# Introduction to the Special Issue on High Performance Memory Systems

[1]Haldun Hadimioglu, David Kaeli and Fabrizio Lombardi

While microprocessor designs have continued to increase in speed and complexity, dynamic random access memories (DRAMs) have failed to keep pace. This trend has created a widening gap in performance between microprocessors and their supporting memory systems. While hierarchical memories (i.e., caches) have been used to bridge this gap in the past, the distance (in terms of cycles) between caches and DRAMs continues to grow. Our inability to design memory systems that can keep pace has warranted us to look for new approaches bridging the *memory wall*.

In June 2000, a new workshop was held with the 27th International Symposium on Computer Architecture (ISCA27), titled: "Solving the Memory Wall Problem." Dr. Maurice V. Wilkes provided the keynote talk on "The Memory Gap," and has also provided us with his perspective on "High Performance Memory Systems" in this issue. Many of the papers submitted to this issue were more mature versions of the work presented at this new Workshop. A second workshop was held with ISCA28 in Goteborg, Sweden.

This special issue of IEEE Transactions on Computers in devoted to papers focused on the issue of "Advances in High Performance Memory Systems." Forty-two papers were submitted to this issue, with each paper receiving a minimum of 3 reviews. More than 160 reviews were completed by the referees. We would like to acknowledge the work of these referees in producing an outstanding set of ten papers for this issue. These papers provide a good cross-section of the research underway both in industry and academia on addressing the memory wall problem. We have organized this issue around three broad approaches:

1.) hardware approaches,
2.) architectural approaches, and
3.) compiler/operating system approaches.

**Hardware approaches**

The first paper, "Cache-Memory Interfaces in Compressed Memory Systems" by Benveniste, Franaszek and Robinson, describes a compressed random access memory system that is being used in IBM's Memory Expansion Technology. In the second paper, John Carter and his colleagues from the University of Utah describe the Impulse Memory Controller. Impulse provides us with the ability to perform dynamic optimization at the memory controller interface. The third paper, "High-Performance DRAMs in Workstation Environments" by Cuppa, Jacob, Davis and Mudge, presents a performance study of some of more recent advances in DRAM design. This study evaluates 8 different design features in terms of latency and bandwidth. The fourth paper in this issue titled "Hardware and Software Techniques for Controlling DRAM Power Modes" presents a different aspect of the memory wall problem: power management. The memory system can consume as much as 90% of the overall system energy dissipation. This paper presents a number of hardware and compile-time techniques to manage power more effectively.

**Architectural approaches**

Next we look at three papers that address the memory wall problem utilizing advanced

[1] Appeared in IEEE Transactions on Computers, Introduction to the special issue devoted to "Advances in High Performance Memory Systems," November 2001.

architectural techniques. The next paper in this issue, "Silent Stores and Store Value Locality" by Lepak, Bell and Lipasti, provides an in-depth analysis of the causes of stores which overwrite a value in memory with the same value (i.e., a silent store). By eliminating these stores, we can reduce the demands on the memory system. The sixth paper, "Improving Performance of Large Physically-Indexed Caches by Decoupling Memory Addresses from Cache Addresses" by Min and Hu, proposes using color-indexed physically addressed caches that can reduce conflict misses. The last paper in this group on architectural techniques is titled "Designing a Modern Memory Hierarchy with Hardware Prefetching." In this paper, Lin, Reinhardt and Burger investigate tuning accesses to the second-level cache with the goal to provide sustained bandwidth.

**Compiler/Operating system approaches**

In this section, we have included three papers that address memory performance employing the operating system or advanced compilation techniques. The eighth paper titled "Hardware Compressed Main Memory: Operating System Support and Performance Evaluation" by Abali et al., describes operating system enhancements used to manage the MXT compressed memory system (see the first paper in this issue). In the ninth paper by Barua, Lee, Amarasinghe and Agarwal titled "Compiler Support for Scalable and Efficient Memory Systems," the authors describe the Maps technology used to perform bank disambiguation at compile time. The final paper in this issue titled "Automatic Code Mappng on an Intelligent Memory Architecture" by Solihin, Lee and Torrellas, presents compile-time algorithms for exploiting processor-in-memory architectures.

We would like to thank all authors that submitted paper to this issue. We hope that the reader finds this issue valuable and enjoyable. We would like to thank the students of the Northeastern University Computer Architecture Research Laboratory (NUCAR) for their assistance during the review process and acknowledge the following individuals that served as referees for this special issue:

D. Albonesi, M. Annavaram, H. Aydin, J. Baer, R. Balasubramonian, I. Bahar, S. Bartolini R. Barua, C. Benveniste, L. Bhuyan, D. Burger, M. Burtscher, A. Buyuktosunoglu, B. Calder, J. Carter, J. Chang, M. Charney, F. Chong, B. Cockburn, J. Corbal, V. De La Luz, A. Delis, A. Dominguez, S. Dwarkadas, S. Eggers, D. Elliott, J. Emer, B. Falsafi, M. Farrens, R. Figueiredo, R. Flynn, M. Forsell, H. Franke, E. Gehringer, R. Giorgi, B. Goeman, P. Gonzalez, S. Haga, U. Holze, W. Hong, Y. Hu, R. Ito, B. Jacob, S. Jones, N. Jouppi, J. Kalamatianos, G. Kandiraju, B. Kazar, G. Kedem, D. Keen, K. Keeton, A. Khalafi, P. Kogge, C. Kulkarni, J. Lee, W. Lee, Z. Li D. Lilja, W. Lin, M. Lipasti, C. Luk, M. Martonosi, S. McFarling, F. Meyer, D. Morano, A. Moshovos, T. Mudge, S. Mukherjee, W. Najjar, S. Onder, D. Ortega, N. Park, M. Pearson, V. Piuri, D. Poff, A. Prete, V. Raman, S. Reinhardt, S. Rixner, L. Sadwick, N. Sam, X. Shen, T. Sherwood, A. Sivasubramaniam, Y. Solihin, Y. Song, V. Srinivasan, P. Stenstrom, S. Swarkadas, J. Torrellas, D. Tullsen, G. Tyson, M. Valero, S. Weiss, S. Yang, Q. Yang, E. Yardimci, J. Yi, X. Zhang, Z. Zhang and B. Zorn.

**Summary**

The architecture, design and implementation of high performance memory systems continues to be an area ripe for new advances in technology and algorithms. Researchers will need to aggressively pursue new ideas in this area if we ever intend to close the memory gap.