CRAY

CRAY XT3 DATASHEET

Cray XT3™ Supercomputer
Scalable by Design

The Cray XT3 system offers a new level of scalable computing where:

- a single powerful computing system handles the most complex problems
- every component is engineered to run massively parallel computing applications reliably to completion
- the operating system and management system are tightly integrated and designed for ease of operation at massive scale
- scalable performance analysis and debugging tools allow for rapid testing and fine tuning of applications
- highly scalable global I/O performance ensures high efficiency for applications that require rapid I/O access for large datasets

Introducing the third generation massively parallel processor (MPP) system from Cray–the Cray XT3 supercomputer.

Building on the success of its predecessors, the Cray T3D™ and the Cray T3E™ systems, the Cray XT3™ system brings astounding new levels of scalability and sustained application performance to high performance computing (HPC).

Purpose-built to meet the special needs of capability class HPC applications, each feature and function is designed for larger problems, faster solutions, and a greater return on investment.

The Cray XT3 supercomputer scales to support the most challenging HPC workloads.

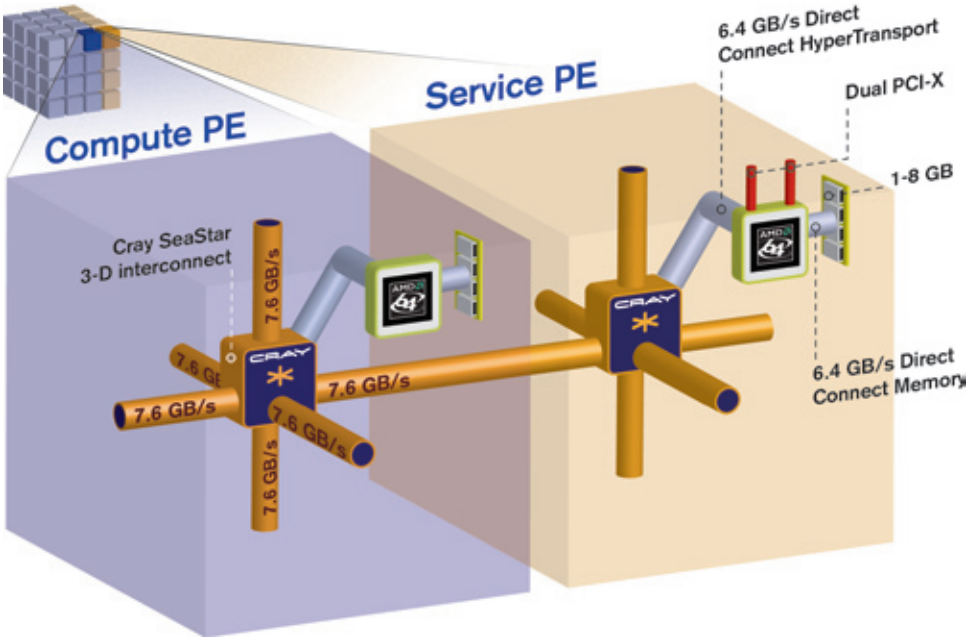
3D Torus Direct Connected Processor Architecture

The Cray XT3 system architecture is designed for superior application performance for large-scale massively parallel computing. As in Cray's previous MPP systems, the design builds upon a single-processor node, or processing element (PE). Each PE is comprised of one AMD Opteron™ processor coupled with its own memory and dedicated communication resource. The Cray XT3 system incorporates two types of processing elements: compute PEs and service PEs. Compute PEs run a light weight kernel that is optimized for application performance. Service PE's run Linux and can be configured for I/O, login, network or system functions.

Each Opteron processor is directly connected to the Cray XT3 interconnect via its Cray SeaStar™ routing and communications chip over a 6.4 GB/s HyperTransport™ path. A powerful communications resource, the Cray SeaStar chip acts as the gateway to the Cray XT3 high bandwidth, low latency interconnect. The router in the Cray SeaStar chip provides six high speed network links to connect to six neighbors in the 3D torus topology.

This architecture directly connects all Opteron processors in the Cray XT3 system, removing PCI bottlenecks and shared memory contention to deliver superior sustained application performance at massive scale.

Cray XT3 Scalable Architecture



Cray XT3 System Sample Configurations

	6 Cabinets	24 Cabinets	96 Cabinets	320 Cabinets
Compute PEs	548	2260	9108	30,508
Service PEs	14	22	54	106
Peak (TFLOPS)	2.6*	10.8*	43.7*	147 *
Max Memory (TB)	4.3	17.7	71.2	239
Aggregate Memory Bandwidth (TB/s)	2.5 TB/s	14.5 TB/s	58.3 TB/s	196 TB/s
Interconnect Topology	6 x 12 x 8	12 x 12 x 16	24 x 16 x 24	40 x 32 x 24
Peak Bisection Bandwidth (TB/s)	0.7	2.2	5.8	11.7
Floor Space (Tiles)	12	72	336	1,200

*** based on 2.4 Ghz AMD Opteron processor**

Cray XT3 System Highlights

Scalable Application Performance

The Cray XT3 supercomputer's high speed 3D torus interconnect, x86 64-bit AMD Opteron™ processors, high speed global I/O, and advanced MPP operating system ensure that applications scale steadily from 200 to 30,000 processors without performance losses from communications bottlenecks, asynchronous processing, or memory access delays.

Scalable Reliability and Management

Each Cray XT3 component, from industrial cooling fans, to disk drives, to the Cray Reliability, Availability and Serviceability (RAS) and Management System, is engineered to operate as part of a highly reliable system at immense scale, ensuring that large, complex jobs run to completion.

Tightly integrated operating and management systems allow administrators to manage hundreds or thousands of processors as a single system, eliminating the administrative effort and problems associated with loosely coupled cluster systems.

Scalable Programmability

The Cray XT3 supercomputer lets programmers focus on their applications instead of designing around processing inefficiencies such as asymmetric processor performance, memory access algorithms, and communication delays. Fully scalable performance analysis and debugging tools enable programmers to rapidly test and fine-tune their applications.

Scalable I/O

The Cray XT3 I/O system uses the highly scalable, open source Lustre™ parallel file system. Highly reliable Fibre Channel disks and controllers provide up to 100 GB/s global I/O performance, ensuring high efficiency for I/O intensive applications and providing the I/O capacity needed for rapid data dumps and user level checkpointing.

Scalable System Upgrades

Cray XT3 systems can be expanded by adding cabinets or by upgrading Opteron processors with faster or dual-core models, or upgrading the Cray SeaStar processor to increase interconnect speeds. This flexible expansion ensures a long system life, maximizing return on investment.

Every aspect of the Cray XT3 system is engineered to deliver superior performance for massively parallel applications, including:

- **scalable processing elements each with their own high performance AMD Opteron processors and memory**
- **high bandwidth, low latency interconnect**
- **MPP optimized operating system**
- **standards-based programming environment**
- **sophisticated RAS and system management features**
- **high speed, highly reliable I/O system**

Scalable Processing Elements

Like previous MPP systems from Cray, the basic building block of the Cray XT3 System is a PE. Each PE contains an Opteron 64-bit processor, dedicated memory and a HyperTransport link to a dedicated Cray SeaStar communications engine. This design eliminates the scheduling complexities and asymmetric performance problems associated with clusters of SMPs. It ensures that performance is uniform across distributed memory processes–an absolute requirement for scalable algorithms.

30,000

200 - 30,000 processors in a single system

Each Cray XT3 compute blade includes four compute PEs for high scalability in a small footprint. Service blades include two service PEs and provide PCI-X connectivity.

AMD Opteron Processor

The industry leading Opteron microprocessor offers a number of advantages for superior performance and scalability.

The Opteron processor's on-chip, highly associative 1 MB cache supports aggressive out-of-order execution and can issue up to nine instructions simultaneously. The integrated memory controller eliminates the need for a separate Northbridge memory controller chip, providing an extremely low latency path to local memory–less than 60 nanoseconds. This is a significant performance advantage, particularly for algorithms that require irregular memory access. The 128-bit wide memory controller provides 6.4 GB/s local memory bandwidth per processor, or more than one byte per FLOP. This balance brings a performance advantage to algorithms that stress local memory bandwidth.

HyperTransport technology enables a 6.4 GB/s direct connection between the processor and the Cray XT3 interconnect, removing the PCI bottleneck inherent in most interconnects.

Memory

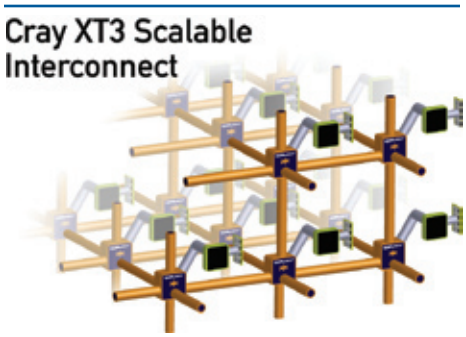
Each Cray XT3 PE can be configured with from 1 to 8 GB of memory. All memory in the Cray XT3 system is protected with Chipkill™ technology, which increases memory reliability by two orders of magnitude when compared with standard error checking and correction (ECC) alone, enabling high memory reliability even in systems with tens of thousands of DIMMs.

Scalable Interconnect

The Cray XT3 system incorporates a high bandwidth, low latency interconnect, comprised of Cray SeaStar chips and high speed links based on HyperTransport and proprietary protocols. The interconnect directly connects all processing elements in a Cray XT3 system in a 3D torus topology, eliminating the cost and complexity of external switches. This improves reliability and allows systems to economically scale to tens of thousands of nodes–well beyond the capacity of fat-tree switches. The backbone of the Cray XT3 system, the interconnect carries all message passing traffic as well as all I/O traffic to the global file system.

Cray SeaStar Chip

The Cray SeaStar chip combines communications processing and high speed routing on a single device. Each communications chip is composed of a HyperTransport link, a Direct Memory Access (DMA) engine, a communications and management processor, a high-speed interconnect router, and a service port.



Interconnect router – The router in the Cray SeaStar chip provides six high-speed network links which connect to six neighbors in the 3D torus. The peak bidirectional bandwidth of each link is 7.6 GB/s with a sustained bandwidth in excess of 4 GB/s. The router also includes reliable link protocol with error correction and retransmission.

DMA Engine – The Cray SeaStar chip features a DMA engine and an associated PowerPC™ 440 processor. These work together to off-load message preparation and demultiplexing tasks from the Opteron Processor, leaving it free to focus exclusively on computing tasks. The DMA engine and the Cray XT3 operating system work together to minimize latency by providing a path directly from the application to the communication hardware without the traps and interrupts associated with traversing a protected kernel.

Interconnect Reliability Features

Each link on the chip runs a reliability protocol that supports Cyclic Redundancy Check (CRC) and automatic retransmission in hardware. In the presence of a bad connection, a link can be configured to run in a degraded mode while still providing connectivity. This protocol also enables routing tables to be dynamically reconfigured while jobs continue to run, leading to increased system efficiency.

The Cray SeaStar chip provides a service port that bridges between the separate management network and the Cray SeaStar local bus. This service port allows the management system to access all registers and memory in the system and facilitates booting, maintenance, and system monitoring.

Scalable Operating System

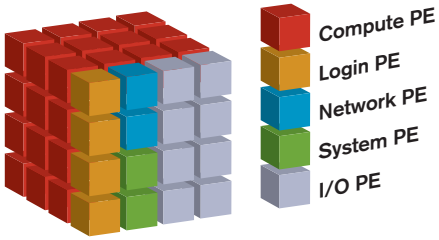
The Cray XT3 operating system UNICOS/lc is designed to run large complex applications and scale efficiently to 30,000 processors. As in previous generation MPP systems from Cray, UNICOS/lc consists of two primary components–a microkernel for compute PEs and a full-featured operating system for the service PEs.

The Cray XT3 Catamount microkernel runs on the compute PEs and provides a computational environment that minimizes system overhead–critical to allowing the systems to scale to thousands of processors. The microkernel interacts with an application process in a very limited way, including managing virtual memory addressing, providing memory protection and performing basic scheduling. This proven microkernel architecture ensures reproducible run-times for MPP jobs, supports fine grain synchronization at scale, and ensures high performance, low latency MPI and SHMEM communication.

Service PEs run a full Linux™ distribution. Service PEs can be configured to provide login, I/O, system, or network services.

Login PEs offer the programmer the look and feel of a Linux-based environment with full access to the programming environment and all of the standard Linux utilities, commands, and shells to make program development both easy and portable. Network PEs provide high-speed connectivity with other systems. I/O PEs provide scalable connectivity to the global, parallel file

UNICOS/lc Architecture



system. System PEs are used to run global system services such as the system database. System services can be scaled to fit the size of the system or the specific needs of the users.

Jobs are submitted interactively from login PEs using the Cray XT3 job launch command, or through the PBS Pro™ batch program, which is tightly integrated with the system PE scheduler. Jobs are scheduled on dedicated sets of compute PEs and the system administrator can define batch and interactive partitions. The system provides accounting for parallel jobs as single entities with aggregated resource usage.

The Cray XT3 system maintains a single root file system across all nodes, ensuring that modifications are immediately visible throughout the system without transmitting changes to each individual PE. Fast boot times ensure that software upgrades can be completed quickly, with minimal downtime. In addition, the Cray XT3 system provides a set of administration tools for tracking and rolling back modifications to the root file system.

Scalable Programming Environment

Designed around open system standards, the Cray XT3 is easy to program. The system's single PE architecture and microkernel-based operating system ensure that system-induced performance issues are eliminated, allowing the user to focus exclusively on their application.

The Cray XT3 programming environment incudes tools designed to complement and enhance each other, resulting in a rich, easy-to-use programming environment that facilitates the development of scalable applications. The Opteron processor's native support for 32-bit and 64-bit applications and full x86-64 compatibility makes the Cray XT3 system compatible with a vast quantity of existing compilers and libraries, including optimized C, C++, and Fortran90 compilers and high performance math libraries such as optimized versions of BLAS, FFTs, LAPACK, ScaLAPACK, and SuperLU.

Communication libraries include MPI and SHMEM. The MPI implementation is compliant with the MPI 2.0 standard and is optimized to take advantage of the scalable interconnect in the Cray XT3 system, offering scalable message passing performance to tens of thousands of PEs. The SHMEM library is compatible with previous Cray systems and operates directly over the Cray SeaStar chip to ensure uncompromised communications performance.

Cray Apprentice^{2}™ performance analysis tools are also included with the Cray XT3 system. They allow users to analyze resource utilization throughout their code and can help uncover load-balance issues when executing in parallel.

Scalable RAS & Administration

The Cray RAS and Management System (CRMS) 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 CRMS monitors and manages all of the major hardware and software components in the Cray XT3 system. In addition to providing recovery services in the event of a hardware or software failure, CRMS controls power-up, power down, and boot sequences, manages the interconnect, and displays the machine state to the system administrator.

45.6

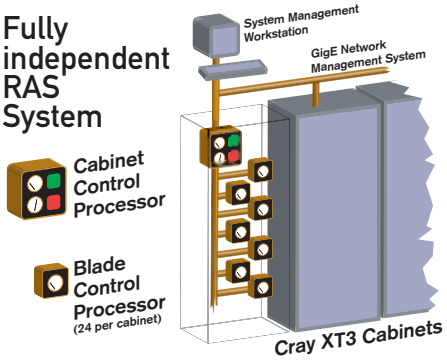
GB/s switching per Cray SeaStar

CRMS is an independent system with its own processors and supervisory network. The services CRMS provides do not take resources from running applications. When a component fails, CRMS can continue to provide fault identification and recovery services and allow the functional parts of the system to continue operating.

The Cray XT3 system is designed for high reliability. Redundancy is built in for critical components and single points of failure are minimized. For example, the system could lose an I/O PE, without losing the job that was using it. A Cray SeaStar chip could fail, and yet jobs routed through that processor can recover and continue. The system boards contain no moving parts, further enhancing overall reliability.

The Cray XT3 processor and I/O boards use socketed components wherever possible. The SeaStar chip, the RAS processor module, the DIMMs, the voltage regulator modules (VRMs), and the Opteron processors are all field

Fully independent RAS System

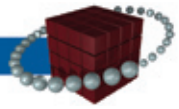


replaceable and upgradable. The Cray XT3 system backplane is designed to allow modules to be hot-swapped to replace faulty components without the need for a system shutdown, enhancing system availability. All components have redundant power, including redundant VRMs on all system blades.

Scalable I/O

The Cray XT3 I/O subsystem scales to meet the bandwidth needs of even the most data intensive applications. The I/O architecture consists of data RAIDs connected directly to I/O PEs which reside on the high-speed interconnect. The Lustre file system manages the striping of file operations across these RAIDs. This highly scalable I/O architecture enables customers to configure the Cray XT3 with desired bandwidth by selecting the appropriate number of RAIDs and service PEs. It gives users and applications access to a filesystem with filenames that are uniform and global without the need for costly FibreChannel fabric and switches.

To maximize I/O performance Lustre is integrated directly into applications running on the system microkernel. Data moves directly between applications space and the Lustre servers on the I/O PEs without the need for an intervening data copy through the lightweight kernel. The Cray XT3 combines the scalability of a microkernel based operating system with the I/O performance normally associated with Large-scale SMP Servers.



CPU	64-bit AMD Opteron 100 series processors; up to 96 per cabinet
Cache	64K L1 instruction cache, 64K L1 data cache, 1 MB L2 cache per processor
FLOPS	460 GFLOPS per cabinet (96 processors @ 2.4 GHz)
Main Memory	1-8 GB Registered ECC SDRAM per processor. Supports Chipkill™
Memory Bandwidth	6.4 GB/s per processor
Interconnect	1 Cray SeaStar routing and communications ASIC per Opteron processor
	6 switch ports per Cray SeaStar chip, 7.6 GB/s each (45.6 GB/s switching capacity per Cray SeaStar chip)
	3 dimensional torus interconnect
	3 microsecond MPI latency between PEs
External I/O	2 independent 64-bit 133 MHz PCI-X buses per service PE
	Gigabit Ethernet PCI-X card (copper and optical)
	Dual-Port 2 GB/s Fibre Channel Host Bus Adapter (optical)
	10 Gigabit Ethernet card (Optical)
Disk	4 and 8 port Fibre Channel RAID controllers
	Configurable Fibre Channel RAID drive sets
File System	Lustre File System
System Administration	Cray System Management Workstation (SMW)
	Graphical and command line system administration
	Single system view for system administration
	PBS Pro job management system
	Hot Swap support for system blades
	System software rollback capability
Reliability Features (Hardware)	Cray RAS and Management Subsystem (CRMS) with independent 100Mbits/s management fabric between all system blades and cabinet level controllers. Over 50 measurement points monitored per Cray XT3 system blade
	Full ECC memory protection from memory to system registers. Chipkill memory protection on all system memory DIMMs
	Full ECC protection in the Cray SeaStar chip
	Cray XT3 interconnect routes around failures, supports graceful link degradation in case of partial link failure
	Redundant power supplies
	Redundant voltage regulator modules (VRMs)
	Redundant paths to all system RAID
	Variable speed blowers with integrated pressure and temperature sensors
	Hot Swap system blades
Reliability Features (software)	Simple, microkernel-based software design
	CRMS 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
Operating System	UNICOS/lc–Components include SUSE™ Linux™, Cray Catamount Microkernel, CRMS and SMW software
Message Passing Libraries	MPI 2.0, SHMEM
Compilers	Fortran 77, 90, 95; C/C++
Power	14.8 kVA (14.5 kW) per cabinet. Circuit Requirement: 80 AMP at 200/208 VAC (3 Phase & Ground), 63 AMP at 400 VAC (3 Phase, Neutral & Ground)
Cooling Requirement	Air Cooled, Air Flow: 3000 cfm (1.41 m3/s) Intake: bottom, Exhaust: top.
Dimensions (cabinet)	H 80.50 in. (2045 mm) x W 22.50 in. (572 mm) x D 56.75 in. (1441 mm)
Weight (maximum)	1529 lbs per cabinet (694 kg)
Acoustical Noise Level	75 dba at 3.3 ft (1.0 m)
Regulatory Compliance	UL 60950-1, CAN/CSA – C 22.2 No 60950–1, CB Scheme Investigation to IEC/EN 60950-1
Safety	FCC Class A, DOC Class A, VCCI Class, CISPR 22, EN 50022 Class A, AS/NZS 3548, EN 50082-1, EN 61000-3-2, EN 61000-3-3, Statskontoret 26.2 Category 1

The Supercomputer Company

Global Headquarters:

Cray Inc.
411 First Avenue S., Suite 600
Seattle, WA 98104-2860 USA

tel (206) 701 2000
fax (206) 701 2500

Sales Inquiries:

North America: 1 (877) CRAY INC
Worldwide: 1 (651) 605 8817
sales@cray.com

www.cray.com

0904