BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Illinois Proposal Considerations - 2016

Greg Bauer

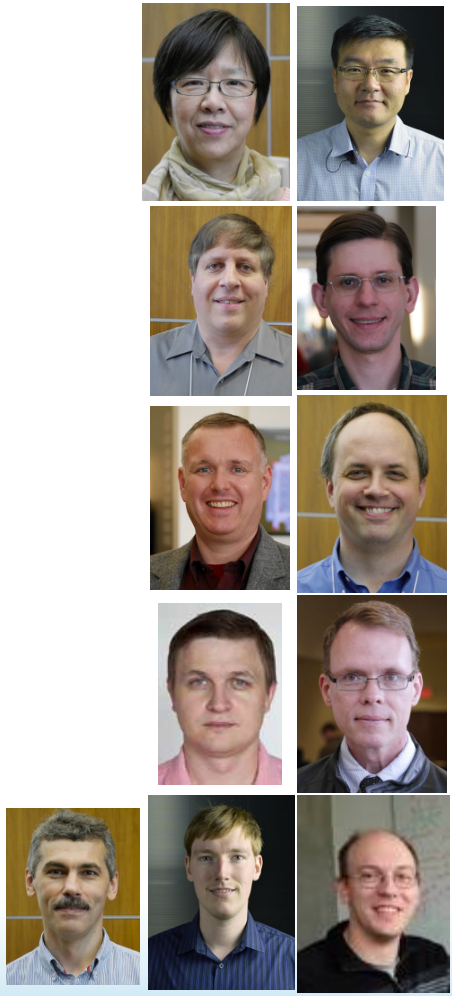
NCSA

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY

Support model

- Blue Waters provides traditional Partner Consulting as part of its User Services. Standard service requests for assistance with porting, debugging, allocation issues, and software requests are handled through the Jira ticket system.
- Advanced Application Support for projects on Blue Waters can be requested via the ticket system. The requests are reviewed and evaluated by the project office for breadth, reach and impact for the project and the community at large.
- Major Science teams (such as NSF PRAC awards) are assigned a Point of Contact (POC) within the Science and Engineering Application Support (SEAS) group. The POC works with the science team on advanced application support areas such as tuning, modeling, IO and optimizing application codes as well as standard service requests.
- In some cases POCs participate in code restructuring, re-engineering or redesign such as implementing GPU functionality via OpenACC or alternatives to MPI collective operations. Such work is tracked via a coordinated work plan that is developed between the POC and the science team to clearly indicate the scope and scale of the work involved including milestones and deliverables. Work plans are reviewed and approved or rejected by the Blue Waters project office.
- Support for workflows, data movement and visualization are provided by the representative groups within Blue Waters.



What makes Blue Waters a Supercomputer?

- Isn't it a really large Linux Cluster?
  - Yes and no. It is running some commodity hardware (processor, memory, GPU) but it has a proprietary interconnect, a tuned Linux OS and a very large and fast parallel file-system. The scale of the system is important.
- My [desktop, local cluster, …] runs my application more quickly than Blue Waters. Why?
  - Consider the following slides

Not All Compute Nodes are Equal

Node	Processor type	Nominal Clock Freq. (GHz)	FPU cores	Peak GF/s per node	Peak Memory GB/s
BlueWaters Cray XE	AMD 6276 Interlagos	2.45	16*	313	102
NICS Kraken Cray XT	AMD Istanbul	2.6	12	125	25.6
NERSC Hopper XE	AMD 6172 MagnyCours	2.1	24	202	85.3
ANL IBM BG/P	POWERPC 450	0.85	4	13.6	13.6
ANL IBM BG/Q	IBM A2	1.6	16*	205	42.6
NCAR Yellowstone	Intel E5-2670 Sandy Bridge	2.6	16*	333	102
NICS Darter Cray XC	Intel E5-2600 Sandy Bridge	2.6	16*	333	102

https://bluewaters.ncsa.illinois.edu/node_core_comparison

Local Resource Comparison

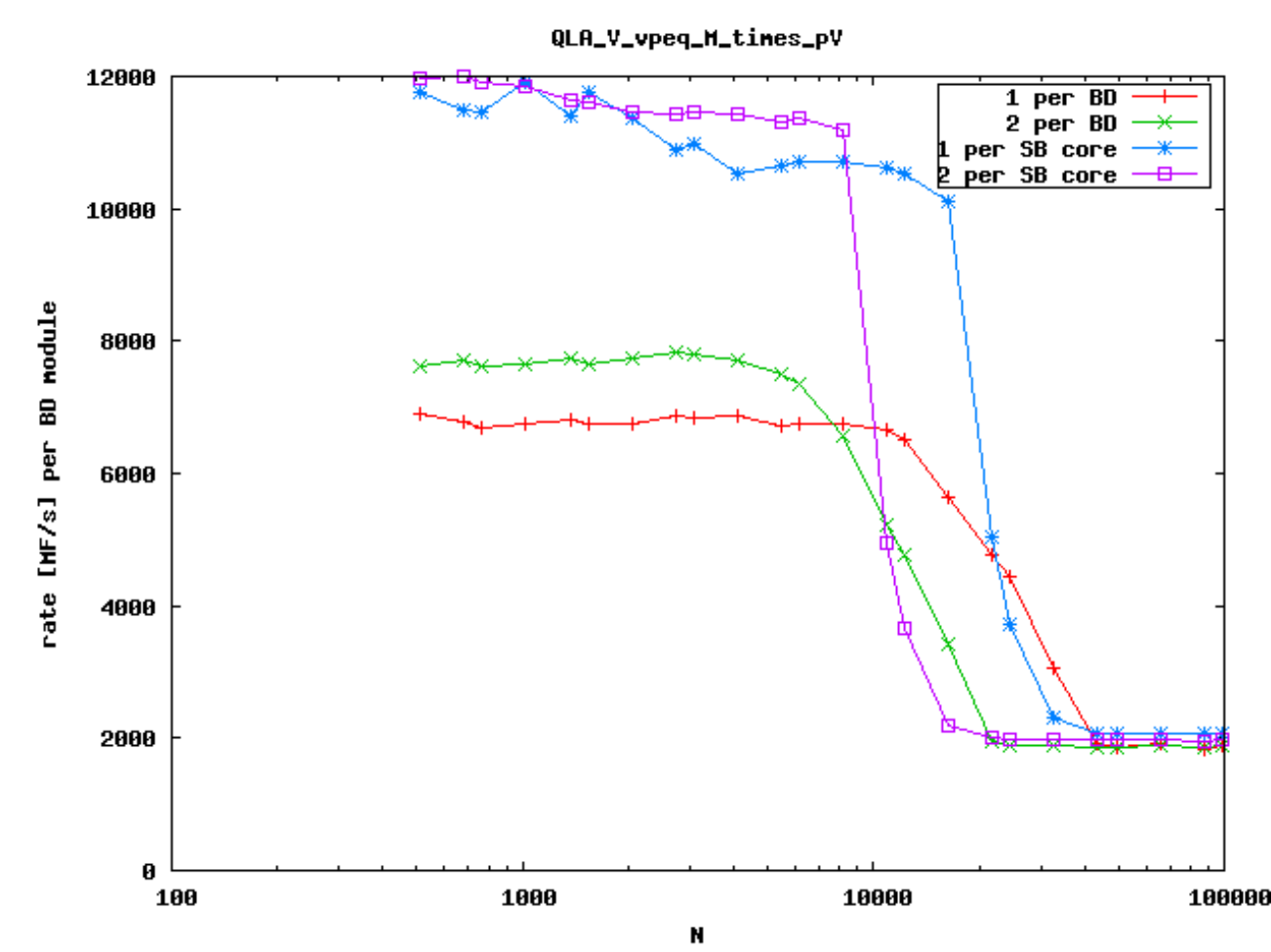
Property	Campus Cluster (Golub)	Blue Waters
# of compute nodes	512	22,640 XE + 4,224 XK
Processors	(2) Intel E5-2670 or (2) E5-2670V2, 2.5 – 2.6 GHz, 20 – 25 MB L3 Cache	(2 or 1) AMD 6276 2.45GHz, 16MB L3 Cache
Memory per Node	32 GB – 256 GB	64 GB or 32 GB
GPUs	NVIDIA M2090, K40	NVIDIA K20x
Interconnect NIC	Mellanox FDR IB 7 GB/s	Cray Gemini 9.6 GB/s
File system	GPFS	Lustre ~ 1 TB/s write
Swap	Local disk	No local disk

Campus Cluster Single Node Performance

- The compute nodes on Campus Cluster are Intel based:
  https://campuscluster.illinois.edu/hardware/
- (2) Intel E5-2670 (Sandy Bridge) 2.60GHz, 20M Cache, 8C
  - This node has peaks of 333 GF/node and 102 GB/s per node. **Performance can be better than XE node**.
- (2) Intel E5-2670V2 (Ivy Bridge) 2.50GHz, 25M Cache, 10C
  - This node has a peak of 400 GF/node (2 CPUs x 2.5 GHz x 10 cores/CPU x 8 Flops/CPU/s) and a peak of 120 GB/s per node. **Performance can be better than XE node**.

One caveat

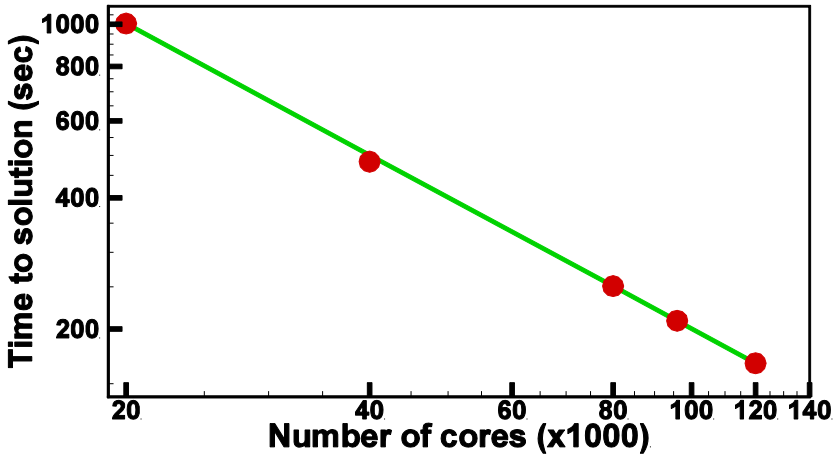
- Intel CPU (SB) L1/L2 caches are faster than AMD CPU (BD) L1/L2 caches.
- Small N fits in L1 and L2 caches. Note the difference in performance rate between SB and BD.
- As N increases, data is in L3 or RAM and the rates become equal.



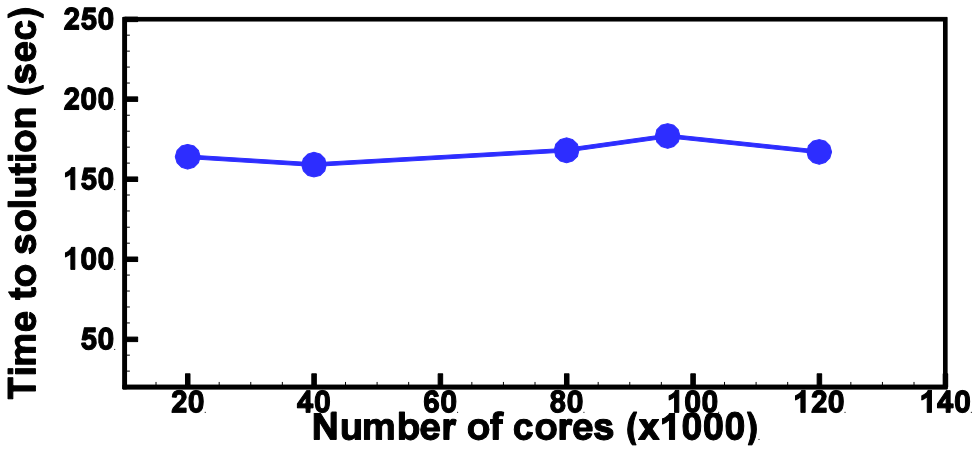
Things to include in a proposal

- Scaling data is useful if you have it.

Strong scaling



Weak scaling



Things to consider in a proposal

- IO Storage requirements
  - Typical file sizes
  - Number of files
  - Residency
- File format(s)
- Application Checkpoint strategy

Runtime and Programming Environment

- No local disk, no swap.
- No ssh access to compute nodes (see CCM).
- Kernel and stack SLES 11 SP3 based.
- CUDA 7.5 (maybe 8 at some point)
- ~~No VM or OS ISO boot support~~. (Shifter)
- Support for Linux cluster compatibility (CCM).
- Current compilers (Cray, PGI, GNU, Intel), Math libraries, MPI-3, Python, R.
- No MATLAB.