Technical Description of the DEC 7000 and DEC 10000 AXP Family

1 Abstract

The DEC 7000 and DEC 10000 products are mid-range and mainframe Alpha AXP system offerings from Digital Equipment Corporation. These machines were designed to meet the needs of large commercial and scientific applications and therefore are high-performance, expandable systems that can be easily upgraded. The DEC 7000 and 10000 systems utilize the DECchip 21064 microprocessor operating at speeds up to 200 MHz. The high-speed chips, large caches, multiprocessor system architecture, high-performance backplane interconnect, and large memory capacity combine to create mainframe-class performance with a cost and size previously attributed to mid-range systems.

The design of the DEC 7000 and 10000 systems provides a high-end platform and system environment for multiple generations of Alpha AXP chips. This platform, combined with a multiprocessor architecture, yields a multidimensional upgrade capability that will allow the system to meet users' needs for several years. System upgrade can take place by adding processors, replacing existing processors with next-generation processors, or both. This upgrade capability ensures stability to the system in terms of the physical and fiscal aspects of the end user's computing environment.

The DEC 7000 and DEC 10000 systems are the logical follow-on products of the highly successful VAX 6000 family.[1] The new systems are capable of supporting either VAX processors or Alpha AXP processors. The capability to upgrade from a VAX processor to an Alpha AXP processor without changes to the system is essential for minimal disruption of large commercial applications. Most features of the VAX 6000 systems have been carried forward to the DEC 7000 and DEC 10000 products, and any deficiencies have been corrected.

The DEC 7000 and DEC 10000 products are derived from the same system design. The DEC 10000 is a more fully configured system and includes an n+1 uninterruptible power system, additional I/O subsystems, and I/O expansion cabinets. The DEC 7000 uses a 182-megahertz (MHz) DECchip 21064 whereas the DEC 10000 uses a 200-MHz DECchip 21064.

A very important goal for the project that encompassed the development of the DEC 7000 and 10000 systems was to provide a similar pair of systems based on a VAX microprocessor. A VAX microprocessor, called NVAX+, was designed to be pin compatible with the DECchip 21064 (the Alpha AXP microprocessor).[2,3] The system was designed to be somewhat microprocessor independent, and both VAX and Alpha AXP versions of the systems were

implemented. The VAX products (VAX 7000 and VAX 10000) were introduced

in July 1992 and can be upgraded to DEC 7000 and DEC 10000 systems by a simple swap of CPU modules.

2 System Architecture

The DEC 7000 system consists of CPU(s), memory, an I/O port controller, and I/O adapters, as shown in Figure 1. The system is configured in a variety of ways, depending on the size and function of the system. A system backplane consists of nine slots and houses CPUs, memory, and an I/O port controller. The I/O port controller resides in a fixed slot, and CPUs and memories occupy the remaining eight slots. The initial system offerings allow up to 6 CPUs. (Architecturally, the system may support up to 16 CPUs.) Up to 14 gigabytes (GB) of memory can be supported if only 1 CPU module is present and all remaining slots contain memory.

The I/O subsystem consists of an I/O port controller and four I/O ports which have been adapted to the XMI or the FutureBus+. The I/O ports are generic and may be adapted to other forms of interconnect in the future. The system backplane, power system, and up to two I/O backplanes are housed in the system cabinet. Additional I/O backplanes (up to a system total of four) may be configured in expansion cabinets.

3 Technology

The DEC 7000 system is built primarily of CMOS (complementary metal-oxide semiconductor) components. The DECchip 21064 microprocessor is built using Digital's 0.75 micrometer CMOS-4 process. All modules utilize LSI Logic LCA100K series gate arrays for the system bus interface and for on-board logic functions. The LSI Logic LCA100K features up to 235K two-input NAND gates. All modules use the same custom I/O driver circuit within their respective gate arrays to drive and receive the system bus. A custom 419-pin pin grid array (PGA) package was developed to house all bus interface gate arrays. Unlike the VAX 6000 series, a common bus driver part is not used in order to minimize the number of levels of buffering in the system.

Module technology is standard 10-layer construction with 4 signal layers, 4 power layers, and top and bottom cap layers. Double-side, surface-mount construction is used extensively throughout the system. Etch width is 5 mils with 7.5-mil minimum spacing. Via sizes down to 15 mils are used. A mixture of physical component technologies is used with all large VLSI (very large scale integration) parts in 100-mil PGA packages. Most standard logic utilizes 50-mil surface-mount technology. Module interconnect to the backplane is made through a 340/420-connection, four-row, 100-mil-spaced pin and socket type connector. Forty-eight-volt power is distributed throughout the system; local regulation is provided on the module for specific voltages required.

4 System Interconnect

The heart of the DEC 7000 system is a high-performance system interconnect, called the LSB, which allows communications between multiple processors, memory arrays, and I/O subsystems. It provides a low-latency, high-bandwidth data path among all components. A common shared view of memory is maintained by means of the system interconnect and cache logic on processor modules.

Three types of modules are defined for the LSB.

- Processor modules, which contain the CPU chip, cache subsystem, and console functions. The initial DEC 7000 design has the capacity for a maximum of six processor modules.

- Memory modules, which contain dynamic random-access memory (DRAM) chips and a memory controller. A system can contain up to seven memory modules, each with a capacity of 64 megabytes (MB) to 2GB.

- I/O interface modules, which provide access to I/O buses and I/O adapters. Only a single I/O port controller module may reside in the system. The I/O port controller module can arbitrate at a higher priority than CPU nodes to improve I/O direct memory access (DMA) latency and provide atomic DMA writes of data less than a cache block in size.

The LSB is a limited-length, non-pended, pipelined, synchronous, 128-bit-wide bus with distributed arbitration. All transactions occur in a set of fixed cycles relative to an arbitration cycle. Up to three transactions can be in the pipeline at a given time, enabling the full capability of the bus to be realized. Arbitration occurs on a dedicated set of control signals and may be overlapped with data transfer. Data and address are multiplexed on the same set of signals. The bus protocol supports write-back caches, and all memory transfers are 64 bytes in length. The cycle time of the bus is 20 nanoseconds (ns), providing an overall data rate of 800MB per second and a utilized system bandwidth of 640MB per second.

The LSB transmits 40-bit physical addresses, providing a physical address space of 1 terabyte. Given the current rate of DRAM technology evolution, the LSB will have a useful life of 8 to 10 years before physical address space is exhausted. A 40-bit physical address was chosen to minimize the data path width in the processor bus control gate array.

A non-pended pipelined bus was chosen instead of a traditional pended bus to allow for simple node interface designs. Transactions start and finish at precisely defined times. A "stall" function may be used if a given transaction cannot be completed within the system timing constraints.

The "stall" function freezes the bus pipeline, maintaining the order of all transactions. Consequently, nodes can be designed with no queuing between the bus interface and local storage (DRAMs for main memory or static RAMs [SRAMs] for cache memory). The maintenance of strict bus transaction ordering also alleviates many potential lockout conditions experienced on pended buses.

Digital's previous mainframe systems have used a switch-based system interconnect instead of a bus. This interconnect was typically required because these systems were based on emitter coupled logic (ECL) with only a small, single-level cache subsystem; therefore, high bandwidth was required between main memory and the processor. The CMOS design of the DEC 7000 allows a large (4MB) second-level cache to complement the 16-kilobyte (KB) on-chip cache. The large amount of cache minimizes the need for memory bandwidth. A bus-based design was chosen over a switch-based design to minimize memory latency, minimize design complexity, and reduce system cost.

All LSB transactions consist of a single command cycle and four data cycles. These five cycles appear in fixed cycles relative to the arbitration cycles. Up to three transactions may be pipelined, as shown in Figure 2.

The LSB uses a distributed arbitration scheme. Ten request wires are driven by the CPUs or the I/O module that wishes to use the bus. Eight request lines are allocated to the eight potential CPU modules. The remaining two request lines are used by the I/O controller module. All modules independently monitor the request wires to determine whether a transaction has been requested, and if so, which module wins the right to send a command cycle to start the transaction.

The arbitration scheme employs a least-recently-used rotating priority algorithm for CPU modules and a fixed high/low scheme for the I/O port controller. The I/O port controller arbitrates using the highest and lowest priority levels, arbitrating high six times then low two times. This arrangement ensures that the I/O port controller can utilize greater than 50 percent of the available system bus bandwidth while still ensuring the CPUs some access to the system bus. The I/O port controller also uses its unique arbitration scheme to ensure atomic read/modify/write sequences on the bus necessary for performing writes of less than a full naturally aligned 64-byte quantity. The I/O port controller does the read at its next scheduled priority and then immediately follows up with the write at highest priority. This scheme ensures that no other node can access the data between the read and the write.

All command/address and control/status register (CSR) cycles are protected with parity. Data cycles to and from memory are protected with error correction code (ECC). Transmit check is used by all modules to verify that what a given module is asserting on the bus is actually being seen on the bus. Transmit check allows the detection of bus collisions and faulty bus drivers or receivers.

The system interconnect is physically implemented as a centerplane which is 350 millimeters (mm) wide and 500 mm high. There are four module

connections on one side, and five on the other. The centerplane-module connection is implemented using a four-row pin and socket connector with connections on a 100-mil grid. Modules are 410 mm high and 340 mm deep. This module size was chosen to allow the maximum module size within the

constraints of an 865-mm-deep cabinet and of the centerplane technology. Modules are spaced on 65-mm centers and are contained within a box that provides customized air flow for each different module design.

The DEC 7000 was designed with a centerplane interconnect to solve the problem of bus length and to meet the need for wide module spacing that allows for the anticipated heat-dissipation requirements of future processor chips. With a centerplane, the number of module slots available for a given length of bus increases by (n*2)-1 where n is the number of slots available in a conventional backplane. A centerplane configuration leaves little space on the backplane for termination networks. Designers solved this problem by adopting a distributed termination scheme with bus terminator networks present on all modules in the backplane.

5 Processor Module

The primary purpose of the processor module is to provide a large second-level cache to the processor chip and to act as an interface to the system bus and memory for missed cache references. The processor module in the DEC 7000 system was designed to use either VAX or Alpha AXP chips. As noted above, a common design is used in the implementation of the VAX and DEC 7000 and 10000 systems, with the only significant differences being the processor chip and the console/diagnostic code. Figure 3 is a block diagram of the processor module.

The processor module provides a 4MB external cache, which is shared by the processor chip and the bus interface chips. The cache is organized as a single set (direct mapped), with a block and fill size of 64 bytes. The external cache conforms to a write-back, conditional update, cache coherency protocol. The processor on-chip data cache is a proper subset of the external cache and uses a write-through protocol.[4]

The structure of the cache is shown in Figure 4. Each cache line consists of 512 bits of data (with 112 bits of ECC), 12 bits of tag (with 1 parity bit), and 3 status bits (with 1 parity bit). The 12 bits of tag data applied to a 4MB cache size sets a processor physical address capability of 16GB. (This is a processor limitation, and future processors will address larger memory sizes.) The control bits contain information that allows the cache and memory systems to maintain coherency. The control bits are defined as follows:

- A valid bit, indicating whether or not this line contains valid data
- A shared bit, indicating whether or not this line may also be resident in another processor's cache in the system
- A dirty bit, indicating whether or not this line has been written to by

this processor

Upon detection of a cache read miss in the processor on-chip cache, the processor accesses the external cache tag to see if the given block is resident. The processor chip contains the tag comparitor and status logic to determine a "hit." If the block is resident in the external cache, the processor then cycles the external cache data store twice, each time reading in 128 bits of data and 28 bits of ECC for a total of 32 bytes (internal processor cache block size is 32 bytes). The external cache cycles at a rate five times the processor chip clock period (and at two times the period for the VAX variant). Upon the detection of a "miss," the processor chip informs the bus interface chips by means of handshake signals and waits until the miss is serviced on the LSB.

Upon a data write by the processor, the data is written through to the external cache. If the data is already resident in the cache, it is updated and conditionally broadcast onto the system bus if marked as shared. If the selected cache line contains a different valid tag, the current (old) cache line is written to memory and replaced by the new tag and data. To improve performance during this operation, the current cache line is stored in a local victim buffer while the new data is read. After the new data has been placed in the cache, the old data is written back to memory as a background operation.

A duplicate set of cache tags (backmaps) are kept by the bus interface logic for both the external cache and the internal processor chip D-cache. These backmaps are accessed by the bus interface logic on all bus references to determine the action necessary to maintain cache/memory coherency.

On bus read requests, the processor bus interface references its external cache backmap and supplies data from the on-board cache if a "dirty" copy of the data is present. On bus writes, a check is performed to see if the data is present in the processor on-chip D-cache. If the data line is present, the updated data is accepted. If the data line is not present but is instead in the external cache, the line is invalidated. This cache update policy is an attempt to minimize false sharing of data by only updating on references to a cache line in the processor on-chip cache, which is small and should contain only freshly referenced data.

False sharing of data is a problem common to multiprocessor systems running fully symmetric operating systems. When a process is migrated from one processor to another, dirty data often remains in the cache of the previous processor. When the new processor requests that data, it becomes "shared," resulting in the need to update all copies by means of bus transactions on all subsequent modifications of the data. Since the process has migrated, there is no need to maintain the state of the data in the cache of the previous processor; doing so slows down execution of the process due to the bus transactions required to update. The write-update policy described

in the previous paragraph provides a means to estimate if "shared" data is still in use by the previous processor and provides a means to flush it from the previous cache if it has not been recently referenced.

The external cache is 128 bits wide with longword ECC protection. The ECC scheme used to protect the external cache is identical to that used on the LSB, which allows flow-through ECC. The processor chip checks and corrects data for all processor refills. The bus interface chips perform lookaside ECC checking for fault isolation purposes but do not perform ECC correction.

The processor module also provides system console functions. The module includes universal asynchronous receiver/transmitters (UARTs) for communication with the console terminal and power subsystems, a time-of-year clock, and 896KB of flash read-only memories (ROMs) for console and diagnostic code. Each processor contains a complete console subsystem, but only one module uses this function in a multiprocessor system. This approach allows static reconfiguration of the system in the event of a module failure.

A 4MB module-level cache was chosen because it was the largest natural implementation using 256K x 4 SRAMs driven by the 128-bit-wide cache data path defined by the DECchip 21064 microprocessor. Denser SRAMs were not available at the necessary speed (10 to 12 ns), and a multiway cache architecture is not easily implemented with the DECchip 21064. The fill size of 64 bytes was selected to efficiently use the 16-byte-wide system bus and provide 80 percent bus data efficiency.

Figure 5 shows a photograph of side 1 of a processor module. Additional cache RAMs and drivers reside on side 2.

NOTE

Figure 5 (Processor Module, Major Components Highlighted) is a photograph and is unavailable.

6 Memory Module

The memory subsystem of the DEC 7000 comprises one to seven memory array modules with a single module capacity of 64 to 2048MB. The primary functions of the memory array modules are to respond to bus read/write functions, refresh the memory RAMs, and maintain ECC data for the memory. The design supports either 4MB or 16MB DRAMs, on-board interleaving on modules with greater than 64MB, and multimodule interleaving under many conditions.

The DEC 7000 memory modules run synchronous with the LSB. Memory transactions occur in fixed cycles relative to the system bus. All memory space transfers consist of 64-byte blocks that are transferred 16 bytes at a time over four contiguous data cycles. Read and write data wrapping is done on 32-byte naturally aligned boundaries. The DRAMs are 4-bit-wide

parts, and an entire 64-byte block is read or written in parallel and buffered for bus transmission.

Data wrapping is a method used to provide a lower latency return of the data required by a read command. The bus contains an extra address bit that indicates in which half of a 64-byte block the requested data lies. The memory controller returns the half block containing the target data first, allowing faster resumption of processing. Data wrapping has no benefit on write transactions but is done to simplify the design of the system.

DEC 7000 memory modules are protected with a quadword ECC algorithm. The chosen ECC implementation allows detection and correction of single-bit failures, detection of all 2-bit failures, and detection and correction of any error wholly contained within a 4-bit-wide DRAM. Memory modules convert LSB longword (32-bit) ECC into quadword (64-bit) ECC that is stored with LSB data on writes. During LSB reads, quadword ECC is converted to longword ECC. Quadword ECC allows for higher packing densities on the memory module with fewer DRAM components. Longword ECC is used on the system bus because the DECchip 21064 microprocessor dictates the use of longword ECC in its external caches, and the timing of the external cache will not allow a conversion to a different ECC for bus transactions.

The memory module contains a hardware-based self-test that checks each bit on the module to be sure it can be set to either a 0 or a 1 state and initializes the memory to a known good ECC state. All memory modules execute self-test in parallel upon system initialization at a rate of approximately 35MB per second. This approach results in substantial savings in boot time as compared to a system that tests memory with initialization code executed by the processor. Moreover, the self-test provides excellent error isolation in the event of a failure.

DEC 7000 memory is designed in 64MB, 128MB, 256MB, 512MB, and 2GB modules. The 64MB, 128MB, and 256MB modules use 4MB DRAMs, double-side surface mounted. The 512MB modules use 4MB DRAMs mounted on soldered-in single in-line memory modules (SIMMs). (PC-style socketed SIMMs proved unreliable for large configurations.) The 2GB modules use 16MB DRAMs mounted on soldered-in SIMMs.

7 I/O Subsystem

The DEC 7000 I/O subsystem consists of an I/O port controller and four high-speed parallel ports. The I/O controller provides an interface between the system bus and the parallel ports. Additional modules provide the interface between the high-speed parallel ports and specific standard I/O buses. To date, interfaces have been designed for the XMI, which is used as the I/O bus on the VAX 6000 and VAX 9000 systems, and for the FutureBus+, which is an IEEE standard high-performance bus definition.

The I/O port controller and specific bus adapter architecture was adopted to allow a flexible bus strategy that can evolve over time, as well as

to accommodate the physical separation of processor and I/O subsystems necessary in an expandable system with multiple I/O channels. The I/O port controller cable(s) will function to a maximum cable length of 3 meters.

This length allows I/O expansion cabinets to be placed on either side of the main system cabinet.

The aggregate bandwidth of the I/O port controller is 256MB per second. Each parallel port is capable of operating at a maximum of 135MB per second for data flowing from the I/O subsystem to memory and at 88MB per second for data flowing from memory to the I/O subsystem.

The I/O port controller module with its four parallel ports is a standard part of all DEC 7000 systems and resides in a dedicated system backplane slot. Various system configurations are available that contain between one and four XMI I/O buses. The FutureBus+ subsystems will be available when FutureBus+ components become available in the computer industry.

The I/O port controller provides a "mailbox" interface between the processor and I/O devices. A processor instruction cannot directly access a register in an I/O device, as was possible on previous VAX implementations. To use the "mailbox" interface, a processor creates a work descriptor packet in memory and then issues a command to the I/O port controller to execute the command. Command completion is asynchronous and the processor may choose to do other work while the command is executed. The "mailbox" interface between processors and I/O devices was created to allow relatively slow I/O devices to interface to a high-speed, non-pended system bus. If a processor were allowed to access the I/O device directly, the system bus would be stalled for large portions of time.

Clearly the mailbox communications method is more complicated than traditional direct access. Fortunately the mailbox is used only when a processor needs to directly access an I/O device. The I/O device can directly access main memory (or possibly a CPU cache) with all necessary buffering done by the I/O port controller. Most modern high-performance I/O adapters use high-level, packet-based protocols, which require very little direct access of the I/O adapter by the processor.

A typical CPU-initiated I/O transaction to an intelligent disk controller on an XMI bus to read from the disk would have the following steps.

- The CPU places a disk controller command packet requesting a disk read into system memory.

- The CPU sets up an I/O mailbox structure with a command to inform the disk controller that there is a command packet in memory, writes a register in the I/O port controller to inform it that there is a mailbox transaction to complete, and then spins on a done bit in the mailbox structure.

- The I/O port controller fetches the mailbox structure from memory,

generates an XMI write command to the disk controller, and sets the done bit in the mailbox structure. The CPU sees the assertion of the done bit and goes on to other work.

- The disk controller receives the mailbox data and then generates an XMI request to read its command packet from memory.

- The I/O port controller reads the specified command packet from memory 64 bytes at a time and sends it back to the disk controller 32 bytes at a time.

- The disk controller decodes the command packet, reads the requested data from disk, and starts writing to system memory in 32-byte segments.

- The I/O port controller buffers the 32-byte writes from the disk controller into 64-byte segments and writes the data to system memory.

- The disk controller signals an interrupt on the XMI to indicate that the requested operation is complete, which is received by the I/O port controller. The I/O port controller signals an interrupt to the CPU.

8 Console and Diagnostics

Like many previous VAX systems, the DEC 7000 system employs an embedded console. The console function is performed by code run on the processors within the system rather than by a dedicated, detached front-end processor.

Unlike the strategy for previous VAX systems, a unified console and diagnostic strategy was adopted for the DEC 7000 and 10000, VAX 7000 and 10000, and DEC 4000 systems. A single code base not only provides the basic console functions but also extends diagnostic support for manufacturing and field firmware upgrade support. This unified strategy has reduced the total development effort and promoted a common "look and feel" across the different systems.

The console development also differed from that of previous VAX systems. The primary implementation language was C, with only various architecture-specific code in Alpha AXP (or VAX) assembly language. The console and processor diagnostic code was simulated prior to the arrival of hardware. This simulation greatly simplified early hardware debug; the console had basic functionality after a single debug session.

At power-up, each processor acts independently to execute processor-specific diagnostics and console initialization. The processors then select a console primary, which then proceeds to test and configure the memory and I/O subsystems. The console primary also retains control of the console terminal line; console secondaries communicate with the primary through memory-resident messages. After initialization, diagnostic or other console tasks can be assigned to any processor in the configuration. One benefit of this arrangement is that system diagnostics and exercisers can be run in parallel.

Like previous DECsystem consoles (that is, systems based on MIPS Co. chips), the DEC 7000 console provides a set of services, or callbacks,

to the operating system. These services can be used to control automatic bootstrapping across operating system crashes as well as primitive I/O services used by the operating system during bootstrap and system crash. The latter simplifies the operating system device support by providing simple read/write functions common to all devices.

A feature of the power of the console is the field firmware update utility. Field upgrade of all system firmware (console and I/O adapters) is accomplished by the DEC 7000 firmware update utility (LFU). LFU is really a dedicated console image which is distributed on CDROM. The system console is used to boot LFU, which is then used to update all system firmware.

9 System Packaging

The DEC 7000 system cabinet is 1700 mm high by 800 mm wide by 865 mm deep. The cabinet houses the system backplane, up to two I/O subsystems, and disk arrays or batteries for the system battery-backup function. Expansion is possible by using one or two I/O expander cabinets, each of which houses up to two additional I/O subsystems and additional disk arrays. Further mass storage expansion is possible with Digital's standard line of mass storage cabinets connected by CI, DSSI, or SI interconnects.

The DEC 7000 cabinetry has been designed for easy system upgrade and servicing. The system backplane assembly, power system, and I/O subsystems are modular and easily replaced by field personnel. The process of future upgrades can be accomplished more quickly and reliably through the use of modular subassemblies.

As shown in Figure 6, the DEC 7000 main system cabinet contains a central air mover with logic assemblies above and below it. The air mover is a single motor with a large molded vane assembly and can pull air through both the upper and the lower logic assemblies. An air flow of approximately 900 cubic feet per minute with velocities up to 1800 linear feet per minute is maintained through the upper logic assembly, which contains the processor and memory subsystems. Although not necessary for the DECchip 21064, this large volume of air movement was designed into the machine to allow upgrades through several generations of processor chips. By using standard air-cooling techniques and customized module "boxes" that optimize local air flow, it is possible to cool processor chips of up to 70 watts in the DEC 7000 system cabinet.

Above the air mover are the system backplane and the modular power subsystem. Below the air mover are four modular spaces for I/O bus backplanes, disk drives, or batteries.

I/O, disk, and battery subsystems occupy varying amounts of the four modular spaces. The XMI subsystem occupies two spaces and is oriented front to back because of its rear-exit cabling scheme. The FutureBus+ subsystem occupies a single rear space. Disk subsystems consisting of up to six 5.25-inch (DSSI or SCSI [small computer system interface]) or fourteen 3.5-inch (SCSI only) drives may occupy any of the modular spaces. Batteries for the uninterruptible power system occupy two modular spaces, which may be oriented either front to back (for XMI-based systems) or side to side (for

FutureBus+ systems).

The expander cabinet is identical to the main system cabinet, with two exceptions: disks may be packaged in the area occupied by the system backplane, and there is no control panel. Up to two XMI or FutureBus+ subsystems may be placed in an expander cabinet.

10 Power Subsystem

The power subsystem of the DEC 7000 family has a highly modular, hierarchical design. The basic power system provides 48-volt direct current (VDC) to all subassemblies which in turn further regulate to necessary voltages. Each module in the system backplane contains on-board regulation. This feature will allow the system to easily evolve with changing voltage requirements as CMOS technology moves to lower voltages to reduce power consumption. Voltage tolerances can be tightly controlled since transmission drops are negated; a precise voltage level can be set at the time of module manufacture. The voltage and tolerance to a high-performance CMOS processor must be very tightly controlled in order to extract maximum performance. The XMI, FutureBus+, and disk subsystems all regulate the 48 VDC to lower voltages at a subsystem-wide level, not at the module level.

The 48-VDC modular power system consists of one to three parallel regulators, each of which produces 2400 watts of power. A maximally configured cabinet needs no more than two power regulators. An additional regulator can be configured into the system to provide an n+1 capability for higher availability.

The power system also includes a battery standby function that provides 48 VDC throughout the system in the event of an AC power failure. Unlike earlier VAX systems in which power was maintained only to system memory, the DEC 7000 keeps the entire system powered, including in-cabinet mass storage. Depending on the system configuration, power is maintained for a minimum of 20 minutes in an n+1 power configuration. N+1 power with full battery backup is standard on all DEC 10000 systems.

The DEC 7000 system employs a highly intelligent power subsystem with microprocessors in all 48-volt regulators, which report status to processor modules by means of a serial interconnect. System software can therefore monitor a wide range of power system operating parameters, including voltage output, AC input, efficiency, and battery charge state. In a large configuration with optional expander cabinets, the expander cabinet power systems also communicate with the system processors to provide system-wide power status.

11 Performance

The DEC 7000 and DEC 10000 systems are the fastest uniprocessor and multiprocessor, microprocessor-based computer systems in the world as of their introduction date (10 November 1992) and as defined by SPEC89 and SPEC92 benchmark data. For compute-intensive benchmarks, the DEC 10000 is approximately 10 percent faster than the DEC 7000, based entirely on the difference in processor clock speed.

The base performance of the DEC 7000 and DEC 10000 systems is determined by the speed of the processor chip and is heavily influenced by cache, memory, and I/O subsystems. The design goal for the DEC 7000 and DEC 10000 systems was to extract the maximum possible performance from the DECchip 21064 by providing an electrical and physical environment capable of supporting 200MHz processor operation as well as large caches, a large and fast memory subsystem, and multiple I/O subsystems.

While full system-level performance data is still being collected, the very high speed processor performance measured on the SPEC benchmarks combined with the very high performance cache, memory, and I/O subsystems of the DEC 7000 and DEC 10000 systems should yield very impressive overall system performance. See Table 1.

Table 1: DEC 7000 and DEC 10000 System Performance Measurements

	DEC 7000	DEC 10000
SPECmark89	167.4	184.1
SPECint89	95.1	104.5
SPECfp89	244.2	268.6
SPECint92	96.9	106.5
SPECfp92	182.1	200.4
SPECthroughput89 (4 CPUs)	604.4	654.6
LINPACK double-precision		
100x100 (MFLOPS)	38.6	42.5
1000x1000 (MFLOPS)	102.1	111.6

12 Design Process

The DEC 7000 system was specified, designed, and tested by a group of approximately 200 people in Boxboro, Massachusetts. The system design team was responsible for all aspects of the design except the DECchip 21064 microprocessor.

Conceptual work on a system to follow the VAX 6000 family was started in early 1989, although at that time design work was focused on implementations using VAX and MIPS R4000 processors. In the latter part of 1989, the decision was made to pursue the Alpha AXP strategy, and earlier concepts were reworked to incorporate much higher levels of performance to accommodate the proposed Alpha AXP chip.

In October-December 1989, a core team of approximately 10 engineers was assembled to firmly define system architecture and to produce specifications for all subassemblies. By July 1990 all specifications were complete, and implementation was started. The first processor module was powered up in June 1991, followed by a full system power-up in September

1991. The VMS operating system was booted on a DEC 7000 system on September 9, 1991, and OSF was booted in November 1991.

A minimal DEC 7000 system includes 430,000 gates of logic contained in gate arrays, whereas a minimal VAX 6000 Model 200 includes 94,000 gates. Despite more than four times the gate count, the design portion of the DEC 7000 program was completed in approximately 9 months as compared to 12 months for the VAX 6000 program. This reduction in design time was achievable in part because of the maturing of the engineering population (many of the DEC 7000 engineers had worked on various VAX 6000 implementations), as well as advances in design tool technology and the availability of significantly

more powerful computers for design simulation. At its peak, the DEC 7000 program was consuming 1500 VAX units of performance, or VUPs, of compute power (primarily multiprocessor VAX 6000 Model 500 systems) and used over 325,000 hours of CPU time for simulations.

13 Conclusion

The DEC 7000 and DEC 10000 systems are the second generation of highly configurable and expandable systems produced by Digital Equipment Corporation. These are the first systems expressly designed to accommodate multiple-processor architecture types. As computer technology moves forward at an ever-increasing pace, this type of design will be demanded by computer users and will be necessary to manage engineering costs.

The DEC 7000 and DEC 10000 system platform will accommodate new VAX and Alpha AXP processors for several years. Over that time, this platform will span a performance range of greater than 50:1. It will provide computer users with a stable system environment that should help minimize the changes caused by the continued development of new processor chips. While this level of flexibility incurs additional initial engineering and product costs, it does provide a very cost-effective way to deal with the inexorable forward march of technology.

14 Acknowledgments

The following engineers formed the system architecture team of the project that produced the DEC 7000 and DEC 10000 and VAX 7000 and VAX 10000 products: Frank Bomba, Reinhard Schumann, Mike Callander, Steve Polzin, Kathy Harrington, Dave Mayo, Catharine van Ingen, Vicky Triolo, Bob Dickson, Dave O'Keefe, Jim Leahy, Hansel Collins, Jim Stegeman, Darrel Donaldson, Dave Hartwell, Charlie Barker, Mark Stefanski, Brian Allison. Various parts of this text originated within engineering specifications written by this team.

15 References

1. Digital Technical Journal, vol. 2, no. 2, featuring papers on the VAX 6000 Model 400 (Spring 1990).

2. G. Uhler et al., "The NVAX and NVAX+ High-performance VAX Microprocessors," Digital Technical Journal, vol. 4, no. 3 (Summer 1992): 11-23.

3. D. Dobberpuhl et al., "A 200-MHz 64-bit Dual-issue CMOS Microprocessor," Digital Technical Journal, vol. 4, no. 4 (1992, this issue): xx-xx.

4. A.J. Smith, "Cache Memories," Computing Surveys, vol. 14, no. 3 (September 1982).

16 Trademarks

Alpha AXP, AXP, DEC 7000 AXP, DEC 10000 AXP, DECchip 21064, VAX, VAX 6000, VAX 7000, and VAX 10000.

The following are third-party trademarks:

MIPS is a trademark of MIPS Computer Systems, Inc.

LSI Logic is a trademark of LSI Logic Corporation.

SPEC, SPECfp, SPECint, and SPECmark are registered trademarks of the Standard Performance Evaluation Cooperative.

17 Biographies

Brian R. Allison Brian Allison is a senior consultant engineer for Digital's mid-range VAX/Alpha AXP systems group and is the system architect responsible for the coordination of the VAX and DEC 7000 and 10000 system definition and design. Prior to this work, he served as system architect for the VAX 6000 product. Brian holds a B.S.E.E. and a B.S.C.S. from Worcester Polytechnic Institute (1977).

Catharine van Ingen A consulting software engineer, Catharine van Ingen was co-system architect for the VAX and DEC 7000 products. Catharine is currently on leave from Digital and is working on engineering document management in large heterogeneous systems. Before joining Digital in 1987, she worked on data acquisition systems for two large physics detectors at the Fermi National Accelerator Laboratory and Stanford Linear Accelerator Center. She holds several degrees in civil engineering, including a B.S. and an M.S. from the University of California and a Ph.D. from the California Institute of Technology.