The Encore Continuum: a complete distributed work station–multiprocessor computing environment

by C. GORDON BELL, HENRY B. BURKHART III, and STEVE EMMERICH
Encore Computer Corporation
Wellesley Hills, Massachusetts

ANTHONY ANZELMO, RUSSELL MOORE, and DAVID SCHANIN
Hydra Computer
Natick, Massachusetts

ISAAC NASSI
Encore Languages
Wellesley Hills, Massachusetts

and CHARLÉ RUPP
Resolution Systems, Inc., an Encore Company
Marlboro, Massachusetts

ABSTRACT

The Encore Continuum is a UNIX-compatible computing environment designed to provide a range of computing styles from distributed work stations to host multis. The *multi,* meaning multiple microprocessor, is a new computer class that spans an order of magnitude performance range covering large micros to small mainframes. The multis' parallelism can be exploited for transparent timesharing, transaction processing, and real time control. Multis are also likely to be the basis for parallel processing.

INTRODUCTION

Companies have traditionally used multiple bit-slice technologies to produce computer "families" that cover a particular price/performance range. Powerful MOS and CMOS microprocessors are likely to change this strategy. Some companies have built proprietary microprocessors to implement new computers in their range (e.g., IBM's PC/370, DEC's MicroVAX and Data General's Micro-Eclipse). Others have built new computer families that leverage off of the performance range of standard microprocessor families (e.g., IBM's PC and AT).

A multiprocessing approach with up to four processors to achieve a performance range has been used by a few mainframe vendors over the last two decades, but the high interconnection costs between processor, memory and input/output components has limited the success of multiprocessing in range and applicability.

The Encore Continuum is a full-range computer family consisting of multis, distributed processing servers, high-resolution terminals, and work stations—all interconnected by local area networks. Microprocessors offer sufficient performance to support today's general-purpose computing at nearly all price/performance levels provided that the microprocessors are organized in groups. The Encore Continuum architecture is designed to exploit this rapidly improving technology; its multi scales linearly in price and performance over an order of magnitude range providing a high performance computation and database server, and its distributed work stations similarly scale over a factor of several hundred and provide both local processing and access to shared multis.

GOALS OF THE ARCHITECTURE

Major objectives of the hardware architecture include

1. Cost-effective multi-user computing with incremental scalability of an order of magnitude range for processor, memory, mass storage, and input/output computing resources.
2. Hardware-independence of microprocessor architecture, data formats (e.g., floating point), and communications technologies in order to track transitions to new technologies without rendering the architecture or user's system obsolete.
3. Arbitrarily large virtual memories to support memory-intensive applications, and the accommodation of very large physical memories for high-performance memory applications and input/output (i.e., file) buffering.
4. Ability to expand or rearrange systems easily to respond to changing requirements, new applications, and new computer and communications technology.
5. Access to the Continuum both within a local area and from long distances using industry-standard local and wide area network protocols from a variety of industry-standard user-interface devices including terminals and personal computers (PCs).
6. High-quality human interface devices with full-page display formats, integral graphics, distributed windowing, and industry-standard connections, protocols, and application software environments.

The goal of the Encore Continuum is to provide a complete and industry-standard compatible software environment that

1. Is UNIX-compatible for multiprogramming in order to achieve long-lived and stable system interfaces and provide for the acquisition of a large flow of compatible and competitive software from many sources. By having a single standard interface, applications software can be written to run on all hardware. Thus, the software industry competes to provide better products.
2. Is location independent so that a user can compute or store data with minimal communication costs, system responsiveness or performance is enhanced, and data and equipment security is increased.
3. Supports multiprogramming, distributed computing, and parallel processing applications with minimal degradation of throughput and response-time due to systems software overhead.
4. Provides identical programmer and user interfaces across the Continuum in order to protect investments in applications programs and personnel training, and enforce security restrictions uniformly, where needed.
5. Allows sharing of resources among nodes in the Continuum and other vendor equipment.

ENCORE PRODUCT OVERVIEW

Figure 1 shows the hardware products of the Encore Continuum as described below:

Multimax: A multiprocessor that spans a processing performance range of 1.5 to 15 million instructions per second (Mips) in increments of 1.5 Mips. Input/output throughput can be expanded in increments of 1.5 megabytes per second to 15 megabytes per second. Memory can be expanded in increments of 4 megabytes to 32 megabytes.

Annex (Ancillary Network Exchange computers): Intelligent, low-cost terminal and PC concentrator and gateway

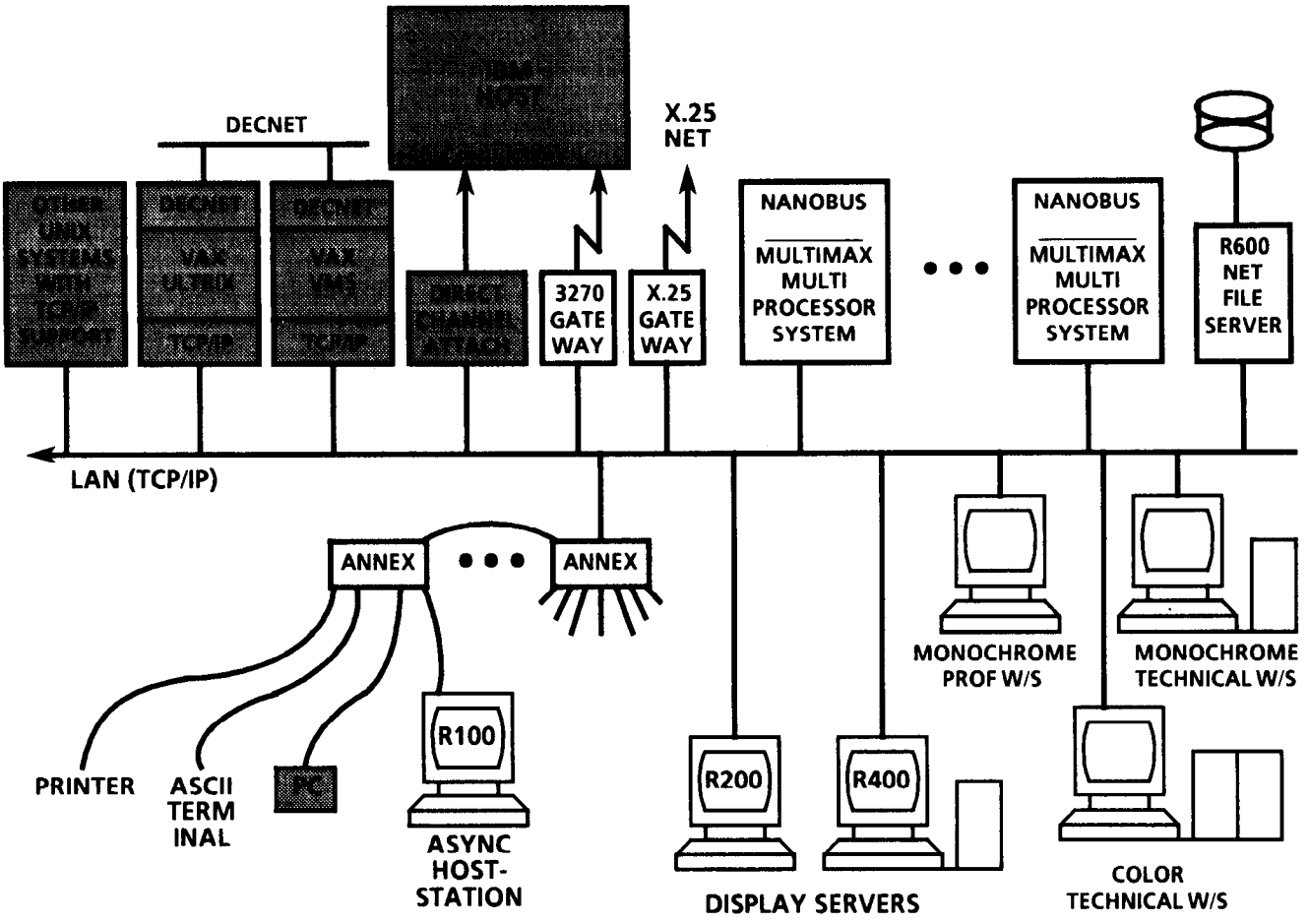

Figure 1—The Encore Continuum, a distributed processing and multiprocessor computing environment

computers for Multimax systems and Resolution stations. Functions such as terminal access to the Continuum and gateWaying to PCs, public data networks, and external computing environments such as SNA take place via Encore's Annex computers.

Resolution R100 Host stations and R500 Work stations: The R100 is a high resolution, large screen, multiple-window, multiple-host access station designed for host-based computer access. The R500 is a compatible work station with local processing, primary memory, and mass storage.

Local Area Network (LAN): Ethernet (IEEE 802.3) LAN is used to interconnect all computing elements and provide computer-to-computer intercommunication, common gateways to other computers and public communications networks, and common access to terminals and PCs.

Interconnections with other computers which support the TCP/IP protocols.

UMax 4.2 and UMax V: Two UNIX software environments derived from the University of California at Berkeley BSD 4.2 and AT&T UNIX are provided. These are primarily for technical and commercial applications, respectively. Software includes the 200 general-purpose tools provided with UNIX, a full complement of user productivity tools, editors, languages and debuggers, and a relational database with Ally, Encore's 4th-generation language.

THE MULTIMAX

Encore's Multimax computer uses multiple microprocessors sharing a common memory to achieve a scalable large performance range, lower price/performance ratio over range, and increased reliability and availability over uniprocessor systems.

Multimax's power is derived from the Nanobus which interconnects 20 Multimax cards within a backplane (see Figure 2). Every 80 nanoseconds, a 32-bit address (giving a system-wide address of up to 4 billion bytes) and 64-bits of data can be transmitted. Thus, Nanobus has a data carrying capacity of 100 million 8-bit bytes per second. (By comparison, standard and emerging buses for multis are usually one-tenth to one-fourth as fast.) All cards can operate online, offline, or on a completely standalone basis for complete self-diagnosis.

The Nanobus provides the key to product longevity, as it is able to accept new higher-speed processors that will evolve with CMOS VLSI. Real-time data can be processed at up to

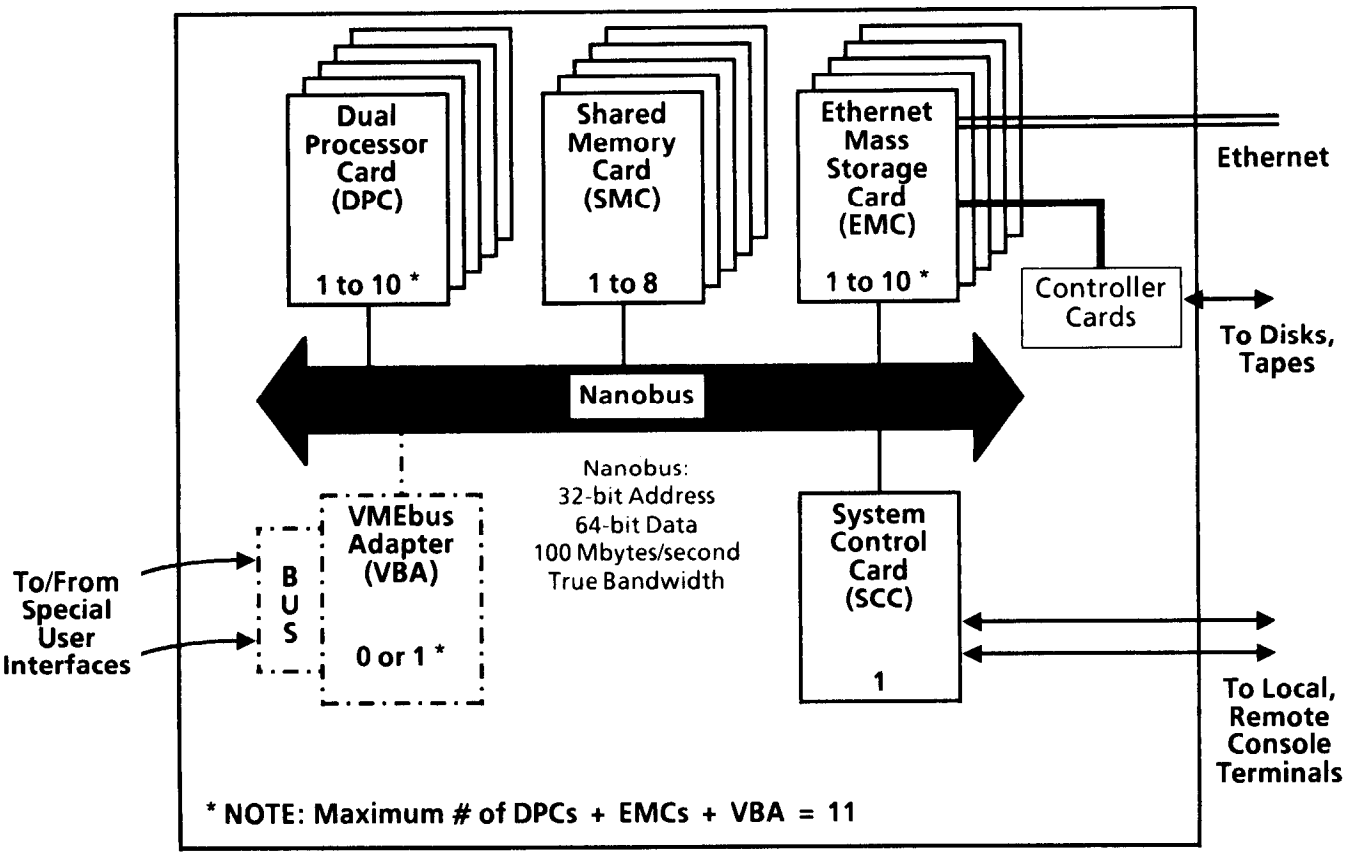

Figure 2—The Encore Multimax multi(ple) microprocessor computer

full bus bandwidth (100 megabytes per second) using direct memory access (DMA) via priority-based programmed interrupts (40,000 events per processor/second maximum) and via direct program control.

At least one of each of the following card-type options is required:

1. Dual Processor Card (DPC)—two National 32032 processors share a common 32-kilobyte cache. A high-performance floating-point option using special chips is provided for arithmetic-intensive applications. Encore rates this processor at 0.75 million instructions per second. With 10 DPCs, a single Multimax can process up to 15 million instructions per second.
2. Ethernet/Mass Storage Card (EMC)—interfaces to Ethernet and to disks. This card contains a 32032 for managing input/output transfers and diagnostics. It has sufficient capability to operate as a LAN-based file service computer. Up to 11 DPC or EMC cards can be placed in one system.
3. Shared Memory Card (SMC)—four megabytes of memory organized in two independent banks with error detection and correction codes, an SMC uses 256-Kbit memory chips. Eight SMCs can be placed in a single system providing up to 32 megabytes of memory. An on-card computer can completely check and diagnose the memory offline.
4. System Control Card (SCC)—performs bus arbitration, logs errors, provides system diagnosis, and communicates with an operator and a remote console.

Peripheral Options

Multimax also offers battery backup—fixed and removable disks of 520 and 300 megabytes, respectively, and 1600/6250 bpi magnetic tape options.

THE LOCAL AREA NETWORK

Encore currently uses the most widely accepted standard (IEEE Standard 802.3—Ethernet) to interconnect its computing nodes at a rate of 10 million bits per second. Other standard LANs will be adopted in response to market requirements. The function of the LAN is for

1. Computer-to-computer communication for distributed processing, file transmission, and virtual terminal access among computers
2. Common access to other networks via Annex gateway computers
3. Common access from terminals and PCs via Annex concentrator computers

4. Formation of a fully distributed computing environment using Encore's powerful single-user work stations
5. Connection to existing personal computers, minicomputers, and mainframes

A LAN is not required for basic system operation.

ANNEX TERMINAL AND ACCESS (CONCENTRATOR) COMPUTER

Each Annex concentrator computer attaches up to 16 terminals and printers along the LAN in a fully distributed fashion permitting up to several thousand terminals to access all computers within a single LAN. Five Annex computers can be connected to a single LAN port, or it can be directly connected to a Multimax if there is no LAN. Annex roughly doubles the processing power of the Continuum, since roughly one Annex computer is used with each Multimax processor. Wiring from terminals to computer is simplified by distributing the physical connections to the Annex concentrators, unlike most terminal architectures, which require all RS-232C terminal lines to be connected to a particular computer. Since the LAN is basically a distributed switch, any serial port on an Annex can communicate with any Multimax or Resolution on the LAN. Annex incorporates a general-purpose remote procedure call facility to communicate with Multimax systems and Resolution work stations.

Annex is programmed to perform time-consuming functions such as character processing on input from, and screen updating on output to, terminals that require no host or central database interaction. Annex contains a National 32016 microprocessor with 128 kilobytes of memory.

On terminal initialization, a switch program asks the user to request a host. The Annex "notifies" the selected host, the terminal becomes connected to it, and the host runs the standard log-on process. Connection switching allows users to connect to other hosts and then to switch among them. Connection binding allows site managers to fix (bind) a port on a given Annex to a given host; this facility binds users to a particular application, and binds dedicated peripherals such as line-printers to particular computers.

Annex has options for both asynchronous and synchronous communications and direct and modem connections. Standard terminals and Encore host stations communicate with hosts at up to 38.4 kilobaud transmission rate per terminal. Hard copy options include a 200 character per second matrix printer and 800, 1200 and 1800 lines-per-minute printers.

ANNEX GATEWAY COMPUTER

The Annex gateway provides access to various communications and industry networks using protocol conversion hardware and software. The protocols include: IBM SNA, IBM Block Mode Terminals (3270), IBM PC and X.25. Presentation-level services associated with gatewaying generally run in the host, which communicates via the remote-procedure call communications architecture with protocol-conversion software running in the gateway.

RESOLUTION COMPUTING STATIONS

The Resolution stations use a 19-inch screen size to give an unscaled, ledger-sized 11″ × 14″ page at high resolution using 1056 × 864 pixels. A ledger sheet of 176 columns and 86 rows can be displayed. Keyboard and pointing device (e.g., mouse) input are provided. The stations (without keyboard) occupy a desk space of 16.5 square inches. The text and graphics protocols provided, which allow existing and future software to be run without modification, include VT100, ANSI 3.64, Tektronix 4010/4014, Regis, and VDI for GKS.

The stations are designed to address a variety of applications, including the station of choice for the professional programmer, text and typographic input, engineering, business and accounting where computational power and large screens are required, and special functions such as translation where side-by-side text is required.

Resolution Host Station (R100)

The R100 is a single, but universal host station because it can communicate with as many as three computers through separate windows. For example, the R100 can simultaneously access Hydra, a traditional host (e.g., IBM 370 or VAX-11), and a PC AT for personal computer software. All functions of the R100 are carried out under the program control of a National 32016 microprocessor. The R100 is also designed to be used as a remote slave station to conventional work stations (i.e., a user can have a work station at home or at a second office).

The R100 can be upgraded to become an R500.

Resolution Work Station (R500)

The R500 is a self-contained computer system with a primary memory of two megabytes and disk memory of 20 megabytes. The processor, a National 32016, is completely compatible with other computers in the Encore Continuum. Thus, software can be run either within the work station, among work stations, or among host stations and Multimax systems in a completely flexible and transparent fashion.

THE UMAX 4.2 AND V DISTRIBUTED SOFTWARE ENVIRONMENT

The software environment in the Encore Continuum is UNIX-compatible and enhanced to support both distributed and parallel processing. Distributed processing support is provided by a communications architecture that provides for cooperative, efficient interprocessing between networked Multimax systems, Resolution stations and Annex computers. Language constructs for assigning task forces of processors to a single process for support of parallel processing are also provided.

UMax 4.2 and UMax V constitute Encore's standard operating systems. Programs that run under either UNIX System V or UNIX BSD 4.2 are compatible and portable to the corresponding Multimax and Resolution systems.

In addition to UNIX standards, the Encore Continuum extends UNIX

1. To take full advantage of demand-paged virtual memory, multiprocessor performance, and distributed terminal architecture.
2. To provide data sharing and synchronization mechanisms between user processes (UMax 4.2). Additional system calls and library subroutines support these new multiprocessor functions.
3. By unifying language standards and language-related data formats across both operating systems, to simplify portability of applications between environments.

UMax Performance on the Multimax

UMax 4.2 and UMax V incorporate three strategies for high performance that are inherent in the Encore Continuum: (1) symmetrical multiprocessing, (2) scalability to a large number of processors, and (3) distributed intelligent peripheral control.

Symmetrical multiprocessing (multithreading) achieves maximal performance in the Multimax by ensuring that any processor can execute any user process or part of the operating system. This ensures that there are no inherent bottlenecks. One copy of the operating system supports all the processors, memory, and input/output computers. In order to allow multiple processors to gain simultaneous access to operating system services, concurrent access must be controlled to each process and operating system routine.

Controlled concurrent access to internal UMax resources is achieved with the following mechanisms:

1. Spin locks—accomplish synchronization by executing tight instruction loops until the expected condition occurs (used only for critical short-duration events).
2. Semaphores (Dijkstra style)—accomplish synchronization by putting requesting processes "to sleep" until the requested resource is available.
3. Read/write locks—specialized forms of semaphores; provide access to data structures for a single writer or multiple readers.

Scaling performance to many processors and very large memories is a major performance consideration. Multithreaded operation alone will not realize the performance potential inherent in the multiple resources of a Multimax. Two additional performance enhancements have been added to accommodate a large processing load: (1) shared data in the operating system is minimized, and (2) the UNIX terminal driver has been redistributed to the Annex concentrator computers. The first method caches frequently used in-memory resources such as file and directory entries. For resources controlled by tables, it is generally appropriate to lock individual entries rather than the whole table. In other cases, kernel tables have been divided into subpools of entries, linked together, and located by hashing. This minimizes search times and allows for locking of subpools rather than whole tables.

High-Level Languages and Debuggers

C. This language is supported by an optimizing compiler. Traditional assembly languages for system-level and time-critical applications programs are minimized.

FORTRAN-77. Fully conformant to the ANSI standard using an optimizing compiler.

Pascal. ISO standard.

COBOL 74. FIPS intermediate-level 2.

A source-level debugger for local debugging of user mode code, written in the C and FORTRAN-77 languages, is available. A remote debugger of supervisor mode programs that facilitates remote Encore diagnosis of system problems by the Encore service organization is also provided.

CONCLUSIONS

The indefinite expandability goals of the architecture are satisfied by allowing almost unlimited numbers of each system to be added to a local area network. Multimax and Resolution are incrementally upgradeable to higher levels of performance by adding processors, memory, and mass storage over an order of magnitude range. Standardization, portability, and ease of use are inherent with UNIX.

APPENDIX A
THE MULTI AS A NEW COMPUTER CLASS

The multi is an emerging computer class made possible by recently developed powerful micros that have the speed and functionality of mid-range super minicomputers. In contrast to computer families implemented from a range of technologies, a multi is scalable, thereby permitting the building of a single computer which spans a performance range. The multi is a significant alternative to conventional micros, minis, and mainframe families.

Multis can be used today, without redesigning or reprogramming of applications, because computer systems operate on many independent processes. With multis, it is possible to operate on many of these processes in a parallel fashion that is transparent to the user (with each process on an independent processor). Thus, the multi is likely to be the path to a fifth generation of computers based on parallel processing.

Historical and Technological Basis

Computer systems with multiple processors have existed since the second generation (the Burroughs B5000, a dual symmetrical processor, was introduced in 1961). Most mainframe vendors and some minicomputer suppliers currently offer systems with two to four processors. However, these structures have been expensive to build, due to the high cost of typical processors. Hence, they have found application mostly for high-availability computing (e.g., communications, banking, airline reservations).

The modern 32-bit microprocessor's function, perfor-

mance, size, and relatively negligible cost create a new potential for multiprocessors. With 32-bit addressing, hardware support for page organized, virtual memory, and complete instruction sets with integer, floating, decimal, and character operations, these chips offer performance levels comparable to mid-range superminis such as the VAX-11/750.

The multi is a multiprocessor structure designed to take advantage of these new micros. It employs an extended UNIBUS-type interconnect whereby all arithmetic and input/output processor modules can access common memory modules. Cache memories attached to each processor handle approximately 95% of its requests, thereby limiting traffic on the common bus. With these local caches, more processors can be attached before saturating the common bus.

With proper attention to the design of critical elements (e.g., the common bus), large multis using current-technology micros can outperform high-end superminis and some mainframes in terms of total performance. This advantage should continue to grow. The performance of MOS and CMOS microprocessors has improved (and is expected to continue to improve) at a 40% per year rate while the TTL and ECL bipolar technologies (on which most traditional minis are based) have shown roughly a 15% per annum improvement.

When compared to traditional uniprocessor designs, the multi delivers improved performance, price, and price/performance as follows:

Configurability Range. Through modular design, the multi allows the user to "construct" the correct level of performance or price without having to choose among a limited number of computer family members.

Availability. The multi has inherent reliability through redundancy because it is built from a small number of module types (four, typically). With appropriate software support, faulty modules which are replicated can be taken out of service, thereby allowing continued operation with minimum downtime.

Design and Manufacture. Because the multi is composed of multiple copies of a small number of modules rather than the large number of unique boards in a typical minicomputer, it is faster and less expensive to design. Individual module types, produced in larger volumes, result in improvements of 30% in manufacturing costs over conventional uniprocessors.

Evolutionary Technology Upscaling. With appropriate design, multis allow long-term performance upscaling through evolution. As key components of the processor and memory cards improve over time, the computer can be upgraded without replacement, in an evolutionary fashion. In addition, increased cache sizes through denser parts and improved cache management disciplines will permit substantially greater numbers of processors to be installed without saturating the common bus (provided that the bus design has allowed for this performance growth). All of this will permit graceful and cost-effective evolution in processor performance, input/output throughout, and memory size over a range of one to two orders of magnitude over a ten-year period.

Applications

Multis will be widely used for many applications because they can provide the most cost-effective computation except in cases where the power of a single large processor is required on a single sequential program. Because of the rapid rate of microprocessor evolution, the percentage of applications requiring single-stream performance in excess of that delivered by each of the multi's processors is already small and will continue to shrink. On the other hand, the emergence of the multi will lead to parallel processing.

We can better understand how multis may be applied by classifying the degrees of parallelism achievable. *Grain size* (see Table I) is the period between synchronization events for multiple processors or processing elements. Synchronization is necessary in parallel processing to initialize a task, parcel out work, and merge results. The multi exploits coarse- and medium-grain parallelism within an application (not the fine-grain, which is the focus of vector pipelined computers such as Cray 1 on wide word microprogrammed array processors).

Table I—Constructs for parallelism, synchronization, and supporting Encore structures across various grain sizes

Grain Size	Construct for Parallelism	Synchronization Interval (Instructions)	Encore Computer Structures to Support Grain
Fine	Parallelism inherent in single instruction or data stream	1	Specialized processors (e.g., *systolic* or *array*) added to Multimax
Medium	Parallel processing or multitasking within a single process	20–200	Multimax
Coarse	Multiprocessing of concurrent processes in a multi-programming environment	200–2000	Multimax
Very Coarse	Distributed processing across network nodes to form single computing environment	2000–1M	Multiple Multimaxes, Encore work stations, and other machines, on Ethernet

Groups of multis and conventional work stations can interact over networks to implement very coarse granularity.

As all modern operating systems are multiprogrammed, whereby each job in the system is at least a single process, and many support multitasking or subprocesses, most current applications are already designed to take advantage of the multi at the coarse-grain level. When used in a timesharing or batch environment, each processor of a multi can be assigned to a separate job to exploit the parallelism inherent in the work load. The UNIX pipe mechanism allows multiple processes to be used concurrently on behalf of a single user or job to achieve parallelism in reading a file, computing, and outputting to one or more files. Transaction processing is inherently a pipeline of independent processes.

The multi can be a more efficient multiprogramming computer than the traditional uniprocessor because the number of context switches (and lost time) is reduced. Additional parallelism is in the operating system itself. Execution of operating system code often accounts for 25% or more of available processing time when file, database, and communications subsystems are included. By restructuring the operating system, multiple independent system functions can be executed on independent processors.

If reprogramming of subsections of the application is possible, multis permit additional parallelism to be realized at the medium-grain level (i.e., parallel processing) by segmenting a problem's data for parallel manipulation by independent processors. This has been shown to be quite effective on simulation, scientific modeling, and analysis problems (such as matrix operations, linear programming, and partial differential equation solution, etc.) that permit data elements to be processed in segments.

Finer granularity of parallelism is achievable in the framework of the multi through specialized processors installed into its common bus. This is most effective when the algorithms are known a priori as in certain signal processing applications.

Multiprocessors, augmented by both programmable pipeline (i.e., systolic) and specialized processors for fine-grain parallelism, will cover the widest range of problems of any computing structure.