The CRAY T3E System

Steve Reinhardt, Cray Research, Inc., 655-F Lone Oak Drive, Eagan, Minnesota 55121

ABSTRACT: *Pioneers began using scalable computing systems over 10 years ago. Slowly the potential audience for scalable systems has grown as programming methods and systems have matured. The CRAY T3D system enabled production users to use scalable systems. Now, the CRAY T3E system is pushing beyond the CRAY T3D base by delivering the most advanced scalable technology at lower price points, and to users of industrial applications.*

1 Introduction

In 1988, MPPs were suitable for a small group of dedicated pioneers, armed with hardware manuals and microcode templates. By 1991, MPPs were still suitable for the same small group of users, though an infusion of government money sometimes made the machines appear otherwise. In 1994, with the delivery of large CRAY T3D systems, the potential audience for MPPs grew dramatically, as the CRAY T3D was the first stable production platform for parallel programs. This enabled many organizations to commit significant money to scalable computing as part of their production workload. Now, in 1996, we no longer speak of "MPP", but rather of "scalable computing." And with the delivery of the CRAY T3E, the potential audience for scalable computing grows significantly once again, as the CRAY T3E will bring scalable computing to users of industrial applications and to users of smaller systems.

2 Motivation for Scalable Computing

Popular wisdom has it that SMP systems are easier to program than scalable systems; furthermore, the growth of single processor speeds (both high-end vector processors and superscalar, cache-based processors) makes these SMP systems extremely powerful in their own right. Why would a programmer or user with access to these computing engines want to endure the (alleged) pain of working with scalable systems?

Consider the following figure, which illustrates the growth in the number of processors which can be purchased for $10M. (This price point has been chosen to make the early data points relevant; any price point will have growth on the same slope.) Note that the vertical axis is a log scale. Within Cray's vector SMP systems, the number of PEs you can buy for $10M has increased by a factor of 5 over the last 10 years. Including Cray's scalable systems, the growth has been a factor of 125 over the same period. This growth is fueled by the smaller device features which are possible by developments in device lithography. Semiconductor industry documents [SIA94] anticipate that this trend will continue for at least the next decade, and thus we can expect PEs per dollar to continue exponential growth for this period.

Of course, based on this simple hardware picture, scalable systems would have taken over the high performance computing market already. This has not happened because the applications which people use to solve problems on computers represent a significant inertia of their own. These applications have typically moved more slowly than the hardware systems to exploit large numbers of PEs. Consider Figure 2, which shows the growth in parallelism of the leading application within a discipline. The applications in some disciplines, such as structural analysis and chemical processing, have moved slowly to exploit large numbers of processors, and hence users of those

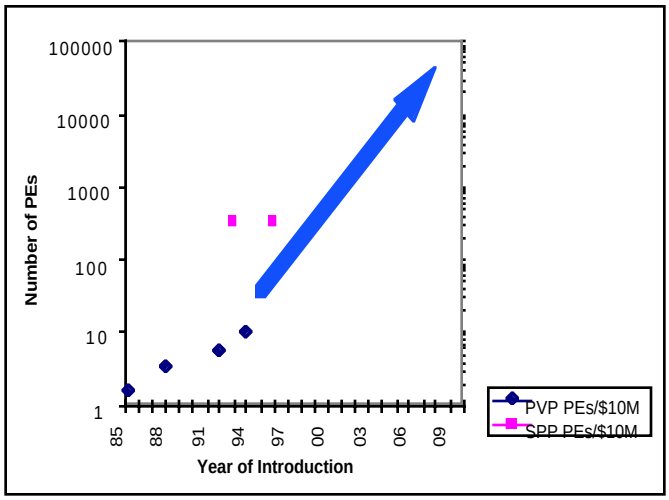

Figure 1.

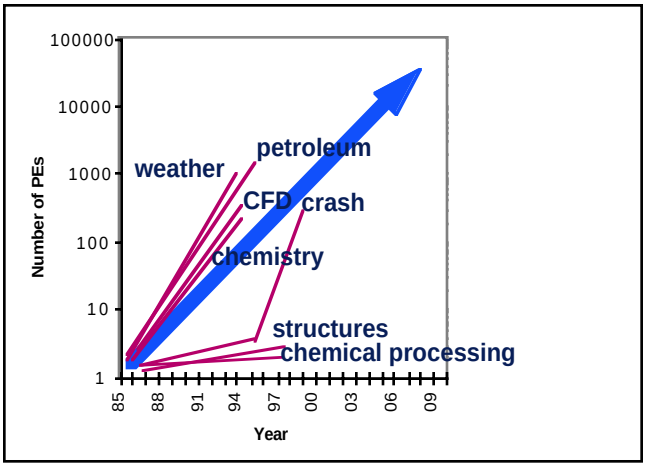

Figure 2.

applications depend on systems with the highest single processor performance. Other disciplines, such as weather and seismic processing, have moved quickly to scalable systems. Vendors of many industrial codes are working on scalable versions of those codes, and we are now seeing the fruits of that work with the availability of applications for the CRAY T3E.

3 Macro-/microarchitecture

As we undertook the CRAY T3D project in 1990, we were uncertain about the best architecture for scalable systems (recall that people were still debating SIMD v. MIMD). Because of this uncertainty and the inertia of applications, we chose a two-level view of system architecture for our scalable systems.

Programmers use the *macroarchitecture*, which must evolve slowly to preserve applications development investment. The Cray scalable macroarchitecture includes as primary features:

- physically distributed, globally addressable memory
- MIMD operation with efficient emulation of SIMD constructs
- low-latency, high-bandwidth interconnection network

By contrast, the *microarchitecture* can change from generation to generation to allow system designers to incorporate the latest developments in microprocessor and communication mechanisms. The microarchitecture is by its nature tactical. We have chosen for each generation of Cray scalable systems one of the fastest processors available and have conformed the rest of the microarchitecture to that processor. [Scott96]

4 Lessons from the CRAY T3D

We now know things about scalable computing as a direct result of the experience with the CRAY T3D.

Applications in numerous disciplines scale efficiently to 1000 processors. Prior to the T3D, kernels of applications had been scaled well, but entire applications had not, typically due to limitations in the interconnection network of earlier scalable systems. Fifteen significant applications ran on a 1024-PE CRAY T3D system in the fall of 1994 with excellent scaling.

The low latency network permits new applications. The ability to communicate between processors of the CRAY T3D at small granularity, with good performance, is a qualitative difference from earlier scalable systems. An example is a particle-in-cell (PIC) code from LANL, which was converted from a serial version to run on the CRAY T3D, and did so efficiently. This code had not previously run effectively on a scalable system.

A stable system (particularly OS and hardware) permits focus on the difficult application task of structuring problems to run on distributed memory. Because of the stability of the CRAY T3D, programmers could devote their attention to many of the applications which will now appear in production form on the CRAY T3E.

5 CRAY T3E Value Proposition

Cray's basic value proposition is "We solve your big problems fast, reliably, affordably, and conveniently." The CRAY T3E system incorporates advances in each of these areas which collectively deliver outstanding value for high performance computer users.

5.1 Fast

The interconnection network is a three-dimensional torus, as was that of the CRAY T3D; adaptive routing has been added to avoid transient hot-spots which can occur. The low latency has been preserved; a program running on all PEs of a 512-PE system, simultaneously making random accesses to memory of other PEs, will see average latency under two microseconds. The payload bandwidth per link has increased significantly to 480 MB/s. This gives network bandwidth per PE about 6 times that of the CRAY T3D, as a network node now supports a single PE rather than the pair of PEs on the CRAY T3D. The *libsma* ("get/put") interface allows programmers to use the network in a simple, high performance way.

The processor is the DECchip 21164 or "EV5" processor, in its 300 MHz version [DEC95]. The chip delivers 600 MFLOPS and 1.2GB/s of off-chip memory bandwidth as peak performance, and a LINPACK rating of 405MFLOPS. We expect to incorporate faster versions of the chip as they become available.

Local memory streams enhance the performance of the single PE. Intelligent hardware in the memory controller recognizes a vector of small-stride memory references and pre-fetches data from DRAM to buffers so that it is available when the processor requests it. This uses the off-chip processor bandwidth as effectively as possible.

Large physical memories accelerate the performance of some problems. Memories up to 2GB per processor are available, both for direct program use and for use as large I/O caches, much as SSD memory is used on Cray vector SMP systems.

Cray's mature optimizing compilers generate efficient code for the EV5 microarchitecture. Additionally, specific optimizations restructure loops to exploit the stream buffers.

Scientific libraries form the building blocks of applications for many users, performing common operations at near-optimal speeds. The CRAY T3E benefits enormously from the work done on the CRAY T3D for parallel scientific libraries, which will all be available for production shipments. ScaLAPACK for dense systems, 2D and 3D FFTs (both in-core and out-of-core), and selected solvers will be available in parallel form. A small set of routines has been hand-optimized; the higher-level routines build on these core routines.

Input/output performance reaches levels unknown in a production computing system. We designed the UNICOS/mk operating system to deliver scalable I/O services, and have implemented innovative solutions to some typical I/O bottlenecks [Broner96]. The GigaRing I/O channel gives excellent speed from even a single channel (960MB/s to a CRAY T3E), and can scale to over 100 GigaRing channels on a single CRAY T3E system [Johnson96].

5.2 Reliably

We based the UNICOS/mk operating system on the UNICOS operating system, a highly reliable, production, UNIX-based operating system. By reusing an overwhelming portion of code from UNICOS, we exploit the reliability work done on that system over the last several years. In addition to UNICOS, we have added reliability features suitable to a highly scalable system. A failing PE can be mapped out, and a redundant PE mapped in its stead on a running system. In the event of an OS failure on a compute PE, the OS on that PE can be restarted without interrupting the entire system.

We designed reliability into the CRAY T3E interconnection network. All packets on the network incorporate error checking. All but the smallest configurations contain redundant PEs. A failing network link can be disabled, while allowing surrounding PEs to continue operation. A module with a failing PE can be logically and electrically isolated from the rest of the system, removed and replaced with a spare module, and the PEs reintroduced to the system, without interrupting the operation of the system as a whole.

The GigaRing I/O channel supports reliable operation by permitting an I/O node to be logically and electrically isolated to permit repair. The "folding" operation done to isolate the failing node preserves device connectivity during the maintenance action.

5.3 Affordably

The CRAY T3E is built from commodity components - microprocessor, DRAM, and disks. For large configurations, this potentially good price performance is realized only if large problems can be solved. The scalability of the CRAY T3E ensures that this is so. For small configurations, many customers focus on low entry price. The smallest CRAY T3E (an air-cooled 6 PE "AC6" system) has a list price below $500K, and additionally can be upgraded in increments of a single module (4 PEs). Price-performance for '96 will be very competitive at $65-85 per peak MFLOPS. Improvements for '97 will be in line with Moore's Law, which predicts a doubling of price-performance every 18-24 months.

5.4 Conveniently

Programmers will exploit a variety of standard programming interfaces on the CRAY T3E: Fortran 90, ANSI C, the draft C++ standard, PVM, MPI, and ScaLAPACK. A recent addition to this list is the HPF-CRAFT implicit parallel programming model, which is being jointly developed by the Portland Group, Inc., and Cray. This continues Cray's tradition of working within standard interfaces to provide high performance. [MacDonald96]

The UNICOS/mk operating system presents a single system image of a CRAY T3E system, regardless of the number of PEs [Harrell96]. The single system image provides great flexibility in placement of data and programs. There aren't any "local" disks; all disks are local to all PEs. The UNICOS programming interface (which is POSIX compliant) is preserved on UNICOS/mk. On the CRAY T3E, UNICOS/mk will look like UNICOS with a large number of processors. Users will log in like UNICOS, run programs like UNICOS, and migrate files like UNICOS. The standard parallel programming mechanisms (MPI, for example) layer on top of the UNICOS interfaces.

Programmers will use the Cray TotalView debugger and MPP Apprentice performance tool for program development on the CRAY T3E. The developers of Cray TotalView have worked closely with users over the last several months. The recent 2.0 release is a stable product which has improved mechanisms for viewing parallel programs. The MPP Apprentice has been widely hailed as a breakthrough in the ability to tune large, long-running parallel applications.

For many users, third party applications represent the essence of convenience. Our plan for the CRAY T3E is to focus on a small set of scalable industrial applications, and push a few of those into industrial production use before we expand our horizon to more codes. We are working on all the codes listed in Table 1 today, and expect most of them to be available on the CRAY T3E by year-end '96.

6 Product Status

We have several small CRAY T3E systems in check-out currently (March 25, 1996) and have confirmed the correct operation of the components of the system (CRAY T3E hardware, GigaRing hardware, UNICOS/mk operating system, and CF90 compiling system, for example).

We are working closely with the Pittsburgh Supercomputing Center to stabilize the system quickly for parallel applications use. To accelerate that work, we have shipped a small CRAY T3E system to PSC. The focus of the work on that system will be porting and tuning of parallel applications, and system exposure to parallel applications.

We expect to ship production machines in the second quarter of this year, and to be shipping at full volume in the third quarter of this year.

Table 1. Third Party Applications for CRAY T3E

academic/ chemistry	CHARMM	GAUSSIAN94	AMBER
	DISCOVER	UNICHEM	GAMESS
seismic	FOCUS	GEOVECTEUR	
reservoir	FALCON	ECLIPSE	VIP
automotive/ aerospace	STAR-HPC	FLUENT/UNS	FLO67
	RAMPANT	AIRPLANE	
	LS-DYNA3D	PAM-CRASH	
environmental	HIRLAM	NOAA/GCM	MM5
	PCCM	POP	RADM
electronics	FAIM		
rendering	SoftImage	RenderMan	

7 Summary

Scalable computing takes a significant step forward with the advent of the CRAY T3E system. The CRAY T3D system sold typically in large configurations to research labs and government agencies, who commonly write their own applications. The audience for the CRAY T3E will include many users of third-party industrial applications and users at organizations who wish to purchase smaller scalable systems.

8 References

[Broner96] "Scalability of the UNICOS/mk Operating System", Proceedings of the Spring '96 CUG.

[DEC95] Alpha 21164 Microprocessor Hardware Reference Manual, Digital Equipment Corporation, 1995.

[Harrell96] "The UNICOS/mk Operating System", Proceedings of the Spring '96 CUG.

[Johnson96] "New I/O Hardware", Proceedings of the Spring '96 CUG.

[MacDonald96] "The HPF-CRAFT Programming Model", Proceedings of the Spring '96 CUG. test

[Scott96] "Synchronization and Communication in the T3E Multiprocessor", submitted to Architectural Support for Programming Languages and Operating Systems, 1996.

[SIA94] The National Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, 1994.