Is 3D-stacked memory the solution for HPC?

BSC leads third chapter of the memory wall trilogy

Barcelona, 5th October 2015 – Emerging 3D-stacking technology enables DRAM devices that support much higher bandwidths than traditional DIMMs. The first commercial products Hybrid Memory Cube and High Bandwidth Memory will soon hit the market, and some of the publicity surrounding these emerging memory devices suggests that they will bring significant performance improvements. Barcelona Supercomputing Center (BSC) researchers have analysed how 3D-stacked DRAMs will affect performance of high-performance computing (HPC) applications, and they concluded that simple replacement of conventional DIMMs with 3D-stacked devices may not lead to announced performance improvements. In order to properly exploit the benefits of the novel high-bandwidth memory solutions, BSC computer architects suggest rethinking about the design of the overall computer systems, mainly processors and memory controllers.

This analysis was presented during the MEMSYS 2015 conference as a result of the paper titled *Another Trip to the Wall: How Much Will Stacked DRAM Benefit HPC?*, written in collaboration with experts from Chalmers University and Lawrence Livermore National Laboratory. This paper is a third chapter of the *memory wall* trilogy following Hitting the Memory Wall: Implications of the Obvious (1995) and Reflections on the memory wall (2004). *Memory wall* refers to the fact that memory latency is so large that most of the time processors are waiting for data from memory. "Technological evolutions and revolutions notwithstanding, the memory wall has imposed a fundamental limitation to system performance for over 20 years", says Petar Radojkovic, Memory systems team lead at BSC.

In the paper, BSC experts have recalled that the memory wall has always been defined in terms of main memory latency, not bandwidth. Higher bandwidth may lower memory latency, provided that the selected applications offer sufficient memory-level parallelism (MLP) and that processors can exploit it. But higher bandwidth cannot guarantee better performance because 3D-stacked DRAMs will not reduce idle-system memory latency. Therefore they will not improve the performance of applications with limited MLP. How well the available bandwidth we can be exploited, ultimately depends on the inherent MLP in our targeted workloads.

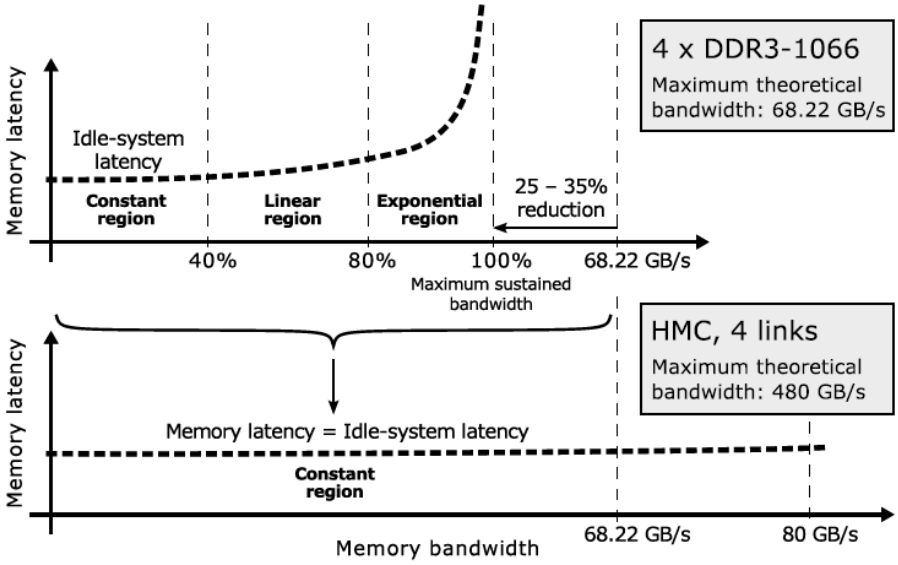

Figure 1: Memory latency depends on the used memory bandwidth, and the memory latency-bandwidth curve has three regions --- constant, linear and exponential. Moving from conventional DDRx (upper figure) to high-bandwidth memory solutions (lower figure) will significantly reduce memory latency *only* for workloads located in the exponential region of the DDRx latency-bandwidth curve.

"Like the initial memory wall paper, this study points out *something that most of computer architects "knew" without really understanding*. And in order to fully exploit the potential of 3D-stacked DRAMs, we have to *really* understand what we can and what we cannot expect from these devices", says Sally A. McKee, Professor at Computer Science Engineering Department of Chalmers University of Technology.

"Also, Hybrid Memory Cube (HMC) and High Bandwidth Memory (HBM) are much more than high-bandwidth memory devices. The logic layers in the HMC and HBM offer possibilities for in-memory processing and sophisticated memory controller functionality. Finding a way to use this innovation to build high-performance systems, however, will take time", concludes Petar Radojkovic.

This paper is the first outcome of the collaboration between BSC and Samsung Electronics Co., Ltd. that started in 2013 in the context of memory technologies which are in line with Samsung's high density memory solution including 3D-TSV technology for HPC systems. On one side, the collaboration focuses on analyzing how production HPC applications exercise the current DRAM memory system and evaluating the frequency and locality of memory errors in exiting DDR3 technologies. On the other side, the collaboration pursues the proposal of new architectures and management algorithms to exploit the upcoming non-volatile STT-MRAM memory technologies in HPC systems.

About the Barcelona Supercomputing Center (BSC)

Barcelona Supercomputing Center (BSC) is the national supercomputing centre in Spain. BSC specialises in high performance computing (HPC) and its mission is two-fold: to provide infrastructure and supercomputing services to European scientists, and to generate knowledge and technology to transfer to business and society. BSC is a Severo Ochoa Center of Excellence and a first level hosting member of the European research infrastructure PRACE (Partnership for Advanced Computing in Europe). BSC also manages the Spanish Supercomputing Network (RES). More information on page www.bsc.es

Press contact

Barcelona Supercomputing Center
Renata Giménez-Binder
Tel: (+34) 93 4015864
communication@bsc.es