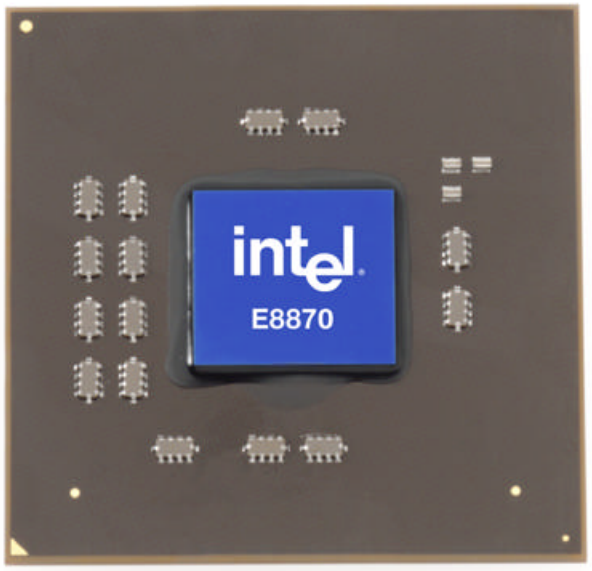


Intel® E8870 System Chipset: The Cost-Effective Solution for Scalable Servers

Intel E8870 system components permit the efficient manufacture of small and large configurations. Advantages built into the chipset ensure reliability, serviceability and performance.

Introduction

The Internet's evolution has generated complex requirements for enterprise server systems. There is increased emphasis on cost-effective servers that consume little power. At the same time, the market is requiring database servers that provide high levels of performance, scalability and availability.

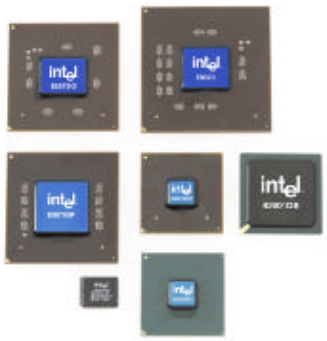
A number of system vendors have responded to these requirements by using standard components for entry-level servers and proprietary components for midrange and high-end servers. This dual-solution approach tends to drive up engineering costs. Intel has an alternative.

The Intel® E8870 chipset offers a scalable architecture for next-generation Intel Itanium® processor-based servers. Building blocks in the chipset allow designers to build multiprocessor server systems that scale from 2-16 processors. The E8870 architecture also permits system vendors to design components that scale beyond server systems with 16 processors; it potentially allows for complex systems that incorporate up to 512 processors.

Chipset Building Blocks

Intel E8870 architecture is based on the following building blocks:

- Scalable node controller (SNC)
- Scalable port switch (SPS)
- I/O hub
- I/O bridge
- Scalability port (SP)



Scalable Node Controller (SNC)

The SNC provides the required interface to processors, memory subsystems and firmware hubs. Its features include:

- Support for up to four processors
- 200-MHz DDR SDRAM support through a DDR memory hub
- Support for 32 dual in-line memory modules (DIMMs) resulting in support for up to 128 Gbytes of memory per SNC when using 1-Gbyte DDR devices
- Dual scalability ports (SPs) to connect to the scalability port switch (SPS) or the I/O Hub.

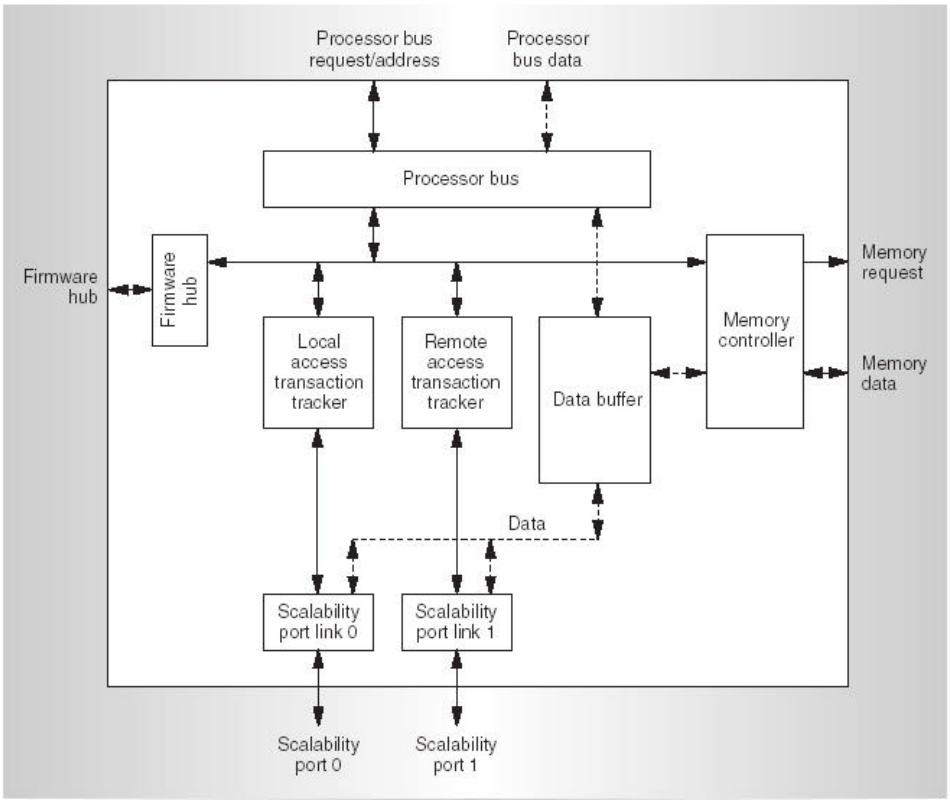


Figure 1: SNC high-level micro-architecture

Figure 1 shows major SNC components. The local-access transaction tracker monitors processor requests, converts processor requests to SP or memory controller requests, and returns responses to the processors. The remote-access transaction tracker follows inbound SP transactions until the necessary snoops and/or memory accesses are complete. The data buffer transports and holds data for the processor bus, memory and SP interfaces.

Scalability Port Switch (SPS)

The SPS is a coherent interconnect switch that connects scalability node controllers (SNC) and I/O hubs that use scalability ports (SPs). Its features include:

- Six SPs with an aggregate peak bandwidth of 38.4 Gbytes/s
- An integrated snoop filter that tracks the state of all cache lines in processor and I/O hub caches, reduces snoop probes to remote nodes, and supports the SP cache consistency protocol
- An internal interconnect comprised of a crossbar and network of buses for critical coherent traffic.

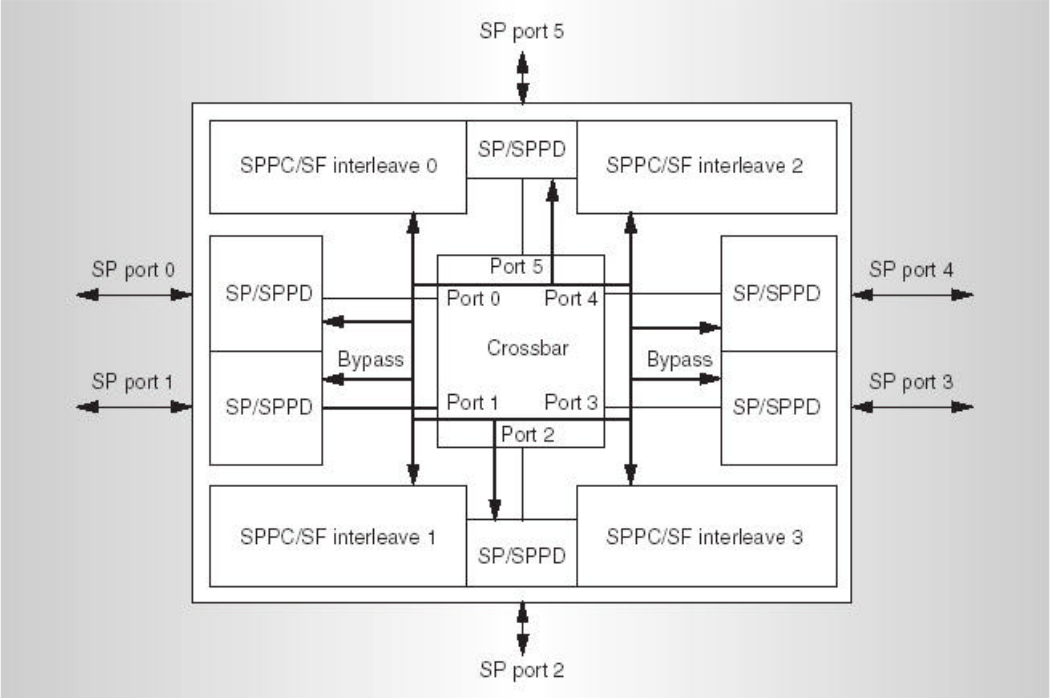


Figure 2: SPS high-level micro-architecture

Figure 2 shows the SPS micro-architecture.

Each of the six SPs implements the physical, link, and part of the SP protocol distributed layers (SPPD); the SPPD performs address/request packet routing and controls data transfers between ports.

Four SP protocol centralized (SPPC) snoop-filter units interleave for improved throughput and ease of physical design. They use a programmable protocol engine that processes requests, processes responses, and spawns transactions. The units also handle global ordering and contain anti-starvation logic that guarantees fairness between nodes.

I/O Hub

The I/O hub enables attachment to any type and number of I/O devices while providing a configurable I/O subsystem. Its features include:

- A prefetch engine
- Read caches to deliver full bandwidth on data return
- Two SP interfaces to connect to SPSs or SNCs
- Four hub interfaces with peak bandwidths of 1 Gbyte/s each
- A legacy I/O controller hub
- A PCI/PCI-X bridge.

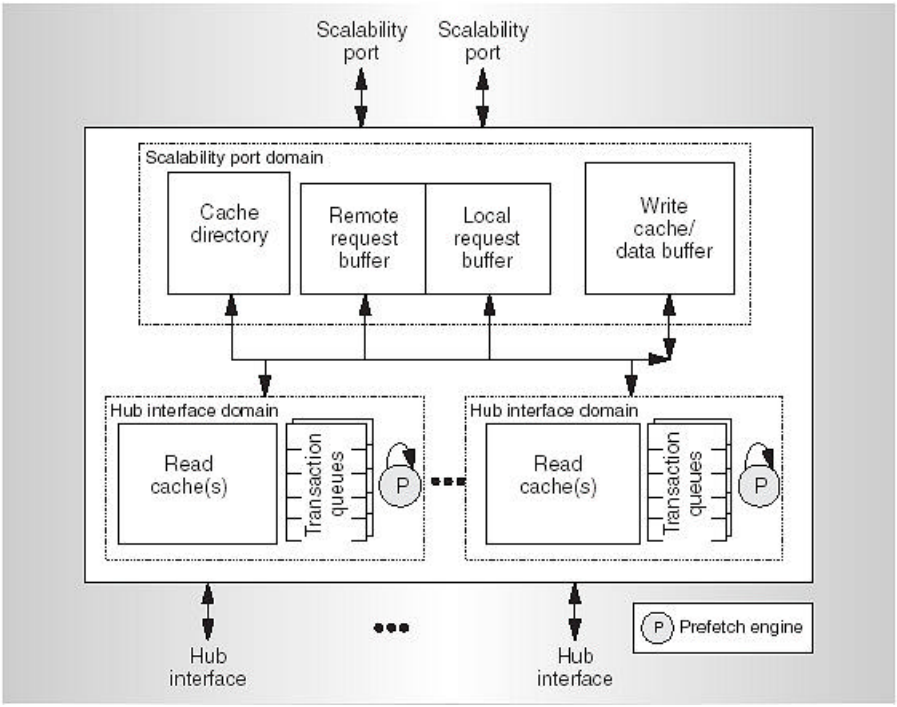


Figure 3: I/O hub micro-architecture

Each hub has a dedicated 4-Kbyte read cache (see Figure 3). Fully coherent read caches allow the use of an aggressive prefetching algorithm without exposure to stale data delivery. There is one write cache. Coherent write caching promotes the combining of write data into a cache line granularity, increasing efficiency and decreasing snoop overhead on the system.

A cache directory tracks the cache lines held in the multiple read caches and the write cache. It tracks duplicate entries of shared lines.

Buffers track coherent transactions issued by the I/O hub (specifically, the local-request buffer) and issued by other components (such as the remote-request buffer). The buffers are used to detect access conflicts and enforce cache consistency.

Scalability Port (SP)

The SP is a point-to-point cache-consistent interface designed to overcome the limitations of shared-bus-based architectures. It consists of three layers of abstraction: the physical, link, and protocol layers.

The physical layer uses a pin-efficient, simultaneous, bidirectional-signaling technology. The interface's transmitter sends clock information along with the data. The receiver uses the clock information to sample the data.

The link layer supports virtual channels and provides flow control and reliable transmission. It uses two virtual channels to build independent request and response interconnects on a single physical link. Flow control uses a credit-based scheme. The layer detects transmission errors and relies on a retry scheme for recovery.

The consistency protocol lets cache lines have modified, exclusive, shared, or invalid state at the caching agent. The protocol is built on the snoop-filter sparse-directory concept; it tracks lines in the caches rather than all lines in memory. The method increases performance by allowing storage of entire snoop filters on the same component as that of the directory state machine (not possible with a conventional directory).

Example Configurations

The chipset architecture supports two classes of shared-memory multiprocessor system architectures: the single-bus shared-memory architecture (scalable from 2- 4 processors systems) and the distributed, shared-memory architecture (scalable from 8-16 processors).

Single-Bus Shared-Memory Architectures

Figure 4 shows a four-processor server configuration that uses the E8870 chipset. All processors and memory controllers attach to a common bus. Each processor has a private cache and uses its internal bus interface unit to monitor memory accesses on the bus.

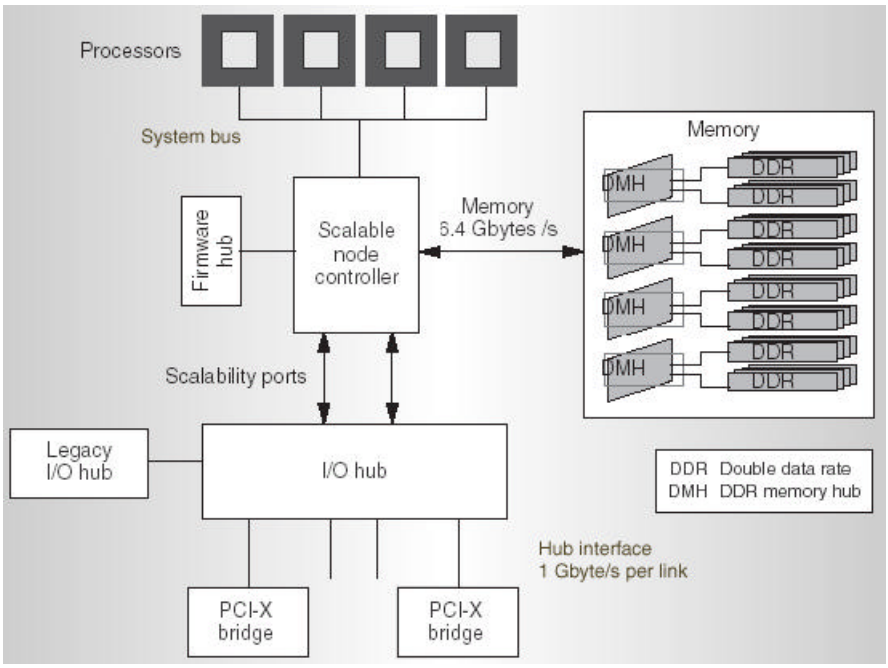


Figure 4: Single-bus shared-memory configuration with four processors

The main components are the scalable node controller (SNC) and the I/O hub. Note that:

- The SNC interfaces directly to the system bus.
- The main memory controller supports four memory channels; the double-data-rate (DDR) memory hub on each memory channel can control eight DDR dual in-line memory modules (DIMMs).
- The firmware hub serves as the system's boot ROM.
- The connection to the I/O hub is through a pair of the scalability ports (SPs); each SP provides 3.2 Gbytes/s of bandwidth in each direction.
- The I/O hub supports four hub interfaces to connect to peripheral component interconnect (PCI) / PCI-X devices. A narrower version of the hub interface supports legacy I/O devices.

Distributed, Shared-Memory Architectures

To scale beyond four processors, the Intel E8870 chipset uses a multi-node scheme. Nodes of four-processor subsystems interconnect using a pair of scalability port switches (SPS).

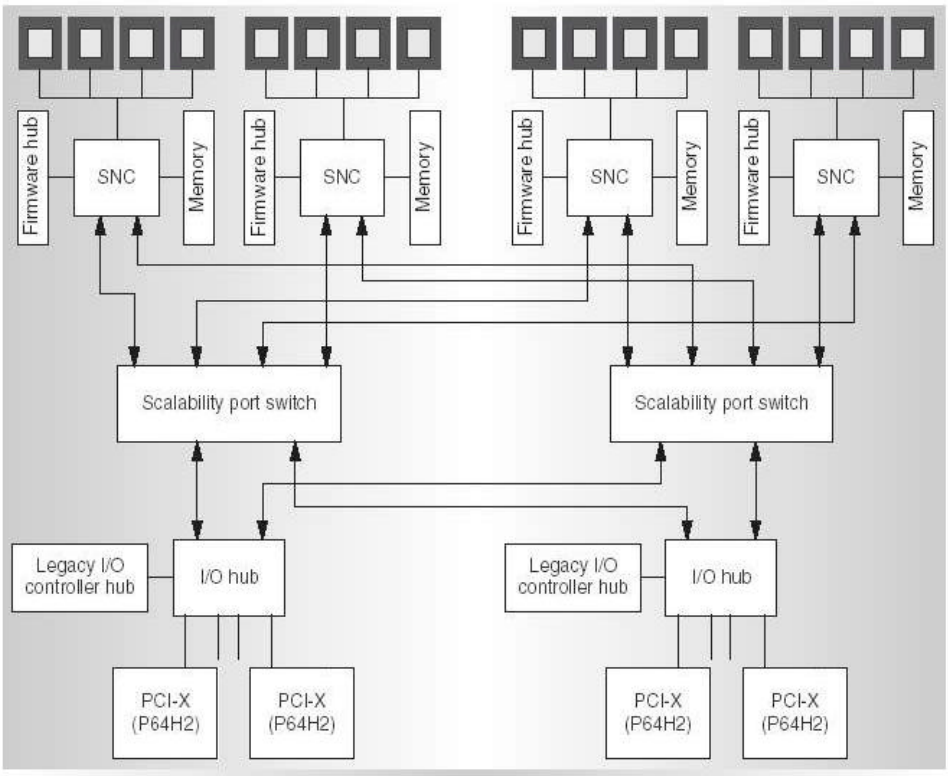


Figure 5: System configuration with 16 processors; the total scalability port bandwidth is 3.2 Gbytes/s x 6 x 2 = 38.4 Gbytes/s per direction (76.8 Gbytes/s total)

The configuration shown in Figure 5 distributes physical memory across nodes. Logical memory, however, is visible from all processors as a single physical space.

Older distributed-memory architectures exhibited significant differences in latency between local and remote accesses. Such architectures typically required software optimizations to mitigate latencies. Because of advantages internal to Intel E8870 chipset, the ratio of remote-to-local latency in the configuration shown above is about 2.2. This ratio does not generally require optimization for scalable performance except in very large systems.

When configuring large systems, system designers should use the hot-page mechanism. This technology facilitates software optimization and allows designers to bring average memory latency close to local latency in systems with 16 processors.

Reliability, Availability, and Serviceability

The Intel E8870 chipset provides a 24/7 computing environment for enterprise applications. Some of the chipset's advantages are listed below.

Error Detection

Error detection logic is built into the chipset; over 50 unique errors are detected. Parity-checking or error-correcting code protects all data paths. ECC protects the snoop-filter contents and the protocol engine flags protocol violations on primary interfaces (including the SP). Both error typing and signaling are compatible with the Itanium processor machine-check architecture.

Fault Resilience

The chipset enables fast reset and reboot in a degraded mode after a component or interconnect failure. For example, if an SP fails, the system controller resets and reconfigures the system to use only one SPS switch. This multi-path architecture protects against a single point of failure.

Failure Tolerance

The chipset implements ECC in the memory subsystem in a manner that tolerates SDRAM failures. That is, a device failure is a correctable error. Each component provides a SMBus-2.0-compliant interface with access to all internal registers.

Hot Replacement

The chipset enables adding, removing, or replacing a processor/memory node or an I/O node in a running system. The SP supports the software sequencing needed for hot replacement including connecting/disconnecting the interface during runtime and signaling (via interrupts) on connection/initialization events.

Performance Optimizations

Memory, caching, and tuning optimizations give E8870-based solutions big system performance. Some optimizations are discussed below.

Memory Interleaving and Reordering

Optimized memory interleaving and reordering maximizes memory throughput. Reordering around page replace and turnaround penalties improves maximum sustained bandwidth by 12 to 30 percent; the actual numbers depend on the mix of read/write transactions.

I/O Caching and Speculative Memory Prefetching

Coherent I/O caches and prefetching hide I/O read latency, even in large multinode configurations. Our tests show these technologies can improve PCI performance by up to 68%.

The E8870 chipset uses a speculative memory prefetch algorithm to minimize memory access latencies. Read requests that correspond to cache content do not incur additional upstream latency. The prefetch engine monitors real time traffic and modifies the prefetch method depending on I/O mode. For example, a single data stream will result in the prefetch of up to eight cache lines; if there are two data streams, prefetch will adjust to a maximum of four cache lines for each stream.

SPS Snoop Filter

The E8870 snoop filter provides an efficient cache coherency implementation. The conventional method was to track data usage in an entire memory range. The Intel solution increases efficiency by reducing both the size of the monitoring task and the required imprint of the snoop filter logic. The smaller imprint allows the filter to be stored on the same component with the directory state machine.

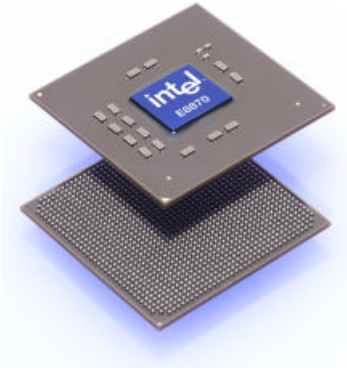
The resulting separation of the snoop filter from the memory agent allows more cost effective systems to be designed; system designers can use only SNCs and I/O hubs when an SPS is not required.

Hot page Technology

Large multi-node configurations have shorter latency for local-memory accesses. On these systems, software that is tuned to favor local-memory accesses will provide optimal performance. To aid in optimizing software for local memory accesses, the SNC component tracks accesses to each address location or range (the granularity is programmable). The tracking mechanism, called hot page, tracks local and remote accesses. Software developers can use hot page to identify hot spots in memory (pages frequently accessed by remote nodes) and optimize software to move such accesses to a local node. Hot-page can also aid other forms of software optimizations.

In Closing…

The Intel E8870 chipset provides flexibility to system designers. Its building blocks permit system designs that scale from 2 to 512 processors. System manufacturers can offer large configurations based on 4-processor subsystems that interconnect via scalability ports. Such design flexibility lets manufacturers offer differentiated products while minimizing development costs.



Reliability, availability, and serviceability features have traditionally only been available from proprietary chipsets used in large, high-end servers. With the E8870 chipset, system manufacturers can deliver a broad range of server products that are reliable, serviceable and highly available.

For more information on the Intel E8870 chipset, visit: http://developer.intel.com/design/chipsets/e8870

Disclaimers

Information in this document is provided in connection with Intel products. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by this document. Except as provided in Intel's Terms and Conditions of Sale for such products, Intel assumes no liability whatsoever, and Intel disclaims any express or implied warranty, relating to sale and/or use of Intel products including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright or other intellectual property right. Intel products are not intended for use in medical, life saving, or life sustaining applications. Intel may make changes to specifications and product descriptions at any time, without notice.

The Intel E8870 chipset may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's website at http://www.intel.com.