BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Why RAIT for BW at NCSA?

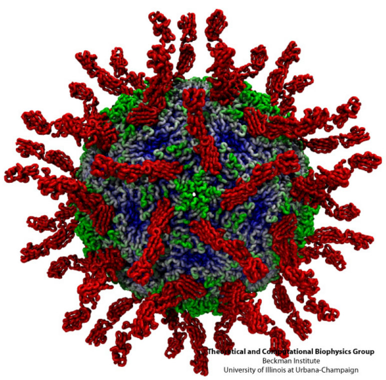
Michelle Butler
mlbutler@illinois.edu

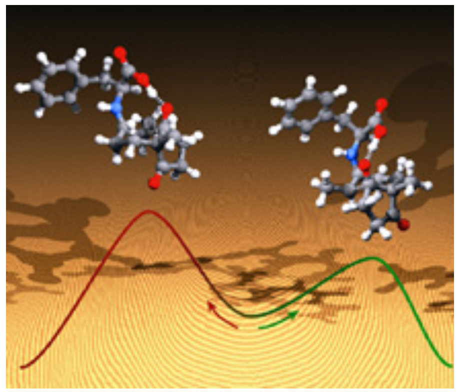
GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

Science & Engineering on Blue Waters

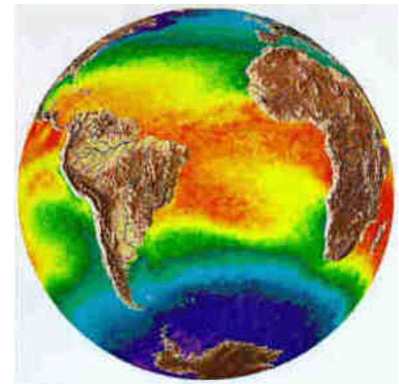
Blue Waters will enable advances in a broad range of science and engineering disciplines. Examples include:

Molecular Science



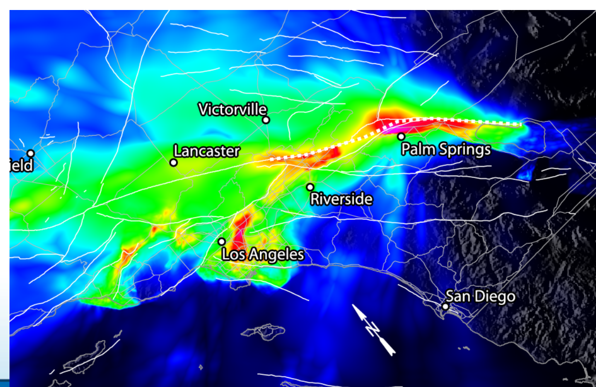


Weather & Climate Forecasting

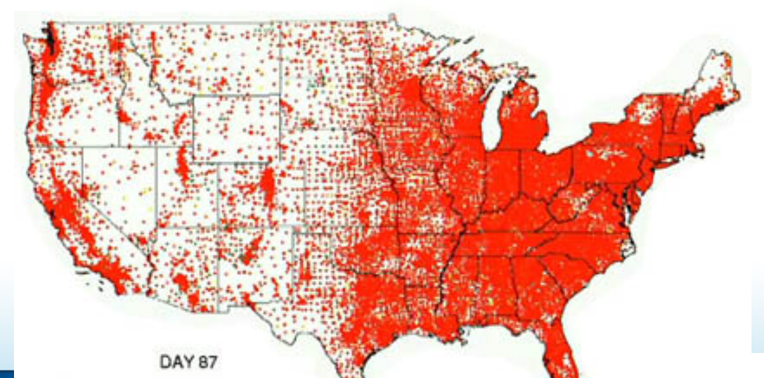


Astronomy

Earth Science



Health



Diverse Large Scale Science

Science areas	Current Number of Teams	General Purpose Balanced System	High Speed CPU	High Performance Memory	High Interconnect Bandwidth	Large Memory Capacity	Low Interconnect Latency/ Acceleration	High Storage and Network Bandwidth
Nano/ Material Science	2		X	X		X		
Chemistry	3	X	X	X	X	X	X	X
Biophysics	2	X	X	X	X		X	
GeoScience	3	X		X		X	X	X
Climate/Weather	3	X		X		X	X	X
Turbulence	1	X		X	X			X
Astrophysics/ Cosmology/ Astronomy	6	X		X		X	X	X
Life Science	2	X		X	X	X	X	
Nuclear/QCD	1	X	X	X		X	X	
Plasma	1	X				X	X	X
System Balance Tests	Total 24	ALL	PS-NAMD PS-MILC WRF PARATEC HPL	PS-DNS3D PS-NAMD NSF-MILC WRF PARATEC STREAM	PS-DNS3D PS-NAMD PARATEC	HPL	PS-MILC PS-NAMD	IOR PS-DNS3D

National Petascale Computing Facility

Partners

- EYP MCF/ Gensler
- IBM
- Yahoo!

- **Modern Data Center**
  - 90,000+ ft^2 total
  - 30,000 ft^2 raised floor
    20,000 ft^2 machine room gallery
- **Energy Efficiency**
  - LEED certified Gold
  - Power Utilization Efficiency = 1.1–1.2

Building Blue Waters

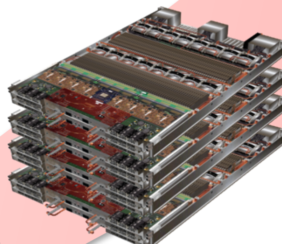
Blue Waters will be the most powerful computer in the world for scientific research when it comes on line in 2011-2.

Blue Waters

≳10 PF Peak
~1 PF sustained
≳300,000 cores
≳1 PB of memory
>25 PB of disk storage
500 PB of archival storage
≥100 Gbps connectivity

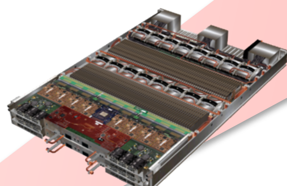
Blue Waters 3-Rack Building Block

32 IH server nodes
  256 TF (peak)
  32 TB memory
  128 TB/s memory bw
4 Storage systems (>500 TB)
10 Tape drive connections



IH Supernode

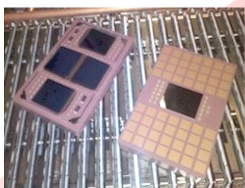
4 IH Server Nodes
  1024 cores
  Up to 32 TF (*peak*)
41 TB memory
  16 TB/s bw
32 Hub chips
  36 TB/s bw



IH Server Node

8 QCM's (256 cores)
  Up to 8 TF (*peak*)
1 TB memory
  4 TB/s bw
8 Hub chips
  9 TB/s bw
Power supplies
PCIe slots

Fully water cooled

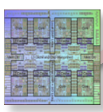


Quad-chip Module

4 Power7 chips
  Up to 1 TF (*peak*)
128 GB memory
  512 GB/s bw

Hub Chip

1.128 TB/s bw



Power7 Chip

8 cores, 32 threads
L1, L2, L3 cache (32 MB)
Up to 256 GF (peak)
128 Gb/s memory bw

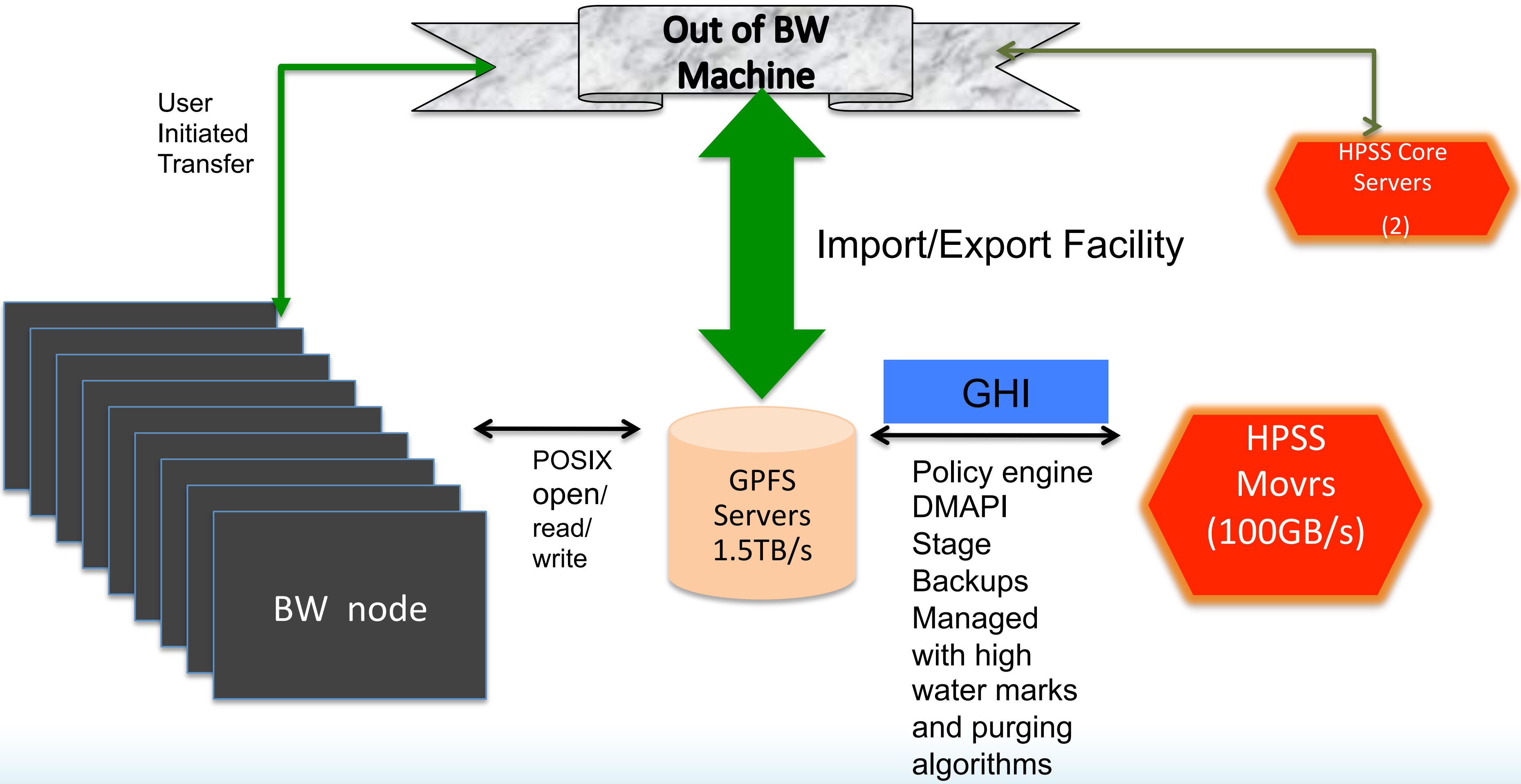
45 nm technology

Blue Waters is built from components that can also be used to build systems with a wide range of capabilities—from desk side to beyond Blue Waters.

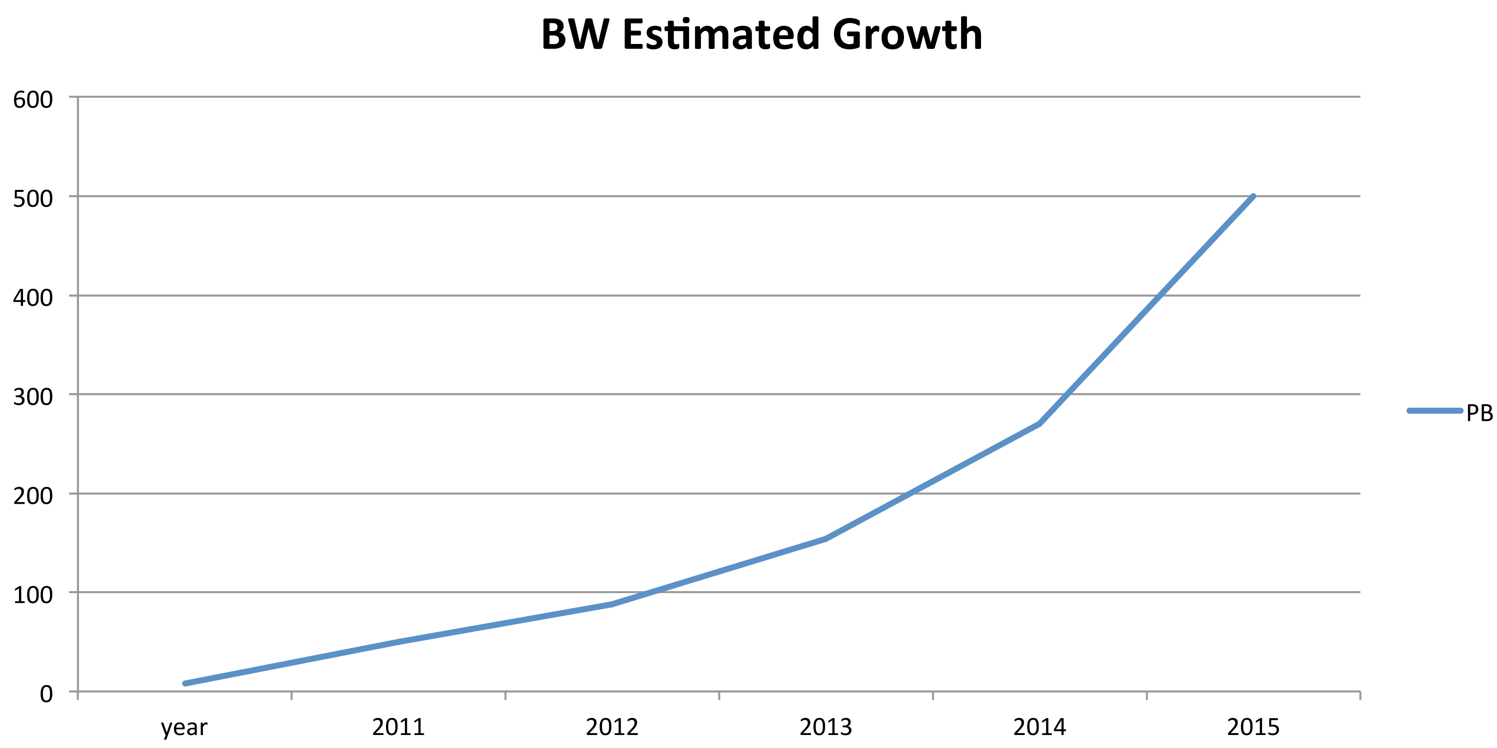
Blue Waters Computing System

System Attribute	JAGUAR		Blue Waters
Vendor	CRAY XT5		IBM
Processor	AMD OPTERON		IBM Power7
Peak Performance (PF)	2.3	4x	>10
Sustained Performance (PF)	≤1		≥1
Number of Cores/Chip	6	1.3X	8
Number of Processor Cores	224,256	1.2X	>300,000
Amount of Memory (TB)	299	4X	1,200
Memory Bandwidth (PB/s)	.478	10X	>5
Interconnect Bisection BW (TB/s)	~2	2X	~1
Interconnect HW Latency (μs)		>>	
Amount of Disk Storage (PB)	5	3X	18
I/O Aggregate BW (TB/s)	.24	6X	>1.5
Amount of Archival Storage (PB)	20	25X	>500
External Bandwidth (Gbps)			100-400

I/O Software environment



5 Years of Growth



Why RAIT at NCSA

- Tape is still viable solution:
  - At the scale that we are at (500PB in 5 years)
    - To date no reliable, scalable disk solution that has a cheaper TOC
      - TOC is total cost of ownership
        - Power & cooling required
        - Floor Space required
        - Length of time for use/viable solution
        - Rolling forward for use of 10 years?
- Tape is still the name of the game
  - 2TB archive site – 20 years ago is a different story

Why RAIT at NCSA

- Primary reason is for Data Protection.
  - At BW scale, we could NOT afford to duplicate copy this data which is current practice at NCSA today.
  - A redundant array of tapes with 8 data and 2 parity can survive the loss of 2 tapes at a cost of only 25% more tapes than unmirrored, single tape
- In last 25 years NCSA has lost 2 user files.
  - We have seen on numerous occasions the need for the second tape.
    - Firmware being one of the worst occasions LTO1/LTO2
    - Library drop tape, tape drive eat tape are more rare, but still occasionally happen

Why RAIT at NCSA

- HPSS already had striped data on tape.
  - NCSA and HPSS collaboration are adding RAIT engine to the overall HPSS environment.
    - Will generate the data that is required to be written in parity.
    - Depending on environment sites will need mulitple RAIT Engines and should not be the bottleneck to the tape device. Parity takes COMPUTE CYCLES!
    - Up to 16 wide devices and 8+8 is the highest level of parity (8 levels).
      - D+P <= 16; D >=P; D>=2, (P<=8)