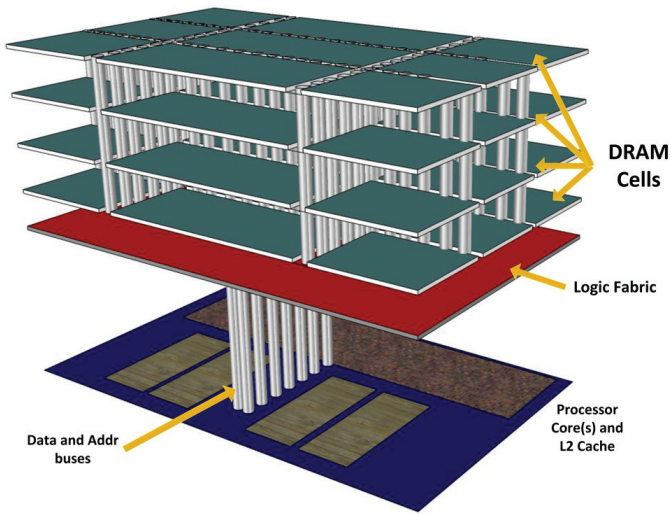
Net-Centric Software & Systems Center (NCSS)

A CISE-funded Center

University of North Texas, Krishna Kavi, Director, 940.369.7216, kavi@cse.unt.edu
University of Texas at Dallas, Farokh B. Bastani, 972.883.2299, bastani@utdallas.edu
Arizona State University, Andreas Spanias, 480.965.1837, spanias@asu.edu
Missouri University of Science and Technology, Sanjay Madria, 573.341.4856, madrias@mst.edu
Southern Methodist University, Jeff Tian, 214.768.2861, tian@lyle.smu.edu

Center website: http://netcentric.cse.unt.edu

Die-Stacked 3D DRAM Processor-Associated Large Memory Technology



The figure shows how 3D DRAM may be manufactured. While some advocate placing 3D DRAM on top of processor cores, the heat dissipation presents challenges.

An emerging DRAM technology, die-stacked 3D DRAM, reduces access latencies for large amounts of memory placed close to processor cores. In addition, very large non-volatile memory technologies such as Flash and Phase-Change Memories (PCM) can be used in place of magnetic disks. These new memory technologies have several advantages over current systems. They present very fast accesses, very large capacities, and potentially lower power requirements.

Before these new memory systems could become commonplace in computer systems, it was necessary to investigate design alternatives on how they can be integrated within a processor memory hierarchy. The low memory latencies and transfer delays required changes to how virtual-to-physical address translation is implemented.

In this breakthrough, NCSS project researchers explored the use of cache-like set-associative indexing to speedup virtual-to-physical address translation and limited the number of page tables needed for address translation.

Researchers evaluated three design alternatives for using 3D DRAM and PCM in memory hierarchy: 3D DRAM as main memory, 3D DRAM as last level cache (LLC) with PCM as main memory, and both 3D DRAM and PCM as main memories. Each of these organizations used cache-like addressing for main memory. In other words, main memory was assumed to behave both as a cache and as a traditional main memory; researchers refer to this organization as Cache Main Memory. The three alternatives presented different benefits: 3D DRAM as LLC provided best performance but consumed more energy, while 3D DRAM as main memory required lower power but executed slower than the other organizations. Using both 3D DRAM and PCM as main memory offered a large design space for exploration.

Current computer systems use hierarchical page tables for virtual to physical address translations, sometimes requiring the system to consult four to seven tables before obtaining the physical address in DRAM. With very low latency 3D DRAMs this approach is unacceptable. Future computer systems, particularly in the server systems, will consist of several 3D DRAM stacks to support the memory requirements of processing cores. The new NCSS approach is leading to more efficient utilization of the new memory technologies. This will result in higher performance and lower power requirements for applications running in Cloud-based computing systems.

> **Economic Impact**: Energy budgets of large data centers are a major factor in the overall cost of supporting "Big Data" applications. This work will positively impact the development of exa-scale computers as well as Big Data applications. The Department of Energy is looking to build next generation super-computers with very tight energy budgets, and is requiring innovative and radical changes to processor and memory systems. It is anticipated that service providers such as Google, Amazon, Microsoft, Hewlett-Packard will realize significant cost savings and efficiencies because they must support very large amounts of data with extremely fast access. The NCSS approach to address translation will aid in reducing execution times and in the process reduce energy requirements. For this reason the die-stacked 3D DRAM technology can be expected to play a significant role in reducing what can only be characterized as major costs associated with addressing data in large data centers.

For more information contact Krishna Kavi, 940.565.2767, Krishna.Kavi@unt.edu.