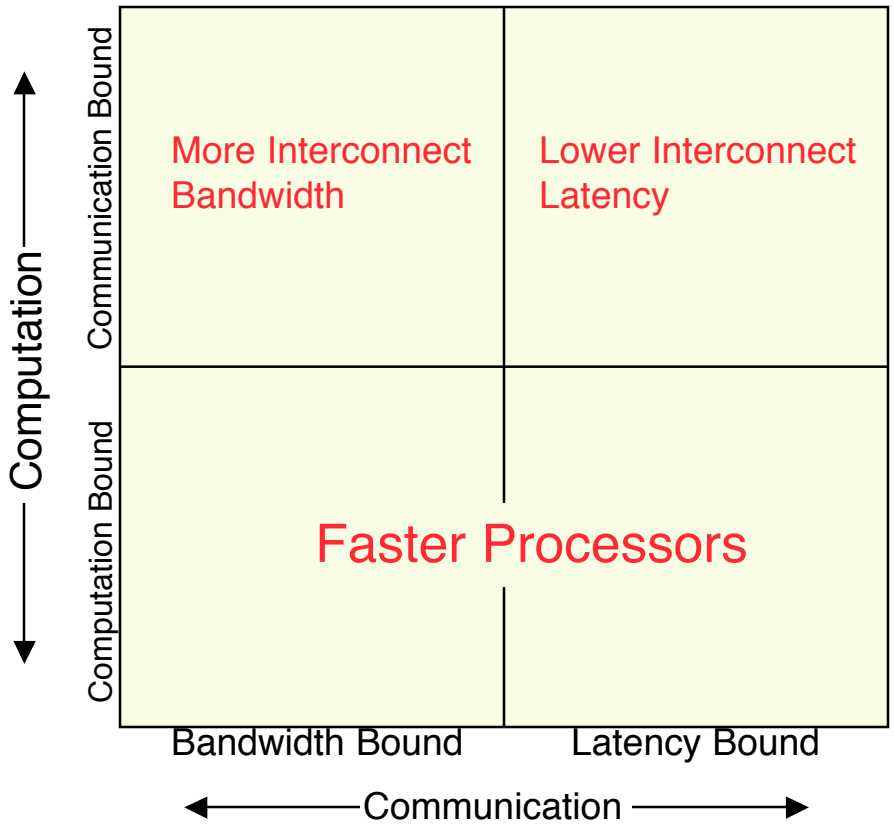
BERKELEY LAB

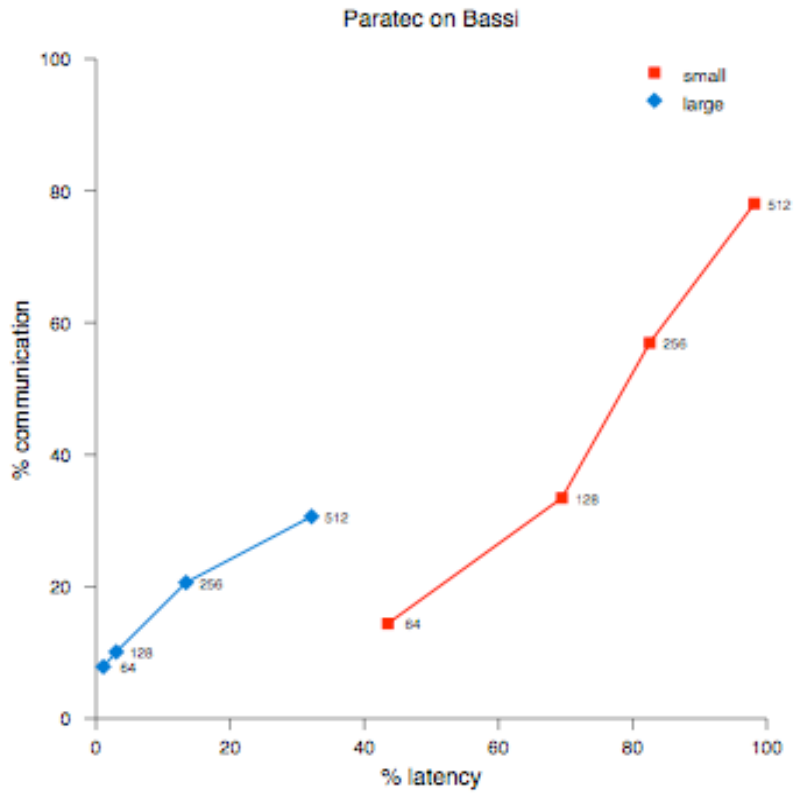


U.S. Department of Energy
Office of Science

Latency/Balance Diagrams

BERKELEY LAB





NERSC

Revisiting Original Questions

- Topology
  - Most codes require far less than full connectivity
    - PARATEC is the only code requiring full connectivity
    - Many require low degree (<12 neighbors)
  - Codes with low topological degree of communication not necessarily isomorphic to a mesh!
    - Non-isotropic communication pattern
    - Non-uniform requirements
- Bandwidth/Delay/Overhead requirements
  - Scalable codes primarily bandwidth-bound messages
  - Average message sizes several Kbytes
- Collectives
  - Most payloads less than 1k *(8-100 bytes!)*
    - Well below the bandwidth delay product
    - Primarily latency-bound *(requires different kind of interconnect)*
  - Math operations limited primarily to reductions involving sum, max, and min operations.
  - Deserves a dedicated network *(significantly different reqs.)*

Algorithm Tracking

- Can't track an algorithm until you define what constitutes an algorithm that is worth tracking
  - Which algorithms or libraries are important?
  - Workload analysis precedes drill-down into algorithm

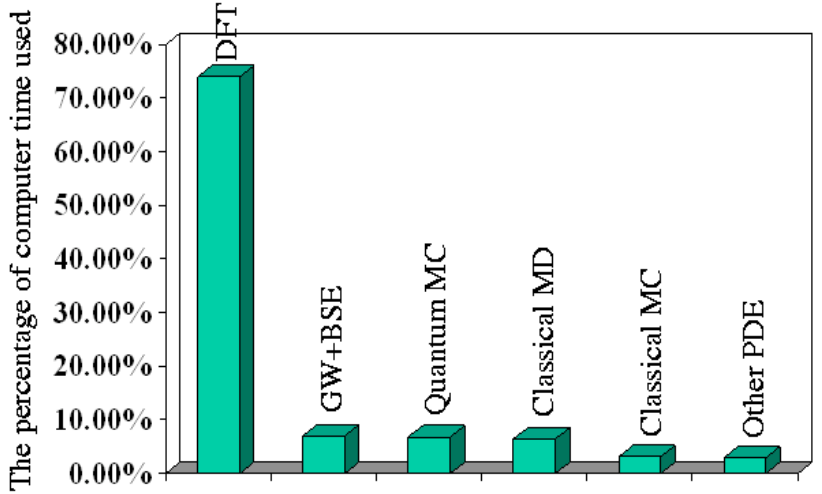
U.S. Department of Energy

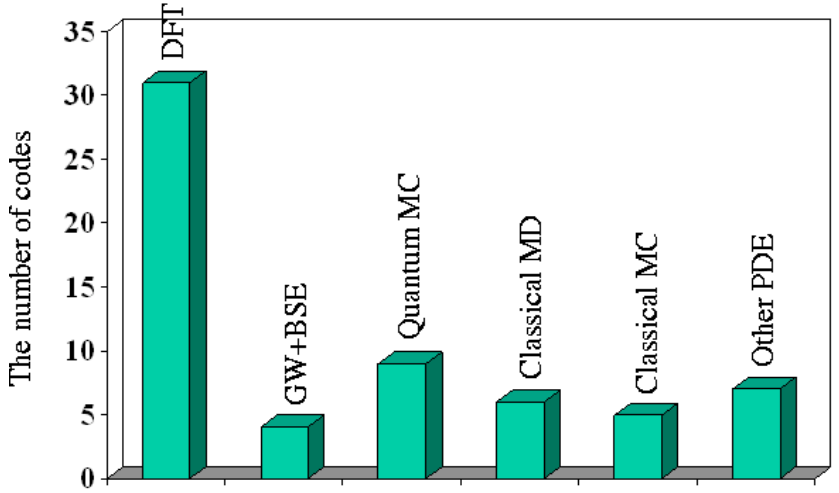
Office of Science

Materials Science Workload

BERKELEY LAB

rank	N-user	HPC(KH)	Code name	Narrative
1	23	704.5	VASP	Vienna, planewave DFT
2	3	632.4	LSMS	ORNL, local mutiple scattering DFT
3	4	216.0	PEscan	LBNL, planewave Folded Spectrum Method
4	3	186.0	Scarlet	Berkeley, quantum transport, atomic basis
5	2	151.0	ALCMD	Classical MD
6	6	142.5	Paratec	Berkeley, planewave DFT
7	2	125.0	Qbox	LLNL, massively parallel planewave DFT
8	3	124.0	CMAT	NREL, Coulomb calc, and CI
9	6	116.5	PWSCF	Planewave DFT
10	2	109.0	FLAPW	NW Univ, FLAPW
11	4	89.00	GW	Berkeley, GW
12	2	84.0	PEtot	LBNL, planewave DFT
13	4	79.0	SIESTA	Spain, local basis/orbital DFT
14	3	78.75	TBMD	Tight-binding Molecular Dynamics
15	2	78.75	BO-LSD-MD	Planewave DFT MD
16	1	70.0	CASINO	Quantum MC
17	3	59.0	DLPOLY	Classical MD
18	4	55.5	NWchem	NWchem DFT, planewav and local basis
19	1	50.0	CHAMP	Quantum Monte Carlo
20	2	43.75	Multigrid	North Carolina, real space DFT
21	1	43.75	XqmmmX	QM/MM, PW for QM, Amber for MM
22	1	40.0	SSEqmc	Quamtum MC for lattice spin
23	3	37.5	Parsec	Minnesota, real space DFT
24	3	37.5	RGWBS	Minnesota, G-space TDLDA
25	1	35.0	LMTO-SIC	LMTO, DFT, Self-interaction correction
26	1	35.0	Psi-Mag	Classical spin-MC
27	1	35.0	SGF	Surface green's function
28	1	35.0	Moldy	Force field classical MC
29	1	35.0	OLCAO	LCAO TB like DFT
30	3	33.0	NAMD	Classical force field MD
31	3	33.0	BSE	Berkeley, Bethe-Salpeter Eq.
32	2	30.0	FDTDn	Genetic Algorithm, classical el-mag, energy
33	1	30.0	Flair	FLAPW LDA
34	1	30.0	CF-Monte-Carlo	Quatum Monte Carlo for model H
35	2	30.0	Abinit	Planewave DFT
36	1	28.0	Dmol3	Atomic orbital DFT
37	1	28.0	FLMTO	LMTO DFT
38	1	26.0	Polycrys	Plastic Deformation, grain boundary
39	1	25.0	LAMMPS	Particle method molecular dynamics
40	1	25.0	PWDFT	Planewave DFT
41	1	25.0	QMC-DCA	Quantum MC and dynamic cluster
42	1	25.0	SPF	Classical spin-phonon-Fermion MC
43	1	20.0	Qdot	Quantum MC for 2D electron
44	1	20.0	FEFF	Multi-scattering Green's func. electronic st
45	1	20.0	becmw	Ordinary Differential Equation.
46	1	20.0	Qmhubbard	Quantum Mechanics, hubbard model
47	1	15.0	Freepar_x	Special algorithm Quantum MC
48	1	14.0	TransG98	Green's func. transport, Gaussian basis
49	1	12.5	AndyS	George Tech. Planewave DFT
50	1	12.5	CL-GCMD	Classical MD, complex liquid
51	1	12.5	Hollicita	2D vertices PDE
52	1	12.0	Mol-dyn	London Eq. MD, particle method
53	1	10.0	Moment	Maxwell Eq, FFT, photonics
54	1	10.0	TMM	Transfer matrix for Maxwell Eq, photonic
55	1	10.0	BEST	Planewave DFT
56	2	8.0	AMber	Classical force field MD
57	1	8.0	MC	Classical MC for vortices
58	1	6.0	ARPEsmpi	Multiple scattering photoemission
59	1	5.0	AFQMC	William Mary, Quantum Monte Carlo
60	1	5.0	WIEN2K	Vienna, FLAPW
61	1	3.75	Hartree	Real-space Hubbard Model, FH, LDA
62	1	2.0	Ginger	Classical particle
Total		4100.0		





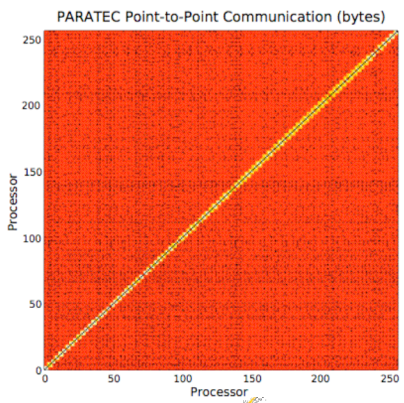
NERSC

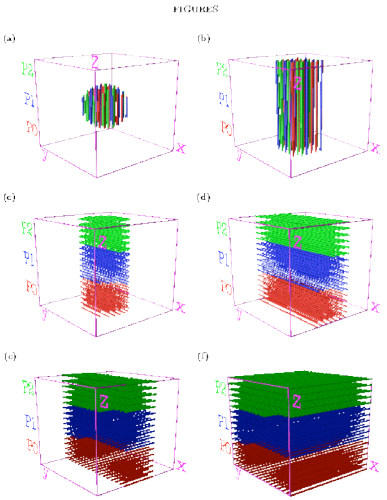
U.S. Department of Energy
Office of Science

Materials Science Workload

BERKELEY LAB

- Materials Science Workload
  - Lin-Wang Wang ERCAP analysis
  - PARATEC is a good proxy for MatSci apps.
  - A massively parallel future (Petascale) may push us to methods that exhibit more Spatial Locality in their communication patterns
  - Are real-space methods a good replacement, or is it just going to waste more CPU cycles to get the same quality answer?





NERSC