Hybrid Memory Cube
New DRAM Architecture Increases Density and Performance

Joe Jeddeloh, Brent Keeth
Minneapolis, MN; Boise, ID, USA
jjeddeloh@micron.com, bkeeth@micron.com

Abstract

Multi-core processor performance is limited by memory system bandwidth. The Hybrid Memory Cube is a three-dimensional DRAM architecture that improves latency, bandwidth, power and density. Through-silicon vias (TSVs), 3D packaging and advanced CMOS performance enable a new approach to memory system architecture. Heterogeneous die are stacked with significantly more connections, thereby reducing the distance signals travel.

Introduction

Multi-core processors are limited by "the memory wall." High-end servers idle 3 out of 4 clocks. High-performance computing performance is often less than 10% of peak. Processor cores and threads are increasing geometrically, resulting in greater demand for memory density. Evolutionary standards do not address the memory wall. System power limits constrain performance. The Hybrid Memory Cube (HMC) is a solution for memory bandwidth, power and scalability. See Fig 1. HMC System Diagram.

Hybrid Memory Cube

The HMC is a stack of heterogeneous die. A standard DRAM building block can be combined with various versions of application-specific logic. Each 1 Gb DRAM layer is optimized for concurrency and high bandwidth. As shown in Fig. 1, the HMC device uses through-silicon via (TSV) technology and fine pitch copper pillar interconnect. Common DRAM logic is off- loaded to a high-performance logic die. DRAM transistors have traditionally been designed for low cost and low leakage. The logic die, with high-performance transistors, is responsible for DRAM sequencing, refresh, data routing, error correction and high-speed interconnect to the host. TSVs enable thousands of connections in the Z direction. This greatly reduces the distance data needs to travel, resulting in improved power. The stacking of many dense DRAM devices produces a very high-density footprint. The HMC was constructed with 1866 TSVs on ~60µ pitch (Fig. 2). Energy per bit was measured at 3.7 pj/bit for the DRAM layers and 6.78 pj/bit for the logic layer, netting 10.48 pj/bit total for the HMC prototype. This compares to 65pj/bit for existing DDR 3 modules.

DRAM

The HMC DRAM is a 68mm^2, 50nm, 1Gb die segmented into multiple autonomous partitions. Each partition includes two independent memory banks for a total of 128 banks per HMC. Each partition supports a closed page policy and full cache line transfers, 32 to 256 bytes. Memory vaults are vertical stacks of DRAM partitions (Fig. 2). This structure allows more than 16 concurrent operations per stack. ECC is included with the data to support reliability, availability and serviceability (RAS) features. The efficient page size and short distance traveled results in very low energy per bit. The TSV signaling rate was targeted at less than 2 Gb/s to ensure yield. This resulted in more robust timing while reducing the need for high-speed transistors in our DRAM process. Each partition consists of 32 data TSV connections and additional command/address/ECC connections. The tiled architecture and large number of banks results in lower system latency under heavy load. The DRAM is optimized for random transactions typical of multi-core processors. The DRAM layers are designed to support short cycle times, utilize short data paths and have highly localized addressing/control paths. Each partition, including the TSV pattern, is optimized to keep the on-die wiring and energy-consuming capacitance to a minimum.

As depicted in Fig. 3, the DRAM data path extends spatially only from the data TSVs to the helper flip flop (HFF) and write driver blocks, which are tucked up against the memory arrays. The address/command TSVs are located along the right side of Fig. 3. The address/command circuits reside in very close proximity to the partition memory arrays. The memory arrays are designed with 256-byte pages.

Logic Die

Many of the traditional functions found in a DRAM have been moved to the logic layer, including the high-speed host interface, data distribution, address/control, refresh control and array repair functions. The DRAM is a slave to the logic layer timing control. The logic layer contains adaptive timing and calibration capabilities that are hidden from the host. This allows the DRAM layers to be optimized for a given logic layer. Timing, refresh and thermal management are adaptive and local.

The logic layer contains a DRAM sequencer per vault. Local, distributed control minimizes complexity. The local sequencers are connected through a cross bar switch with the host interfaces. Any host link interface can connect to any local DRAM sequencer. (Fig. 3) The switch-based interconnect also enables a mesh of memory cubes for expandability.

Host Interface

The logic layer is implemented with high-performance logic transistors. This makes the high-speed SERDES (15 Gb/s) possible. HMC uses a simple abstracted protocol versus a traditional DRAM. The host sends read and write commands versus the traditional RAS and CAS. This effectively hides the natural silicon variations and bank conflicts within the cube

and away from the host. The DRAM can be optimized on a per-vault basis. Out-of-order execution resolves resource conflicts within the cube. This results in nondeterministic and simpler control for the host, similar to most modern system interfaces. The abstracted protocol enables advanced DRAM management.

RAS

There are extensive RAS capabilities within the HMC. Onboard ECC allows for local error detection and correction. Soft errors and variable retention time errors (TRV) can be dynamically repaired while the system is operating. Redundant interconnect resources are also available, such that failing TSV connections can be mapped out. Redundant links allow multiple links to fail while still allowing access to the entire array. Self test and repair is built into the HMC. Power- on diagnostic capabilities enable the HMC to determine failures and make in-field repairs.

Design Process

The design of the HMC involved many engineering teams within Micron: DRAM design, logic layer design, packaging, assembly, process R&D and test. Each function fed data into an extensive simulation and modeling environment. Architecture, performance analysis, stress analysis, TSV development, thermal analysis, power delivery, signal integrity continually helped to refine models that verified assumptions or demonstrated new directions.

Our goal for the power delivery network (PDN) for this device was to keep peak supply noise below 50mV for the power and ground networks, respectively. This is extremely critical given the use of 1.2V supplies for the DRAM and active power consumption approaching 5 watts. Using a combination of commercially available and in-house PDN analysis tools, we developed a set of power delivery models for the DRAM layers, the TSV interconnections, the logic layer and the package. These models were merged with a DRAM circuit-level netlist, within an analog test bench, to facilitate concurrent simulation of both circuits and the power delivery network. Supply noise initially pushing beyond 100mV during a read operation. We tried significantly widening the on-die power busing, but it produced negligible improvement. Ultimately, we had to add more power/ground TSVs to the partition, especially in close proximity to the HFF block, reallocate circuits to different supply domains, increase on-die decoupling, and modify circuit timings. The result of these changes was that peak supply noise was reduced to an acceptable 30mV.

Future

Now that the baseline technology has been verified, additional variations of HMC will follow. DRAM process geometry and cell size will continue to shrink to half the size it is today. This and improved stacking will allow greater density for a given cube bandwidth and area. HMC devices will extend beyond 8GB. The number of TSV connections will double to create a cube capable of 320 GB/s and beyond.

New interconnects will be optimized for a given system topology. Short-reach SERDES are being developed that are capable of less than 1 pj/bit. Medium-reach SERDES will serve 8–10 inches of FR4. Silicon photonics will extend the reach to 10 meters and beyond.

Atomic memory operation, scatter/gather, floating point processors, cache coherency and meta data are natural candidates for inclusion in the HMC. Data manipulation closest to where the data resides is the highest-bandwidth, lowest-power solution.

Illustrations

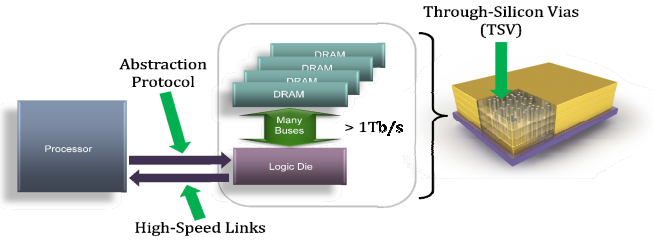

Fig. 1. HMC System Diagram

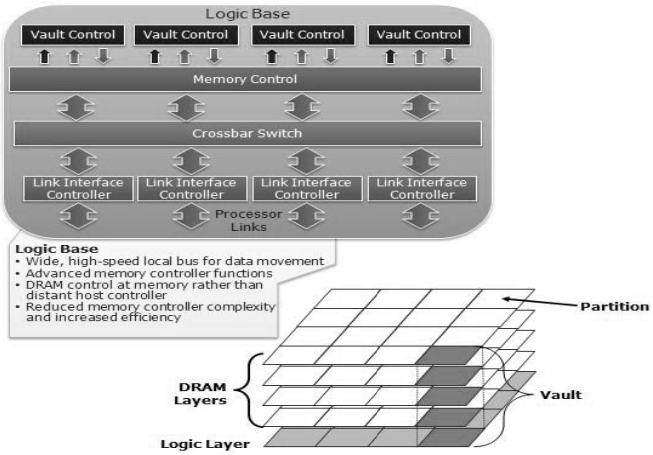

Fig. 2. HMC Block Diagram

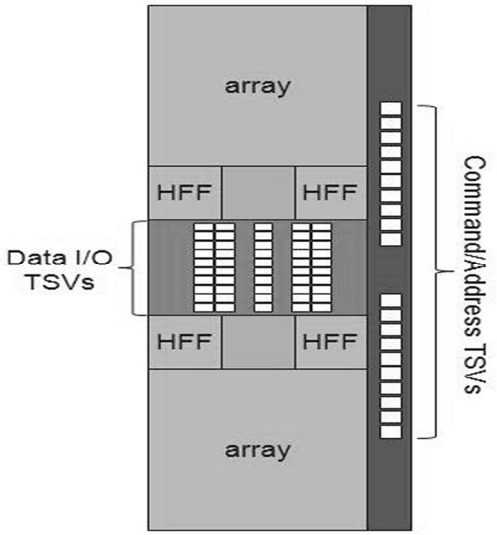

Fig. 3. HMC DRAM Partition Floor Plan