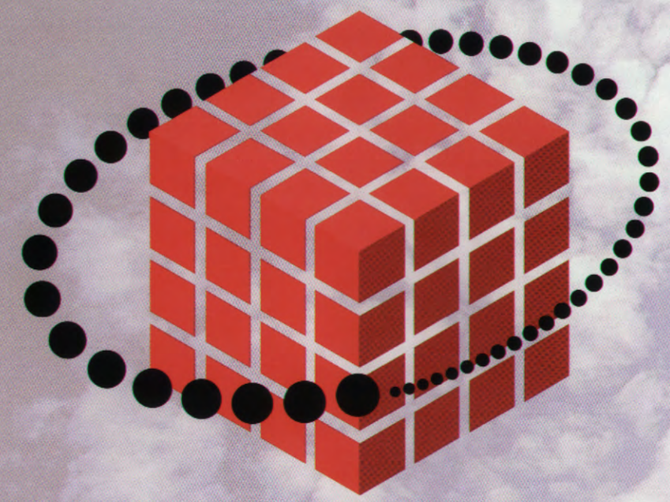
Cray Research, Inc.

MPP Technology Preview

CRAY RESEARCH, INC.



3-D torus interconnect network

Cray Research MPP systems feature a 3-D torus interconnect network to transfer system data and control information between nodes. The nodes in each dimension of the network form a torus ring, which allows communications to pass from one node, through all of the nodes in the same dimension, and back to the original node in a circular fashion. This design minimizes network distances and provides high-bandwidth, low-latency communications.

The world's first production-oriented MPP system

Cray Research is using its unmatched supercomputing experience to build the world's first production-oriented massively parallel system. Based on high-speed supercomputer technology combined with a high-performance software base, our approach to massive parallelism will deliver an unprecedented level of accessible performance to a broad spectrum of users.

Cray Research MPP systems are based on the same balanced approach to supercomputing that gives our parallel-vector systems the highest level of sustained performance in the industry.

By providing an optimal level of computational capacity, high-speed synchronization, and high-bandwidth low-latency communications, Cray Research MPP systems are designed to deliver the highest level of sustained MPP performance available.

Realizing MPP potential using supercomputer technology

Massively parallel systems offer exciting potential performance on highly parallel problems, but experience to date has yielded less than spectacular sustained-peak performance ratios. Clearly, realizing the true potential of MPP requires more than just connecting large numbers of microprocessors; it requires a balanced approach using true supercomputing technology.

Today's RISC processors lack the communication, memory, and synchronization features necessary to build efficient MPP systems. Cray Research circumvents these shortcomings by surrounding the RISC chips with supercomputer-class communication and synchronization hardware. At the same time, our MPP systems take advantage of our proven supercomputer packaging and cooling techniques. In this way, Cray Research is transforming thousands of independent RISC processors into a supercomputer-class MPP system—one that can address terabytes of memory, minimize communication overhead, and provide excellent lightweight synchronization.

Far left, F-18 flowfield visualized using surface flow patterns, off-surface particle traces, and helicity density contours. Left, atmospheric modeling for weather prediction and climate research.

The road to sustained teraflops performance

Cray Research's objective is to provide affordable, sustained teraflops-level performance for a wide range of complete scientific and engineering applications. We expect to achieve this objective within five years through the introduction of a series of compatible MPP systems with increasing performance and capacity.

Cray Research's first generation MPP system, scheduled for availability in mid-1993, is scalable to over 300 GFLOPS peak performance in a 2048-processor configuration. The second generation MPP systems will be based on a newer generation of RISC microprocessor technology that will allow 1 TFLOPS (peak) systems to be available mid-decade. The following generation of this technology is expected to be two to four times more powerful per processor, making possible a system with sustained 1 TFLOPS performance by 1997.

Phase	Performance rating	Time frame
1	300 GFLOPS peak	1993
2	1 TFLOPS peak	mid-decade
3	1 TFLOPS sustained	1997

Application focus

To ensure that customers have a rich selection of MPP application software, Cray Research MPP systems are designed to allow existing MPP codes to be ported easily, typically with improved performance. The MPP systems support message-passing, data-parallel, and HPF-like programming models.

Our applications efforts focus on the following key areas:

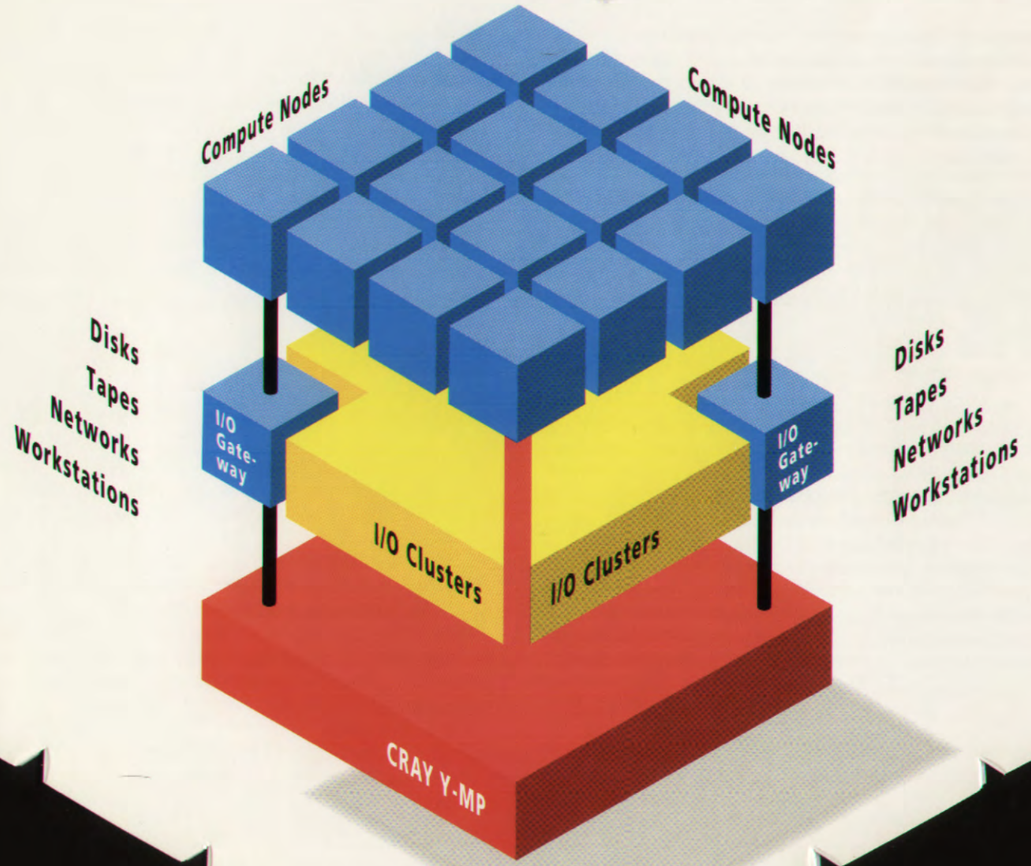
- Seismic data processing for petroleum exploration
- Atmospheric modeling for weather prediction and climate research
- Computational fluid dynamics and structural analysis for the aerospace and automotive industries
- Computational chemistry for drug design and materials science applications
- Computational electromagnetics

Heterogeneous architecture for real workloads

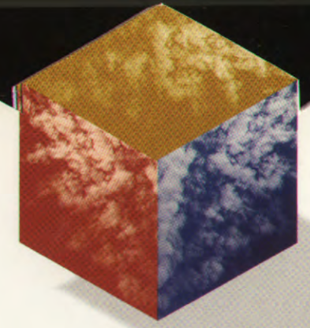
MPP systems are designed to run highly parallel problems. However, most real-world applications typically comprise a mix of scalar, vector, parallel, and highly parallel codes. To provide the highest levels of applications performance, Cray Research MPP systems will be coupled closely to the parallel-vector-scalar architecture of the industry-leading CRAY Y-MP supercomputer.

A key advantage of the Cray Research heterogeneous architecture is that it provides flexibility to meet the needs of individual user workloads. Customers can choose from a wide variety of system solutions with different mixes of parallel-vector and MPP computational capabilities.

Our proven UNICOS system software offers our MPP users a complete supercomputing environment with parallel, vector, and scalar processing integrated into a single system. The parallel-vector component offers production quality I/O, proven software tools, high-speed networking, and the ability to distribute a single application between a vector-scalar environment and a closely coupled highly parallel environment. This total computational capability simply does not exist on any MPP or clustered workstation system in the market today.



Cray Research MPP heterogeneous architecture



MPP macroarchitecture

The heart of the Cray Research MPP design is a balanced, scalable macroarchitecture that combines the fastest available microprocessors with a high-bandwidth, low-latency interconnect network that is an order of magnitude faster than networks in currently available MPP systems. The macroarchitecture has the following characteristics:

- *Multiprocessor multiple-instruction, multiple-data (MIMD) architecture.* The MIMD architecture allows the independent, parallel execution of different program threads, and even entirely distinct programs, on different processors. The Cray Research MPP system is also capable of emulating single-instruction, multiple-data architectures (also called SIMD or data parallel architectures) with high efficiency through the use of special synchronization hardware.

- *High-speed interprocessor communications.* The system processing elements (PEs) are connected by a very fast interconnection network used to access and redistribute global data. Using the same high-performance switch technology as the CRAY Y-MP processor-to-memory interface, the network operates at the same clock speed as the PEs, 150 MHz, which allows extremely fast remote memory access.

- *3-D torus interconnect topology.* The interconnect network is three-dimensional, which increases bandwidth and minimizes network distances. The 3-D torus uses high-performance switch nodes that allow interprocessor communications to occur without interrupting the PEs. Each switch node can operate bidirectionally in each dimension.

- *Globally addressable, physically distributed memory.* Because the memory is logically shared, any PE can access the memory of any other processing element without explicit message passing or involving the remote PE. As a result, the system can be scaled to address terabytes of memory. This design provides ease of use, high memory bandwidth, and low memory latency.

- *Latency hiding.* To help sustain high performance, special communications hardware allows data in remote PEs to be moved into a local PE before it is needed.

- *Fast synchronization.* The hardware provides a rich set of synchronization primitives for both SIMD and MIMD and data-driven programming styles.

- *High bandwidth, parallel I/O.* To balance high computational performance with high-performance I/O, Cray Research MPP products incorporate Model E I/O subsystems through multiple high-speed channels (200 Mbytes per second in each direction, per channel). This I/O technology scales with the number of processors and is capable of sustaining multiple gigabytes of I/O transfers. Our MPP systems support a wide range of I/O and networking devices including fast disks, IBM 3490 tapes, D2 tapes, HIPPI, and FDDI.

- *Scalability.* The interconnect design allows customers to scale up easily from hundreds to thousands of PEs.

- *Easy porting to future systems.* To protect customer application software investments, programs written for our initial MPP systems will port easily to our future MPP systems.

- *Reliability.* Software-configurable redundant hardware is included so that processing can continue, without hardware maintenance, should a PE fail.

MPP PERFORMANCE: EXPERIENCE COUNTS

Since Cray Research introduced its first system in 1976, two fundamental factors have determined supercomputer performance: fast processors and fast processor-to-memory connections. With the introduction of our parallel supercomputers in 1982, a third important factor was added: fast interprocessor synchronization. The balance among these three factors will remain key to delivered performance from supercomputers in the 90s, including MPPs.

To continue its leadership of the supercomputing industry, Cray Research will continue to excel in the following areas with CRAY Y-MP C90 and MPP systems:

- ***Fast processors.*** **Cray Research will use the fastest microprocessors available in each of the three phases of its MPP plan. The first phase will use the 150 MFLOPS (peak) Alpha chip from DEC.**

- ***Fast processor-to-memory connections.*** **Cray Research is using proven, high-speed CRAY Y-MP C90 interconnect technology to provide the fastest MPP interprocessor connections in the industry. Cray Research's MPP has higher global memory bandwidth and lower latency than any other MPP system. Supercomputing hardware experience counts.**

- ***Fast interprocessor synchronization.*** **Cray Research's advanced CRAY Y-MP C90 technology also yields the fastest MPP synchronization in the industry. Parallel processing experience counts.**

Other factors are also important for a production-level MPP supercomputer:

- ***Ease of use.*** **Cray Research's MPP system uses productive and efficient MPP languages. A key part of programming an MPP is communicating efficiently between the physically distributed processors. Cray Research supplies a combination of support for message passing, data parallel, and work sharing styles of MPP programming. These styles give programmers the tools to write code with the minimum effort, without sacrificing the ability to optimize time-critical codes. Our existing modular compiling systems will allow us to introduce and optimize these new styles quickly. Supercomputer compiler experience counts.**

- ***Applicability.*** **Cray Research will widen the range of applications on MPPs. To ensure the highest levels of sustained performance, Cray Research will mask much of the inherent performance penalties of the distributed memory architecture. We will do this with a combination of our excellent compilers, optimized scientific libraries, and fast global memory (high bandwidth, low latency). This ability to communicate between processors with quick short messages will allow levels of fine-grain parallelism that are not feasible on other MPPs. Supercomputing application experience counts.**

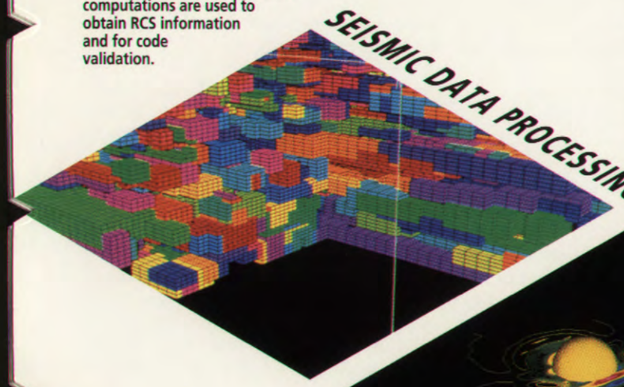
- ***Heterogeneous computing.*** **Some codes run well on distributed memory machines (MPPs) and some run better on general purpose machines (uniform-memory CRAY Y-MP C90 systems). For this reason, Cray Research provides a closely coupled system of general purpose (CRAY Y-MP C90) and distributed memory (MPP) parallel supercomputers. Production supercomputing experience counts.**

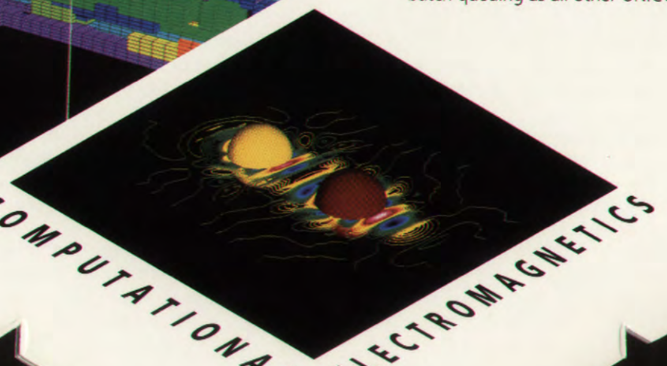
- ***Scalable and high bandwidth I/O.*** **To balance high computational performance with high-performance I/O, Cray Research MPP products incorporate Model E I/O subsystems. This I/O technology scales with the number of processors and is capable of sustaining multiple gigabytes of I/O transfers. Our MPPs support a wide range of I/O and networking devices including fast disks, IBM 3490 tapes, D2 tapes, HIPPI, and FDDI.**

- ***Scalability and reliability.*** **Cray Research's MPP systems are available in a wide range of configurations and include redundant hardware in Cray Research's highly reliable packaging. Supercomputing packaging experience counts.**

Below, petroleum reservoir simulation used for recovery profile predictions.

Below right, scattered electrical fields around a double sphere at one instant of time. These field computations are used to obtain RCS information and for code validation.





Flexible microarchitecture

The MPP macroarchitecture uses a variable microarchitecture that will allow Cray Research to incorporate the fastest microprocessors available for each generation of MPP systems. Our first MPP microarchitecture uses the Alpha microprocessor from Digital Equipment Corporation, a reduced instruction set computer (RISC) chip capable of 150 MFLOPS performance. The Alpha microprocessor is cache-based with pipelined functional units, issues multiple instructions per cycle, and supports IEEE standard 64-bit floating-point arithmetic.

Cray Research's MPP operating system

The MPP operating system is a microkernel-based implementation of Cray Research's UNICOS operating system. Each PE runs a minimal microkernel to handle frequently used functions, such as interprocessor communications (IPCs) and memory management. Higher level UNIX system calls are sent by way of an IPC mechanism to a closely coupled UNIX agent running on the CRAY Y-MP host system. This arrangement provides maximum UNIX functionality for our initial MPP offering. As a result, programs running on our MPP systems can access the same file systems, network protocols, and batch queuing as all other UNICOS applications.

Cray Research's MPP programming environment

The Cray Research MPP programming environment provides tools that programmers can use to minimize communication overhead and maximize the execution rates of their programs. MPP systems require careful programming to handle the communication necessary between the distributed memories and processing elements. Our programming languages provide the additional constructs necessary to program this communication efficiently. Our goal is to make this communication implicit in the algorithms, minimizing the burden on programmers. Our programming model also allows programmers to code the communication explicitly when they need full control of the communication. This mix of explicit and implicit communication techniques gives programmers the tools to write code with minimum effort without sacrificing the ability to optimize the time-critical sections.

Our MPP system supports several programming styles: message passing, data parallel, and work sharing. These styles are available in the following languages:

- *Message passing (explicit communications only).* In the message-passing model, programmers must explicitly send messages to request information from other PEs. Our message-passing library is based on the Parallel Virtual Machine (PVM 3.0) developed at the Oak Ridge National Laboratory. It will enable easy ports of existing message-passing codes from other MPP systems. Cray Research views message passing as a key programming method on MPP systems and is giving it full support, including special fast message passing hardware. We expect message-passing codes written for other MPP systems to port easily to our MPP system, typically with improved performance. Cray Research's PVM supports Fortran, C, and C++.
- *High Performance Fortran (HPF) model (mostly implicit communications).* Cray Research is a strong supporter of the emerging HPF programming standard and is an active participant in the HPF Forum. HPF uses Fortran 90 data parallel constructs to allow programmers to avoid explicitly coding the communication. HPF uses Fortran-D-like data distribution constructs to divide the arrays among the distributed memories. The HPF Forum is finishing the specification of HPF. Cray Research is evaluating this preliminary specification. We are working with our customers to decide how and when to implement HPF efficiently.
- *Cray Research's MPP Fortran model (implicit or explicit communications).* MPP Fortran is similar to HPF, with more options for code optimization. It allows programmers to write HPF-like code using Fortran 90 data parallel programming, thus avoiding explicit communications. Programmers can distribute data among the processors, much like in HPF. In addition, MPP Fortran includes work sharing features that allow work to be shared among the processors using standard FORTRAN 77 DO loops. The work sharing model supports dynamic distribution of work for load balancing. MPP Fortran includes enhancements that give programmers greater control of communication and synchronization, allowing them to code the communication explicitly, if necessary.

Programmers can mix the data parallel, work sharing, and message-passing styles of programming within an application. Because our MPP system offers choices rather than dictating one particular model, users can decide which approach is best for their application. This variety of models will also facilitate the porting of existing programs from other massively parallel systems and from Cray Research parallel-vector systems.

Powerful programming tools

Cray Research will provide a rich set of programming tools to aid in tuning and debugging MPP codes. These tools work in the UNICOS 7.0 integrated programming window environment, which integrates the compilers, editors, debuggers, browsers, analyzers, and emulators to provide a highly productive MPP programming environment.

- *TotalView* is a powerful debugger designed specifically for MPPs. Cray Research purchased the rights to TotalView from Bolt, Beranek, and Newman (BBN) and has customized and extended Totalview to be the best MPP debugger in the industry.
- *Apprentice* is a performance analyzer tool written by Cray Research, similar to Cray Research's ATExpert, but with features specific to programming distributed memory MPP systems. Apprentice will provide the profiling and performance monitoring information crucial to optimizing MPP codes.
- *The MPP emulator* is a performance analysis tool designed to help customers develop programs for Cray Research's MPP system before its release. The emulator executes code written in Cray Research's MPP Fortran or the PVM message-passing programming model and provides feedback to help developers write more efficient parallel code. The emulator allows users to study data layout, data locality, and data reference patterns.

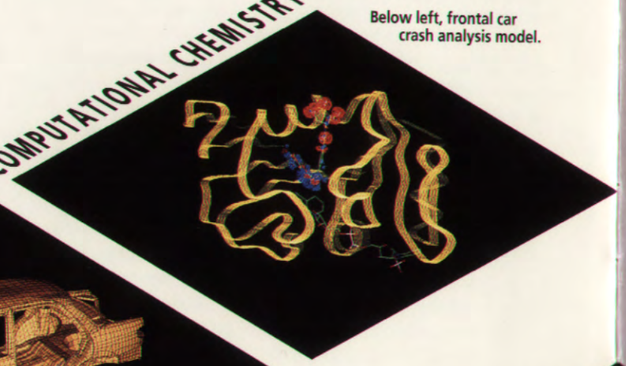
Setting new standards for supercomputing excellence

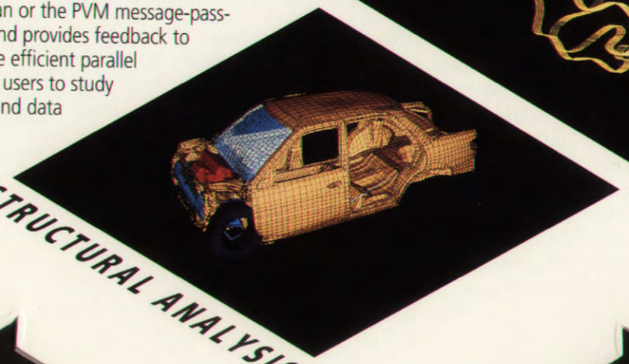
The Cray Research approach to MPP technology is based on over 20 years of supercomputing leadership. By addressing the real hardware and software challenges of MPP, Cray Research will provide a new generation of science and engineering solutions for the next century.

Below, charge polarization in a folate molecule upon binding to the enzyme dihydrofolate reductase (DHFR).

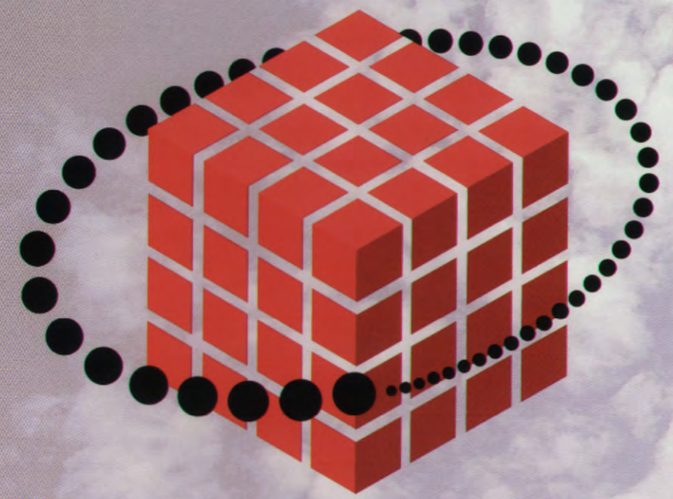
Below left, frontal car crash analysis model.

COMPUTATIONAL CHEMISTRY





STRUCTURAL ANALYSIS



Credits: Petroleum reservoir slide courtesy of British Petroleum. Weather forecast simulation courtesy of European Centre for Medium-Range Weather Forecasts. F-18 flowfield solution obtained by Farhad Ghaffari. Car crash model courtesy of Kia Motors.

MCPF-2421192

CRAY RESEARCH, INC.
655-A Lone Oak Drive
Eagan, MN 55121
(612) 683-3801