The Memory System and You
A Love/Hate Relationship

Bruce Jacob
Electrical & Computer Engineering
University of Maryland at College Park
blj@umd.edu
Good Quotes from Tues.

Al Geist: Apps get bigger and more complex

Thomas Schulthess: Memory BW primary limiter to solving superconductivity equations

Karl-Heinz Winkler: Memory = 50% power (cf. Zia spec: proc=214W, proc+mem=230W)

Chuck Moore: image/video as data types => larger working sets, larger data types

Bill Camp: “DRAM sucks—that’s the real problem”
Some Trends

Storage per CPU socket has been flat:

(per-core capacity decreased as # cores/CPU increased)

(the capacity problem, Brinda Ganesh’s thesis)
Some Trends

Required BW per core (~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 …
Some Trends: FB-DIMM

- JEDEC DDRx: ~1W/DIMM, ~10W total
- FB-DIMM: 5–10W/DIMM, ~350W total
Some Perspective

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

\[ t_{RP} = 15\text{ns} \quad t_{RCD} = 15\text{ns}, \quad t_{RAS} = 37.5\text{ns} \]

- Bank Precharge
- Row Activate (15ns) and Data Restore (another 22ns)

Column Read
DATA (on bus)

\[ \text{CL} = 8 \quad \text{BL} = 8 \]
Bottom Line

DRAM is performance limiter: HPC apps will always exceed your cache (if you could run it on a laptop, you would)

Memory system determines, to large extent:

- Your performance
- Your power dissipation
- Your system cost

PROBLEM:

- Nobody models it accurately
What It Looks Like

CPU/$

Read A

Outgoing bus request

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

MC

CPU/$

read data

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

MC

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

MC

read data

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

MC

read data

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

MC

read data

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

MC

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

MC

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

MC

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

MC
How It Is Represented

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

... even in simulators with “cycle accurate” memory systems—no lie
Some Cases In Point

Prefetching, Multicore, etc.
Some Cases In Point

Prefetching, Multicore, etc.

SLIDE 12

![Graph showing IPC Difference Normalized to a Cycle Accurate Model for TPCC with different core counts and various models: SILM, SALM, QILM, QALM.](image-url)
Some Cases In Point

Prefetching, Multicore, etc.

SOS-13
March 2009

![Bar chart showing IPC difference normalized to a cycle-accurate model across different core configurations](chart.png)
Your Choice

Today, DRAM is primary limiter. If you really want to predict system behavior

- performance
- power
- cost

you have to model the memory system, accurately.

... so what do you get if you do?
DRAMsim

System Parameters
DRAMsim

Device Parameters

The image shows a screenshot of the DRAMsim tool with Device Parameters window open. The parameters include:

- **Physical**
  - DEVICE_WIDTH: 8
  - NUM_BANKS: 8
  - NUM_COLS: 1024
  - NUM_ROWS: 16384
  - REFRESH_PERIOD: 7800

- **Power**
  - IDD0: 85
  - IDD1: 100
  - IDD2N: 40
  - IDD2P: 7
  - IDD2Q: 40
  - IDD3N: 55
  - IDD3Pf: 30
  - IDD3Ps: 10
  - IDD4R: 135
  - IDD4W: 135
  - IDD5: 215
  - IDD6: 7
  - IDD6L: 5
  - IDD7: 280

- **Timing**
  - AL: 0
  - BL: 4

The context is likely related to memory systems education, as indicated by the text at the top of the page.
DRAMsim

Overlay Graphs from Multiple Runs
DRAMsim Results: Epetra

bandwidth, open page

bandwidth, closed page
DRAMsim Results: Epetra

latency, closed page

latency, open page
DRAMsim Results: Epetra

Bandwidth (GB/s)

BANDWIDTH, addr2
DRAMsim Results: Epetra

Bandwidth (GB/s)

BANDWIDTH, addr4
DRAMsim Results: Epetra

Power Dissipation (watts)

Average Power (watts)

POWER, addr2

Time (milliseconds)
DRAMsim Results: Epetra

Power Dissipation (watts)

Average Power (watts)

POWER, addr4

Time (milliseconds)
Read Response Time vs. Organization

Configuration

- 1 x 8 x 25
- 1 x 8 x 50
- 1 x 16 x 25
- 1 x 16 x 50
- 2 x 8 x 25
- 2 x 8 x 50
- 2 x 16 x 25
- 2 x 16 x 50
- 4 x 8 x 25
- 4 x 8 x 50
- 4 x 16 x 25
- 4 x 16 x 50

Figure 8: Flash SSD Organizations. (a) Single I/O bus is shared - 1, 2, or 4 banks; (b) dedicated I/O bus: 1, 2, or 4 buses and single bank per bus; (c) multiple shared I/O channels - 2 or 4 channels with 2 or 4 banks per channel.
... and tons more

Address-mapping policies (2–10x)

Scheduling policies (2–10x)
(e.g., improved Cray MC by 20% sustained BW)

Paging policies (2x)

System organization (10+x)

Blocking, cache interactions, OS interactions
(e.g. virtual memory), file system interactions,
queueing mechanisms, device architectures, …

If you want to reduce power, this is where to look
If you want to increase performance, this is where to look