Cray X1E™ Supercomputer

Unrivalled Vector Processing and Scalability

for Extreme Performance

The Cray X1E system delivers maximum performance for the world's most challenging computational problems.

- Advanced processor architecture combines vectorization with hardware-enabled processor coupling for extreme performance

- Scalable system architecture with thousands of processors able to share each other's memory

- True single system image operating system minimizes system administration and simplifies application development

- Mature programming environment includes vectorizing compilers and support for a variety of highly optimized parallel programming models

The Cray X1E supercomputer combines the processor performance of traditional vector systems with the scalability of microprocessor-based architectures. High performance interconnect and memory subsystems allow the Cray X1E system to scale from 16 to 8,192 processors, delivering up to 147 TFLOPS in a single system. The Cray X1E supercomputer and its predecessor, the Cray X1™ supercomputer, are the first vector systems designed to scale to thousands of processors in a single system image.

The Cray X1E system builds on the success of the Cray X1 system by adding advanced, binary compatible, dual-core processors to deliver better performance and greater density. It offers a significant improvement in price/performance over its predecessor.

System Architecture

The Cray X1E distributed shared memory (DSM) system architecture gives each processor access to memory anywhere in the system through the Cray X1E system's high performance interconnect and memory subsystems. By removing memory contention and interconnect bottlenecks, the Cray X1E system's DSM architecture enables exceptional sustained application performance at extreme scale.

The CPU for the Cray X1E system is the multi-streaming processor (MSP). Each MSP provides 18 GFLOPS and 2 MB cache and can access memory at a rate of up to 34 GB per second.

The physical and scalable building block of the Cray X1E system is the compute module. A Cray X1E compute module contains four multi-chip modules (MCM), 16 or 32 GB of memory, routing logic and connections to other modules and to the I/O subsystem. The major advance between the Cray X1 and Cray X1E systems is in the implementation of the MSPs. Advanced ASIC technology on the Cray X1E system allows two MSPs to be implemented on one MCM, compared to one MSP in the Cray X1 MCM, for a total of eight MSPs per module. Doubling the processing density leads to significant price/performance improvements.

Cray X1E Compute Module
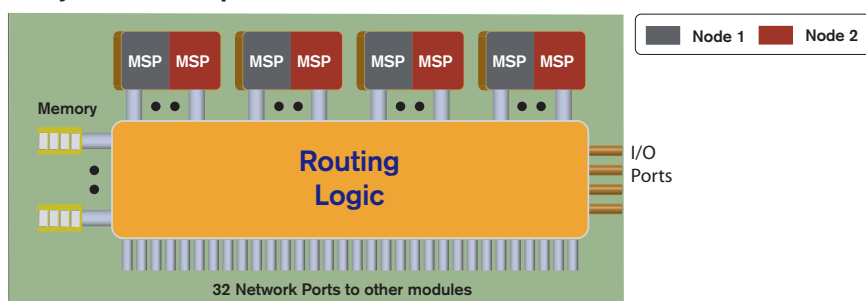
The eight MSPs on a module are organized into two logical nodes of four MSPs operating as symmetric multi-processors (SMPs). Each node uses half the module's memory–8 or 16 GB– as its local cached memory and shares the compute module's network and I/O ports.

With this DSM architecture, the MSPs are able to reference memory associated with other nodes or located on other modules using standard load and store instructions. This is made possible by a close coupling of Cray X1E system's high performance interconnect and memory subsystems, combined with scalable cache coherence and address translation protocols. Routing on the compute module handles processor and I/O memory references, improving performance and allowing the interconnect to scale naturally with the number of processors.

The DSM architecture provides the high performance important for application scalability. In addition, the architecture enables a true single system image operating system and a global I/O model–both simplify the developer's view of the system and reduce administration effort. The Cray X1E system includes embedded systems to handle compiles, builds, and other application development work, system networking functions, and administrator and operator management activities. By off-loading this work, more Cray X1E resources are made available for HPC applications.

Cray X1E System Highlights

High Performance CPU

The Cray X1E CPU uses vectorization and streaming to deliver peak performance of 18 GFLOPS per CPU. High memory bandwidth allows the processor to achieve a higher percentage of peak performance for applications than in most HPC systems.

Scalable Interconnect and Memory

The Cray X1E system is designed to scale beyond what any vector supercomputer has before. This scalability is made possible by extremely high-performance interconnect and memory subsystems. While the Cray X1E memory is physically distributed on individual modules, any MSP can logically share any memory. Using standard load and store instructions, an MSP can use a remote memory address to read or write memory located on a separate module. The system supports very high concurrency to allow applications to tolerate global network latencies.

True Single System Image OS

UNICOS/mp™, the Cray X1E operating system, is a true single system image operating system (OS). Regardless of system size, system administrators need only manage and configure a single OS – significantly reducing the time and effort needed for software upgrades and other tasks.

High Performance Application Support

The Cray X1E programming environments support multiple levels of parallel programming, with a wide choice of programming models. Optimizing Fortran, C, and C++ compilers perform automatic vectorization and streaming, and support OpenMP programming within a node, and Co-Array Fortran or UPC across the system. Optimized libraries are included for MPI and SHMEM, as well as for a large number of key scientific functions. To assist programmers in developing and optimizing applications, the TotalView® debugger and powerful performance analysis tools are available for the Cray X1E system.

Cray X1E Supercomputer Sample Configurations

	1 AC* Cabinet	1 LC* Cabinet	4 LC Cabinets	8 LC Cabinets
MSPs	32	128	512	1,024
Peak Performance	576 GFLOPS	2.3 TFLOPS	9.2 TFLOPS	18.4 TFLOPS
Maximum Memory	128 GB	512 GB	2 TB	4 TB
Aggregate Peak Memory Bandwidth	800 GB/s	3.2 TB/s	12.8 TB/s	25.6 TB/s

*Air Cooled (AC) Liquid Cooled (LC)

The Cray X1E system is designed for maximum performance. Key features include:

- **High performance CPU**
- **Scalable interconnect and memory subsystems**
- **Scalable I/O**
- **True single system image operating system**
- **Programming environments designed to deliver maximum performance**

High Performance CPU

The Cray X1E CPU is a multi-streaming processor (MSP) that uses the instruction set introduced with the Cray X1 system. The instruction set architecture (ISA) includes a complete set of powerful vector arithmetic and logical operations supplemented by a full set of scalar and memory reference instructions. With large numbers of registers, support for both 32- and 64-bit computations, and scalable synchronization features, the Cray X1E system can deliver excellent performance to a wide range of applications.

Multi Chip Module (MCM)
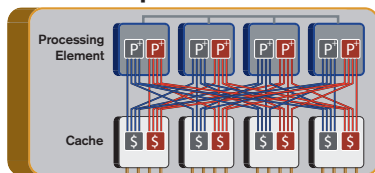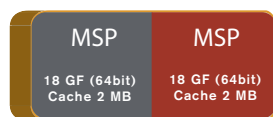
Among the unique features in the ISA is support for streaming. The MSP is actually made up of four single-streaming vector processors (SSPs). These SSPs have special hardware features that allow them to be operated collectively with a second, low-latency level of parallelism called streaming. Compilers can recognize and schedule standard Fortran, C, or C++ code to take advantage of both vectorization and stream parallelism.

18
GFLOPS per CPU

The combination of these features results in a very high performance CPU. The Cray X1E MSP delivers peak performance of 18

GFLOPS, an increase of over 40 percent from the peak performance of the Cray X1 MSP. This is achieved while maintaining full binary compatibility—applications developed for the Cray X1 system can run, as is, on the Cray X1E system.

Scalable Interconnect and Memory

Excellent CPU performance must be balanced by memory and interconnect performance to ensure that CPUs are fully utilized and not left waiting for data. The Cray X1E system is designed with tightly-coupled memory and interconnect subsystems that deliver excellent, scalable performance.

34
GB/s memory bandwidth per CPU

Unlike most other systems, the Cray X1E system has a DSM architecture that integrates the memory and interconnect subsystems to allow memory references to be efficiently routed directly to the appropriate local or remote memory. Memory is physically distributed on individual modules, but all memory is directly addressable to and accessible by any MSP in the system through the use of load and store instructions.

While each node functions like a traditional SMP node, its processors can also directly address memory on any other node—remote memory accesses go over the interconnect to request processors, bypassing the local cache. This mechanism is more scalable than traditional shared memory and it provides very low latencies and unprecedented inter-processor bandwidths. Each processor can have up to 2,048 outstanding memory references, allowing applications to tolerate global network latencies.

The interconnect supporting these remote references includes routing logic and network ports on each compute module, and separate routing modules. A novel design effectively implements 16 independent 2D torus topologies within the interconnect, each called a slice.

Processor or I/O memory references are first handled by the routing logic of the appropriate slice. If the address is local to the node, the routing logic accesses the node's local memory. If the memory address is on a remote node, the request is routed using the network, and the routing logic on the remote node handles the request as a local reference. Each slice of the machine independently handles all memory accesses and routing for addresses that map to that slice.

Each compute module accesses the network through a total of 32 network ports, two per slice, each of which supports 1.6 GB/second peak per direction. For large systems, half of these ports are connected to router modules which are connected to other compute or router modules to build up the interconnect. For small systems with up to four compute modules, the network ports are connected to each other.

The scalability of this interconnect is further enhanced by the following features:

Scalable cache coherence protocol

Cache coherence is critical for a shared memory, but many SMP protocols scale very poorly. The Cray X1E system uses a protocol designed for scalability without the overhead associated with SMP coherence protocols. Local memory references are cached, remote memory references are not, reducing the overhead normally associated with SMP coherence protocols.

Scalable address translation

The Cray X1E system translates addresses at the destination node, requiring each node only keep track of translation information for its local memory. This DSM architecture scales to very large configurations without impacting performance.

Efficient hardware synchronization

Parallel processes regularly need to synchronize with each other. With a combination of shared memory and atomic memory operations, these types of operations can be performed on Cray X1E systems with a minimum of delay.

Scalable I/O

A key requirement for a truly scalable system is scalable I/O. The Cray X1E I/O subsystem has been designed to balance computing power with I/O performance and maintain this balance as the system grows. Each compute module has four I/O ports that connect to the Cray X1E I/O cabinet, and from there to storage devices or the Cray Network Server (CNS) which handles external network connections. Since each port provides peak channel bandwidth of 1.2 GB per second per direction, each additional compute module can add an additional 4.8 GB per second of peak I/O bandwidth to the system. These I/O ports are connected to the memory and interconnect subsystems, making each I/O channel accessible to all processors and allowing I/O operations to read from or write to memory anywhere in the system.

True Single System Image OS

The operating system for the Cray X1E supercomputer, UNICOS/mp, is a true single system image operating system. UNICOS/mp takes advantage of the DSM architecture of the system to simplify administration, to simplify the I/O architecture, and to provide a single login.

> ## System-wide services run on a Single Node
>
> One system to install
> One system to upgrade
> One system to configure
> One system to manage

With UNICOS/mp, each node on the system runs a kernel responsible for managing resources local to the node, but system-wide services, such as scheduling and file system management, are run on a single node designated as the system node. Multiple system nodes can be configured for very large systems. The kernels on each node communicate using low-level message-passing protocols that exploit the DSM architecture for high performance.

The result is a system that allows administrators to manage a Cray X1E system, regardless of size, as if it were a single node. One system to install, one system to upgrade, one system to configure, one system to manage.

In addition to being a true single system image OS, UNICOS/mp offers a number of additional capabilities that allow applications and customers to exploit the power of the Cray X1E system, including:

- Sophisticated application scheduling, with algorithms to allow parallel applications, single-processor user commands, and operating system processes to share the Cray X1E system. Included is the option for programmers to use one SSP (instead of one MSP) per MPI process or OpenMP thread, which can result in better performance for some applications.

- System partitioning allows a customer to divide a system into two or more separate systems, each with an independent operating system.

- System-level checkpoint/restart improves the resiliency of long-running applications to system or environmental problems and allows customers to make major changes to the operational environment.

- Two powerful file systems are supported. The 64-bit journaling XFS file system for direct-attached disk, and the StorNext™ File System from Advanced Digital Information Corporation (ADIC™) is used as a global shared file system.

- Batch and interactive job processing is supported by third-party products such as PBS Pro™ from Altair™ Engineering and Platform LSF from Platform™ Computing. Special interfaces allow these products to coordinate their activity with that of the UNICOS/mp scheduler, resulting in greater system utilization.

- The Cray Workstation Server (CWS) provides management and maintenance for the Cray X1E system including disks and peripheral systems.

UNICOS/mp provides standard features that customers expect from high-performance operating systems today, including standard networking protocols, accounting, UNIX commands, security (including ACLs), and system monitoring and management tools.

High Performance Application Support

To allow application developers to exploit the powerful features of the Cray X1E system, Cray offers a full set of compilers, libraries, and tools. This software is run on the Cray Programming Environment Server (CPES), off-loading development workload from the primary compute processors.

Compilers

Cray X1E compilers are proven products that provide automatic vectorization and streaming capabilities while adhering to a wide range of formal and de facto industry standards.

All the compilers support the OpenMP interface for SMP-based parallel programming. The Cray Fortran compiler supports the Co-Array Fortran interface for distributed memory programming, and the Cray C and C++ compilers support the similar UPC (Unified Parallel C) programming model.

All the compilers are designed to deliver maximum performance from the Cray X1E system. Vectorization and streaming are performed automatically by the compiler, and the developer can provide additional hints through compiler directives to further improve performance. A sophisticated set of optimizations and scheduling is performed to deliver even greater performance to the application.
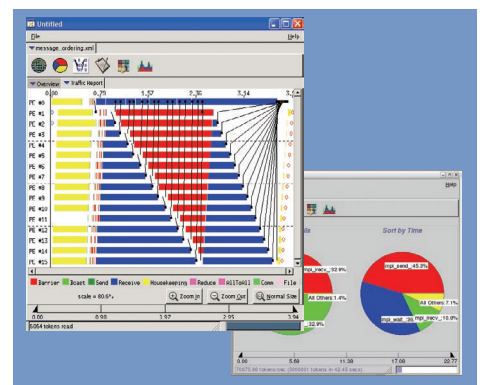
Libraries

Cray provides a full set of subroutine libraries that offer additional capabilities and performance. These libraries include:

- High performance scientific library (LibSci), which includes industry standard libraries such as BLAS, LAPACK, BLACS, and ScaLAPACK.

- Optimized MPI library, providing support for the full MPI-1 interface for distributed parallel programming, as well as most features from the MPI-2 interface.

- Optimized SHMEM library, implementing the efficient one-way message passing introduced with the Cray T3D™ and Cray T3E™ MPP systems.

- Language support libraries for Fortran, C, and C++. This includes Cray's flexible file I/O (FFIO) system for source-independent I/O optimizations.

- System libraries providing standard UNIX functions.

Tools

Along with compilers and libraries, Cray has made some exceptional tools available—such as the scalable Etnus TotalView® debugger—allowing developers to debug and optimize their applications.

CrayPat™ and Cray Apprentice[2]™ performance analysis tools together provide extremely powerful features to allow application developers to optimize application performance. CrayPat uses several types of hardware- and software-based performance analysis methods to gather data on application performance. Cray Apprentice[2] is a graphical tool that presents information in formats such as call graphs and time lines to help developers better understand and fix performance bottlenecks in their applications.

CPU	64-bit Cray X1E Multistreaming Processor (MSP); 8 per compute module
	8 vector pipes per MSP
	Scalar operations overlapped with vector operations
	IEEE floating point compatible
	Bit matrix multiply, pop count, and integer add/subtract with carry
Cache	2MB cache per MSP
FLOPS	18 GFLOPS theoretical peak performance per MSP (1.13 GHz vector clock speed)
SMP	4-way SMP node
Main Memory	16 GB or 32 GB RDRAM memory per compute module
Memory Bandwidth	200 GB/s per compute module peak bandwidth; 34 GB/s per processor peak bandwidth
Interconnect	Custom high performance interconnect providing distributed shared memory access through entire system
	16 parallel 2D torus networks
	51 GB/s per compute module peak bandwidth
	5 microsecond MPI latency between processors
External I/O	Peak bandwidth of 4.8 GB/s through dedicated I/O channels from each compute module
	Shared memory allows any processor to perform I/O to any I/O channel
	Separate Cray Network Server (CNS) supports network I/O
File System	XFS (direct-attached disk)
	ADIC StorNext File System (SAN-attached disk)
System Administration	Cray Workstation Server (CWS) provides management and maintenance for Cray X1E system, including disks and peripheral systems
	Cray Programming Environment Server (CPES) off-loads programming environment workload
	PBS Pro and LSF workload management systems
	Optional partitioning of system into independent units with separate I/O infrastructures
Reliability Features	System is resilient to hardware or software failures in application processors
	System can be run in degraded mode if necessary due to hardware failures
Operating System	UNICOS/mp
Parallel Programming	MPI 1.2, SHMEM, OpenMP, Co-Array Fortran, UPC
Cray Fortran Compiler	Fully adheres to Fortran 95 standard (ISO/IEC 1539-1:1997 Part 1); supports selected features from proposed Fortran 2003 standard
Cray C & C++ Compilers	Adhere to industry standards for C (ISO/IEC 9899:1999 (C99)) and C++ (ISO/IEC 14882:1998)

	Liquid-cooled (LC) cabinet	**Air-cooled (AC) cabinet**
Cabinets	up to 16 compute modules (128 processors)	up to 4 compute modules (32 processors)
Maximum Configuration	Up to 64 LC cabinets (8,192 processors)	Up to 4 AC cabinets (128 processors)
Standard Configuration*	Up to 8 LC cabinets (1,024 processors)	1 AC cabinet (32 processors)
Power (cabinet)	65 kW, 200-208 VAC	17 kW, 200-208 VAC
Footprint (cabinet)	50.75 in. x 103 in. (1.3 m x 2.6 m)	35.5 in. x 59.75 in. (.9 m x 1.5 m)
Weight (cabinet)	5,754 lbs. (2,610 kg)	1,973 lbs. (895 kg)

* Configurations exceeding these limits, up to the design parameters of the system, are available by special order.

CRAY

The Supercomputer Company