FlexRAM: Toward an Advanced Intelligent Memory System

A Retrospective Paper

Josep Torrellas
Department of Computer Science
University of Illinois
torrellas@cs.uiuc.edu
http://iacoma.cs.uiuc.edu

I. Context of the Work

The work that lead to our ICCD-1999 FlexRAM paper [4] started in 1996. At that time, there was great interest in the potential of integrating compute capabilities in large DRAM memories — an architecture called Processing-In-Memory (PIM) or Intelligent Memory. Prof. Kogge at the University of Notre Dame had been an early and persistent proponent of the technology since his EXECUBE work [6]. Prof. Patterson at UC Berkeley had been leading the Berkeley IRAM project [11], and co-organized a workshop on these architectures in June 1997 [12]. In addition, Dr. Lucas from DARPA was outlining plans for an effort in this area. Finally, some chip manufacturers were investing in a DRAM technology that could be compatible with high-speed logic — e.g., IBM's CMOS 7LD and Mitsubishi's ERAM.

At Illinois, my research group was trying to understand what a PIM computer architecture might entail. Our initial emphasis was on understanding the needs of the driving applications for this technology. In May 1998, I started a sabbatical at IBM T.J. Watson Research Center, which ended up being very useful for this project. My host, Dr. Pattnaik (also a co-author of the ICCD-1999 paper) joined the project, and put me in touch with IBM experts on DRAM technology and potential applications for PIM systems.

II. Developing the Ideas

We proposed a design that places many tiny cores in the memory system of an otherwise commodity machine. The main processor(s) of the machine is left unmodified, since it works well for legacy applications. The memory cores are small, to minimize losses in memory integration, and numerous, to extract high bandwidth. Moreover, they are general-purpose and not designed to pattern-match any particular algorithm. We also detailed a possible layout of the intelligent memory chips for 180nm. We carefully described how several classes of applications would be mapped to the machine.

As we developed our ideas, other groups proposed somewhat similar designs. In particular, Oskin *et al.* at UC Davis developed Active Pages [10], and Hall *et al.* at ISI developed DIVA [2].

In retrospect, we were naive to assume that it would be reasonable to ask for such a deep integration of many simple cores in the main-memory chips. The DRAM memory industry was and still is focused on low cost and standardization. Thus, one has to find a way to make the design less intrusive on the DRAM memory. On the other hand, two other aspects of the work appear broadly in line with current thinking, namely refraining from hard-wiring the design for particular applications, and starting off with an analysis of the needs of a broad range of potential application domains.

We concluded the paper reflecting that "it is hard to justify a chip like FlexRAM given today's low DRAM prices. However, we believe that, with the fast growth of chip density... [PIM] is the only way to alleviate the memory bottleneck, and possibly the best way to exploit the huge number of transistors available" [4]. These thoughts ring true today.

The FlexRAM project was a major effort in my research group for several years. It involved the work of about 10 graduate students, collaboration with several faculty at Illinois (Profs. Padua, Chien, Reed, and Huang), and interaction with visiting scholars (Profs. Fraguela, Lee, and Feautrier) and IBM researchers (Drs. Pattnaik, Ekanadham, and Lim). We published about 15 papers. They included a revised design of the FlexRAM architecture [15] and its energy analysis [3]. We invested a substantial effort on designing a programming environment for the FlexRAM system [1]. It was composed of: (i) a family of compiler directives called *CFlex* that annotate source code like OpenMP for irregular codes, and (ii) a library of Intelligent Memory Operations (IMOs) that encapsulate operations to be performed in memory processors, such as a vector reduction. Additionally, we examined compiler techniques to map code on a heterogeneous machine with both conventional and memory processors [7]. Finally, we proposed Memory-Side prefetching, where a PIM prefetches data into the processor's cache [14].

III. Intelligent Memory: What Has Happened?

The excitement about PIM lasted for a few years past 2000. DARPA started a program in this area, and there were more projects, such as Stanford's Smart Memories [8]. Moreover, some companies continued to invest on PIM products. For example, Mitsubishi's M32Rx/D was a microcontroller with 4MB of DRAM designed for the embedded market. Micron Technology was working on the Yukon Active Memory prototype [5], which included a large embedded DRAM and 256 simple processors in the same chip.

However, it is fair to say that the excitement fizzled soon after that. The economies of building specialized PIM systems,

with suboptimal DRAM integration and non-standard interfaces were unattractive to industry. With no company pursuing PIMs commercially, research in academia also largely dried out.

Instead, less disruptive technologies such as Multi-Chip Modules (MCM) were used to bring together processor dies and large cache-memory dies. With MCMs, each die is fabricated separately, with its own optimal technology, and then they are integrated into a single package. The result is reduced communication cost and energy. For example, IBM's POWER5 [13] was introduced in 2004, and was packaged in an MCM containing four POWER5 dies and four 36-Mbyte L3 cache dies.

Of course, within the supercomputing research and development community, interest in PIM never died down. Researchers realized that successive semiconductor generations inexorably bring about more and more chip integration. The cost of reaching out from the many processors to the outer memory system quickly increases (relatively speaking) in terms of latency and, most importantly, energy. It therefore makes sense to avoid communication as much as possible, and compute "in place" where the data is stored — hence the motivation for PIM.

Unfortunately, even in this market segment, and despite significant research results showing the merits of PIM, commercial considerations kept preventing PIM technology from emerging. As an example, in the recent High Productivity Computing Systems (HPCS) program from DARPA, there were many discussions and plans for PIM components in the selected designs. However, as the designs have evolved, there are no such components in the final systems (at least publicly disclosed).

IV. Intelligent Memory: The Future

The technical rationale for PIM or Intelligent Memory is compelling. Therefore, it is likely that this technology will appear and be used in the future in some form. In fact, we may already be watching it take shape. Specifically, a few companies such as Micron and Samsung have announced 3-D integrated circuits that stack one or more dies of memory over a die of logic — effectively creating a PIM.

Micron's Hybrid Memory Cube (HMC) [9] has received considerable publicity since it was announced in August 2011. HMC is a memory chip that contains a die of logic sitting below a stack of 4 or 8 DRAM dies, connected using through-silicon-via (TSV) connections. The DRAM dies only store data, while the logic die handles the DRAM control. Currently, the logic die only includes advanced memory controller functions plus self-test and error detection, correction, and repair. The HMC stack can be placed in an MCM with processor dies.

For conventional memory use, this design improves bandwidth, latency and energy characteristics — without changing the high-volume DRAM design. However, it is easy to imagine how to augment the capabilities of the logic die to support Intelligent Memory Operations. These can consist of preprocessing the data as it is read from the DRAM stack into the processor chip. They can also involve performing operations in place on the DRAM data.

It is significant that a memory company has decided to integrate logic with DRAM. It will likely lead to a change in thinking among processor manufacturers. For example, Intel announced in September 2011 that it is working with Micron on HMC development. We may be watching the beginning of true commercial PIMs.

Acknowledgments

The FlexRAM project benefited from the efforts of many people; the author thanks them for their contributions. Thanks also to Peter Kogge, Michael Huang, David Padua, and Jose Renau for discussions on this paper.

References

[1] B. Fraguela, P. Feautrier, J. Renau, D. Padua, and J. Torrellas, "Programming the FlexRAM Parallel Intelligent Memory System," in *International Symposium on Principles and Practice of Parallel Programming (PPoPP)*, June 2003.

[2] M. Hall *et al.*, "Mapping Irregular Applications to DIVA, a PIM-based Data-Intensive Architecture," in *Supercomputing*, November 1999.

[3] M. Huang, J. Renau, S. Yoo, and J. Torrellas, "Energy/Performance Design of Memory Hierarchies for Processor-in-Memory Chips," in *Lecture Notes in Computer Science (Vol. 2107), Springer-Verlag*, 2001.

[4] Y. Kang, M. Huang, S. Yoo, Z. Ge, D. Keen, V. Lam, P. Pattnaik, and J. Torrellas, "FlexRAM: Toward an Advanced Intelligent Memory System," in *International Conference on Computer Design (ICCD)*, October 1999.

[5] G. Kirsch, "Active Memory: Micron's Yukon," in *International Parallel and Distributed Processing Symposium (IPDPS)*, April 2003.

[6] P. Kogge, "The EXECUBE Approach to Massively Parallel Processing," in *International Conference on Parallel Processing*, August 1994.

[7] J. Lee, Y. Solihin, and J. Torrellas, "Automatically Mapping Code on an Intelligent Memory Architecture," in *International Symposium on High-Performance Computer Architecture (HPCA)*, January 2001.

[8] K. Mai, T. Paaske, N. Jayasena, R. Ho, W. Dally, and M. Horowitz, "Smart Memories: A Modular Reconfigurable Architecture," in *International Symposium on Computer Architecture (ISCA)*, June 2000.

[9] Micron Technology, Inc., "Hybrid Memory Cube," 2011, http://www.micron.com/products/hybrid-memory-cube.

[10] M. Oskin, F. Chong, and T. Sherwood, "Active Pages: A Computation Model for Intelligent Memory," in *International Symposium on Computer Architecture*, June 1998, pp. 192–203.

[11] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "A Case for Intelligent DRAM," in *IEEE Micro*, March/April 1997, pp. 33–44.

[12] D. Patterson and M. Smith, "First Workshop on Mixing Logic and DRAM: Chips that Compute and Remember," June 1997.

[13] B. Sinharoy, R. Kalla, J. Tendler, R. Eickemeyer, and J. Joyner, "POWER5 System Microarchitecture," in *IBM Journal of Research and Development*, 2005.

[14] Y. Solihin, J. Lee, and J. Torrellas, "Using a User-Level Memory Thread for Correlation Prefetching," in *International Symposium on Computer Architecture (ISCA)*, May 2002.

[15] S. Yoo, J. Renau, M. Huang, and J. Torrellas, "FlexRAM Architecture Design Parameters," Center for Supercomputing Research and Development (CSRD), Tech. Rep. 1584, October 2000.