ABSTRACT

According to Moore’s law, the number of transistors on chip will increase two fold every couple of years. This results in an exponential increase in microprocessor performance. However, memory systems do not show such an enhancement in their performance hence creating a gap between the performance of the CPU and the memory system. Also, as the CPU has to go off-chip to fetch data, the latency encountered is more. High frequency signal propagation on the PCB transmission lines involves a whole set of signaling issues including but not limited to latency, cross talk, noise etc.

In this report, we study and analyze the various problems involved in signaling and off-chip data fetch. We then propose a solution to this problem in the form of heterogeneous integration of DRAM chip and CPU. We compare this approach with another possible solution namely embedded DRAM or eDRAM.
1. INTRODUCTION

The rate of improvement in microprocessor speed exceeds that of DRAM. Hence computer system designers are faced with an increasing Processor - Memory Performance Gap [1]. This is referred to as the memory wall (Fig. 1) between CPU and memory.

As seen in the above figure, this gap tends to increase in the coming years. Owing to this reason, the available potential bandwidth between CPU and memory is not fully exploited. Furthermore, accessing data from the DRAM involves signaling off-chip. We are then faced with problems of cross talk, noise, the skin effect in transmission lines and other signaling issues. These factors become prominent as signaling frequency increases. We have addressed these problems by proposing a new interface between CPU and memory. A widely observed trend nowadays, is to move the memory controller on-chip [3][4]. We
suggest integrating the processor and memory onto a 3D-IC to alleviate off-chip signaling issues and get better performance.

The rest of the report is organized as follows. Section 2, discusses the organization of modern memory systems and performance issues. Section 3 will discuss the signaling problems on transmission lines. An existing solution to some of these problems, i.e., eDRAM, is covered in section 4. In section 5, we will propose our solution of DRAM on CPU and compare it with the eDRAM. Section 6 will conclude the paper.

2. INTRODUCTION TO MODERN MEMORY SYSTEM

2.1 Memory System Organization

An ideal memory system is practically not implementable as the three features of memory (capacity, speed, and cost) are in direct opposition. The solution is a memory hierarchy organized in several levels, each larger, slower and less expensive per byte than the next as we move away from the CPU.

Closest to the CPU you have registers and then the on-chip cache (SRAM). The total cache is usually a combination of a L1 cache and a larger L2 cache. Large L3 caches are also seen on modern microprocessor systems today. When the processor does not find data in cache, it goes to DRAM. Traditionally the DRAM is off-chip and by design its access latency is much larger. But DRAM is cheaper, denser, and less power hungry than SRAM, which is a 6-transistor/bit structure [2] while DRAM is a 1-transistor-capacitor/bit structure.
Table 1. Access latencies in memory system

<table>
<thead>
<tr>
<th></th>
<th>L1 cache</th>
<th>L2 cache</th>
<th>L3</th>
<th>DRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sizes</td>
<td>32-128Kb</td>
<td>128-512Kb</td>
<td>1-2Mb</td>
<td>64Mb-2Gb</td>
</tr>
<tr>
<td>Latency</td>
<td>3ns</td>
<td>11ns</td>
<td>24ns</td>
<td>45-100ns</td>
</tr>
</tbody>
</table>

A standard configuration of the memory system is as below:

Trends [3] [4] indicate the relocation of the memory controller on to the CPU chip. The integration of a PCI controller and memory controller onto the CPU chip reduces the board complexity in terms of both logic design and the components required. In our discussion here, we will consider a memory system with the controller on-chip.

2.2 Memory Access Mechanism

When the CPU first tries to read from a memory location, it looks for that memory address in the cache. On a cache miss it looks for the address in the DRAM. The address is sent to the
memory controller where it is converted to command sequences. These commands are forwarded off-chip to the DRAM via the transmission lines on the PCB. First the row address is decoded by the DRAM and a row is activated and forwarded to the sense-amps. This time is called $t_{\text{RAS}}$. Then the column address is decoded and that column is sent back to the memory controller via the transmission-line again. This time is called $t_{\text{CAS}}$. The total time to get data from the DRAM is given by the Random Access Delay $t_{\text{RAC}} = t_{\text{RAS}} + t_{\text{CAS}}$ [5].

2.3 The DRAM Interface and Performance Parameters

In this study we will be concentrating on the interconnection between the Memory Controller and the DRAM. The performance of a memory system can be measured in terms of two parameters, latency and bandwidth. Access latency is dependent on process technology and memory topology. Memory Bandwidth is defined as the rate at which the memory system can serve requests from the processor i.e., the amount of data that is transferred in each access, as measured in Mb/s.

The schematic model [6] of the Data bus (DQ) and Data bus strobe (DQS) paths for DDR2 SDRAM is shown in fig 3.
From this example, the longest path between the memory controller and the DRAM I/O buffers is around 14.7cms. Propagation delay of a PCB trace is around 180ps/in [7]. Hence, for this length, we can approximate the delay to be around 1ns. So its round-trip contribution to DRAM access latency is around 2ns. Commodity DRAM $t_{RAC}$ values are around 45ns [8]. But high end DRAMs like FCRAM claim $t_{RAC}$ values of around 25ns. As technology improves, access times of around 10ns will be visible. But the interface latency will still remain the same. So its contribution towards the total latency will increase significantly.

Memory bandwidth can be attacked in two ways: speeding up the system clock and increasing the bus width. Clock rate scaling is difficult as it relies on precise component and PCB modeling. Bandwidth, on the other hand, is limited by the width.
of the data bus and consequently the number of pins on the DRAM. Since the cost of the DRAM is determined to a large extent, by the number of pins on the package, the bandwidth of a memory system is limited. A technology roadmap for semiconductors [9] indicates that the number of transistors on chip will scale the fastest but number of pins on chip, the slowest i.e., adding logic is not going to be an issue but justifying the cost related with extra pins will.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU MHz</td>
<td>3900</td>
<td>8740</td>
<td>12000</td>
<td>19000</td>
<td>29000</td>
</tr>
<tr>
<td>Logic Transistors/cm^2</td>
<td>77.2</td>
<td>154.3</td>
<td>309</td>
<td>617</td>
<td>1235</td>
</tr>
<tr>
<td>High Perf chip pin count</td>
<td>2263</td>
<td>3012</td>
<td>4009</td>
<td>5335</td>
<td>7100</td>
</tr>
<tr>
<td>High Performance chip cost (cents/pin)</td>
<td>1.88</td>
<td>1.61</td>
<td>1.68</td>
<td>1.44</td>
<td>1.22</td>
</tr>
<tr>
<td>Memory pin cost (cents/pin)</td>
<td>0.34-1.39</td>
<td>0.27-0.84</td>
<td>0.22-0.34</td>
<td>0.19-0.39</td>
<td>0.19-0.33</td>
</tr>
<tr>
<td>Memory pin count</td>
<td>48-160</td>
<td>48-160</td>
<td>62-208</td>
<td>81-270</td>
<td>105-351</td>
</tr>
</tbody>
</table>

Table 2. Number of transistors compared with number of pins

Another method of increasing bandwidth is to operate the bus at higher frequencies. Signaling at higher frequencies involves resolving a lot of problems. We will discuss signaling systems in the next section.

3. SIGNALING SYSTEM

3.1 Transmission Line Modeling

A signaling system consists of a Transmitter, Transmission Line, Receiver, Terminator, and a clock. A transmission line can be modeled as an RLC circuit as shown below [10]:
In many situations it is reasonable to set \( G = 0 \). So the behavior of a transmission line will depend on the relationship between \( L \) and \( R \). Every line has a characteristic frequency, \( f_0 = \frac{R}{2\pi L} \), below which it behaves as an RC circuit and above which as an LC circuit. A typical on-chip wire has \( f_0 = 32\text{GHz} \), whereas a typical PC-board has \( f_0 = 2.5\text{MHz} \). Hence, the RC model works for on-chip wires and the LC model for PC board wires.

3.2 Effects of signaling noise

Noise sources in a digital system limit the operational frequency [11]. These include:

1. Power Supply Noise: In particular, single supply noise is of interest here. Single supply noise is the spatial variation in a single supply between two points in the system.

<table>
<thead>
<tr>
<th>Distance</th>
<th>Single Supply Noise (mV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip, local bus, 30um</td>
<td>10</td>
</tr>
<tr>
<td>On-chip, global bus, 1mm</td>
<td>50</td>
</tr>
<tr>
<td>On-chip, global bus, 7mm</td>
<td>350</td>
</tr>
<tr>
<td>Between chips on a PCB</td>
<td>250</td>
</tr>
<tr>
<td>Between PCB, 1m</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 3. Single supply noise variation with the distance
It is quite clear that on-chip wires are less prone to single supply noise as compared to longer PCB wires.

2. Cross talk: Cross talk is the signal interference that arises between two lines running close to one another. On-chip cross talk is primarily due to parasitic capacitance between the lines. It increases the delay. A signal transition on an off-chip transmission line may induce traveling waves on nearby transmission lines owing to parasitic capacitance and mutual inductance between the lines. The noise resulting from crosstalk depends on termination at the transmitter and receiver ends. In the absence of proper termination, off-chip lines show increased crosstalk. On-chip termination is not generally required.

3. Intersymbol Interference: Intersymbol interference is the LC circuit ringing, or extraneous charge storage that occurs due to unrelated values/symbols stored on a channel that sum with the current signal to produce interference. This interference is generally due to reflections resulting from mismatched termination resistors or transmission line discontinuities, which are a common occurrence on long off-chip transmission lines.

4. Timing Noise: Both Skew and Jitter affect the phase rather than the magnitude of the signal. Skew is due to mismatched line lengths and jitter due to power supply noise or additive noise moving the point when the transition is detected.

Timing Closure: Timing Closure problems occur when timing estimates computed during logic synthesis don’t match with timing estimates computed from the layout of the circuit. Timing Closure problems occur mainly due to the difficulty in
accurately predicting interconnect delay during logic synthesis [12].

I/O Buffers: When a signal goes off-chip, the I/O buffers have to drive a significant amount of current to ensure that the signal will be received correctly with sufficient amplitude and signal integrity. Power to drive these I/O buffers is large and is one of the main components of power consumption in DRAMs.

Skin effect: Skin effect is referred to the phenomenon due to which current in the conductor travels only through its surface. This results in increased resistance of the conductor as its effective cross section decreases. Increased current density along the surface can also cause reliability issues. Transmission lines at high frequencies are prone to the skin effect.

The wire delay of a transmission line increases quadratically with its length. In order to reduce this delay both characteristic Resistance and Capacitance need to be reduced. However, since it is difficult to decrease both of them at the same time, the alternative is to have shorter wires.

4. EXISTING SOLUTION: EMBEDDED DRAM

We observe that signaling problems for long off-chip wires are a magnitude more as compared to those faced by the shorter on-chip wires. A solution to achieve both reduced latency and increased bandwidth is to eliminate the off-chip PCB wires from the memory system and move the DRAM on-chip.

Keeping this concept in mind and that DRAM is denser than SRAM; embedded DRAM (eDRAM) was first introduced for use in
high-end system designs.

In an eDRAM system implementation, you can have a small fixed quantity of DRAM on-chip (up to 128Mbit) with or without external DRAM. Generally speaking, eDRAM is applicable to System-on-Chip (SoC) designs because it integrates memory and logic on a single die. For applications where memory requirements are low and no external DRAM is necessary, total chip count in the system and power consumption reduces while performance improves [13].

On-chip DRAMs with very low access latencies (10-20ns) are available. But recently NEC unveiled its eDRAM having access times in the order of 3.5ns, which is comparable to cache access times of 3ns [14].

Though eDRAM may seem like the optimal solution, it has some distinct disadvantages. There are many technological hurdles in integrating DRAM into the same logic process technology as microprocessors [15]. Manufacturing has been particularly troublesome for IC suppliers, as the chips with eDRAM have to be transported between a DRAM fab line and a logic line, where there are wide differences in factors such as annealing temperatures and materials used. Efforts to eliminate this problem are costly and tend to reduce wafer yields [16]. Logic-based eDRAM designs enjoy high-speed and easy compatibility but are very expensive.
Table 4. Die area occupied by eDRAM

Applications that are larger than the on-chip DRAM can negate DRAM advantages [17]. eDRAM significantly increases die area, and hence there is a limit on the amount of DRAM you can
place on-chip. As indicated in the table [18] above, eDRAM takes upto 20 to 50% of die area. Quoting Hideya Horikawa, a design manager for NEC, more than 70% of SoC (system-on-a-chip) die area is expected to be occupied by memory by the year 2005. On-chip DRAM, which can take up a lot of die area, seems like a viable solution for replacing L2 or L3 cache only and not main memory. Most of the advantages of eDRAM are lost if external DRAM is required.

5. PROPOSED SOLUTION: DRAM ON 3D-IC

5.1 DRAM ON 3D-IC

A 3-D chip design strategy [19] exploits the vertical dimension to alleviate interconnect related problems. We propose that by using this technology to integrate the DRAM and the processor, communication between the two will improve drastically.

---

3D ICs are realized by placing multiple dies in a single package. Vertical connectors, called vias, complete paths between different dies through small pads.
Here, the pads required to connect from one layer to the next are very small in area as compared to conventional bonding pads, which are about 100um by 100um [20]. This is because bonding wires and pins are eliminated from the chip-to-chip path. You can achieve large bandwidth between the chips, as you can have thousands of vias.

5.2 Comparison With Embedded DRAM

By placing DRAM on another die and integrating it into a 3D IC, it will have all the advantages of embedded DRAM and eliminate some of its limitations too. Tremendous bandwidth is made available through the many via connections. Latency is reduced by not having to go off-chip and having the address lines of the DRAM accessed by the row decoders hence leading to smaller address setup time [21].
3D DRAMs do not face die size limitations. One can stack multiple DRAM wafers on top of the processor chip and get a large quantity of memory with fast access i.e., no need for external DRAM. Large applications will not lower performance. Also integrating DRAM into the same logic process technology as microprocessors is not an issue here as DRAM can be fabricated using existing DRAM-based designs. Also routing with 3D DRAMs is easier and wire lengths are reduced leading to less signaling issues.

Mainstream adoption of 3D ICs has been sluggish because the fabrication process has proven difficult [22]. But IBM claims to have developed a “wafer-level bonding approach”. This 3D process uses lower temperatures and mechanical stresses, which aids in extending the chip’s life. They employ 130nm silicon-on-insulator devices.

5.3 Some limitations in the development process

There are problems to be overcome. At present, cost of developing 3D chips is still high. Another problem is associated with the densely packed layers of interconnect lines that form the 3D array. The self-heating of interconnect lines within such an array could be significantly more severe due to thermal coupling between neighboring lines. Also upgrading memory is not possible after integration. Hence 3D DRAMs are focused more on embedded applications rather than general purpose computing.

6. CONCLUSIONS AND FUTURE WORK

In this report we have addressed the memory controller-DRAM interface and associated signaling problems. We believe that with the increase in operating frequency, signaling problems
will play a major part in system design. In order to overcome these problems and to increase memory bandwidth while reducing latency, we have proposed a modification in the existing memory controller – DRAM interface. Our approach of DRAM on 3D IC, will address most signaling and bandwidth problems. It can also serve as an attractive alternative approach to embedded DRAM architecture. In the future, we would like to focus on quantifying the parameter values and concentrate on heating related issues.

7. LIST OF REFERENCES


[8] ENEE 759H: Lecture 2 slide 17


