20th Anniversary of PDC

Exascale Computing, a new challenge ahead

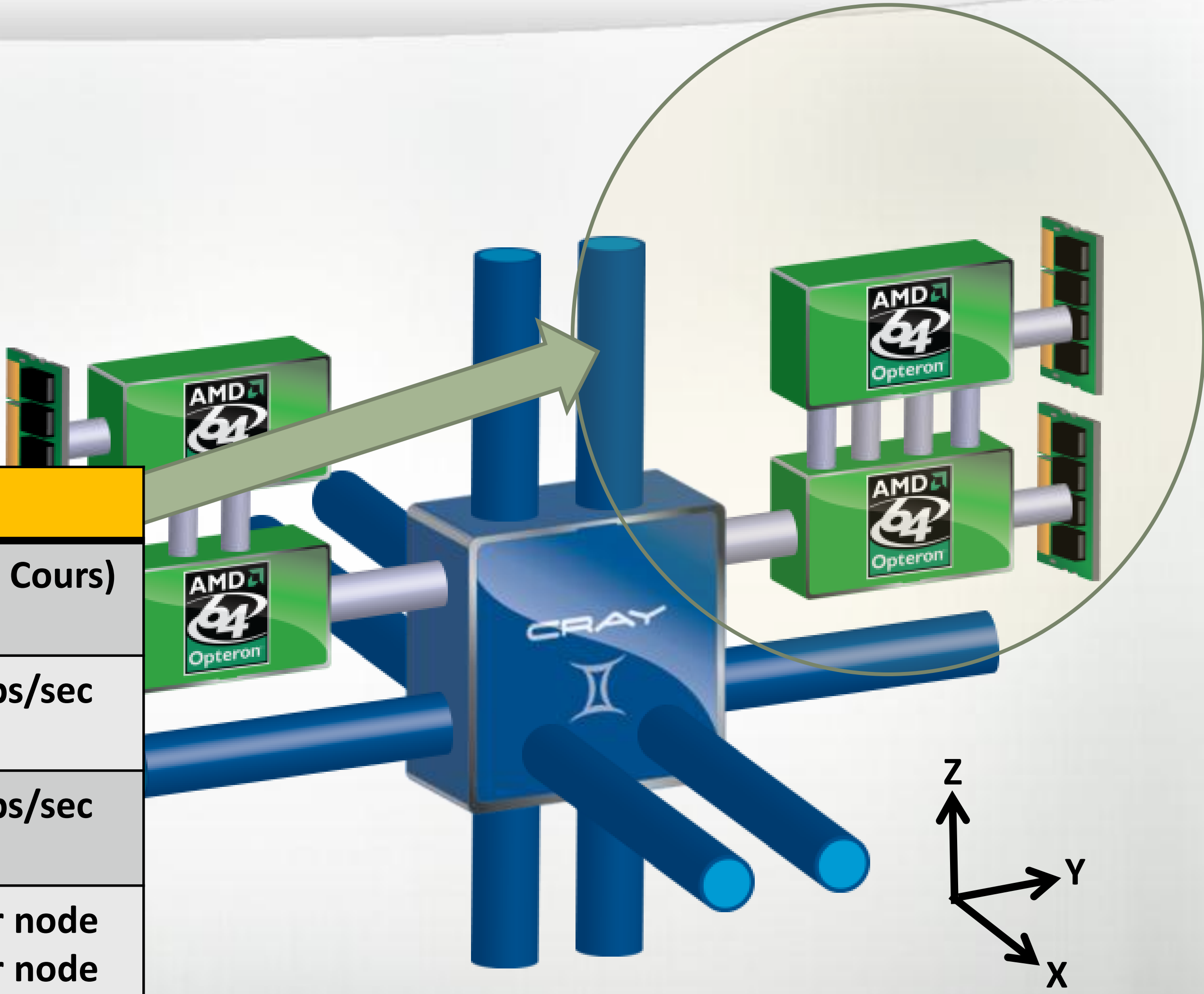
Mario Mattia
Cray Europe

Cray XE6 System

- System announced at CUG in Edinburgh, May 2010
- Over $200M in booked orders
  - Over 5 PF's Cray XE6 systems ordered
- Deliveries starting in July (really on June 11th)
- Key New Technologies
  - Gemini interconnect
  - Series 6 Processor blade to support AMD's new 6100 series Opteron
  - New XIO blade for I/O
  - CLE 3 Operating System

Cray XE6 Node

Node Characteristics	
Number of Cores	24 (Magny Cours)
Peak Performance MC-12 (2.2)	211 Gflops/sec
Peak Performance MC-8 (2.4)	153 Gflops/sec
Memory Size	32 GB per node 64 GB per node
Memory Bandwidth (Peak)	83.5 GB/sec

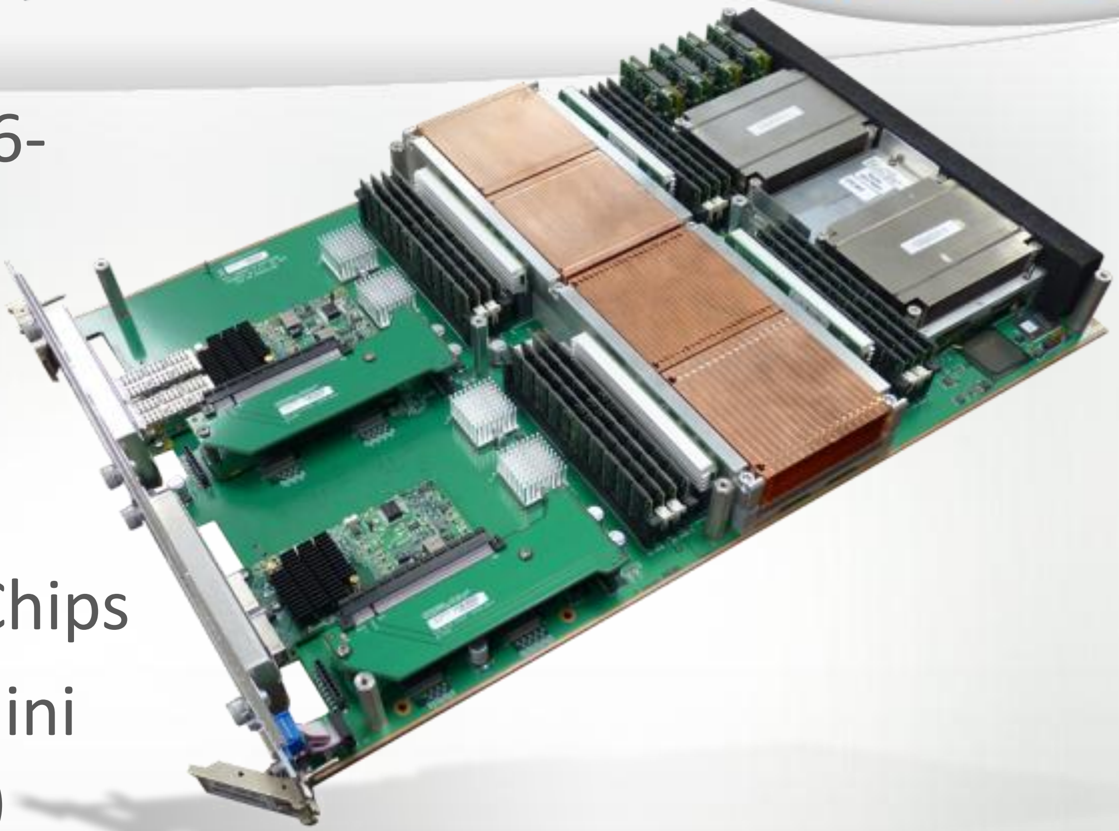


Gemini Advanced Features

- Globally addressable memory provides efficient support for UPC, Co-array FORTRAN, Shmem and Global Arrays
- Atomic memory operations
  - Provides fast synchronization needed for one-sided communication models
- Pipelined global loads and stores
  - Allows for fast irregular communication patterns
- Compared to Cray SeaStar
  - 100x improvement in message throughput
  - 3x improvement in latency

Cray XE6 I/O Blade Summary

- 4 Single Socket Nodes (6-core processors)
- 4 DDR2 DIMMs per Node
  - 4GB DIMMs supported
- 4 AMD SR5670 Bridge Chips
- Will only ship with Gemini Interconnect (XE Series)



Blade Feature	XT3 SIO Blade	FSIO Blade	FSIO Difference
# of cores	4	24	6x
Max. memory size	16GB	128GB	8x
Memory Bandwidth	12.8 GB/s	51.2 GB/s	4x
Sustained I/O Bandwidth	2.7 GB/s	10.8 GB/s	4x

CLE3, An Adaptive Linux OS designed specifically for HPC

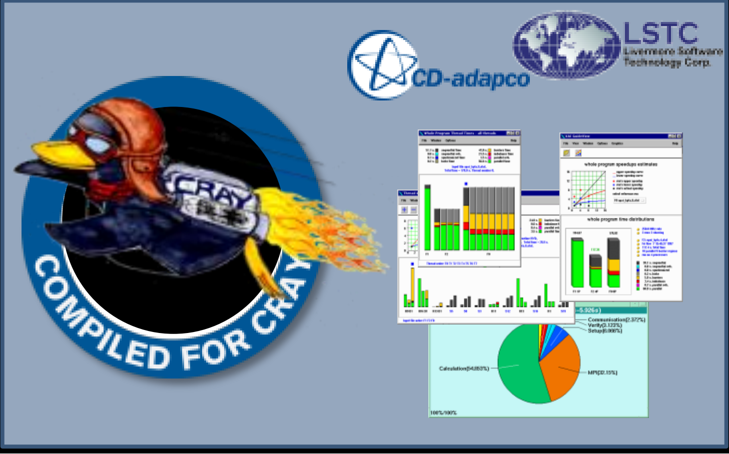
CRAY LINUX ENVIRONMENT CLE3

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

On the road to Exaflop

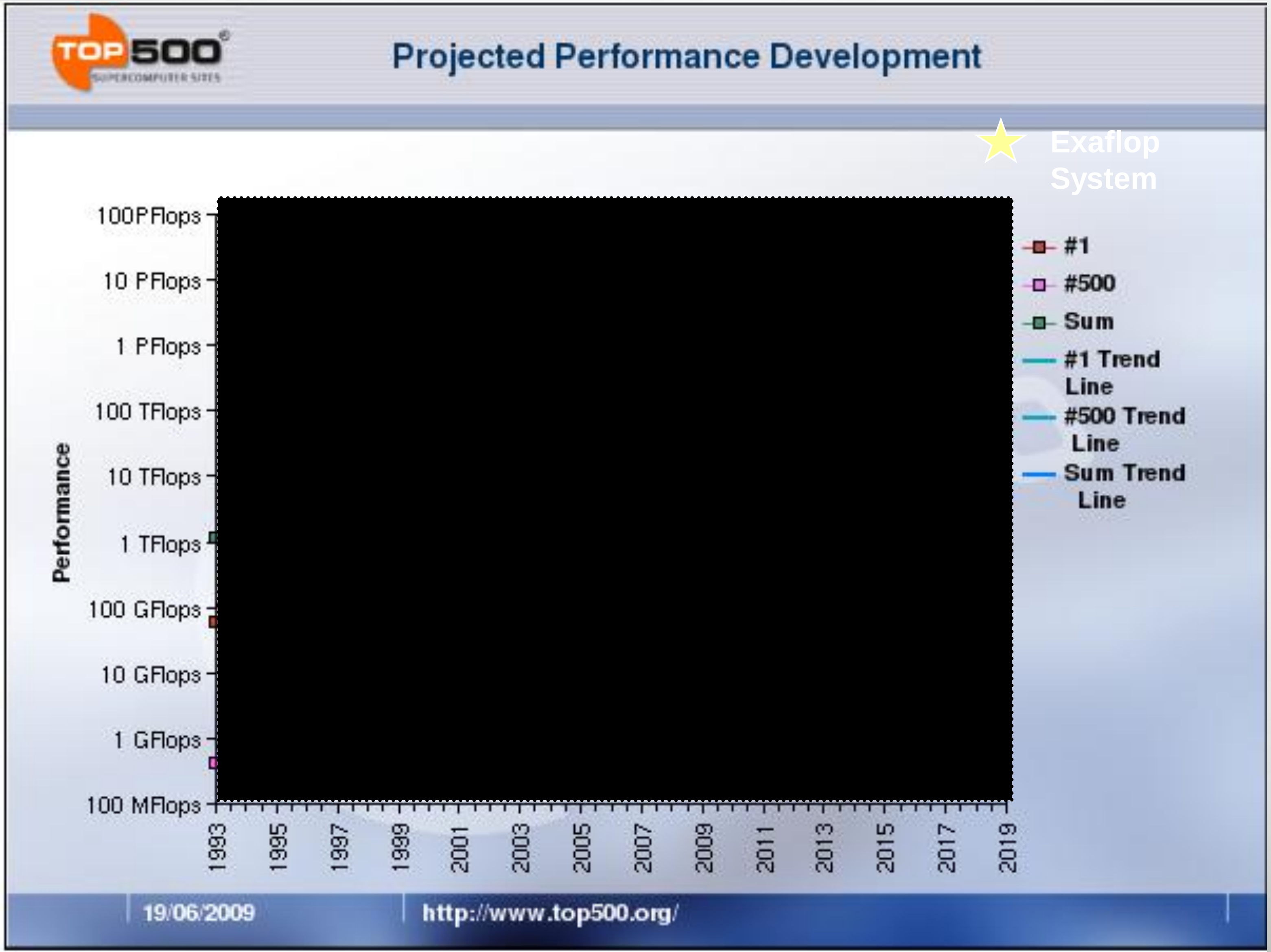
It's All About the Exascale....

1,000,000,000,000,000,000

floating point operations per second

sustained

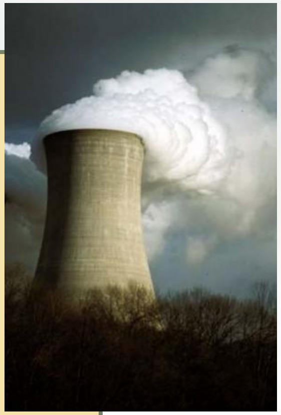
Expectations of an Exaflop by ~2019


[redacted]

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

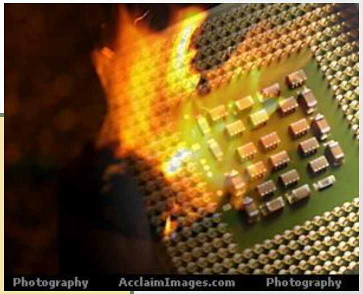
- A billion operations per clock
- Billions of refs in flight at all times
- Will require *huge* problems
- Need to exploit *all* available parallelism

Programming Difficulty

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

Resiliency

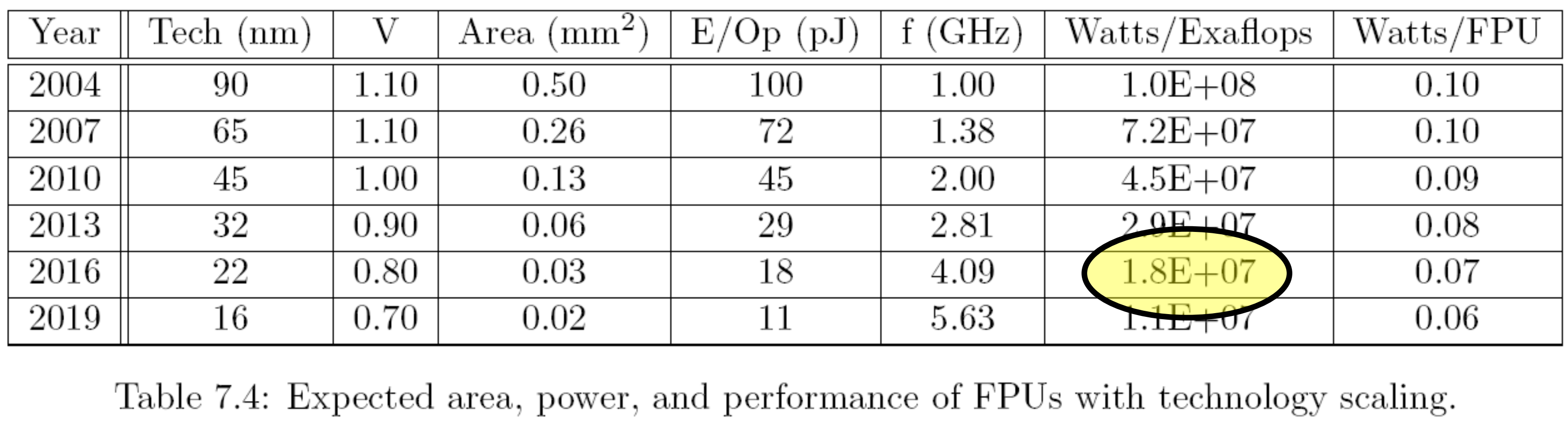
- Many more components
- Components getting less reliable
- Checkpoint bandwidth not scaling



The Power Problem

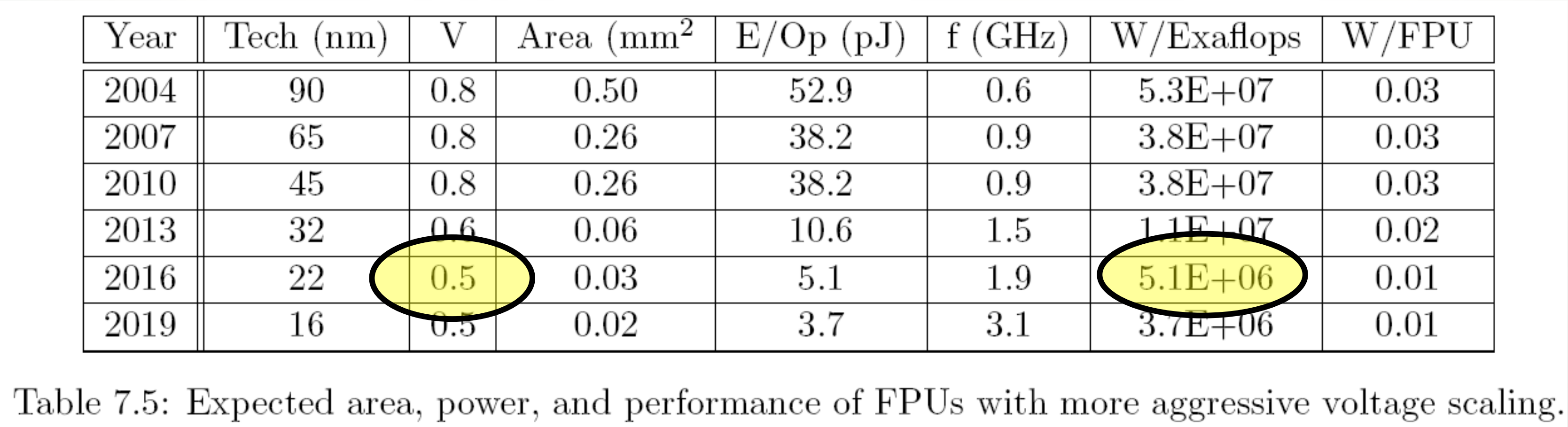
- The most power-efficient standard processors today can achieve ~400 MF/watt on HPL
  - This corresponds to ~2.5 MW per Petaflop
  - Or about 2.5 GW for an Exaflop!
- DARPA UHPC goal: 50 GF/watt in ~8 years
  - Corresponds to 20 MW for an Exaflop
  - *Need a factor of over 100 improvement*

Projections for Exaflop Power (FPUs only)



Year	Tech (nm)	V	Area (mm^2)	E/Op (pJ)	f (GHz)	Watts/Exaflops	Watts/FPU
2004	90	1.10	0.50	100	1.00	1.0E+08	0.10
2007	65	1.10	0.26	72	1.38	7.2E+07	0.10
2010	45	1.00	0.13	45	2.00	4.5E+07	0.09
2013	32	0.90	0.06	29	2.81	2.9E+07	0.08
2016	22	0.80	0.03	18	4.09	1.8E+07	0.07
2019	16	0.70	0.02	11	5.63	1.1E+07	0.06

Table 7.4: Expected area, power, and performance of FPUs with technology scaling.



Year	Tech (nm)	V	Area (mm^2)	E/Op (pJ)	f (GHz)	W/Exaflops	W/FPU
2004	90	0.8	0.50	52.9	0.6	5.3E+07	0.03
2007	65	0.8	0.26	38.2	0.9	3.8E+07	0.03
2010	45	0.8	0.26	38.2	0.9	3.8E+07	0.03
2013	32	0.6	0.06	10.6	1.5	1.1E+07	0.02
2016	22	0.5	0.03	5.1	1.9	5.1E+06	0.01
2019	16	0.5	0.02	3.7	3.1	3.7E+06	0.01

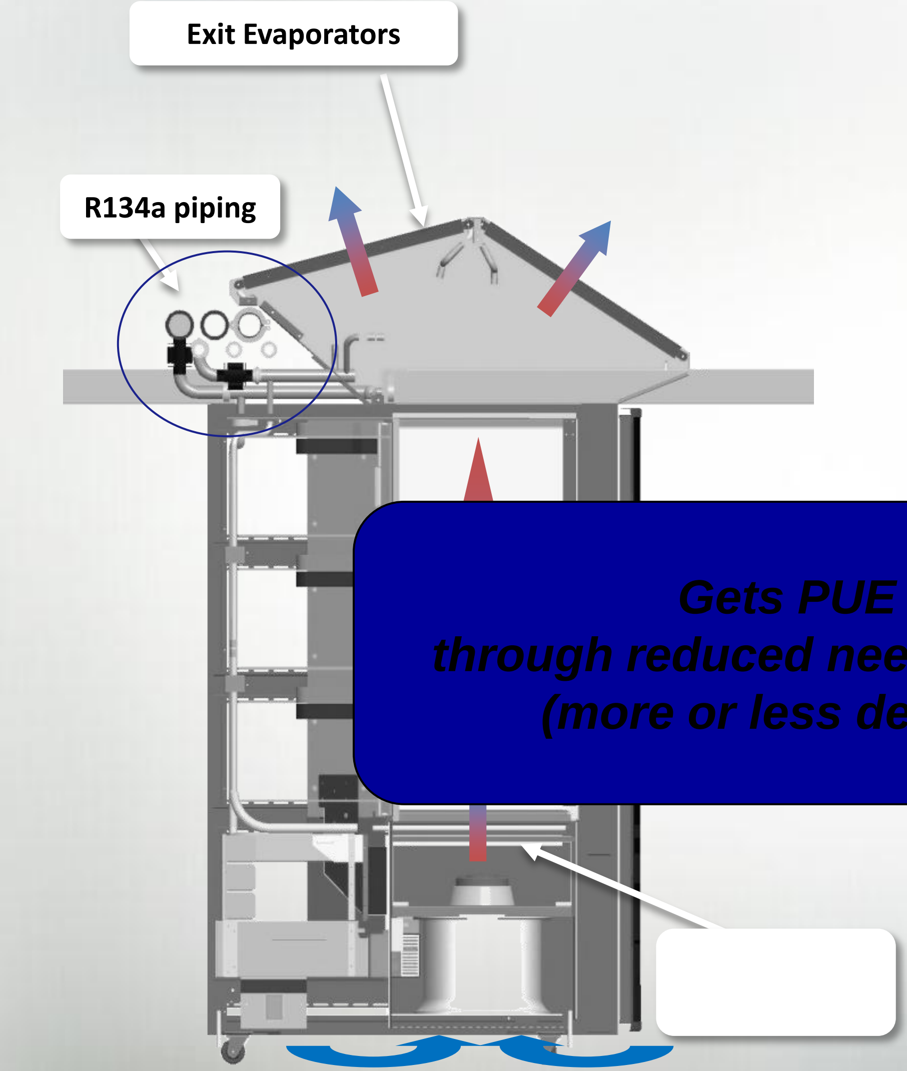
Table 7.5: Expected area, power, and performance of FPUs with more aggressive voltage scaling.

- With more aggressive voltage scaling than assumed in ITRS roadmap, can get power of FPUs for an Exaflop down to ~5MW in 22nm IC technology
- *FPUs are just one of many consumers of power*

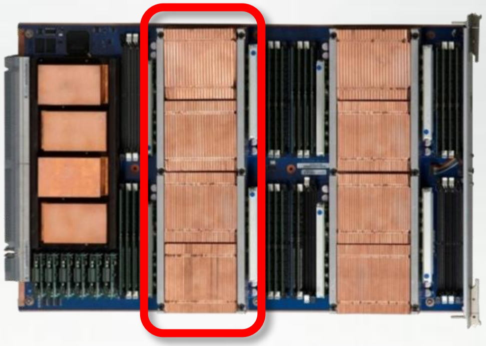
Three Steps to Power Efficiency

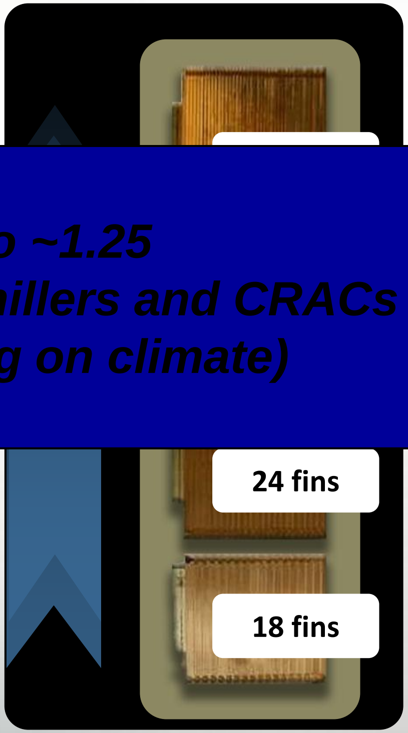
1. Power and cool the system efficiently
   - PUE (ratio of facility power to machine power) should be as close as possible to 1
   - Power delivery efficiency *inside* the cabinet is important too
   - Spend most of the energy on the computer itself, not on power delivery and cooling infrastructure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

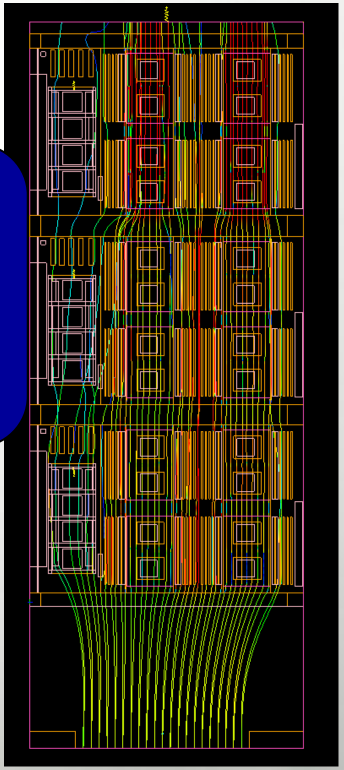
ECOphlex and Progressive Airflow Technology in the Cray XT Cabinet



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







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…..

- 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….

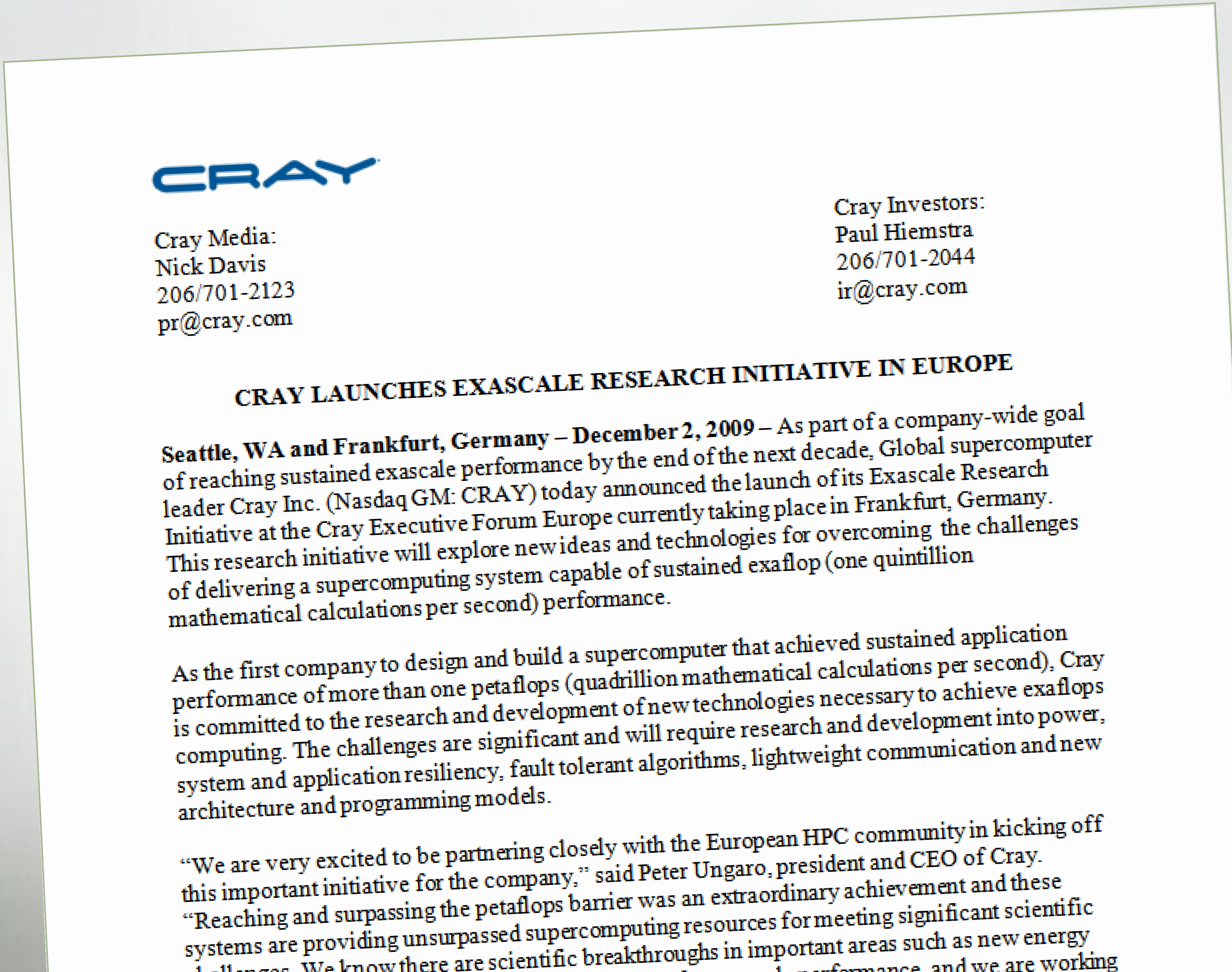
Programming Models for Future Processors

- 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

Application Resiliency

- Stopgap measures will hold us for a while
  - Checkpoint/Restart with Flash or similar non-volatile memory
  - Resilient communication protocols, better ECC,…
- We need *community* engagement and collaboration
  - Need better understanding of classes of apps and resiliency attributes
  - Need standards for applications
    - How to specify variable resiliency requirements (e.g.: reliability critical sections)
    - APIs for the system to provide failure information to the application
    - APIs for the applications to specify actions to the system (e.g.: restart *this* piece)
- May be potential for (semi-)automatic application resiliency
  - Use compiler techniques to decompose applications into sufficiently constrained work items
    - Some such decomposition already occurs as part of parallelization
    - Expect that user directives will be needed to make this work well enough
  - Use runtime techniques to reliably execute these work items
    - Work distribution is already done by some runtimes
    - Need to add reliability and encapsulation aspects
    - In-memory checkpoints and transactional techniques may apply

Exascale Research Initiative



Cray Media:
Nick Davis
206/701-2123
pr@cray.com

Cray Investors:
Paul Hiemstra
206/701-2044
ir@cray.com

CRAY LAUNCHES EXASCALE RESEARCH INITIATIVE IN EUROPE

Seattle, WA and Frankfurt, Germany – December 2, 2009 – As part of a company-wide goal of reaching sustained exascale performance by the end of the next decade, Global supercomputer leader Cray Inc. (Nasdaq GM: CRAY) today announced the launch of its Exascale Research Initiative at the Cray Executive Forum Europe currently taking place in Frankfurt, Germany. This research initiative will explore new ideas and technologies for overcoming the challenges of delivering a supercomputing system capable of sustained exaflop (one quintillion mathematical calculations per second) performance.

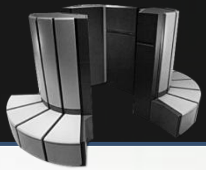
As the first company to design and build a supercomputer that achieved sustained application performance of more than one petaflops (quadrillion mathematical calculations per second), Cray is committed to the research and development of new technologies necessary to achieve exaflops computing. The challenges are significant and will require research and development into power, system and application resiliency, fault tolerant algorithms, lightweight communication and new architecture and programming models.

"We are very excited to be partnering closely with the European HPC community in kicking off this important initiative for the company," said Peter Ungaro, president and CEO of Cray. "Reaching and surpassing the petaflops barrier was an extraordinary achievement and these systems are providing unsurpassed supercomputing resources for meeting significant scientific challenges. We know there are scientific breakthroughs in important areas such as new energy performance, and we are working

Sustained Performance Milestones

1 GF – 1988: Cray Y-MP; 8 Processors

- Static finite element analysis

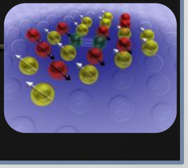


1 TF – 1998: Cray T3E; 1,024 Processors

- Modeling of metallic magnet atoms

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

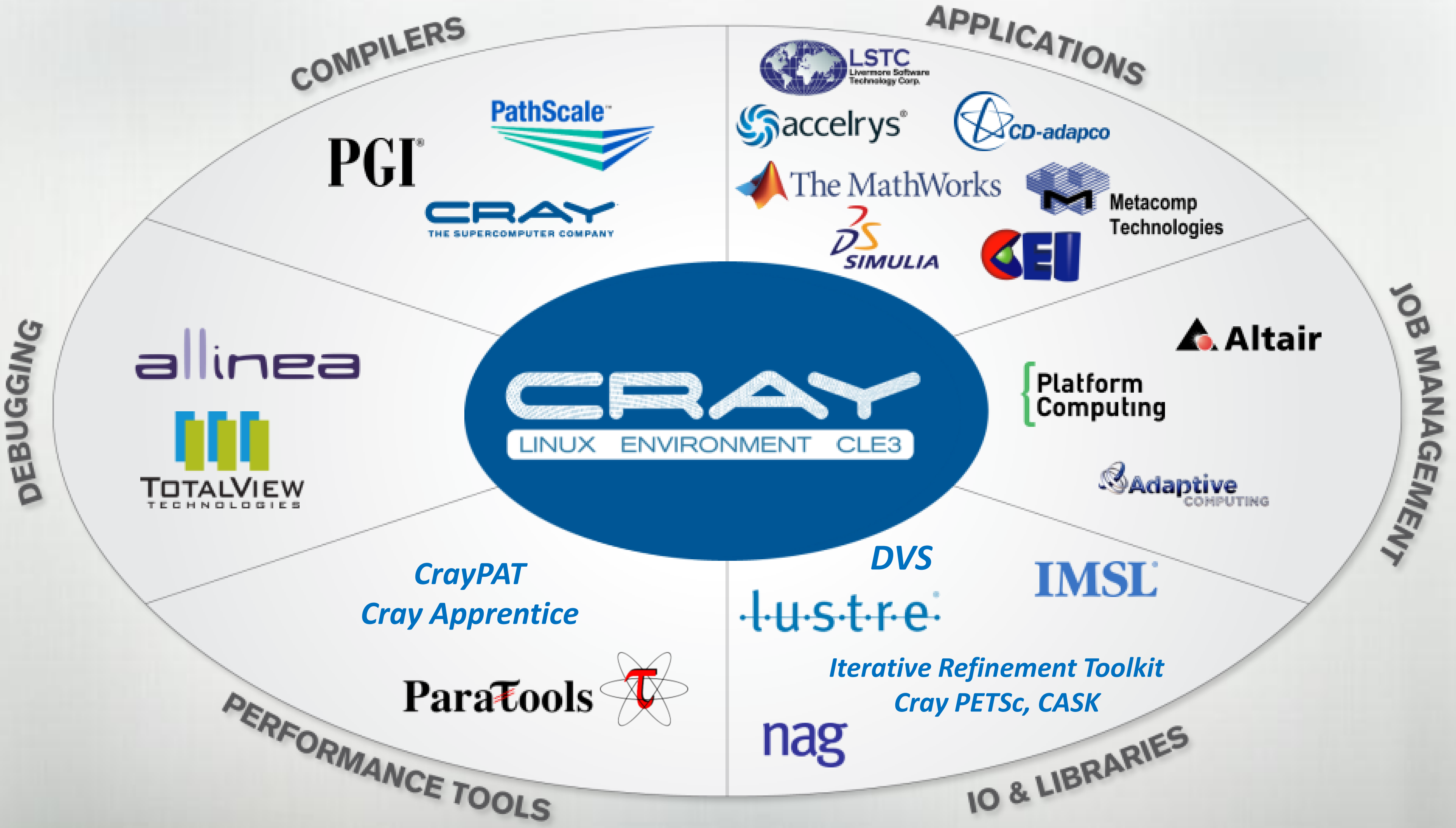
- Superconductive materials



1 EF -- ~2018: Cray ____; ~10,000,000 Processors

Thank you!

Cray Software Ecosystem



COMPILERS

PGI

PathScale

CRAY THE SUPERCOMPUTER COMPANY

APPLICATIONS

LSTC Livermore Software Technology Corp.

accelrys

CD-adapco

The MathWorks

Metacomp Technologies

SIMULIA

CEI

JOB MANAGEMENT

Altair

Platform Computing

Adaptive COMPUTING

IO & LIBRARIES

DVS

lustre

IMSL

Iterative Refinement Toolkit
Cray PETSc, CASK

nag

PERFORMANCE TOOLS

CrayPAT
Cray Apprentice

ParaTools

DEBUGGING

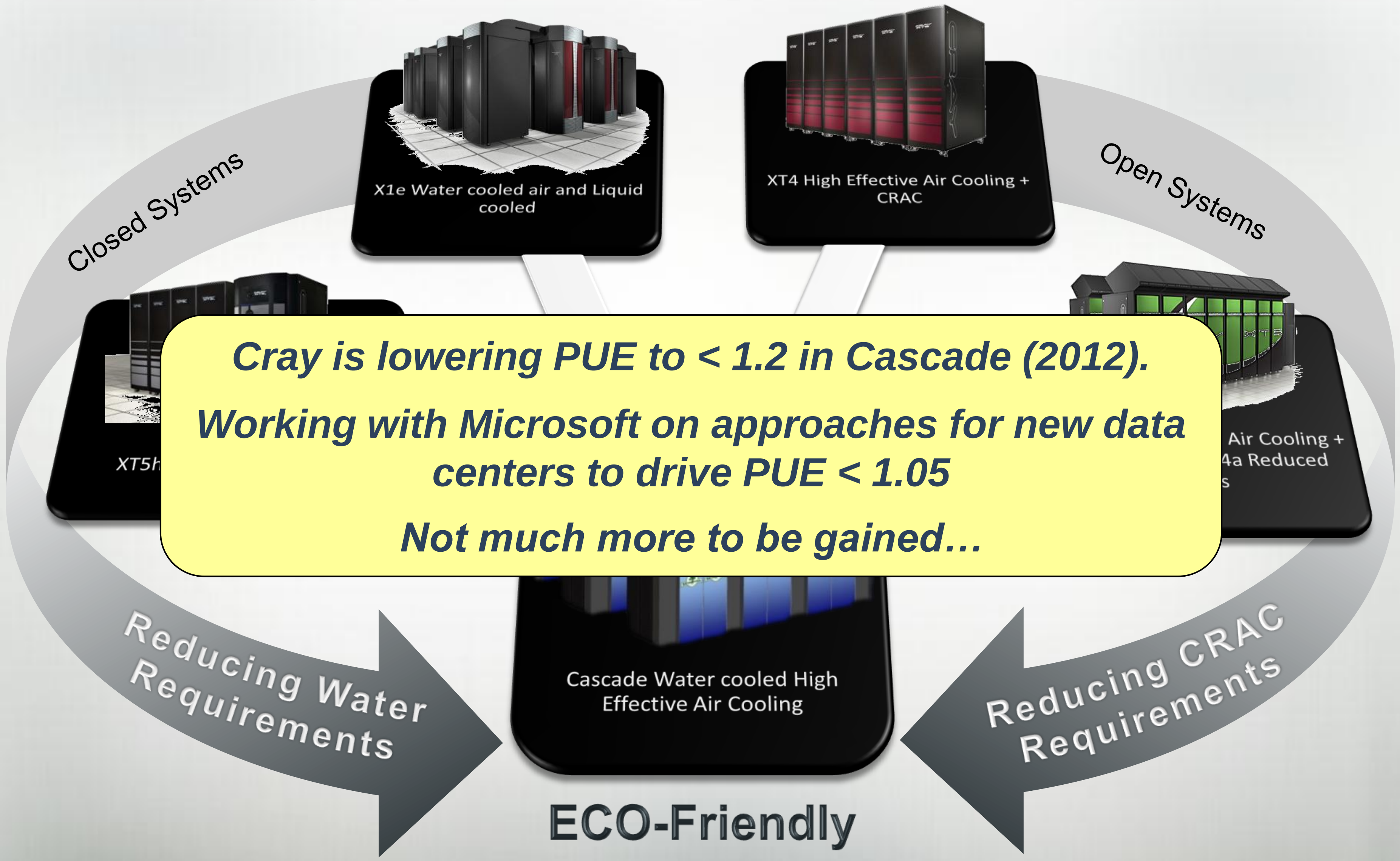
allinea

TOTALVIEW TECHNOLOGIES

CRAY LINUX ENVIRONMENT CLE3

Copyright 2010 Cray Inc. - 20th Anniversary of PDC

Next Generation Cooling Infrastructure



Closed Systems

X1e Water cooled air and Liquid cooled

XT4 High Effective Air Cooling + CRAC

Open Systems

XT5h

Air Cooling + 4a Reduced s

Cascade Water cooled High Effective Air Cooling

Reducing Water Requirements

Reducing CRAC Requirements

ECO-Friendly

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

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

Not much more to be gained…

Hardware Reliability?

- Yesterday's approach was to make the underlying system reliable through good engineering and redundancy
- This is simply not going to work at the Exascale
  - Component counts going way up
  - Underlying components getting less reliable
- Of course, we'll still use redundancy extensively
  - Processors, networks, memory, etc.
- Most important thing to focus on in hardware is avoiding silent data corruption