Memory technology evolution

technology profile, 3rd edition

System memory is arguably the most important component in servers and workstations. Processors use system memory to temporarily store the operating system, mission-critical applications, and the data they use and manipulate. Applications cannot effectively use the full computing power of the processor if the processor must wait for data from system memory. Therefore, both the performance of the applications and reliability of the data are intrinsically tied to system memory. These factors have driven the evolution of system memory from asynchronous dynamic random access memory (DRAM) technologies to high-bandwidth synchronous DRAM (SDRAM) technologies. Still, system memory bandwidth has not kept pace with improvements in processor performance.

This technology profile describes SDRAM and identifies some of the newest memory technologies that HP is evaluating for servers and workstations. For more information, read the HP technology brief "Memory technology evolution: an overview of system memory technologies" available on www.hp.com.

SDRAM

The Joint Electronic Device Engineering Council (JEDEC)—the electronics industry standards agency—developed the SDRAM standard to speed up system memory. By using a memory bus clock to synchronize the input and output signals on the memory chips, SDRAM simplifies the memory controller and reduces the latency from CPU to memory. SDRAM accelerates data retrieval and increases memory capacity also by use of burst mode access, multiple memory banks, greater bandwidth, and register logic chips (registers).

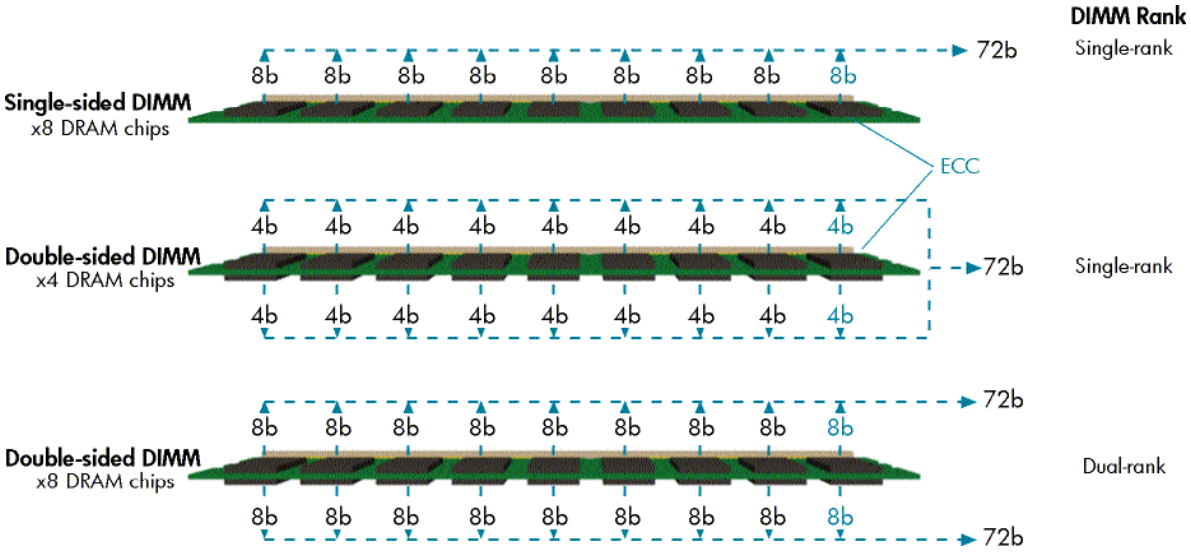
Although use of SDRAM improves overall system performance, there remains a growing performance gap between the memory and processor. That gap must be filled by more advanced memory technologies, including three generations of Double Data Rate (DDR) SDRAM, Rambus DRAM, and a new bi-directional serial interface called Fully Buffered DIMM. These faster memory technologies, along with higher server memory capacities, have increased the importance of memory ranks when adding or upgrading DIMMs. The following section defines memory ranks and explains the importance of using only HP-certified DIMMs for each ProLiant server.

DIMM Configurations

Each DRAM chip on a DIMM provides either 4 bits or 8 bits of a 64-bit data word. Chips that provide 4 bits are called x4 (by 4), and chips that provide 8 bits are called x8 (by 8). It takes eight x8 chips or sixteen x4 chips to make a 64-bit word, so at least eight chips are located on one or both sides of a DIMM. However, a standard DIMM has enough room to hold a ninth chip on each side. The ninth chip is used to store 4 bits or 8 bits of Error Correction Code (ECC). An ECC DIMM with all nine DRAM chips on one side is called single-sided, and an ECC DIMM with nine DRAM chips on each side is called double-sided (Figure 1). A single-sided x8 ECC DIMM and a double-sided x4 ECC DIMM each create a single block of 72 bits (64 bits plus 8 ECC bits).

In addition to single-sided and double-sided configurations, DIMMs are classified as single-rank or dual-rank. A memory rank is defined as an area or block of 64-bits created by using some or all of the DRAM chips on a DIMM. For an ECC DIMM, a memory rank is a block of 72 data bits (64 bits plus 8 ECC bits). A single-rank ECC DIMM (x4 or x8) uses all of its DRAM chips to create a single block of 72 bits, and all the chips are activated by one chip-select signal from the chipset (top two illustrations in Figure 1). A dual-rank ECC DIMM produces two 72-bit blocks from two sets of DRAM chips on the DIMM, requiring two chip-select signals. The chip-select signals are staggered so that both sets of DRAM chips do not contend for the memory bus at the same time.

Figure 1. Single-sided and double-sided DDR SDRAM DIMMs and corresponding DIMM rank.



The role of memory ranks has become more critical with the advent of new chipset and memory technologies and growing server memory capacities. The chipset considers each rank as an electrical load on the memory bus. At slower bus speeds, the number of loads does not adversely affect bus signal integrity. However, for faster memory technologies such as DDR333 and DDR2-400, there is a maximum number of these loads (typically 8) that the chipset can drive. If the total number of ranks in the populated DIMM sockets exceeds the maximum number of loads the chipset can support, the server may not boot properly or it may not operate reliably. Some systems check the memory configuration while booting to detect invalid memory bus loading. When an invalid memory configuration is detected, the system stops the boot process, thus avoiding unreliable operation. To prevent this and other memory-related problems, customers should only use HP-certified DIMMs available in the memory option kits for each ProLiant server.

Double Data Rate SDRAM

DDR is advantageous for systems that require higher bandwidth than SDRAM can provide. To develop DDR SDRAM, designers made enhancements to the SDRAM core to increase the data rate. These enhancements include prefetching, double transition clocking, strobe-based data bus, and SSTL_2 low voltage signaling. At 400 MHz, DDR increases memory bandwidth to 3.2 gigabytes per second (GB/s)—400 percent more than original SDRAM.

DDR-2 SDRAM is the second generation of DDR SDRAM. It offers data rates of up to 6.4 GB/s, lower power consumption, and improvements in packaging. DDR-2 SDRAM achieves this higher level of performance and lower power consumption through faster clocks, 1.8-V operation and signaling, and a simplified command set.

JEDEC is currently developing the third-generation DDR SDRAM technology, DDR-3, which will further improve bandwidth and power consumption. For example, DDR-3 will use 1.5-V signaling compared to 1.8 V for DDR-2 and 2.5 V for DDR SDRAM. JEDEC is expected to approve the final DDR-3 specification by the end of 2005. DDR-3 volume production is expected in 2007.

Rambus DRAM

Rambus DRAM (RDRAM) allows data transfer through a bus operating in a higher frequency range than DDR SDRAM. In essence, Rambus moves small amounts of data very fast, whereas DDR SDRAM moves large amounts of data more slowly. With only an 8-bit-wide command bus and an 18-bit data bus, RDRAM has the lowest signal count of all of the memory technologies. RDRAM uses a packet protocol and is capable of operating at 800 MHz and providing a peak bandwidth of 2.4 GB/s. Although it offers a lower pin count than SDRAM and DDR SDRAM, RDRAM has larger dies and is produced in more limited quantities. As a result, RDRAM costs up to twice as much as SDRAM.

Fully Buffered DIMMs

As memory speed continues to increase with DDR-2 and DDR-3 SDRAM, the number of DIMMs supported per channel decreases. This decrease is related to the parallel stub-bus topology in which electrical signals travel along 72 data lines (64 for data bits and 8 for error checking bits) from the memory controller to every DIMM on the bus. Signal degradation at bus-pin connections and latency resulting from complex routing of data lines cause the error rate to increase as the bus speed increases. To achieve higher bus speeds with DDR technology, designers must sacrifice capacity by decreasing the number of DIMMs per channel.

JEDEC is currently developing the Fully-Buffered DIMM (FB-DIMM) specification, a new bi-directional serial interface that eliminates the parallel stub-bus topology and allows higher memory bandwidth while maintaining or increasing memory capacity. The FB-DIMM can achieve a maximum speed of 4.8 Gb/s. It features 69 pins (less than one-third the pin count of DDR-2 SDRAM). The reduced pin count greatly simplifies board design due to fewer signal traces and the ability to use traces of unequal length. FB-DIMMs also feature an advanced memory buffer chip that transmits signals between the memory controller and memory modules using a point-to-point architecture. The bi-directional interface allows simultaneous reads and writes, thus eliminating delays between data transfers.

TC050803TP, 08/2005