

SeaMicro SM10000 System Overview

Anil Rao, June 2010
anil@seamicro.com, www.seamicro.com

Overview

The SeaMicro SM10000 brings together in a single standards-based system, compute, storage, switching, server management, and load balancing. SeaMicro has reconceived the volume server as a single box cluster computer built around a parallel array of independent, low power CPUs with a traffic controller front end. The result is a system that uses 1/4 the power and 1/4 the space of the best in class volume server.

The SM 10000 is comprised of 64 compute cards, 8 storage cards, and 8 Ethernet uplink cards, all tied together with a high bandwidth, low latency super computer style fabric. The SM10000 is 10 rack units tall and draws less than two kilowatts of power, enabling users to deploy up to 2,048 CPUs in a standard 8KW rack. The SM10000 is a standards-based x86 system. It is 30 inches deep and fits in a standard 19 inch server rack. It is plug and play, and runs existing operating systems, applications, and management tools without the need for custom drivers or any software modifications or recompilation.

Compute: In the SM10000, each of the 64 compute cards holds eight single socket, dual threaded 1.6 GHz Intel Atom CPUs bringing the system total to 512 Atom CPUs. Each Atom CPU can be configured with 1 GbE or 2 GbE DDR2 memory. An Atom CPU can be carved into smaller units of compute using standard virtualization software.

Storage: Each of the eight storage cards supports eight, hot-swappable, 2.5 inch SATA HDD/SSD drives. The system can be configured to operate without disks. The drives can be bonded together in a RAID configuration.



Ethernet: Each of the eight Ethernet uplink cards supports either 8 x 1 GbE ports or 2 x 10 GbE ports. The basic I/O configuration is either 8 x 1 GbE or 2 x 10 GbE; a fully loaded system can be configured with either 64 x 1 GbE or 16 x 10 GbE uplinks.

Fabric: In the SeaMicro SM10000, the 512 CPUs, the storage cards, and the Ethernet uplink cards are linked together with a 1.28 terabit per second super computer style fabric. The fabric provides each Atom CPU with up to 2.5 Gbps of bandwidth. All intra SM10000 communication is done over this fabric.

System and Redundancy: The SM10000 provides both hardware

and software redundancy. All modules including compute cards, Ethernet uplink cards, storage cards, disks, power supplies, and fan trays can be hot-swapped. The system can be configured with multiple redundant management modules that can operate either in an active/standby or an active/active mode.



System Details

Compute: In the SM10000, the basic building block is a credit card-sized compute block, comprised of an Intel Atom CPU and its chipset, DRAM, and a custom SeaMicro ASIC. All the other components normally found on a standard motherboard have been removed by a patented hardware-based CPU I/O virtualization technology. Eight of these credit card-size blocks fit on a 5 x 11 inch motherboard (see figure above); 64 of these hot-pluggable motherboards are packed into a 10 rack unit SM10000, for a total system density of 512 Atom CPUs.

Fabric Technology: SeaMicro has developed interconnect fabric technology capable of linking together hundreds of these credit card-sized compute building blocks. The fabric is constructed by linking SeaMicro ASICs in a multi-dimensional torus. The fabric is FLIT-based and wormhole routed, with integrated virtual channel technology. The fabric has the ability to transmit and receive different classes of traffic enabling both loss sensitive storage and loss insensitive Ethernet traffic to be transferred on a single infrastructure. These technologies combine to produce resilient, low-latency, high-bandwidth links at very low cost while providing 1.28 terabits-per-second of fabric bandwidth.

Storage and Disk: The SeaMicro SM10000 can be configured with 0 to 64 2.5 inch SATA hard disk drives (HDD) or solid state drives (SSD). The 512 CPUs in the system can be allocated portions of a disk or whole disks. A physical disk (HDD or SSD) can be divided into multiple virtual disks – from 2GB to the maximum capacity of the disk – and assigned to one or more CPUs. Data resiliency is maintained by marking a disk to be part of a RAID pool or by assigning multiple disks to a CPU. The system can be configured to run with or without disk, ensuring the flexibility to appropriately provision storage for the desired applications.

A disk (either physical or virtual) can be configured to be in “write” mode for one server and “read-only” mode for the other 511

servers, allowing users to store “read-only” copies of data to be shared among multiple processors. Multiple “read-only” disks can be created for redundancy or for different pools of CPUs. This “read-only” disk enables data center operators to update software for the entire system by updating the software just once – a huge operational advantage over existing systems where an update is required for each individual server.

Ethernet Uplinks: SeaMicro Ethernet uplink cards provide uplinks from the supercompute fabric to the external data center network via 1 Gigabit Ethernet or 10 Gigabit Ethernet interfaces. The interfaces support link aggregation (IEEE 802.3ad) with LACP both within and across I/O cards, providing redundancy against link and I/O card failure. In the SM10000, inter-server communication traffic remains on the fabric and does not consume any uplink bandwidth, enabling customers to configure uplinks to their core switches based on the desired bandwidth.

The uplink cards also include a management processor and provide a range of value-added features never before included in a server. For example, algorithms running on the management processor and dedicated hardware can implement MAC aggregation and hiding. This enables the system to be configured to expose as few as one and as many as 64 MAC addresses while continuing to provide an IP address and a MAC address to each CPU (or virtual machine instance).

Load Balancing and Dynamic Compute Allocation Technology™ (DCAT). DCAT integrates CPU management and stateful load balancing, enabling intelligent load distribution across the CPUs. The SeaMicro system software constantly polls the health of, and the workload on, all the CPUs in the system. Based on this information, SeaMicro’s DCAT transparently provisions the CPUs and dynamically programs the load balancing hardware to direct traffic to one group of CPUs and away from another. The load balancing technology works by creating virtual IP addresses, which can be assigned to pools of compute as small as one CPU and as large as 512 CPUs (or 4,096 virtual machines in a virtualized environment). The stateful hardware load balancer then distributes flows across these pools of servers using user selected load-management algorithms including round-robin, least connections, and max connections. CPUs can be added or removed from a virtual IP pool dynamically based on predetermined rules. For example, traffic can be directed to a pool of CPUs to ensure they are operating in the maximally efficient range, while allowing other CPUs to enter deep sleep mode or even be turned off. Similarly, a utilization threshold for a pool of compute can be set, and if met, CPUs can be dynamically added or removed from the pool.

System Management and Terminal Server: The SeaMicro SM10000 simplifies operations by providing a simple software management environment. The management software runs on separate and redundant control plane processors on the Ethernet uplink modules (not on one of the 512 CPUs available for user software) and provides value added features including:

- *Integrated DHCP Server:* SeaMicro’s integrated DHCP can automatically assign and manage IP addresses to all the 512 CPUs (or logical interfaces created therein) and also provides for IP persistence and automatic DHCP renewal. The assigned IP addresses can be either private or public.

- *PXE Install and Boot:* Using a single command line option, all 512 CPUs can be installed with an off-the-shelf operating system and application software. The CPUs can be configured to PXE boot each time, or can be configured to PXE install once and thereafter boot from local disk.
- *Built-in Terminal Server:* The SeaMicro system incorporates a built-in terminal server concentrator for all of the 512 CPUs for remote console management and monitoring. All functions normally accessible through a console session are available to the customer using the built-in terminal server, eliminating cumbersome external access concentrators and their associated cables.
- *System Management:* The SeaMicro operating environment provides users the ability to monitor and manage the servers, storage and Ethernet infrastructure using a rich array of user interfaces including CLI, SNMP, Syslog, and/or IPMI.

Redundancy and Reliability: The SeaMicro system implements redundancy in both hardware and software. At the hardware level, all major subsystems are redundant and hot swappable, including compute cards, disks, network interface cards, power supplies, and fans.

At the software layer, customers can configure the system to run active/standby software on two separate management cards or across up to eight management cards providing resiliency against multiple failures. In the event of a failure, standby software will assume the responsibility of managing the system without any manual intervention. SeaMicro’s software also manages the internal terabit fabric and has the intelligence to configure the hardware to route traffic around failure using multiple alternative fabric paths. SeaMicro’s modular management software provides process isolation and modularity with each major process operating in its own address space – thereby increasing system availability and reliability.



SM10000 Summary

The SM10000 is a standards-based X86 system that is designed to replace 40 1 RU dual socket quad core servers, the top of rack switch, the terminal server, and the load balancer. The SM10000 is built from 512 Intel Atom processors and draws 1/4 the power and 1/4 the space of the best in class server on the market, while requiring no modifications to software.