other devices of the module will be in the Standby state (which is the primary state of all RDRAM devices) or another state with low-power consumption. The RDRAM devices provide several power management features, which are superior to those provided by DDR SDRAM devices. Therefore, Rambus DRAM is a superior solution for portable computers, but special chipsets are needed to take advantage of the power saving modes. An improper power management can cause higher latencies.

The Rambus DRAM devices include a large number of memory banks compared to that of the DDR SDRAM devices. For example, the 128-Mbit and 256-Mbit Rambus DRAM devices contain 32 memory banks versus 4 in the DDR SDRAM devices. This means that a larger number of pages are open at any time and the percentage of page hits is higher. The large number of open pages can decrease the average latency of the Rambus memory module, even if the initial latency is higher.

By incorporating the Rambus memory controller as a part of the CPU itself, much of the current latency problems of this technology will be eliminated. Examples of embedded memory controller solutions are Sun Microsystems' MAJC processor and Compaq Computer's upcoming EV7 (Alpha 21364), which uses 8 Rambus channels to increase bandwidth without increase latencies.

In conclusion, there is not a clear answer whether Rambus DRAM or DDR SDRAM is a better solution. Both technologies have their favorite application environments. In a large number of cases, DDR SDRAM exceeds the performance of Rambus DRAM, but in environments where substantial multi-threading operations and heavy bus loading occurs, both are serious contenders. However, for most applications DDR SDRAM is a better and more attractive solution than Rambus DRAM, at least for the time being and the near future. It has a lower cost, offers better latency, the same bandwidth, and has more industry support.

4.4.6.13. IRAM

Principle of IRAM

IRAM (*Intelligent* RAM) is the name of a chip which is under development at the Berkeley University, consisting of a processor and a large amount of DRAM. This processor is designed in a memory fabrication process, instead of a conventional logic fabrication process. The chip was called *Intelligent* RAM because most of the transistors on this chip are dedicated to memory. The reason to place the processor in DRAM rather than increasing the on-processor SRAM is that DRAM technology allows in practice approximately 20 times higher density than SRAM. This ratio is much larger than the transistor ratio because the DRAM technology uses 3D structures to reduce cell size. Thus, IRAM enables a much larger amount of on-chip memory than is possible in a conventional architecture.

The development of IRAM is based on several observations on today's computer architectures. One of them is the increase of performance gap between processors and memory. While processor performance is improving at a rate of 60% per year, the access time to DRAM memory is improving at a rate of just 7% per year. To

compensate this gap, usually a multilevel cache memory hierarchy is introduced. Unfortunately, this makes the memory latency even higher in the worst case. To build this hierarchy, an increasing fraction of the area within microprocessor chips is devoted to static RAM (SRAM) cache memories. For instance, almost half of the area in the Alpha 21164 processor is occupied by cache memories, used only for hiding memory latency. This cache memory is just a redundant copy of information that would not be necessary if main memory would have enough speed. Nevertheless, some applications show modest access locality, resulting in low performance even with large caches.

Other latency-tolerance techniques include combining large cache memories with some form of out-of-order execution and speculation. This solution requires a disproportionate increase in chip area and complexity, which does not result in a corresponding increase of performance. Other architecture alternatives, like superscalar and VLIW (*Very Long Instruction Word*), have drawbacks such as implementation complexity, low utilization of resources, and insufficiently developed compiler technology.

Beyond the single-processor, the possibility exists for the integration of multiple processors on a single chip, but this integration would place even greater demands on the memory system. For any given chip size, placing more processors on a chip will result in less on-chip memory for each, thus increasing the number of slow, off-chip memory accesses. In general, increasing the computing resources without a corresponding increase in the on-chip memory will lead to an unbalanced system. Functional units will often be idle waiting for data because of the high latency and limited bandwidth with off-chip memory.

The IRAM approach is to use the on-chip area for DRAM memory instead of SRAM cache memories. This on-chip memory can be treated as main memory instead of a redundant copy. In many cases the entire application could be loaded into the on-chip memory. When the application requires more memory than it is available on the IRAM chip, the off-chip memory will be used.

Advantages of IRAM

Researchers from the Berkeley University consider that the integration of a processor and DRAM not only will narrow or remove the processor-memory performance gap, but it will have the following additional benefits:

- Improves memory latency by factors of 5 to 10 and memory bandwidth by factors of 50 to 100. This can be obtained by redesigning the memory interface and exploiting the proximity of on-chip memory.
- Reduces the energy consumption of memory by factors of 2 to 4 by using DRAM as on-chip memory instead of SRAM. Since DRAM consumes less energy than SRAM, on-chip accesses are more energy-efficient. DRAM allows a much higher density than SRAM, and an IRAM will have many fewer external accesses, which consume a large amount of energy.

- Allows a more flexible memory size and organization. Instead of being limited by powers of 2 for the number of memory words or the word size, IRAM designers can specify exactly the number of words and their width. This flexibility can reduce the cost of IRAM solutions versus memories with conventional organization.
- Reduces the board area by a factor of 4 by integrating many components on a single chip. This is an important advantage in applications where board area is critical, such as portable computers or cellular phones.
- Improves 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.

The Berkeley IRAM Project

The IRAM project team, conducted by Prof. David A. Patterson, proposes to design, fabricate, and evaluate single-chip systems for data-intensive applications. The single-chip IRAM system will combine a processor and high capacity DRAM to deliver a floating-point and memory performance comparable to that of vector supercomputers, but at substantially reduced power. The goal is to demonstrate that a single chip with a simple processor and very high local memory bandwidth can be faster than conventional systems on memory-intensive applications. Given that conventional systems have separate chips for the processor, main memory, external cache memory, and networking, an IRAM would be smaller, use less power, and be less expensive. The IRAM architecture will be scalable, allowing the processing power to vary with memory size or power consumption, without changes to the architectural specification. IRAM will also be easily programmable using traditional high-level languages. Another goal is to develop a compiler technology to utilize efficiently the available high bandwidth.

A multi-chip system project which is also undergoing at the Berkeley University, called ISTORE, investigates the problems involved in connecting several IRAM-type processing nodes, combined with a disk drive. The goal is to demonstrate that an ISTORE system with processing power at each disk can deliver very high performance for memory-intensive and I/O-intensive applications. Instead of an expensive centralized processing element and interconnect, ISTORE uses less expensive processors that gain performance through their high bandwidth access to data.

The Berkeley V-IRAM Architecture

Placing a conventional superscalar microprocessor in an IRAM device does not necessarily lead to a very high performance. Therefore IRAM needs a new architecture. Researchers from the Berkeley University have chosen a vector architecture to be integrated into the IRAM device. A vector architecture has several advantages. For instance, the specification of many parallel operations in a single instruction helps to reduce power without affecting performance. Another advantage is that the multime-

dia support needed for video games or portable computers is an ideal application for vector architectures.

The architecture proposed by the Berkeley research group is called Vector IRAM (V-IRAM), and it is an attempt to design an architecture that meets the requirements of the mobile personal computing environment. Figure 4.32 shows a possible floorplan of the V-IRAM chip. The architecture consists of a vector execution unit integrated with a scalar processor. The vector unit consists of two load, one store, and two arithmetic units, each with eight 64-bit pipelines running at 1 GHz. Each pipeline is split into multiple 8-bit pipelines for multimedia operations. The peak performance of the V-IRAM implementation is 16 GFLOPS (at 64 bits per operation) or 128 GOPS. The scalar CPU is a dual-issue superscalar processor and first-level instruction and data cache memories.

The memory system consists of 96 MB of DRAM used as main memory. It is organized in a hierarchical manner as 32 sections, each comprising 16 banks and eight sub-banks per bank, connected to the scalar and vector unit through a crossbar switch. Assuming a pipelined SDRAM-like interface with 20-ns latency and a 4-ns cycle, the memory system can meet the bandwidth demands of the vector unit at 192 GB/s.

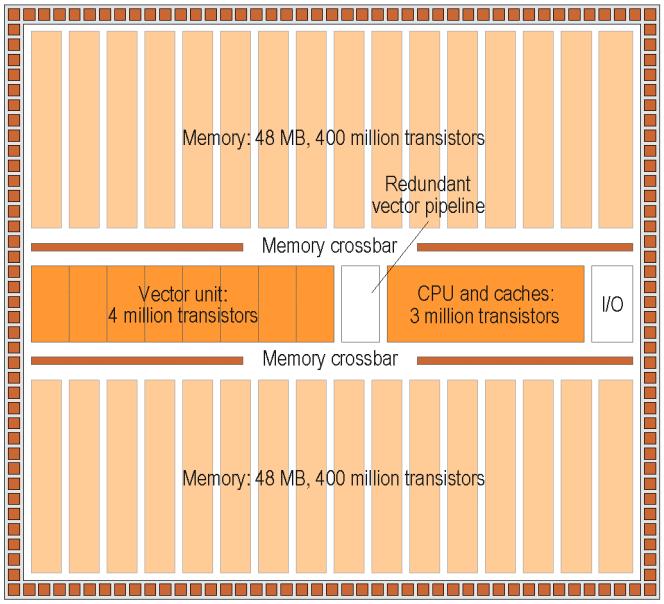

Figure 4.32. Possible floorplan of the Berkeley V-IRAM chip.

External I/O operations use high-speed serial lines (instead of parallel buses) which connects directly to the on-chip memory. The serial lines operate at speeds in the Gbit/s range.

From a programming point of view, V-IRAM is comparable to a vector or SIMD processor. Since most multimedia functions are based on algorithms working

on vectors of pixels or samples, a vector unit can deliver the highest performance. The code size of programs written for vector processors is small compared to that of other architectures. This compactness is possible because a single vector instruction can specify whole loops.

V-IRAM includes a number of DSP features like high-speed multiply-accumulate instructions. It also provides auto-increment addressing, a special addressing mode often found in DSP architectures. V-IRAM is designed to be programmed in high-level languages, unlike most DSP architectures. This is accomplished by avoiding special features and complex instructions that do not allow an efficient compilation. Like all general-purpose processors, V-IRAM provides virtual memory support, which DSP processors do not offer.

V-IRAM has a high energy-efficiency. Each vector instruction specifies a large number of independent operations. Hence, no energy needs to be wasted for fetching and decoding multiple instructions or checking dependencies and making various predictions. In addition, the execution model is strictly in order. Therefore, the control logic is simple and power efficient. In CMOS logic, energy increases with the square of the voltage, so lower voltages can considerably improve energy efficiency. The same performance can be obtained at lower clock rates – and lower voltages – as long as more functional units are added. The hierarchical structure of the on-chip memory provides the ability to activate just the sub-banks containing the necessary data.

Potential Applications of IRAM

IRAM has two important application domains. The first domain is represented by multimedia processing: image processing, video processing, voice recognition, 3D graphics, animation, digital music, encryption. These applications use narrow data types and require real-time response. The second domain is represented by portable and embedded systems: notebooks, PDAs, cellular phones, digital cameras, game consoles. These applications are required to use a limited number of chips and limited power.

Some of the most promising applications of IRAM are presented next.

- *Intelligent Disk*. A large number of magnetic disks are produced each year, and they include integrated circuits with a track cache memory and logic to calculate the error correction codes for each block. The size of the cache memory grows with the increasing linear density of a track. The new interfaces for disks increase bandwidth demands. An IRAM with high-speed serial interfaces could easily supply the required memory capacity and 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 disk may improve scalability and bandwidth over conventional systems.

- *Intelligent PDA*. Palm-top PDAs are becoming increasingly popular. PDAs require the user to 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 speech input to a PDA, the device would be much more useful. In such a case the stylus would be used to correct the errors, usually selected from a list of potential words. At 90% to 95% word accuracy, speaking into a PDA could be as fast as typing on a full-sized keyboard. An IRAM with sufficient performance and enough memory to hold the dictionary, combined with the advantages of energy-efficiency and small board area, could be an attractive building block for the next generation of PDAs.
- *Intelligent Video Game.* Millions of Nintendo and other video players are sold each year. Nintendo video games are based on four chips: one 64-bit MIPS processor, one graphic accelerator chip, and two Rambus memory chips. 3D graphics needs high memory bandwidth and floating point performance. An IRAM combining the processor, graphics accelerator, and 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.

Potential Disadvantages of IRAM

In addition to the advantages, several questions must be answered for IRAM to succeed. With respect to the fabrication process, the major concerns are the speed of transistors and noise. Current developments in the DRAM industry, such as more than two layers of metal, suggest that future DRAM transistors will approach the performance of transistors in logic processes. Noise introduced into the memory array by the fast switching logic can be addressed by using separate power lines for the two system components and by adopting low-power design techniques.

Another concern is the impact of increasing bandwidth on the area and power of DRAM array. Standard DRAM arrays are designed with few, highly multiplexed I/O lines to reduce area and power. To make effective use of a DRAM's internal bandwidth, more I/O lines must be added. The area increase will affect the cost per bit of IRAM.

A more serious consideration is the restricted amount of DRAM that can fit on a single IRAM chip. An amount of 96 MB may be sufficient for portable computers, but not for high-end workstations. A possible solution is to add external DRAM to the IRAM chip, using the off-chip memory as secondary storage with pages swapped between on-chip and off-chip memory. Alternatively, multiple IRAM chips could be interconnected with a high-speed network to form a parallel computer.