CRAY XE6

Cray built the world's first production petaflops system with the Cray XT5™ supercomputer in 2009. Now with the Cray XE6™ system, Cray is raising the expectations for high performance computing once again.

Redefining Supercomputing

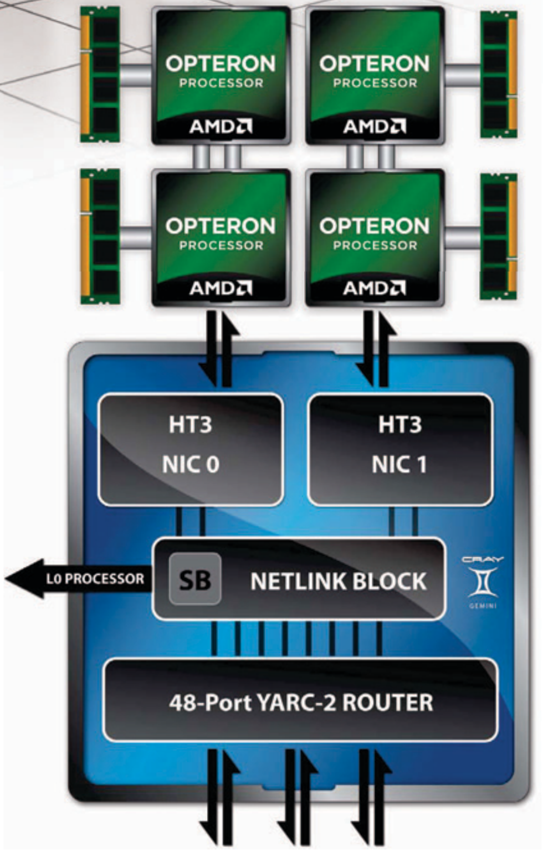
The Cray XE6 supercomputer takes the proven Cray XT5 infrastructure and incorporates it with two innovative new technologies: AMD's powerful multi-core processors and the revolutionary Gemini™ interconnect. The result is a system that brings production petascale to a wider HPC community and fundamentally changes how Cray systems communicate. Designed to scale to over 1 million processor cores, every aspect of the Cray XE6 supercomputer – from its industry-leading resiliency features to its host of scalability-boosting technologies – has been engineered to meet science's ever-toughening demands for scalability, reliability and flexibility.

Scalable Performance

Gemini Scalable Interconnect

The Gemini interconnect is the heart of the Cray XE6 system. Capable of tens of millions of MPI messages per second, the Gemini ASIC is designed to complement current and future massively multi-core processors. Each dual-socket node is interfaced to the Gemini interconnect through HyperTransport™ 3.0 technology. This direct connect architecture bypasses the PCI bottlenecks inherent in commodity networks and provides a peak of over 20 GB/s of injection bandwidth per node. The Gemini router's connectionless protocol scales from hundreds to hundreds of thousands of cores without the increase in buffer memory required in the point-to-point connection method of commodity interconnects.

The Cray XE6 network provides industry-leading sub-microsecond latency for remote puts and 1-2 microsecond latency for most other point-to-point messages. An internal block transfer engine is available to provide high bandwidth and good overlap of computation and communication for long messages. Advanced features include support for one-sided communication primitives and support for atomic memory operations. The proven 3-D torus topology provides powerful bisection and global bandwidth characteristics as well as support for dynamic routing of messages.



Scalable Programming Paradigms

In addition to supporting MPI over the standard programming languages of C, C++ and Fortran, the Gemini interconnect has direct hardware support for partitioned global address space (PGAS) programming models including Unified Parallel C (UPC), Co-array Fortran and Chapel. Gemini allows remote references to be pipelined in these programming models which can result in orders-of-magnitude performance improvement over library-based message passing models. This feature brings highly scalable performance to communication-intensive, irregular algorithms which until now have been limited by the MPI programming paradigm.

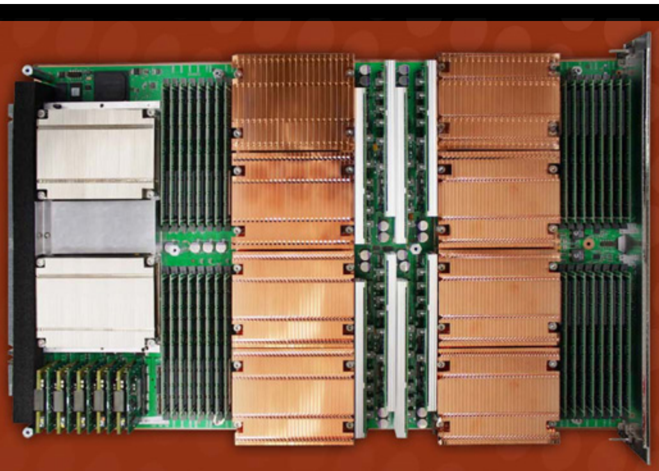
Each Cray XE6 system includes a fully integrated Cray programming environment with tools designed to maximize programmer productivity and application scalability and performance. This feature-rich, easy-to-use programming environment facilitates the development of scalable applications. Supported parallel programming models include MPI, Cray SHMEM™, UPC, Co-Array Fortran and OpenMP. The MPI implementation is compliant with the MPI 2.0 standard and is optimized to take advantage of the Gemini interconnect in the Cray XE6 system. Cray's performance analysis tools CrayPat™ with Cray Apprentice2™ allow users to analyze resource utilization throughout their code at scale and eliminate bottleneck and load imbalance issues.

Scalable Software

The Cray XE6 system ships with Cray Linux Environment™ v3 (CLE3), a suite of high performance software including a SUSE™ Linux-based operating system designed to run large, complex applications and scale to more than 1 million processor cores. The Linux® environment features Compute Node Linux (CNL), a compute kernel. When running highly scalable applications, CNL runs in Extreme Scalability Mode (ESM) which ensures operating system services do not interfere with application scalability. Real world applications have proven this optimized design scales to more than 200,000 cores and is capable of scaling to more than 1 million cores on the Cray XE6 supercomputer.

The Cray XE6 system provides for tightly integrated, industry-standard batch schedulers and parallel job accounting with aggregated resource usage. Supported workload managers include Altair PBS Professional®, Moab Adaptive Computing Suite™ and Platform LSF®; compilers from PGI, PathScale and Cray; debuggers from TotalView Technologies and Allinea and many open source programming tools.

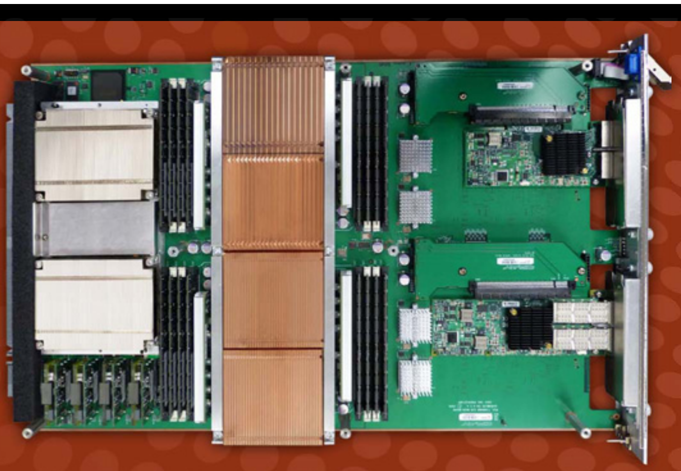
Additionally, the Cray XE6 supercomputer can use a variety of high performance Fortran, C and C++ compilers and libraries including PGI, PathScale and the Cray Compiler Environment with support for optimized C, C++ and Fortran 90, UPC and Co-Array Fortran, as well as high performance-optimized math libraries of BLAS, FFTs, LAPACK, ScaLAPACK, SuperLU and Cray Scientific Libraries.



Scalable Compute Nodes

Each Cray XE6 blade includes four compute nodes for high scalability in a small footprint – up to 96 processor cores per blade or 2,304 processor cores per cabinet. Each compute node has two AMD Opteron™ 6100 Series processors (eight or 12 core) coupled with its own memory and Gemini communication interface and is designed to efficiently run up to 24 MPI tasks. Alternately, it can be programmed to run OpenMP within a compute node and MPI between nodes.

Each Cray XE6 node can be configured with 32 GB or 64 GB DDR3 memory. Memory on compute nodes is registered and memory controllers provide x4 device correction, ensuring reliable memory performance while retaining the upgradeability, serviceability and flexibility of a socketed component.



Scalable I/O

The Cray XE6 I/O subsystem scales to meet the bandwidth needs of even the most data-intensive applications. The newly designed Cray XIO service blades provide four I/O nodes, each with a six-core AMD Opteron Series 2000 processor coupled to 16 GB of DDR2 memory and a PCI-express GEN2 interface. The XIO service blade provides 32 GB/s of peak I/O bandwidth and supports connectivity to networks and storage devices using Ethernet, Fibre Channel or InfiniBand interfaces. The Cray user data storage architecture consists of RAID6 arrays connected directly to Cray XIO nodes or via external SANs with complete multi-path failover. The Oracle® Lustre file system manages the striping of file operations across these arrays. This highly scalable I/O architecture allows customers to configure bandwidth and data capacity by selecting the appropriate number of arrays and service nodes.

OPTERON PROCESSOR AMD

The AMD Opteron 6100 Series processor's highly associative on-chip data cache supports aggressive out-of-order execution. The integrated memory controller eliminates the need for a separate Northbridge memory chip and provides a high-bandwidth path to local memory – 85 GB/sec per dual-socket compute node or over 8 TB/s per cabinet. This design brings a significant performance advantage to algorithms that stress local memory bandwidth and plenty of headroom for processor upgrades.

Production Reliability

Integrated Hardware Supervisory System

Cray's Hardware Supervisory System (HSS) integrates hardware and software components to provide system monitoring, fault identification and recovery. An independent system with its own control processors and supervisory network, the HSS monitors and manages all major hardware and software components in the Cray XE6 supercomputer. In addition to providing recovery services in the event of a hardware or software failure, HSS controls power-up, power-down and boot sequences, manages the interconnect, reroutes around failed interconnect links, and displays the machine state to the system administrator. The Cray XE6 system also supports a warm swap capability allowing a system operator to remove and repair system blades without disrupting an active workload.

Cray XE6 System Resiliency Features

The Gemini interconnect is designed for large systems in which failures are to be expected and applications must run to successful completion in the presence of errors. Gemini uses error correcting code (ECC) to protect major memories and data paths within the device. The ECC combined with the Gemini adaptive routing hardware (which spreads data packets over the four available lanes which comprise each of the torus links) provide improved system and applications resiliency. In the event of a lane failure, the adaptive routing hardware will automatically mask it out. In the event of losing all connectivity between two interconnects, the HSS automatically reconfigures it to route around the bad link.

Additionally, CLE3 features NodeKARE™ (Node Knowledge and Reconfiguration). If a user's program terminates abnormally, NodeKARE automatically runs diagnostics on all involved compute nodes and removes any unhealthy ones from the compute pool. Subsequent jobs are allocated only to healthy nodes and run reliably to completion.

Finally, the Cray XE6 system features redundant power supplies and voltage conversion modules to increase reliability at scale. The Lustre file system can be configured with object storage target failover and metadata server failover. Software failover is provided for all critical system software functions.

Adaptive Supercomputing

Extreme Scale and Cluster Compatibility in One System

The Cray XE6 system provides complete workload flexibility. For the first time, users can buy a single machine to run both a highly scalable custom workload and industry-standard ISV workload. CLE3 accomplishes this through the new Cluster Compatibility Mode (CCM). CCM allows out-of-the-box compatibility with Linux/x86 versions of ISV software – without recompilation or relinking – and allows for the use of various versions of MPI (e.g., MPICH, Platform MPI™). At job submission, the user can request the CNL compute nodes be configured with CCM, complete with the necessary services to ensure Linux/x86 compatibility. The service is dynamic and available on an individual job basis.

Support for Other File System and Data Management Services

Customers can select the Lustre parallel file system or another option, including connecting to an existing parallel file system. The Cray Data Virtualization Service allows for the projection of various other file systems (including NFS, GPFS™, Panasas® and StorNext®) to the compute and login nodes on the Cray XE6 supercomputer. The Cray Data Management group can also provide Cray XE6 customers with solutions for backup, archiving and data lifecycle management.

Cray Efficiency with ECOphlex Cooling

With a standard air- or liquid-cooled High Efficiency cabinet and optional ECOphlex™ technology, the Cray XE6 system can reduce cooling costs and increase flexibility in datacenter design. Each High Efficiency cabinet can be configured with inline phase-change evaporator coils which extract virtually all the heat imparted to the airstream as it passes through the cabinet. Coolant is recondensed in a heat exchange unit connected to the building chilled water supply.

ECOphlex technology accommodates a range of building water temperatures, so a modern datacenter can operate chillers and air handlers less often, reducing electrical costs. In fact, during much of the year in many climates a system fitted with ECOphlex operating at full capacity needs only cooling towers.

Investment Protection

The Cray XE6 supercomputer is engineered for easy, flexible upgrades and expansion, a benefit that prolongs its productive lifetime – and the customer's investment. As new technologies become available, customers can take advantage of these next-generation compute processors, I/O technologies and interconnect without replacing the entire Cray XE6 system.

Cray XE6 Specifications

Processor	Eight or 12-core 64-bit AMD Opteron 6100 Series processors; up to 192 per cabinet
	64K L1 instruction cache, 64K L1 data cache, 512 KB L2 cache per processor core, 12 MB shared L3 cache
Memory	32 GB or 64 GB registered ECC DDR3 SDRAM per compute node
	Memory Bandwidth: 85.3 GB/s per compute node
Compute Cabinet	Cores : 1,536 or 2,304 processor cores per system cabinet
	Peak Performance : 12.2 to 20.2 teraflops per system cabinet
Interconnect	1 Gemini routing and communications ASIC per two compute nodes
	48 switch ports per Gemini chip, (160 GB/s internal switching capacity per chip)
	3-D torus interconnect
System Administration	Cray System Management workstation
	Graphical and command line system administration
	Single-system view for system administration
	System software rollback capability
Reliability Features (Hardware)	Cray Hardware Supervisory System (HSS) with independent 100 Mb/s management fabric between all system blades and cabinet-level controllers
	Full ECC protection of all packet traffic in the Gemini network
	Redundant power supplies; redundant voltage regulator modules
	Redundant paths to all system RAID
	Variable speed axial turbofan with integrated pressure and temperature sensors
Reliability Features (Software)	HSS system monitors operation of all operating system kernels
	Lustre file system object storage target failover; Lustre metadata server failover
	Software failover for critical system services including system database, system logger and batch subsystems
	NodeKARE (Node Knowledge and Reconfiguration)
Operating System	Cray Linux Environment (components include SUSE Linux SLES11, HSS and SMW software)
	Extreme Scalabiliity Mode (ESM) and Cluster Compatibility Mode (CCM)
Compilers, Libraries & Tools	PGI compilers, Cray Compiler Environment, PathScale Support for Fortran 77, 90, 95; C/C++, UPC, Co-Array Fortran MPI 2.0, Cray SHMEM, other standard MPI libraries using CCM Cray Apprentice, Cray PAT and Cray Compiler included with systems
Job Management	PBS Professional job management system Moab Adaptive Computing Suite job management system Platform LSF job management system
External I/O Interface	InfiniBand, 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet
Disk Storage	Full line of FC-attached disk arrays with support for FC and SATA disk drives
Parallel File System	Lustre, Data Virtualization Service allows support for NFS, external Lustre and other file systems
Power	45-54.1 kW (45.9 – 55.2 kVA) per cabinet, depending on configuration Circuit requirements: three-phase wye, 100 AMP at 480/277 and 125 AMP at 400/230 (three-phase, neutral and ground)
Cooling	Air-cooled, air flow: 3,000 cfm (1.41 m3/s); intake: bottom; exhaust: top
	Optional ECOphlex liquid cooling
Dimensions (Cabinet)	H 93 in. (2,362 mm) x W 22.50 in. (572 mm) x D 56.75 in. (1,441 mm)
Weight (Maximum)	1,600 lbs. per cabinet (725 kg) air cooled; 2,000 lbs. per cabinet (907 kg) liquid cooled
Regulatory Compliance	UL 60950-1, CAN/CSA – C 22.2 No. 60950-1, CE-mark, RoHS, WEEE
Safety	FCC Class A, VCCI Class A, ICES-003, EN 50022:2006 Class A, AS/NZS CISPR 22:2006, EN 55024: 1998 +A1:2002 +A2:2003

CRAY

THE SUPERCOMPUTER COMPANY