

Implementing Storage Class Memory with HLNAND™

July 2009



INTRODUCTION

Replacing or augmenting rotating magnetic disk storage with solid state storage promises vast improvements in computer performance and power consumption. This technology has been termed Storage Class Memory or SCM¹. While there are many emerging memory technologies with potential for SCM application, this paper focuses on proven NAND Flash technology. Current mainstream NAND Flash memory devices have a slow 40MB/s interface that does not allow many devices to be connected to a single channel. Today's HLNAND devices feature a 266MB/s interface and support a virtually unlimited number of devices on a channel, while also offering lower interface power. Using HLNAND, Storage Class Memory is viable today using proven NAND Flash technology.

STORAGE CLASS MEMORY

Today's mainstream computer memory is organized hierarchically as shown in Figure 1. A pyramid with CPU located at the top indicates that there are smaller amounts of memory located near the CPU and greater amounts located further away.

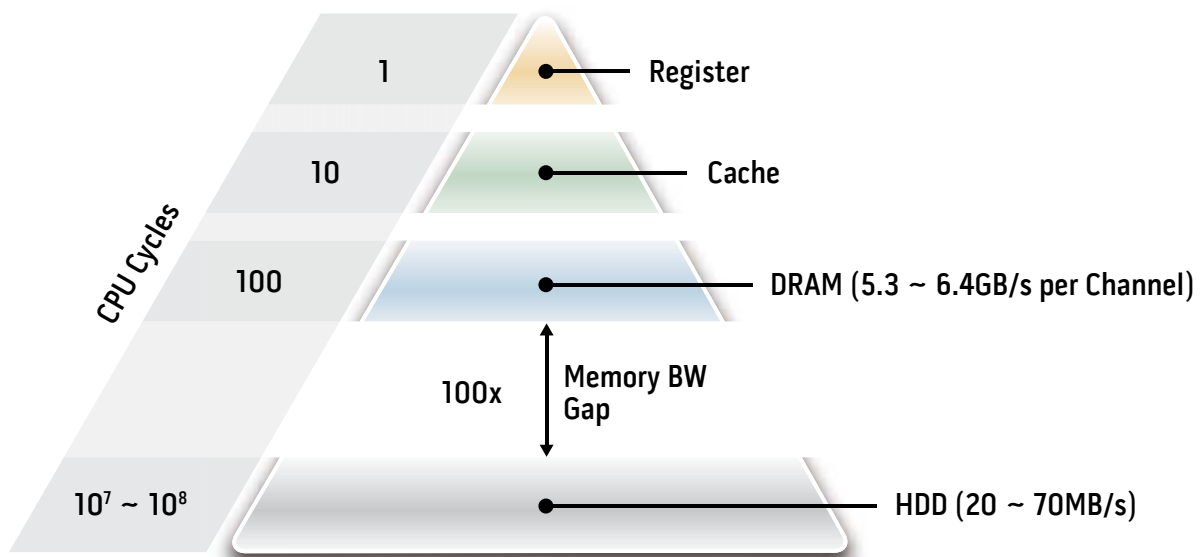


Figure 1. Mainstream Memory Hierarchy

Closest to the CPU are registers within the CPU itself. These registers can be accessed by the CPU within a single CPU clock cycle, typically several hundred pico-seconds at today's multi-GHz clock frequencies. Several levels of cache memory, L1, L2, and possibly L3 cache, are also located within the CPU and can be accessed within approximately 10 clock cycles. The next level of hierarchy is DRAM main memory which has an access latency of 50ns or roughly 100 CPU clock cycles. Mainstream DDR2-800 SDRAM modules provide up to 6.4GB/s bandwidth. To this point, the memory hierarchy is well balanced, with each level down providing several orders of magnitude more memory with a single order of magnitude increased latency. However, below the DRAM layer there is a huge gap in both latency and bandwidth. Rotating magnetic Hard Disk Drives (HDD) require several milli-seconds seek time for the head to swing into position over the desired track. This translates to more than 10 million

1. R.Freitas and W.Wilcke, "Storage-class memory: The next storage system technology", IBM J. RES. & DEV. VOL. 52 NO. 4/5 JULY/SEPTEMBER 2008.

CPU cycles. Clearly if the CPU requires data from the hard disk it will be waiting for an eternity. Once the head is positioned over the appropriate track, the data transfer from HDD to DRAM will also be slow due to the rotational speed of the disk and the track density. As a result, the HDD provides only 1% of the bandwidth available from DRAM.

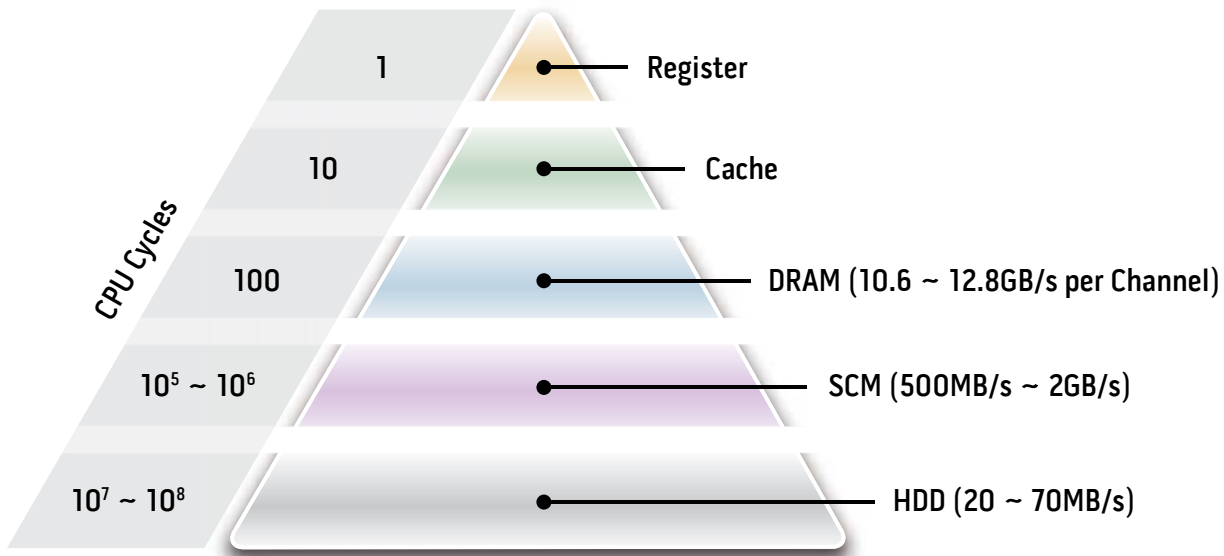


Figure 2. Memory Hierarchy with NAND Flash based Storage Class Memory

Figure 2 shows a memory hierarchy with HLNAND-based SCM filling the gap between DRAM main memory and HDD. Here the DRAM has been upgraded to DDR3-1600 where a single module can provide up to 12.8GB/s. Single bit-per-cell (SLC) NAND Flash has a page read time of 25µs and a page program time of 200µs, while two bit-per cell (MLC) NAND Flash delivers a page of read data in 60µs and programs a page in 600µs. This results in a latency ranging between 10^5 and 10^6 CPU clock cycles depending on the cell technology and whether a read or write operation is desired. HLNAND-based SCM provides a latency improvement of two orders of magnitude over HDD for accesses that cannot be serviced by the DRAM main memory layer. Furthermore, the bandwidth is better matched. SCM based on HLNAND can provide up to 2GB/s bandwidth depending on the number of HyperLink channels. Also, depending on the system, the HDD may be eliminated entirely. For example, portable computers may not require HDD.

SYSTEM TOPOLOGY

A typical system topology is shown in Figure 3. A CPU along with a chipset comprising a Northbridge and a Southbridge provide connections to system peripherals. The Northbridge is connected to the CPU via the Front Side Bus (FSB) and provides high bandwidth connections to DRAM main memory and a PCI Express (PCIe) interface. The DRAM interface may support dual channel (2 x 64bit) DDR2 or DDR3 modules delivering 12.8GB/s to 25.6GB/s. The PCIe interface is comprised of multiple bidirectional serial links, running 5GT/s transfer rate each in v2.0 of the PCIe standard. Due to coding overhead, each link or lane provides a net 4Gb/s (500MB/s) data rate. A Northbridge PCIe interface typically supports up to 16 lanes to deliver 8GB/s. The most common use of the PCIe interface is for high-performance graphics cards, although other peripheral devices can be supported.

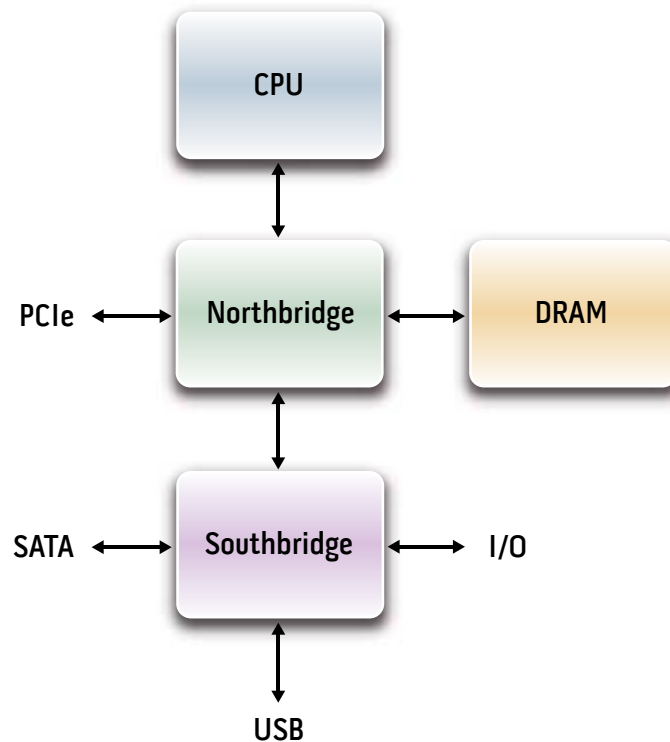


Figure 3. Computer System Topology

The Northbridge is connected to the Southbridge via the Direct Media Interface (DMI) bus to access lower bandwidth peripherals. The Southbridge supports a variety of I/O interfaces including USB and ports for keyboard, mouse, and other low-speed peripherals. The Southbridge also includes the HDD interface which is most commonly Serial-ATA or SATA. The current version of SATA specifies a single 3Gb/s serial link while the next version will double this to 6Gb/s. Similar to PCIe, coding overhead reduces the SATA physical bandwidth by 80%, thereby providing 300MB/s to 600MB/s usable bandwidth. These levels of performance are more than adequate for HDD, where bandwidth is limited by the mechanical constraints of rotating magnetic disks, but they are not sufficient for SCM.

Solid State Drives (SSD) employing NAND Flash memory have been available for several years. SSDs commonly employ established HDD interfaces, such as SATA, which connect via the Southbridge. As such, they are limited in bandwidth and latency by the HDD interface and the distance from the CPU. A typical SSD may contain 64 or more individual NAND die. Assuming just the conventional asynchronous 40MB/s NAND Flash interface, these 64 die together are capable of delivering 2.5GB/s, 10 times the capacity of SATA 3Gb/s. By utilizing the conventional HDD interface, clearly a lot of performance is left on the table. There are many other examples of an emerging technology being misapplied to directly replace an existing standard.

Another problem in implementing high-performance SCM with NAND Flash is the difficulty in connecting multiple NAND die to a controller chip. Conventional NAND Flash employs a multi-drop bus architecture which suffers loading problems and reduced speed if more than 8 devices are connected to the same channel. As a result, SSD controllers typically support at least 8 parallel channels, consuming several hundred pins and dissipating significant power to deliver 320MB/s. HLNAND employs the HyperLink interface where devices are interconnected in a serial point-to-point topology. Since the loading is point-to-point regardless of how many devices are interconnected, the full

266MB/s bandwidth per channel can be maintained. Figure 4 shows a Multi-Chip-Package including 4 conventional NAND devices and a bridge chip interfacing to an external HyperLink ring. Each channel requires 24 active signals for both input and output.

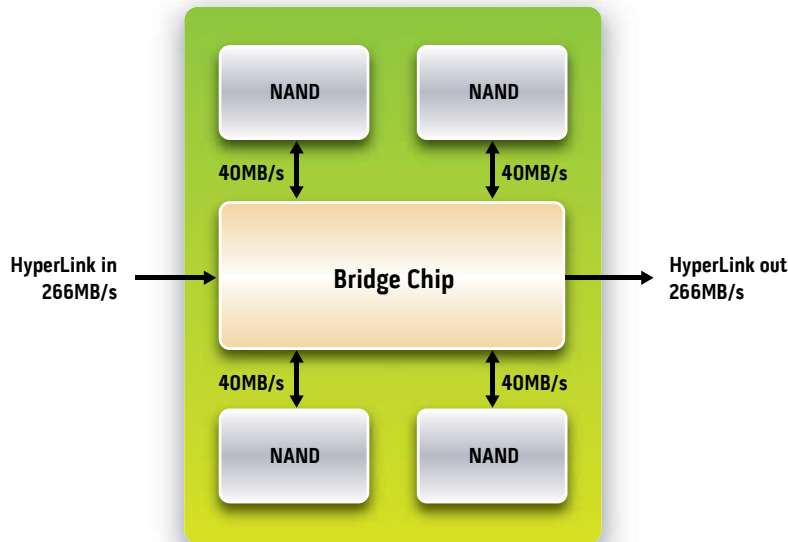


Figure 4. HLNAND MCP

To fully leverage the inherent performance of NAND Flash technology, SCM employing HLNAND devices should be connected through the Northbridge rather than the Southbridge. The multi-lane PCIe connection available here offers an order of magnitude increase in bandwidth. A controller with 4 parallel HyperLink channels requires fewer pins than 8 conventional NAND channels, while delivering more than 1GB/s bandwidth with reduced I/O power. This would be appropriate for a 4 lane PCIe interface. Larger configurations capable of filling the bandwidth of 8 lane and 16 lane PCIe interfaces can also be contemplated.

A final requirement for Storage Class Memory is the ability to upgrade non-volatile memory capacity. Today's SSD products have fixed capacity due to the use of soldered-on memory components. The multi-drop bus architecture of conventional NAND Flash memory cannot support multiple removable modules because the load of additional modules will reduce channel bandwidth. HLNAND supports removable modules that allow user configurable SCM.

HLNAND MODULE

Figure 5 shows a block diagram of a module incorporating 8 HLNAND MCPs, for a total of 32 NAND die. With 16Gb individual NAND die, the module represents 64GB non-volatile memory capacity. With 8 stacked die per module, 128GB capacity can be realized. The module employs the widely used 200 pin DDR2 SO-DIMM form factor.

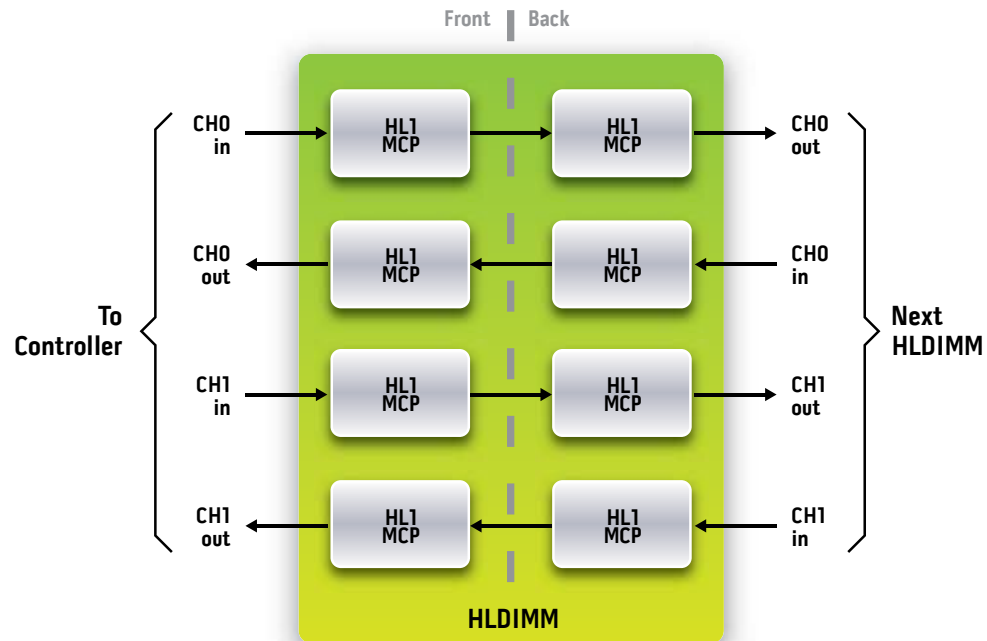


Figure 5. HLDIMM Block Diagram

A configuration of 4 HLDIMMs organized in a 2-channel configuration is shown in Figure 6. Each module only has to communicate with the neighboring module to keep signal paths short and allow high-speed operation. A loop-around connection located on the motherboard at the module farthest away from the controller completes the ring. If the full memory configuration is not required then the HLDIMMs should be located close to the controller and a loop-around module should be plugged into the next socket. This configuration offers 533MB/s read bandwidth and 533MB/s write bandwidth for an aggregate bandwidth of 1066MB/s.

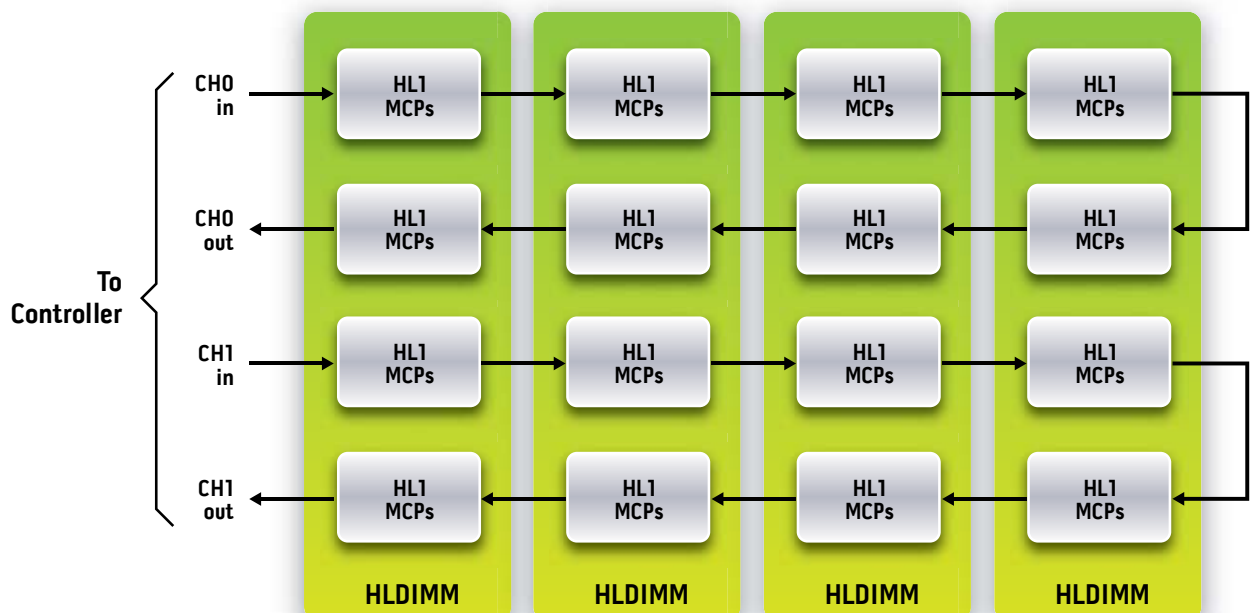


Figure 6. Two Channel Configuration using 4 HLDIMMs

If more data throughput is desired, an interleaved configuration using identical modules can be implemented as shown in Figure 7. This 4 channel configuration offers 1066MB/s read bandwidth and 1066MB/s write bandwidth for an aggregate bandwidth of 2133MB/s.

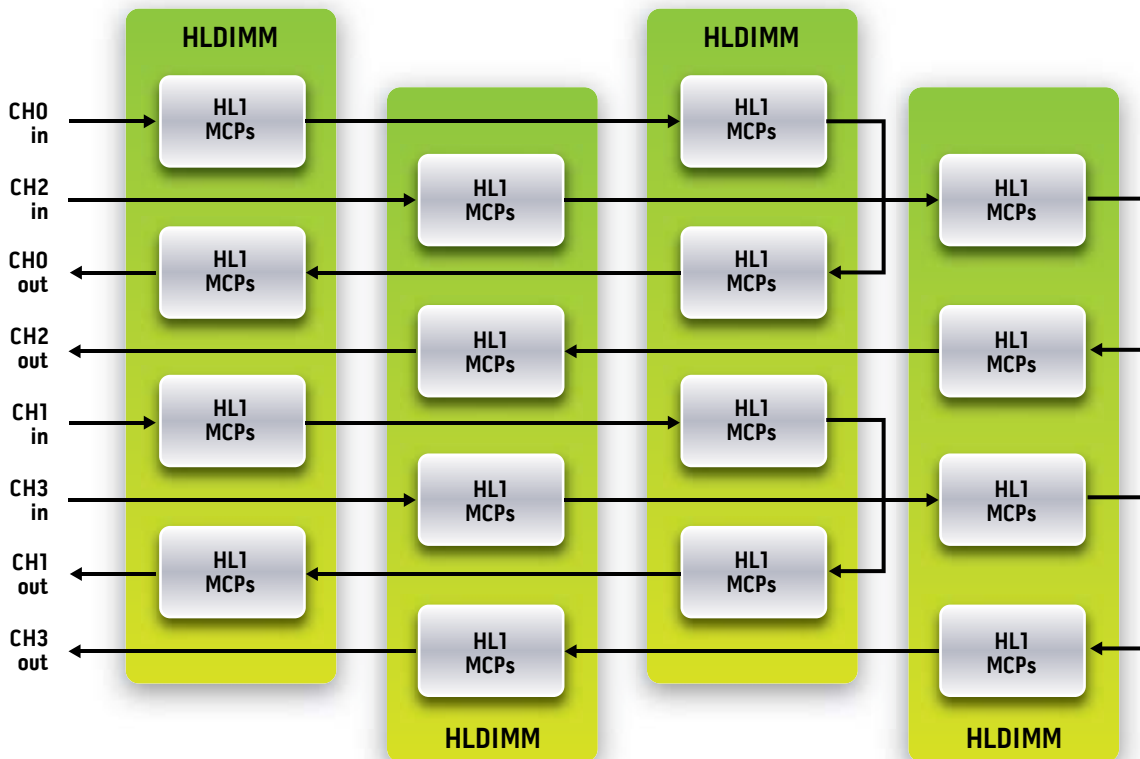


Figure 7. Four Channel Configuration using 4 HLDIMMs

Figure 8 shows a motherboard layout with a controller and 8 HLDIMM sockets. The controller interfaces to the system with a high bandwidth interface such as 4-lane or 8-lane PCIe. Aggregate NAND bandwidth of 1GB/s or 2GB/s to match the PCIe system interface can be achieved with non-interleaved or interleaved channel configurations. Fully populated memory capacity of 512GB or 1TB is achieved using 64GB and 128GB HLDIMM modules within only 60cm² of motherboard area. For comparison with a system using conventional NAND components to achieve similar capacity and performance, 25 separate NAND channels operating at 40MB/s would be required. This will consume close to 400 active pins on the controller in addition to power and ground pins.

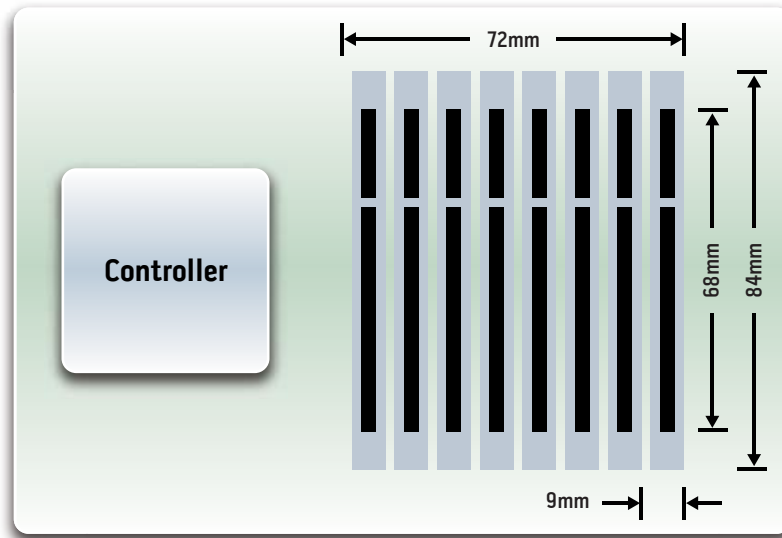


Figure 8 Motherboard layout with 8 HLDIMM sockets

HLDIMM modules allow user upgradeable SCM systems. Compared to conventional SSDs using soldered-on NAND components, the module approach offers significant benefits to the supplier. The supplier does not have to commit to fixed configurations during board manufacturing; rather the appropriate number of modules can be plugged in during order fulfillment. This saves significant inventory since memory represents 80 – 90% of the cost of such systems.

Another benefit of HLDIMM modules versus soldered-on conventional NAND components is that even a minimally populated system provides full bandwidth. In a partially populated conventional system each channel may support only 2 NAND packages. Thus, there will be unused channels if the memory is configured to less than half of full capacity. Figures 9 and 10 show that both memory bandwidth and IOPS remain high even with minimally populated memory configurations.

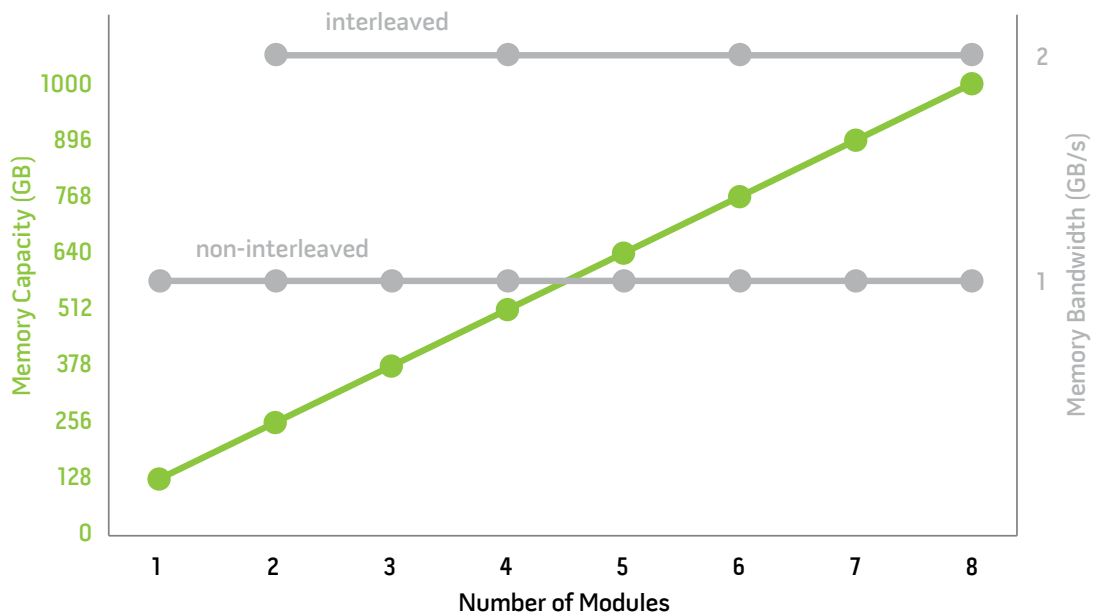


Figure 9 Memory Capacity and Bandwidth as a function of number of HLDIMM modules

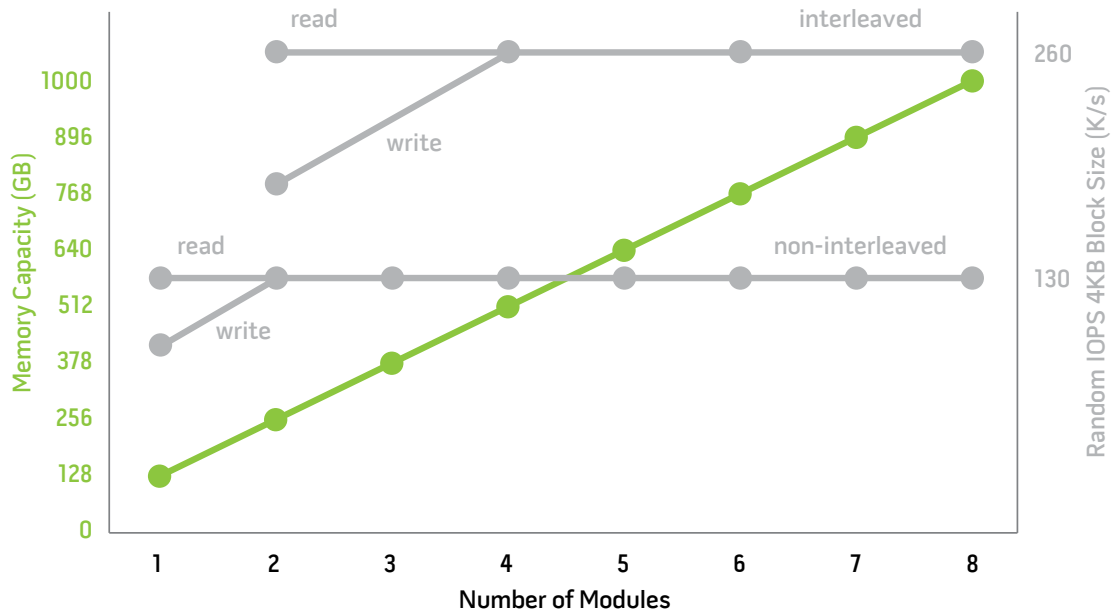


Figure 10 MLC Memory Capacity and IOPS as a function of number of HLDIMM modules

CONCLUSION

NAND Flash technology has the potential to realize true Storage Class Memory, the first significant development in computer memory hierarchy in decades. Rather than simply replacing the functionality and performance levels of conventional rotating magnetic hard disks, there is an opportunity to achieve orders of magnitude higher performance by connecting NAND Flash memory through the Northbridge via an interface such as PCIe, rather than traditional HDD Southbridge interfaces such as SATA. HLNAND enables NAND Flash technology for SCM applications through a high-performance point-to-point interface that is scalable to large memory configurations without bandwidth degradation. HLDIMM modules provide flexibility in supporting a wide range of end-user configurable memory subsystems.

NOTES

NOTES



MOSAID Technologies Incorporated
Corporate Headquarters
11 Hines Road
Kanata, ON
Canada K2K 2X1
www.MOSIAD.com

Information relating to products and services furnished herein by MOSAID Technologies Inc. or its subsidiaries is believed to be reliable.
The products, their specifications, services and other information appearing in this publication are subject to change by MOSAID without notice.

MOSAID, the MOSAID logo and HLNAND are trademarks of MOSAID Technologies Inc.

© 2009, MOSAID Technologies Inc. All Rights Reserved. Publication Number 9MT150