**The Gap between Processor and Memory Speeds**

**RAVITEJA YADA**

**NYIT**

**ID:1079869**

**CSCI 641-M02-2016SP-S**

**ASSIGNMENT -4**

**The Gap between Processor and Memory Speeds**

When we are comparing with the Gap between the processor and the memory speed processor is rising to peaks than the memory because communication addresses the recent past and current efforts to attenuate their disparity, namely memory hierarchy strategies, improvement of bus controllers and the development of smarter memories. The reason behind this is the division of the semiconductor industry into microprocessor and memory fields. As their technology is headed in different ways, first one is increase the speed and latter increasing the capacity. As a result of this every year 60% has improved in the microprocessor performance. On the other is the access to DRAM has been increasing at less than 10%/year.

For recognizing where is the problem, let us take a hypothetical computer with a processor that operates at 100 MHz (a Pentium III, for instance), connected to a memory through a 10 MHz bus (SDRAM PC-100). Let us consider that this processor manipulates 100 million items (instructions and/or data) per second and that the memory achieves a 28 ICCA’02 debit (sending or receiving) of 10 million items per second. In this computer, for each single memory access, 8 processor clock cycles have elapsed. This way 7 in each 8 clock cycles are wasted, waiting for items. That represents a very high cost.

The performance of the processor-memory interface is characterized by two parameters: the latency and the bandwidth. The latency is the time between the initiation of a memory request, by the processor, and its completion. Maximum performance is done by the zero lattices and infinite bandwidth.

More recently, Cuppu et al. indicate that the DRAM industry invested some efforts that frequently improved bandwidth, such as

1. Synchronous DRAM (SDRAM),
2. Enhanced SDRAM (ESDRAM),
3. Double data rate DRAM (DDR), and
4. Ram-bus DRAM (RDRAM)

A way to analyze the performance of a memory-hierarchy is through the average memory access time (medium latency), using the following expression:

Average memory access time = hit time + miss rate \* miss penalty

A memory unit is designated as Random Access Memory (RAM) if any location can be accessed in some fixed amount of time that is independent of the location address. The location cells are stored in the form of the array in this cell can store only one bit of information. DRAM has several times the capacity of the SRAM and is cheaper. These are the underlying reasons why DRAMs are widely used in the memory units of computers. The are some DRAM are available such as follow.

Conventional DRAM: In this interface, the address bus is multiplexed between row and column components.

Fast Page Mode DRAM (FPM DRAM): FPM DRAM is advance to conventionl improvement on conventional DRAM in which the row address is held constant while data from multiple columns is read from MDR using several column-addresses.

Extended Data out DRAM: This stores the output data and keeps it stable for the time required for it to be read through the bus.

Ram-bus DRAM (RDRAM): This is the latest DRAM which is mostly used in the market because it is divides the large memory into the bank based on the division. It also transfer the data from several banks and the data transfer is too fast because of electromagnetic interface.

There is one recently one model is proposed named as processor centric architecture, in this we scale is maintain until the two condition holds first, that the processing core has sufficient work to do, to mitigate cache miss latencies; second, that the processor has enough bandwidth to load changes to the cache set without excessive delay. This model has received several denominations: smart memories, intelligent memories, intelligent RAM (IRAM), merged DRAM/Logic (MDL), processor in memory (PIM), etc. To increase the amount of integrated storage space, most of the smart memories proposals use DRAM instead of SRAM.

Processor-in-Memory is a simple approach and the theoretically capable to achieve the good performance, but there is some drawbacks though architecturally simple, serious complications arise in the actual design and production, as most DRAM cores are highly optimized, and can only be modified with difficulty

Vector DRAM has received wider attention to date is the IRAM project and this strategy integrates a complete vector processor on a DRAM chip. Unlike the PIM strategy, the computation is out of the DRAM array, what reduces the peak throughput. The IRAM architecture has a much lower peak performance than the PIM architecture due to it is smaller number of parallel functional units.

Multiprocessor-on-a-Chip is lies between the PIM an IRAM. In this the data is data is stored multiple, simpler, processors onto the same chip offers a number of potential advantages. It maintains the bottleneck architecture process. In this high-level programmability of these designs, they are more easily programmed for maximum parallelism than the other smart memory designs.

Overall is a continuous growing gap between the processors and memoirs speeds is increasing. To this problem some of them had a solutions were oriented to the use of caches: memories of small size, high speed and high cost, that accelerate other memories of high dimension, high speed and reduced cost, which lead to the concept of memory hierarchy and also by taking advantage of the inherent characteristics the this architecture, several hardware and software techniques have been proposed and implemented, making possible the optimization of his operation, answering somehow to the continuous improvement of the processors performance and lessening someway to the discrepancy. Nowadays and in the future continue to take advantage of the memory hierarchy optimization techniques for performance and of the recent progresses in the bandwidth DRAMs field, maintaining the classic separation between processors and memories. By taking the advantage of the parallel processing we can reduce the gap and by placing the several conductors and the transistors in a single chip we can reduce the cost.