# Intelligent RAM (IRAM): the Industrial Setting, Applications, and Architectures

David Patterson, Krste Asanovic, Aaron Brown, Richard Fromm,
Jason Golbus, Benjamin Gribstad, Kimberly Keeton, Christoforos Kozyrakis,
David Martin, Stylianos Perissakis, Randi Thomas, Noah Treuhaft, and Katherine Yelick
Computer Science Division, University of California, Berkeley CA 94720-1776

**Abstract**

*The goal of Intelligent RAM (IRAM) is to design a cost-effective computer by designing a processor in a memory fabrication process, instead of in a conventional logic fabrication process, and include memory on-chip.*

*To design a processor in a DRAM process one must learn about the business and culture of the DRAMs, which is quite different from microprocessors. We describe some of those differences, and then our current vision of IRAM applications, architectures, and implementations.*

## 1. Potential and Challenges of IRAM

Intelligent RAM (IRAM) may lead to a different style of computer than those based on conventional microprocessors. IRAM technology offers the following potential:

- Improve memory latency by factors of 5 to 10 and memory bandwidth by factors of 50 to 100, by redesigning the memory interface and exploiting the proximity of on-chip memory [1][2];
- Improve energy efficiency of memory by factors of 2 to 4, primarily by going off-chip less frequently [3][4];
- Reduce design effort tenfold by filling the die with replicated memory rather than with custom logic [5];
- Make the memory size and organization fit the intended workload;
- Reduce board area by factors of 4 or much greater by integrating many components on a single chip; and
- Improve I/O bandwidth by factors of 4 to 8 by replacing the conventional I/O bus with multiple high-speed, point-to-point, serial lines. [6-8]

This list makes IRAM an exciting opportunity.

One IRAM challenge is matching the performance of microprocessors. Performance obstacles include:

- IRAM is fabricated in a process that has been oriented towards small memory size and low charge leakage rather than fast transistor speed;
- This same DRAM fabrication process offers fewer metal layers than a logic process to lower costs since routing speed is less of an issue in a memory;
- DRAMs are designed to work in plastic packages and dissipate less than 2 watts, while desktop microprocessors dissipate 20 to 50 watts using ceramic packages;
- DRAM refresh rates go up with operating temperature, approximately doubling for every 10 degrees C raise;
- Some applications may not fit within the on-chip memory of an IRAM, and hence IRAMs must access either conventional DRAMs or other IRAMs over a much slower path than on-chip accesses.

Another major IRAM challenge is matching the cost of DRAM memory. Cost obstacles include:

- DRAMs include redundant memory so that fabrication flaws can be circumvented to improve yield and therefore lower cost. Microprocessors traditionally have no redundant logic to improve yield. Hence the on-chip logic may effectively determine the yield of the IRAM.
- Testing time affects chip costs. Given both logic and DRAM on a the same die, an IRAM die may need to be tested on both logic and memory testers.
- To help close the performance gap for logic in a DRAM process, merged-logic DRAM processes are being created with faster transistors and more metal layers which increasing the cost per wafer by 10% to 30%.

The business model for IRAM also has challenges. Although an IRAM may be classified as a single chip computer and sold like desktop or embedded microprocessors, the initial companies most interested in pursuing IRAMs are DRAM companies, and they generally have little experience in the microprocessor market. Some challenges are:

- DRAMs are “generic” parts, used in many places without impacting the software. Putting a processor in the DRAM limits the software that can run on the IRAM.
- The DRAM economic model depends on producing a very high volume of parts—billions of DRAMs are

made each year—while some microprocessors sell less than a million per year.

- DRAMs companies do not need to worry about a supply of support and application software for their chips. IRAM would change that requirement.

This paper first goes into more depth on the DRAM industry to motivate initial solutions to the IRAM challenges, looks at potential IRAM applications and architectures, and then concludes with our target implementation alternatives that could be taped out in 1999.

## 2. DRAM and Microprocessor Industries

Figure 1 highlights some of the differences between the DRAM industry and the desktop microprocessor industry. DRAM companies agree on new standard interfaces for new generations and configurations of DRAMs. These standards include almost everything: pinout, package, addressing, refresh rates, and so on.

Each microprocessor manufacturer generally sets their own instruction set standards to ensure software compatibility with prior generations, but is free to invent new interfaces with different packages and pins, different memory interfaces, and so on. Whereas microprocessors follow their own architecture standards with varying implementations over time, DRAM manufacturers standardize at the package level and innovate in the size of the memory cell and in efficiency of manufacturing process.

| | DRAM | Microprocessor |
|---|---|---|
| Standards | pinout, package, refresh rate, addressing, capacity, width, fast transfer mode, failure rate | binary compatibility, IEEE 754 Floating Point, I/O bus |
| Sources | multiple | single |
| Key figures of merit | 1) capacity, cost/bit<br>2) bandwidth<br>3) latency | 1) performance on standard benchmarks<br>2) cost |
| Rate of improvement | 1) 60%/year, 25%/year<br>2) 20%/year<br>3) 7%/year | 1) 60%/year<br>2) little change |

**Figure 1. Business models of DRAM and desktop microprocessor industries.**

### 2.1. Differing Design Targets

Not only do the multiple source versus single source business model affect the design of the chips, the figures of merit vary between the two cultures. DRAM designers pride themselves on improving storage capacity per chip by fourfold every three years (60%/year) and by having the smallest memory cell so as to have the lowest cost per bit. The capacity increases are generally achieved by reducing cell size by about a factor of 2.5 and increasing the die size by a factor of 1.5. The increasing die size is a major reason that the cost per bit changes more slowly than the capacity per chip. A secondary design target is bandwidth in the fast access mode, and the trailing concern is latency to access a random bit in memory.

A DRAM company's business goal is typically to supply 10% of a single DRAM generation. As there were 6.25 billion DRAMs shipped in 1996, such an apparently modest target can lead to hundreds of millions of chips.

Desktop microprocessor designers tend to have a design cost target, expressed as die size, and then build the fastest chip they can for that size. Microprocessor volumes have more to do with an instruction set target than with the actual final performance, but given that microprocessor designers generally do not get to pick the instruction set, they aim for the highest performance. Of secondary importance is the cost. Recently the power dissipation has become so high that it is now a concern.

Embedded microprocessor designers have much lower cost targets and power budgets, and more likely to sacrifice performance to ensure meeting the cost/power budgets than designers of desktop microprocessors.

The different figures of merit for memory designers and microprocessor designers have resulted in a performance gap between processor and memory in computer systems. The primary approach to bridging this gap has been increasing the amount of SRAM on a microprocessor to act as a cache. Today, many microprocessors dedicate between one-third and two-thirds of the area on chip to these caches.[2] Moreover, today there are often external SRAM chips to build secondary caches. Such chips add to cost and increase board area.

### 2.2. Differing Generation Strategies

Traditionally, DRAM manufacturers would design a new memory cell and a new fabrication process simultaneously. The company then produces tens of thousands of "engineering samples" until both the fabrication process and memory cell design are fully "characterized." Characterization means that the resulting dies will operate at minimum refresh rates over the full temperature range supplying data with acceptable bit error rates.

Once characterized, the subsequent chips are at the "first customer ship" milestone. There may also be a separate milestone of "mass production" when the part achieves the high volume that DRAM manufacturers strive for. Given that all DRAM manufacturers use the same semiconductor fabrication equipment and same wafers, the time to these milestones can determine what share of the market a company will achieve.

The size of the die, testing time, and yield determine profit of a company that has a sizeable market share. As a result, DRAM manufacturers are much more secretive about Spice parameters and design rules than microprocessor companies. To lower costs they shrink the die to increase the number of chips per wafer and improve the fabrication process to improve yield. As they better understand a process, they will reduce the testing time and may even reduce the number of spare rows and columns to get slightly smaller dies. DRAMs typically go through 3 to 4 generations of die sizes over a 4 to 6 year lifetime.

Recently, DRAM manufactures have separated the process and memory cell size from the capacity of the die. Hence the same line might make third generation 64 Mbit and first generation 256 Mbit parts depending on the demands of the market. Today, it makes more sense today to talk about the generations of memory cell size and process rather than just the generation of, say, a 64 Mbit part.

Once in mass production, DRAM die yields below 60% are considered disastrous. Such high yields comes from small die, low defect density, and using redundant rows and columns to repair some flaws. Although real yields are closely guarded secrets, yields of 80% or 90% are apparently achieved by some efficient manufacturers.

Microprocessor manufacturers generally are not as tightly tied to the fabrication process as are DRAM designers. In fact, there are several “fabless” microprocessor manufacturers, but no major “fabless” DRAM manufacturers. Microprocessor designers tend to not worry as much about fully characterizing a design. The key milestones tend to be tape out, booting the operating system on an early chip, and then mass production occurs when the system using the chip is also shipped. Intel, which ships 10 to 100 times the volume of other microprocessor manufacturers, spends much more time on design verification and process tweaking to improve yield.

While every chip designer desires high yield, microprocessor designers typically design chips that almost fill the full reticle and hence may be very happy with initial yields of 20%.

The die is shrunk once as the technology scales, thereby improving yield and increasing clock rate. Companies with high volumes like Intel have a shrink team at work before the die is originally taped out, and will go through more generations of the die than lower volume manufacturers.

### 2.3. Differing Profits

Between 1994 and early 1996, DRAM price per megabyte did not decline by its historical 25% per year. Since technology continued to improve and thus costs continued to decline, the DRAM industry became increasingly profitable.

The economic law of supply and demand was invoked in 1996, as DRAM companies increased production and new companies entered the market. Between January 1996 and December 1996 the price of a 16 Mbit DRAM fell from about $40 per chip to $6 per chip, below the historical 25% per year price decline. Stated alternatively, overall DRAM sales fell from $16.5B in 3Q95 to $7B in 1Q97. And although prices rose to $8 per 16 Mbit DRAM in March 1997, they returned to $6 in August 1997.[9]

At the same time Intel was posting record profits. In 1996 Intel’s net revenue was $20 billion, with a ten year growth rate of 30% per year. In recent quarters about a third of Intel’s income was profit.

In addition to the interesting potential of the IRAM technology, DRAM companies are hoping that IRAM would enable profits per wafer to be more like recent microprocessors wafers than like recent DRAM wafers.

## 3. Potential IRAM applications

For DRAM manufacturers to enjoy the profits of an Intel, they need to find potential IRAM applications that sell in the millions. The first three applications could meet that goal. The last two applications are predicated on the success of one or more of these first three, as they are unlikely to achieve such high volumes.

### 3.1. “Intelligent” Video Game

Nintendo sold 2.6 million of its latest video player for $150 in its first year. Each is based on a four-chip set: one 64-bit MIPS processor chip, one graphic accelerator chip, and two RAMBUS memory chips. Graphics and sound have always needed as much performance as possible, with 3D graphics being especially needy in memory bandwidth and floating point performance.

An IRAM combining the processor, graphics accelerator, and 4 to 16 megabytes of memory could exploit the orders of magnitude in memory bandwidth and small board area advantages of IRAM to offer an attractive chip for the next generation of video games.

### 3.2. “Intelligent” PDA

Palm-top PDAs are becoming increasingly popular. For example, 1 million Palm Pilots were sold in its first year, each for about $300. The Palm Pilot requires the user to learn a new alphabet and then enter the characters with a stylus on a touch sensitive screen. Other PDAs offer miniature keyboards.

If an IRAM could include sufficient computing power to enable speaker trained, isolated-word speech input to a PDA, the device would be much more useful. In such a machine the stylus would be used to correct the errors, usually selected from a pop-up list of potential words. At

90% to 95% word accuracy, achieved by systems like Dragon Dictate, and if 80% to 90% of the time the correct word is found in the popup error menu, then speaking into a PDA could be as fast as typing on a full-sized keyboard.

An IRAM with sufficient performance and 4 to 16 MB of memory to hold the dictionary, when combined with the advantages of energy efficiency and small board area, could be an attractive building block for the next generation of PDAs.

### 3.3. "Intelligent" Disk

Tens of millions of magnetics disks are made each years, and they include integrated circuits with memory for a track cache and logic to calculate the error correction codes for each block. The track cache grows with the increasing linear density of a track, or about 1.3X per year. For example, the 9-GB Seagate Cheetah drive comes with a 0.5 Mbyte track cache and offers a 2.0 Mbyte cache as an option. The new Fibre Channel serial interfaces for disks increase bandwidth demands, requiring transfer rates to the cache be 100 Mbytes/second over two ports.

An IRAM with high-speed serial interfaces could easily supply the required memory capacity and network bandwidth. With sufficient computing power, in addition to calculating error correction codes, it could handle the network and security protocols. Such a disk could attach directly to a local area network, thereby avoiding a server. Such a network-attached secure disk may improve scalability and bandwidth over conventional systems.[10]

As disks will dissipate between 5 and 20 watts, an IRAM for an Intelligent disk must be power efficient. Disks also value small board area very highly, as the chips must fit on the back of 2.5 inch or 3.5 diameter disks.

An attractive chip for disk manufacturers might be a low-power IRAM with 4 to 16 MB of memory for disk caches and networking code plus serial I/O for the interface to disk and local area networks.

### 3.4. Scalable, Low-Cost, Data-Server Cluster

If IRAM proves successful in such high volume markets as those above, such chips may be available to construct much more cost-effective cluster-based servers than those based on conventional desktop microprocessors.

One example comes from the commercial world. One I/O benchmark is Minute Sort, which copies data from disk, sorts it, and then stores it back to disk. This application places the same demands on servers as decision support systems. The current world record is 8.6 GBytes using a cluster of 95 Sun Ultra 1 workstations connected via 160 Mbyte per sec links through switched-based local area network.[11]

Using the serial lines to connect to disks should allow a single IRAM in two to three years to sort more than the current record. Using a few serial lines to connect a cluster of 16 to 32 IRAMs via a switch for network communication and other serial lines connect them to disks could allow this cluster to sort more than 100 GB in a minute. Given that the high volume applications above need inexpensive IRAMs, the cost of 16-32 IRAMs would likely be much less than 10% of the disk infrastructure cost.[7]

Greg Papadopolous, Chief Technical Officer of Sun Microsystems Computing Corporation, observed a trend in data mining. [12] While processors are doubling performance every 18 months, customers are doubling data storage every 5 months. Customers would like to "mine" this data overnight to shape their business practices, but data is being accumulated faster than affordable computers can process the information. Combining Intelligent Disks with an IRAM cluster might lead to scalable processing for data mining that can keep up with "Greg's Law" at a fraction of the costs of the disks.

### 3.5. Low-Cost TeraFLOPS Cluster

A traditional but even lower volume market is supercomputing. Using the same serial networks to connect IRAMs via cross bar switches, hundreds of small, low power IRAMs could be placed on a few small boards. If IRAMs for video games could compute at 1 GFLOPS, then in 2 to 3 years 1000 IRAMs and the disk system needed for the sorting above could offer TeraFLOPS computing for less than $500,000. Figure 2 compares key parameters to the $55,000,000 ASCI Red machine.

Note the smaller memory and higher I/O bandwidth of the IRAM cluster. The sort benchmark was able to trade off higher I/O bandwidth for smaller memory. Whether this would be true for supercomputing remains to be seen.

Even adjusting cost/performance of ASCI Red by a factor of 4 to 6 improvement for technological advances between 1996 and 2000, an IRAM cluster might be attractive for supercomputing.

| | ASCI Red [13] | IRAM cluster |
|---|---|---|
| Processors | 9000 Pentium Pros | 1000 IRAMs |
| Memory | 600 GB | 16-24 GB |
| Disk | 2000 GB | 2100 GB |
| Peak Perf. | 1.8 TeraFLOPS | 1.0 TeraFLOPS |
| I/O speed | 450 GB/s | 2000 GB/s |
| Floor space | 1600 sq. ft. | <10 sq. ft. |
| Cost | $55,000,000 | <$500,000 |
| Year | 1996 | 2000 |

**Figure 2. Supercomputing clusters.**

## 4. IRAM Architectures and Implementations

Putting a conventional cache-based, superscalar microprocessor in an IRAM does not lead to exciting performance.[7][14] Hence IRAM needs a new architecture.

If an architecture requires programmers to rewrite their programs, then it needs advantages of factors of at least 10 and as much as 50.[15] The reason for this high threshold is that software development is slow, and with conventional microprocessor performance doubling every 18 months, there must still be a large advantage after the programming is completed. Otherwise programmers will just wait, as in the long run novel machines are often unsuccessful commercially.

Given the silicon budgets of the next five or so years, its unlikely that any alternative will have that large an advantage over conventional microprocessors for a large set of programs. Keep in mind the DRAM vendors want designs that can be fabricated in the millions, so it is likely that IRAMs will be targeted at many applications.

Hence, in selecting a new architecture, the key is finding a design that exploits the memory bandwidth potential of IRAM while leveraging software developed for traditional computing. Thus an architecture that has offers mature compiler technology is at an advantage. A secondary consideration is energy efficiency. Given the applications in section 3, architectures that reduce power while preserving performance are very attractive for IRAM. Another consideration is small code size to reduce the amount of memory occupied by programs in IRAM.

We see four architectural alternatives: SIMD, VLIW, MIMD on a chip, or vector. While SIMD is a good match to the IRAM technology when the logic is distributed with memory modules, it has never been a general purpose solution. It also has received little compiler development for traditional programming languages. So we rejected it.

VLIW is very popular today in the architecture research community, but it has three negatives. The first is that the compiler technology has not been successful commercially, although it is an area of active compiler research. The second is that VLIW architectures traditionally have the largest code size of the alternatives. The third is object-code compatibility across multiple generations.

MIMD on a chip is a plausible direction for IRAM, and many have taken or are taking this track.[16-18] The MIMD commercial successes have been servers, where the performance is number of tasks per hour rather than time for a single task. While servers are found in section 3, they probably will not have the volumes to justify IRAM. Hence one question is whether a specific MIMD organization lends itself to compiler technology to automatically parallelize an application to run well on all processors with a single chip. A second question is the energy efficiency of fetching four independent instructions streams.

We selected a vector architecture for four reasons. The first is the compiler technology is the most mature of the options, increasing the chances that programs would run on an IRAM with little or no change.

The second reason is that the specification of many parallel operations in a single instruction helps in the power-performance trade-off. Since the power is reduced by the square of a voltage reduction, two techniques allow us to lower power while maintaining performance: deeper pipelines and multiple pipes or lanes. Deeper pipelines make more sense in a vector architecture because the vector operation specifies 64 or 128 operations without a branch. Multiple pipes or lanes means that by including, say, 2 ALUs and cutting clock rate in half we can maintain performance while reducing voltage to lower power.

The third reason is that the multimedia support suggested by video games, PDAs, or data mining is an ideal application for vector architectures. Compared to multimedia extensions such as MMX, vectors are a more elegant way of specifying multiple subword operations. We can simply divide vector registers into smaller elements.

The fourth reason is that the use of multiple pipes or lanes gives the IRAM the ability to have redundant logic that can be discarded to improve yield. With four ALUs, for example, it may cost little in overall area but significantly reduce costs to include a fifth ALU as a spare

## 5. Conclusion

Figure 3 shows the 1999 merged logic-DRAM technology, available from several companies, and parameter estimates of two potential vector IRAMs: low power and high performance. We believe the low power option. is a

| Target | Low Power | High Performance |
|---|---|---|
| Technology | 0.18-0.20 micron, 5-6 metal layers, fast xtor | |
| Die size | ≈200 $mm^2$ | |
| Memory | 16-24 MB | |
| Vector lanes | 4 64-bit (or 8 32-bit or 16 16-bit or 32 8-bit) | |
| Serial I/O | 4 lines @ 1 Gbit/s | 8 lines @ 2 Gbit/s |
| Power | ≈2 w @ 1-1.5 v logic | ≈10 w @ 1.5-2 v |
| $Clock_{univers.}$ | $200_{scalar}/100_{vector}$ MHz | $250_s/250_v$ MHz |
| $Clock_{industry}$ | $400_{scalar}/200_{vector}$ MHz | $500_s/500_v$ MHz |
| $Perf_{university}$ | 0.8 $GFLOPS_{64}$-6 $G_8$ | 2 $GFLOPS_{64}$-16 $G_8$ |
| $Perf_{industry}$ | 1.6 $GFLOPS_{64}$-12 $G_8$ | 4 $GFLOPS_{64}$-32 $G_8$ |

**Figure 3. Low power and high performance Vector IRAM goals to be taped out in 1999.** The two clock rates are for the scalar unit and the vector unit, and the range of the performance is between 64-bit floating point and 8-bit integer.

better match to high volume applications such as video games, PDAs, or disks.

We believe our small, academic design team can build an IRAM with half the performance of a larger and more experienced industrial team. Yet even this design would demonstrate the potential of IRAM to offer an interesting combination of performance, power, memory capacity, board space, and cost.

Several characteristics make IRAM an exciting research topic: large advantages on many dimensions, the design challenges that make success not obvious, the need to rethink the computer design for IRAM, its availability in a fairly standard manufacturing process, and its potential impact on two large industries. Only time can tell us the impact of this intriguing opportunity.

## 6. Acknowledgments

This research was supported by DARPA (DABT63-C-0056), the California State MICRO program, and by research grants from Intel, Samsung, Silicon Graphics/Cray Research, and Sun Microsystems.

## 7. References

[1] Patterson, D.; Anderson, T.; Cardwell, N.; Fromm, R.; Keeton, K.; Kozyrakis, C.; Thomas, R.; Yelick, K. "Intelligent RAM (IRAM): chips that remember and compute," *1997 IEEE International Solids-State Circuits Conference. Digest of Technical Papers*, San Francisco, CA, USA, 6-8 Feb. 1997. p.224-5.

[2] Patterson, D.; Anderson, T.; Cardwell, N.; Fromm, R.; Keeton, K.; Kozyrakis, C.; Thomas, R.; and Yelick, K. "A case for intelligent RAM", *IEEE Micro*, vol.17, (no.2), March-April 1997. p.34-44.

[3] Fromm, R.; Perissakis, S.; Cardwell, N.; Kozyrakis, C.; McGaughy, B.; Patterson, D.; Anderson, T.; Yelick, K. "The energy efficiency of IRAM architectures," *24th Annual International Symposium on Computer Architecture*. (ISCA '97.), Denver, CO, USA, 2-4 June 1997. p.327-37.

[4] Shimizu, T.; et al. "A multimedia 32 b RISC microprocessor with 16 Mb DRAM." *ISSCC Digest of Technical Papers*, San Francisco, CA, USA, 8-10 Feb. 1996 p. 216-17, 448.

[5] Perissakis, S.; Kozyrakis, C.; Anderson, T.; Asanovic, K.; Cardwell, N.; Fromm, R.; Golbus, J.; Gribstad, B.; Keeton, K.; Patterson, D.; Thomas, R.; Treuhaft, N.; and Yelick, K. "Scaling Processors to 1 Billion Transistors and Beyond: IRAM," To appear in *IEEE Computer*, September 1997.

[6] Yang, C.K.K.; Horowitz, M.A. "A 0.8- mu m CMOS 2.5 Gb/s oversampling receiver and transmitter for serial links." *IEEE Journal of Solid-State Circuits*, 31:12, Dec. 1996. p.2015-23.

[7] Keeton, K.; Arpaci-Dusseau, R; and Patterson, D; "IRAM and SmartSIMM: Overcoming the I/O Bus Bottleneck," *Workshop on Mixing Logic and DRAM: Chips that Compute and Remember*, Denver, CO, USA, 1 June 1997. (http://iram.cs.berkeley.edu/isca97-workshop/w2-120-draft.ps)

[8] Saulsbury, A.; Nowatzyk, A. "Missing the memory wall: the case for processor/memory integration." *ISCA'96: The 23rd Annual International Conference on Computer Architecture*, Philadelphia, PA, USA, 22-24 May 1996. p.90-101.

[9] Achilles Corporation; "DRAM Market Price Information in Japan," 1 August 1997. (http://pweb.aix.or.jp/~maski-na/index1-1EG.html)

[10] Gibson, G.A.; Nagle, D.F.; Amiri, K.; Chang, F.W.; Feinberg, E.M.; Gobioff, H.; Lee, C.; Ozceri, B.; Riedel, E.; Rochberg, D.; Zelenka, J. "File server scaling with network-attached secure disks." *1997 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS 97)*, Seattle, WA, USA, 15-18 June 1997. p.272-84.

[11] Arpaci-Dusseau, A.C.; Arpaci-Dusseau, R.H.; Culler, D.E.; Hellerstein, J.M.; Patterson, D.A. "High-performance sorting on networks of workstations." SIGMOD 1997: *ACM SIGMOD International Conference on Management of Data*, Tucson, AZ, USA, 13-15 May 1997. p.243-54.

[12] Papadopolous, G. "The Future of Computing." Unpublished talk at NOW Workshop, Lake Tahoe, CA USA, 27 July 1997.

[13] Rowell, J. "Intel Ships 20 Gflops Teraflops Installment to Sandia," May 9, 1996,(http://www.ssd.intel.com/tflop1.html)

[14] Bowman, N.; Cardwell, N.; Kozyrakis, C.; Romer, C.; and Wang, H. "Evaluation of Existing Architectures in IRAM Systems," *Workshop on Mixing Logic and DRAM: Chips that Compute and Remember*, Denver, CO, USA, 1 June 1997. (http://iram.cs.berkeley.edu/isca97-workshop/w2-114.ps)

[15] Weems, C. "Considerations Leading to an Asynchronous SIMD Architectural Approach for Exploiting Mixed Logic and Memory," *Workshop on Mixing Logic and DRAM: Chips that Compute and Remember*, Denver, CO, USA, 1 June 1997. (http://iram.cs.berkeley.edu/isca97-workshop/w2-108.ps)

[16] Kogge, P.M.; Sunaga, T.; Miyataka, H.; Kitamura, K.; and others. "Combined DRAM and logic chip for massively parallel systems." *Proceedings. 16th Conference on Advanced Research in VLSI*, Chapel Hill, NC, USA, 27-29 March 1995, p. 4-16.

[17] Murakami, K.; Shirakawa, S.; Miyajima, H. "Parallel processing RAM chip with 256 Mb DRAM and quad processors." *1997 IEEE International Solids-State Circuits Conference. Digest of Technical Papers*, San Francisco, CA, USA, 6-8 Feb. 1997, p.228-9, 528.

[18] Yamauchi, T., Hammond, L. and Olukotun, K. "Evaluation of Existing Architectures in IRAM Systems," *Workshop on Mixing Logic and DRAM: Chips that Compute and Remember*, Denver, CO, USA, 1 June 1997. (http://iram.cs.berkeley.edu/isca97-workshop/w2-106.ps)