Accelerators
Light at the end of a tunnel
or
a Train

John M. Levesque
CTO Office
Director – Cray's Supercomputing CoE
Cray Inc

Outline

- Why do we have accelerators
  - Power Savings
    - More flops/watt
  - Cost Savings
    - More flops/$$
- Problems with HPL and HPCC
- Where are WE going?

What About Power Savings

- More things need to be considered than the power required to run the system
  - What about the energy required by the code developers

Power Savings

- The power used by Roadrunner 2345.5 Kw
- However
  - The rumored time to code HPL was a year by 20 people. I think this must be high; however, Andy White gave a talk where he stated that a typical application might need 1 man year, so how much energy was required to develop HPL for Roadrunner?

From “The Matrix” we can determine the amount of energy required

One Man-Year of Energy

From “The Matrix”

From the script:

[Morpheus] "The human generates more bio-electricity than 120-volt battery and over 25,000 BTUs of body heat. Combined with a form of fusion, the machines have found all the energy they would ever need. There are fields...endless fields, were human beings are no longer born. We are grown. For longest time, I wouldn't belive it...and then I saw the fields with my own eyes. Watch them liquefy the dead, so they could be fed intravenously to the living. And standing there, facing the pure horrifying precision, I came to realize the obviousness of the truth. What is The Matrix? Control. The Matrix is a computer generated dream world, built to keep us under control in order to change a human being into this.

[Morpheus holds up a battery to Neo]

Power Savings

- The power used by Roadrunner is 2345.5 Kw
- However
  - The rumored time to code HPL was a year by 20 people. I think this must be high; however, Andy White gave a talk where he stated that a typical application might need 1 man year, so
    - The human body generates 25,000 BTU of heat or 2.931×10^{-4} kWh = $2.931 \times 10^{-4} * 2.5 \times 10^{4} * 365 * 8$ = 21396 Kw – almost 10 time the system power

Clearly the Energy required to program the Cell is far more than the Energy to run the system

What About Power Savings

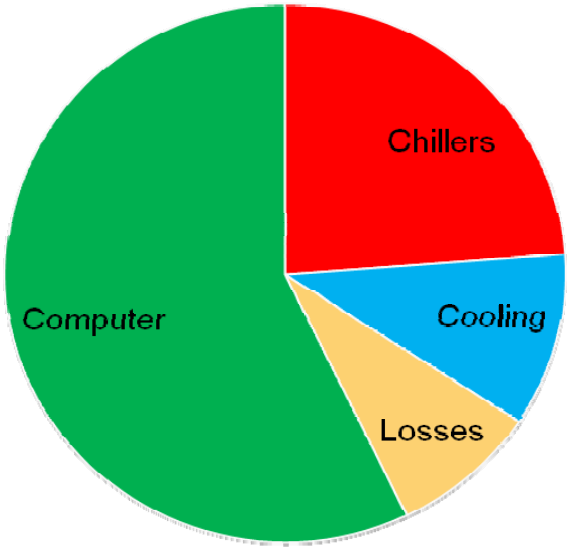
- More things need to be considered than the power required to run the system
  - What about the energy required by the code developers
  - What about the energy to run the rest of the computer room

Facilities and Efficiency

- Power in a DataCenter
  - 30-35% to Cooling
    - Chillers (This is the BIG HITTER)
    - Computer Room Air Conditioner (CRAC) Units
      - Typically only 60-80% efficient in extracting heat
    - Motors, Fans, Pumps
  - 10-15% Electrical Losses
    - AC→DC conversion
    - Inefficiency of systems
  - 50-60% to the Computer
- PUE (Power Usage Effectiveness) Typically PUE=1.8 for most datacenters today.
  - PUE = Total Facility Power / IT Power

the green grid™
get connected to efficient IT

Datacenter Power Usage



Cray's Latest Product Line

Cray's New Computer Room Chiller

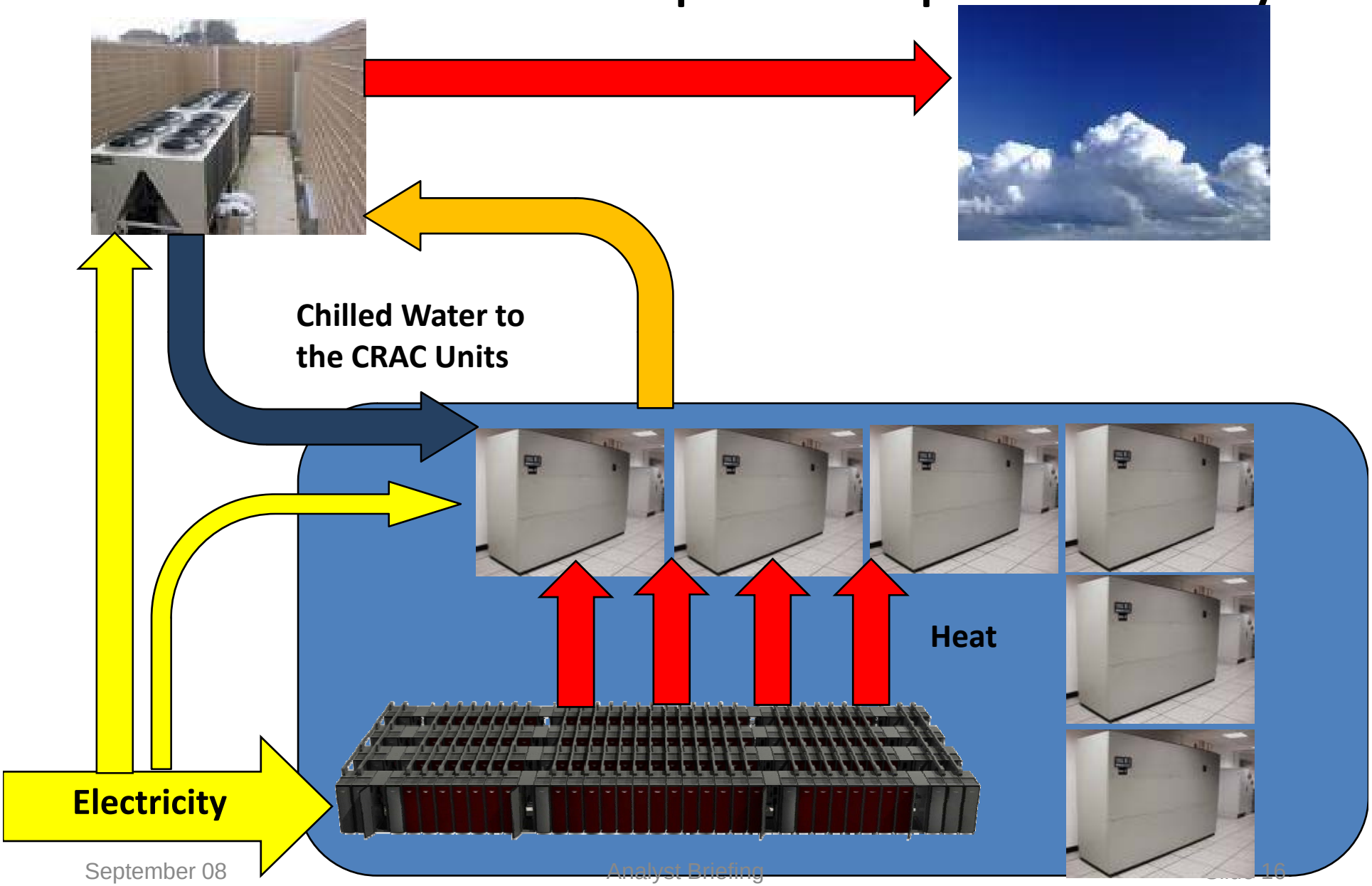
The Cray High Efficiency Cabinet with ECOphlex Technology

(PHase change Liquid EXchange(PHLEX))

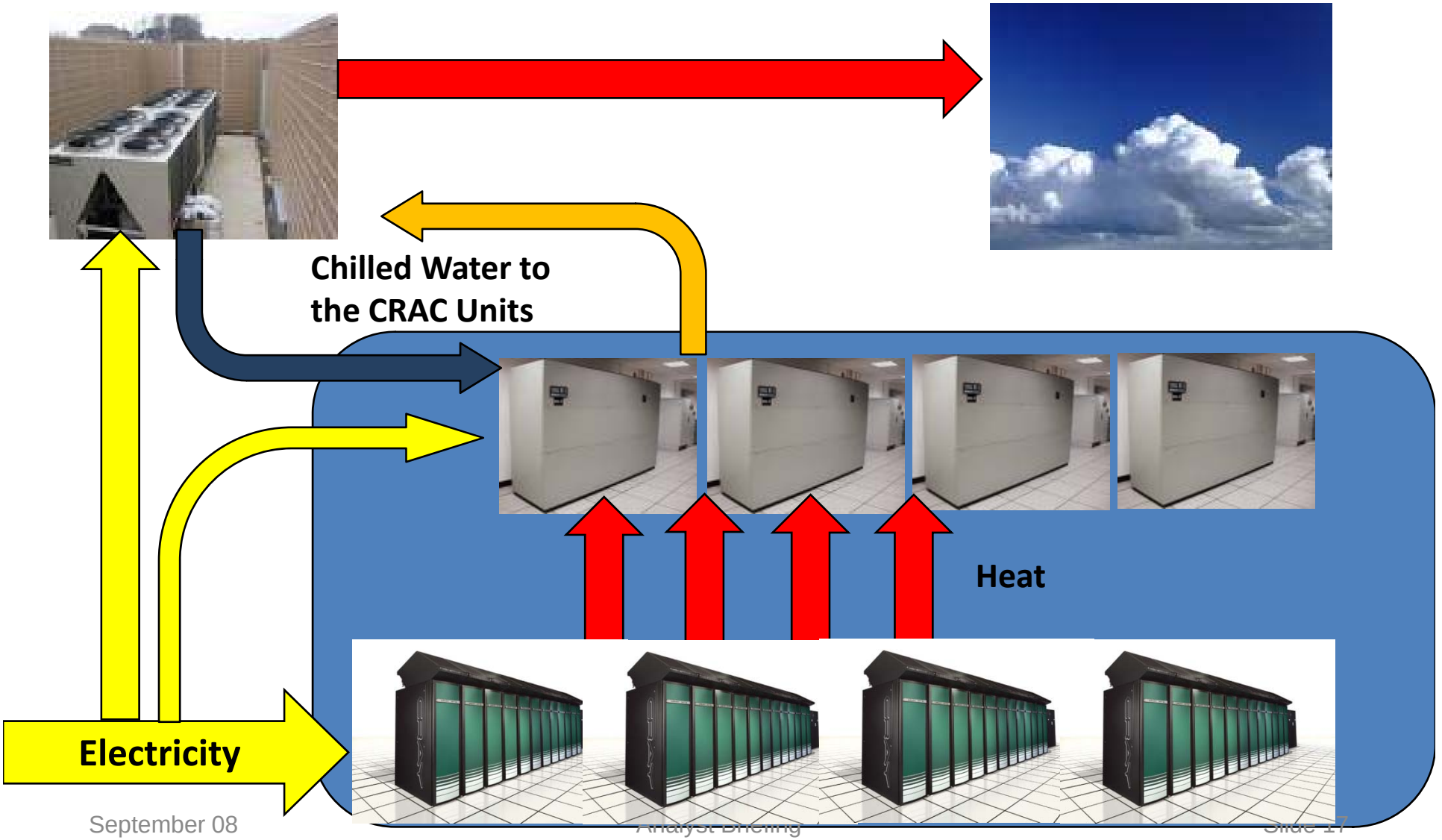
12 HE Cabinet System With XDPs (Front View)

12 HE Cabinet System With XDPs (Rear View)

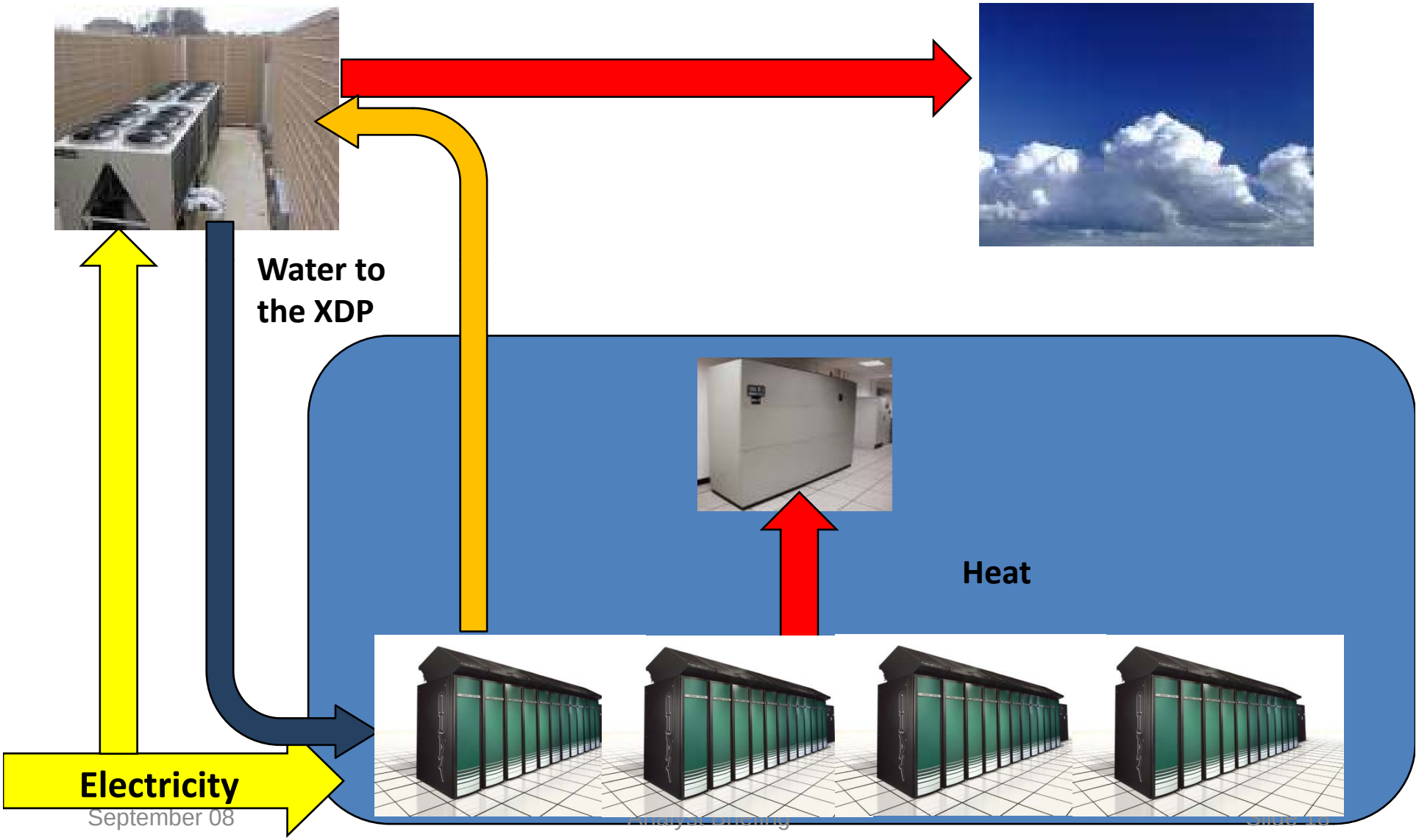
An Air-Cooled SuperComputer Today



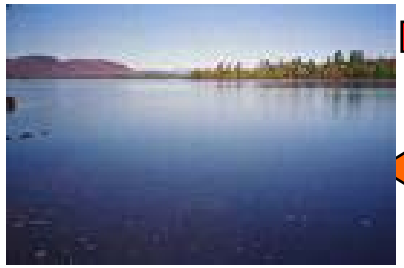
ECOphlex- A Flexible, Efficient Air-Cooled Supercomputer



ECOphlex- A Flexible, Efficient Liquid-Cooled Supercomputer



A “Green” Supercomputer

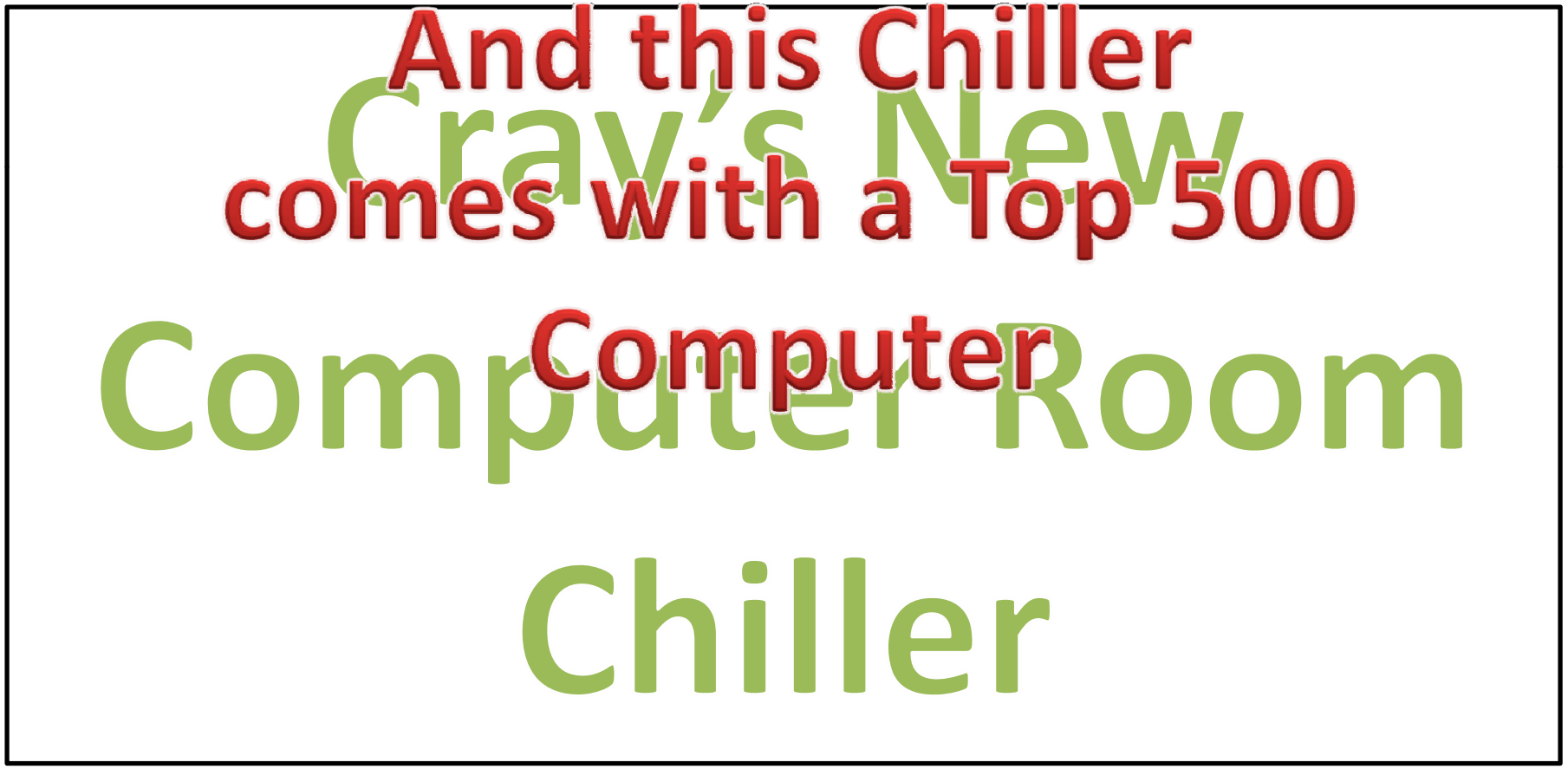


Warm
Water
Cooling

Heat

Electricity

Cray's Latest Product Line



So what about low power processors?

- Some vendors have taken the path of reducing the power and performance of the processor to save on overall power

Cray on the other hand has taken the approach that application developers need the performance to get their science done in a timeframe that is conducive to their research.

So what about low power processors?

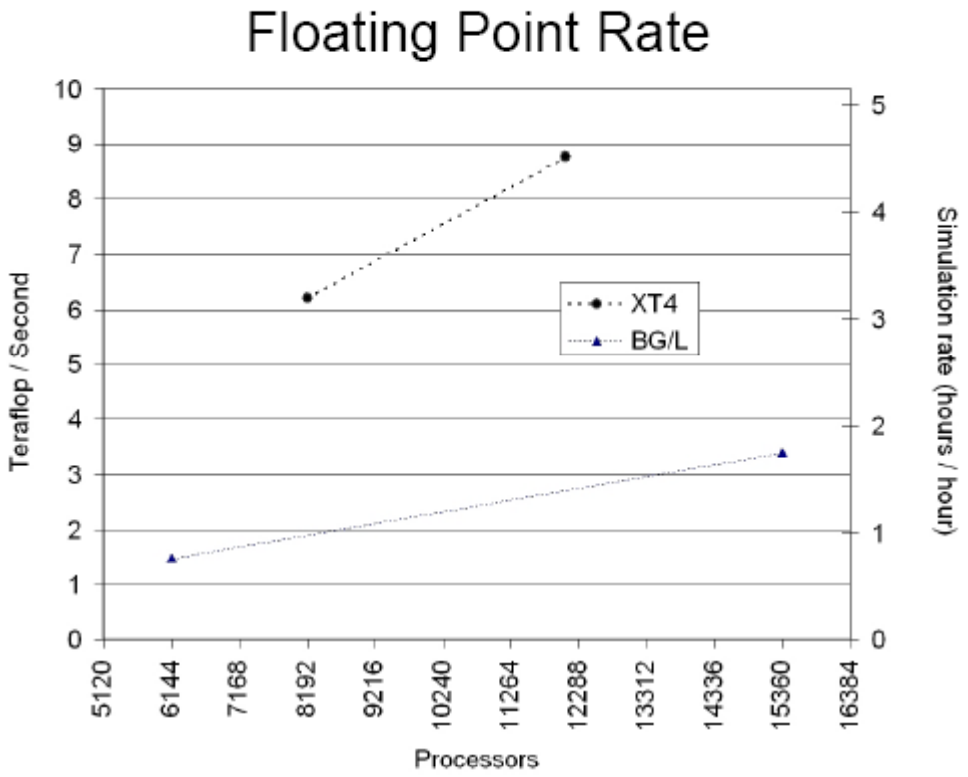
- Some vendors have also taken the path of reducing the memory of the node to reduce the cost of the system and the overall power

Cray on the other hand has taken the approach that application developers need at least 2GB of memory to get their science done

And Since the HPL Number looks at Sustained Performance/Kw

WRF Nature Run by Allan Snavely

XT has four times the memory which uses a significant chunk of the power



XT's result rate is 3 * 14000/12000 = 3.5 faster

And this is before we add PHLEX liquid cooling

BlueGene's Power consumption is 1/5 that of XT

And What about Price

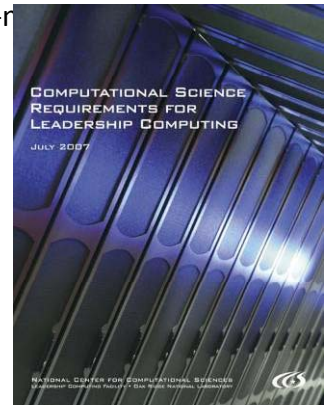
- Reduction of facility cost with PHLEX
- Reduction of programming cost with superior compilation
  - The Granite processor in Cascade will be used by Fortran, C and C++ using the Cray Compilation System
- And of course the ability to produce Break-through Science is

"priceless"

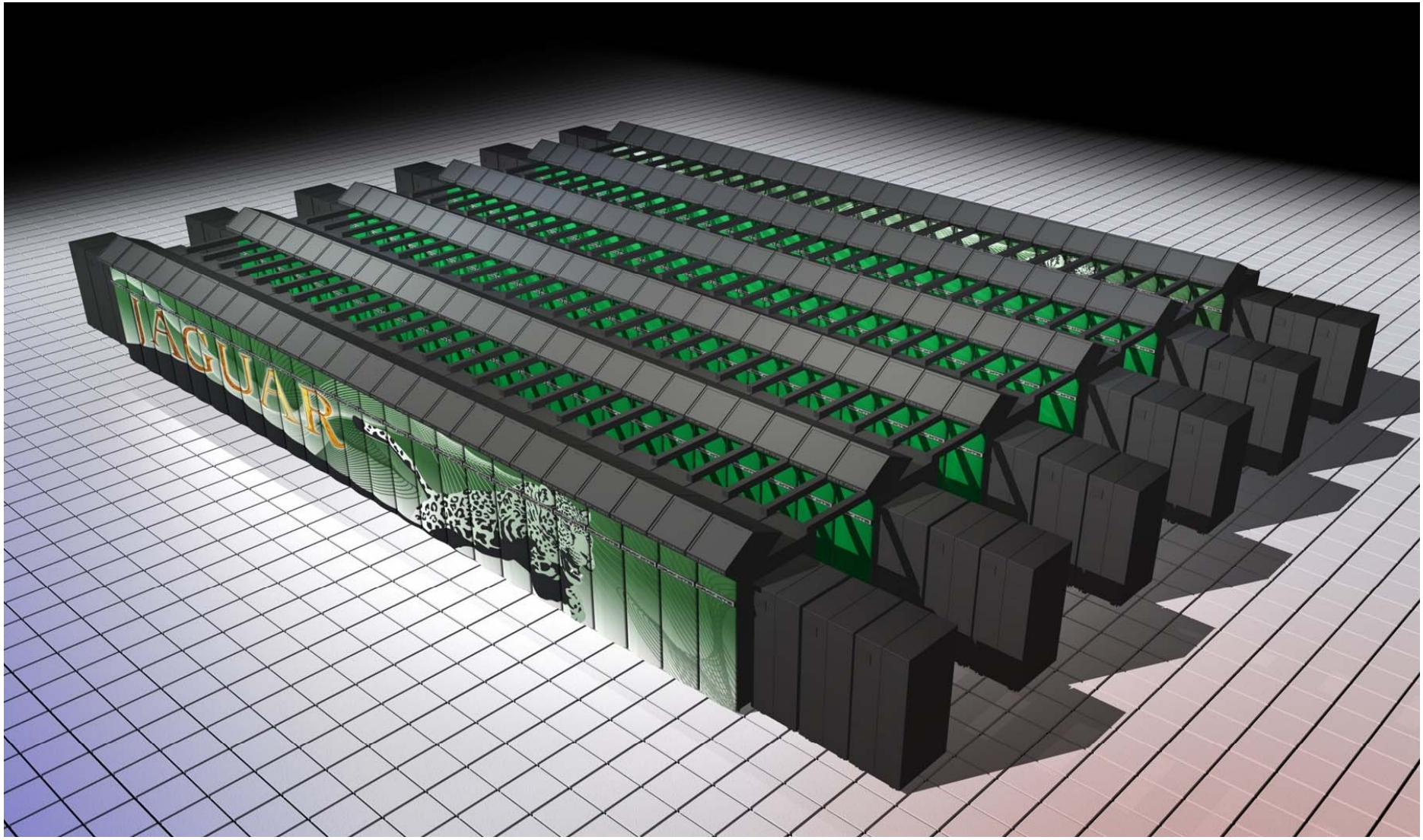
Pioneering Applications

250 TF Selection Process

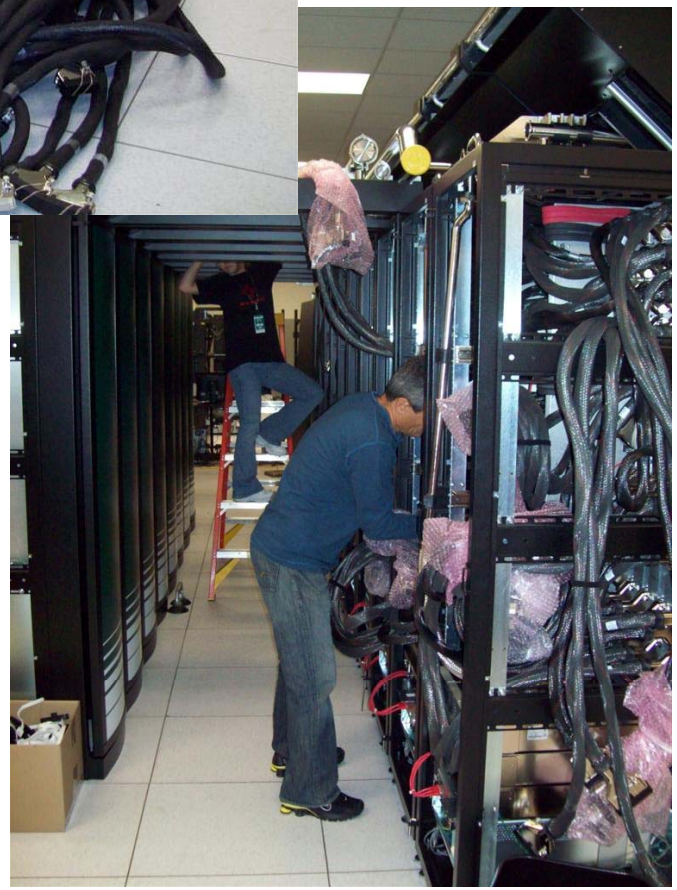
- ORNL LCF collected data in an open call from science application teams
  - Physics models
    - What physical models are in your code and what changes are planned in the near future?
  - Algorithms
    - What algorithms are in your code and what changes are planned in the near future?
  - Scaling
    - How does your code currently scale and what bottlenecks preclude improved performance?
  - If chosen for acceptance
    - How might your code be used to test and accept a leadership system?
  - If chosen for science on day one
    - What science would you explore and what simulations would you do with a 250 TF-r
  - Functional software requirements
    - What system software and math libraries are required by your code?
- Over 20 application teams delivered written responses
  - Broad email requests sent out to user groups
  - Predominant response from INCITE, SciDAC, and NSF Projects
  - Documented in Appendix E of NCCS 2007 Requirements Document
    - *Computational Science Requirements for Leadership Computing*
  - Data delivered in fall 2006 to DOE/ASCR for decision
- Pioneering applications for the 1 PF T2O period
  - Web-based form available online by 12/31/07; accept applications through Spring 07
  - Each application could potentially access 50M hours during the 1 PF T2O period!



ORNL Petaflop System



7M
Anacondas



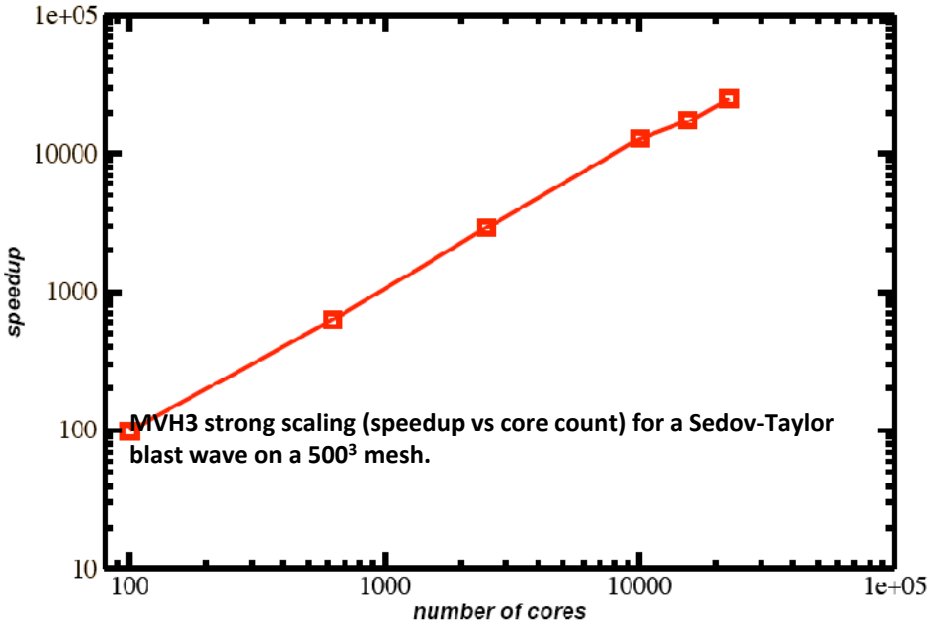
Pioneering Application: CHIMERA

Code Readiness, Scalability, and Performance

Readiness Activities

- Physical Models
  - Alpha network
- Algorithms
  - Spherical polar coordinate singularity workaround
  - Poisson solver
- Scalability & performance
  - Multi-core ray-by-ray solves
  - Replace domain decomposition from slab to pencil
  - Parallel I/O
  - Joule metric benchmark studies

LCF liaison contributions
- **Implementing efficient, collective I/O**
- **Pencil decomposition of 3D flow algorithm**
- **Preconditioning of the neutrino transport equation**



Scalability/Performance

- **Good weak and strong scaling**
- **Initial Barcelona quad-core testbed performance promising**
  - **Currently using 1 MPI task/core, with plans to implement OpenMP for threading of transport and nuclear burning solves**

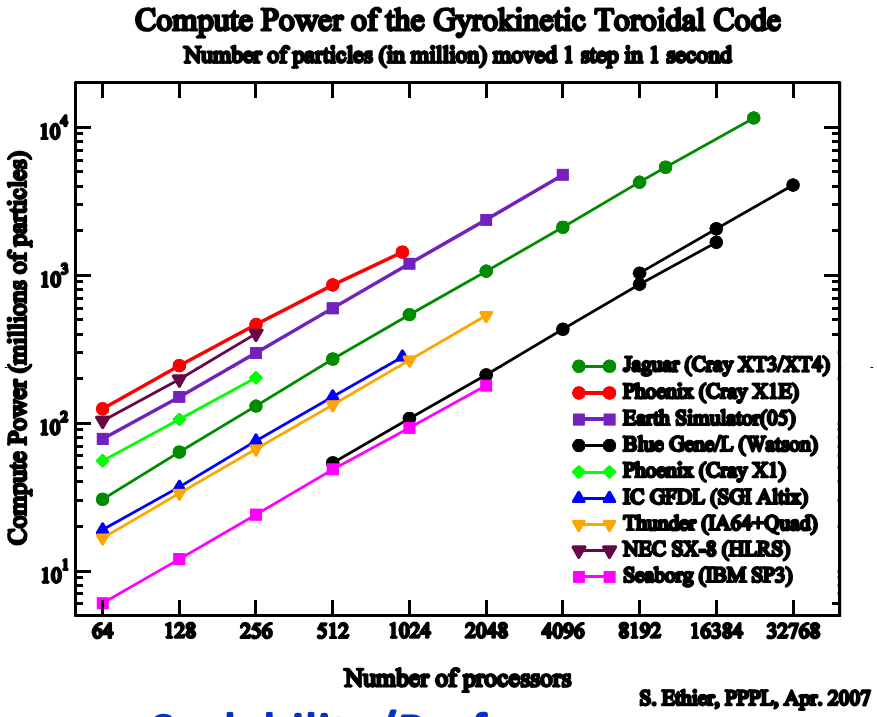
Pioneering Application: GTC

Code Readiness, Scalability, and Performance

Readiness Activities

- Physical Models
  - Implement split-weight scheme for kinetic electrons in shaped plasma component (GTC-S)
- Algorithms
  - Port and optimize GTC-S
- Scalability & performance
  - Implement radial and particle domain decomposition in GTC-S
  - Implement asynchronous I/O
  - Data flow automation
  - Joule metric benchmark studies

LCF liaison contributions
- **Asynchronous I/O**
- **Automated end-to-end workflow**
- **Porting/scaling new shaped plasma version**



Scalability/Performance

- **Excellent full system weak scaling with ~20% of peak performance realized**
  - **Parallelized with MPI and OpenMP**
- **Initial Barcelona quad-core testbed performance promising**
  - **OpenMP threads perform well**
  - **Reduced memory B/W may not be an issue**
- **Needs to vectorize better**

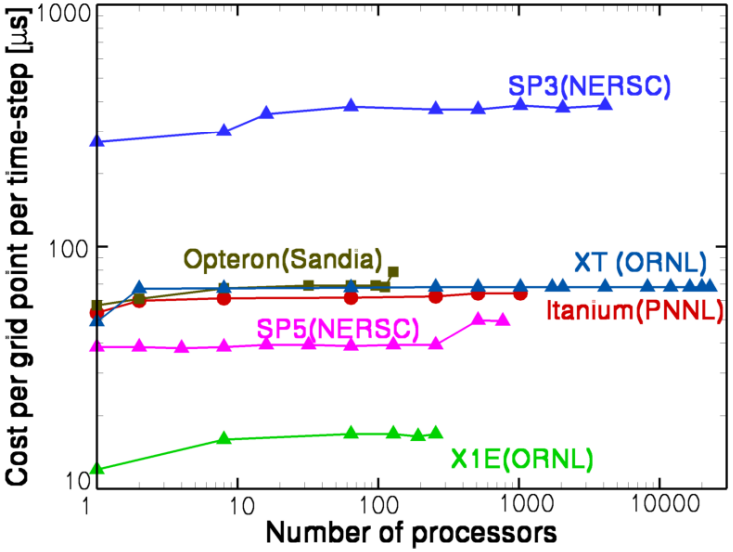
Pioneering Application: S3D

Code Readiness, Scalability, and Performance

Readiness Activities

- Physical Models
  - Develop reduced chemical mechanism for n-heptane and ethylene; developed reduced efficient transport model
- Algorithms
  - Test n-heptane model for stiffness; develop additive RK integration scheme if stiffness limits integration time step
  - Implement massless Lagrangian tracers
- Scalability & performance
  - Tune multi-core performance
  - Develop and test collective I/O
  - Finalize run parameters (e.g. spatial resolution, domain size)
  - Joule metric benchmark studies

LCF liaison contributions
- **Implement Lagrangian tracers**
- **I/O rework with NW University**
- **Scaling studies identified processors burdened by memory corrections**



Scalability/Performance

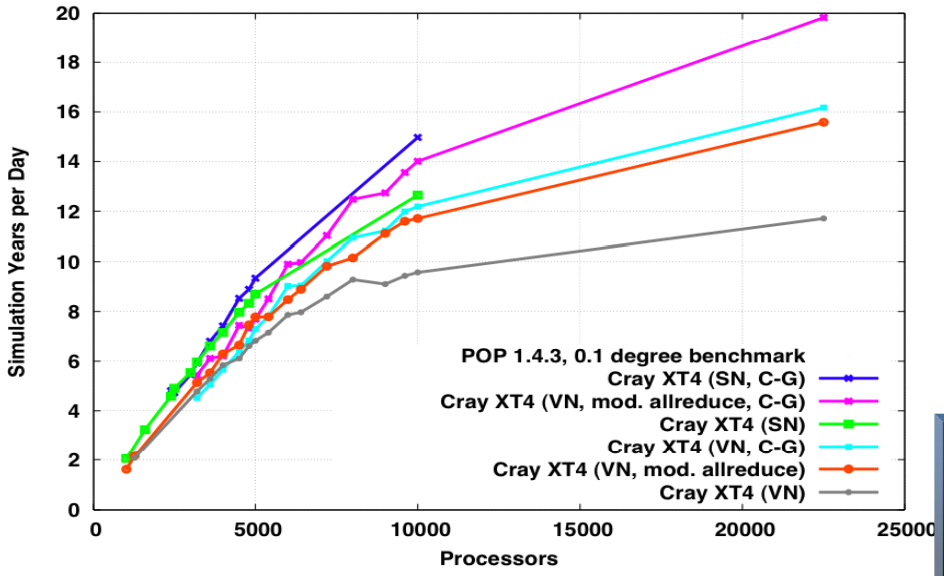
- **Excellent full system weak scaling with ~15% of peak performance**
- **Initial Barcelona quad-core testbed performance promising**
  - **Good vectorization**
  - **Reduced memory B/W may not be an issue**
  - **Addition of OpenMP threads still of interest**
- **Efforts of SciDAC-PERI and Cray COE @ ORNL helpful**

Pioneering Application: POP

Code Readiness, Scalability, and Performance

Readiness Activities

- Algorithms
  - Implement more scalable barotropic solver with improved CG preconditioner
    - Block Jacobi (additive Schwartz), with plans for multi-level enhancement
    - Trade extra flops for more iterations
- Scalability & performance
  - Tune for SSE and OpenMP parallelism
  - Implement parallel I/O and test



Scalability/Performance

- **Ever-improving strong scaling with ~10% of peak performance**
  - **Tackle scalability-limiting barotropic solver dominated by MPI all-reduce latency with new block Jacobi preconditioner**
  - **Should benefit more from QC SSE instructions**
- **New preconditioner in barotropic solve is 1.78x faster on 15,000 cores**
  - **Full benchmark 1.38x faster**
- **Initial Barcelona quad-core testbed perf**
  - **Good vectorization**
  - **Memory B/W an issue unless high processor counts are used to ensure small subgrid size**
  - **Improved speedup needed w/ OpenMP threads**
- **Addition of biogeochemistry creates more independent work, improving scalability**
- **Issue with global gather for I/O on CNL**
  - **Currently being addressed in multiple ways**

LCF liaison contributions

- **New preconditioner for barotropic solver**
- **Contributed bug fixes to POP 2.0**
- **Represent needs at OBER/ESNET meeting**

Pioneering Application: DCA++

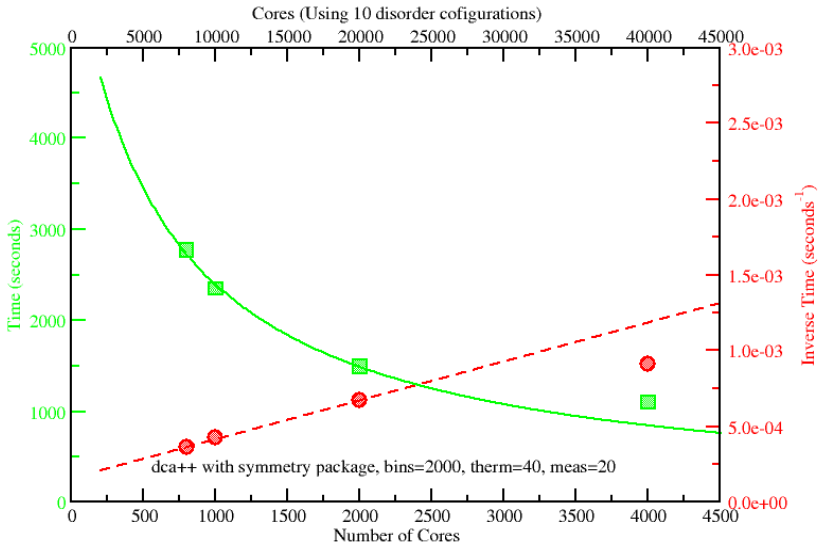
Code Readiness, Scalability, and Performance

Readiness Activities

- Physical Models
  - Develop space group package for 2D/3D symmetry
  - Develop multi-band Hamiltonian concept and DFT
- Algorithms & Software
  - Rewrite current QMC/DCA code
- Scalability & performance
  - Implement additional parallelization over disorder configurations (order 10^2)
  - Additional parallelizable loop over disorder configuration lies between the outer most self-consistency loop of the DCA and the Monte Carlo sampling loop
  - Enables ~10 disorder configurations in parallel on a total of up to 20K cores
    - Assuming individual QMC runs scale to 2000 cores at near optimal speedup

Scalability/Performance

- **Good weak scaling**
- **Single-node performance relies on efficient execution of DGEMM on long thin rectangular matrices**



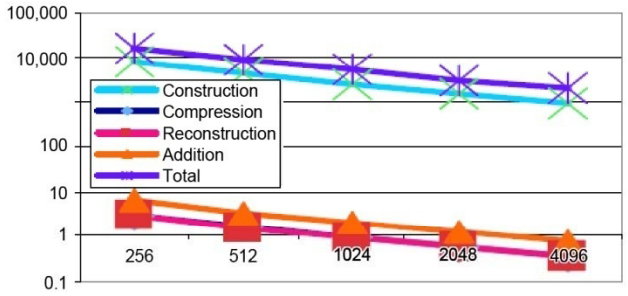
Time to solution and speedup (inverse time) for a prototype DCA++ run of the 2D Hubbard model with 16 sites, 80 time slices, and 40,000 measurements, and two steps of MC updates between measurements

Pioneering Application: MADNESS

Code Readiness, Scalability, and Performance

Readiness Activities

- Dynamic load-balancing
  - Testing data redistribution
  - Commencing development on work stealing
- Multi-core
  - Testing design choices for threading of task queue
- Applications
  - Density functional theory – migrating from prototype to implementation
  - Dynamics – evaluating new time evolution scheme



Scalability/Performance

- **Runtime objective: scalability to 1+M processors ASAP**
- **Runtime responsible for**
  - **scheduling and placement,**
  - **managing data dependencies,**
  - **hiding latency, and**
  - **Medium to coarse grain concurrency**
- **Compatible with existing models**
  - **MPI, Global Arrays**
- **Borrow successful concepts from Cilk, Charm++, Python**
- **Performance examples**
  - **Small matrix BLAS in x86 assembly**
    - **Tuned for target problems**
    - **2-6x faster than existing libraries (ACML, ATLAS, Goto, MKL)**
    - **50-87% of theoretical peak FLOP/s speed**
  - **Parallel scalability**
    - **Tested for correctness and performance on 4096 cores under CNL. Also functions on BG**

Current Planned Pioneering Application Runs

Cursory Look at the Simulation Specs

Code	Quad-Core Nodes	Global Memory Reqm (TB)	Wall-Clock Time Reqm (hours)	Number of Runs	Local Storage Reqms (TB)	Archival Storage Reqms (TB)	Resolution and Fidelity
CHIMERA	7824 4045	16 8	100 100	1 1	13	50	256x128x256 or 256x90x180 20 energy groups, 14 alpha nuclei
GTC-S GTC-C	3900 3900	40 60	36 36	2 2	350	550	600M grid points, 60B particles 400M grid points, 250B particles
S3D	7824	10	140	1	50	100	1B grid points, 15 μm grid spacing 4 ns time step, 23 transport vars
POP	2500	1	400	1	1	2	3600x2400x42 tripole grid (0.1°) 20-yr run; partial bottom cells; first with biogeochemistry at this scale
MADNESS	7824	48	12 2	10 12	5	50	600B coefficients
DCA++	2000 6000	16 48	12 to 24	20	1	1	Lattices of 16 to 32 sites 80 to 120 time slices $O(10^2\text{-}10^3)$ disorder realizations

Astrophysics — Fusion — Combustion — Climate — Chemistry — Materials Science

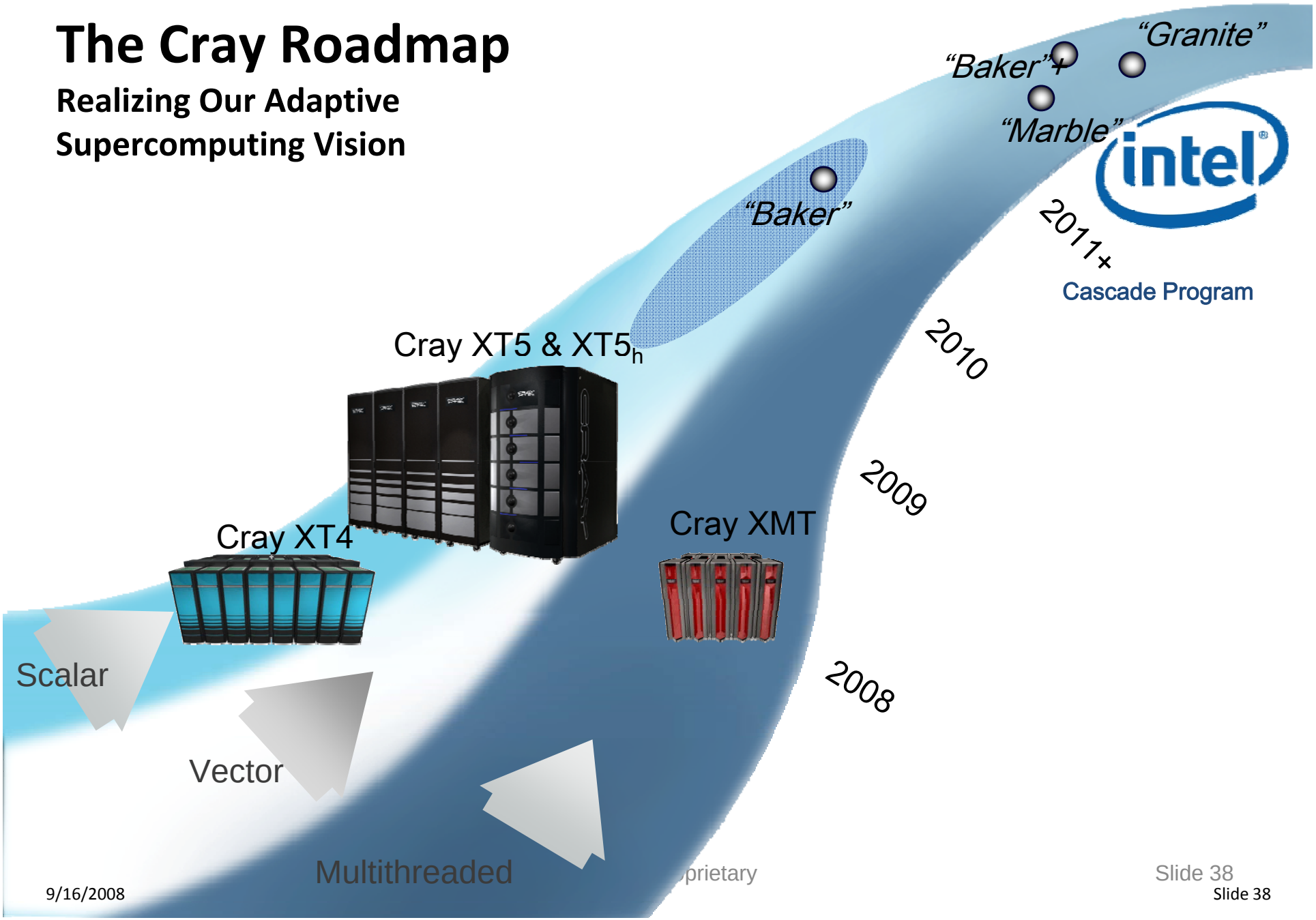
What's wrong with HPL & HPCC

- HPL is only measuring Peak Flops
- HPCC measures everything
  - Only two of vendors in the June 2008 Top 100 has submitted HPCC.
    - IBM and Cray
  - No non-custom interconnect system over 128 nodes have been submitted
- To get high percentage of peak, one must use all the memory available, which means longer running time, which quickly becomes a MTTF test
- Once accelerators are 64 bit capable, they will dominate the Top 500, because they can use fewer faster processors
- Maybe we should require all Top 500 entries should come from HPCC

	Manufacturer	Interconnect	Processor Count			Cores per Chip	HPL Processes	Threads		MPI Processes
			Total	from Form	Derived			from Form	from Code	
2006-04-06	IBM	Custom	65536	65536	65536	1	65536	1	1	65536
2005-11-02	IBM	Custom Torus / Tree	131072	131072	65536	1	65536	1	1	65536
2005-11-02	IBM	Custom Torus / Tree	131072	131072	65536	1	65536	1	1	65536
2005-11-02	IBM	Custom Torus/Tree	65536	65536	65536	1	65536	1	1	65536
2005-11-04	IBM	Blue Gene Custom Interconnect	40960	32768	32768	2	32768	1	1	32768
2007-11-06	Cray Inc.	Seastar	12960	12960	25920	2	25920	1	1	25920
2006-11-10	Cray Inc.	Cray custom	12960	12960	25920	2	25920	1	1	25920
2006-11-10	Cray Inc.	Cray custom	12960	12960	25920	2	25920	1	1	25920
2007-11-06	Cray Inc.	Seastar	12960	12800	25600	2	25600	1	1	25600
2005-11-04	IBM	Blue Gene Custom Interconnect	40960	32768	16384	2	16384	1	1	16384
2006-11-06	Cray Inc.	Cray SeaStar	10424	10404	10404	2	10404	1	1	10404
2006-11-06	Cray Inc.	Cray SeaStar	10424	10404	10404	2	10404	1	1	10404
2006-01-11	Cray Inc.	Seastar	10368	10350	10350	1	10350	1	1	10350
2005-11-04	IBM	HPS	10240	10240	10240	1	10240	1	1	10240
2008-05-14	Cray Inc.	Cray Seastar	8608	8464	8464	4	8464	1	1	8464
2006-08-02	IBM	HPS	12288	8192	8192	1	8192	1	1	8192
2005-11-09	IBM	HPS	8192	8192	16384	1	8192	2	2	8192
2007-04-09	Cray Inc.	Cray Seastar	8192	8190	8190	2	8190	1	1	8190
2005-11-12	Cray Inc.	Cray Seastar	5212	5208	5208	1	5208	1	1	5208
2005-11-12	Cray Inc.	Cray Seastar	5212	5208	5208	1	5208	1	1	5208
2005-11-10	Cray Inc.	Cray Seastar	5212	5208	5208	1	5208	1	1	5208
2005-11-10	Cray Inc.	Cray Seastar	5212	5208	5208	1	5208	1	1	5208
2005-08-01	Cray Inc.	Cray XT3 MPP Interconnect	5200	5200	5200	1	5200	1	1	5200
2005-11-13	Cray Inc.	Seastar	4178	4128	4128	1	4128	1	1	4128
2005-09-19	Cray Inc.	Cray XT3 MPP Interconnect	4096	4096	4096	1	4096	1	1	4096
2005-06-21	Cray Inc.	Cray XT3 MPP Interconnect	3748	3744	3744	1	3744	1	1	3744
2007-12-19	SGI	InfiniBand	512	512	2048	4	2048	1	1	2048

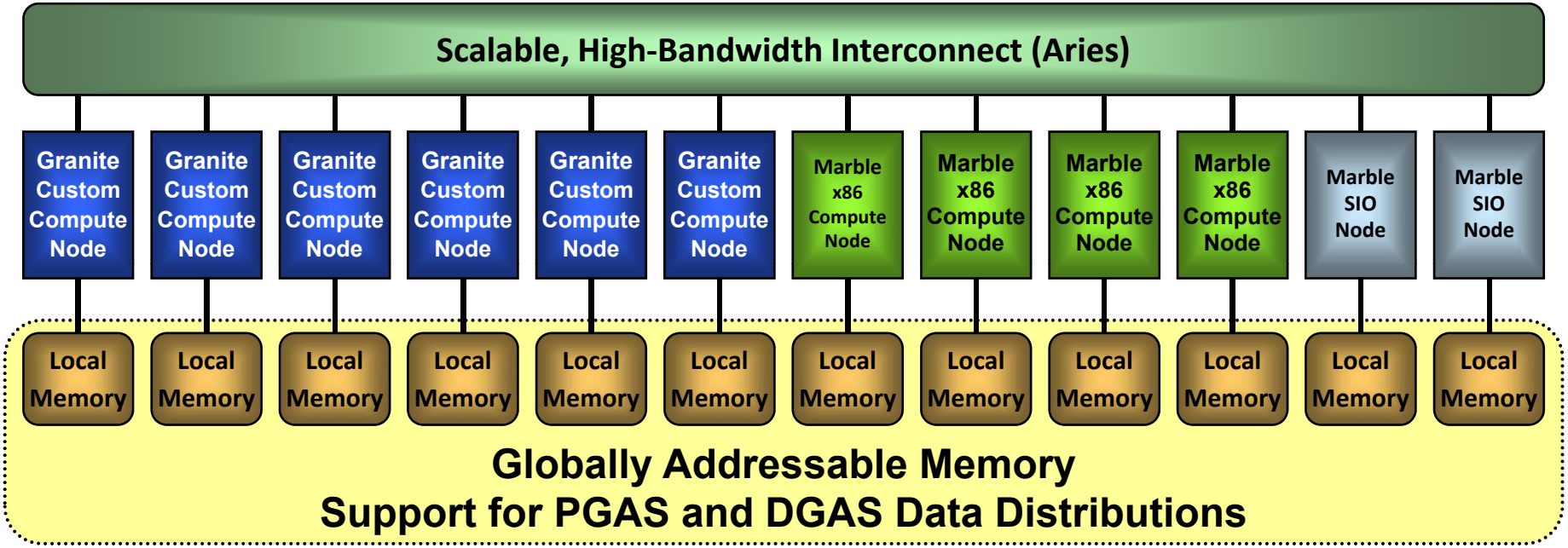
The Cray Roadmap

Realizing Our Adaptive Supercomputing Vision



Cray Cascade Project

Cray High Productivity Computing Systems



- Tightly integrated hybrid computing
- Configurable network, memory, processing and I/O
- Globally addressable memory
- Very high performance communication and synchronization

Path Forward