BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Performance Expectations and Experience at Scale

Gregory Bauer

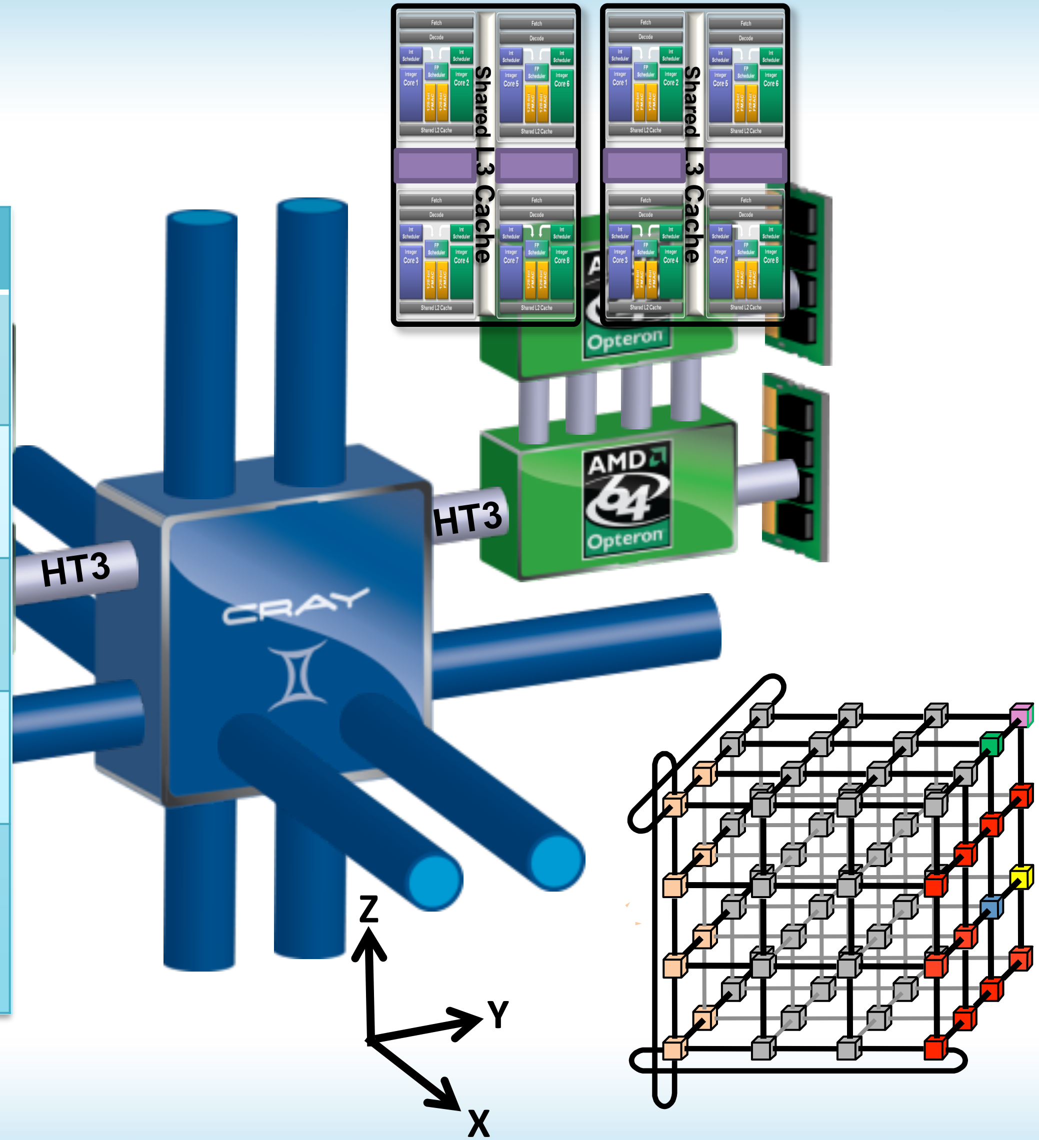
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Outline

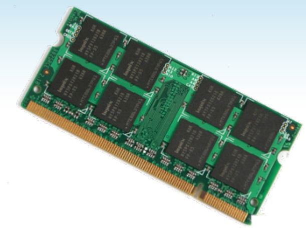
- Performance Expectations
  - Peak
  - Realized
- Application Performance
  - SPP
  - Optimizations and Settings

Peak Performance

Node Characteristics	
Number of Core Modules*	16
Peak Performance	313 Gflops/sec
Memory Size	64 GB per node
Memory Bandwidth (Peak)	102 GB/sec
Interconnect Injection Bandwidth (Peak)	9.6 GB/sec per direction



Memory Subsystem Performance



- Stride-1 word load/store/copy (32 MiB data):
  - 1 int core r/w/c: 3.8 / 4 / 3 GB/s
  - 16 int cores (1 IL) r/w/c: 32 / 16 / 9.6 GB/s
- CL latency (random pointer chase, 1 GiB data):
  - 1 int core : 110 ns
  - 16 int cores (1 IL): 257 ns
  - 32 int cores (2 IL): 258 ns

Measured with Netgauge 2.4.7, pattern memory: stream and pchase

- STREAM Triad
  - 1 core : 13 GB/s
  - 8 cores (1 IL): 34 GB/s (32 GB/s with 4 cores)
  - 16 cores (2 IL): 68 GB/s

Compute performance

- Single core (two integer cores)

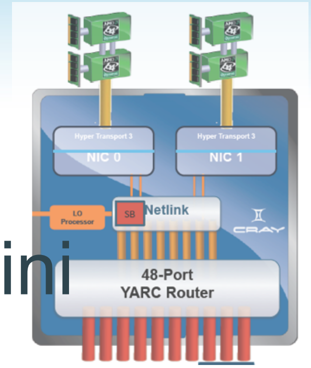
MF/s	1 thread	2 threads	peak
DGEMM (FMA4)	13127	16353	18400

- Single IL processor (MPI tasks x OMP threads)

MF/s	16 x 1	2 x 8	1 x 16
DGEMM FMA4)	126335	125292	120517

- Representative and not optimal performance

Network performance



- IMB PingPong between nodes sharing gemini
  - latency < 2 us
  - bandwidth > 6 GB/s
- Randomly Ordered Ring on 25710 nodes
  - latency ~ 5 us
  - aggregate bandwidth ~ 3 TB/s
    - somewhat less than realized bi-section bandwidth

Compute Kernel Performance

Table 1: Performance characteristics for simple kernels

kernel	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
triad s	300	407	3958	1.1	0.1	0.1	2.3
triad l	241	156	1574	1.0	0.1	0.1	2.6
stencil s	1089	2508	9172	1.4	0.3	0.5	2.3
stencil l	181	458	1684	1.4	0.3	0.1	2.6
dgemm l	3690	7940	3297	5.0	2.4	1.6	2.3
reg int	2000	0	0	0.0	0.0	0.8	2.6

- s=small, l=large
- CI=Computational Intensity, AI=Algorithmic Intensity
- Hardware performance counter measurements are per integer core.
- MiBPS is from LL_CACHE_MISSES are L2 misses, impacted by prefetching.
- Stream Triad 1919 MiBPS / core with 16 cores.
- AMD Processor adjusts clock frequency between P states depending on thermal/power levels: between 2.3 and 2.6 GHz. Cases that fit in cache or are cache-blocked cause the lower clock state to be used.

Sustained Petascale Performance (SPP)

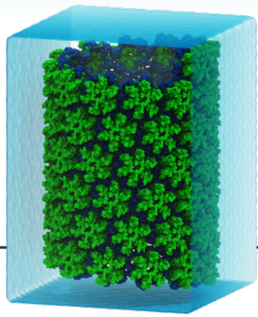
- Full application based benchmark modeled after the NERSC SSP.
- Still a FLOP based benchmark. Validated hand-counts and hardware counts.
- Includes time for IO such as defensive checkpointing, start-up and data.
- Composed of CPU and GPU applications that represent the average workload on the system.

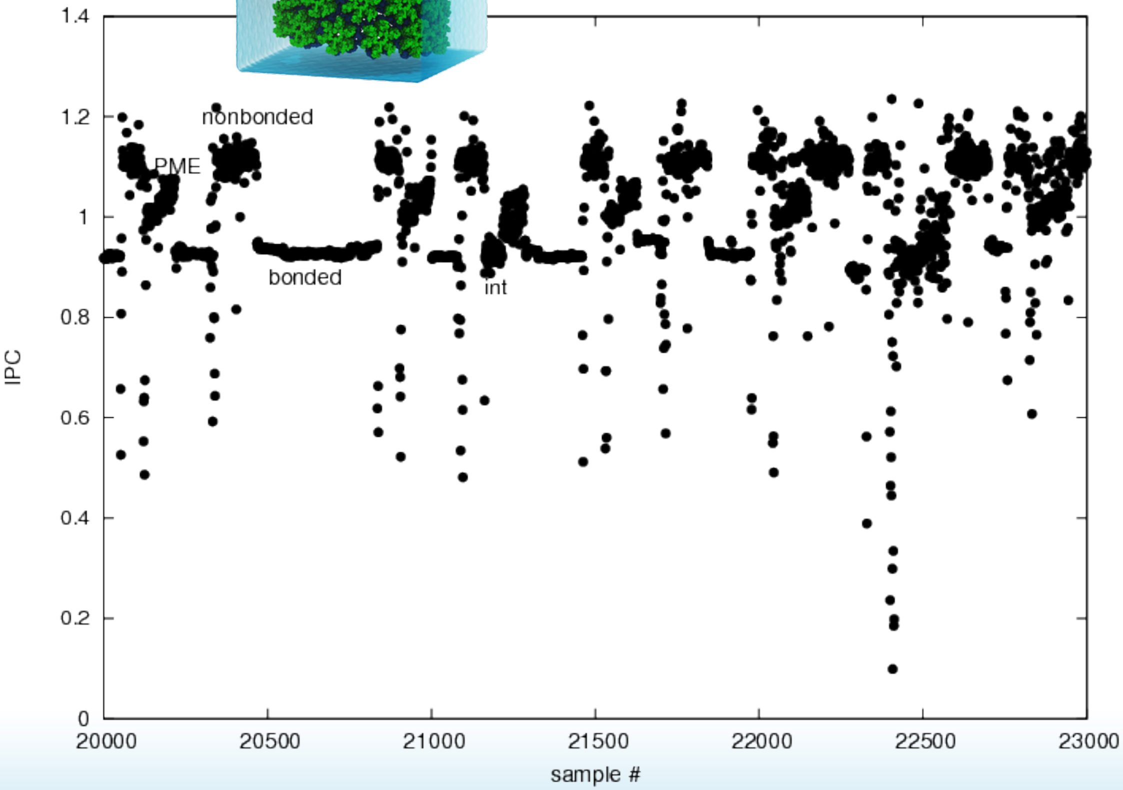
SPP Application Summary

Application	Field of Science	CPU	GPU	Program Model	Compiler	Note
NAMD	Bio-molecular dynamics	✔	✔	Charm++	GNU/C++	ASM, CUDA
QMCPACK	Materials Science	✔	✔	MPI + OpenMP	GNU/C++	Vec.Instrins., CUDA
MILC	Lattice QCD	✔		MPI	GNU/C	ASM
NWCHEM	Quantum Chemistry	✔		GA	PGI/F90,C	
PPM	Astrophysics	✔		MPI + OpenMP	Cray/F90	
SPECFEM3DGLOBE	Geophysics	✔		MPI	Cray/F90	
VPIC	Plasma Physics	✔		MPI + OpenMP	GNU/C	Vec.Instrins.
WRF	Weather	✔		MPI + OpenMP	Cray/F90	
Chroma	Lattice QCD		✔	MPI	GNU/C++	QUDA
GAMESS	Quantum Chemistry		✔	MPI	Cray/F90	OpenACC

NAMD

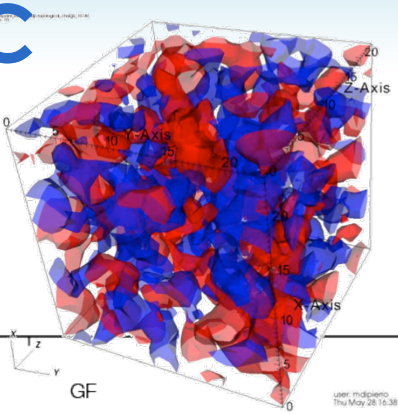
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
nonbonded	2460	1377	7506	1.1	0.2	1.1	2.3
PME	1772	1408	3299	1.7	0.4	0.8	2.3
bonded	1617	723	1821	0.8	0.4	0.7	2.3
integrate	1394	581	4573	0.8	0.1	0.6	2.3



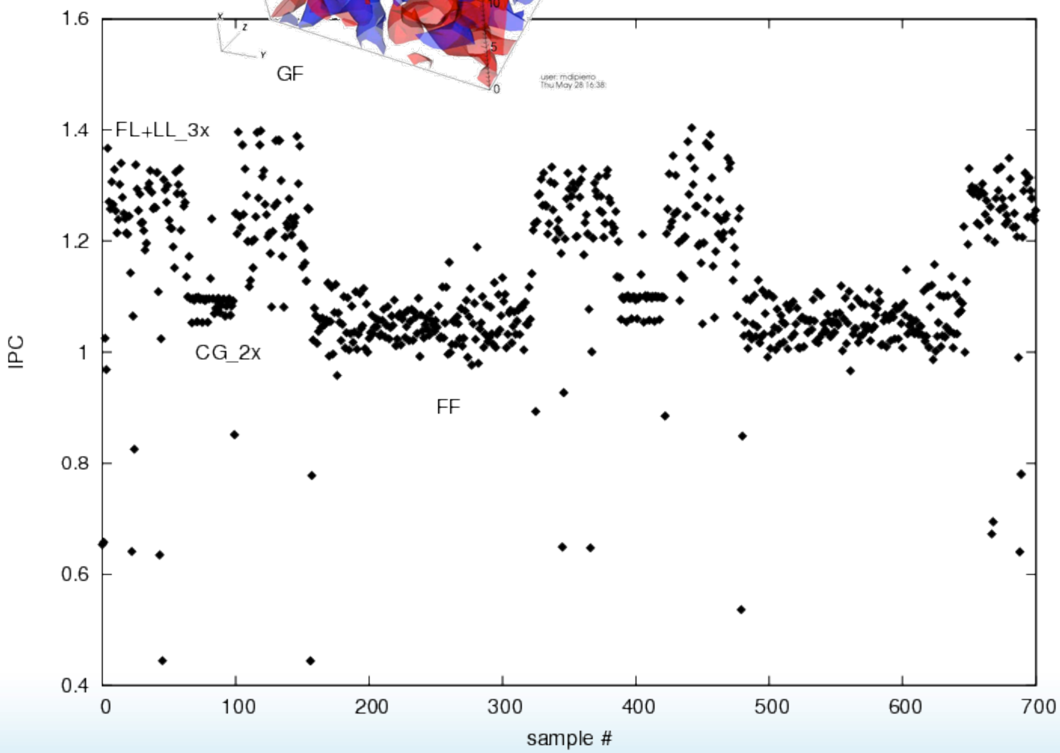


- Dynamic scheduling complicates model
- Excellent cache locality
- PME performs well but will slow down at scale (alltoall)
- Good IPC

MILC



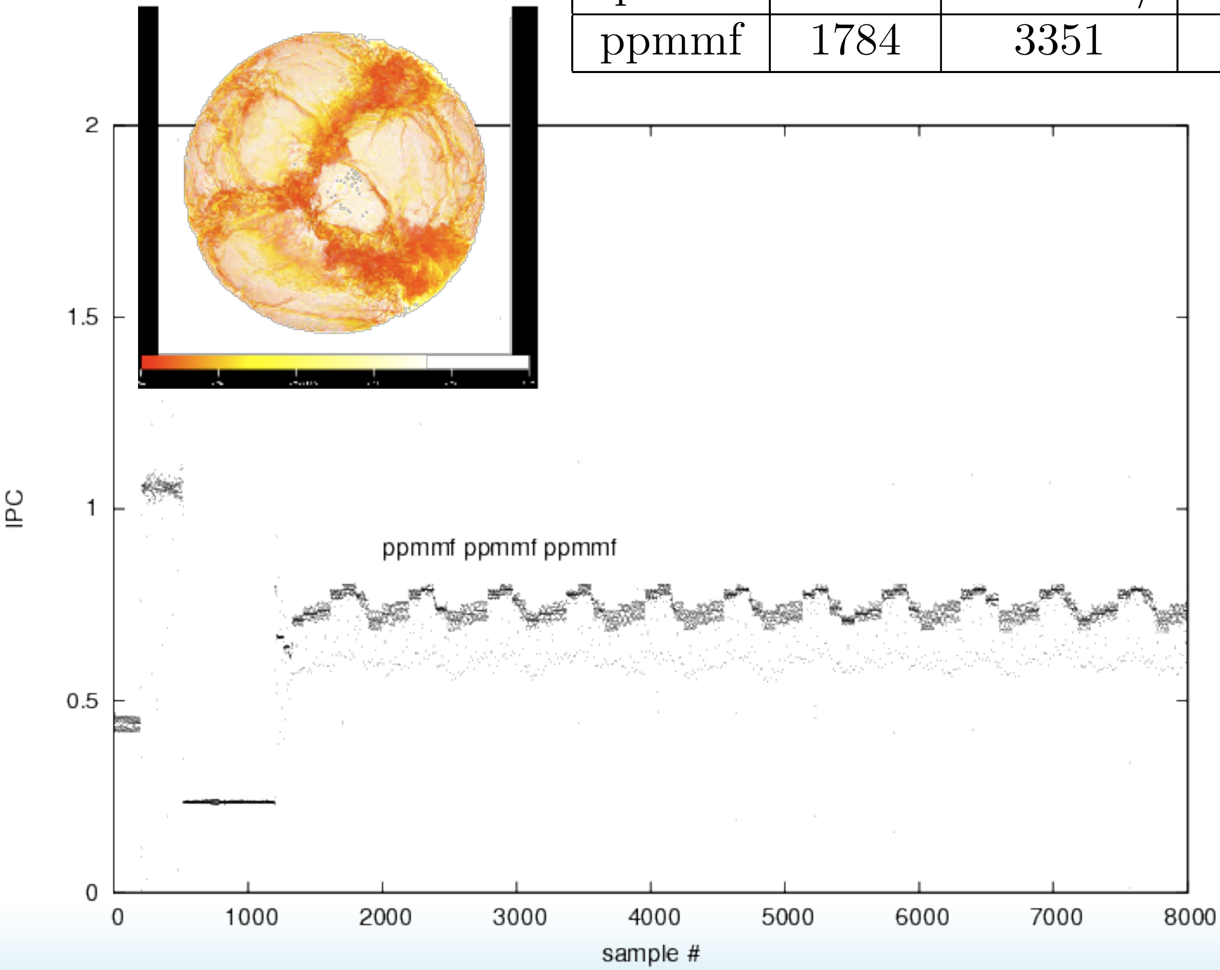
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
LL	1123	707	3179	1.1	0.2	0.5	2.2
FL	1475	1425	3233	1.9	0.4	0.6	2.4
FF	1305	1057	2055	1.2	0.5	0.5	2.4
GF	1414	1087	3719	1.4	0.3	0.6	2.4
CG	1353	1082	3051	1.7	0.4	0.6	2.5



- Five phases, CG most critical at scale
- Low FLOPs and IPC
  - Turbo boost seems to help here!
- Low FLOPs are under investigation (already using SSE)

PPM

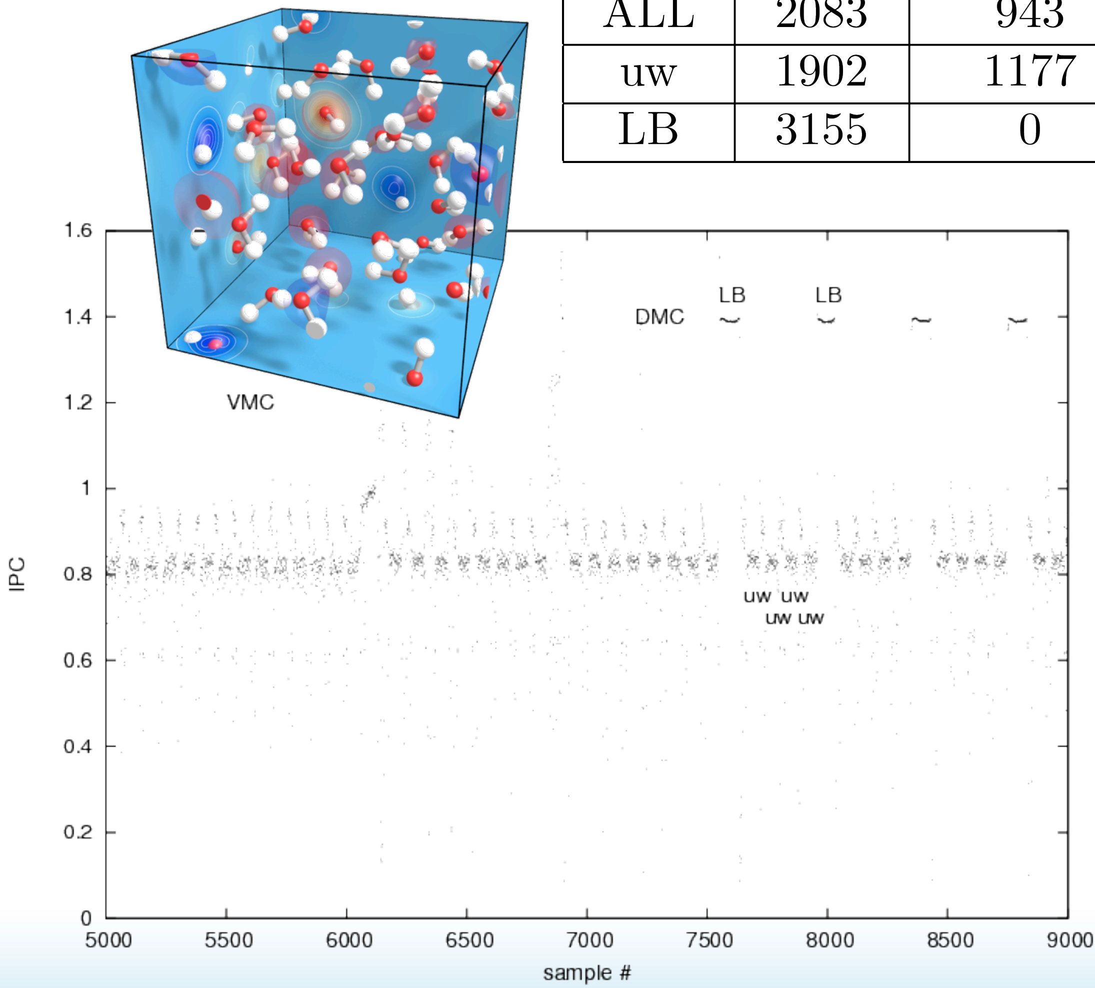
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
ppmmf	1784	3351	2839	3.0	1.1	0.7	2.4



- Many micro-phases
- Hard to instrument
- Very highly optimized by science team
  - Cache blocking
  - High FLOP rate
  - High locality

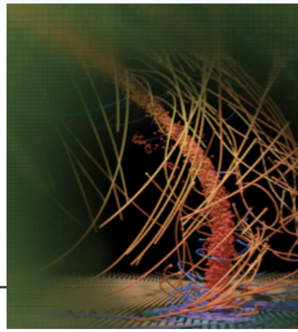
QMCPACK

phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
ALL	2083	943	1933	1.1	0.5	0.9	2.3
uw	1902	1177	2433	1.5	0.5	0.8	2.3
LB	3155	0	18	0.0	0.0	1.4	2.3

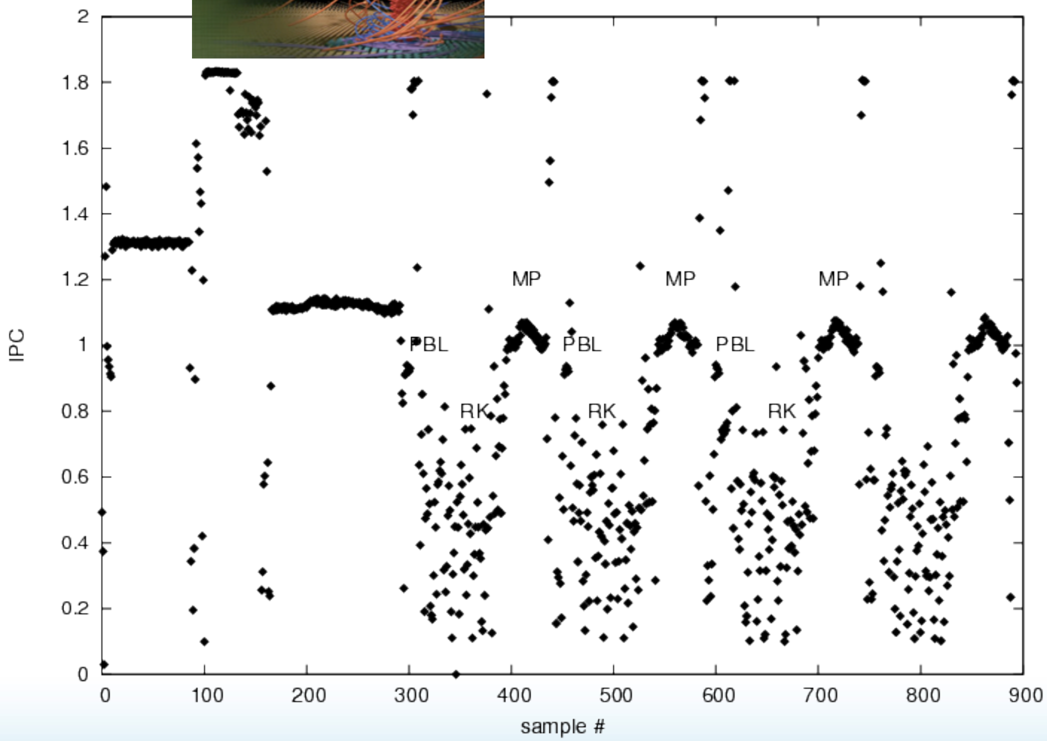


- Variational Monte Carlo initializes
- Performance issues are investigated
- Diffusion Monte Carlo:
  - load balance (LB)
  - update walker (uw)

WRF



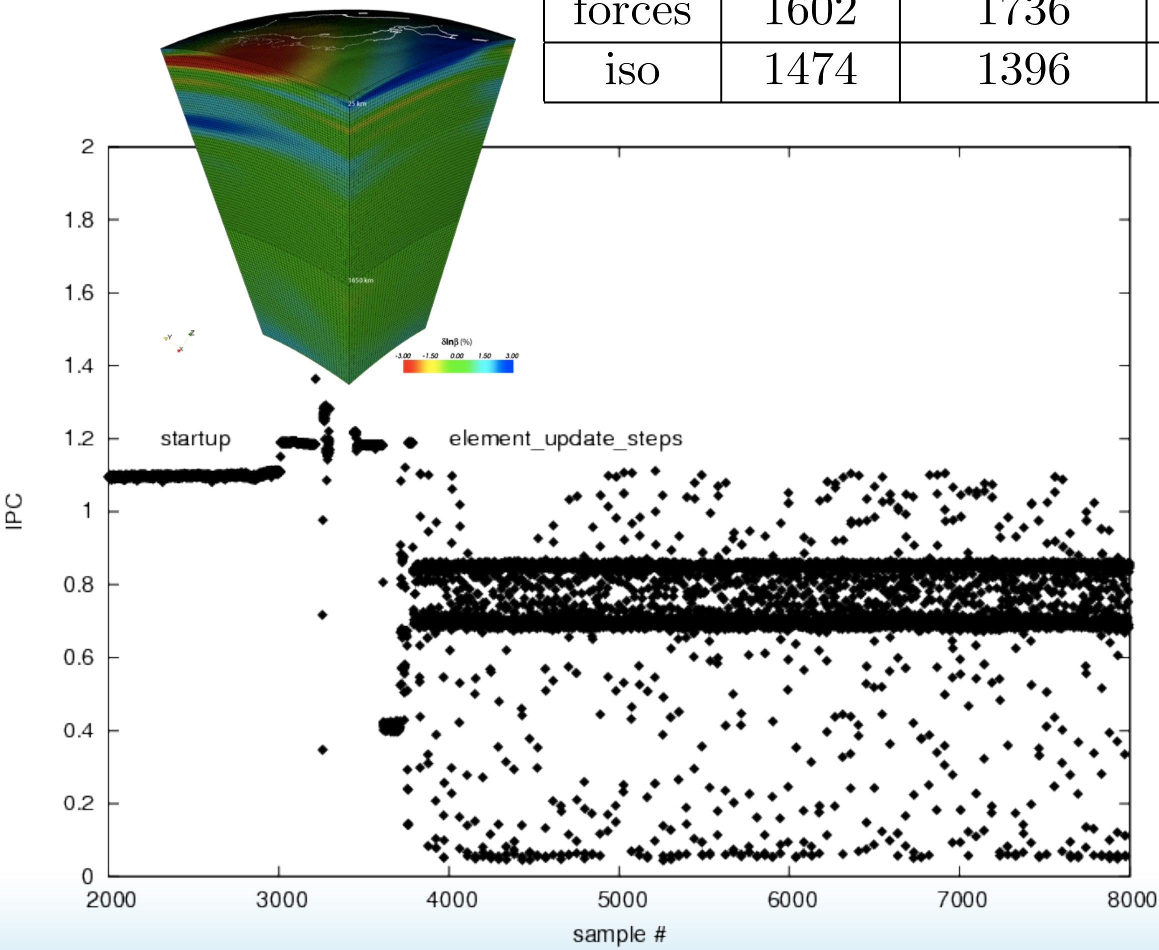
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
MP	2647	590	1288	0.5	0.5	1.0	2.6
PBL	2197	566	4511	0.5	0.1	0.9	2.6
RKt	1328	2695	11842	2.0	0.2	0.6	2.3
RKs	1764	1120	4967	0.8	0.2	0.7	2.5



- Microphysics dominates
  - Low performance, many branches
- Planet Boundary Layer also problematic
  - Turbo Boost helps!
- Runge Kutta is fast
  - High locality

SPECFEM3D

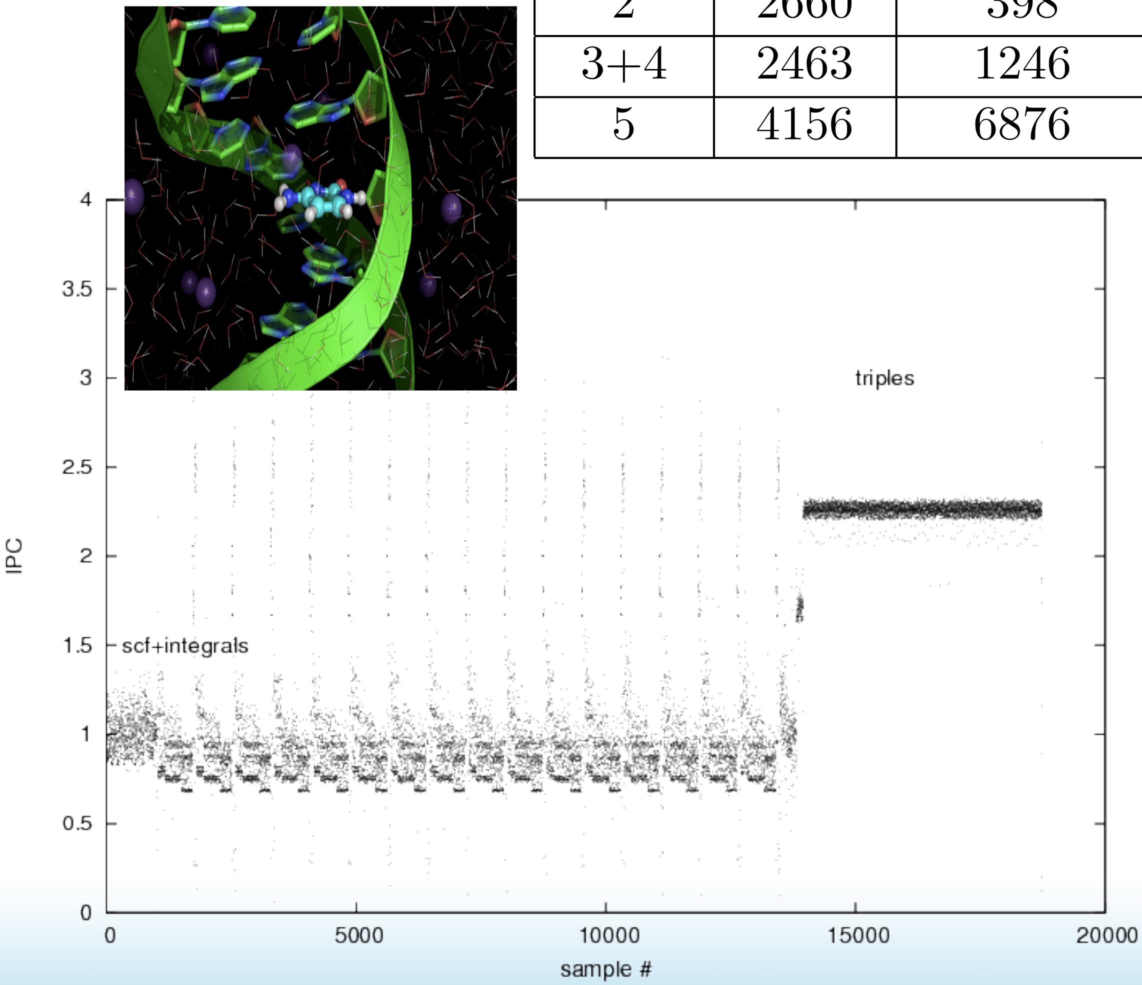
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
tiso	1973	2010	1197	1.9	1.8	0.8	2.3
forces	1602	1736	4577	1.5	0.4	0.7	2.3
iso	1474	1396	1617	1.6	0.9	0.6	2.3



- Two phases, both do small mat-mat mult
- Internal forces perform well

NWCHEM

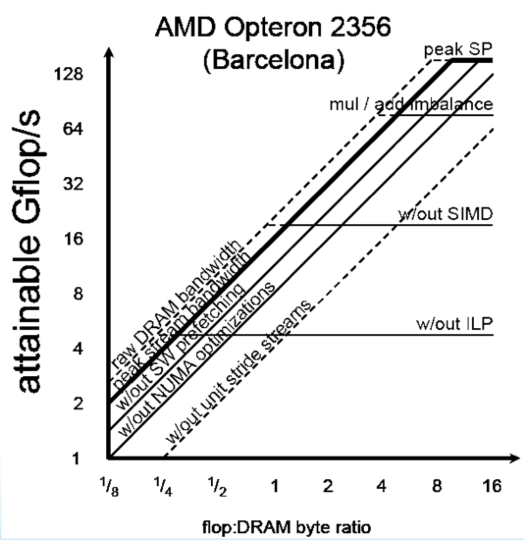
phase	MIPS	MFLOPS/s	MiBPS	CI	AI	IPC	effGHz
1	2616	431	5464	0.3	0.1	1.0	2.6
2	2660	398	4818	0.3	0.1	1.0	2.6
3+4	2463	1246	6030	0.9	0.2	1.0	2.6
5	4156	6876	15583	3.5	0.4	1.6	2.6



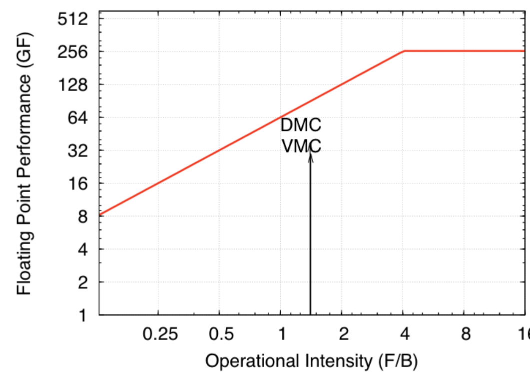
- Highly optimized
  - Even running in turbo boost!
- Very good locality
- Steps 3+4 decent
- Step 5 close to peak!

Roofline models

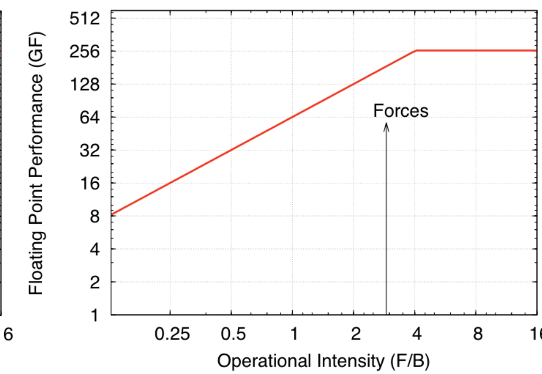
- Current performance and opportunity for improvement.
- Ceilings of the roofline model suggest which optimizations to take.
- Flat roofline is compute-bound, otherwise memory bandwidth limited.



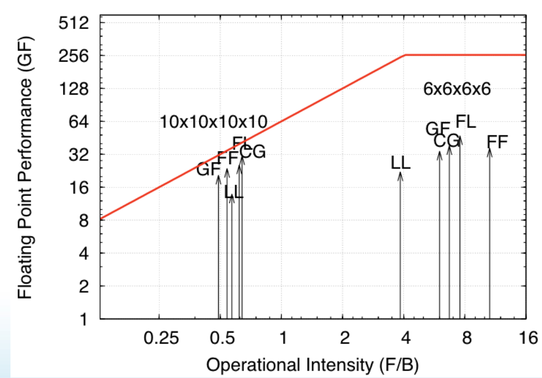
QMCPACK



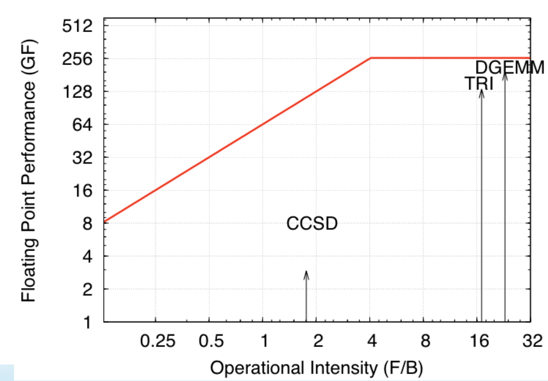
SPECFEM3DGLOBE



MILC



NWCHEM



Optimizations

GNU compiler

- SSE4 compiler intrinsics

Cray compiler

- Reorder code to improve compiler vectorized code
- Compiler directives to control aggressive loop optimization

Programming Model

- CAF for Alltoall, UPC for MILC 4D halo exchange (NERSC)

Runtime Options

- Rank reordering
  - grid_order: Chroma, SPECFEM3DGLOBE, WRF
  - custom: MILC, PPM, VPIC
  - random: NWCHEM
- Hugepages
  - 2MB pages: MILC, PPM, VPIC
  - 8MB pages: NAMD(Charm++), NWCHEM(GA)
- aprun options
  - Core specialization: aprun -r
  - NUMA node memory containment : aprun –ss
- MPI runtime
  - MPICH_COLL_OPT_OFF
  - MPICH_ALLREDUCE_NO_SMP (large messages)
  - MPICH_GNI_MAX_EAGER_MSG_SIZE

Applications at the PF as of 02/25/2013

Full system runs

- VPIC

  3072x3072x2464 cell domain with 7.44103E+12 particles run on 22,528 nodes with 180,224 MPI ranks with 4 OMP threads/rank, and achieved **1.25 PF/s** sustained over 2.5 hrs.

- PPM

  7040^3 zone mesh run on 21,417 XE nodes with 85,668 MPI ranks with 8 threads/rank, and achieved **1.23 PF/s** sustained over 1 hour. 121 nodes were used for I/O and 14 TB of data was written. Recently sustained **1.5 PF/s** with newer code and I/O strategy.

- QMCPACK

  432-atom high-pressure Hydrogen run on 22,500 XE nodes with 4 MPI ranks per node with 8 OpenMP threads per rank and achieved **1.037 PF/s** for 1 hour.

- SPECFEM3DGLOBE

  2720x2720x6 surface element run on 21,675 XE nodes with 693,600 MPI ranks and achieved just over **1 PF/s** sustained.

Honorable Mention

- NWCHEM

  A **0.6 PF/s** on XXX nodes with YYY tasks per node.

- WRF

  Hurricane Sandy grid of 9120x9216x48 with 4 billion points run on 11,400 XE nodes with 16 MPI tasks per node and 2 OpenMP thread/rank, and achieved **0.250 PF/s.**

GPU Applications

- ACM article in progress
- Chroma, NAMD and QMCPACK use CUDA
- GAMESS used OpenACC
  - CUDA Proxy
- Relative performance of XK to XE
  - Speed up of 1.8 – 2.7 on ~ 700 nodes.
  - See http://developer.download.nvidia.com/GTC/PDF/GTC2012/PresentationPDF/Wen-meiHwu_UIUC_BlueWaters_SC12.pdf