Mastering EDA Environments with High Performance Memory Technology

by

Chris Bell, Scott Clark & Ryan Radcliff

Deopli Corporation

Executive Summary

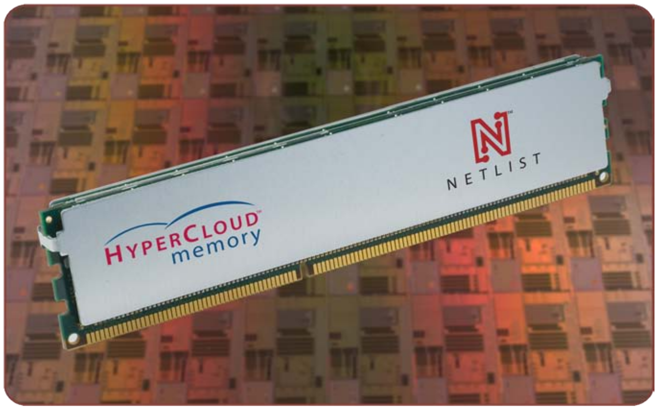
As semiconductor companies continue to evolve with the development of new technology, Moore's law (i.e., transistor count doubling every two years) continues to hold true. The trend is forcing the software and systems used to design these semiconductors to expand their use of memory in alignment with the increased gate count.

As an example, Electronic Design Automation (EDA) jobs that used to require 4GB of memory now require 8GB just to complete, while jobs that once needed 128GB of memory to run successfully now require 256GB and beyond. Compound never-ending growth in memory demand with steadfast developments in processor technology (as processor count scales with multi-core technology), and we can conclude that systems are woefully behind in memory capacity, as well as performance. This lag represents a growing risk for EDA environments, as costs stand to balloon to geometric proportion.

Constraints in advancing memory design are the result of many factors. We will explore how memory technology is currently designed, how its cost is derived, and the barriers that exist in designing larger, and more cost effective, capacity. Throughout this process, the inherent need for denser, and faster memory will become clear.

Moreover, this paper shows how HyperCloud™ memory from Netlist, Inc., can help overcome these constraints by adding 288GB DRAM running at 1333 MT/s. This can enable 15% improvement in an EDA job runtime.

To put it plainly, expensive EDA tools that are typically memory hungry will run faster with HyperCloud. A workload run with Hypercloud memory will create opportunities to either use fewer licenses (i.e., saving cost), or allow more workload to run in the same time.



EDA's Thirst for Better Memory

To date, memory technology development follows a pattern very similar to Intel's "tick-tock" evolutionary cycle: semiconductor companies perform what is called a "geometry shrink," fitting transistors and connectors closer to create an additional level of capacity. This results in the same technology fitting a smaller footprint on the same chip, but with increased programming potential. Historically, geometry shrinking has been accomplished in one-year cycles, while a second year is spent leveraging the newly available programming space (i.e., transistors) for additional features and capabilities. The cycle then repeats with a geometry shrink - the "tick-tock" describing the "shrink-add" part of the process.

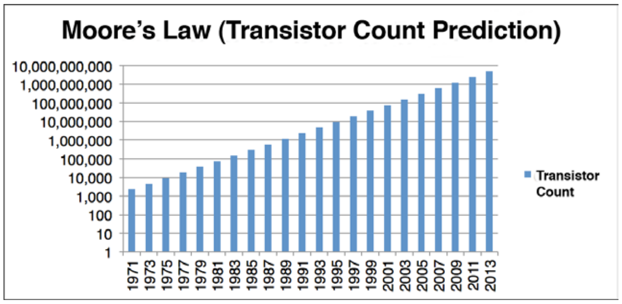


Figure 1. Moore's Law illustrated: Transistor counts double every 24 months.

Though Moore's original prediction of transistor count doubling every 18 months was long ago adjusted to every 24 months, Moore's Law is alive and very present in technology development today. The graph above reflects this - from 2300 transistors in 1971 to the anticipated 1.2 billion-transistor mark in 2009 (Opteron - Istanbul, Power7, Westmere). Looking forward, the prediction aims at 2.5 billion-transistor projects this year (2011) and ~5 billion transistors in 2013.

> *To support this trend, surrounding factors in EDA technology must adapt to Moore's Law and expand as well, and memory is no exception.*

When a geometry shrink is performed, there is a nonlinear correlation based on chip dimension. Shrinking from 130nm to 65nm seems like a doubling of available capacity, but in actuality, available capacity quadruples. This means that there is four times the work required, provided the process of designing the chip remains the same (in actuality, more work will be required). Furthermore, going from 130nm to 90nm would equate to twice the workload, and require twice the capacity, across the board (e.g., storage, memory, processing capability resources).

Next, observe the performance gap that continues to develop between processor and memory technology. In Figure 2, processor performance continues to grow at 60% per year, and while (there are limitations in DRAM capacity as a result of the 8-rank limitation on the CPUs memory channel; you need a 2R RDIMM to populate all 3 DIMM slots in a channel. Hard Drive capacities keep pace, memory performance only grows at 9% per year. (*Computer Architecture*, Henenessey and Patterson, third edition, p. 391)

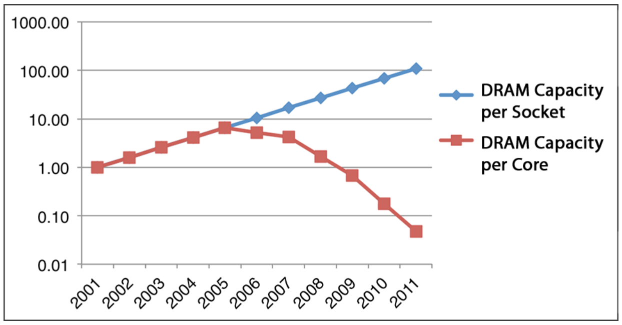


Figure 2. Memory capacity per socket and core illustrating the continuing performance gap between processor and memory technology.

> *This discrepancy indicates that we cannot afford to miss an opportunity to pull as much performance out of memory subsystems as possible.*

The bad news does not stop there: the following chart shows that core counts have practically gone viral. Thus, demand has feverishly grown for memory capacities, and their performance, to follow suit.

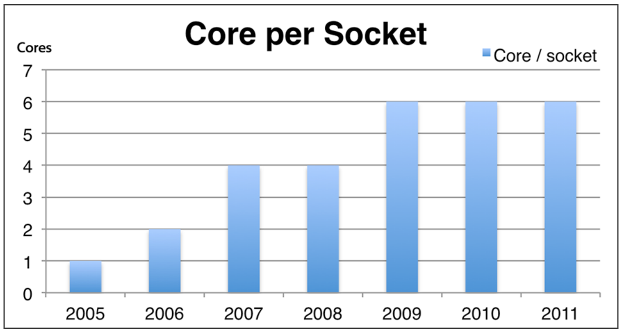


Figure 3. Increasing number of processor cores per socket on newer chipsets.

Not only must memory capacities keep pace, but they also need to multiply by the core count per socket as shown in Figure 3.

From a visual perspective, under-performing memory has obvious cost and performance implications. We'll explore constraints in memory design at a deeper level, focusing on DIMMs (i.e., dual in-line memory modules). More specifically we will focus on, Intel's Westmere-EP 64-bit architecture, because it is currently the best performing processor for overall EDA workloads. A server motherboard architecture typically consists of 3 memory channels per socket and 3 DIMMs per channel (DPC) for a maximum supported memory footprint.

Note: By specification from Intel, 1 & 2DPC run at 1333 MT/s and 3DPC runs at 800 MT/s.

The Driving Components of Memory Performance

Distance and Loading Restrictions - Due to space and wire-length restrictions on motherboards, the number of DIMM slots is limited to 3 for each channel.

Speed Limitation - A memory bandwidth limitation of 800MT/s is encountered with 3DPCs due to channel loading. When more DIMMs are populated per channel the digital signal begins to distort, due to the increased loading, causing the CPU to decrease the speed thereby minimizing the distortion.

Cost - To obtain larger memory configurations, it is logical to increase the density of the individual chips (e.g., one 16GB DIMM has twice the capacity of a 8GB DIMM) using the same socket. However, memory prices are not linear - doubling the density typically quadruples the price.

Registered or Unbuffered Types - Registered memory has an additional register to buffer the control signals, facilitating larger configurations. Unbuffered memory has slightly lower latency (i.e., less than 1%) when used with only one DIMM per channel, but cannot be combined more than two ranks deep. Using two DIMMs per channel actually reverses that penalty with interleaving, and the RDIMMS are faster.

Swap Space - If a process needs more memory than the system has available in populated RAM, the operating system's virtual memory manager will allocate more space using disks for swap space. For a high performance system, this is a very last resort, since the memory speed we have focused on is in the 6400 to 10600 MB/s range. A typical 7200-RPM hard disk can only sustain about 3 MB/s under optimum conditions. Seek time to find the location to read/ write in RAM is instantaneous, where a hard disk can be tens of milliseconds.

Rank Limitation - Each channel contains up to 8 'ranks' of memory. A rank is a group of chips that are addressed together as a set 64 bits wide, with an additional 8 bits for ECC error correction. The CPU can only address 8-ranks per channel which limits a channel to being populated with two 4-rank DIMMs. Today's industry has switched to the term 'rank' instead of 'bank' when referring to DIMM modules.

A Closer Look at DIMM Density

Among constraints in memory design, DIMM density presents the largest hurdle. First, memory manufacturers stack multiple groups of DRAMs on one DIMM to increase density. Two groups are referred to as 'dual-rank,' while four is 'quadrank.' Each rank has a separate pin for the 'Chip Select' function to access that rank. Therefore, the sockets must be uniquely connected. The first socket (i.e., farthest from the CPU, and populated first) occupies chip selects 0-3; the second socket 4-7, and the third socket reuses 2 and 3. Thus, two quad-rank DIMMs use all 8 chip selects, leaving the third socket completely unused.

The individual chips that go onto a DIMM are available in various densities and organizations. The density is defined as the quantity of memory bits, while the organization is how the arrays of bits are arranged (e.g., a 1G chip could be arranged as 256Mx4,128Mx8 or 64Mx16). The DIMM has a databus width of 64 bits (plus another 8 for ECC if used). Therefore, a rank of memory needs sixteen x4 chips (plus two for ECC), eight x8 chips (plus 1 for ECC), or four x16 (plus a x8 or two x4 for ECC). Since the x4 example is using 16 chips, they can be half as dense as the x8 case to reach the same total capacity for the DIMM. However, they consume twice the physical space.

In addition, densities are not linear in cost, so a 1Gb chip is more than twice the cost of a 512Mb chip. See Figure 4 for three different ways to arrive at the 1GB total capacity of a DIMM. Furthermore, adding more chips adds to heat produced, and increases the cost of manufacturing. The standard LP form-factor DIMM has room for 18 chips per side, with 36 as the usual maximum per DIMM.

Number of Chips	Chip Organization	Chip Density	Ranks
36	64Mx4	256M	2
18	64Mx8	512	2
18	128Mx4	512M	1

Figure 4. Three 1GB DIMM configurations, showing different ways to produce 1GB total capacity.

One trick used to gain space for more (i.e., less dense) chips is to stack them in a second layer on the DIMM (in addition to both sides), using special chips with a slightly different control pin. Thus, the chips share the same pins. This method adds some cost to the chip, and a lot of extra cost in manufacturing, but allows up to 72 chips on one DIMM.

Traditional ECC can correct a single-bit error in an 8-byte word, and detect a two-bit error. If a whole chip went bad, that would be 4 (or 8 or 16) bits at a time, so the ECC algorithm could possibly miss completely. However, IBM introduced a technology they called "Chipkill" (also called SDDC [Single Device Data Correction] by Intel) that uses the distributed ideas from disk RAID and makes the ECC data/checksum interleaved over 4 words. Now with a x4 chip, only one bit will be used for each calculation, enabling full correction for even an entire chip going bad. However, with x8 or x16 chips, some whole-chip detection ability is lost. The different chip organizations should be comparable in cost, but the x8 are about 10% cheaper due to very high volumes in the consumer PC market.

HyperCloud Solves Memory Bottleneck

Netlist has developed a new memory technology, known as HyperCloud™ memory, that addresses the issues associated with memory bottlenecks. HyperCloud™ memory utilizes an ASIC chipset that incorporates Netlist's patented rank multiplication and load reduction technology shown in Figure 5.

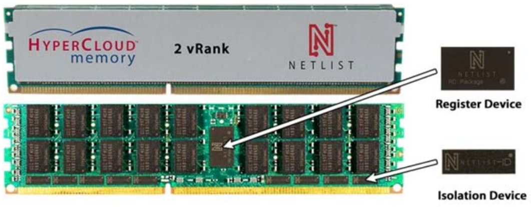


Figure 5. HyperCloud™ Memory provides load reduction and rank multiplication for high speed, high density memory.

The register device contains the rank multiplication functionality and the isolation devices perform the load reduction between the DRAM and the CPU.

Rank Multiplication - Rank multiplication increases memory capacity in servers. The rank multiplication functionality enables 4-physical ranks to be presented as 2 virtual ranks (vRanks) to the CPU. Three, 2 vRank (4 physical rank) DIMMs can be populated per channel with rank multiplication thus enabling population of the third slot on the memory channels.

Load Reduction - Load reduction increases memory bandwidth in servers. The load reduction functionality "cleans" up the distortions in the digital signal due to increased channel loading thereby allowing the CPU to maintain the 1333 MT/s speed with increased loading. Three DIMMs can be populated in a channel while maintaining the 1333 MT/s on each channel. Now 288GB can loaded in a 2P server running 1333 MT/s.

HyperCloud Benchmarks

Netlist's HyperCloud DIMMs were tested as a possible solution for systems responsible for chip design. The memory modules were tested on a Cirrascale VB1325 Server with 288GB HyperCloud RAM running at 800 MTs/ and 1333 MT/s. A tool was selected that exercises a great deal of memory, but with relatively little CPU load (and no disk I/O), to test how memory speed affects compute performance. Figure 6 shows the results. The first choice was the test number, which is related to the mathematical operations used. We chose a simple set that excludes the fancier MMX or SSE instructions, and simply adds and multiplies. The graph displays the block size (in kilobytes) on the X-axis, and the total memory speed (in Megabytes/second) on the Y-axis. The three plots are the number of concurrent processes (2, 4 and 8).

Note: Ramspeed/SMP was written specifically to test memory performance. It contains options to modify the running parameters, so we had to choose an appropriate configuration.

The affect of caches can be observed by varying the block size of the memory tests (as shown on the X-axis, in Kb/block). This particular CPU has three levels of on-chip cache, which can be seen with drops near 64Kb, 512Kb, and 4Mb. In addition, the interaction between multiple cores can be seen, as the relative speeds converge when accessing main memory.

Memory Test Conditions

System:

Cirrascale VB1325
Dual Xeon X5660 CPUs running at 2.80 GHz
Cache: L1 = 384 Kbytes, L2 = 1536 Kbytes,
L3 = 12288 Kbytes

Memory:

HyperCloud-8: Netlist 8Gb 2Rx8 PC3-10600R-9-10-22
HyperCloud-16: Netlist 16Gb 2Rx4 PC3-10600R-9-10-22

Tools:

Ramspeed/SMP version 3.5.0
CentOS 5.5
Linux Distribution

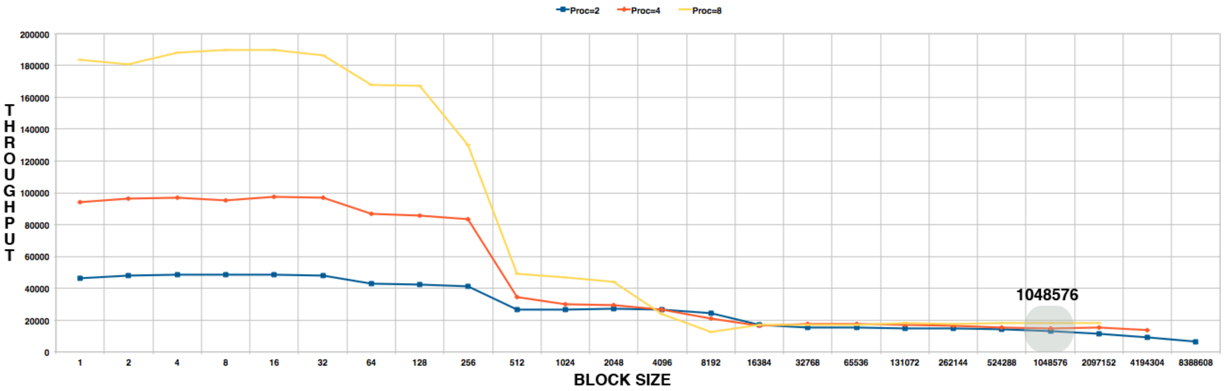


Figure 6. HyperCloud DIMM Test Results. The 1048576 block size is used as the basis for additional test results.

The L1 (32K data) and L2 (256K) cache is dedicated to each core, so as the number of processes increases, the total memory speed increases linearly (on the left-hand side of the graph). The L3 cache (12MB) is shared over the 6 cores in a CPU chip, and the third plateau does not have as much benefit from the 8 cores, since they are all sharing the cache. Above the 16MB block size (16384 on the graph), the effects of cache are overwhelmed, and all three lines settle to a fairly consistent level. This represents the speed of the memory system. Therefore, because high memory use is the intent, a point larger than 16MB is chosen. This ensures no cache was involved, while large memory utilization is represented by selecting 1024K (which is 1048576 on the graph) for the block size. In addition, the highest concurrency (i.e., 8 processes) is chosen to utilize both CPU chips.

Connecting HyperCloud to EDA

To put it plainly: expensive, typically memory hungry EDA tools will run faster as a result of HyperCloud. Since most, if not all, of the major EDA vendors exercise statements in their contracts that prevent publishing benchmark runs of their tools, we opted to use a theoretical memory benchmark tool. As a companion, we note the tools that have memory intensive characteristics, as well as the part of the design flow in which their memory heavy runs appear. Mileage may vary, depending on the specifics of the design and quantity of memory used. Semiconductor design companies should be able to measure the memory I/O of their tool runs. Measure the process listing (i.e., output of the ps command) to determine if the RSS entry for the tool run process is a significant percentage of the total memory size. If so, fit the memory intensive profile to benefit from faster memory at that capacity. A workload running with Hypercloud will create opportunities to either use fewer licenses (i.e., saving cost) or allow more workload to run (e.g., tests that were previously exempted due to lack of time and/ or licenses). This additional workload then delivers a higher yielding part, due to the increased chances of catching mistakes prior to tapeout.

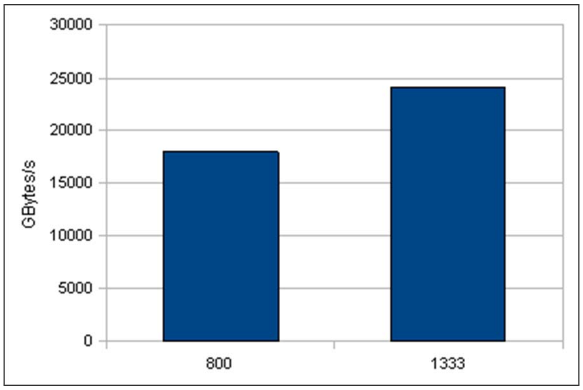


Figure 7. 67% benefit with HyperCloud memory at 1333MT/s versus 800MT/s

HyperCloud enables a system to run at 1333MT/s, which is 67% greater than 800MT/s. Figure 7 shows the results of the Ramspeed benchmark using HyperCloud memory both at the default 800MT/s and the full speed of 1333MT/s with 288GB capacity.

Note: A typical server does not utilize the full 67% advantage but, still runs 34% faster.

The specific facets of EDA design flow that would benefit from larger footprint, faster memory access speeds are defined in Figure 8. The first step toward implementation is to define the overall design flow and the time allocations for each phase:

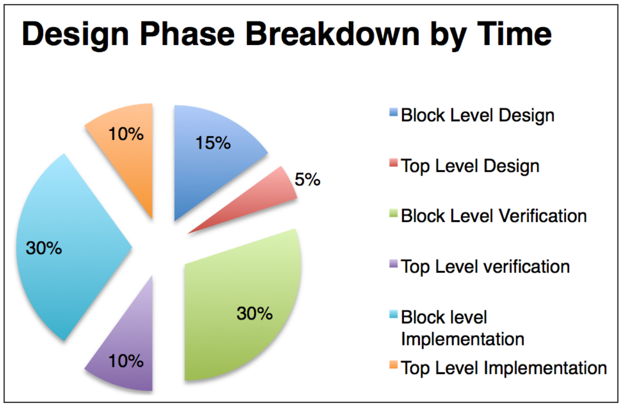


Figure 8. EDA Design Flow Breakdown, with block-level design, verification, and implementation (75% total) being memory hungry phases.

As illustrated, select parts of block-level verification and implementation are memory hungry, as are most facets of Top-Level design, verification, and implementation.

Specific tools that must be considered, from this perspective, are those used in expensive, top-level verification and implementation steps:

- Apache Design Automation Redhawk
- Cadence Affirma
- Cadence Celtic
- Cadence SOC_Encounter
- Cadence UltraSIM
- Mentor Calibre
- Mentor TestKompress
- Synopsys Formality
- Synopsys HSIM
- Synopsys ICC
- Synopsys Primetime
- Synopsys Star_RCXT

These tools have cost vectors in the hundreds of thousands of dollars per seat range, and with ~20% of the design, possess memory intensive work profiles. Given large, required memory footprint, “filling all banks” is usually implied. This traditionally means running at 800 MT/s speeds with modern processors (e.g., Nehalem, Westmere). However, if we speed up to 1333 MT, that would yield a 60% mathematical increase. This is measured to be 34% in a real memory test, and since we have selected only memory intensive runs to speed up, we can assume that they spend >50% of their time accessing data in memory. Therefore, we conservatively estimate improvement to be 15% over the entire run - based on the improvement we have created in the memory architecture.

In closing, this increase is a conservative estimate, and reality may show even greater benefit. With even this estimate, we can look at implications for cost avoidance (i.e. avoiding additional spending on more licenses as workloads continue to increase), that can significantly reduce software costs in an EDA environment.

Summary

By supporting multi-core processors with large capacity high speed memory, HyperCloud gives EDA professionals flexibility to run hardware and EDA software with increased server utilization thereby reducing design time and increasing design productivity. In summary this paper has shown HyperCloud memory can enable:

- 15% improvement in an EDA job runtime
- 288GB RAM in a 2P server running at 1333 MT/s
- Increased design productivity
- Reduced EDA software expenditures.

References

Ramspeed/SMP [http://alasir.com/software/ramspeed/] •

- Intel's Tick-Tock Model [http://www.intel.com/technology/tick-tock/ index.htm] •
- Intel Xeon Processor 5500 Series Datasheet, Volume 2 (April 2009) [http://www.intel.com/Assets/en_US/PDF/datasheet/321322. pdf] \
- Intel Xeon Processor 5600 Series Datasheet, Volume 2 (March 2010) [http://www.intel.com/Assets/en_US/PDF/datasheet/323370. pdf]

The benchmarks were conducted on a Cirrascale VB1325 Server with 288GB HyperCloud RAM running 1333 MT/s.

Cirrascale is a premier developer of build-to-order, independent blade-based high performance computing and storage data center infrastructures.

Company:	Cirrascale
Located:	12140 Community Road Poway, CA 92064
Phone:	+1 888.942.3800
Web:	www.cirrascale.com

About Deopli

Deopli is one of the foremost thought leaders in the EDA infrastructure and cloud computing space. Composed of highly- trained personnel, equipped with technology and experience, operating under principles of self-sufficiency, technical competence, speed, efficiency and close teamwork. Providing advisory and consulting services to EDA companies with respect to their HPC environments, they also conduct specialized operations including reconnaissance, strategy definition, tactical definition and resource training. In addition, Deopli executes non-operational, high-risk tasks to achieve significant strategic objectives. Deopli is headquartered in Irvine, California. For more information, visit **www.deopli.com**.

About Netlist

Founded in 2000 and headquartered in Irvine, California, Netlist is the leading provider of high-performance modular memory subsystems to the world's premier OEMs. Netlist specializes in bridging the widening gap between the system OEM's requirements and the capabilities of the IC manufacturer. Our patented memory subsystem technologies overcome density, performance, and cost limitations, effectively blending commodity components with their inherent deficiencies into highly reliable, optimized memory solutions. Netlist pioneered ideas such as embedding passives into printed circuit boards to free up board real estate, doubling densities via 4-rank double data rate (DDR) technology, and other off-chip technology advances that result in improved performance and lower costs compared to conventional memory. For more information, visit **www.netlist.com**.