Titan Overview

Presented by:
Robert Whitten

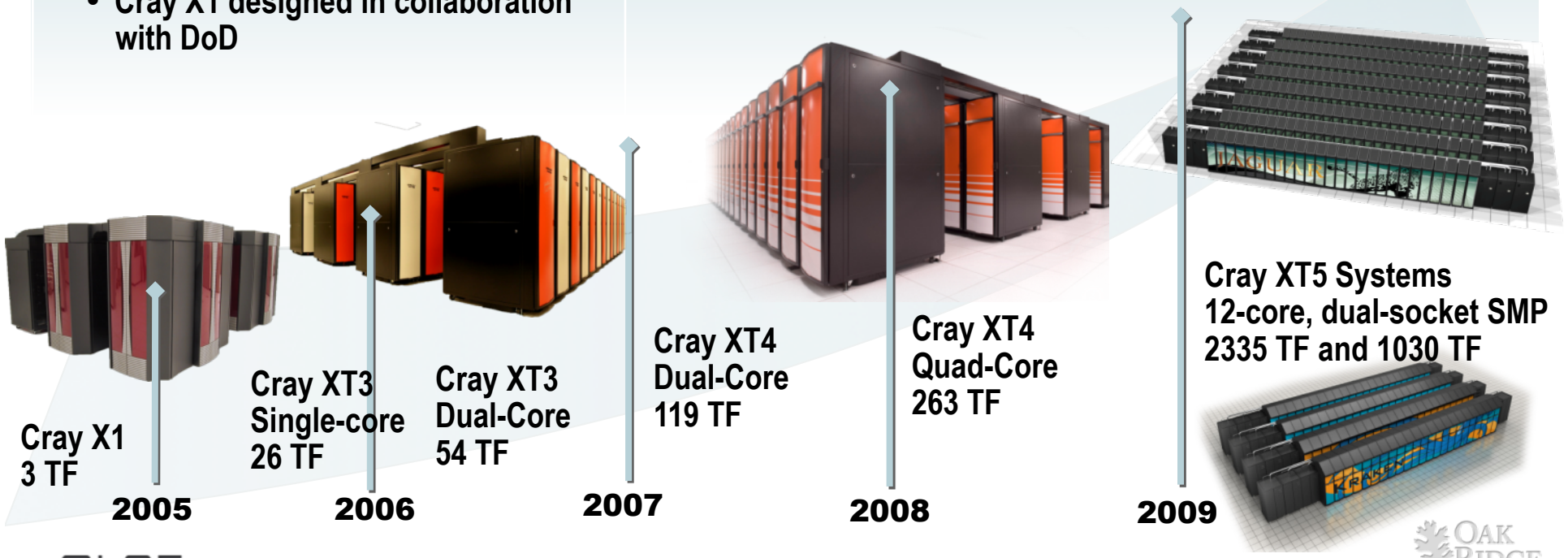
DISCLAIMER

This presentation is the OLCF's current vision for Titan. The strategy outlined in this presentation is awaiting final approval. There is no contract with the vendor at this point. All details are subject to change based on contractual and budget constraints.

We have increased system performance by 1,000 times since 2004

Hardware scaled from single-core through dual-core to quad-core and dual-socket , 12-core SMP nodes	Scaling applications and system software is the biggest challenge

• NNSA and DoD have funded much of the basic system architecture research
  • Cray XT based on Sandia Red Storm
  • IBM BG designed with Livermore
  • Cray X1 designed in collaboration with DoD
	• DOE SciDAC and NSF PetaApps programs are funding scalable application work, advancing many apps
	• DOE-SC and NSF have funded much of the library and applied math as well as tools
	• Computational Liaisons key to using deployed systems

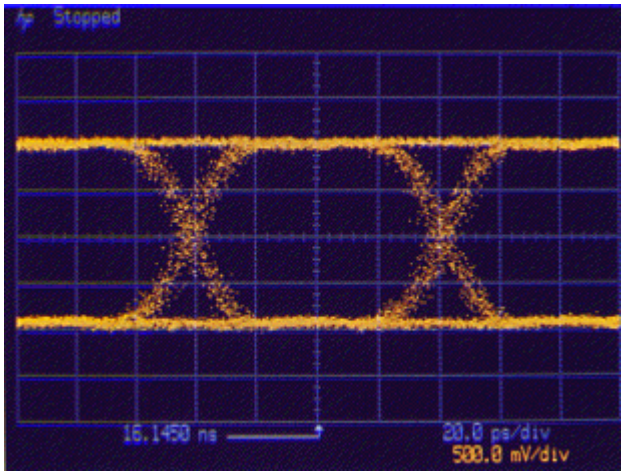


Disclaimer: No contract with vendor is in place

Why has clock rate scaling ended?

Power = Capacitance * Frequency * Voltage^2 + Leakage

- Traditionally, as Frequency increased, Voltage decreased, keeping the total power in a reasonable range
- But we have run into a wall on voltage
  - As the voltage gets smaller, the difference between a “one” and “zero” gets smaller. Lower voltages mean more errors.
  - While we like to think of electronics as digital devices, inside we use analog voltages to represent digital states.



- Capacitance increases with the complexity of the chip
- Total power dissipation is limited by cooling

Disclaimer: No contract with vendor is in place

Power to move data

Energy_to_move_data = bitrate * length^2 / cross_section_area_of_wire

- The energy consumed increases proportionally to the bit-rate, so as we move to ultra-high-bandwidth links, the power requirements will become an increasing concern.
- The energy consumption is highly distance-dependent (the square of the length term), so bandwidth is likely to become increasingly localized as power becomes a more difficult problem.
- Improvements in chip lithography (making smaller wires) will not improve the energy efficiency or data carrying capacity of electrical wires.

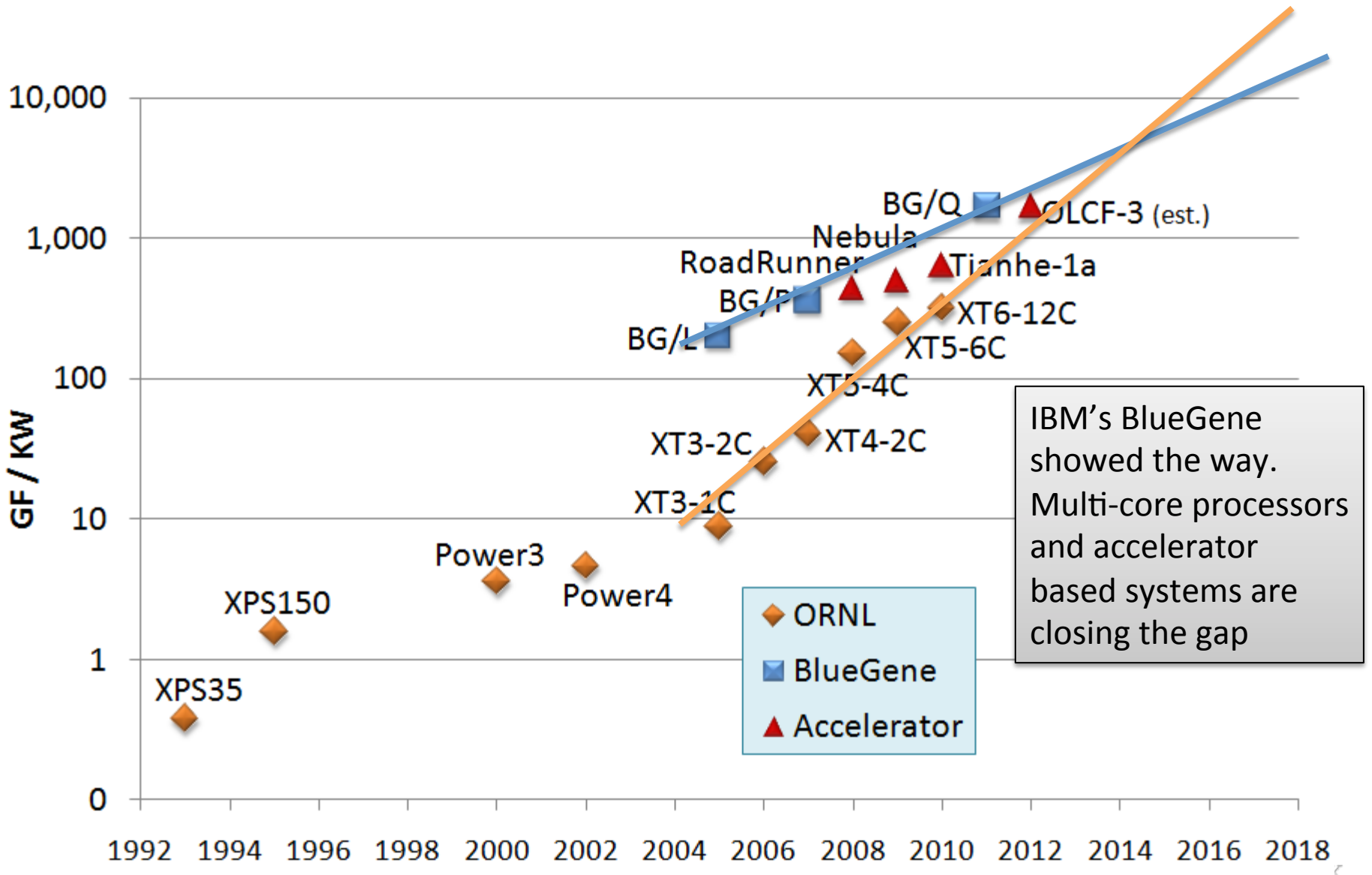
D. A. B. Miller and H. M. Ozaktas, "Limit to the Bit-Rate Capacity of Electrical Interconnects from the Aspect Ratio of the System Architecture," Journal of Parallel and Distributed Computing, vol. 41, pp. 42-52 (1997) article number PC961285.

Implications for future systems

- Clock rates will stay largely the same as today, increasing the parallelism of systems to improve performance
- Energy cost of moving data is very large. We will have to explicitly manage data locality to limit power consumption

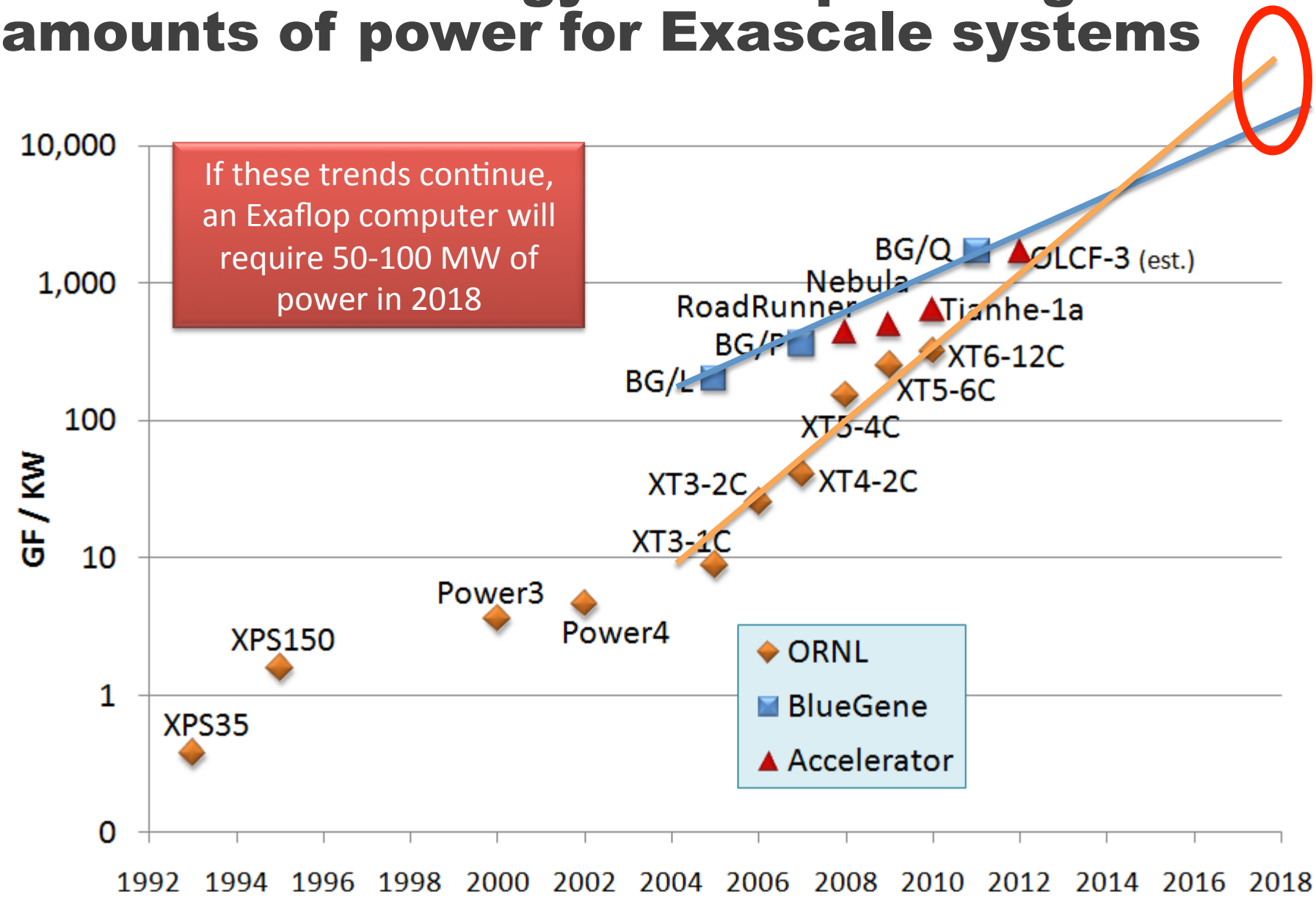
Disclaimer: No contract with vendor is in place

Trends in power efficiency



Disclaimer: No contract with vendor is in place

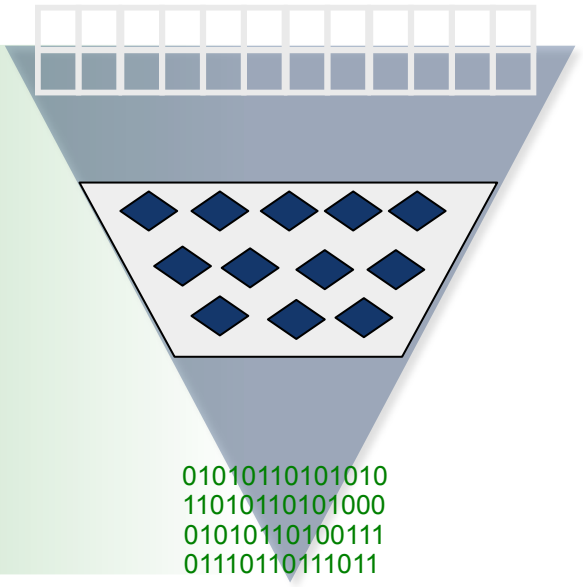
Current Technology will require huge amounts of power for Exascale systems



Disclaimer: No contract with vendor is in place

Hierarchical Parallelism

- MPI parallelism between nodes (or PGAS)
- On-node, SMP-like parallelism via threads (or subcommunicators, or…)
- Vector parallelism
  - SSE/AVX on CPUs
  - GPU threaded parallelism

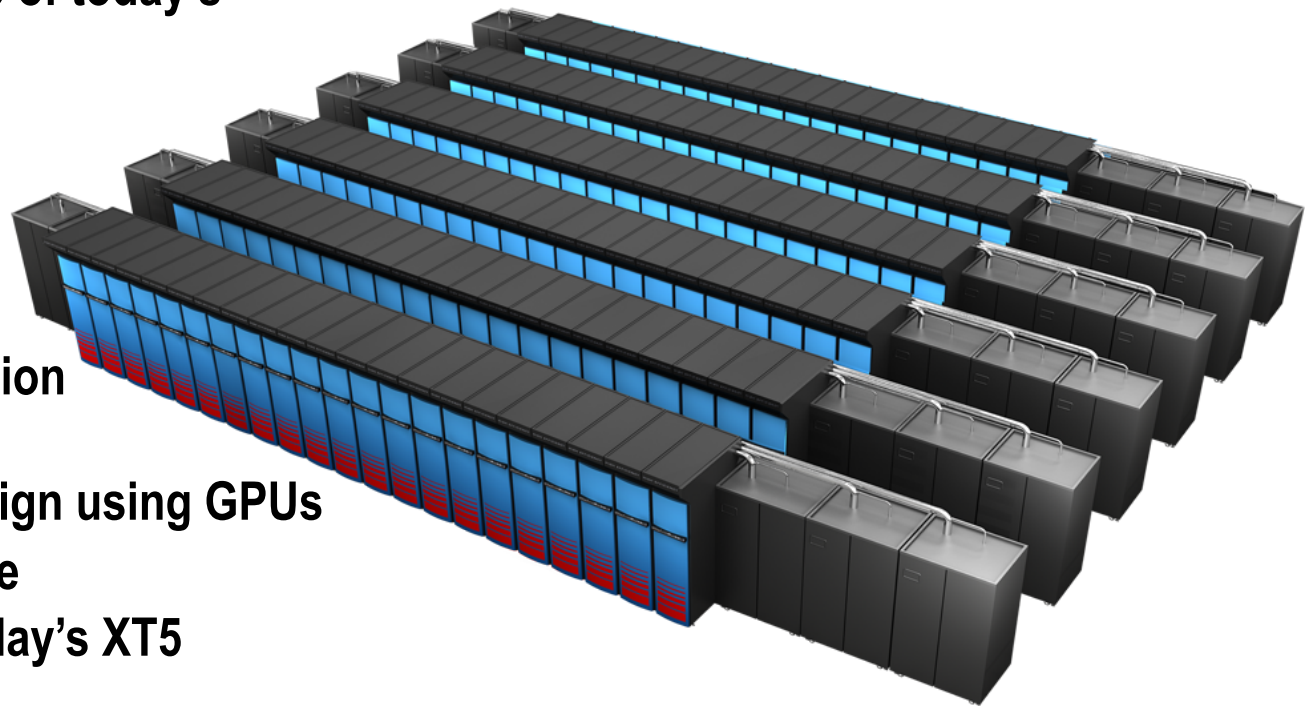


- **Exposure of unrealized parallelism is essential to exploit all near-future architectures.**
- **Uncovering unrealized parallelism and improving data locality improves the performance of even CPU-only code.**

Disclaimer: No contract with vendor is in place

ORNL's "Titan" 20 PF System Goals

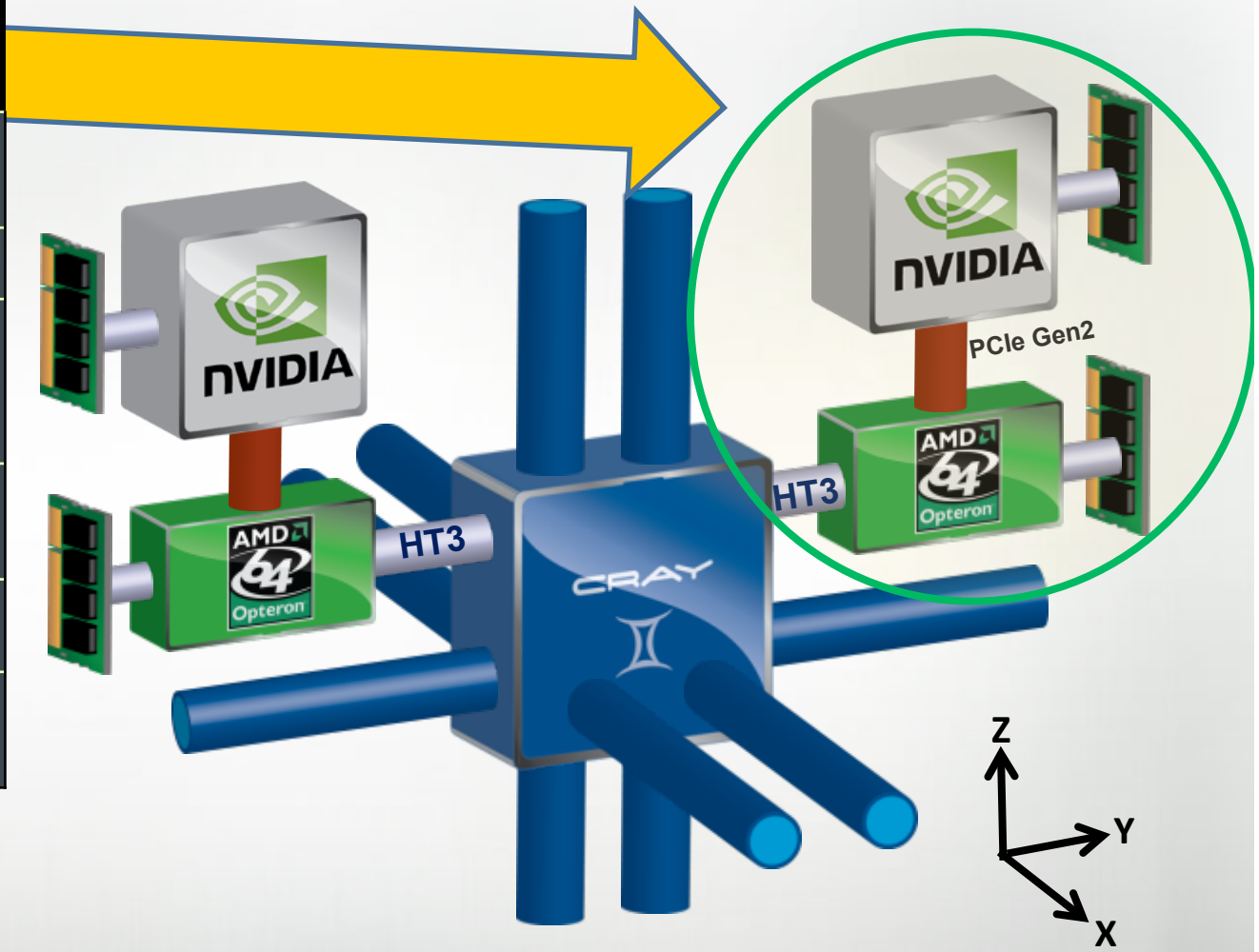
- Designed for science from the ground up
- Operating system upgrade of today's Linux Operating System
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- New accelerated node design using GPUs
- 10-20 PF peak performance
  - 9x performance of today's XT5
- Larger memory
- 3x larger and 4x faster file system



Disclaimer: No contract with vendor is in place

Cray XK6 Compute Node

XK6 Compute Node Characteristics
AMD Opteron 6200 Interlagos 16 core processor
Tesla X2090 @ 665 GF
Host memory 16 or 32GB 1600 MHz DDR3
Tesla X090 memory 6GB GDDR5 capacity
Gemini high speed Interconnect
Upgradeable to NVIDIA's Kepler many-core processor



Slide courtesy of Cray, Inc.

Disclaimer: No contract with vendor is in place

File System

- We will continue to use Lustre for our file system for Titan
- Plan to use Lustre version 2.x
  - Much more scalable metadata
  - Etc. etc. etc.
- Competitive procurement for the storage
  - Expect to get between 400 and 700 Gigabytes per second of bandwidth
  - Expect to add between 10 and 30 Petabytes of storage

Disclaimer: No contract with vendor is in place

Titan is a staged upgrade of jaguar

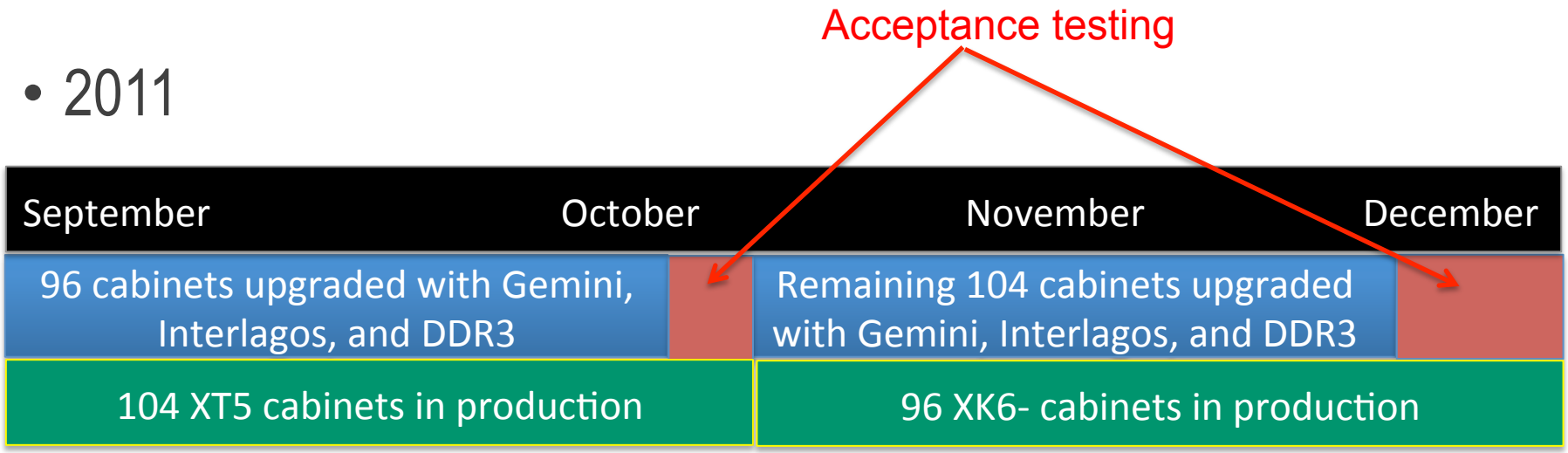
1st Upgrade (Sept. – Dec. 2011)	Final System

• Jaguar will be upgraded in 2 stages
• 1st stage
  • Roughly 3rd week of September through the end of October
  • 96 of jaguar's 200 cabinets will be removed from service and upgraded to XK6, minus the GPUs
• 2nd stage
  • November – December
  • 96 XK6- cabinets returned to service
  • Remaining 104 jaguar XT5 cabinets upgraded to XK6-
  • 2 "halves" joined and acceptance test run in late December – no user access during this time
  • ~10 cabinets will contain Tesla X2090 GPU
	• 2nd half of 2012
	• 2nd socket in each XK6 board populated with Kepler GPU
	• Precise timeline unknown at this time

Disclaimer: No contract with vendor is in place

Jaguar upgrade timeline

- 2011



- Jan 2012 – Full, upgraded, 200-cabinet jaguar system returned to service
- 2012 final upgrade to Titan
  - Roughly mirrors the 2011 initial upgrade path

Disclaimer: No contract with vendor is in place

Upgrade schedule

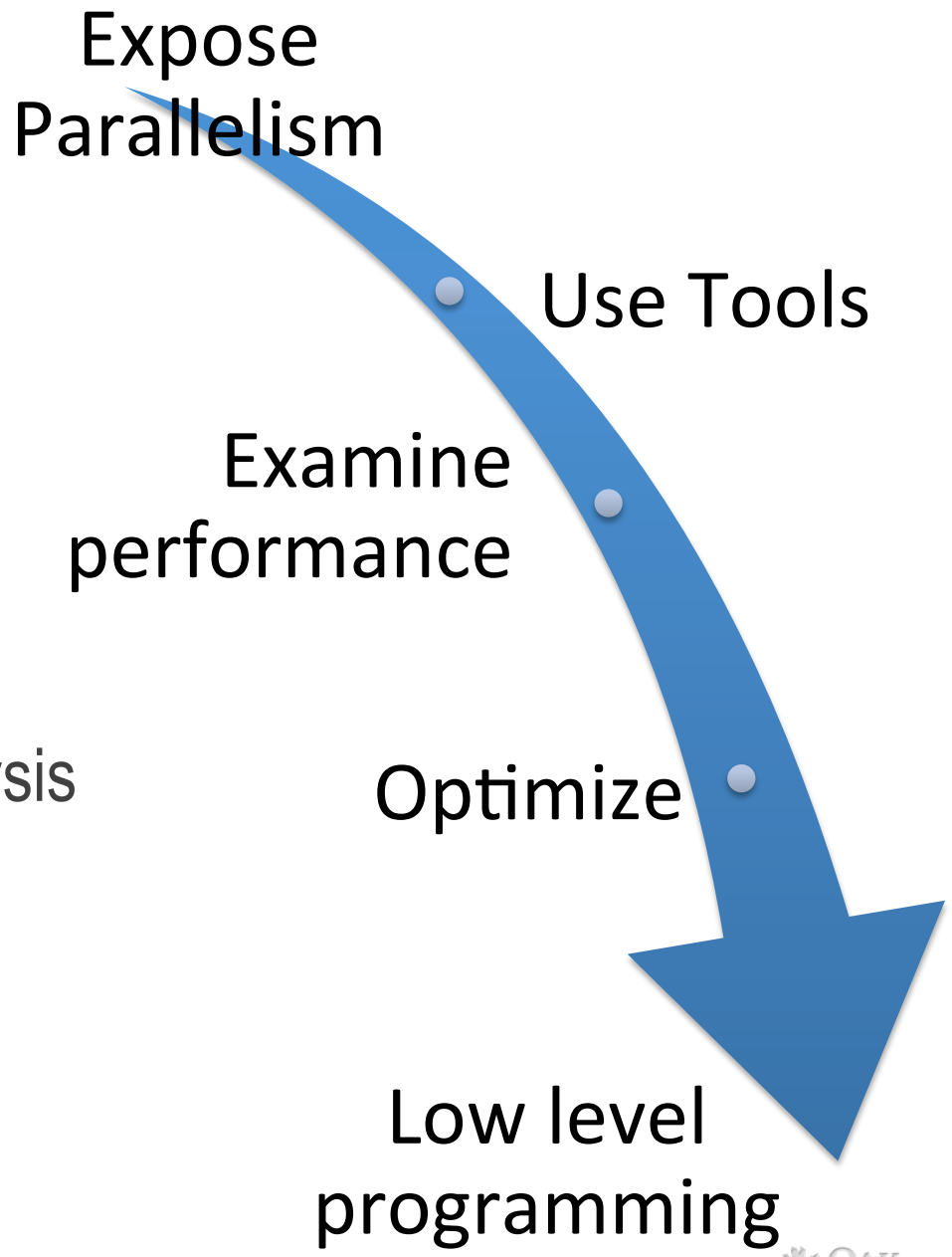
	Phase 0 Current	Phase 1 10/01/11	Phase 2 10/08/11	Phase 3 10/15/11	Phase 4 11/07/11	Phase 5 11/21/11	Phase 6 01/01/12	Phase 7 TBD	Phase 8 01/01/13
Name	Jaguar								Titan
Architecture	XT5	XT5	XT5	XT5	XK6	XK6	XK6	XK6	XK6
Processor	6-Core AMD	6-Core AMD	6-Core AMD	6-Core AMD	16-Core AMD	16-Core AMD	16-Core AMD	16-Core AMD	16-Core AMD
Cabinets	200	144	120	104	96	0	200	0	200
Nodes	18,688	13,824	11,520	9,984	9,216	0	18,688	0	18,688
Cores/node	12	12	12	12	16	0	16	0	16
Total cores	224,256	162,240	134,592	117,120	142,848	0	299,008	0	299,008
Memory/node	16GB	16GB	16GB	16GB	32GB	0	32GB	0	32GB
Memory/core	1.3GB	1.3GB	1.3GB	1.3GB	2GB	0	2GB	0	2GB
Interconnect	SeaStar2+	SeaStar2+	SeaStar2+	SeaStar2+	Gemini	0	Gemini	0	Gemini
GPUs	0	0	0	0	960	0	960	0	18,688

unavailable

unavailable

Application Strategy

- Expose parallelism
  - Case studies
  - Lessons learned
- Compiler / Directive-based
  - Cuda Fortran
  - HMPP
- Debugging / Performance Analysis
- Optimization
- Accelerator programming
  - Cuda
  - OpenCL



Disclaimer: No contract with vendor is in place

Planned Upcoming Events

- Titan Summit - August 15-17, 2011
- CAPS Workshop - September 19-23, 2011
- Fall Training - October 2011
- Titan Workshop - December 2011
- Spring Training - March 2011

Disclaimer: No contract with vendor is in place

In the works - TBD

- Cray Workshop
  - Cray compiler
  - Craypat
- Cuda Workshop
- Vampir Workshop
- Hybrid-programming Workshop
- DDT Workshop
- Totalview Workshop
- Tau Workshhop
- Some will be webcast-only events

Disclaimer: No contract with vendor is in place

Questions?

http://www.olcf.ornl.gov

Disclaimer: No contract with vendor is in place