CRAY XK6

We built the world's first production petaflops system with the Cray XT5™ supercomputer. Then we reinvented high performance networking with the Gemini™ interconnect. Now we bring it all together with many-core technology. Introducing the first production hybrid supercomputing system – the Cray XK6™ supercomputer.

Redefining Supercomputing

The Cray XK6 supercomputer is a trifecta of scalar, network and many-core innovation. It combines Cray's proven Gemini interconnect, AMD's leading multi-core scalar processors and NVIDIA's powerful many-core GPU processors to create a true, productive hybrid supercomputer.

Capable of scaling to 500,000 scalar processors and 50 petaflops of hybrid peak performance, every aspect of the Cray XK6 system – from its resiliency features to its scalability-boosting technologies – has been engineered to meet science's real-world demands. And with the expertise of the industry's only company singularly focused on high performance computing behind it, the Cray XK6 supercomputer brings Cray's trademark reliability, flexibility and scalability to the many-core HPC environment.

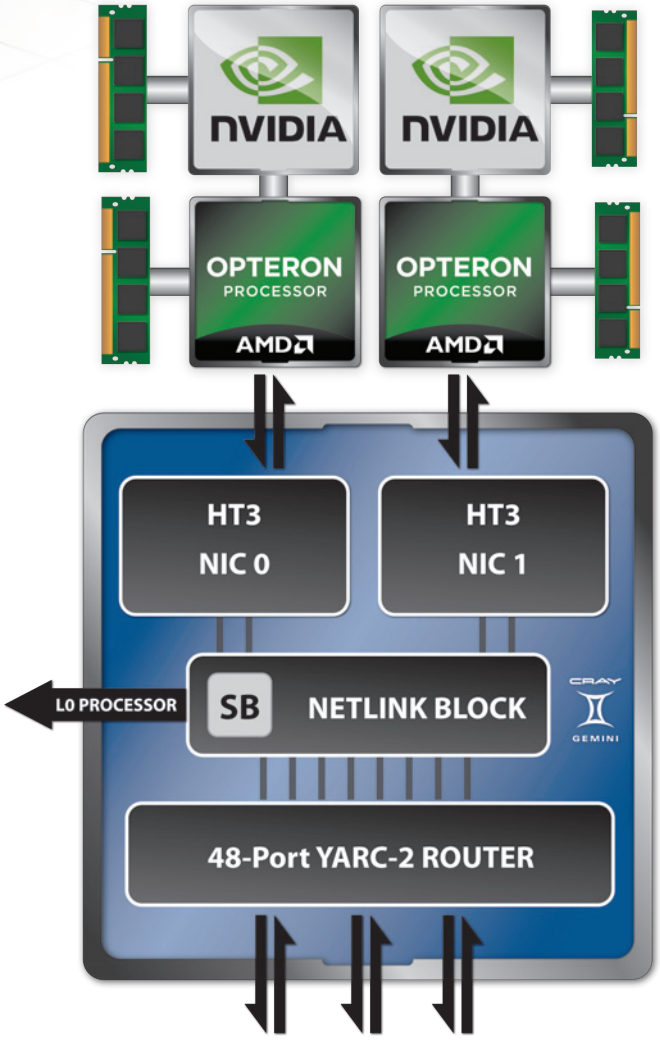
Scalable Many-Core Performance

Adaptive Hybrid Computing

The Cray XK6 compute node combines AMD's 16-core Opteron™ 6200 Series processor and NVIDIA's Tesla X2090 many-core accelerator to create a hybrid unit with the intra-node scalability, power-efficiency of acceleration and flexibility to run applications with either scalar or accelerator components. This compute unit, combined with the Gemini interconnect's excellent inter-node scalability, creates a system geared for any computing challenge.

Gemini Scalable Interconnect

Capable of tens of millions of MPI messages per second, the Gemini ASIC complements current and future massively multi-core and many-core processors. Each hybrid compute 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 XK6 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 Capabilities

The Cray XK6 supercomputer supports a full range of powerful scalar tools, libraries, compilers, operating system and third-party software.

Programming Environment

Each Cray XK6 system includes a fully integrated Cray programming environment with tools designed to maximize programmer productivity and application scalability and performance. This feature-rich, flexible programming environment facilitates the development of scalable applications. For example, the Cray XK6 supercomputer can use a variety of high performance Fortran, C and C++ compilers and libraries for both x86 and GPU workloads. Available compilers include GPU-enabled PGI, CAPS, PathScale and the Cray Compiler Environment for x86 with support for optimized C, C++ and Fortran 90, UPC and Co-Array Fortran. Available high performance-optimized math libraries include BLAS, FFTs, LAPACK, ScaLAPACK, SuperLU and Cray Scientific and Math Libraries.

Programming Models

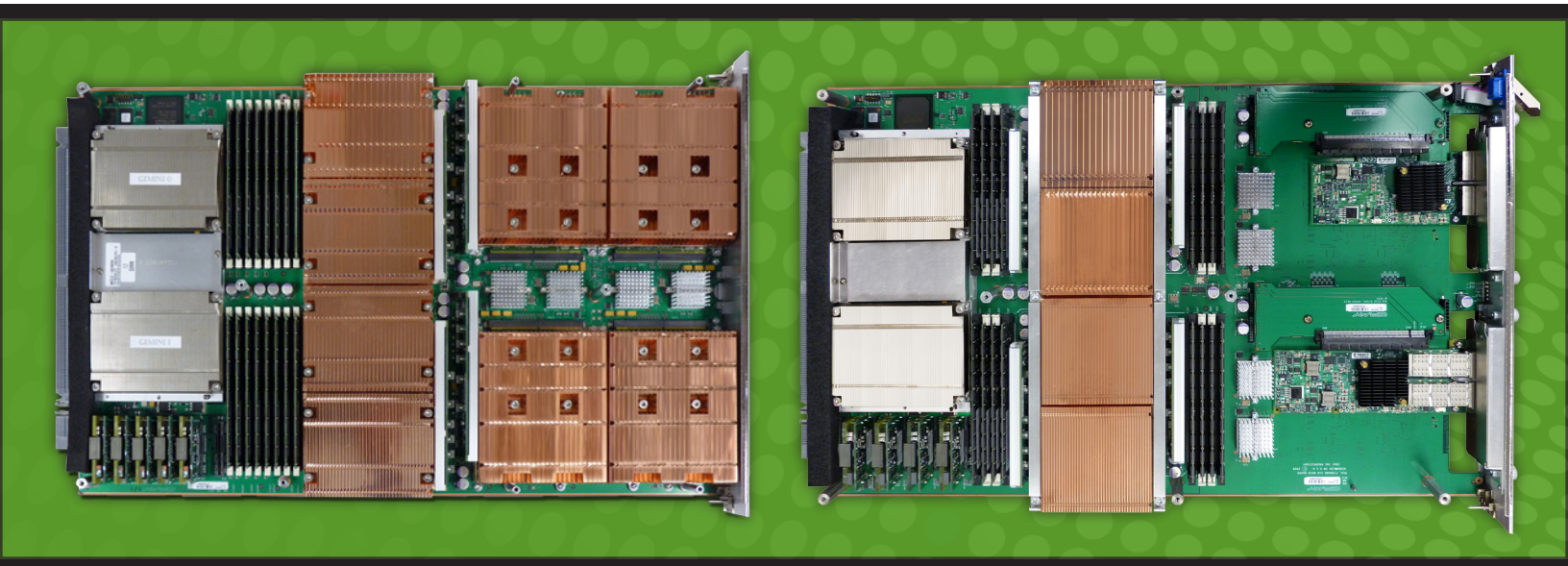
Supported parallel programming models include MPI, CUDA, Cray SHMEM™, UPC, Co-Array Fortran and OpenMP. The MPI implementation is compliant with the MPI 2.0 standard and is optimized to take advantage of the Gemini interconnect. Cray's performance and analysis toolkit CrayPat™ allows users to analyze resource utilization throughout their code at scale and eliminate bottleneck and load imbalance issues.

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.

Scalable Software

The Cray XK6 system ships with the latest Cray Linux Environment™ (CLE), a suite of high performance software including a SUSE™ Linux-based operating system designed to scale efficiently and run large, complex applications. With highly scalable applications, Compute Node Linux (CNL) runs in Extreme Scalability Mode (ESM), ensuring operating system services do not interfere with application scalability. Real-world applications have proven this optimized design scales to 250,000 cores.

Additionally, the Cray XK6 system provides for tightly integrated, industry-standard batch schedulers, compilers and debuggers. Supported workload managers include Altair PBS Professional®, Moab Adaptive Computing Suite™ and Platform LSF®. Supported compilers include PGI, PathScale, CAPS and Cray. Supported debuggers include TotalView Technologies and Allinea as well as many open source programming tools.



Scalable Compute Nodes

Each Cray XK6 blade includes four compute nodes for high scalability in a small footprint with up to 64 AMD processor cores per blade. Each compute node has an AMD Opteron 6200 Series processor with eight channel DDR3 memory and an NVIDIA Tesla X2090 GPU many-core processor with 6GB of GDDR5 memory. The CUDA GPU architecture of the NVIDIA Tesla processor incorporates error correcting code for memories and double precision floating point units.

Each Cray XK6 node can be configured with 16 GB or 32 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 XK6 I/O subsystem scales to meet the bandwidth needs of even the most data-intensive applications. Each Cray XIO service blade provides four multi-purpose 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. Additionally, the Cray 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. Cray systems may be ordered with a Lustre file system that manages the striping of file operations across these arrays. This highly scalable I/O architecture allows for configuring bandwidth and data capacity by selecting the appropriate number of arrays and service nodes.

The AMD Opteron 6200 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 with eight channels of DDR3 memory per dual-socket compute node. This design brings a significant performance advantage to algorithms that stress local memory bandwidth and plenty of headroom for processor upgrades.

The NVIDIA Tesla 20-series is designed from the ground up for high performance computing. Based on the next generation CUDA GPU architecture codenamed "Fermi," it supports many must-have features for technical and enterprise computing. These include ECC protection for uncompromised accuracy and data reliability, support for C++, and double precision floating-point performance.

CRAY LINUX ENVIRONMENT

ESM –
Extreme Scalability Mode

- Uncompromised 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

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 XK6 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 XK6 system also supports a warm swap capability allowing a system operator to remove and repair system blades without disrupting an active workload.

Cray XK6 System Resiliency

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 to route around the bad link.

Additionally, the Cray Linux Environment features NodeKARE™ (Node Knowledge and Reconfiguration). If a 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.

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 XK6 system provides complete workload flexibility. For the first time, you can buy a single machine to run both a highly scalable custom workload and industry-standard ISV workload. CLE accomplishes this through the 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, you 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

You 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 XK6 system. The Cray Data Management group can also provide solutions for backup, archiving and data lifecycle management.

Emphasis on Power Efficiency

Many-core processing is the key to ultimate energy efficiency. Applications using the Cray XK6 many-core GPU processors will experience industry-leading energy efficiency when measured for real application workloads. Combined with our standard air- or liquid-cooled High Efficiency cabinet and optional ECOphlex™ technology, the Cray XK6 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, a system fitted with ECOphlex operating at full capacity needs only cooling towers during much of the year in many climates.

Investment Protection and Blended Systems

The Cray XK6 supercomputer is engineered for easy, flexible up-grades and expansion – prolonging its productive lifetime and your investment. As new technologies become available, you can take advantage of these next-generation compute processors, I/O technologies and interconnect without replacing the entire Cray XK6 system. In addition, Cray XK6 and Cray XE6 systems, support blended configurations on the same Gemini interconnect and share the same power, cooling, I/O and service infrastructure, making it easy for current Cray XE6 users to add Cray XK6 technology.

Cray XK6 Specifications

Processor	16-core 64-bit AMD Opteron 6200 Series processors, up to 96 per cabinet; NVIDIA Tesla X2090 GPU, up to 96 per cabinet
Memory	16 GB or 32 GB registered ECC DDR3 SDRAM and 6 GB GDDR5 per compute node
	Memory bandwidth: 8 channels of DDR3 memory per compute node
Compute Cabinet	AMD processing cores: 1,536 processor cores per system cabinet
	Peak performance: 70+ Tflops 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
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, CUDA, CAPS, support for Fortran 77, 90, 95, C/C++, UPC, Co-Array Fortran MPI 2.0, Cray SHMEM, other standard MPI libraries using CCM
Job Management	PBS Professional, Moab Adaptive Computing Suite, Platform LSF
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