Introduction to the Oak Ridge Leadership Computing Facility for CSGF Fellows

Bronson Messer

Acting Group Leader
Scientific Computing Group
National Center for Computational Sciences

Theoretical Astrophysics Group
Oak Ridge National Laboratory

Department of Physics & Astronomy
University of Tennessee

Outline

- The OLCF: history, organization, and what we do
- The upgrade to Titan
  - Interlagos processors with GPUs
  - Gemini Interconnect
  - Software, etc.
- The CSGF Director's Discretionary Program
- Questions and Discussion

ORNL has a long history in High Performance Computing

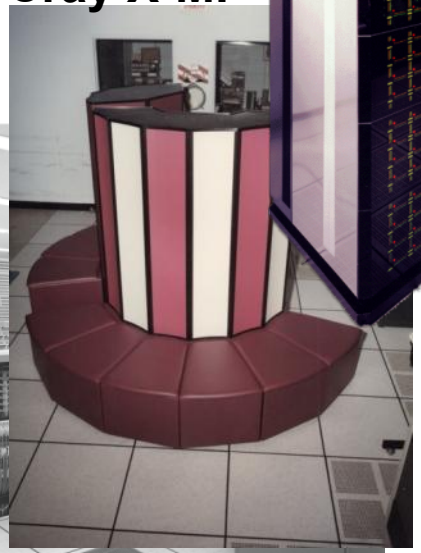
ORNL has had 20 systems on the TOP500® SUPERCOMPUTER SITES lists

2007
IBM Blue Gene/P

1996-2002
IBM Power 2/3/4

1992-1995
Intel Paragons

1985
Cray X-MP

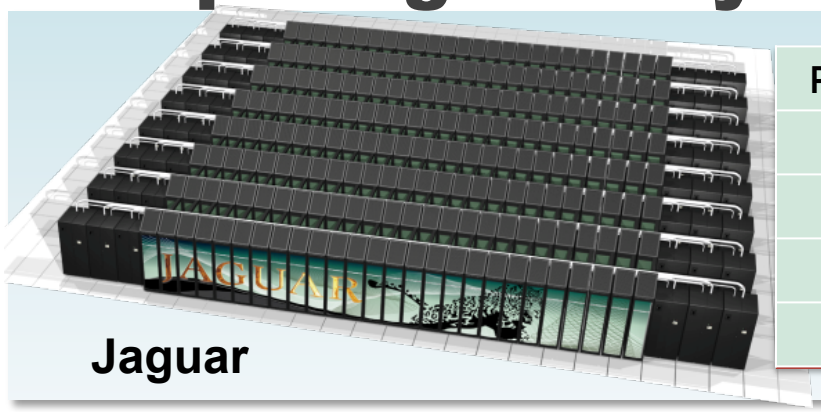


1969
IBM 360/9

1954
ORACLE

2003-2005
Cray X1/X1E

Today, we have the world's most powerful computing facility



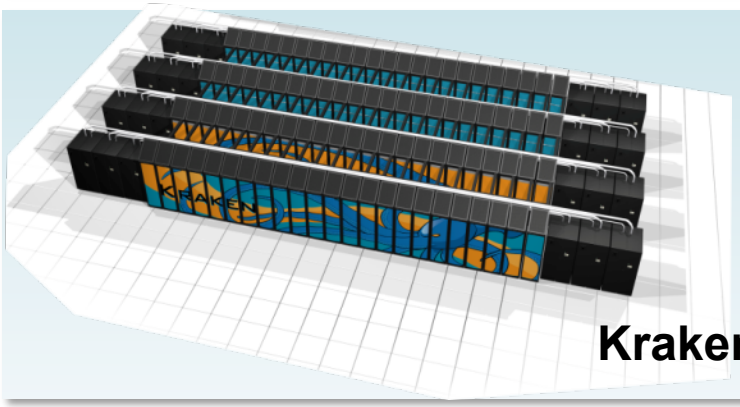
Jaguar

Peak performance	2.33 PF/s
Memory	300 TB
Disk bandwidth	> 240 GB/s
Square feet	5,000
Power	7 MW

TOP500 SUPERCOMPUTER SITES

#2

Dept. of Energy's most powerful computer



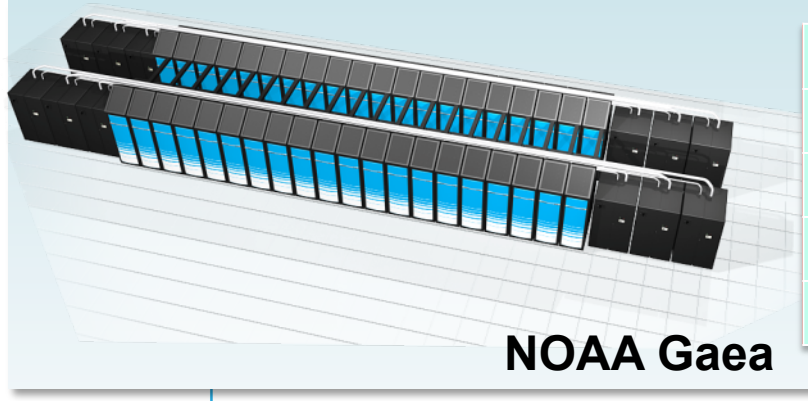
Kraken

Peak performance	1.03 PF/s
Memory	132 TB
Disk bandwidth	> 50 GB/s
Square feet	2,300
Power	3 MW

TOP500 SUPERCOMPUTER SITES

#8

National Science Foundation's most powerful computer



NOAA Gaea

Peak Performance	1.1 PF/s
Memory	248 TB
Disk Bandwidth	104 GB/s
Square feet	1,600
Power	2.2 MW

TOP500 SUPERCOMPUTER SITES

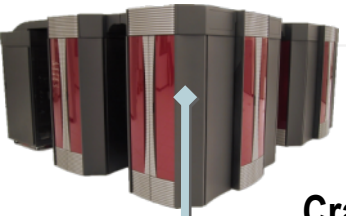
#32

National Oceanic and Atmospheric Administration's most powerful computer

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

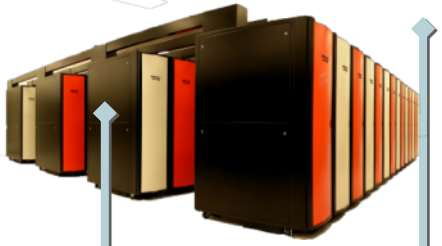
Hardware scaled from single-core through dual-core to quad-core and dual-socket , 12-core SMP nodes	Scaling applications and system software is the biggest challenge

• 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



Cray X1
3 TF

2005



Cray XT3
Single-core
26 TF

2006

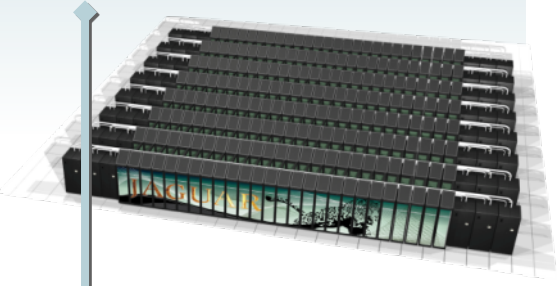
Cray XT3
Dual-Core
54 TF

2007

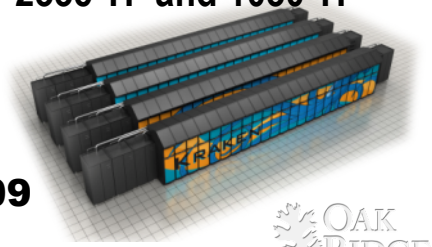
Cray XT4
Dual-Core
119 TF

Cray XT4
Quad-Core
263 TF

2008



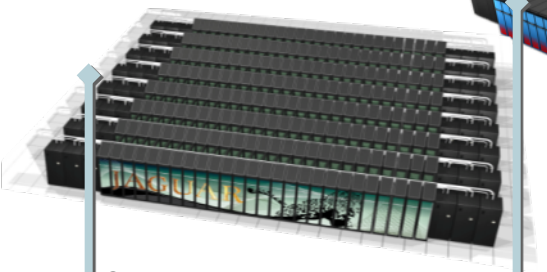
Cray XT5 Systems
12-core, dual-socket SMP
2335 TF and 1030 TF



2009

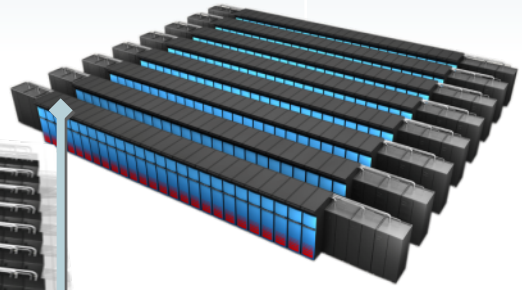
Our science requires that we advance computational capability 1000x over the next decade

Mission: Deploy and operate the computational resources required to tackle global challenges	Vision: Maximize scientific productivity and progress on the largest scale computational problems
• Deliver transforming discoveries in climate, materials, biology, energy technologies, etc. • Ability to investigate otherwise inaccessible systems, from regional climate impacts to energy grid dynamics	• Providing world-class computational resources and specialized services for the most computationally intensive problems • Providing stable hardware/software path of increasing scale to maximize productive applications development



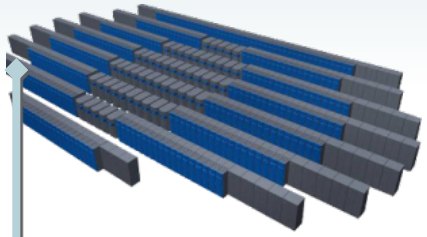
Cray XT5 2+ PF
Leadership system for science

2009



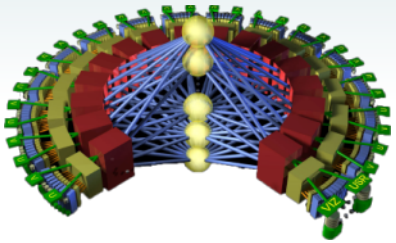
OLCF-3: 10-20 PF
Leadership system with some HPCS technology

2012



OLCF-4: 100-250 PF
based on DARPA HPCS technology

2015



OLCF-5: 1 EF

2018

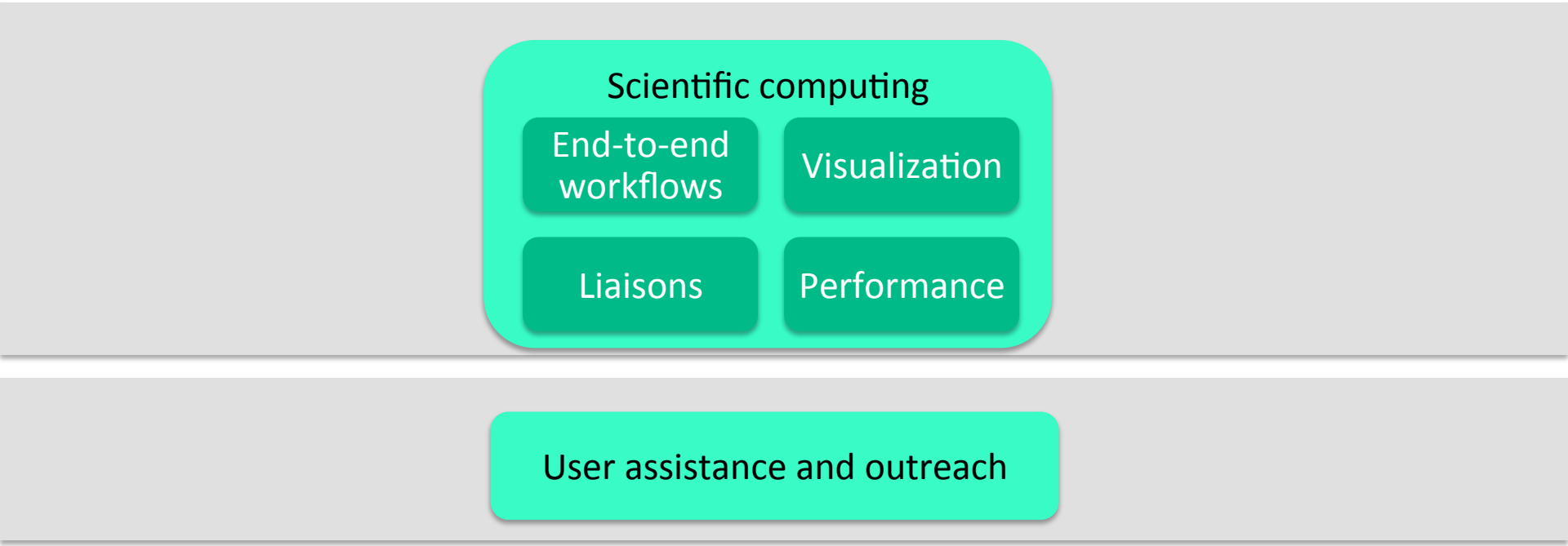
The Oak Ridge Leadership Computing Facility (OLCF)

- One of 2 DOE SC leadership computing facilities
  - Designed to support capability computing

- The National Center for Computational Sciences (NCCS) is the division at ORNL that contains the OLCF
- Four main groups in NCCS
  - HPC Operations
  - Technology Integration
  - User Assistance and Outreach
  - Scientific Computing

Scientific support model

- "Two-Line" support model



Scientific support

- User Assistance groups provide “front-line” support for day-to-day computing issues
- SciComp Liaisons provide advanced algorithmic and implementation assistance
- Assistance in data analytics and workflow management, visualization, and performance engineering are also provided for each project

Enabling capability-class research: Queuing policies

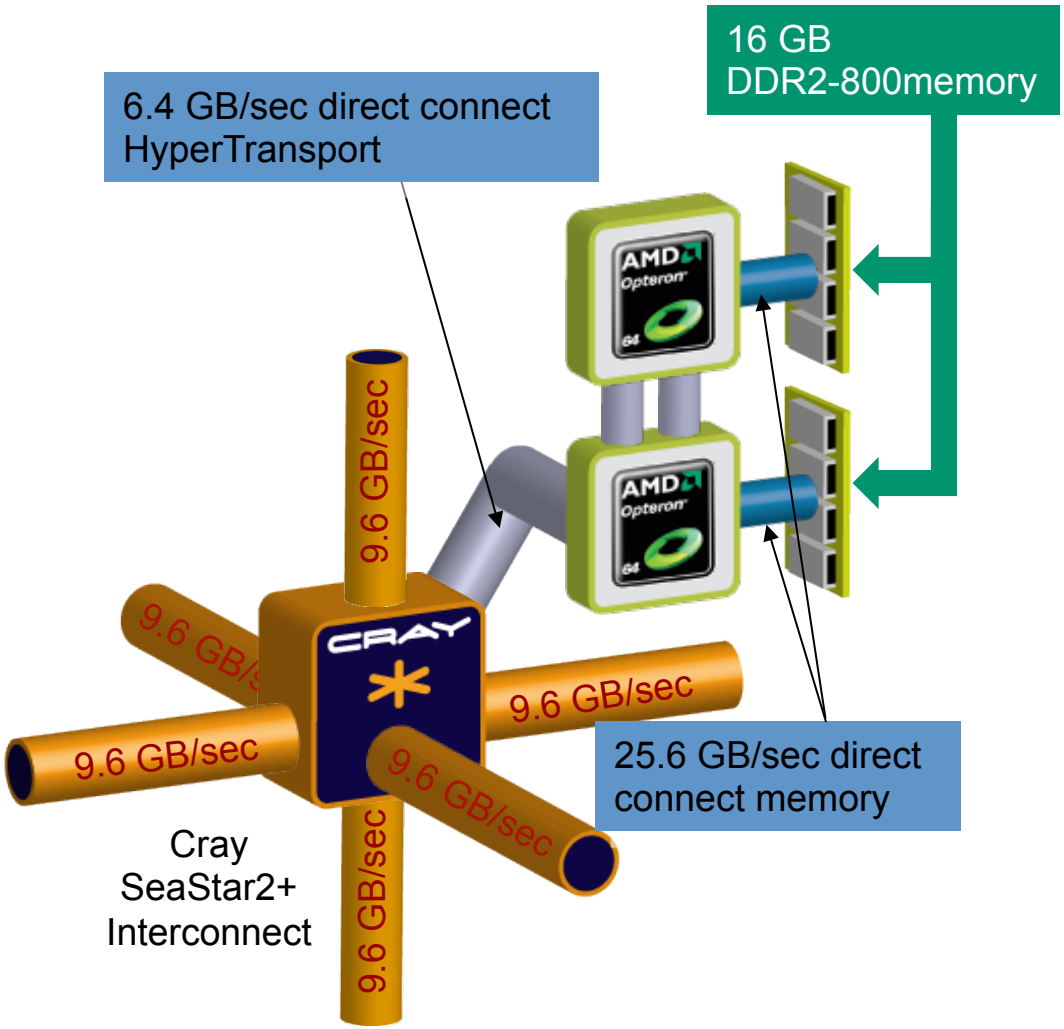
- Default queuing policies are designed to enable capability jobs
  - Capacity jobs are allowed only short runtimes and have lower priority
  - This mode is important for achieving the INCITE goal of enabling breakthroughs even though may lead to lower overall system utilization
- Queue bin edges, run times, and priorities are frequently updated based on usage patterns
  - Resource manager simulators are used to evaluate the impact of possible queue policy changes

ORNL's "Titan" System

- Upgrade of Jaguar from Cray XT5 to XK6
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6274 processors (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2012: 960 NVIDIA x2090 "Fermi" GPUs
  - 2013: 14,592 NVIDIA "Kepler" GPUs
- 20+ PFlops peak system performance
- 600 TB DDR3 mem. + 88 TB GDDR5 mem

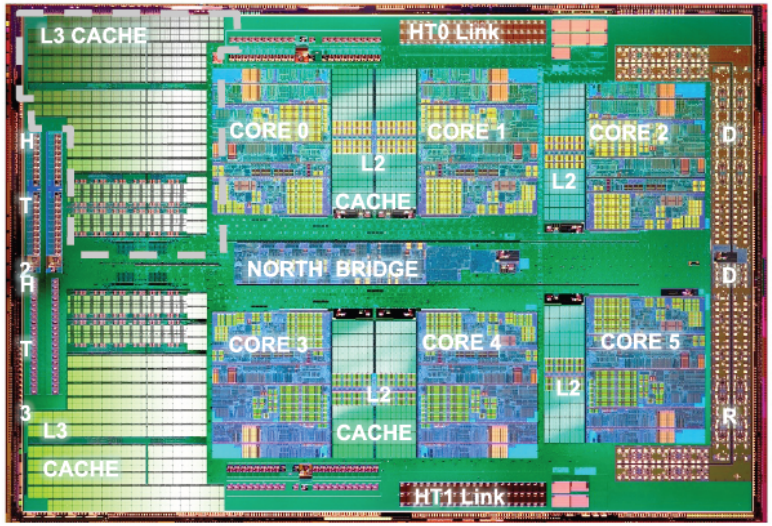
Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
# of Fermi chips (2012)	960
# of NVIDIA "Kepler" (2013)	14,592
Total System Memory	688 TB
Total System Peak Performance	20+ Petaflops
Cross Section Bandwidths	X=14.4 TB/s Y=11.3 TB/s Z=24.0 TB/s

Cray XT5 Compute Node



Cray XT5 Node Characteristics	
Number of Cores	12
Peak Performance	125 Gflops/sec
Memory Size	16 GB per node
Memory Bandwidth	25.6 GB/sec

AMD Opteron 2435 (Istanbul) processors

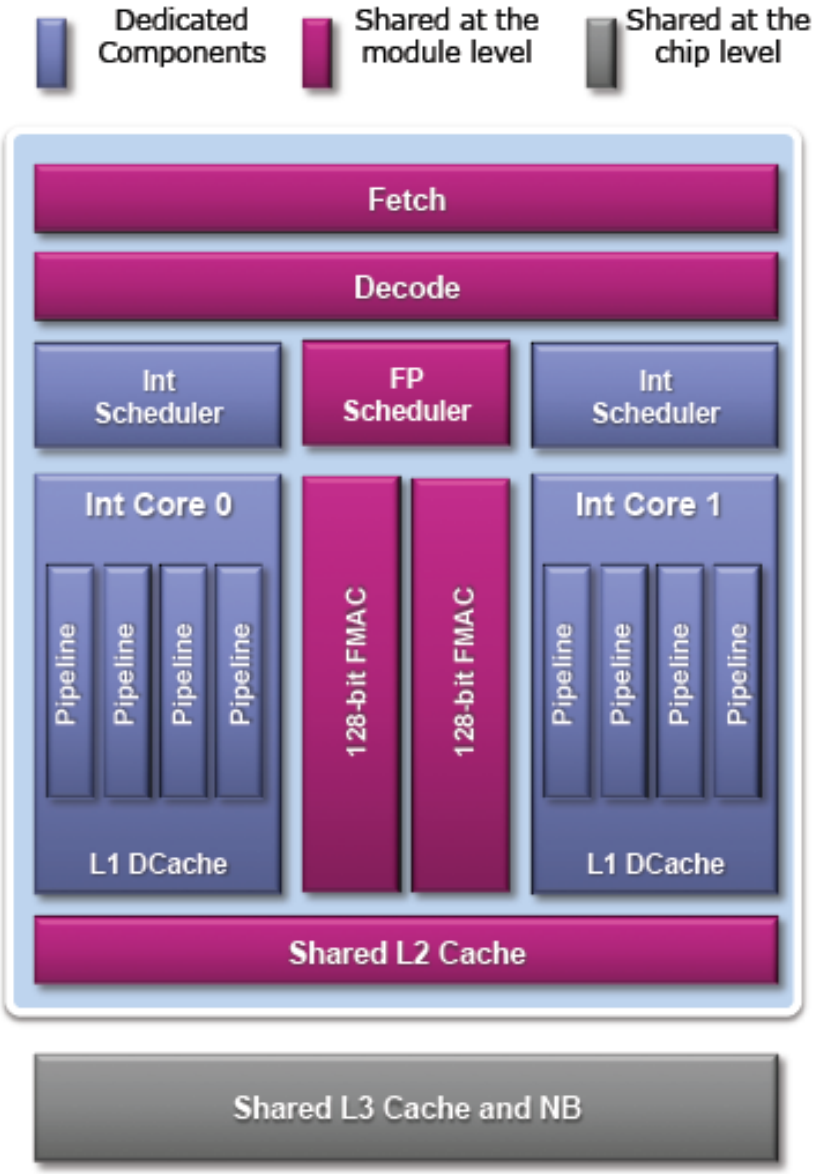


Cray XK6 Compute Node

XK6 Compute Node Characteristics
AMD Opteron 6200 “Interlagos” 16 core processor @ 2.2GHz
Tesla M2090 “Fermi” @ 665 GF with 6GB GDDR5 memory
Host Memory 32GB 1600 MHz DDR3
Gemini High Speed Interconnect
Upgradeable to NVIDIA’s next generation “Kepler” processor in 2012
Four compute nodes per XK6 blade. 24 blades per rack

Interlagos Processor Architecture

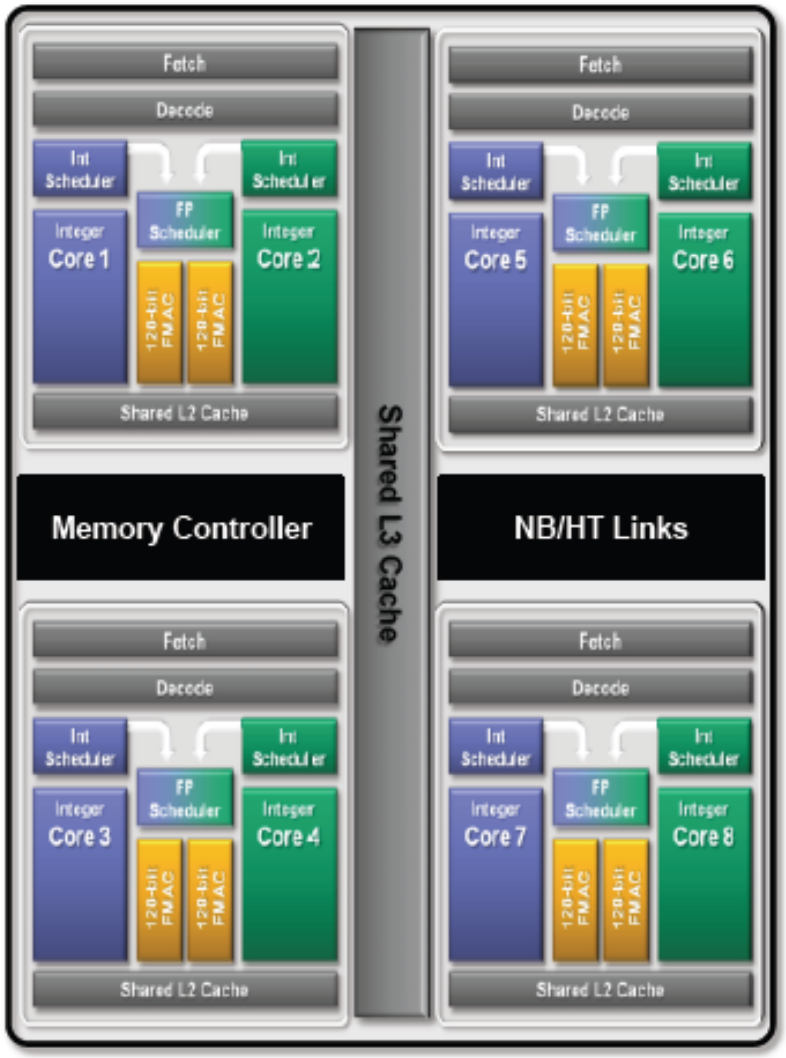
- Interlagos is composed of a number of "Bulldozer modules" or "Compute Unit"
  - A compute unit has shared and dedicated components
    - There are two independent integer units; shared L2 cache, instruction fetch, Icache; and a *shared*, 256-bit Floating Point resource
  - A single Integer unit can make use of the entire Floating Point resource with 256-bit AVX instructions
    - Vector Length
      - 32 bit operands, VL = 8
      - 64 bit operands, VL = 4



Slide courtesy of J. Levesque, Cray

Building an Interlagos Processor

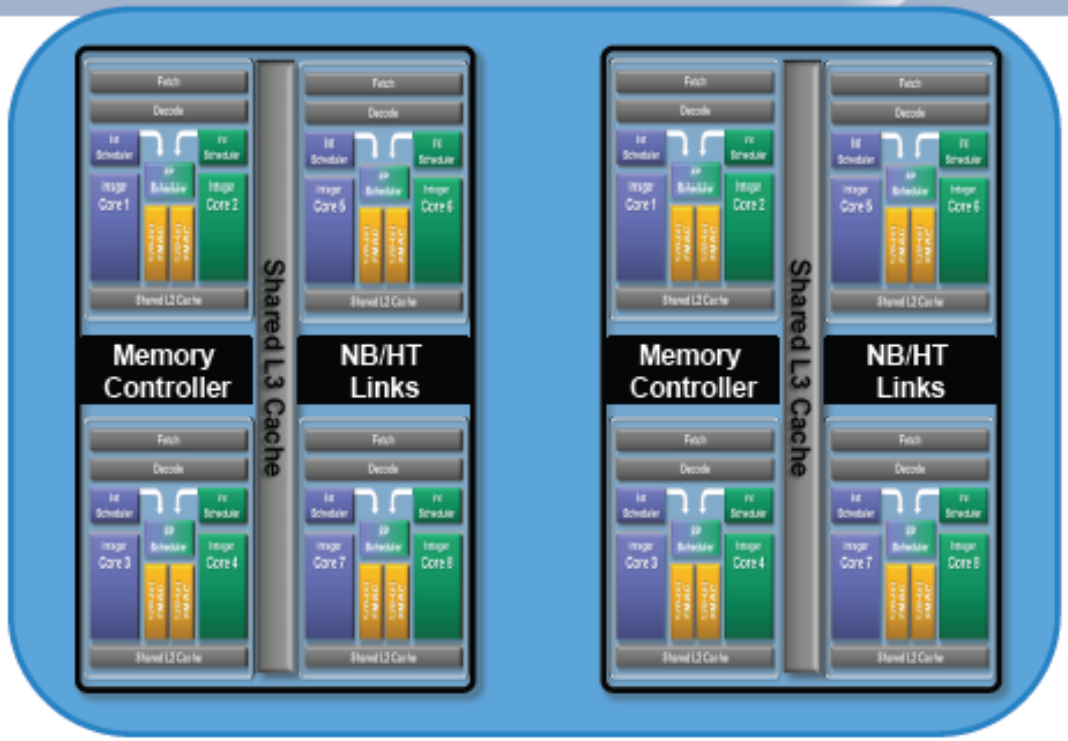
- Each processor die is composed of 4 compute units
  - The 4 compute units share a memory controller and 8MB L3 data cache
    - Each processor die is configured with two DDR3 memory channels and multiple HT3 links

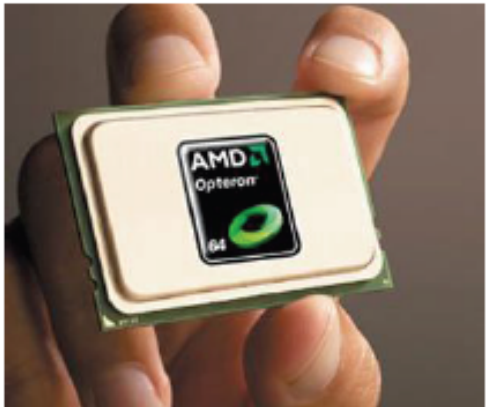


Slide courtesy of J. Levesque, Cray

Interlagos Processor

- Two die are packaged on a multi-chip module to form an Interlagos processor
  - Processor socket is called G34 and is compatible with Magny Cours
  - Package contains
    - 8 compute units
    - 16 MB L3 Cache
    - 4 DDR3 1333 or 1600 memory channels

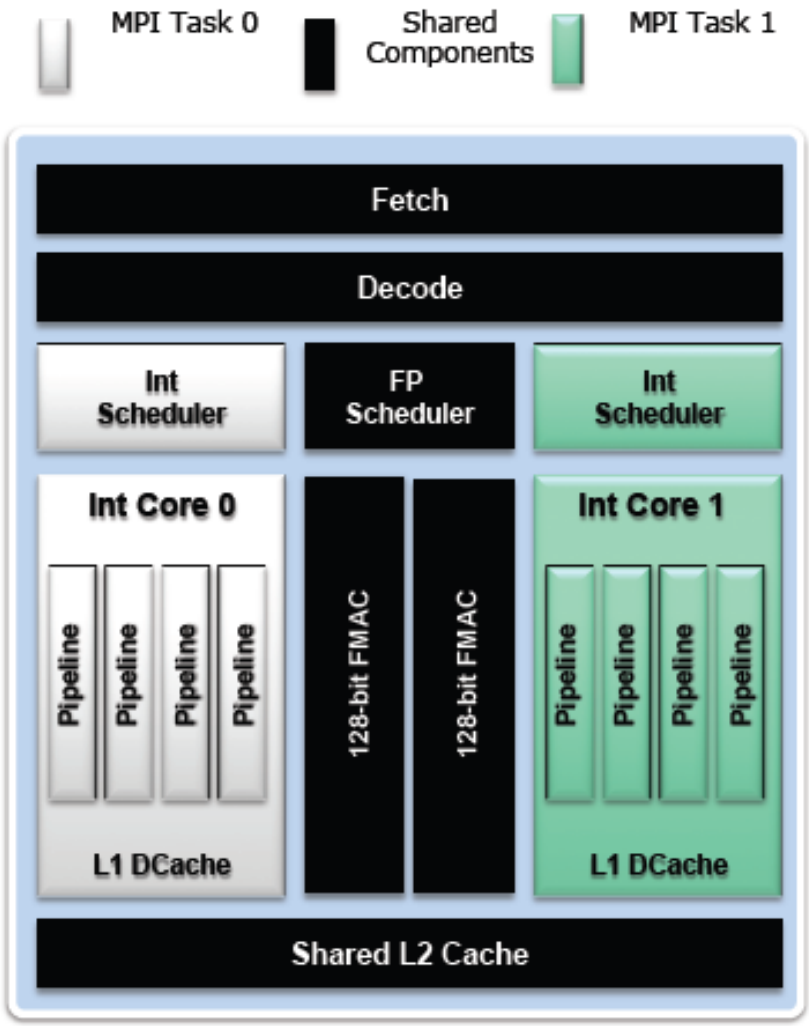




Slide courtesy of J. Levesque, Cray

Two MPI Tasks on a Compute Unit ("Dual-Stream Mode")

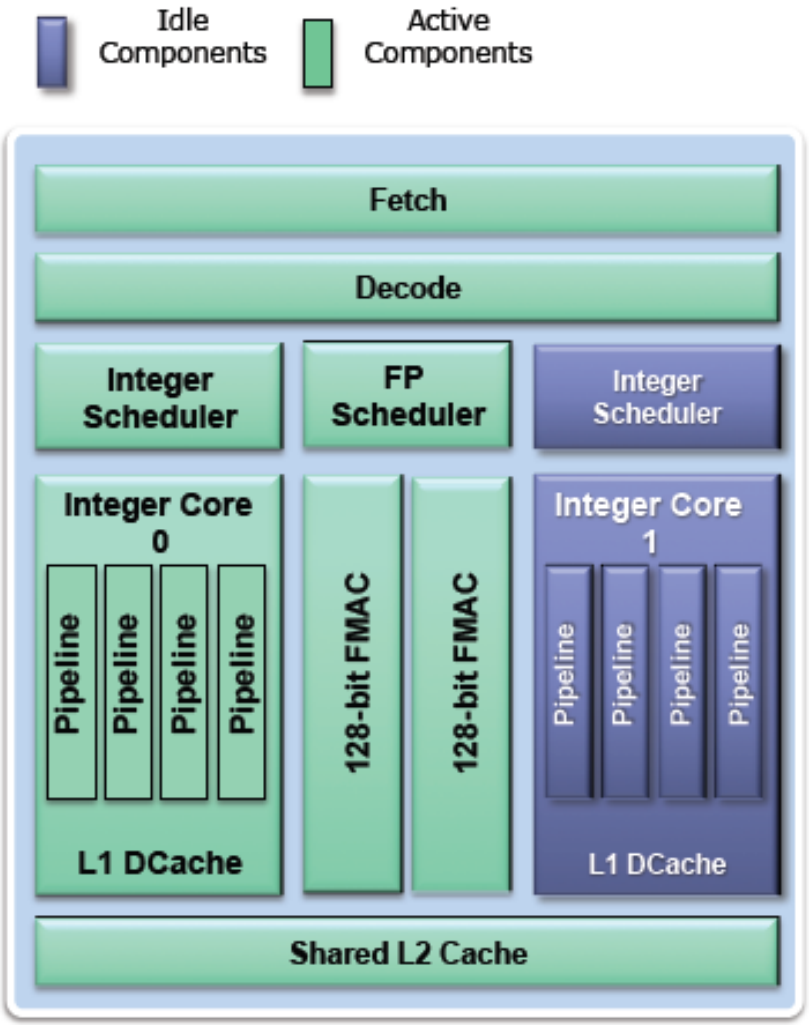
- An MPI task is pinned to each integer unit
  - Each integer unit has exclusive access to an integer scheduler, integer pipelines and L1 Dcache
  - The 256-bit FP unit, instruction fetch, and the L2 Cache are shared between the two integer units
    - 256-bit AVX instructions are dynamically executed as two 128-bit instructions if the 2nd FP unit is busy
- When to use
  - Code is highly scalable to a large number of MPI ranks
  - Code can run with a 2GB per task memory footprint
  - Code is not well vectorized



Slide courtesy of J. Levesque, Cray

One MPI Task on a Compute Unit ("Single Stream Mode")

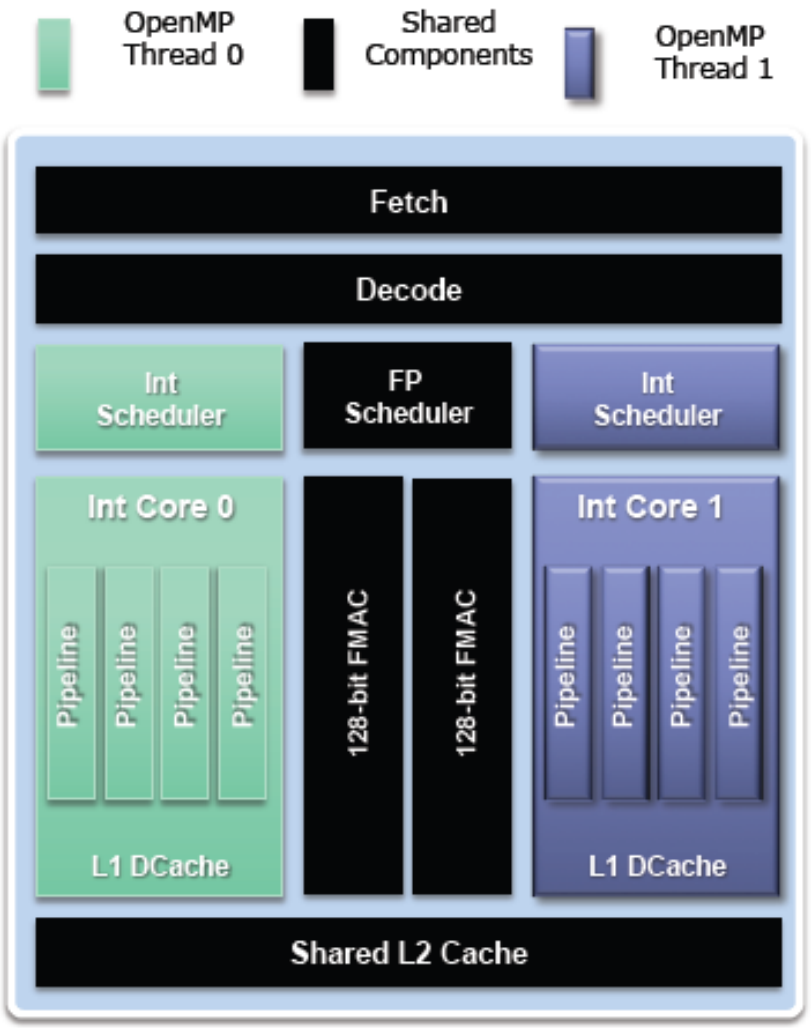
- Only one integer unit is used per compute unit
  - This unit has exclusive access to the 256-bit FP unit and is capable of 8 FP results per clock cycle
  - The unit has twice the memory capacity and memory bandwidth in this mode
  - The L2 cache is effectively twice as large
  - The peak of the chip is not reduced
- When to use
  - Code is highly vectorized and makes use of AVX instructions
  - Code benefits from higher per task memory size and bandwidth



Slide courtesy of J. Levesque, Cray

One MPI Task per compute unit with Two OpenMP Threads ("Dual-Stream Mode")

- An MPI task is pinned to a compute unit
- OpenMP is used to run a thread on each integer unit
  - Each OpenMP thread has exclusive access to an integer scheduler, integer pipelines and L1 Dcache
  - The 256-bit FP unit and the L2 Cache is shared between the two threads
  - 256-bit AVX instructions are dynamically executed as two 128-bit instructions if the 2nd FP unit is busy
- When to use
  - Code needs a large amount of memory per MPI rank
  - Code has OpenMP parallelism at each MPI rank



Slide courtesy of J. Levesque, Cray

Cray 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
- Scalability to 1M+ cores

Aries

- Ask me about it

Slide courtesy of J. Levesque, Cray

Two Phase Upgrade Process

- Phase 1: XT5 to XK6 without GPUs
  - Remove all XT5 nodes and replace with XK6 and XIO nodes
  - 16-core processors, 32 GB/node, Gemini
  - 960 nodes (10 cabinets) have NVIDIA Fermi GPUs
  - Users ran on half of system while other half was upgraded
- Add NVIDIA Kepler GPUs
  - Cabinet Mechanical and Electrical upgrades
    - New air plenum bolts on to cabinet to support air flow needed by GPUs
    - Larger fan
    - Additional power supply
    - New doors ☺
  - Rolling upgrade of node boards
    - Pull board, add 4 Kepler GPUs modules, replace board, test, repeat 3,647 times!
    - Keep most of the system available for users during upgrade
  - Acceptance test of system

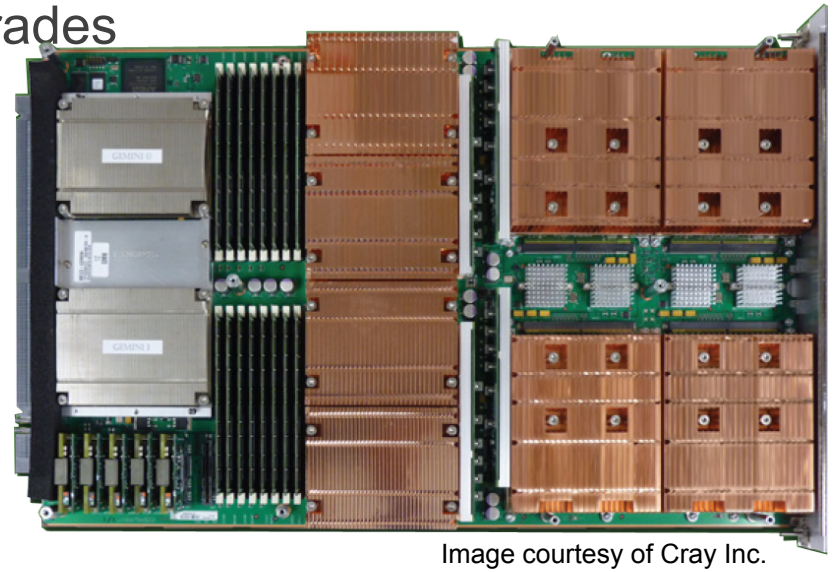
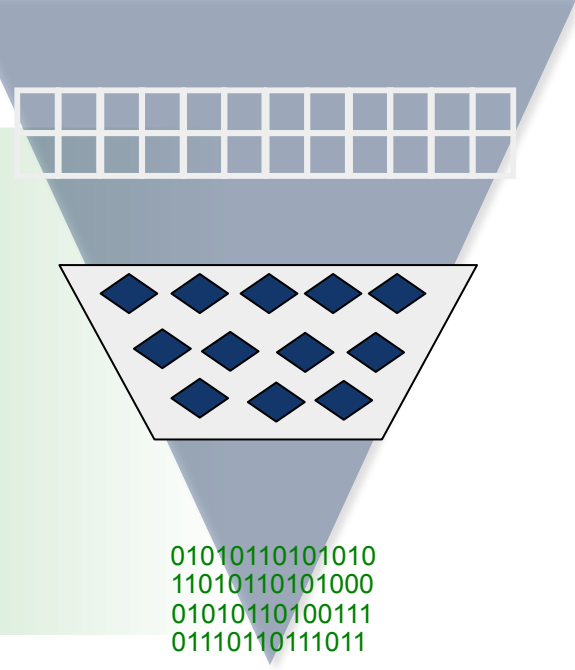


Image courtesy of Cray Inc.

Hierarchical Parallelism

- MPI parallelism between nodes (or PGAS)
- On-node, SMP-like parallelism via threads (or subcommunicators, or…)
- Vector parallelism
  - SSE/AVX/etc 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.**

How do you program these nodes?

- **Compilers**
  - OpenACC is a set of compiler directives that allows the user to express hierarchical parallelism in the source code so that the compiler can generate parallel code for the target platform, be it GPU, MIC, or vector SIMD on CPU
  - Cray compiler supports XK6 nodes and is OpenACC compatible
  - CAPS HMPP compiler supports C, C++ and Fortran compilation for heterogeneous nodes and is adding OpenACC support
  - PGI compiler supports OpenACC and CUDA Fortran
- **Tools**
  - Allinea DDT debugger scales to full system size and with ORNL support will be able to debug heterogeneous (x86/GPU) apps
  - ORNL has worked with the Vampir team at TUD to add support for profiling codes on heterogeneous nodes
  - CrayPAT and Cray Apprentice support XK6 programming

Titan Tool Suite

Compilers	Performance Tools	GPU Libraries	Debuggers	Source Code
Cray PGI CAP-HMPP Pathscale NVIDIA CUDA GNU Intel	CrayPAT Apprentice Vampir VampirTrace TAU HPCToolkit CUDA Profiler	MAGMA CULA Trillinos libSCI	DDT NVIDIA Gdb	HMPP Wizard

Director's Discretionary Program

- 10% of all the time available at the OLCF is allocated via the Director's Discretionary (DD) Program.
- Proposals to the DD program are meant to be short and should delineate a particular computational campaign designed to improve scaling or answer a particular need.
- Typical proposals/awards
  - INCITE scaling runs
  - "Strategic" applications
  - Time-sensitive runs (e.g. natural disasters)
- **In 2011 we instituted a special track for DD proposals from CSGF fellows**

CSGF DD Program

- CSGF fellows are encouraged to contact the CSGF DD coordinator(s) before they submit their proposal
  - Judy Hill (hilljc-at-o-r-n-l-dot-gov) and me (Bronson Messer, bronson-at-o-r-n-l-dot-gov)
- The coordinators will work with the fellow to make sure their proposal has the best possible chance of being awarded an allocation
- This usually doesn't entail changing the intent of the proposal, or even the scope. Rather, it's often just a matter of better estimates for runtime, a better explanation of what the purpose is (i.e. not an "expert" explanation of why the science is important), or other small tweaks.
- After the fellow and the coordinator(s) have converged, the proposal is submitted to the Resource Allocation Council at OLCF.

Questions?

The research and activities described in this presentation were performed using the resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.