THE SUPERCOMPUTER COMPANY

CRAY XT4 DATASHEET

Cray XT4™ Supercomputer

Introducing the latest generation massively parallel processor (MPP) system from Cray—the Cray XT4 supercomputer. Building on the success of the Cray XT3™ system, the Cray XT4 system brings new levels of scalability and sustained performance to high performance computing (HPC).

Engineered to meet the demanding needs of capability class HPC applications, each feature and function is selected in order to enable larger problems, faster solutions, and a greater return on investment. Designed to support the most challenging HPC workloads, the Cray XT4 supercomputer delivers scalable power for the toughest computing challenges.

The Cray XT4 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 to completion, reliably and fast
- 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

3D Torus Direct Connected Processor Architecture

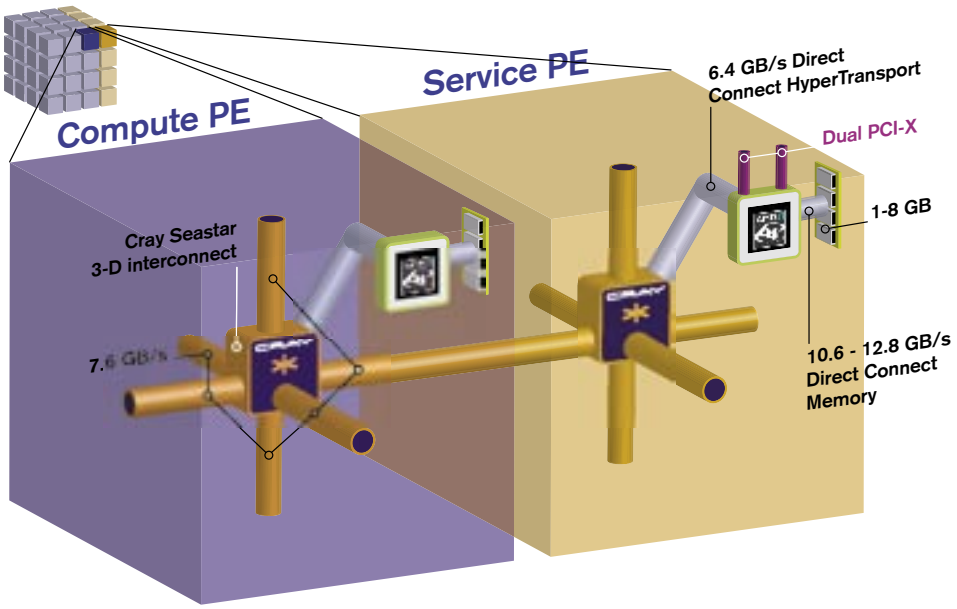
The Cray XT4 system architecture is designed for superior application performance for large-scale massively parallel computing. As in Cray's previous MPP systems, the design builds upon a single processor node, or processing element (PE). Each PE is comprised of one AMD microprocessor (single, dual or quad core) coupled with its own memory and dedicated communication resource. The Cray XT4 system incorporates two types of processing elements:

Compute PEs and service PEs. Compute PEs run a lightweight kernel that is optimized for application performance. Service PE's run standard Linux and can be configured for I/O, login, network or system functions.

Each AMD processor is directly connected to the Cray XT4 interconnect via its Cray SeaStar2™ routing and communications chip over a 6.4 GB/s HyperTransport™ path. A powerful communications resource, the Cray SeaStar2 chip acts as a gateway between the users MPI program and the Cray XT4 high bandwidth, low latency interconnect. The router in the Cray SeaStar2 chip provides six high speed network links to connect to six neighbors in the 3D torus topology.

This architecture directly connects all AMD processors in the Cray XT4 system, removing PCI bottlenecks and shared memory contention to deliver superior sustained application performance at massive scale.

Cray XT4 Scalable Architecture



Cray XT4 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)	5.6*	23.4*	94.6*	318 *
Max Memory (TB)	4.3	17.7	71.2	239
Aggregate Memory Bandwidth (TB/s)	7 TB/s**	29 TB/s**	116 TB/s**	390 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)	1.4	2.9	8.7	19.4
Floor Space (Tiles)	12	72	336	1,200

*** based on 2.6 GHz AMD Dual core Processor**
**** Based on 800 Mhz DDR2 memory system**

Cray XT4 System Highlights

Scalable Application Performance

The Cray XT4 supercomputer's high speed 3D torus interconnect, 64-bit AMD Opteron processors, high speed global I/O and advanced MPP operating system ensure that applications scale steadily from 200 to 120,000 processor cores. All system components are designed to avoid the performance losses associated with communication bottlenecks, asynchronous processing, memory access delays or operating system jitter.

Scalable Reliability and Management

Each Cray XT4 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 XT4 supercomputer lets programmers focus directly on their applications instead of programming around the numerous hardware and operating system inefficiencies found in typical cluster-based machines. Fully scalable and integrated performance analysis and debugging tools enable programmers to rapidly test and fine-tune their applications on extremely large processor counts.

Scalable I/O

The Cray XT4 I/O system uses the highly scalable, open source Lustre™ parallel file system. When combined with high-performance disk storage systems, Lustre provides high-bandwidth and capacity for rapid data dumps, and user level checkpointing.

Scalable System Upgrades

Cray XT4 systems can be expanded by adding cabinets or by upgrading AMD processors with faster or quad-core models. This flexible expansion ensures a long system life, maximizing return on investment.

Every aspect of the Cray XT4 system is engineered to deliver superior performance for massively parallel applications, including:

- **scalable processing elements each with their own high performance AMD 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 Cray MPP systems, the basic building block of the Cray XT4 system is a PE. Each PE is comprised of one AMD processor (single, dual or quad core) coupled with its own memory and dedicated communication resource. 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.

Each Cray XT4 compute blade includes four compute PEs for high scalability in a small footprint. Service blades include two service PEs and provide direct I/O connectivity

AMD Opteron

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

The AMD processor's on-chip, highly associative data 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 10.6 to 12.8 GB/sec local memory bandwidth per AMD Opteron, 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 XT4 interconnect, removing the PCI bottleneck inherent in most interconnects.

Memory

Each Cray XT4 PE can be configured with from 1 to 8 GB DDR2 memory. Memory on compute PEs is unbuffered, which provides applications with the lowest possible memory latency.

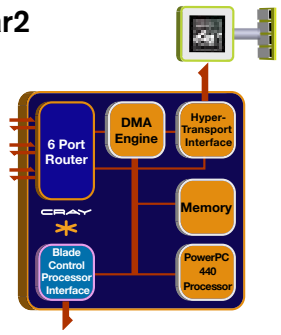
Scalable Interconnect

The Cray XT4 system incorporates a high bandwidth, low latency interconnect, comprised of Cray SeaStar2 chips and high speed links based on HyperTransport and proprietary protocols. The interconnect directly connects all processing elements in a Cray XT4 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. As the backbone of the Cray XT4 system, the interconnect carries all message passing traffic as well as all I/O traffic to the global file system.

Cray SeaStar2 Chip

The Cray SeaStar2 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.

Cray SeaStar2 Architecture



Interconnect router – The router in the Cray SeaStar2 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 6 GB/s. The router also includes reliable link protocol with error correction and retransmission.

Communications Engine - The Cray SeaStar2 chip features a DMA engine and an associated embedded PowerPC™ 440 processor. These work together to off-load message preparation and demultiplexing tasks from the AMD processor, leaving it free to focus exclusively on computing tasks. Logic within the SeaStar2 efficiently matches MPI send and receive operations, eliminating the need for the large, applications-robbing memory buffers required on typical cluster-based systems. The DMA engine and the Cray XT4 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.

The Cray SeaStar2 chip provides a service port that bridges between the separate management network and the Cray SeaStar2 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

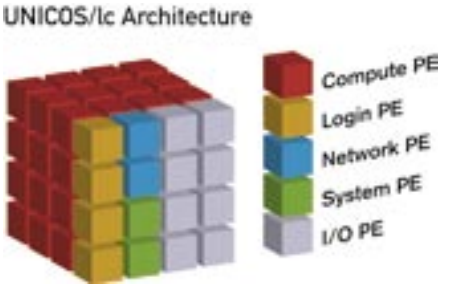
The Cray XT4 operating system UNICOS/lc™ is designed to run large complex applications and scale efficiently to over 120,000 processor cores. As in previous generation MPP systems from Cray, UNICOS/lc consists of two primary components—a microkernel for compute PEs and a full-featured operating system for the service PEs.

The Cray XT4 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. The special lightweight design means that there is virtually nothing that stands between a user's scalable application and the bare hardware. 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 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 XT4 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 XT4 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.

Scalable Programming Environment

Designed around open system standards, the Cray XT4 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 XT4 programming environment includes 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 AMD processor's native support for 32-bit and 64-bit applications and full x86-64 compatibility makes the Cray XT4 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 XT4 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 SeaStar2 chip to ensure uncompromised communications performance.

Cray Apprentice2™ performance analysis tools are also included with the Cray XT4 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 XT4 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.

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.

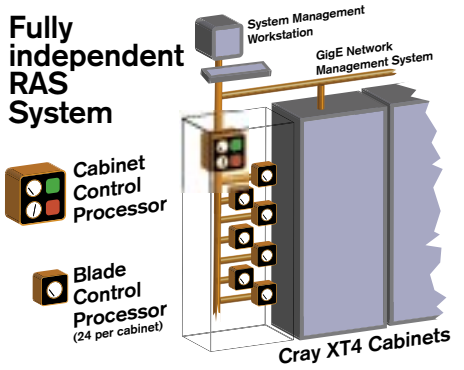
The Cray XT4 system is designed for high reliability. Redundancy is built in for critical components and single points of failure are minimized. For example, the system could lose an I/O PE, without losing the job that was using it. An AMD processor or local memory could fail and yet, jobs routed through that node can continue uninterrupted. The system boards contain no moving parts, further enhancing overall reliability.

The Cray XT4 processor and I/O boards use socketed components wherever possible. The SeaStar2 chip, the RAS processor module, the DIMMs, the voltage regulator modules (VRMs), and the AMD processors are all field replaceable and upgradeable. All components have redundant power, including redundant VRMs on all system blades.

Scalable I/O

The Cray XT4 I/O subsystem scales to meet the bandwidth needs of even the most data intensive applications. The I/O architecture consists of storage arrays 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 arrays. This highly scalable I/O architecture enables customers to configure the Cray XT4 with desired bandwidth by selecting the appropriate number of arrays and service PEs. It gives users and applications access to a high-performance file system with a global namespace.

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 XT4 combines the scalability of a microkernel based operating system with the I/O performance normally associated with large-scale SMP servers.

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