CRAY
THE SUPERCOMPUTER COMPANY

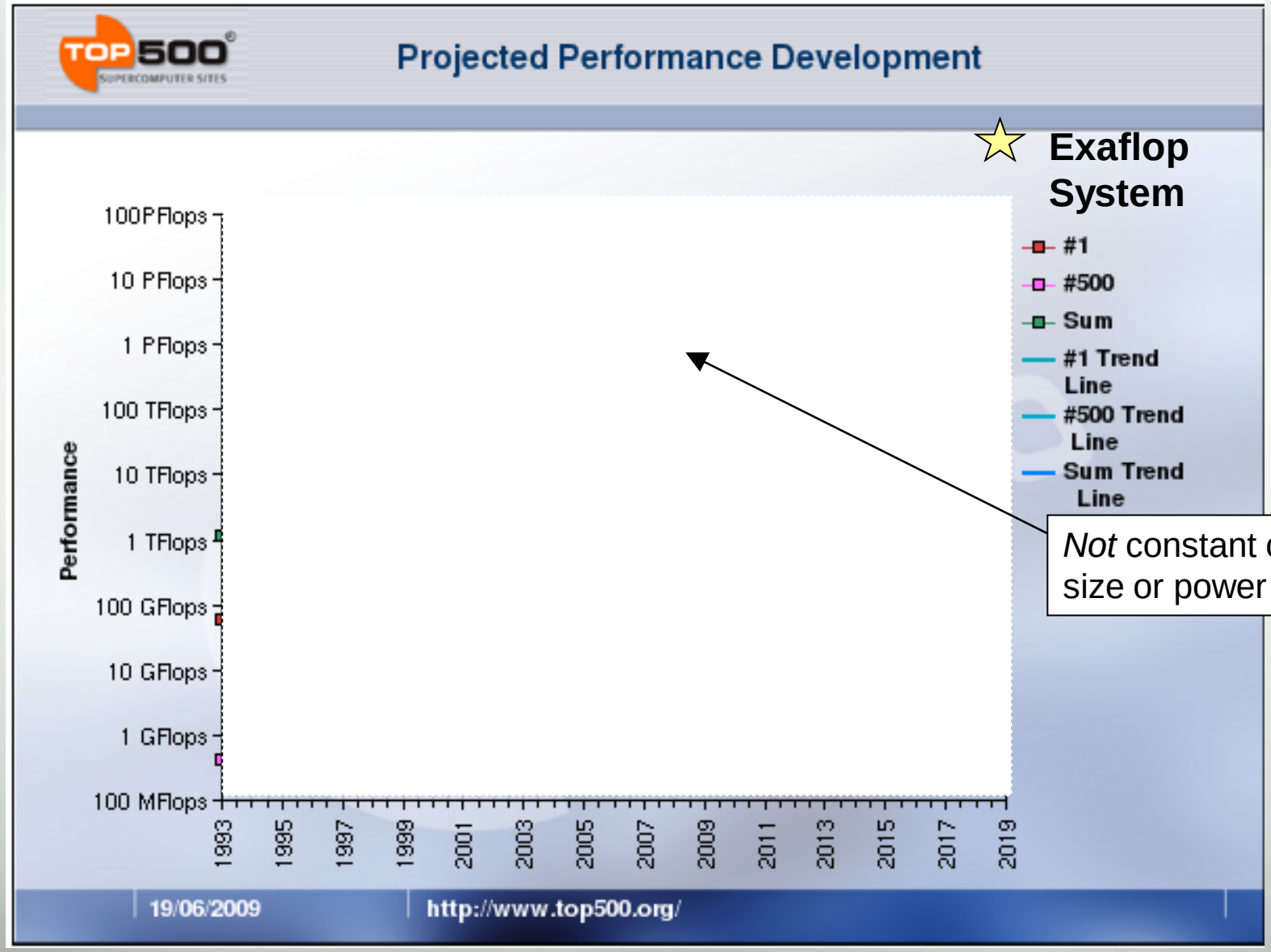
ENA Energy-Aware High Performance Computing HPC

Energy Efficiency Aspects in Cray Supercomputers

September 2010

Vincent Pel (vpel@cray.com)

Expectations of an Exaflop



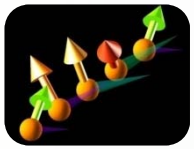
Breaking Sustained Performance Barriers

1 GF – 1988: Cray Y-MP; 8 Processors

- Static finite element analysis

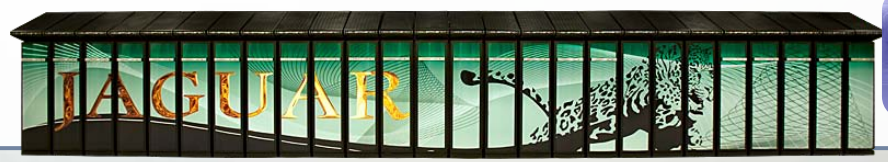
1 TF – 1998: Cray T3E; 1,024 Processors

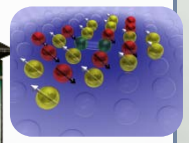
- Modeling of metallic magnet atoms



1 PF – 2008: Cray XT5; 150,000 Processors

- Superconductive materials





1 EF – ~2018: ~10,000,000 Processors

Transitions in HPC

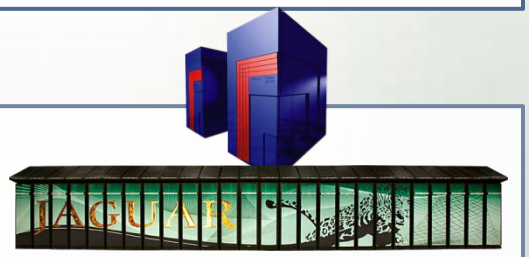
Change	Technologies	Parallelism
Scalar to Vectors	Dependency analysis, vectorizing compilers	1 to 10s

1 GF – 1988: Cray Y-MP; 8 Processors

Change	Technologies	Parallelism
SMP	Multitasking, Microtasking, OpenMP	10s to 100s

Change	Technologies	Parallelism
SMP to MPP	Distributed memory programming, PVM, MPI	100s to millions

1 TF – 1998: Cray T3E; 1,024 Processors



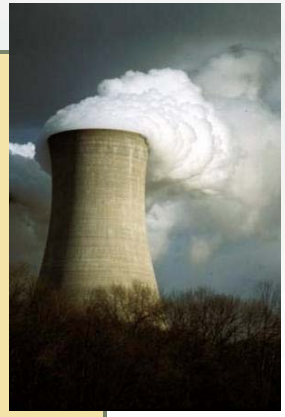
1 PF – 2008: Cray XT5; 150,000 Processors

Change	Technologies	Parallelism
MPP to ??	Accelerators, Manycore, Chapel, X10	Millions to Billions?

1 EF – ~2018: ~10,000,000 Processors

Key Challenges to Get to an Exascale

Power



- Traditional voltage scaling is over
- Power now a major design constraint
- Cost of ownership
- Driving significant changes in architecture

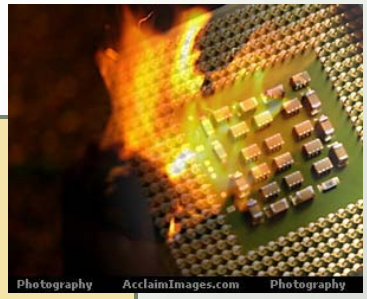
Concurrency

- A billion operations per clock
- Billions of refs in flight at all times
- Will require *huge* problems
- Need to exploit *all* available parallelism

Programming Difficulty

- Concurrency and new micro-architectures will significantly complicate software
- Need to hide this complexity from the users

Resiliency



- Many more components
- Components getting less reliable
- Checkpoint bandwidth not scaling

The Power Problem

- The most power-efficient standard processors today can achieve ~400 MF/watt on HPL
  - This corresponds to ~2.5 MW per Petaflop
  - Or about 2.5 GW for an Exaflop!
- DARPA UHPC goal: 50 GF/watt in ~8 years
  - Corresponds to 20 MW for an Exaflop
  - *Need a factor of over 100 improvement*

Three Steps to Power Efficiency

1. Power and cool the system efficiently
   - PUE (ratio of facility power to machine power) should be as close as possible to 1
   - Power delivery efficiency *inside* the cabinet is important too
   - Spend most of the energy on the computer itself, not on power delivery and cooling infrastructure

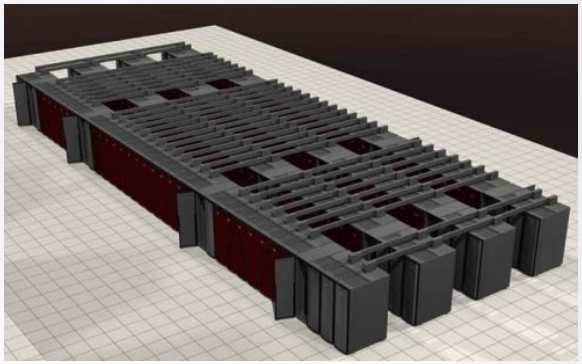
2. Architect system (processors, memory, network) to maximize power efficiency
   - Spend most of the computer's power on actual computation
   - Minimize energy spent on data movement and control overhead

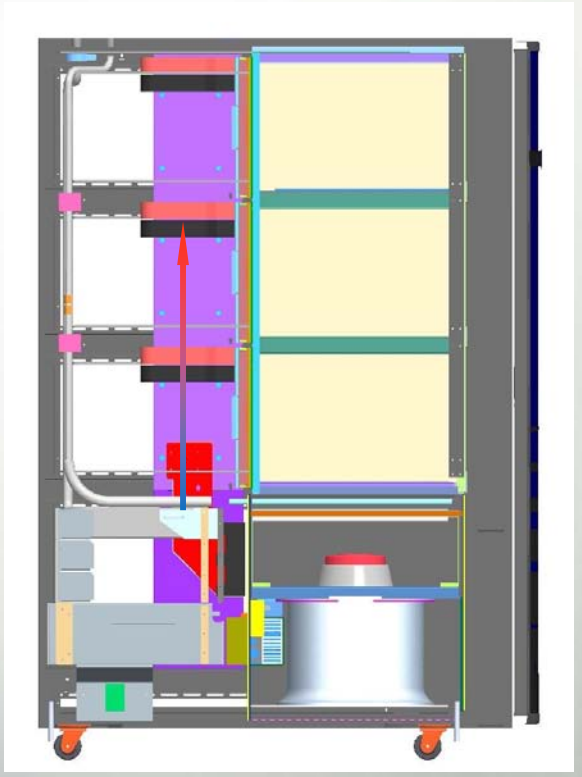
3. Sustain a high fraction of peak performance
   - Eliminate bottlenecks; don't leave performance on the floor
   - *Sustained* flops/watt is what matters, *not* peak flops/watt

Cray XT Vertical Air Cooling

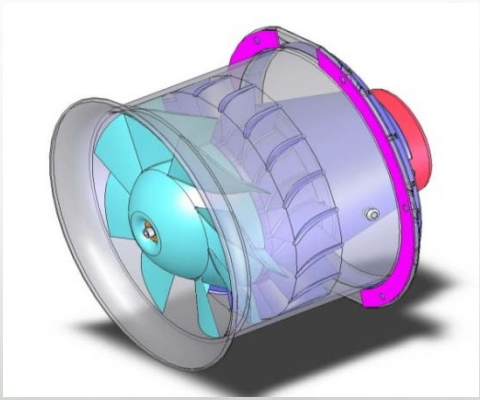
First introduced on XT3 & Red Storm in 2004.

- Eliminates air flow through network cabling.
- Eliminates hot air recirculation.
- Allows efficient exhaust transport to ceiling plenum.
- Eliminates hot aisle working environment.
- Allows implementation of single, high efficiency industrial blower.
- Air cooling allows for simplified blades and better upgradeability
- Much more power efficient than muffin fans

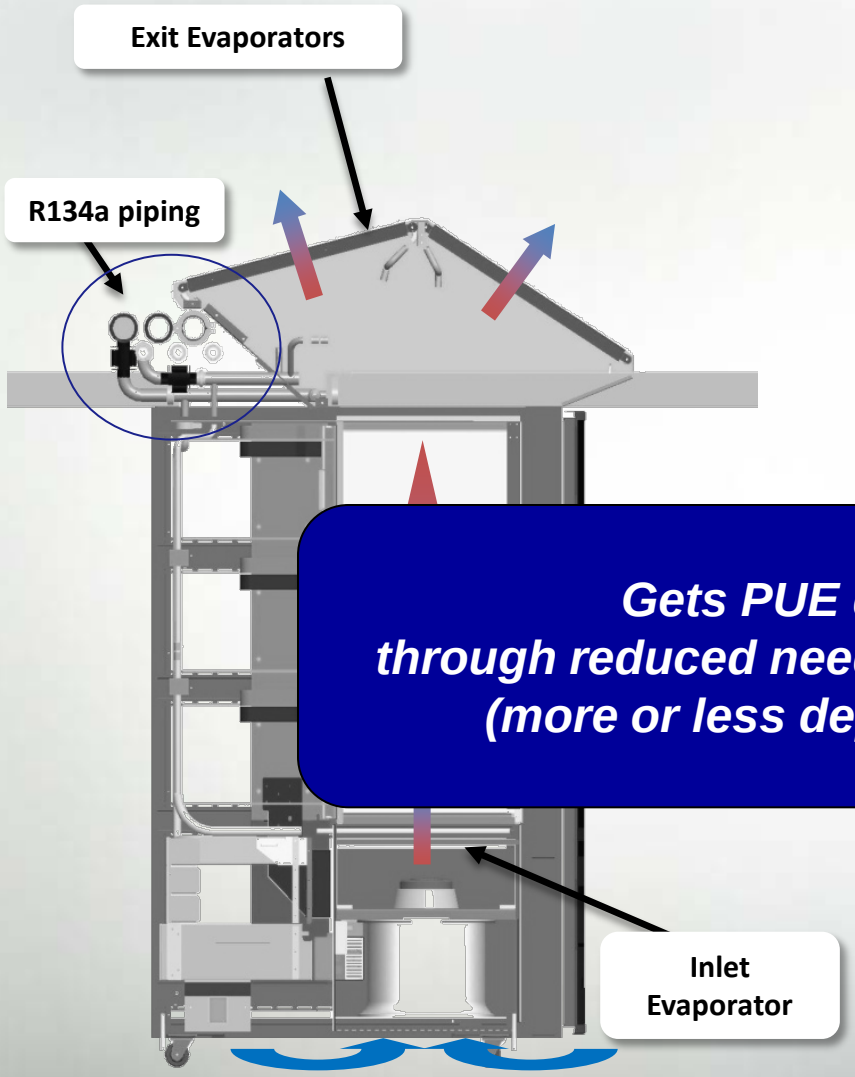


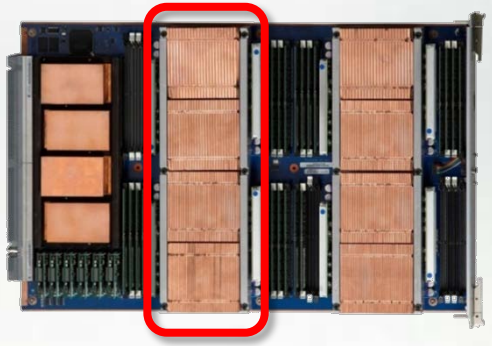


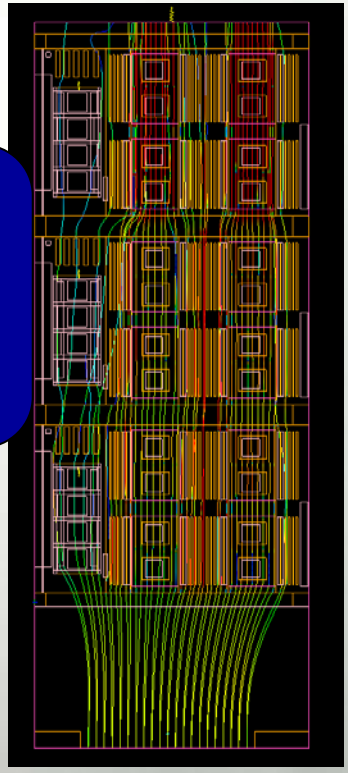
- Questions:
  - How does the PUE take blower efficiency in consideration ?
  - Is PUE a good indication of Power efficiency ?



ECOphlex and Progressive Airflow Technology in the Cray XT Cabinet

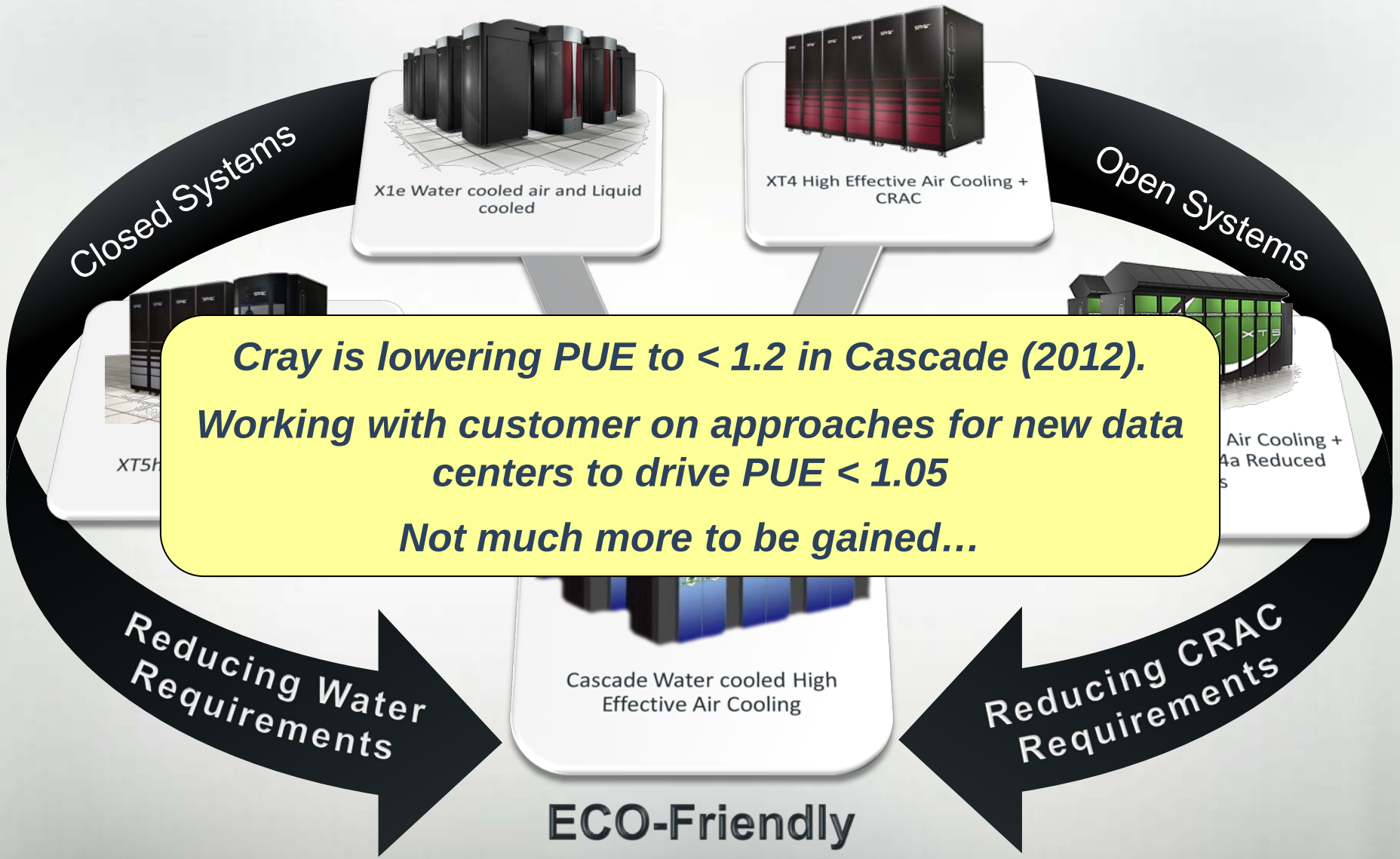






Gets PUE down to ~1.25 through reduced need for chillers and CRACs (more or less depending on climate)

Next Generation Cooling Infrastructure



Cray is lowering PUE to < 1.2 in Cascade (2012).

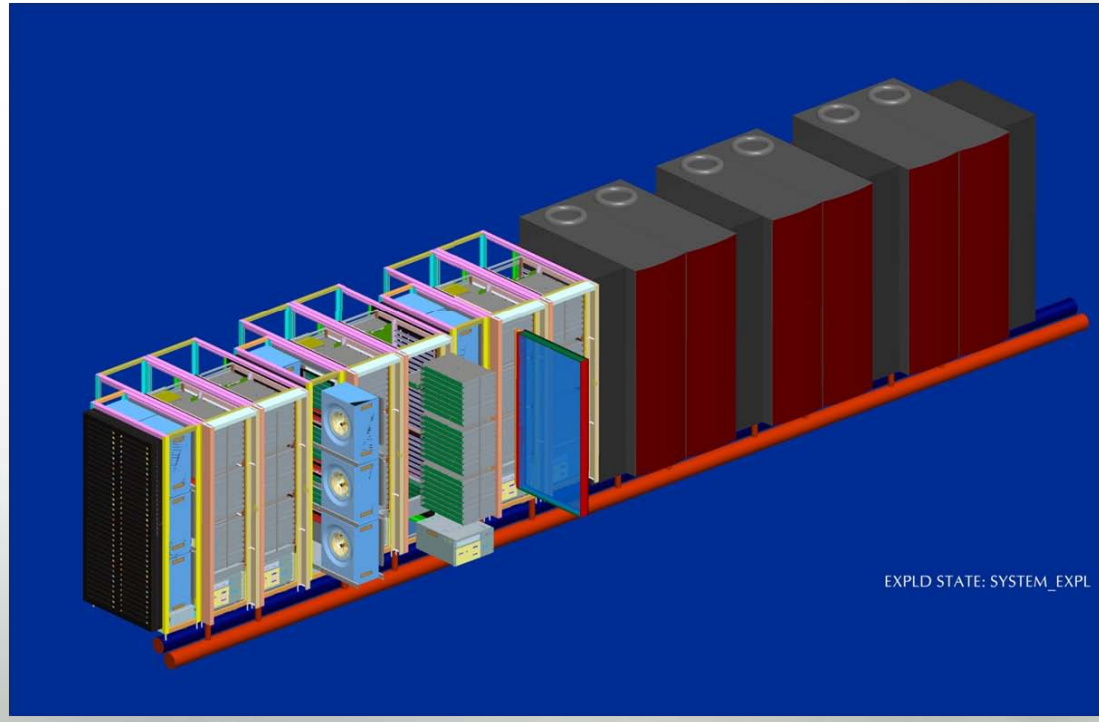
Working with customer on approaches for new data centers to drive PUE < 1.05

Not much more to be gained…

Cascade Transverse Cooling

Moving cooling air transversely, this architecture allows safe cooling water internal to rack.

- Eliminates intermediate heat transfer step in EcoPlex (refrigerant /water): A greener cooling solution.
- Allows expanded operating humidity envelope: Less energy for dehumidification.
- Uses less fan power: Lower cost of ownership.
- Improves resiliency
- Uses the largest face and allows higher density



Further reducing PUE: customer initiatives

- HECToR has some months of free cooling with a water circuit on the roof
  - Because of the nice fresh Edinburgh climate
  - Because of the ECOphlex flexibility in terms of temperature
- CSCS is building a new computing center with cold water directly taken from the lake
- Another customer is planning to reuse the hot air exiting their XT system to heat some buildings
  - Because of the bottom to top air cooling concept
- Several ways to improve PUE by infrastructure work, but often vendor dependent !

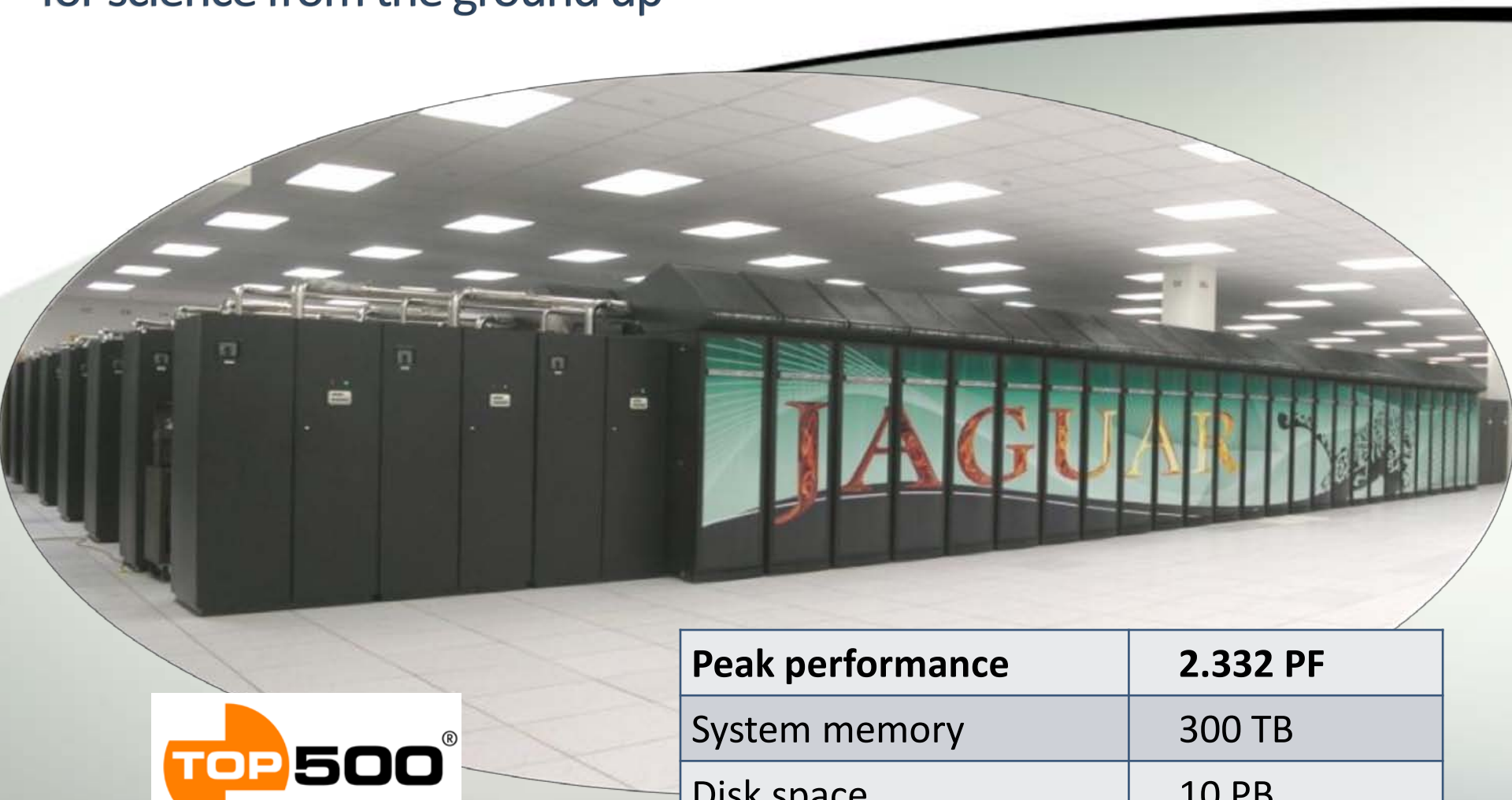
CRAY
THE SUPERCOMPUTER COMPANY

Better usage of hardware resources

Cray has fully entered the petascale era

Jaguar: World's most powerful computer—Designed for science from the ground up

CRAY
THE SUPERCOMPUTER COMPANY



TOP500®
SUPERCOMPUTER SITES

#1 Nov. 2009
#1 June 2010

Peak performance	**2.332 PF**
System memory	300 TB
Disk space	10 PB
Disk bandwidth	240+ GB/s
Interconnect bandwidth	374 TB/s

Kraken World's most powerful academic computer

CRAY THE SUPERCOMPUTER COMPANY

TOP500 SUPERCOMPUTER SITES

#4 June 2010

Peak performance	**1.03 petaflops**
System memory	129 terabytes
Disk space	3.3 petabytes (raw)
Disk bandwidth	30 gigabytes/second

A UT/ORNL PARTNERSHIP
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
NICS

More Petascale Systems in 2010

CRAY
THE SUPERCOMPUTER COMPANY

Better usage of hardware resources

Which Takes More Energy?

Performing a 64-bit floating-point FMA:

$$
\begin{aligned}
& 893{,}500.288914668 \\
\times\ & 43.90230564772498 \\
\hline
= \ & 39{,}226{,}722.78026233027699 \\
+ \ & 2.02789331400154 \\
\hline
= \ & 39{,}226{,}724.80815564
\end{aligned}
$$

Or moving the three 64-bit operands 20 mm across the die:

This one takes over 3x the energy!

And loading the data from off chip takes > 10x more yet

Flops are cheap, communication is expensive.
Exploiting data locality is *critical* for energy efficiency.

Processor Architecture: Power vs. Single Thread Performance

- Multi-core architectures are a good first response to power issues
  - Performance through parallelism, not frequency
  - Exploit on-chip locality
- However, conventional processor architectures are optimized for single thread performance rather than energy efficiency
  - Fast clock rate with latency(performance)-optimized memory structures
  - Wide superscalar instruction issue with dynamic conflict detection
  - Heavy use of speculative execution and replay traps
  - Large structures supporting various types of predictions
  - Relatively little energy spent on actual ALU operations
- Could be much more energy efficient with multiple simple processors, exploiting vector/SIMD parallelism and a slower clock rate
- But serial thread performance is really important (Amdahl's Law):
  - If you get great parallel speedup, but hurt serial performance, then you end up with a niche processor (less generally applicable, harder to program)

Exascale Conclusion: Heterogeneous Computing

- To achieve scale and sustained performance per {$,watt}, must adopt:
  - …a *heterogeneous* node architecture
    - fast serial threads coupled to many efficient parallel threads
  - …a deep, explicitly managed memory hierarchy
    - to better exploit locality, improve predictability, and reduce overhead
  - …a microarchitecture to exploit parallelism at all levels of a code
    - distributed memory, shared memory, vector/SIMD, multithreaded
    - (related to the "concurrency" challenge—leave no parallelism untapped)
- This sounds a lot like a GPU accelerators…
- NVIDIA Fermi™ has made GPUs feasible for HPC
  - Robust error protection and strong DP FP, plus programming enhancements
- Expect GPUs to make continued and significant inroads into HPC
  - Compelling technical reasons + high volume market
- Programmability remains primary barrier to adoption
  - Cray is focusing on compilers, tools and libraries to make GPUs easier to use
  - There are also some structural issues that limit applicability of current designs…
- Technical direction for Exascale:
  - Unified node with "CPU" and "accelerator" on chip sharing common memory
  - Very interesting processor roadmaps coming from Intel, AMD and NVIDIA….

Cray's Network Evolution

SeaStar

- Built for scalability to 250K+ cores
- Very effective routing and low contention switch

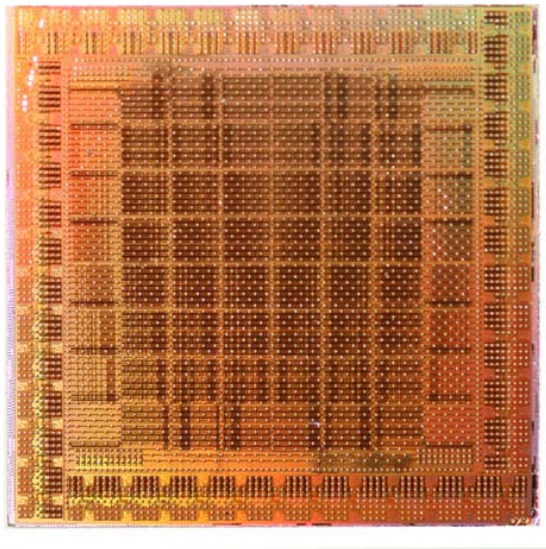
Gemini

- 100x improvement in message throughput
- 3x improvement in latency
- PGAS Support, Global Address Space in hardware
- Scalability to 1M+ cores, Improved Reliability in SW and HW

Aries

- Low Diameter, High Bandwidth Network
- Very effective routing and low contention switch
- Electro-Optical Signaling, Improved Reliability in SW and HW

Power-Efficient Networks



64 port YARC router in Cray X2

- Cray pioneered the use of high radix routers in HPC
  - Becoming optimal due to technology shift
    - Router pin bandwidth growing vs. packet length
    - Reduces serialization latency of narrow links
  - Reduced network diameter (number of hops)
    - Lowers network latency
    - Lowers network cost
  - But higher radix network require longer cable lengths
    - Limits electrical signaling speed
- Advent of cost-effective optics allows longer cable lengths
  - Optics are now cost effective above ~7 meters (and dropping)
  - Cost, bandwidth and power are relatively insensitive to cable length
  - Opens the door to some innovative new topologies
- Future Cray systems will be based on hybrid, electrical-optical networks
  - Cost-effective, scalable global bandwidth
  - Very low network diameter (small number of hops) ⇒ very energy efficient
- Lower power electrical and optical links are critically important
  - Optics directly off chip package provide potential for much higher bandwidth

Cray CLE3, An Adaptive Linux OS specifically for HPC

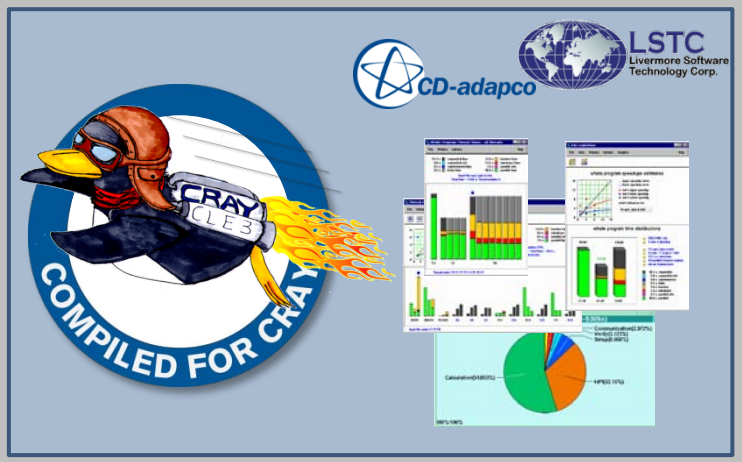
CRAY LINUX ENVIRONMENT CLE3

ESM – *Extreme Scalability Mode*

- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

CCM –*Cluster Compatibility Mode*

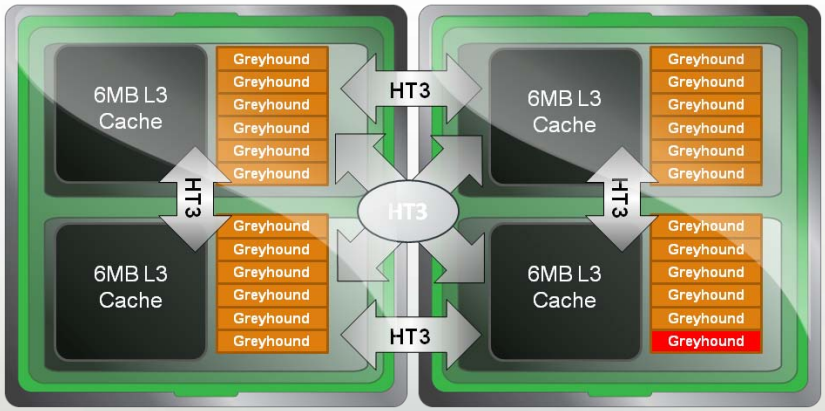
- No compromise *compatibility*
- Fully standard x86/Linux
- Standardized Communication Layer
- Out-of-the-box ISV Installation
- ISV applications simply install and run



CLE3 run mode is set by the user on a job-by-job basis to provide full flexibility

An Example of Pre-Exascale Linux Scaling Core Specialization

- Benefit: Eliminate noise with overhead (interrupts, daemon execution) directed to a single core
- Rearranges existing work
  - Overhead is confined, giving app exclusive access to remaining cores
- Helps some applications, needs to be adaptive
  - Future nodes with larger core counts will see even more benefit
- *Like the CCM, this feature is adaptable and available on a job-by-job basis*



- *Have observed up to 30 % improvement on some applications*
- *Can be converted in Power savings !*

Programming Models for Future Processors

- Programming model and tools will be critical to achieving practical Exaflops
- Need a single programming model that is portable across machine types, and also forward scalable in time
  - Portable expression of heterogeneity and multi-level parallelism
  - Programming model and optimization should not be significantly difference for “accelerated” nodes and multi-core x86 processors
- Need to shield user from the complexity of dealing with heterogeneity
  - High level language with good complier and runtime support
  - Optimized libraries
- Directive-based approach makes sense
  - A Cray employee is co-chairing OpenMP group on accelerators
  - Plan to have “accelerator” directives in 4.0
- Identifying the parallelism is the hard part, not the mechanics
  - Provide tools to sophisticated users to make this easier
  - Compiler and runtime can map the parallelism onto the hardware

Conclusions

- We can still improve PUE
- We can still improve internal efficiency
- We can partner with customers to design the most efficient infrastructure and supercomputer
- Bust most of the potential is in the application efficiency.
- We need to provide efficient tools to allow users to make a better usage of hardware resources
- It is all about Sustained Performance

Thank You. Questions?