

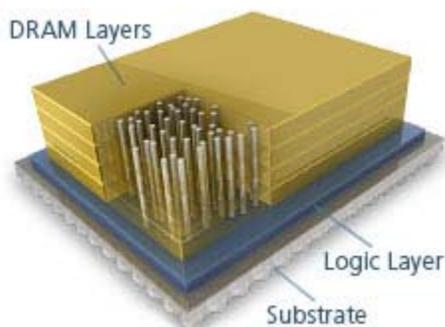
HPCwire

Micron Readies Hybrid Memory Cube for Debut

January 17, 2013 | Tiffany Trader

The next-generation memory-maker Micron Technology was one of the many innovative companies demonstrating its wares on the Supercomputing Conference (SC12) show floor last November. Micron's General Manager of Hybrid Technology Scott Graham was on hand to discuss the latest developments in their Hybrid Memory Cube (HMC) technology, a multi-chip module (MCM) that aims to address one of the biggest challenges in high performance computing: scaling the memory wall.

Memory architectures haven't kept pace with the bandwidth requirements of multicore processors. As microprocessor speeds out-accelerated DRAM memory speeds, a bottleneck developed that is referred to as the memory wall. Stacked memory applications, however, enable higher memory bandwidth.



The Hybrid Memory Cube (HMC) is a new memory architecture that combines a high-speed logic layer with a stack of through-silicon-via (TSV) bonded memory die that enables impressive advantages over current technology. According to company figures, a single HMC offers a 15x performance increase and uses 70 percent less energy per bit when compared to DDR3 memory, and takes up 90 percent less space than today's RDIMMs. The Cube is also scalable per application, which is not possible with DDR3 and DDR4. System designers have the option of employing the HMC as near memory for best performance or in a scalable module form factor, as far memory, for optimum power efficiency.



Micron HMC demo at SC12 - screen shot

This is a huge leap forward from a technology perspective, noted Graham, compared to DDRx and other boutique memory products that are out there.

As *HPCwire* editor Michael Feldman explained in an earlier review of the technology:

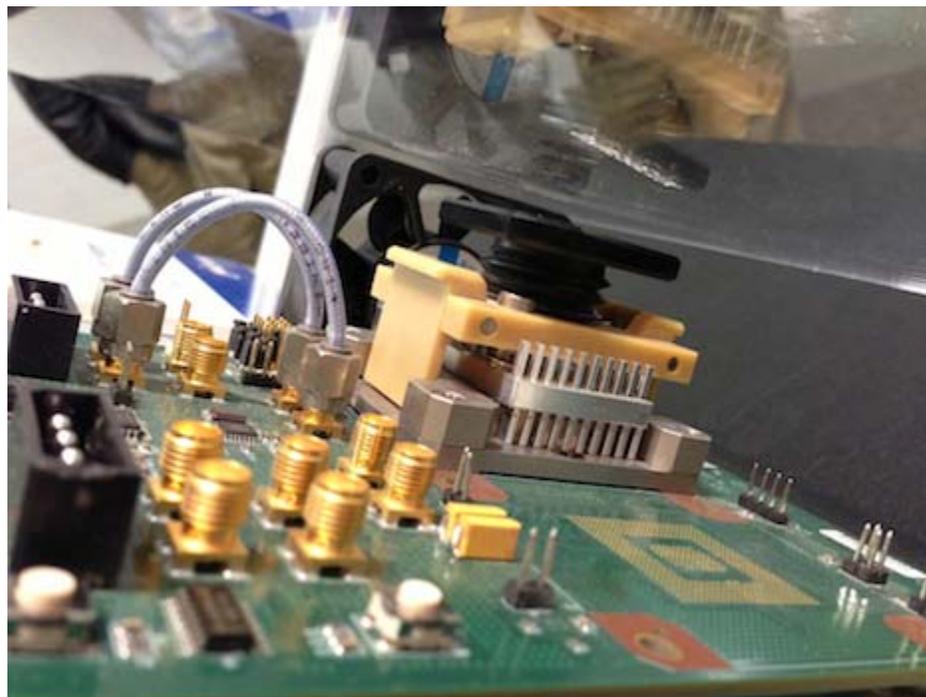
The speedup and better energy efficiency is achieved principally through parallelism. Because the memory chips are stacked, there is more space for I/O pins through the TSVs. Thus each DRAM can be accessed with more (and/or wider) channels. The end result is that the controller can access many more banks of memory concurrently than can be accomplished with a two-dimensional DIMM. And because the controller and DRAM chips are in close proximity, latencies can be extremely low.

Judging by the degree and caliber of community involvement, Micron's HMC technology represents a real breakthrough in how memory is used. In October 2011, Micron together with Samsung Electronics Co., Ltd., formed the Hybrid Memory Cube Consortium, tasked with developing an open industry standard that facilitates HMC integration into a wide variety of systems, platforms and applications.

The consortium is managed by a group of ten developers (Altera, ARM, HP, IBM, Micron, Microsoft, Open Silicon, Samsung, SK Hynix, and Xilinx), which have equal voting power on the final specification, along with an additional 75 adopters. The members are currently reviewing a draft specification, scheduled to be released next month, that details the communication interface between the Cube and the processor – CPU or GPU or FPGA.

Speaking to the initial set of targeted applications, the driving body notes that the "Hybrid Memory Cube represents the key to extending network system performance to push through the challenges of new 100G and 400G infrastructure growth. Eventually, HMC will drive exascale CPU system performance growth for next generation HPC systems."

Companies are eager to get their hands on this product and Micron is working with an aggressive roadmap to meet that demand, Graham told *HPCwire*. The Gen1 demo, on display at SC12, was real silicon, and engineering samples for the Gen2 device are due out this summer. If all goes as planned, the Hybrid Memory Cube will be in full production at the end of this year or early 2014. In fact, contracts are already in place for the 2014 timeframe.



Side shot of the HMC demo at SC12. The actual Cube is on the far corner of the board.

High-speed networking vendors have signed up for the first productized version – with an HPC-centric product not far behind. Micron was not at liberty to identify these initial customers, but a look at the list of consortium partners turns up candidates like Lawrence Livermore, Cray, NEC, and T-Platforms.

The first couple of HMC implementations will be straight DRAM, but Micron and others are researching alternative memory combinations, for example multi-memory stacks that employ NAND flash and DRAM.

"There are all kinds of things you can do within the logic layer to pull different types of functionality, that are maybe off-chip today, into the logic layer and innovate further, with more functionality, better performance, and lower energy," noted Graham.

Micron is framing this as an aggressive technology, emphasizing that this is the first time that all three of the largest memory makers (Micron, Samsung, and presumably SK Hynix) have teamed up.

"All of the memory manufacturers face the same challenge of being able to scale beyond 20nm, so we're all coming up to even a bigger memory wall eventually," explained Graham "Beyond 20nm, you're going to have to move to some kind of management layer in order to continue with DRAM technology."

The logic layer developed for the Cube allows for different flavors of technology – like spin-torque, memristor and others – to extend the viability of DRAM memory. "This means we can be really innovative with different types of cell technology and process technology so we can bring a more standard memory into the marketplace without these major shifts," said Graham.

According to the Micron rep, they've also seen interest from companies outside the consortium. Most notably absent from the member roster are AMD and Intel, but that by no means implies a lack of involvement. Intel, for its part, demonstrated a prototype HMC device during the fall Intel Developer Forum in September 2011, deeming it the fastest and most efficient DRAM ever built. As Graham put it, these companies have opted not to be involved in the open standard in order to develop their own way of using the technology.

Still a year away from production, pricing for Cube products has not been announced, but early adopters should expect to pay a premium for the benefits of increased performance, power efficiency and space savings. The exascale community, in particular, will be paying close attention. If they're to realize their goal of a 10^{18} FLOPS machine within a 20MW power envelope, they'll need all the help they can get.

Copyright © 1994-2011 **Tabor Communications**, Inc. All Rights Reserved.

HPCwire is a registered trademark of Tabor Communications, Inc. Use of this site is governed by our Terms of Use and Privacy Policy.

Reproduction in whole or in part in any form or medium without express written permission of Tabor Communications Inc. is prohibited.

Powered by **Xtenit**