Efficient Data Access in Future Memory Hierarchies

Rajeev Balasubramonian

School of Computing
Research Buffet, Fall 2010
Acks

- Terrific students in the Utah Arch group
- Collaborators at HP, Intel, IBM
- Prof. Al Davis, who re-introduced us to memory systems
Current Trends

• Continued device scaling
• Multi-core processors
• The power wall
• Pin limitations
• Problematic interconnects
• Need for high reliability
Anatomy of Future High-Perf Processors

The Core

Hi-Perf

The Core

Lo-Perf

- Designs well understood
- Combo of hi- and lo-perf
- Risky Ph.D.!!
Anatomy of Future High-Perf Processors

- Large shared L3
- Partitioned into many banks
- Assume one bank per core
Anatomy of Future High-Perf Processors

- Many cores!
- Large distributed L3 cache
Anatomy of Future High-Perf Processors

- On-chip network
- Includes routers and long wires
- Used for cache coherence
- Used for off-chip requests/responses
Anatomy of Future High-Perf Processors

- Memory controller handles off-chip requests to memory
Anatomy of Future High-Perf Processors

- Multi-socket motherboard
- Lots of cores, lots of memory, all connected
Anatomy of Future High-Perf Processors

- DRAM backed up by slower, higher capacity emerging non-volatile memories
- Eventually backed up by disk… maybe
Life of a Cache Miss

- Miss in L1 (Core L1)
- Look up L2/L3 bank
- On- and off-chip network
- Wait in MC queue
- Access DRAM (DIMM)
- Access PCM (Non-Volatile PCM)
- Access disk (DISK)
- on-chip network
Research Topics

Core L1

Miss in L1

on-chip network

Look up L2/L3 bank

On- and off-chip network

Wait in MC queue

Non Volatile PCM

Access PCM

Access DIMM

Access DRAM

DISK

Access disk

Not very hot topics!
Research Topics

1. Access DRAM
2. Access disk
3. Look up L2/L3 bank
4. On- and off-chip network
5. Wait in MC queue
6. Access PCM

- Core L1
- Non Volatile PCM
- DISK
- DIMM
- MC

Miss in L1
on-chip network

Look up L2/L3 bank

Access DRAM
Access PCM
Access disk
Problems with DRAM

- DRAM main memory contributes $1/3^{rd}$ of total energy in datacenters
- Long latencies; high bandwidth needs
- Error resilience is very expensive
- DRAM is a commodity and chips have to be compliant with standards
  - Initial designs instituted in the 1990s
  - Innovations are evolutionary
  - Traditional focus on density
Time for a Revolutionary Change?

• Energy is far more critical today
• Cost-per-bit perhaps not as relevant today
• Memory reliability is increasing in importance
• Multi core access streams have poor locality
• Queuing delays are starting to dominate
• Potential migration to new memory technologies and interconnects
Key Idea

It’s worth a small increase in capital costs to gain large reductions in operating costs.

And not 10X, just 15-20%!
DRAM Basics

- Array
- 1/8th of the row buffer
- One word of data output
- Bank
- DRAM chip or device
- Rank
- DIMM
- Memory bus or channel
- On-chip Memory Controller
DRAM Operation

One bank shown in each chip
New Design Philosophy

• Eliminate overfetch; activate a single chip and a single small array $\rightarrow$ much lower energy, slightly higher cost

• Provide higher parallelism

• Add features for error detection

[ Appears in ISCA’10 paper, Udipi et al. ]
Single Subarray Access (SSA)
SSA Operation

Sleep Mode (or other parallel accesses)
Consequences

• Why is this good?
  ▪ Minimal activation energy for a line
  ▪ More circuits can be placed in low-power sleep
  ▪ Can perform multiple operations in parallel

• Why is this bad?
  ▪ Higher area and cost (roughly 5 – 10%)
  ▪ Longer data transfer time
  ▪ Not compatible with today’s standards
  ▪ No opportunity for row buffer hits
Narrow Vs. Wide Buses?

• What provides higher utilization? 1 wide bus or 8 narrow buses?

• Must worry about load imbalance and long data transfer time in latter

• Must worry about many bubbles in the former because of dependences

[ Ongoing work, Chatterjee et al. ]
Methodology

- Tested with simulators (Simics) and multi-threaded benchmark suites
- 8-core simulations
- DRAM energy and latency from Micron datasheets
Moving to close page policy – 73% energy increase on average
Compared to open page, 3X reduction with SBA, 6.4X with SSA
Results – Energy – Breakdown

- Termination Resistors
- Global Interconnect
- Bitlines
- Decoder + Wordline + Senseamps
• Serialization/Queuing delay balance in SSA - 30% decrease (6/12) or 40% increase (6/12)
Results – Performance – Breakdown

- **BASELINE (OPEN PAGE, FR-FCFS)**
- **BASELINE (CLOSED ROW, FCFS)**
- **SBA**
- **SSA**

- **Data Transfer**
- **DRAM Core Access**
- **Rank Switching delay (ODT)**
- **Command/Addr Transfer**
- **Queuing Delay**
Error Resilience in DRAM

- Important to not only withstand a single error, but also entire chip failure – referred to as *chipkill*

- DRAM chips do not include error correction features -- error tolerance must be built on top

- Example: 8-bit ECC code for a 64-bit word; for chipkill correctness, each of the 72 bits must be read from a separate DRAM chip → significant overfetch!

72-bit word on every bus transfer
Two-Tiered Solution

• Add a small (8-bit) checksum for every cache line
• Maintain one extra DRAM chip for parity across 8 DRAM chips
• When the checksum flags an error, use the parity to re-construct the corrupted cache line
• Writes will require updates to parity as well
Research Topics

1. Access DRAM
2. Access disk
3. Look up L2/L3 bank
4. On- and off-chip network
5. Wait in MC queue
6. Access PCM

Core L1

Miss in L1

on-chip network

$ Look up L2/L3 bank

Non Volatile PCM

Access PCM

DIMM

Access DRAM

MC
Topic 2 – On-chip Networks

- Conventional wisdom: buses are not scalable; need routed packet-switched networks

- But, routed networks require bulky energy-hungry routers

- Results:
  - Buses can be made scalable by having a hierarchy of buses and Bloom filters to stifle broadcasts
  - Low-swing buses can yield low energy, simpler coherence, and scalable operation

[ Appears in HPCA’10 paper, Udipi et al. ]
• In a large distributed NUCA cache, or in a large distributed NUMA memory, data placement must be controlled with heuristics that are aware of:
  - capacity pressure in the cache bank
  - distance between CPU and cache bank and DIMM
  - queuing delays at memory controller
  - potential for row buffer conflicts at each DIMM

[ Appears in HPCA’09, Awasthi et al. and in PACT’10, Awasthi et al. (Best paper!) ]
Topic 4: Silicon Photonic Interconnects

- Silicon photonics can provide abundant bandwidth and makes sense for off-chip communication
- Can help multi-cores overcome the bandwidth problem
- Problems: DRAM design that best interfaces with silicon photonics, protocols that allow scalable operation

[ On-going work, Udipi et al. ]
Topic 5: Memory Controller Design

• Problem: abysmal row buffer hit rates, quality of service

• Solutions:

  ▪ Co-locate hot cache lines in the same page

  ▪ Predict row buffer usage and “prefetch” row closure

  ▪ QoS policies that leverage multiple memory “knobs”

[ Appears in ASPLOS’10 paper, Sudan et al. and on-going work, Awasthi et al., Sudan et al. ]
Emerging memories (PCM):
- can provide higher densities at smaller feature sizes
- are based on resistance, not charge (hence non-volatile)
- can serve as replacements to DRAM/Flash/disk

Problem: when a cell is programmed to a given resistance, the resistance tends to drift over time → may require efficient refresh or error correction

[ On-going work, Awasthi et al. ]
Title

• Bullet