HPC wire

Stacking Stairs Against the Memory Wall

April 02, 2013 | Nicole Hemsoth

Last year at Intel's Developer Forum, the chipmaker unfurled a plan it cooked up with Micron to release "real" stacked memory into the world. The concept was too appealing to ignore—a potential solvent to dissolve the parallel DRAM performance and efficiency wall, not to mention backed by well-known vendors.

The need in the market isn't easy to ignore either. With ever-mounting CPU advancements that promise superior performance, the blame for lousy delivery on those chip promises lies squarely on memory. Finding a way to refine how quickly, not to mention efficiently, memory can feed dashing CPUs is a priority--and its one that several vendors have been eyeing.

While Micron and Intel weren't the first or only to propose clicking together 3D memory, the news cycle around the technology did heat up. Since then, several vendors have stacked their chips on the betting table, including NVIDIA with its recent Volta announcement.

Not long ago we sat down with former Cray scientist, Steve Scott, who is now CTO for the Tesla division at NVIDIA. During our chat, which went beyond GPU computing, we talked about the memory wall and how approaches like stacked DRAM are still "not ready for primetime." He noted that the challenges of stacking memory are limited by some pure physics barriers—at least if users are going to see bandwidth in the arena of one terabyte per second, as it claims will be possible with its stacked solution by around 2015 with the rollout of their own spin on the stack, Volta.

Of course, not everyone agrees with the assertion that the approach isn't ready for primetime, particularly those who have been pushing 3D memory concepts—and getting some sizable industry backing in the process. Count among that crowd veteran semiconductor giant, Micron, which is staking a portion of its future on being the foundational force behind ultra-dense stacked memory with its Hybrid Memory Cube.

Just as with other tech that promises to revolutionize performance and efficiency, stacked DRAM could put a big dent in the network and high performance computing side of the memory wall says Mike Black, who serves chief technology strategist for Micron's Hybrid Memory Cube (HMC). Black helped spearhead an effort announced today that is set to standardize on this technology and make Micron, Samsung and its partners the rally point for the big takeoff they predict for the technology.

Sloughing off the assertion about the HMC not being ready for real production, the Hybrid Memory Consortium is out to bolster stacked memory via an open standard that a vendor community can build on. Supported by founding partner Samsung and a bevy of developers and adopters including Altera, IBM, ARM, Xilinx and others, the effort extended today to set a fresh global standard to bring the dense-memory approach into motion.

The team behind the standard has been working for close to two years to deliver on the final specifications for HMC interface standards. As Jim Elliot, VP of Memory Planning at Samsung said today, the ideal outcome is that "system designers and manufacturers will be able to get new green memory solutions that outperform other memory options offered today."

As the group noted:

The achieved specification provides an advanced, short-reach (SR) and ultra short-reach (USR) interconnection across physical layers (PHYs) for applications requiring tightly coupled or close-proximity memory support for FPGAs, ASICs and ASSPs, such as high-performance networking, and test and measurement. The next goal for the consortium is to further advance standards designed to increase data rate speeds for SR from 10, 12.5 and 15Gb up to 28Gb. Speeds for USR interconnections will be driven from 10 up to 15Gb. The next level of specification is projected to gain consortium agreement by the first quarter of 2014.

For Micron and the partners they've rallied for the Hybrid Memory Cube Consortium (an effort they launched with Samsung), stacked memory is a current reality--and one that presents a viable production opportunity, not to mention a fifteen-fold boost per cube over standard approaches and stated 70% energy usage reduction per bit than average DDR3 DRAM.

And besides, notes Micron's Black, there wasn't a lot of value for a company like Micron to be floating around in niche offering territory. To lead in this still immature space they need the backing of a larger community of users, developers and vendors.

There is little doubt that there's a wall to be scaled on performance and efficiency fronts when it comes to traditional approaches to DRAM, claims Black. The Hybrid Memory Cube technology that his company has led seeks to scale the wall via a memory architecture that matches a high-speed logic layer with a stack of through-silicon via (TSV) bonded memory die.

Micron points to the reality of this technology, noting that the stacked architecture has been shown to do everything but wash the dishes by kicking out the slower parallel interfaces of current DDR3s to create shorter, more efficient pathways and using 90% less energy than current RDIMMs via the more dense memory approach.

In an effort to get a production-ready offering, Micron launched a demo platform a couple of years ago to vet out some of the challenges that come with trying to cobble together a heterogeneous package of DRAM and logic devices on the same slice.

Black said that when it came to addressing challenges during that test phase, the biggest hurdle was in the TSV technology itself—more specifically, making it viable and ready for mass production. "There was a lot of physics and learning as we developed from that point," he noted. "We had to ask what was missing, how small in diameter is it possible to make a TSV, what's the best conductive material, how to negotiate the radio between the TSV diameter and the aspect ratio of the thickness of the silicon wafer—more learning had to happen as we kept stacking."

Micron sees a day on the horizon when they'll be able to stack beyond the four and eight-high DRAM units that are a current possibility. When they release this fall, they'll be able to offer 160 gigabits of bandwidth—so basically a terabit of bandwidth per cube. Of course, once users start blocking multiple cubes together it's possible to further increase total bandwidth. By the 2016 point, Black says they'll likely double the bandwidth they're able to offer now.

Micron plans to start sampling their stacked DRAM device this fall and will ramp it into production in the first half of next year.

Related Articles

Volta Adds Charge to GPU Roadmap

MetaRAM Launches New Memory Technology

Penguin Pushes Envelope on Compute Density
