Memory Systems — DRAM, etc.

Prof. Bruce Jacob
Keystone Professor & Director of Computer Engineering Program
Electrical & Computer Engineering
University of Maryland at College Park
Today’s Story

• DRAM
  (the design space is huge, sparsely explored, poorly understood)

• Disk & Flash
  (flash overtaking disk, very little has been published)

• For each, a quick look at some of the non-obvious issues
Perspective: Performance

~10 IPC

~0.001 IPC

~0.0000001 IPC

~0.1 IPC
Perspective: Power

~100 W

~10 W per DIMM

~10 W per Disk

~100–400 W
DRAM
Perspective

DDRx@800Mbps = 6.4GB/s
(x4 DRAM part: 400MB/s, 100mA, 200mW)

Entry system: 2x 3GHz CPU
(2MB cache each), 1GB DRAM, 80GB disk (7.2K)

CPU = $300
DIMM = $30
DRAM = $3
Some Trends

Some Trends

- Storage per CPU socket has been relatively flat for a while

- Note: per-core capacity decreases as # cores increases
Some Trends

• Required BW per core is roughly 1 GB/s

• Thread-based load (SPECjbb), memory set to 52GB/s sustained

• Saturates around 64 cores/threads (~1GB/s per core)

• cf. 32-core Sun Niagara: saturates at 25.6 GB/s
Some Trends

Commodity Systems:

• Low double-digit GB per CPU socket

• $10–100 per DIMM

High End:

• Higher (but still not high) double-digit GB per CPU socket

• ~ $1000 per DIMM

Fully-Buffered DIMM:

• (largely failed) attempt to bridge the gap …
Fully Buffered DIMM

JEDEC DDRx
~10W/DIMM, 20 total

FB-DIMM
~10W/DIMM, ~400W total
The Root of the Problem

Cost of access is high; requires **significant effort** to amortize this over the (increasingly short) payoff.
“Significant Effort”

Outgoing bus request

Read A
Read B
Write X, data
Write Q, data
Write A, data
Read W
Read Z
Read Y

MC

CPU/$

read data

CPU/$

Read A
Read data

Read Z
Read Y

Read Z

Read A

Read W

read data

Read

PRE

ACT

RD

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT

PRE

WR

ACT
**System Level**

One **DRAM device** with eight internal **BANKS**, each of which connects to the shared I/O bus.

One **BANK**, four **ARRAYS**

One **DRAM array** is comprised of many **DRAM ARRAYS**, depending on the part’s configuration. This example shows four arrays, indicating a x4 part (4 data pins).

One **DIMM** can have one **RANK**, two **RANKs**, or even more depending on its configuration.

**Side View**

- Package Pins
- Memory Controller
- Edge Connectors
- DRAMs
- DIMMs

**Top View**

- DIMM 0
- DIMM 1
- DIMM 2
- PCB Bus Traces
- Memory Controller
- Rank 0, Rank 1
  - or
  - Rank 0, Rank 1
  - or even
  - Rank 0/1, Rank 2/3
  - ...

**Internal**

- MUX
- I/O
- One BANK, four ARRAYS
Device Level

- Data I/Out Buffers
- Column Decoder
- Sense Amps
- Row Decoder
- Memory Array

Storage Cell and its Access:
- Word Line
- Bit Line or Digitline
- A transistor
- A capacitor
Issues: Palm HD

- 1920 x 1080 x 36b x 60fps = 560MB/s (~1GB/s incl. ovhd)

- 3 x4 DDR800 = 1.2GB/s, 600mW

- Power budget = 500mW total (DRAM 10–20%)
Issues

**Cache-Bound ≤ 10M***
Much SPECint (not all), etc.
Embedded: mp3 playback

**DRAM-Bound ≤ 10G***
SpecJBB, SPECfp, SAP, etc.
Embedded: HD video

**Disk-Bound ≥ 10G***
TPCC, Google

* Desktop; scale down for embedded
Issues: Cost is Primary Limiter

- CPUs: die area (& power)
  Systems: pins & power
  (desktop: power is cost
   embedded: power is limit)

- FB-DIMM (Intel’s solution to the capacity problem) observed former at cost of latter … R.I.P. FBD

- Whither PERFORMANCE w/o limits? 10x at least
Issues: Education

```c
if (L1(addr) != HIT) {
    if (L2(addr) != HIT) {
        sim += DRAM_LATENCY;
    }
}
```

- Because modeling the memory system is hard, few people do it; because few do it, few understand it
- Memory-system analysis domain of architecture (not circuits)
- Computer designers are enamored w/ CPU … R.I.P. [insert company]
How It Is Represented

if (cache_miss(addr)) {
    cycle_count += DRAM_LATENCY;
}

... even in simulators with “cycle accurate” memory systems—no lie
Issues: Accuracy

- Graphs compare
  - fixed latency
  - queueing model (from industry)
  - “real” model

- Using simple models gives inaccurate insights, leads to poor design

- Inaccuracies scale with workload (this is bad)
Issues: Accuracy

SAP w/ prefetching
Trends …

Jacob, Ng, & Wang: Memory Systems, 2007.
Trends …

Table Ov.4  Cross-comparison of failure rates for SRAM, DRAM, and disk

<table>
<thead>
<tr>
<th>Technology</th>
<th>Failure Rate&lt;sup&gt;a&lt;/sup&gt; (SRAM &amp; DRAM: at 0.13 μm)</th>
<th>Frequency of Multi-bit Errors (Relative to Single-bit Errors)</th>
<th>Expected Service Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>100 per million device-hours</td>
<td>10–20%</td>
<td>Several years</td>
</tr>
<tr>
<td>DRAM</td>
<td>1 per million device-hours</td>
<td></td>
<td>Several years</td>
</tr>
<tr>
<td>Disk</td>
<td>1 per million device-hours</td>
<td></td>
<td>Several years</td>
</tr>
</tbody>
</table>

Table 30.2  Reported SER (for DRAMs)

<table>
<thead>
<tr>
<th>Reported by</th>
<th>Device Gen</th>
<th>Reported FIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>256 KB</td>
<td>27,000 – 160,000</td>
</tr>
<tr>
<td>IBM</td>
<td>1 MB</td>
<td>205 – 40,000</td>
</tr>
<tr>
<td>IBM</td>
<td>4 MB</td>
<td>52 – 10,000</td>
</tr>
<tr>
<td>Micron</td>
<td>16 MB</td>
<td>97 – ?</td>
</tr>
<tr>
<td>Infineon</td>
<td>256 MB</td>
<td>11 – 900</td>
</tr>
</tbody>
</table>

Jacob, Ng, & Wang: Memory Systems, 2007.
Trends …

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi generation (nm)</td>
<td>90</td>
<td>65</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>High perf. device pin count</td>
<td>2263</td>
<td>3012</td>
<td>4009</td>
<td>5335</td>
<td>7100</td>
</tr>
<tr>
<td>High perf. device 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><strong>Memory device pin count</strong></td>
<td><strong>48–160</strong></td>
<td><strong>48–160</strong></td>
<td><strong>62–208</strong></td>
<td><strong>81–270</strong></td>
<td><strong>105–351</strong></td>
</tr>
<tr>
<td>DRAM device 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>
</tbody>
</table>

Jacob, Ng, & Wang: *Memory Systems*, 2007.
**Trends ...**

---

**Table 12.3** Quick summary of SDRAM and DDRx SDRAM devices

<table>
<thead>
<tr>
<th></th>
<th>SDRAM</th>
<th>DDR SDRAM</th>
<th>DDR2 SDRAM</th>
<th>DDR3 SDRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supply voltage</strong></td>
<td>3.3 V</td>
<td>2.5&lt;sup&gt;a&lt;/sup&gt; V</td>
<td>1.8 V</td>
<td>1.5 V</td>
</tr>
<tr>
<td><strong>Signaling</strong></td>
<td>LVTTL</td>
<td>SSTL-2</td>
<td>SSTL-18</td>
<td>SSTL-15</td>
</tr>
<tr>
<td><strong>Bank count</strong></td>
<td>4&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4</td>
<td>4&lt;sup&gt;c&lt;/sup&gt;</td>
<td>8</td>
</tr>
<tr>
<td><strong>Data rate range</strong></td>
<td>66–133</td>
<td>200–400</td>
<td>400–800</td>
<td>800–1600</td>
</tr>
<tr>
<td><strong>Prefetch length</strong></td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td><strong>Internal datapath width</strong></td>
<td>×4</td>
<td>4</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>×8</td>
<td>8</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>×16</td>
<td>16</td>
<td>32</td>
<td>128</td>
</tr>
</tbody>
</table>

<sup>a</sup>400-Mbps DDR SDRAM standard voltage set at 2.6 V.

<sup>b</sup>16-Mbit density SDRAM devices only have 2 banks in each device.

<sup>c</sup>256- and 512-Mbit devices have 4 banks; 1-, 2-, and 4-Gbit DDR2 SDRAM devices have 8 banks in each device.
Trends …

\[ t_{\text{FAW}} \text{ vs. bandwidth (Dave Wang’s thesis)} \]
DISK & FLASH
Disk

Chapter 17 THE PHYSICAL LAYER

RAMAC to the washing machine-size disk drives of the 1970s and 1980s and, finally, to the palm-size disk drives of the 1990s and today. Today's disk drives all have their working components sealed inside an aluminum case, with an electronics card attached to one side. The components must be sealed because, with the very low flying height of the head over the disk surface, just a tiny amount of contaminant can spell disaster for the drive.

This section very briefly describes the various mechanical and magnetic components of a hard disk drive [Sierra 1990, Wang & Taratorin 1999, Ashar 1997, Mee & Daniel 1996, Mamun et al. 2006, Schwaderer & Wilson 1996]. The desirable characteristics of each of these components are discussed. The major physical components are illustrated in Figure 17.8, which shows an exposed view of a disk drive with the cover removed. The principles of operation for most components can be fully explained within this chapter. For the servo system, additional information will be required, and it will be described in Chapter 18.

17.2.1 Disks

The recording medium for hard disk drives is basically a very thin layer of magnetically hard material on a rigid circular substrate [Mee & Daniel 1996]. A flexible substrate is used for a flexible, or floppy, disk. Some of the desirable characteristics of recording media are the following:

- Thin substrate so that it takes up less space
- Light substrate so that it requires less power to spin
- High rigidity for low mechanical resonance and distortion under high rotational speed; needed for servo to accurately follow very narrow tracks
- Flat and smooth surface to allow the head to fly very low without ever making contact with the disk surface
- High coercivity ($H_c$) so that the magnetic recording is stable, even as areal density is increased

---

**FIGURE 17.8:** Major components of today's typical disk drive. The cover of a Hitachi Global Storage Technologies UltraStar™ 15K147 is removed to show the inside of a head-disk assembly. The actuator is parked in the load/unload ramp.
Forget everything you knew about rotating disks. SSDs are different. SSDs are complex software systems. One size doesn't fit all.

Magnet structure of voice coil motor

Spindle & Motor

Disk

Actuator

Flash memory arrays

Circuit board

ATA Interface

Flash SSD
Disk Issues

• Keeping ahead of Flash in price-per-GB is difficult (and expensive)

• Dealing with timing in a polar-coordinate system is non-trivial
  • OS schedules disk requests to optimize both linear & rotational latencies; ideally, OS should not have to become involved at that level

• Tolerating long-latency operations creates fun problems
  • E.g., block-fill not atomic; must reserve buffer for duration; Belady’s MIN designed for disks & thus does not consider incoming block in analysis

• Internal cache & prefetch mechanisms are slightly behind the times
Flash SSD Issues

• Flash does not allow in-place update of data (must block-erase first); implication is significant amount of garbage collection & storage management

• Asymmetric read [1x] & program times [10x] (plus erase time [100x])

• Proprietary firmware (heavily IP-oriented, not public, little published)

  • Lack of models: timing/performance & power, notably Flash Translation Layer is a black box (both good & bad)
  Ditto with garbage collection heuristics, wear leveling, ECC, etc.

  • Result: poorly researched (potentially?)
  E.g., heuristics? how to best organize concurrency? etc.
SanDisk SSD Ultra ATA 2.5” Block Diagram
Flash SSD Organization & Operation

- Numerous Flash arrays
- Arrays controlled externally (controller rel. simple, but can stripe or interleave requests)
- Ganging is device-specific
- FTL manages mapping (VM), ECC, scheduling, wear leveling, data movement
- Host interface emulates HDD
Flash SSD Organization & Operation

- 2 KB Page
- 128 KB Block
- 2 μs page read
- 200 μs page program
- 3 ms block erase
- 32 GB total storage

Flash Memory Bank

2K bytes
Data Reg
Cache Reg
1 Page = 2 K bytes

1 Block
1 Blk = 64 Pages

1024 Blocks per Device (1 Gb)
Flash SSD Timing

**Read 8 KB (4 Pages)**
- **I/O [7:0]**
  - **Cmd**
  - **Addr**
  - **R/W**
  - **Rd0**
  - **Rd1**
  - **Rd2**
  - **Rd3**

- **Xfer from data to cache register**
  - 25 us
  - 3 us

- **Subsequent page is accessed while data is read out from cache register**
  - 2048 cycles
  - 81.92 us

**Write 8 KB (4 Pages)**
- **I/O [7:0]**
  - **Cmd**
  - **Addr**
  - **DI0**
  - **DO0**
  - **DO1**
  - **DO2**
  - **DO3**

- **Xfer from cache to data register**
  - 3 us
  - 200 us

- **Page is programmed while data for subsequent page is written into cache register**
  - 200 us
Some Performance Studies

![Diagram of Flash SSD Organizations]

(a) Single channel  (b) Dedicated channel for each bank  (c) Multiple shared channels

(response time vs. organization graph)

Some Performance Studies
I/O Access Optimization

- Access time increasing with level of banking on single channel
- Increase cache register size
- Reduce # of I/O access requests
I/O Access Optimization

- Implement different bus-access policies for reads and writes

Reads: Hold I/O bus between data bursts

Writes do not need I/O access as frequently as reads