Great Ideas in Computer Architecture

Warehouse Scale Computers, MapReduce

Instructor: Nick Riasanovsky

Oh no you giga dumbo you’ve just got
Review of Last Lecture

• OpenMP as simple parallel extension to C
  – Synchronization accomplished with critical/atomic/reduction
  – Pitfalls can reduce speedup or break program logic

• Cache coherence implements shared memory even with multiple copies in multiple caches
  – The protocol we learned was MOESI
  – False sharing renders a block useless! A ping-pong chain of invalidation

  • Coherence misses are the fourth cache miss type
Agenda

• Warehouse Scale Computers
• Administrivia
• Cloud Computing
• Request Level Parallelism
• MapReduce
Great Idea #4: Parallelism

- **Parallel Requests**
  Assigned to computer
  e.g. search “Steven Ho”

- **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. add a pair of 6 words

- **Hardware descriptions**
  All gates functioning in parallel at the same time

---

**Software**

**Hardware**

- Warehouse Scale Computer
- Leverage Parallelism & Achieve High Performance

**Computer**

- Core
- …
- Core
- Memory
- Input/Output

**Instruction Unit(s)**

**Functional Unit(s)**

- $A_0 + B_0$
- $A_1 + B_1$
- $A_2 + B_2$
- $A_3 + B_3$

**Cache Memory**

**Logic Gates**
Warehouse Scale Computers

• Massive scale datacenters: 10,000 to 100,000 servers + networks to connect them together
  – Emphasize cost-efficiency
  – Attention to power: distribution and cooling
  – (relatively) homogeneous hardware/software

• **Single gigantic** machine

• Offer very large applications (Internet services): search, voice search (Siri), social networks, video sharing

• Very highly available: < 1 hour down/year
  – Must cope with failures common at scale

• “…WSCs are no less worthy of the expertise of computer systems architects than any other class of machines” (Barroso and Hoelzle, 2009)
Design Goals of a WSC

• Unique to Warehouse-scale
  – Ample parallelism:
    • Batch apps: many independent data sets with independent processing (Data-Level and Request-Level Parallelism)
  – Scale and its Opportunities/Problems
    • Relatively small number of WSC make design cost expensive and difficult to amortize
    • But price breaks are possible from purchases of very large numbers of commodity servers
    • Must also prepare for high component failures
  – Operational Costs Count:
    • Cost of equipment purchases $<<$ cost of ownership
Google’s Oregon WSC
Containers in WSCs

Inside WSC

Inside Container
Equipment Inside a WSC

Server (in rack format):
1 ¾ inches high “1U”, x 19 inches x 16-20 inches: 8 cores, 16 GB DRAM, 4x1 TB disk

7 foot Rack: 40-80 servers + Ethernet local area network (1-10 Gbps) switch in middle (“rack switch”)

Array (aka cluster):
16-32 server racks + larger local area network switch (“array switch”) 10X faster => cost 100X: cost f(N²)
Server, Rack, Array
Coping with Performance in Array

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
Coping with Workload Variation

- Online service: Peak usage 2X off-peak
Impact of latency, bandwidth, failure, varying workload on WSC software?

• WSC Software must take care where it places data within an array to get good performance
  – Latency & bandwidth impact Performance

• WSC Software must cope with failures gracefully
  – High failure rate impact Reliability Availability

• WSC Software must scale up and down gracefully in response to varying demand
  – Varying workloads impact Availability

• More elaborate hierarchy of memories, failure tolerance, workload accommodation makes WSC software development more challenging than software for single computer
Power vs. Server Utilization

- Server power usage as load varies idle to 100%
- Uses \( \frac{1}{2} \) peak power when idle!
- Uses \( \frac{2}{3} \) peak power when 10% utilized! 90%@ 50%!
- Most servers in WSC utilized 10% to 50%
- Goal should be *Energy-Proportionality*: % peak load = % peak energy
Power Usage Effectiveness

- Overall WSC Energy Efficiency: amount of computational work performed divided by the total energy used in the process

- Power Usage Effectiveness (PUE):
  \[
  \frac{\text{Total Building Power}}{\text{IT equipment Power}}
  \]
  - Power efficiency measure for WSC, *not* including efficiency of servers, networking gear
  - Power usage for non-IT equipment increases PUE
  - 1.0 is perfection, higher numbers are worse
  - Google WSC’s PUE: 1.2
PUE in the Wild (2007)

FIGURE 5.1: LBNL survey of the power usage efficiency of 24 datacenters, 2007 (Greenberg et al.)
High PUE: Where Does Power Go?

- Uninterruptable Power Supply (battery)
- Chiller cools warm water from Air Conditioner
- Power Distribution Unit
- IT Equipment 30%
- PDU 5%
- Lighting 1%
- Transformers / Switchgear 1%
- Computer Room Air Conditioner
- CRAC 9%
- Humidifier

Servers + Networking
Google’s Belgium WSC PUE: 1.09

1. Careful air flow handling
   – Don’t mix hot & cold; containers
2. Elevated cold aisle temperatures
   – Server reliability still OK if not too cool
3. Use of free cooling
   – Location climate, large bodies of water
4. Per-server 12-V DC UPS
5. Measured vs. estimated PUE, publish PUE, and improve operation
**Question:** Which statement is TRUE about Warehouse Scale Computers?

(A) Idling electronic equipment consumes almost no power

(B) Reducing lighting costs will *decrease* the Power Usage Effectiveness (PUE)

(C) Qatar (summer temps > 110°F) will one day be a hotbed for WSC housing

(D) Using cheaper components is more expensive because of the higher failure rate
Question: Which statement is TRUE about Warehouse Scale Computers?

(A) Idling electronic equipment consumes almost no power

(B) Reducing lighting costs will decrease the Power Usage Effectiveness (PUE)

(C) Qatar (summer temps > 110°F) will one day be a hotbed for WSC housing

(D) Using cheaper components is more expensive because of the higher failure rate
Agenda

- Warehouse Scale Computers
- Administrivia
- Cloud Computing
- Request Level Parallelism
- MapReduce
Administrivia

• HW6 Due 7/30
• Proj4 Released today at some point
• Guerilla session is now Sunday 2-4pm, @Cory 540AB
• “Lost” discussion will be tomorrow 4-6pm, @Cory 540AB!
Agenda

• Warehouse Scale Computers
• Administrivia
• Cloud Computing
• Request Level Parallelism
• MapReduce
Why Cloud Computing Now?

- Build-out of extremely large datacenters (10,000’s of *commodity* PCs)
  - Build-out driven by growth in demand (more users)
  - Infrastructure software and Operational expertise
- Discovered economy of scale: 5-7x cheaper than provisioning a medium-sized (1000 servers) facility
- More pervasive broadband Internet so can access remote computers efficiently
- Commoditization of HW & SW
- Better tooling for standardizing software
Cloud Distinguished by...

- Shared platform with illusion of isolation
  - Collocation with other tenants
  - Exploits technology of VMs and hypervisors (next lectures!)
  - At best “fair” allocation of resources, but not true isolation

- Attraction of low-cost cycles
  - Economies of scale driving move to consolidation
  - Statistical multiplexing to achieve high utilization/efficiency of resources

- Elastic service
  - Pay for what you need, get more when you need it
  - But no performance guarantees: assumes uncorrelated demand for resources
Cloud Services

• SaaS: deliver apps over Internet, eliminating need to install/run on customer's computers, simplifying maintenance and support
  — E.g., Google Docs, Win Apps in the Cloud
• PaaS: deliver computing “stack” as a service, using cloud infrastructure to implement apps. Deploy apps without cost/complexity of buying and managing underlying layers
  — E.g., Hadoop on EC2, Apache Spark on GCP
• IaaS: Rather than purchasing servers, software, data center space or net equipment, clients buy resources as an outsourced service. Billed on utility basis. Amount of resources consumed/cost reflect level of activity
  — E.g., Amazon Elastic Compute Cloud, Google Compute Platform
### July 2018 AWS Instances & Prices

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>vCPU</th>
<th>ECU</th>
<th>Memory (GiB)</th>
<th>Storage (GB)</th>
<th>Usage Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>t2.nano</td>
<td>1</td>
<td>Variable</td>
<td>0.5</td>
<td>EBS Only</td>
<td>$0.0077 per Hour</td>
</tr>
<tr>
<td>t2.micro</td>
<td>1</td>
<td>Variable</td>
<td>1</td>
<td>EBS Only</td>
<td>$0.015 per Hour</td>
</tr>
<tr>
<td>t2.small</td>
<td>1</td>
<td>Variable</td>
<td>2</td>
<td>EBS Only</td>
<td>$0.031 per Hour</td>
</tr>
<tr>
<td>t2.medium</td>
<td>2</td>
<td>Variable</td>
<td>4</td>
<td>EBS Only</td>
<td>$0.061 per Hour</td>
</tr>
<tr>
<td>t2.large</td>
<td>2</td>
<td>Variable</td>
<td>8</td>
<td>EBS Only</td>
<td>$0.122 per Hour</td>
</tr>
<tr>
<td>m4.large</td>
<td>2</td>
<td>6.5</td>
<td>8</td>
<td>EBS Only</td>
<td>$0.117 per Hour</td>
</tr>
<tr>
<td>m4.xlarge</td>
<td>4</td>
<td>13</td>
<td>16</td>
<td>EBS Only</td>
<td>$0.234 per Hour</td>
</tr>
<tr>
<td>m4.2xlarge</td>
<td>8</td>
<td>26</td>
<td>32</td>
<td>EBS Only</td>
<td>$0.468 per Hour</td>
</tr>
<tr>
<td>m4.4xlarge</td>
<td>16</td>
<td>53.5</td>
<td>64</td>
<td>EBS Only</td>
<td>$0.936 per Hour</td>
</tr>
<tr>
<td>m4.10xlarge</td>
<td>40</td>
<td>124.5</td>
<td>160</td>
<td>EBS Only</td>
<td>$2.43 per Hour</td>
</tr>
<tr>
<td>m3.medium</td>
<td>1</td>
<td>3</td>
<td>3.75</td>
<td>1 x 4 SSD</td>
<td>$0.077 per Hour</td>
</tr>
<tr>
<td>m3.large</td>
<td>2</td>
<td>6.5</td>
<td>7.5</td>
<td>1 x 32 SSD</td>
<td>$0.154 per Hour</td>
</tr>
<tr>
<td>m3.xlarge</td>
<td>4</td>
<td>13</td>
<td>15</td>
<td>2 x 40 SSD</td>
<td>$0.308 per Hour</td>
</tr>
<tr>
<td>m3.2xlarge</td>
<td>8</td>
<td>26</td>
<td>30</td>
<td>2 x 80 SSD</td>
<td>$0.616 per Hour</td>
</tr>
</tbody>
</table>
Agenda

• Warehouse Scale Computers
• Administrivia
• Cloud Computing
• Request Level Parallelism
• MapReduce
Request-Level Parallelism (RLP)

• Hundreds or thousands of requests per sec
  – Not your laptop or cell-phone, but popular Internet services like web search, social networking, …
  – Such requests are largely independent
    • Often involve read-mostly databases
    • Rarely involve strict read–write data sharing or synchronization across requests
• Computation easily partitioned within a request and across different requests
Google Query-Serving Architecture
Anatomy of a Web Search

Search Query: dank memes

- Dank Memes - Reddit
  [Link](https://www.reddit.com/r/dankmemes/)
  (This includes using the Impact font in a non-ironic way.)
  XIII: No posts where the title is the meme caption.
  Dank? · What's going on... · Expectation != reality · enis to [Pagina]

- Dank Memes ♻️arence (@FreeMemesKids) · Twitter
  [Link](https://twitter.com/FreeMemesKids)
  Sliding in to DMs like
  pic.twitter.com/jtXUN0u...
  22 hours ago · Twitter

- Dank Memes Vine Compilation V15 - YouTube
  [Link](https://www.youtube.com/watch?v=dprHL6nm18o)
  Jun 1, 2016 · Uploaded by Emisoccer
  From now on, expect me to upload weekly because I'm going to do extra work on
  my videos in order to not get ...

Images for dank memes
Anatomy of a Web Search (1 of 3)

• Google “dank memes”
  – Direct request to “closest” Google Warehouse Scale Computer
  – Front-end load balancer directs request to one of many arrays (cluster of servers) within WSC
  – 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 contain the search words, “dank”, “memes”, may use location of search as well
  – Return document list with associated relevance score
Anatomy of a Web Search (2 of 3)

• In parallel,
  – Ad system: run ad auction for bidders on search terms
  – Get images of dank memes and trash posts

• 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 of 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 hot spots, e.g. “UCBMFET”
  – Increases opportunities for request-level parallelism
  – Makes the system more tolerant of failures
Agenda

• Warehouse Scale Computers
• Administrivia
• Cloud Computing
• Request Level Parallelism
• MapReduce
Great Idea #4: Parallelism

- **Parallel Requests**
  Assigned to computer e.g. search “Chand”

- **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. add of 4 pairs of words

- **Software**

- **Hardware**
  Warehouse Scale Computer
  Leverage Parallelism & Achieve High Performance

- **Computer**
  Core ... Core
  Memory
  Input/Output

- **Instruction Unit(s)**
- **Functional Unit(s)**
  $A_0 + B_0$, $A_1 + B_1$, $A_2 + B_2$, $A_3 + B_3$

- **Cache Memory**

- **Logic Gates**

- **Smart Phone**

- **Hardware descriptions**
  All gates functioning in parallel at same time
Data Level Parallelism (DLP)

• SIMD
  – Supports data-level parallelism in a single machine
  – Additional instructions & hardware
  – e.g. Matrix multiplication in memory

• DLP on WSC
  – Supports data-level parallelism across multiple machines
  – MapReduce & scalable file systems
  – e.g. Training CNNs with images across multiple disks
MapReduce

• Simple data-parallel programming model and implementation for processing large dataset

• Users specify the computation in terms of
  – a map function, and
  – a reduce function

• Underlying runtime system
  – Automatically parallelize the computation across large scale clusters of machines.
  – Handles machine failure
  – Schedule inter-machine communication to make efficient use of the networks

• Invented at Google
MapReduce Uses

• At Google:
  – Index construction for Google Search
  – Article clustering for Google News
  – Statistical machine translation
  – For computing multi-layers street maps

• At Yahoo!:
  – “Web map” powering Yahoo! Search
  – Spam detection for Yahoo! Mail

• At Facebook:
  – Data mining
  – Ad optimization
  – Spam detection
Example: Facebook Lexicon

www.facebook.com/lexicon (no longer available)
Map & Reduce Functions in Python

- Calculate: \[ \sum_{n=1}^{4} n^2 \]

```python
list = [1, 2, 3, 4]
def square(x):
    return x * x
def sum(x, y):
    return x + y
reduce(sum, map(square, list))
```
MapReduce Programming Model

- **Map**: \((in\_key, in\_value) \rightarrow list(interm\_key, interm\_val)\)

  ```python
  def map(in_key, in_val):
    # DO WORK HERE
    emit(interm_key, interm_val)
  
  - Slice data into “shards” or “splits” and distribute to workers
  - Compute set of intermediate key/value pairs
  ```

- **Reduce**: \((interm\_key, list(interm\_value)) \rightarrow list(out\_value)\)

  ```python
  def reduce(interm_key, list(interm_val)):
    # DO WORK HERE
    emit(out_key, out_val)
  
  - Combines all intermediate values for a particular key
  - Produces a set of merged output values (usually just one)
MapReduce Word Count Example

- **Map** phase: (doc name, doc contents) \( \rightarrow \) list(word, count)
  
  ```
  // “I do I learn”” \( \rightarrow \) [“(I”,1), (“do”,1), (“I”,1), (“learn”,1)]
  ```

  ```python
  map(key, value):
      for each word w in value:
          emit(w, 1)
  ```

- **Reduce** phase: (word, list(count)) \( \rightarrow \) (word, count_sum)
  
  ```
  // (“I”, [1,1]) \( \rightarrow \) (“I”,2)
  ```

  ```python
  reduce(key, values):
      result = 0
      for each v in values:
          result += v
      emit(key, result)
  ```
MapReduce Implementation
MapReduce Execution

(1) Split inputs, start up programs on a cluster of machines
MapReduce Execution

(2) Assign map & reduce tasks to idle workers
(3) Perform a map task, generate intermediate key/value pairs
(4) Write to the buffers
MapReduce Execution

(5) Read intermediate key/value pairs, sort them by its key.
MapReduce Execution

(6) Perform a reduce task for each intermediate key, write the result to the output files
MapReduce Processing Time Line

- Master assigns map + reduce tasks to “worker” servers
- As soon as a map task finishes, worker server can be assigned a new map or reduce task
- Data sort begins as soon as a given Map finishes
- Reduce task begins as soon as all data sort finish
- To tolerate faults, reassign task if a worker server “dies”
Big Data MapReduce Engine: Spark

• Fast and general engine for large-scale data processing.
• Originally developed in the AMPlab at UC Berkeley
• Running on HDFS
• Provides Java, Scala, Python APIs for
  – Database
  – Machine learning
  – Graph algorithms
• MUCH faster and easier to use compared to predecessor Hadoop
// RDD: primary abstraction of a distributed collection of items
file = sc.textFile("hdfs://...")
// Two kinds of operations:
// Actions: RDD → Value
// Transformations: RDD → RDD
// e.g. flatMap, Map, reduceByKey
file.flatMap(lambda line: line.split())
  .map(lambda word: (word, 1))
  .reduceByKey(lambda a, b: a + b)
MapReduce Word Count Example

- **Map** phase: (doc name, doc contents) → list(word, count)
  
  ```
  // “I do I learn”" → [("I",1),("do",1),("I",1),("learn",1)]
  ```

  ```python
  map(key, value):
      for each word w in value:
          emit(w, 1)
  ```

- **Reduce** phase: (word, list(count)) → (word, count_sum)
  
  ```
  // ("I", [1,1]) → ("I",2)
  ```

  ```python
  reduce(key, values):
      result = 0
      for each v in values:
          result += v
      emit(key, result)
  ```
Word Count in Spark’s Python API

// RDD: primary abstraction of a distributed collection of items
file = sc.textFile("hdfs://...")
// Two kinds of operations:
// Actions: RDD \to \text{Value}
// Transformations: RDD \to RDD
// e.g. flatMap, Map, reduceByKey
file.flatMap(lambda line: line.split())
 .map(lambda word: (word, 1))
 .reduceByKey(lambda a, b: a + b)
MapReduce Word Count Example

- **Map** phase: (doc name, doc contents) → list(word, count)
  
  `// “I do I learn”” → [ (“I”,1), (“do”,1), (“I”,1), (“learn”,1) ]`

  ```python
  map(key, value):
      for each word w in value:
          emit(w, 1)
  ```

- **Reduce** phase: (word, list(count)) → (word, count_sum)
  
  `// (“I”, [1,1]) → (“I”,2)`

  ```python
  reduce(key, values):
      result = 0
      for each v in values:
          result += v
      emit(key, result)
  ```
Word Count in Spark’s Python API

// RDD: primary abstraction of a distributed collection of items
file = sc.textFile("hdfs://...")

// Two kinds of operations:
// Actions: RDD → Value
// Transformations: RDD → RDD
// e.g. flatMap, Map, reduceByKey

file.flatMap(lambda line: line.split())
  .map(lambda word: (word, 1))
  .reduceByKey(lambda a, b: a + b)
MapReduce Word Count Example

- **Map** phase: (doc name, doc contents) → list(word, count)
  // “I do I learn” → [ (“I”, 1), (“do”, 1), (“I”, 1), (“learn”, 1) ]
  
  ```python
  map(key, value):
    for each word w in value:
      emit(w, 1)
  ```

- **Reduce** phase: (word, list(count)) → (word, count_sum)
  // (“I”, [1,1]) → (“I”, 2)
  
  ```python
  reduce(key, values):
    result = 0
    for each v in values:
      result += v
    emit(key, result)
  ```
Summary

• Warehouse Scale Computers
  – Supports many of the applications we have come to depend on
  – Software must cope with failure, load variation, and latency/bandwidth limitations
  – Hardware sensitive to cost and energy efficiency

• Request Level Parallelism
  – High request volume, each largely independent
  – Replication for better throughput, availability

• MapReduce
  – Convenient data-level parallelism on large dataset across large number of machines
  – Spark is a framework for executing MapReduce algorithms