Cisco Unified Computing System Extended Memory Technology Overview

What You Will Learn

To address rising data center operations costs and increasing workloads, businesses look for ways to increase server computing platform efficiency through consolidation and virtualization. As line-of-business (LOB) applications, corporate databases, online transaction processing (OLTP), and web server applications continue to evolve, the memory capacity of server systems plays an important role.

The demand for more memory is propelled by 64-bit applications, operating systems, and the mainstream adoption of virtualization. This demand creates a need for simplified server architectures that provide centralized scale-out computing, with advanced management capabilities and innovation in x86 server memory architectures to support new applications with large memory requirements. This simplified architecture helps reduce ongoing acquisition, operating, and maintenance costs.

Cisco® Extended Memory Technology was developed to address these needs in x86-based computing environments. A crucial innovation of the Cisco Unified Computing System, Cisco Extended Memory Technology provides more than twice as much memory (384 GB) as traditional two-socket servers, increasing performance and capacity for demanding virtualization and large-data-set workloads. Alternatively, this technology offers a more cost-effective memory footprint for less-demanding workloads.

Introduction

The processor, memory, and I/O are three of the most important subsystems in a server. All are critical to the workload performance and reliability of a server. At any given point, one of these subsystems tends to become a performance bottleneck. When an application is described as CPU, memory, or I/O bound, that subsystem is likely to create a bottleneck for that application.

Memory is central to support for large corporate databases, virtualization, online transaction processing, and business resilience for important line-of-business applications such as enterprise resource planning (ERP) and customer relationship management (CRM). Therefore, the amount of system memory available significantly affects overall system performance. The reliability and performance of the operating system, mission-critical applications, and data are intrinsically tied to the memory capacity.

As servers have evolved, they have grown from single core to multicore, offering more powerful and energy-efficient processors. Research from Mindcraft Labs, however, has revealed that adding memory rather than processors represents a more cost-effective solution to the challenge of improving web server and database management

system (DBMS) server performance across operating platforms. With this view, cost-effective memory extension technology is the key to providing data center managers with the opportunity to scale server workloads and increase processor usage. These memory subsystems must be capable of keeping up with the low access time required by modern processors and the high capacity required by today's applications.

Options for Increased Server Workloads

Server workloads will continue to increase over time, forcing data center managers to address workload increases while balancing price, performance, power, cooling, reliability, and availability. There are multiple approaches to these challenges:

- Increase the number of servers
- Increase the number of sockets per server
- Increase the memory capacity per server

Increasing the number of servers drastically increases operating expenses (OpEx), management, and maintenance costs due to increased power, cooling, and infrastructure needs. This option results in a high total cost of ownership (TCO) and may not be the optimal solution for today's businesses.

Increasing the number of sockets per server increases the amount of memory available, as each socket offers more addressable memory. Increasing the number of sockets, however, may also incur hidden software licensing costs in addition to increases in capital expenditures (CapEx) and OpEx to support the server application environment. Again, the result is higher TCO.

In contrast, increasing the memory capacity per server is a viable, cost-effective, and scalable solution that can offer a better return on investment (ROI) and lower TCO for servers, enabling them to meet or exceed the goal of achieving midrange and high-end enterprise server workloads.

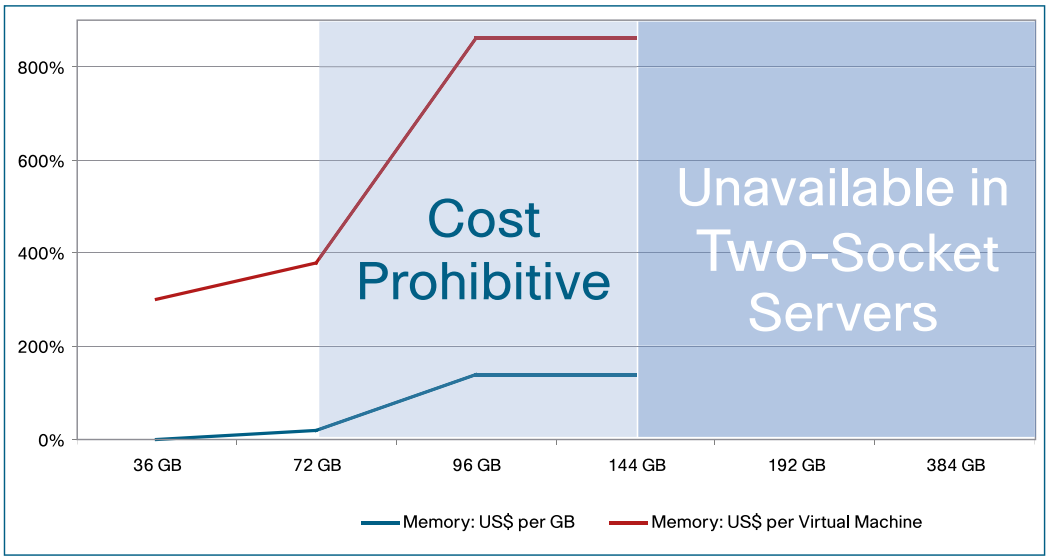
Memory on Virtualized Servers

Many companies are virtualizing servers, running multiple server workloads encapsulated in virtual machines on one physical server. The amount of memory typically installed in a server may be insufficient in a virtualized environment, as multiple virtual machines all requiring a memory footprint run on the same server. Each virtual machine consumes memory based on its configured size, plus a small amount of additional overhead memory for virtualization.

Virtualization management software allows administrators to specify a minimum and maximum amount of memory per machine; the minimum is guaranteed to be available, and the maximum can be reached depending on the memory use of the other virtual machines.

In practice, most production environments use from 2 to 4 GB per virtual machine. Using this assumption, Figure 1 shows the relationship between memory cost and total system memory as well as the exponential cost increases incurred based on an average amount of memory per virtual machine (based on two-socket Intel Xeon 5500 Series processors and estimated pricing of double-data-rate-3 (DDR3) memory in various sizes; the baseline is 36 GB of system memory). As the number of virtual machines increases, so does the amount of required memory per server. Even with the latest memory technology in servers, the desired memory configuration rapidly becomes either cost prohibitive or nonexistent. As a result, more or larger physical servers are typically deployed for a given set of virtualized workloads.

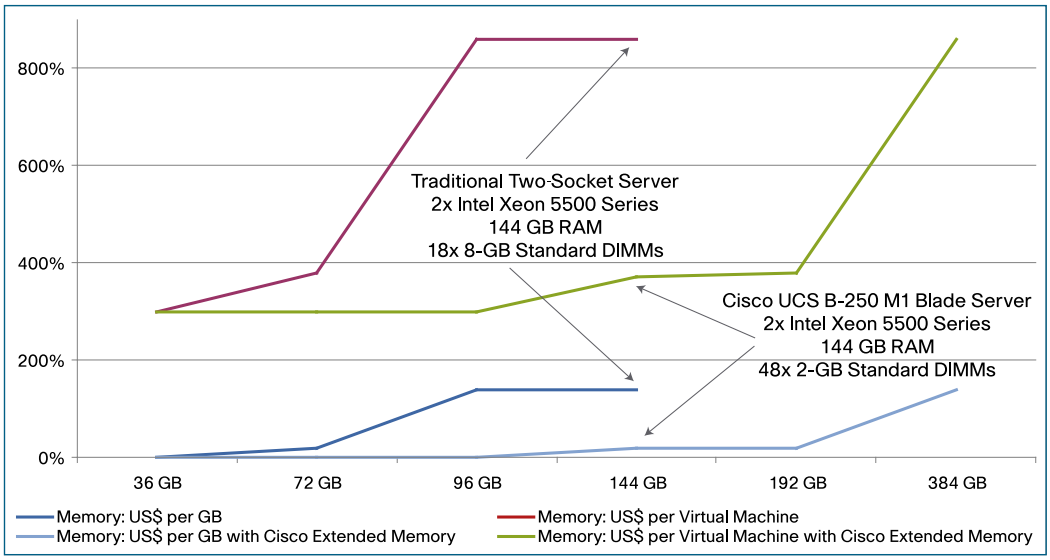
Figure 1. Two-Socket Memory Economics



Cisco Extended Memory Technology Innovation

Building on the power of the Intel Xeon 5500 Series processors in the Cisco Unified Computing System, Cisco's patented Extended Memory Technology enables up to 384 GB of memory on a single server (available on the Cisco UCS B250 M1 Blade Server and the Cisco UCS C250 M1 Rack-Mount Server). This technology provides more than double the industry-standard memory footprint when compared even to other Xeon 5500 Series processor-based systems. Cisco Extended Memory Technology enables memory scalability decoupled from the traditional cost. With reduced costs and larger-than-ever memory footprints, IT departments can now consolidate more applications and virtual machines more economically. Figure 2 shows the increased capacity and decreased cost possible with a two-socket system with Cisco Extended Memory Technology (based on two-socket Intel Xeon 5500 Series processors and estimated pricing of DDR3 memory in various sizes; the baseline is 36 GB of system memory).

Figure 2. Cisco Extended Memory Economics



Cisco Extended Memory Technology provides cost savings in addressing server workloads through:

- Reduced expenditure on additional servers for memory-bound applications
- Reduced software license costs, enabled by increased consolidation and use of two-socket systems
- Reduced power and cooling requirements, through deployment of fewer servers
- Improved server density efficiency in the data center
- Lower system capital expenditures through more economical memory costs
- Reduced capital, operational, and maintenance costs through deployment of fewer servers with additional capacity

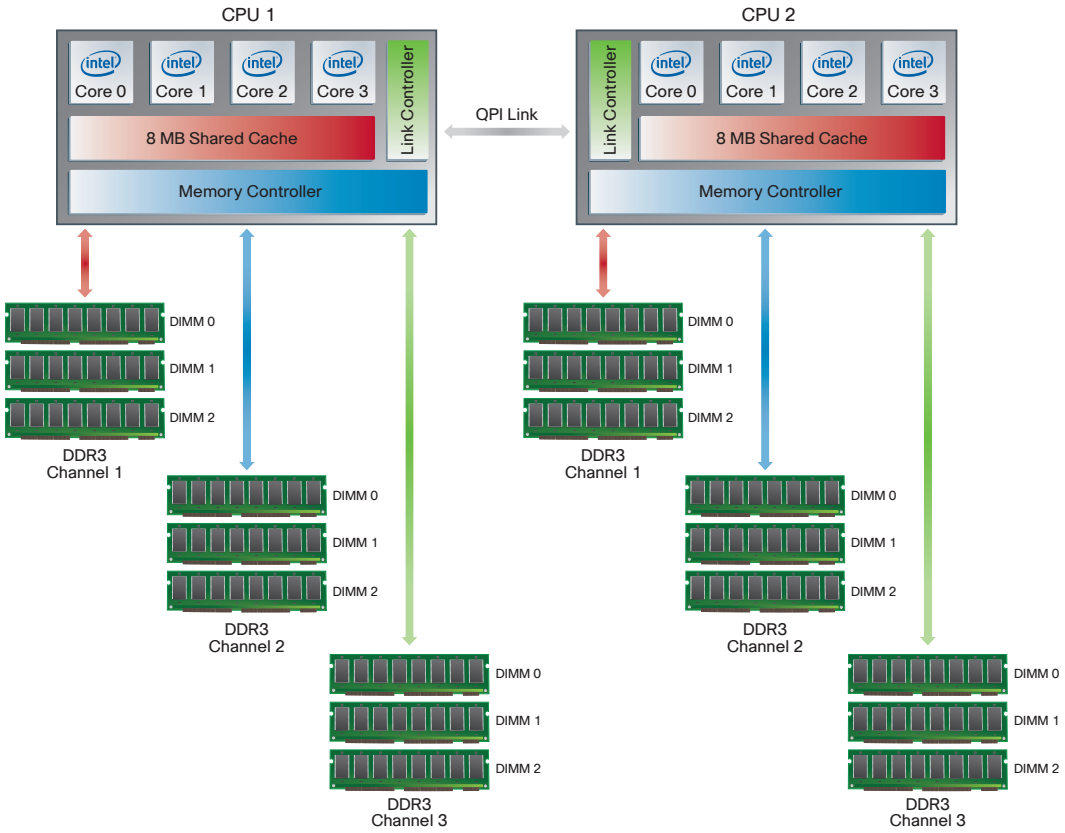
Intel Xeon 5500 Series Processor System: Memory Architecture

The development of the Intel Xeon 5500 Series processor microarchitecture comes with a new system architecture called Intel QuickPath Technology. This architecture integrates a memory controller into each microprocessor, dedicates specific areas for system memory to each processor, and connects processors and other components with a new high-speed interconnect.

The quad-core Intel Xeon 5500 Series processor CPU has an integrated memory controller with three channels and two Intel QuickPath Interconnects (QPIs). Intel QPI is a coherent point-to-point protocol introduced by Intel to provide communication between processors and I/O devices. Intel QPIs can be used for either CPU-to-CPU communication or CPU-to-I/O controller hub communication. Three DDR3 channels are supported per socket, and two sockets are connected through Intel QPI. Each Intel QPI link is capable of up to 25.6 Gbps or 6.4 gigatransfers per second (Gtps).

The Intel Xeon 5500 Series processor system has three DDR3 channels per socket. The dual-socket architecture Intel Xeon 5500 Series processor has two sets of memory controllers instead of one, leading to a 3.4X bandwidth increase compared to the previous Intel platform. The integration of the memory controller onto the chip also reduces latency by 40 percent. Power consumption is also reduced, since DDR3 is 1.5-volt technology, compared to 1.8 volts for DDR2. Figure 3 shows the Intel Xeon 5500 Series processor 18-DIMM memory system.

Figure 3. Intel Xeon 5500 Series Processor 18-DIMM Memory System



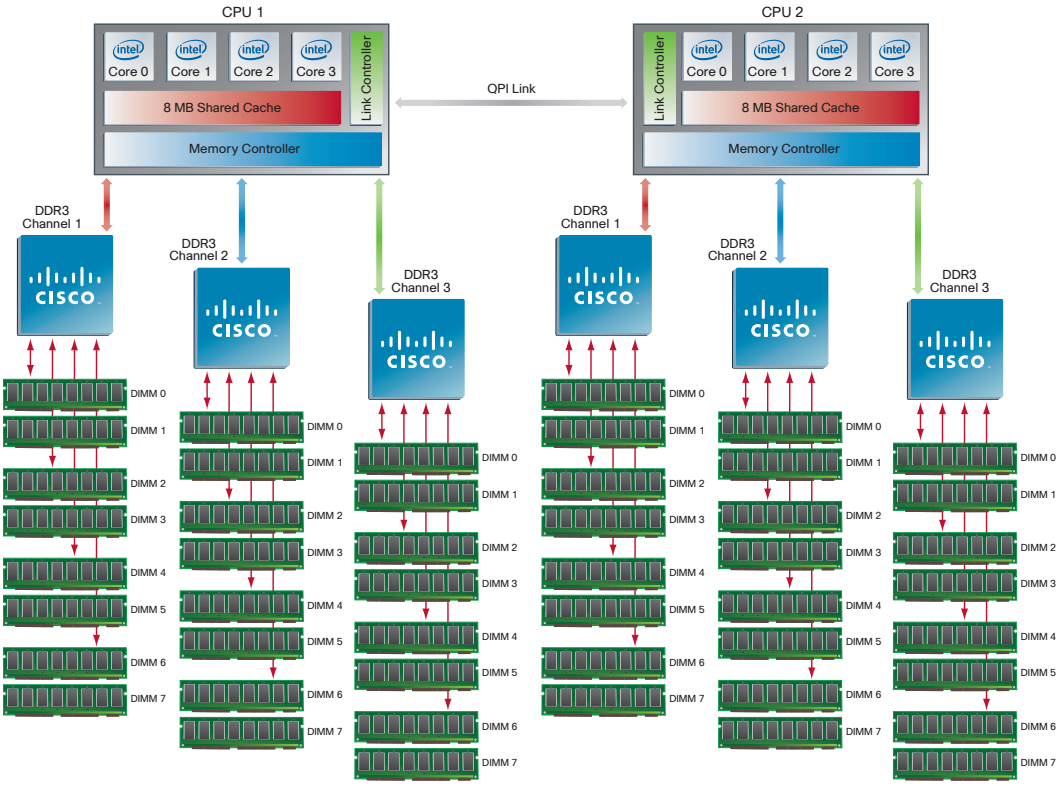
Cisco Extended Memory Technology

Cisco Extended Memory Technology provides connectivity for up to 32 ranks on eight RDIMMs, for up to 64 GB per DDR3 channel. With three channels per socket and both sockets populated, this equates to a maximum amount of memory of 384 GB using industry-standard DDR3 8-GB DIMMs. Each DDR3 channel is buffered and expanded to four DDR3 subchannels. Instead of directly connecting one DDR3 channel to the memory controller, four DDR3 subchannels are connected to the memory controller indirectly. Each DDR3 subchannel can support two single-, dual-, or quad-rank DDR3 RDIMMs (see Figure 4). The extended memory is transparent to software.

Cisco Extended Memory Technology overcomes the electrical issues associated with high DIMM counts using Cisco application-specific integrated circuits (ASICs) interposed between the processor and the memory DIMMs. This technology enables an increase in the memory capacity of conventional two-socket systems to up to 384 GB using industry-standard DDR3 memory.

Cisco accomplishes its memory extension by using custom ASICs to expand the number of DIMM sockets per single DDR3 channel and by bypassing the control signal limitations in the Intel Xeon 5500 Series processor standard design. This extension is performed at the electrical level, and it is completely transparent to the operating system and its applications. The BIOS is extended to initialize and monitor the ASICs and to perform error reporting.

Figure 4. Cisco Extended Memory Technology Architecture



The latency to local memory is marginally higher than in a system without memory expansion, and it is significantly lower than the latency incurred on access to memory on the next socket. This feature can result in significant performance improvement on memory access and can deliver significant latency improvement for disk access. This improvement is in addition to the performance improvements that can be achieved with a larger memory capacity in the server. The improvement in latency improves overall system performance and application response time.

The maximum memory configuration using 1-Gb DRAM on Cisco UCS B250 M1 or C250 M1 Extended Memory Servers is 384 GB (Table 1).

Table 1. Cisco Extended Memory Technology Capacity

Example Used: Cisco UCS B250 M1 Extended Memory Blade Server					
DIMMs per Channel	**DRAM Device Density**	**DIMM Type**	**DIMM Capacity (GB)**	**Channel Capacity (GB)**	**Total Memory (GB)**
8	1 Gb	1Rx4	2	16	96
		2Rx4	4	32	192
		4Rx4	8	64	384

Cisco UCS B-Series Blade Servers

Based on industry-standard server technologies, the Cisco UCS B-Series Blade Servers are crucial building blocks of the Cisco Unified Computing System, delivering scalable and flexible computing for today's and tomorrow's data center while helping reduce TCO.

The Cisco UCS B-Series Blade Servers incoporate numerous innovations that maximize the effectiveness of Cisoc Extended Memory Technology. Together, these innovative features allow data center managers to purchase and operate fewer, less-expensive servers, easily allocate workloads appropriately, and maximize their computing capacity. Ultimately, this scenario helps lower TCO and increase IT responsiveness to rapidly changing demands.

From a pure cost-savings standpoint, the Cisco UCS B250 M1 requires fewer devices and components to acquire, manage, power, and cool, especially for memory-intensive workloads. The typical resolution for memory-bound applications is to spread workloads across additional two-socket devices, or acquire more-expensive four-socket servers. The Cisco UCS B250 M1 provides the best for both scenarios. Each B-Series blade server supports up to two Intel Xeon 5500 Series multicore processors, which offer intelligent performance based on application needs, automated energy efficiency, and flexible virtualization support.

As an integral component of a unified computing system, the Cisco UCS B250 M1 can also reduce the amount of network and compute support infrastructure by more than 50 percent compared to traditional blade server environments.

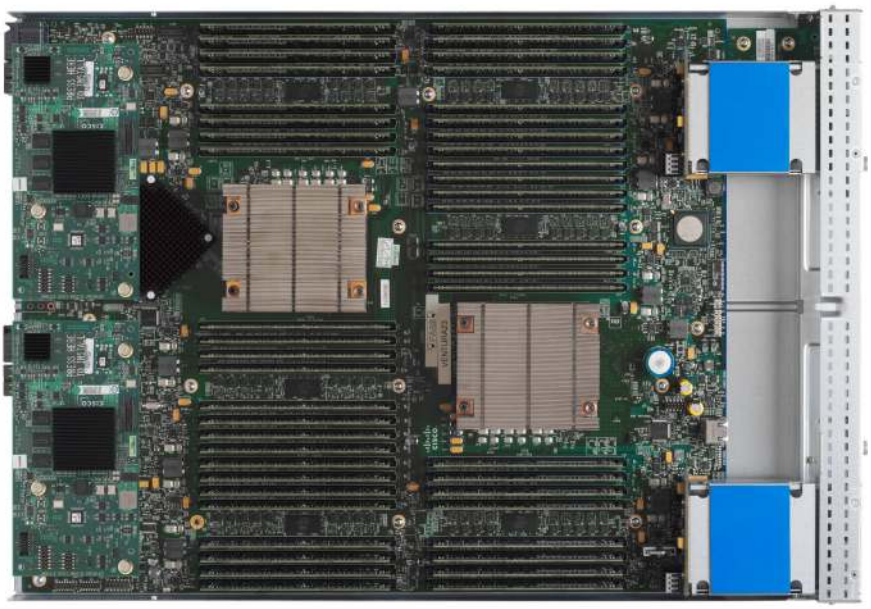
With up to 40 Gbps of redundant I/O throughput through two dual-port mezzanine card connections, the unified-fabric capabilities of the system enable a "wire-once" deployment model. Changing I/O configurations is simplified and no longer means reinstalling adapters and drivers and recabling racks and switches.

Computational scale can be maximized in virtualized environments by combining the extended memory with optional VN-Link capabilities and I/O virtualization. VN-Link allows SAN and LAN connectivity to be centrally configured and managed without introducing additional switching layers into virtualized environments. Now, network policies and configuration follow virtual machines as they are moved from one physical host to another, enabling more consistent and coherent policy enforcement. The ability to hold more data in memory, especially important in transaction-intensive scenarios, further improves processor availability.

Unified, embedded management allows policies defined in the Cisco UCS Manager software for memory-intensive applications to be easily applied through service profiles to selected blades. Different types of workloads can therefore be assigned to the most appropriate hardware. Having a common management domain for as many as 160 discrete servers allows IT managers to create pools of extended memory servers with specific characteristics and policies, simplifying deployment of large, nonvirtualized operating systems and databases.

Figure 5 shows the board layout and DIMM arrangement for an Intel Xeon 5500 Series processor-based Cisco UCS B-Series Blade Server with Cisco Extended Memory Technology (specifically, a Cisco UCS B250 M1 Extended Memory Blade Server). The blade server with extended memory provides cost-effective computing in application environments featuring virtualization or large databases in which memory is a critical factor for scaling application performance.

Figure 5. Cisco UCS B250 M1 Extended Memory Blade Server



Cisco UCS C-Series Rack-Mount Servers

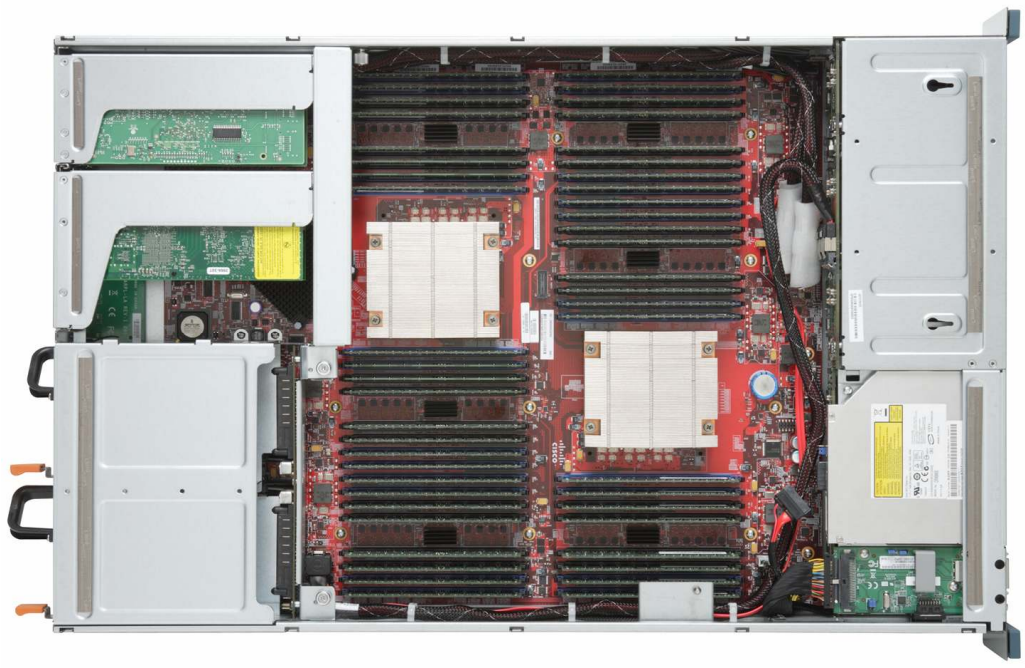
Like the Cisco UCS B250 M1 Blade Server, the Cisco UCS C250 M1 Rack-Mount Server can increase performance and capacity for demanding virtualization and large-data-set workloads. The two-socket, 2RU rack-mount server is designed to operate both in standalone environments and as part of the Cisco Unified Computing System.

The Cisco UCS C-Series Rack-Mount Servers provide additional architectural choice with a built-in migration path to unified computing for IT organizations that wish to take advantage of the Cisco Extended Memory Technology and other innovations, such as:

- Up to two quad-core Intel Xeon 5500 Series multicore processors (Figure 3) automatically and intelligently adjust server performance according to application needs, increasing performance when needed and achieving substantial energy savings when not.
- Increased I/O bandwidth and flexibility is provided by up to 5 PCIe cards in 3 low-profile, half-length x8 and two full-height, half-length x16 slots.
- Optional VN-Link capabilities allow all links to be centrally configured and managed without introducing additional switching layers into virtualized environments. I/O configurations and network policies move with virtual machines, helping increase security and efficiency while reducing complexity.
- When the C-Series is integrated within the Cisco UCS framework, unified, embedded management allows policies defined in the Cisco UCS Manager software for memory-intensive applications to be easily applied through service profiles to the servers.

Figure 6 shows the board layout and DIMM arrangement for an Intel Xeon 5500 Series processor-based Cisco UCS C-Series Rack-Mount Server with Cisco Extended Memory Technology (the Cisco UCS C250 M1). The rack-mount server with extended memory provides additional flexibility in architecting cost-effective computing environments to support virtualization or large databases in which memory is a critical factor for scaling application performance.

Figure 6. Intel Xeon 5500 Series Processor-Based Cisco UCS C-Series Rack-Mount Server with Cisco Extended Memory Technology



Memory Cost Analysis

The price per GB using 2- or 4-GB RDIMMs is not much different; however, 8-GB RDIMMs cost a substantial premium. As previously discussed, customers can scale memory capacity in the Cisco Unified Computing System with current Intel Xeon 5500 Series processors by providing 4-Gb capacity at 1-Gb pricing and availability.

The cost of memory changes dramatically over time. At the time of this writing, current memory costs were calculated for comparative study (see Table 2). The cost of the RDIMM per GB is another important factor. This cost does not increase linearly with RDIMM capacity.

Table 2. DDR3 DIMM List Prices

RDIMM Capacity	List Price	List Price per GB
2 GB	US$125	US$65 per GB
4 GB	US$300	US$75 per GB
8 GB	US$1200	US$150 per GB

Conclusion

Cisco Extended Memory Technology, using Intel Xeon 5500 Series processors, provides CapEx, OpEx, and server performance improvement to support the most demanding applications. It improves the server price/performance ratio, which ultimately yields a quicker and higher ROI with a lower TCO.

Cisco Extended Memory Technology, coupled with the power efficiency of Intel Xeon 5500 Series processors, can provide significant benefits such as:

- Reduced overall server operating costs through reduced power and cooling requirements
- Reduced server configuration costs, with the capability to support medium-size to large memory configurations using inexpensive RDIMMs
- Server memory configuration flexibility, with the capability to build very large memory configurations using the highest-density RDIMMs available on the market
- Reduced software costs for applications that are licensed on a per-socket basis
- Reduced need for server purchases for memory-bound application environments

Cisco Unified Computing Services

Using a unified view of data center resources, Cisco and our industry-leading partners deliver services that accelerate your transition to a unified computing environment. Cisco Unified Computing Services help you quickly deploy your data center resources and optimize ongoing operations to better meet your business needs. For more information about these and other Cisco Data Center Services, visit http://www.cisco.com/go/dcservices.

Why Cisco?

The Cisco Unified Computing System continues Cisco's long history of innovation in delivering integrated systems for improved business results based on industry standards and using the network as the platform. Recent examples include IP telephony, LAN switching, unified communications, and unified I/O. Cisco began the unified computing phase of our Data Center 3.0 strategy several years ago by assembling an experienced team from the computing and virtualization industries to augment our own networking and storage access expertise. As a result, Cisco delivered foundational technologies, including the Cisco Nexus™ Family, supporting unified fabric and server virtualization. The Cisco Unified Computing System completes this phase, delivering innovation in architecture, technology, partnerships, and services. Cisco is well-positioned to deliver this innovation by taking a systems approach to computing that unifies network intelligence and scalability with innovative ASICs, integrated management, and standard computing components.

For More Information

For more information about Cisco UCS B-Series Blade Servers, visit http://www.cisco.com/en/US/products/ps10280/index.html.

For more information about the Cisco UCS C-Series Rack-Mount Servers, visit http://www.cisco.com/en/US/prod/ps10265/rack_mount_promo.html.

You may also contact your local Cisco representative.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-525300-02 11/12