



HP

Technology Brief

servers technology

## HP system interconnect technologies: the DNA (cell controller) family of ASICs

*By reducing an entire system interconnect to a few chips, HP has redefined scalability and performance levels in a new line of servers.*

September, 2001

---

## Introduction

Increasing demands for computer system scalability (consistent price/performance and higher processor counts) have combined with performance increases of individual components to drive systems manufacturers to rethink core system architectures.

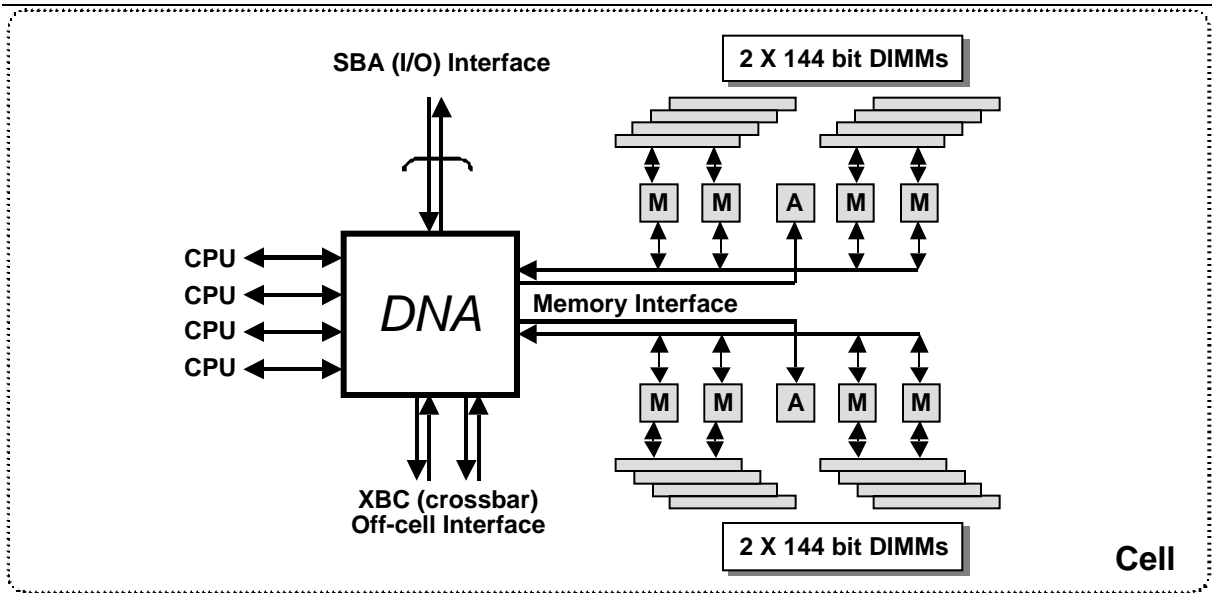
HP has recently introduced multiple server systems that meet these demands for scalability by implementing a new memory subsystem. Leveraging proven experience in semiconductor and systems technology, HP design teams have created a family of ASICs (Application Specific Integrated Circuits) that provide scalability to tens or hundreds of processors, while maintaining a high degree of performance, reliability, and efficiency.

The key ASIC in this new architecture is dubbed the “DNA<sup>1,2</sup>.” The DNA (called the *cell controller* in the high-end Superdome server), is a processor-I/O-memory interconnect, and is responsible for communications and data transfers, cache coherency, and an interface to other hierarchies of the memory subsystem.

### DNA Benefits

The DNA is the heart of the memory subsystem in this new line of scalable servers. The DNA replaces – in a single ASIC – the venerable system bus. By reducing the function of the system bus to a single chip, the system takes on a more modular design comprised of “cells”. Each cell is a small (4 CPUs or less) symmetric multiprocessor (as shown in Figure 1).

Multiple cells are “clicked” together to form larger systems based on a hierarchical memory architecture. The result is a ccNUMA (cache-coherent, non-uniform memory access) design that scales effectively from tens to hundreds of processors, gigabytes of memory and large I/O slot counts.



**Figure 1. The cell architecture featuring the DNA ASIC.**

<sup>1</sup> “DNA” derives its name from its function as the heart of the “cell” building block.

<sup>2</sup> The DNA ASIC is sometimes referred to as the CC (Cell Controller).

Direct benefits of this implementation include:

### ***Performance***

By reducing the length of the physical path taken by processor-memory-I/O transactions within a cell, the effective transaction rate between components is improved significantly. This increases bandwidth (more data per period) and decreases latency (the time to transfer the data), thereby increasing overall system performance while eliminating the most common barriers to scalability. DNA technology also opens opportunities for performance increases beyond raw clock rate (for example, a more efficient implementation of cache coherency).

### ***Reliability***

At the highest level, a single chip is inherently more reliable than a large, multi-layer printed circuit board. The DNA delivers reliability beyond this level through extensive built-in logic to permit parity or ECC-protected transfers on all ports. Further, the chip is capable of detecting and isolating faults that could otherwise corrupt the rest of the system.

### ***Lower Costs***

Component and manufacturing costs are lowered in systems that use the DNA technology. Since multiple systems employ the DNA, lowered part costs are realized through higher volume production. Costs of service (diagnostics, replacement of components, etc.) further improve as a result of the reduction in complexity in the interconnect.

### ***Scalability***

Scalability refers to the ability to realize near-linear performance gains as a result of aggregating components such as processors and I/O channels. For example, when the  $n^{\text{th}}$  processor is added to a system, it should optimally provide as much performance as the  $1^{\text{st}}$  processor. Scalability is directly related to the efficiency of the interconnect – especially under high loads (many concurrent transactions). The high degree of parallelism inherent in the DNA design ensures that a high percent of aggregate peak bandwidth is maintained.

### ***Modularity, Expandability and Upgradability***

By reducing an entire, 4-processor SMP to a single board, the physical characteristics of systems in which this board resides are greatly changed from traditional multiprocessor designs. First, the density of the system is reduced considerably. For example, one of the latest DNA-based systems from HP, the rp8400, fits 16 processors and 64 GB of memory in less than one-half of a standard 2-meter rack. This density is unrivaled in mid- to high-end systems available today.

Similarly, the system may be easily expanded. With HP's high-end Superdome system, the DNA architecture allows customers to start with a minimal configuration, and by plugging in additional cell boards, grow the system up to 16 cells (64 processors) – while maintaining a single, shared-memory system.

Lastly, the DNA/cell technology simplifies upgradability to future generations of processors and/or memory technology. This is especially important as the new standard in processor technology, Intel's Itanium™ Processor Family (IPF), becomes the predicted standard for performance and application availability.

## Technology

While previous generations of HP servers have delivered impressive performance, the one element of high performance computing that has been difficult to achieve is *high-end scalability* – efficiently exploiting many (tens to hundreds) of processors and I/O channels to provide high application performance. With that goal in mind, and additional goals of lowered costs and increased reliability, HP engineers leveraged their extensive experience in semiconductor and system technologies to design the DNA and its companion ASICs.

The DNA ASIC operates at a clock frequency of 250 MHz – with some of the ports taking advantage of double data rate (DDR) signal technology<sup>3</sup>. This technology advance effectively doubles the transaction rate without increasing the actual frequency.

The DNA is the world’s largest ASIC, comprising 24 million transistors implemented with IBM’s new SA-27 CMOS process. This process features an  $L_{\text{eff}} = 0.12 \mu\text{m}$ ,  $L_{\text{drawn}} = 0.16 \mu\text{m}$  and seven levels of metal. The die size is approximately 24mm X 24mm (576 mm<sup>2</sup>), employing a Ceramic Ball Grid Array (CBGA) package with a high pin count of 1,657 leads. At 250 MHz, the DNA clock is the fastest commercially available chip of similar functionality. Figure 2 shows a typical implementation of the DNA in a single-board cell with HP-patented Turbo-cooler heat dissipation devices. Note the physical closeness of the DNA to the critical system components to minimize signal latency.

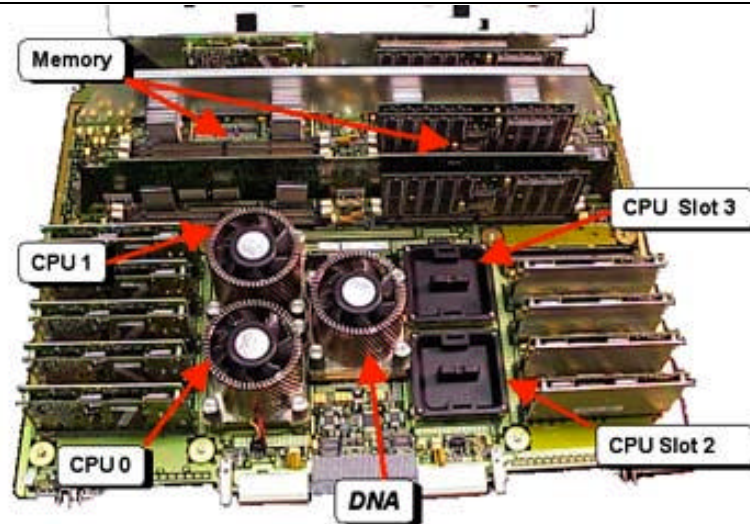


Figure 2. A typical implementation of DNA.

## Theory of operation

The major functional units of DNA are as shown in Figure 3. The core of the ASIC is a switching interface referred to as the Central Data Path (CDP). Connected to the CDP are four Processor Interfaces (PI), the two Memory Units (MU), Processor Dependent Hardware (PDH), I/O controller, and external crossbar interface. Up to four PA-RISC<sup>4</sup> processors

<sup>3</sup> DDR involves superimposing an additional clock signal (a “strobe”) on data lines, resulting in the ability to carry signals on the both the leading and trailing edges of data pulses.

<sup>4</sup> Later implementations will support Intel’s Itanium™ Processor Family bus architecture.

utilize separate point-to-point connections to the PIs. The memory interfaces consist of two replicated memory subsystems that are controlled independently.

As a result of the massive space allocated to control logic and on-chip memory, the DNA is capable of making cache coherency decisions, encaching frequently-used data cache lines and coherency tags, and multiplexing data transfers from various sources to/from the four processors.

System reliability was one of the cornerstone design goals for the DNA. Full SECDED (single-bit error correction, double-bit error detection) ECC is implemented on data busses, parity for addresses, and timers to watch over system components. When failures are detected, error containment is guaranteed and data corruption is prevented<sup>5</sup>. Further, all of the off-cell communication paths are engineered to tolerate on-line addition and deletion (OLAR) of the cell board.

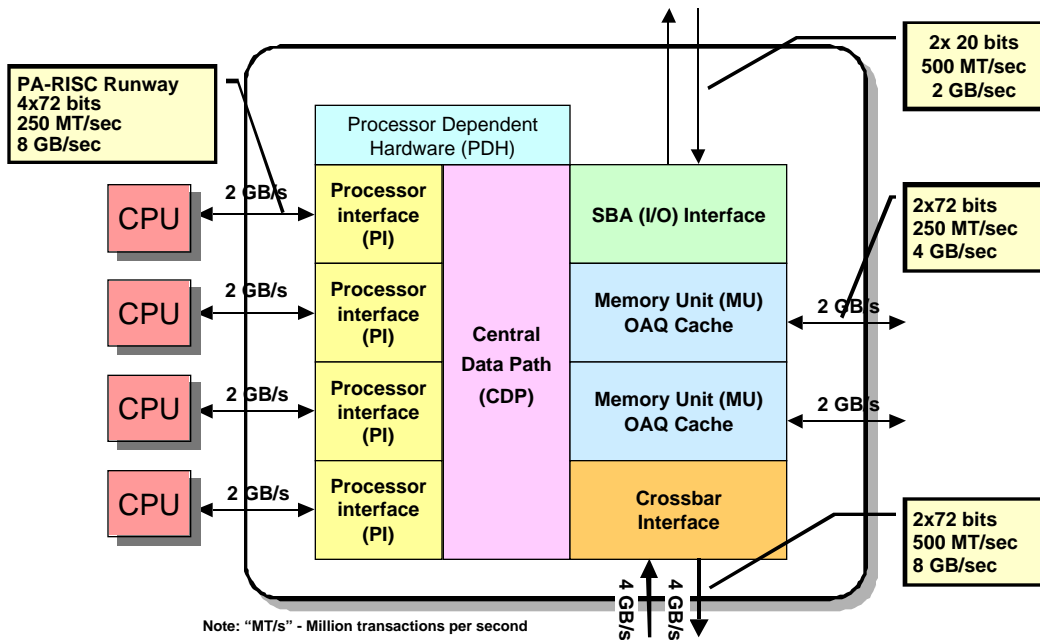


Figure 3. DNA Block Diagram

### Functional Units

#### Processor Interface

The DNA interfaces to as many as four PA-RISC processors via HP’s 64-bit processor interface bus called Runway. Although the Runway bus is designed to support multiprocessing, each processor has its own dedicated interface to the DNA. This improves performance by eliminating the use of Runway bandwidth for coherency checking (coherency is discussed below.) The DNA processor interface utilizes the Runway bus running in double data rate mode, which doubles the peak bandwidth of each Runway bus to 2 GB/sec.

<sup>5</sup> HP has several patents pending on the error containment mechanisms in the DNA.

Split read request transactions and data return transactions are supported by both the PA-RISC processor and the processor interface, enabling multiple outstanding transactions. The processor directly exploits this feature, with the ability to support multiple outstanding cache misses and instruction scheduler support for out-of-order data returns. This capability maximizes bus efficiency and overlaps execution of the on-chip functional units.

A unique feature of the DNA's processor interface is the per-processor address routing logic called the "cell map"<sup>6</sup>. Processor data requests are presented to the interface on the DNA, which uses the cell map to immediately determine where the desired cache line resides. When the cache line resides within the current cell the request is transferred directly to the memory interface. If the cache line is off-cell, the request is sent over the crossbar interface directly to a DNA in another cell.

The DNA is capable of supporting interleaved memory requests (across multiple cells) as well as cell-local memory requests. Interleaved memory optimizes the system resources for large, flat applications such as databases or very large arrays. Localized memory is important to reduce latency of tasks with data structures such as stack and heap and single CPU processes that can fit within a single cell's memory.

### ***Memory Unit***

Each memory unit provides a dual function: management of cache coherency and control over memory transfers.

In all of the systems supported by DNA, cache coherency is maintained using directory-based coherency structures in main memory.. The memory subsystem utilizes a special, highly efficient ECC code that stores both the directory tags and ECC codes with exactly the same overhead (12.5%) as is present in almost every system that employs ECC. Complementing the cost effective ECC is an organization that effectively includes a "spare" DRAM so that any DRAM failure is corrected without loss of data or system downtime.

A directory provides a distinct advantage over a "snoopy bus" system by making cache coherency queries to processors only when absolutely necessary. This prevents valuable address bus bandwidth from being occupied by coherency snoops as is prevalent in competing systems. To maintain a simple programming model, all memory in the system is strongly ordered<sup>7</sup> and implements the MESI coherency protocol, where cache lines can take the state of Modified, Exclusive, Shared, or Idle.

This coherency mechanism minimizes the number of messages sent between the processors and the memory unit. The vast majority of data requested is fetched into the processor's cache and marked as either modified or exclusive. The processor is then free to modify the data without any additional memory traffic. After the processor is done with the data, the cache line is cast-out. When cast-outs of modified data are sent to memory, both the data and the "idle" directory states are written simultaneously. Exclusive cast-outs send a short message to the memory controller indicating that the CPU no longer owns the line, which in turn sends a message to memory to update the tag as idle without pulling the cache-line's data back into the DNA. This conserves bandwidth along the entire path to memory.

---

<sup>6</sup> HP has several patents pending on the cell map logic.

<sup>7</sup> "Strongly ordered" means that cache consistency is maintained automatically by the system – eliminating the need for special compiler or programming features to maintain coherency.

Each MU employs a unique mechanism to increase the efficiency of the management of cache coherency, called an Ordered Access Queue (or OAQ). The OAQ is a 56-entry (per memory unit) cache that contains both directory tags and data. A fetch that requires a coherency query to a third-party owner is held in the OAQ, transferring back to memory only when the transaction is complete. The fully associative OAQ cache resolves multiple requests for the same cache-line quickly and efficiently without cycling the memory SDRAMs. An example of how this improves performance is the case where multiple processors are accessing the same memory location repeatedly – such as access to a semaphore.

Each memory subsystem has separate address and data paths; addresses are sent directly to the memory DIMMs while data is sent and returned through dedicated 2 GB/sec (64 bits of data) bi-directional data busses. The result is extremely low memory latency accesses with near-peak bandwidth to/from memory DIMMs.

### *I/O Interface*

The I/O interface consists of a pair of 500 MT/sec differential, uni-directional point-to-point links between the DNA on a cell board and a system bus adapter ASIC (SBA). The SBA is a companion ASIC to the DNA, specifically designed to support the I/O subsystem – and is usually integrated with the PCI backplane. Differential was chosen because of the relatively long signal runs to external I/O subsystems. The I/O link consists of 21-bits of data/address and a clock strobe pair to provide source synchronous timing. Each link is composed of two sets of signals; one set carries data in one direction and the other set carries data in the opposite direction.

The SBA ASIC contains two caches, is strongly ordered and participates fully in the cache coherency protocol. This design was required to maintain a simple programming model where CPUs can access critical control structures without the special “flush” and “synchronization” operations required with non-coherent or buffered I/O.

Error handling features include CRC and ECC on the link providing for single wire recovery with parity protection on interfaces. Timers watch over PCI cards and detect downstream system failures. Configurable response to errors allows smart PCI drivers to recover from failures

### *Crossbar Interface*

The external crossbar interface is used to connect the cell to the rest of the system while maintaining high per-processor bandwidth and low latency across cells. The bandwidth across the crossbar is maximized by "wave-pipelining" data – that is, multiple data bits are sent down a data wire at the same time and resolved at the destination.

The interface itself logically consists of two 64-bit (72-bit with ECC) unidirectional links clocked at 250 MHz with DDR, yielding an effective rate of 500 MHz, providing an aggregate peak bandwidth between the DNA and the system crossbar(s) of 8 GB/sec.

### *Processor Dependent Hardware (PDH)*

A complete utilities sub-system called Processor Dependent Hardware (PDH) is included in the cell to guarantee independence and fast boot time. The PDH is the module that provides the cell-local resources required to reset a cell and bring it up to a point where it can join

other cells and boot the OS. PDH contains the system boot firmware, which is also used at run-time.

Each PDH contains a Universal Serial Bus (USB) connection into a master USB hub allowing the entire system to be monitored. The PDH interface to the DNA is for diagnostic purposes.

### ***High availability***

The DNA interacts with numerous system-level high availability components. For example, all off-cell ports on the DNA are designed to permit OLAR (On-Line Addition and Replacement) of cell boards. Further, all of the ports of the DNA provide end-to-end error detection and correction mechanisms to ensure that single-bit errors will not effect operation.

With a system that supports multiple instantiations of the operating system (“partitions”), it is imperative that there is isolation between them. In conjunction with the system crossbar, partitions are completely isolated (the DNA prevents even coherency queries from reaching across partitions). This keeps errant transactions in one partition from affecting other partitions and ultimately bringing the entire system down.

## **Conclusion**

HP continues to be a leader in technology that enhances customers’ computing capabilities. The technology delivered with DNA and its companion ASICs represents a breakthrough which allows HP to deliver highly scalable platforms while maintaining cost-effectiveness and high reliability.

The DNA is the key component in providing scalability and leadership performance in HP systems. Maintaining low latencies and introducing overlapped data transfers complements the increasing performance of processor, memory and I/O subsystems that are being incorporated into systems today.

Lastly, the DNA represents an innovative, foundation architecture designed for high availability. Supporting component redundancy, error detection and correction and system partitioning, the technology raises the standard for single-system high availability and reliability.



---

For More Information

Contact any of our worldwide sales offices or HP Channel Partners

(in the U.S. call 1-800-637-7740)

or visit our Web site at: [www.techservers.hp.com](http://www.techservers.hp.com)

Microsoft, Windows, and Windows NT are U.S. registered trademarks of Microsoft Corporation. UNIX is a registered trademark of the The Open Group.

Technical information in this document is subject to change without notice. HP believes information on competitors to be accurate at the time of publication, but does not guarantee its accuracy.

© Copyright Hewlett-Packard Company 2001

**Printed in U.S.A.**  
09/01

---