CS 61C: Great Ideas in Computer Architecture (Machine Structures)

Warehouse-Scale Computing

Instructors:
Nicholas Weaver & Vladimir Stojanovic

http://inst.eecs.berkeley.edu/~cs61c/
Coherency Tracked by Cache Block

• Block ping-pongs between two caches even though processors are accessing disjoint variables
• Effect called *false sharing*
• How can you prevent it?
Review: Understanding Cache Misses: The 3Cs

• **Compulsory** (cold start or process migration, 1\textsuperscript{st} reference):
  – First access to block, impossible to avoid; small effect for long-running programs
  – Solution: increase block size (increases miss penalty; very large blocks could increase miss rate)

• **Capacity** (not compulsory and...)
  – Cache cannot contain all blocks accessed by the program even with *perfect replacement policy in fully associative cache*
  – Solution: increase cache size (may increase access time)

• **Conflict** (not compulsory or capacity and...):
  – Multiple memory locations map to the same cache location
  – Solution 1: increase cache size
  – Solution 2: increase associativity (may increase access time)
  – Solution 3: improve replacement policy, e.g., LRU
Fourth “C” of Cache Misses: 

**Coherence Misses**

- Misses caused by coherence traffic with other processor
- Also known as *communication* misses because represents data moving between processors working together on a parallel program
- For some parallel programs, coherence misses can dominate total misses
  - It gets even more complicated with multithreaded processors: You want separate threads on the same CPU to have common working set, otherwise you get what could be described as *in*coherence misses
New-School Machine Structures (It’s a bit more complicated!)

- **Parallel Requests**
  Assigned to computer
  e.g., Search “cats”

- **Parallel Threads**
  Assigned to core
  e.g., Lookup, Ads

- **Parallel Instructions**
  >1 instruction @ one time
  e.g., 5 pipelined instructions

- **Parallel Data**
  >1 data item @ one time
  e.g., Deep Learning for image classification

- **Hardware descriptions**
  All gates @ one time

- **Programming Languages**

---

**Diagram Description:**
- **Software**
  - Parallel Requests
  - Parallel Threads
  - Parallel Instructions
  - Parallel Data
  - Hardware descriptions
  - Programming Languages

- **Hardware**
  - Core
  - Memory
  - Input/Output
  - Instruction Unit(s)
    - $A_0 + B_0$
    - $A_1 + B_1$
    - $A_2 + B_2$
    - $A_3 + B_3$
  - Functional Unit(s)
  - Cache Memory
  - Logic Gates
Back in 2011

• Google disclosed that it continuously uses enough electricity to power 200,000 homes, but it says that in doing so, it also makes the planet greener.

• Average energy use per typical user per month is same as running a 60-watt bulb for 3 hours (180 watt-hours).

Google’s WSCs

Ex: In Oregon
Containers in WSCs

Inside WSC

Inside Container
Server, Rack, Array
Google Server Internals
Warehouse-Scale Computers

• Datacenter
  – Collection of 10,000 to 100,000 servers
  – Networks connecting them together

• *Single gigantic* machine

• Very large applications (Internet service): search, email, video sharing, social networking

• Very high availability

• “...WSCs are no less worthy of the expertise of computer systems architects than any other class of machines” Barroso and Hoelzle, 2009
Unique to WSCs

• Ample Parallelism
  – **Request-level Parallelism**: ex: Web search
  – **Data-level Parallelism**: ex: Image classifier training

• Scale and its Opportunities/Problems
  – **Scale of economy**: low per-unit cost
  – Cloud computing: rent computing power with low costs (ex: AWS)

• Operation Cost Count
  – Longer life time (>10 years)
  – **Cost of equipment purchases << cost of ownership**
  – Often semi-custom or custom hardware
    • But consortiums of hardware designs to save cost there

• Design for failure:
  – Transient failures
  – Hard failures
  – **High # of failures**
    ex: 4 disks/server, annual failure rate: 4%
    ⇒ WSC of 50,000 servers: 1 disk fail/hour
WSC Architecture

1U Server:
8-16 cores,
16 GB DRAM,
4x4 TB disk
+ disk pods

Rack:
40-80 servers,
Local Ethernet (1-10Gbps) switch
(30$/1Gbps/server)

Array (aka cluster):
16-32 racks
Expensive switch
(10X bandwidth \(\Rightarrow\) 100x cost)
WSC Storage Hierarchy

Lower latency to DRAM in another server than local disk
Higher bandwidth to local disk than to DRAM in another server

1U Server:
DRAM: 16GB, 100ns, 20GB/s
Disk: 2TB, 10ms, 200MB/s

Rack (80 servers):
DRAM: 1TB, 300us, 100MB/s
Disk: 160TB, 11ms, 100MB/s

Array (30 racks):
DRAM: 30TB, 500us, 10MB/s
Disk: 4.80PB, 12ms, 10MB/s
• Online service: Peak usage 2X off-peak
Impact on WSC software

- **Latency, bandwidth** → Performance
  - Independent data set within an array
  - Locality of access within server or rack

- **High failure rate** → Reliability, Availability
  - Preventing failures is effectively *impossible* at this scale
  - Cope with failures gracefully by designing the system as a whole

- **Varying workloads** → Availability
  - Scale up and down gracefully

- More challenging than software for single computers!
Power Usage Effectiveness

• Energy efficiency
  – Primary concern in the design of WSC
  – Important component of the total cost of ownership

• Power Usage Effectiveness (PUE):

\[
\text{Total Building Power} \div \text{IT equipment Power}
\]

– A power efficiency measure for WSC
– Not considering efficiency of servers, networking
– Perfection = 1.0

– Google WSC’s PUE = 1.2
  • Getting pretty close to Amdahl's law limit
PUE in the Wild (2007)

FIGURE 5.1: LBNL survey of the power usage efficiency of 24 datacenters, 2007 (Greenberg et al.)
Where Data Center Power Goes

- Electricity Transformer/UPS: 10%
- Air Movement: 12%
- Cooling: 25%
- Lighting, etc.: 3%
- IT Equipment: 50%
Load Profile of WSCs

- Average CPU utilization of 5,000 Google servers, 6 month period
- Servers rarely idle or fully utilized, operating most of the time at 10% to 50% of their maximum utilization
Energy-Proportional Computing: Design Goal of WSC

- Energy = Power x Time, Efficiency = Computation / Energy
- Desire:
  - Consume almost no power when idle (“Doing nothing well”)
  - Gradually consume more power as the activity level increases
Cause of Poor Energy Proportionality

- CPU: 50% at peek, 30% at idle
- DRAM, disks, networking: 70% at idle!
  - Because they are never really idle unless they are powered off!
- Need to improve the energy efficiency of peripherals
Clicker/Peer Instruction: Which Statement is True

- A: Idle servers consume almost no power.
- B: Disks will fail once in 20 years, so failure is not a problem of WSC.
- C: The search requests of the same keyword from different users are dependent.
- D: More than half of the power of WSCs goes into cooling.
- E: WSCs contain many copies of data.
Administrivia

• Reminder that Project 4 is out...
Agenda

• Warehouse Scale Computing
• Administrivia & Clickers/Peer Instructions
• Request-level Parallelism
  e.g. Web search
Request-Level Parallelism (RLP)

• Hundreds of thousands of requests per sec.
  – Popular Internet services like web search, social networking, ...
  – Such requests are largely independent
    • Often involve read-mostly databases
    • Rarely involve read-write sharing or synchronization across requests
• Computation easily partitioned across different requests and even within a request
Google Query-Serving Architecture
Anatomy of a Web Search

cats

About 650,000,000 results (0.29 seconds)

Black Cats are Good Luck - berkeleyhumane.org
In October, Adopt a Black Cat For only $10. Save a Life Today!

Cat - Wikipedia, the free encyclopedia
The domestic cat (Felis catus or Felis silvestris catus) is a small, typically furry, domesticated, and carnivorous mammal. They are often called house cats when...

African wildcat - Creme Puff - List of cat breeds - Human interaction with cats

Cats (musical) - Wikipedia, the free encyclopedia
Cats is a musical composed by Andrew Lloyd Webber, based on Old Possum's Book of Practical Cats by T. S. Eliot, and produced by Cameron Mackintosh.

Music: Andrew Lloyd Webber
Lyrics: T. S. Eliot; Trevor Nunn (addition...

Cats - Reddit
Your reddit account must be at least 15 days of age to post in /r/cats. Redditors ... The mom cat has a very special mark on her coat that I think you all would like.

Cats: Pictures, Videos, Breaking News - Huffington Post
Big News on Cats. Includes blogs, news, and community conversations about Cats.

Cats on About.com - All About Cats and Kittens
cats.about.com » About Home
Learn all about the care and feeding of cats. Free articles on cat behavior, cat health, pregnancy and birth, vet care and the human bond with cats.

Cat Health Center | Cat Care and Information from WebMD
WebMD veterinary experts provide comprehensive information about cat health care, offer nutrition and feeding tips, and help you identify illnesses in cats.

Cat killer on the loose? Police think so
Detroit Free Press - 17 hours ago
Police say someone has been beating cats to death in Hazel Park two blocks north of Detroit ...

New Study Finds Cats Have The Surface Area Of A Ping Pong Table
Popular Science - 18 hours ago
This breed of cats makes them look just like werewolves
AOL News - 2 days ago

More news for cats

Funny Cats Big Compilation 2015! [NEW] - YouTube
Dec 2, 2014 - Uploaded by Funny Animals Channel
New Crazy compilation of 2014. ENJOY and SUBSCRIBE, Merry Christmas!
Anatomy of a Web Search (1/3)

• Google “cats”
  – Direct request to “closest” Google WSC
    • Handled by DNS
  – Front-end load balancer directs request to one of many arrays (cluster of servers) within WSC
    • One of potentially many load balancers
  – Within array, select one of many Google Web Servers (GWS) to handle the request and compose the response pages
  – GWS communicates with Index Servers to find documents that contains the search word, “cats”
    • Index servers keep index in RAM, not on disk
  – Return document list with associated relevance score
Anatomy of a Web Search (2/3)

• In parallel,
  – Ad system: run ad auction for bidders on search terms
    • Yes, you are being bought and sold in a realtime auction all over the web
    • Page ads are worse than search ads
  
• Use docids (Document IDs) to access indexed documents

• Compose the page
  – Result document extracts (with keyword in context) ordered by relevance score
  – Sponsored links (along the top) and advertisements (along the sides)
Anatomy of a Web Search (3/3)

• Implementation strategy
  – Randomly distribute the entries
  – Make many copies of data (a.k.a. “replicas”)
  – Load balance requests across replicas

• Redundant copies of indices and documents
  – Breaks up search hot spots, e.g. “Taylor Swift”
  – Increases opportunities for request-level parallelism
  – Makes the system more tolerant of failures
Summary

• Warehouse Scale Computers
  – New class of computers
  – Scalability, energy efficiency, high failure rate

• Request-level parallelism
  e.g. Web Search

• Data-level parallelism on a large dataset
  – A gazillion VMs for different people
  – MapReduce
  – Hadoop, Spark