NUMA (non-uniform memory access)

Part of the Hardware glossary:

NUMA (non-uniform memory access) is a method of configuring a cluster of microprocessor in a multiprocessing system so that they can share memory locally, improving performance and the ability of the system to be expanded. NUMA is used in a symmetric multiprocessing (SMP) system. An SMP system is a "tightly-coupled," "share everything" system in which multiple processors working under a single operating system access each other's memory over a common bus or "interconnect" path. Ordinarily, a limitation of SMP is that as microprocessors are added, the shared bus or data path get overloaded and becomes a performance bottleneck. NUMA adds an intermediate level of memory shared among a few microprocessors so that all data accesses don't have to travel on the main bus.

NUMA can be thought of as a "cluster in a box." The cluster typically consists of four microprocessors (for example, four Pentium microprocessors) interconnected on a local bus (for example, a Peripheral Component Interconnect bus) to a shared memory (called an "L3 cache ") on a single motherboard (it could also probably be referred to as a card). This unit can be added to similar units to form a symmetric multiprocessing system in which a common SMP bus interconnects all of the clusters. Such a system typically contains from 16 to 256 microprocessors. To an application program running in an SMP system, all the individual processor memories look like a single memory.

When a processor looks for data at a certain memory address, it first looks in the L1 cache on the microprocessor itself, then on a somewhat larger L1 and L2 cache chip nearby, and then on a third level of cache that the NUMA configuration provides before seeking the data in the "remote memory" located near the other microprocessors. Each of these clusters is viewed by NUMA as a "node" in the interconnection network. NUMA maintains a hierarchical view of the data on all the nodes.

Data is moved on the bus between the clusters of a NUMA SMP system using scalable coherent interface (SCI) technology. SCI coordinates what is called "cache coherence" or consistency across the nodes of the multiple clusters.

SMP and NUMA systems are typically used for applications such as data mining and decision support system in which processing can be parceled out to a number of processors that collectively work on a common database. Sequent, Data General, and NCR are among companies that produce NUMA SMP systems.