COEN-4730/EECE-5730 Computer Architecture

Lecture 10
Warehouse Scale Computing (WSC)

Cris Ababei
Dept. of Electrical and Computer Engineering
MARQUETTE UNIVERSITY
BE THE DIFFERENCE.

Credits: Slides adapted from presentations of Sudeep Pasricha and others: Kubiatowicz, Patterson, Mutlu, Elsevier

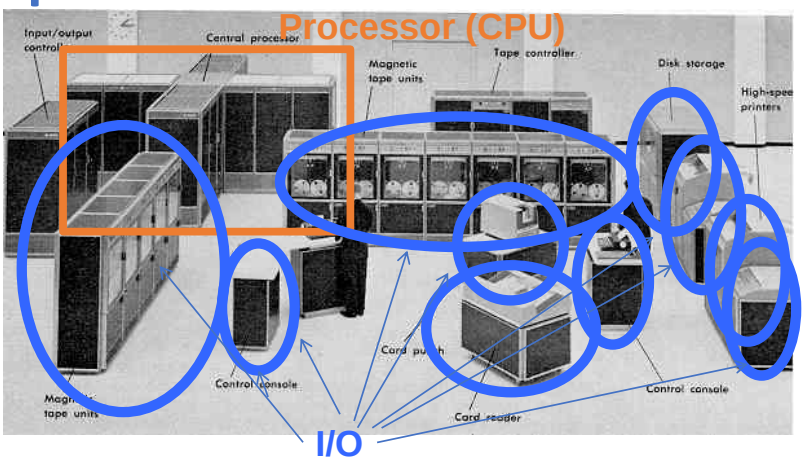
1

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

2

Computer Eras: Mainframe 1950s-60s



"Big Iron": IBM, UNIVAC, … build $1M computers for businesses ➔ COBOL, Fortran, timesharing OS

3

Minicomputer Era: 1970s

Using integrated circuits, Digital, HP… build $10k computers for labs, universities ➔ C, UNIX OS

4

PC Era: Mid 1980s - Mid 2000s

Using microprocessors, Apple, IBM, … build $1k computer for 1 person ➔ Basic, Java, Windows OS

5

Post-PC Era: Late 2000s - Present

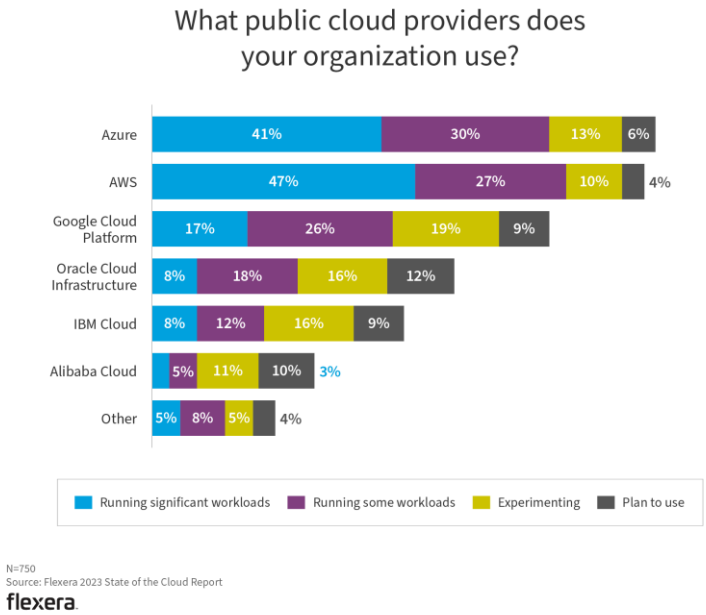
Personal Mobile Devices (PMD): Relying on wireless networking, Apple, Nokia, … build $500 smartphone and tablet computers for individuals

➔ Objective C, Java, Android OS + iOS

Cloud Computing: Using Local Area Networks, Amazon, Google, … build $200M Warehouse Scale Computers with 100,000 servers for Internet Services for PMDs

➔ MapReduce/Spark, Ruby on Rails

6



Source: https://info.flexera.com/CM-REPORT-State-of-the-Cloud

7

Some videos…

- Amazon
  - https://www.youtube.com/watch?v=q6WlzHLxNKI (2021) – beyond 80000 servers, not worth it?
  - https://www.youtube.com/watch?v=cstnNg1_oRo (2023) – retired servers?
- Microsoft
  - https://www.youtube.com/watch?v=80aK2_iwMOs (2021) – half a million servers
  - https://www.youtube.com/watch?v=Rk3nTUfRZmo (2023) – what powers ChatGPT?
- Google data center tour:
  - https://www.youtube.com/watch?v=kNoh9t2oNRg (2023) - Inside YouTube's data center, world's biggest data center
  - https://www.youtube.com/watch?v=9CL3pZfsHbs (2021) - Inside Google's $13 Billion Data Centers
  - https://www.youtube.com/watch?v=kd33UVZhnAA (2023) - Google Data Center Security: 6 Layers Deep
  - http://www.youtube.com/watch?v=zRwPSFpLX8I (2009)
- Facebook
  - https://www.youtube.com/watch?v=_r97qdyQtIk (2015) - Facebook Data Center
  - https://www.youtube.com/watch?v=2l6gI-ksdKs (2019) - What's inside a Facebook Datacenter Open Compute Rack?
- IBM, HP, Dropbox, …

8

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

9

What is a Warehouse-Scale Computer (WSC)?

- A "datacenter" for internet-scale services/cloud computing
  - Examples: Google, Facebook, Yahoo, Amazon web services, Microsoft, Baidu
- Both consumer and enterprise services
  - Windows live, gmail, hotmail, dropbox, bing, google, Webapps, exchange online, salesforce.com, azure platform, Adcenter, GoogleApps, ...

10

Motivation for WSCs

- Some applications need big machines
  - Examples: search, language translation, etc
- User experience
  - Ubiquitous access
  - Ease of management (no backups, no config)
- Vendor benefits (all translate to **lower costs**)
  - Faster application development
    - Tight control of system configuration
    - Ease of (re)deployment for upgrades and fixes
    - Single-system view for storage and other resources
  - Lower cost by sharing HW resources across many users
  - Lower cost by amortizing HW/storage management costs

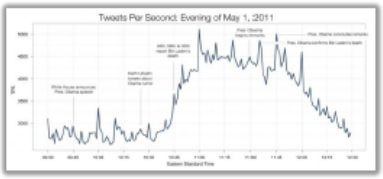
11

WSC design considerations: Scale/Rapid Growth

- More requests, big problem sizes, complex algorithms
  - Microsoft: Windows Live: 500M IDs, Live Hotmail: 355M Active Accounts, Live Messenger: 303M users, Bing: 4B Queries/month, Xbox Live: 25M users, adCenter: 14B Ads served/month, Exchange Hosted Services: 2 to 4B emails/day
  - Zynga Farmville: 1 million players 4 days after launch; 10 million players after 60 days; 75 million players after 270 days (contrast: previous most popular online game: 5 million players)
  - "If facebook were a country, third largest in the world (500million+ users)"

	Aug-04	Mar-06	Sep-07	Sep-09
Number of MapReduce jobs	29,000	171,000	2,217,000	3,467,000
Average completion time (seconds)	634	874	395	475
Server years used	217	2,002	11,081	25,562
Input data read (terabytes)	3,288	52,254	403,152	544,130
Intermediate data (terabytes)	758	6,743	34,774	90,120
Output data written (terabytes)	193	2,970	14,018	57,520
Average number of servers per job	157	268	394	488

Figure 6.2 Annual MapReduce usage at Google over time. Over five years the number of MapReduce jobs increased by a factor of 100 and the average number of servers per job increased by a factor of 3. In the last two years the increases were factors of 1.6 and 1.2, respectively [Dean 2009]. Figure 6.16 on page 385 estimates that running the 2009 workload on Amazon's Cloud Computing Service EC2 would cost $133M.



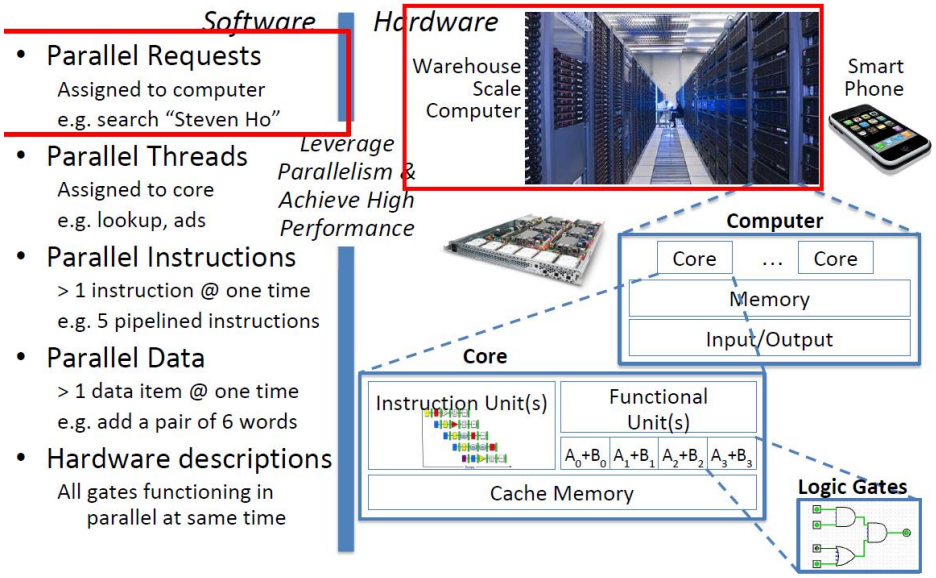
12

WSC design considerations: Request-Level Parallelism (RLP)

- Instruction-level parallelism (ILP)
  - Pipelining, speculation, OOO, …
- Data-level parallelism (DLP)
  - Vectors, GPUs, MMX, …
- Thread-level parallelism (TLP)
  - Multithreading, multi-cores, …
- Request-level parallelism (RLP)
  - Parallelism among multiple decoupled tasks
  - Web servers, "map-reduce", search, email, …
  - Large-scale distributed systems (clusters, NOW, Grids)

13

Parallelism



14

WSC design considerations: Cost, cost, cost…

- High volume needs low costs
  - "multiplier effect" (e.g., mgbility, reliability notes)
- Business models based on low costs
  - How much do you pay for gmail?
  - Key competitive advantage
- Both capital (capex) & operational (opex) costs
  - Power and cooling important component

15

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

16

WSCs ≠ Standard Datacenters

WSCs belong to single organization, relatively homogenous HW and system SW, and common mgmt layer

- Traditional datacenters heterogeneous: number of small and medium applications on dedicated hardware; HPC clusters more special-purpose and batch-centric
- See costs' comparison later

- Large scale: 50K-100K servers; $150M
  - Changes everything!
- Request-level parallelism
  - Internet-scale services, cloud services, large data
- The datacenter is the computer
  - Focus on few key apps with hardware-software codesign
- Enhanced focus on cost efficiency
  - Volume economics important

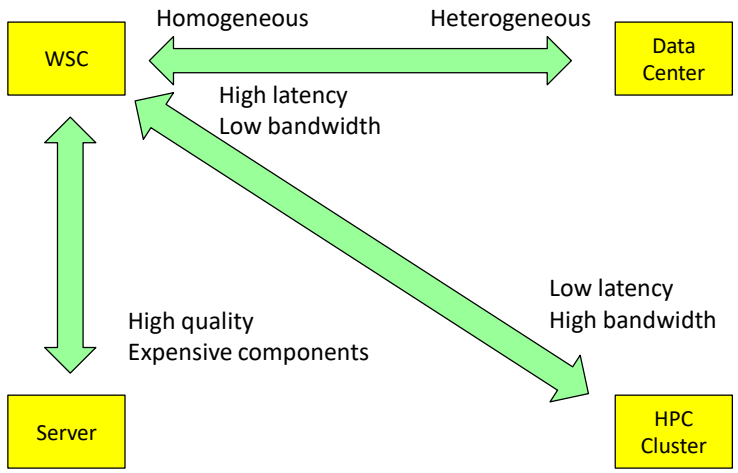
17

WSC vs. Datacenters

- Based on a study in 2006 that compared a WSC with a datacenter with only 1000 servers, WSCs had the following advantages:
  - 5.7X reduction in storage costs—It cost the WSC $4.6 per GByte per year for disk storage versus $26 per GByte for the datacenter.
  - 7.1X reduction in administrative costs—The ratio of servers per administrator was over 1000 for the WSC versus just 140 for the datacenter
  - 7.3X reduction in networking costs—Internet bandwidth cost the WSC $13 per Mbit/sec/month versus $95 for the datacenter
    - Unsurprisingly, you can negotiate a much better price per Mbit/sec if you order 1000 Mbit/sec than if you order 10 Mbit/sec
  - High level of purchasing leads to volume discount prices on the servers and networking gear for WSCs
  - Datacenters have a PUE ~2; WSCs can justify hiring mechanical and power engineers to develop WSCs with lower PUEs ~1.2
  - Datacenter server utilization 10-20%; WSCs around 50% on average as they are open to public

18

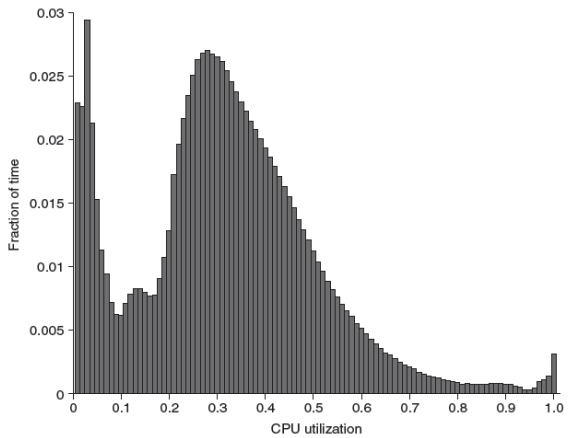
WSC vs Datacenter vs HPC Cluster vs Server



19

Google WSC: Server Load

- Average CPU utilization of more than 5000 servers during a 6-month period at Google



Servers are rarely completely idle or fully utilized, instead operating most of the time at between 10% and 50% of their maximum utilization

20

Summary of WSC Characteristics

- Built from low-cost commodity servers
- Tend to run their own custom software rather than buy third-party commercial software, in part to cope with the huge scale and in part to save money
  - E.g., cost of the Oracle database and Windows operating system doubles the cost of the Dell Poweredge 710 server
  - Google runs Linux operating system on its servers, for which it pays no licensing fees
- Must be able to cope with highly variable load
  - Holidays, weekends, Christmas season bring unique load spikes
- Massive data replication to deal with failure
- Operational costs count in a big way, which alters WSC design
  - Not so much for individual servers

21

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

22

Distributed Programming Models and Workloads for Warehouse-Scale Computers

- WSCs run public-facing Internet services such as search, video sharing, and social networking, as well as *batch applications*, such as converting videos into new formats or creating search indexes from Web crawls
- One of the most popular frameworks for batch processing in a WSC is Map-Reduce and its open-source twin Hadoop
  - E.g. Annual MapReduce usage at Google over time
  - Facebook runs Hadoop on 2000 batch-processing servers of the 60,000 servers it is estimated to have in 2011

	Aug-04	Mar-06	Sep-07	Sep-09
Number of MapReduce jobs	29,000	171,000	2,217,000	3,467,000
Average completion time (seconds)	634	874	395	475
Server years used	217	2002	11,081	25,562
Input data read (terabytes)	3288	52,254	403,152	544,130
Intermediate data (terabytes)	758	6743	34,774	90,120
Output data written (terabytes)	193	2970	14,018	57,520
Average number of servers per job	157	268	394	488

23

MapReduce

- Programming model and runtime for processing large data-sets
  - E.g., Google's search algorithms
  - Gool: make it easy to use 1000s of CPUs and TBs of data
- Inspiration: functional programming languages
  - Programmer specifies only "what"
  - System determines "how"
    - Schedule parallelism, locality, communication,…
- Ingredients
  - Automatic parallelization and distribution
  - Fault-tolerance
  - I/O scheduling
  - Status and monitoring

24

End of the Road for MapReduce?

- Google has abandoned MapReduce, the system for running data analytics jobs spread across many servers.
- It has built a new cloud analytics system: Cloud Dataflow (2014).
- Other approaches: SPARK
  - Spark is a new processing model that facilitates iterative programming and interactive analytics
  - Spark provided in-memory primitive model - loads the data into memory and query it repeatedly; makes Spark well suited for a lot data analytics and machine learning algorithms
  - Note: Spark only defines the distributed processing model. Storing the data part is not addressed by Spark; it still relies on Hadoop (HDFS) to efficiently store the data in a distributed way
  - Spark promises to be 10-100x faster than MapReduce. Many think this could be the end of MapReduce.

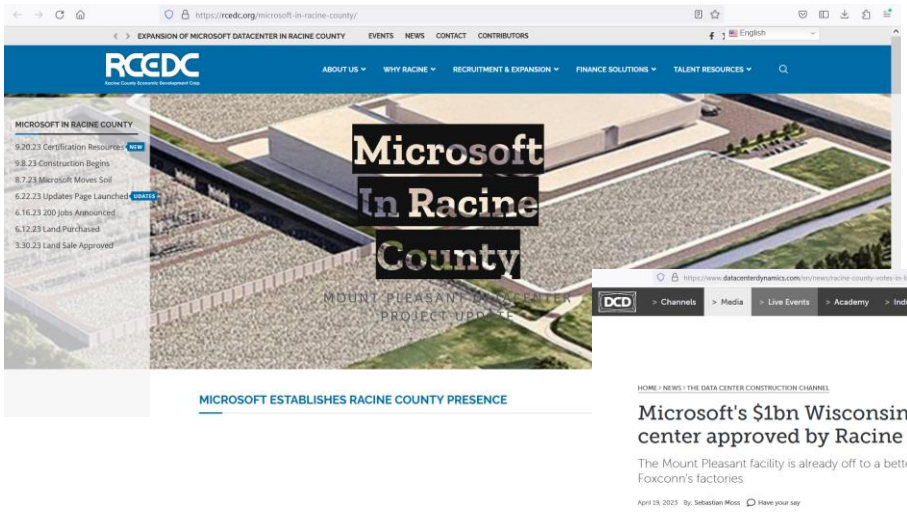
25

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

26

Morning news, WI related…



27

A story….

You work for Facebook, and Mr. Zuckerberg has tasked you with selecting a location for a new warehouse computer in order to serve your users better. Zuckerberg has suggested building a datacenter in California, but Microsoft has just bought the plot of land you wanted, and you must relocate.

You have three choices:

Arizona. The land here is 20% of the cost as land in California, but the hot climate will increase your cooling infrastructure cost by 10%.

Washington. The temperate climate will reduce your power and cooling infrastructure cost by 20%, but your personnel will have to be paid twice as much.

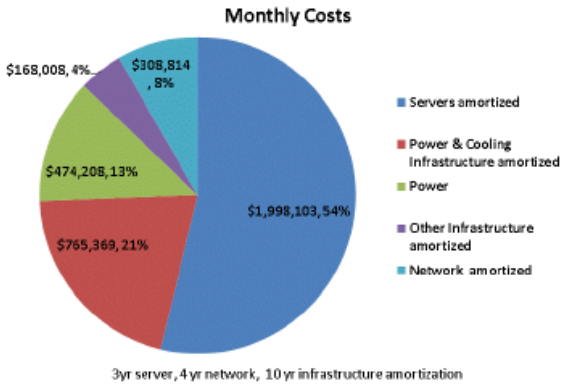
The midwest. The land, personnel, and building costs here are 50% of those of California.

Which location shall you choose?

28

Cost Model

Servers amortized	$1,998,103
Power & Cooling Infrastructure amortized	$765,369
Power	$474,208
Other Infrastructure amortized	$168,008
Network amortized	$308,814
People	$85,410



- Observations
  - 34% costs related to power (trending up while server costs down)
  - Networking high @ 8% of overall costs; 15% of server costs

29

Infrastructure and Costs - Server

- Determining the maximum server capacity
  - Nameplate power rating: maximum power that a server can draw
  - Better approach: measure under various workloads
  - Oversubscribe by 40%
- Typical power usage by component:
  - Processors: 42%
  - DRAM: 12%
  - Disks: 14%
  - Networking: 5%
  - Cooling: 15%
  - Power overhead: 8%
  - Miscellaneous: 4%

30

30

Location of the Warehouse!

- Location
  - Proximity to internet backbone optical fibers
  - Low cost of electricity, property tax rates
  - Low risk from environmental disasters
    - earthquakes, floods, hurricanes
  - Geographic proximity to key user populations

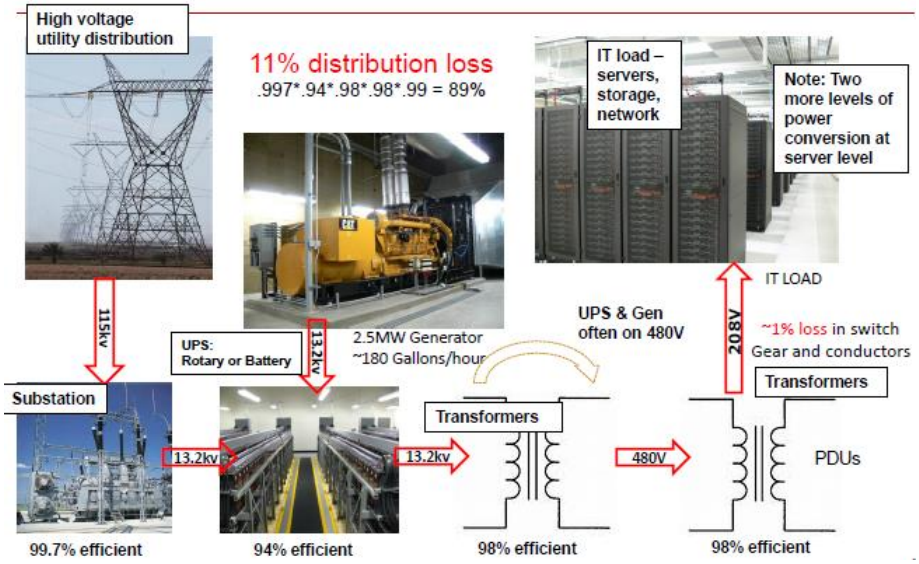
31

Planning the Facility

- Apart from computers & switches, you need:
  - Power infrastructure: voltage converters and regulators, generators and UPSs, …
  - Cooling infrastructure: A/C, cooling towers, heat exchangers, air impellers,…

32

Power Distribution



33

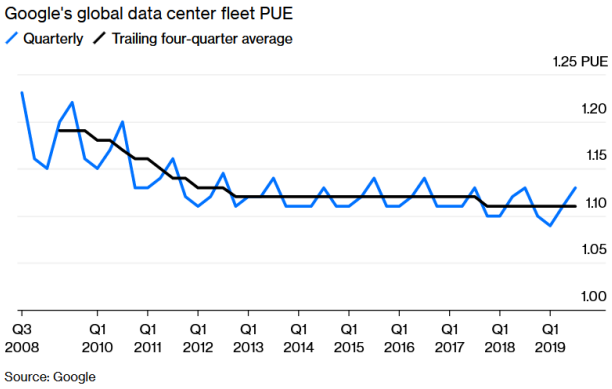
Energy Efficiency

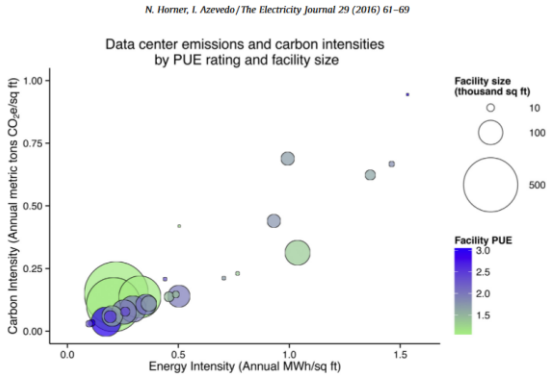
$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \left(\frac{1}{\text{PUE}}\right) \times \left(\frac{1}{\text{SPUE}}\right) \times \left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)$$

- PUE = power usage effectiveness
  - Building power/power of IT (servers, switches etc)
  - Some DCs as bad as PUE = 3
  - Current state of the art PUE = ~1.2
- SPUE = server power usage effectiveness
  - Server power/power for CPUs, DRAM, disk, etc
  - Most servers have SPUE = 1.6
  - State of the art SPUE = 1.2
- If PUE=SPUE=1.2 => 30% of energy is "wasted"

34

Significant Emphasis on Energy Efficiency





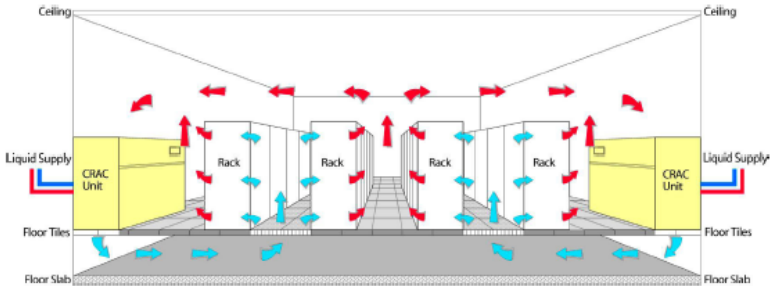
35

Recent Trends: Improving Energy Efficiency in WSC

- Run WSC at higher temperature (71F or 22C)
- Alternating cold/hot aisles, separated by thin plastic sheets
- Leverage colder outside air to cool the water before it is sent to the chillers
- Water-to-water intercooler
  - Google's WSC in Belgium takes cold water from an industrial canal to chill the warm water from inside the WSC
- Airflow simulation to carefully plan cooling

36

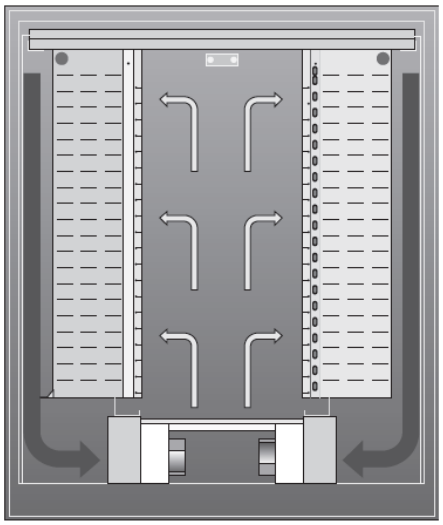
Cooling: Cold/Hot Aisles



- CRAC = computer room air conditioning
  - Cold airs goes through servers and exits in hot aisle
  - Cold aisles ~18-22C, hot aisles ~35C
  - CRAC units consume significant amount of energy!
- 10% reduction in fan speed translates into 27% energy savings
- 20% reduction in fan speed translates into 49% energy savings

37

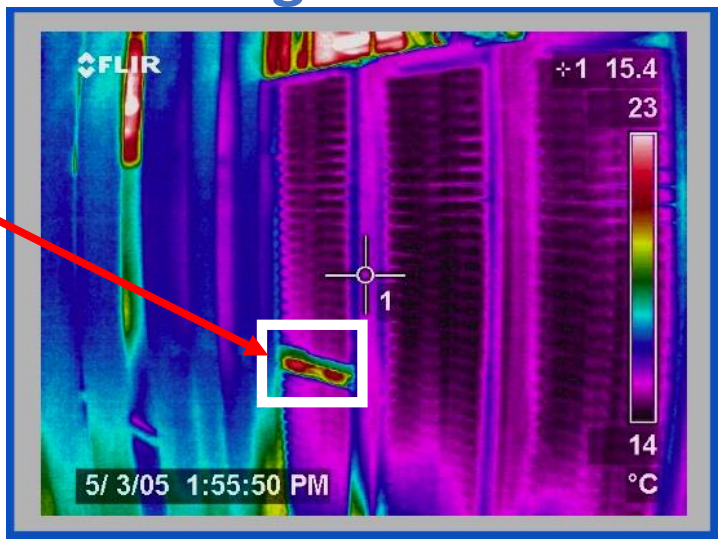
Airflow within Container (for container based WSC)



- Two racks (attached to ceiling) on each side of the container
- Cold air blows into the aisle in the middle of the container (from below) and is then sucked into the servers
  - "cold" air is kept 81°F (27°C)
  - Careful control of airflow allows this high temperature vs. most datacenters
- Warm air returns at the edges of the container
- Design isolates cold and warm airflows

38

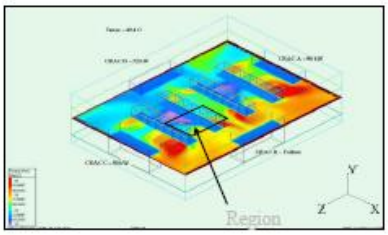
Thermal Image of Cluster Rack

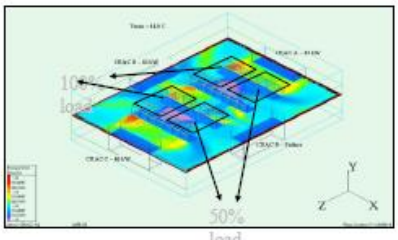


M. K. Patterson, A. Pratt, P. Kumar,
"From UPS to Silicon: an end-to-end evaluation of datacenter efficiency", Intel Corporation

39

Cooling Management: Thermal aware Scheduling





- Example configuration
  - Four homogeneous racks, 75% capacity, one CRAC unit fails
    - Localized hotspots; redline limits breached
- Migrate power from redlined to non-redlined units
  - Region-level power control
    - No hotspots/steep gradients; 15% energy savings

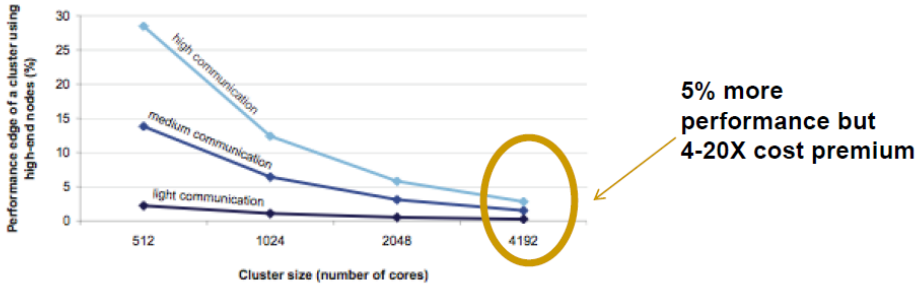
40

WSC Servers

- PC-based servers (e.g., x86)
- 1U or blade form factor are popular
  - E.g., 2 processor sockets, a few GB of DRAM, 2 disks, …
- "Vanity-free"/Custom-designed
- Building block: clusters of low-end servers

41

Performance Advantage over Cluster: Consider Scale



- SMP advantage of 128-core SMP over cluster of 4-core SMPs
  - DC apps are often too large even for largest SMPs (45K servers)
    - 100ns SMP speeds versus 100µs LAN speeds
  - Clusters of commodity (simpler) servers are more cost effective

42

E.g., Facebook circa 2012 OpenCompute.org

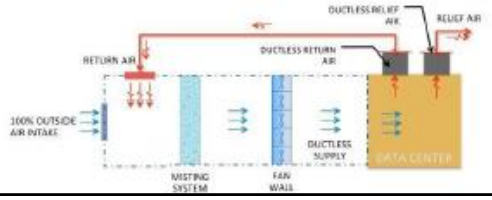
- Server specs
  - 2-socket 2-core 2.2GHz AMD Barcelona processor
  - 8 DIMMs = 8 GB (downclocked from 666MHz to 533MHz)
  - 1-2 SATA disks, 1Gig ethernet NIC
  - Peak power: 160W; idle: 85W
- Only 12V supply to motherboard
  - increased power supply efficiency (92%)
- Battery for "distributed UPS"
  - Enabled by 12V power supply
  - No need for separate battery room
  - battery eff = 99.99% vs UPS eff 94%
  - off-the-shelf UPS for networks
  - BUT space: Size limit to 20 per rack
  - BUT failures, need chargers/monitoring

- Separate storage nodes
  - 10 SATA disks; 300W peak/198Watt idle
  - Storage node takes 2 slots in each rack
  - 2 compute nodes per storage node ~ 5 disks per server (but varies per WSC)

43

- Power and cooling
  - Single transformer rather than four transformer setup
    - No 480VAC to 208VAC but non-commodity PSUs
  - Semi-distributed UPS system
    - No centralized UPS, but single 48V DC UPS integrated with 277 volt AC server power supply
  - Outside air cooling
    - Outside air most of the time
    - Alternatively, use evaporative system
      - MS: Use DX unit, GOOG: use SW
    - Full building ducting with huge plenum areas, no-process-based cooling, mist-based evaporative cooling, large impellers with efficient frequency drive, full-wall low-resistance filtration

- IT systems
  - Custom server "vanity free"
    - 1.5RU, larger fans
    - Intel and AMD motherboards
      - 12V-only designs,
  - Custom "triplet" rack
    - Three columns of thirty servers



44

Servers

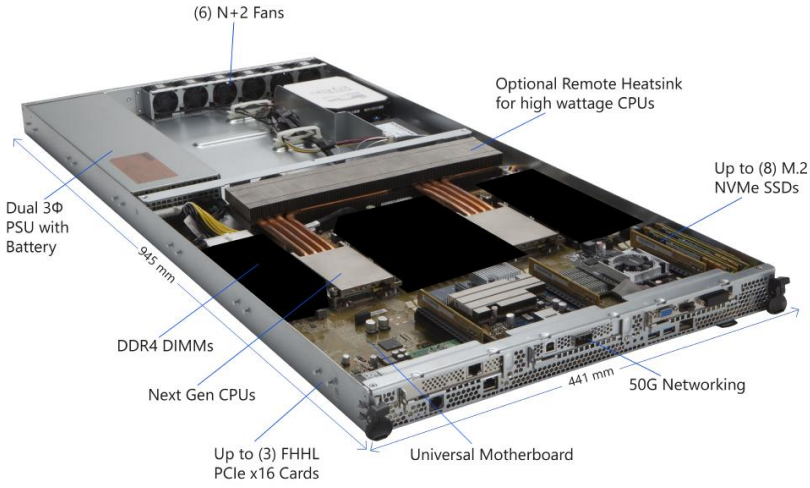
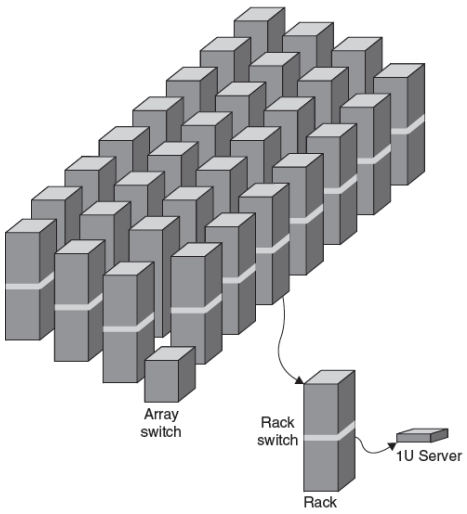


image: Open Compute Project (proposed 2016)

45

WSC Networking

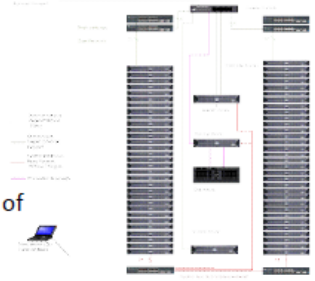
- Connecting 5000+ servers challenging
- Hierarchy
  - Rack switch, array (cluster) switch, L3 switch, border routers
- Switches typically offer 2-8 uplinks, which leave the rack to go to the next higher switch in the hierarchy
  - Bandwidth leaving rack is 6-24x smaller—48/8 to 48/2—than the bandwidth within the rack
  - Ratio is called oversubscription; large oversubscription impacts performance significantly

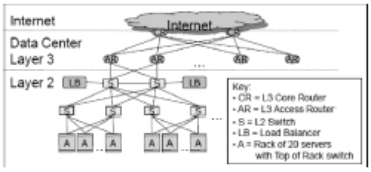


46

Example: Google

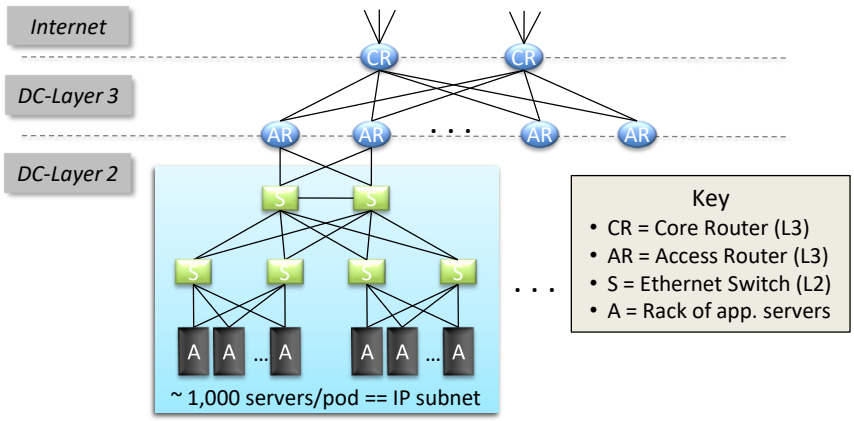
- Rack switch = 48-port ethernet 1Gig switch
  - Commodity switch >= $30 per port
    - Infiniband ~= $500/port
  - One Switch per two racks
  - 40 server ports; 2-8 uplink ports
    - Oversubscription ratio
    - Programmer burden
  - Bandwidth within rack is same irrespective of sender/receiver
- Array switch
  - More expensive: 10X more BW = 100X more $
  - High-end switches feature-rich (mgmt, inspection, CAMs, FPGAs
  - 480 1Gbit links, few 10Gbit ports to datacenter routers
  - Manage oversubscription carefully
- Layer 3 switches, border routers





47

Layer 3 network used to link arrays together and to the Internet



48

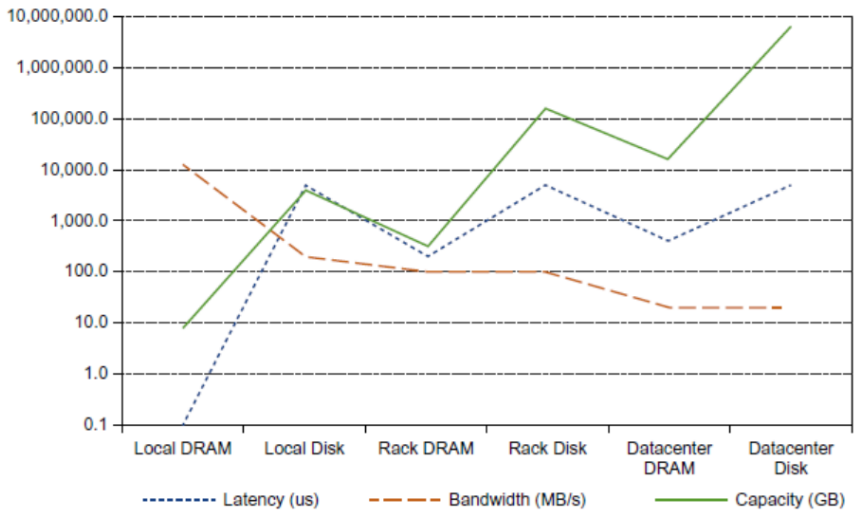
Memory Hierarchy of a WSC

- Servers can access DRAM and disks on other servers using a NUMA-style interface

	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1040	31,200
Disk capacity (GB)	2000	160,000	4,800,000

49

Memory Hierarchy of a WSC



50

50

WSC Storage

- Storage
  - Distributed FS using disks on servers
    - Better fault-tolerance across nodes, lower cost, better scalability
  - Network attached storage (NAS) devices
    - Specialized systems with disk arrays that provide FS storage services and connect directly to the networking fabric
    - Better fault tolerance within device (e.g., RAID), easier management
    - More expensive
- Google circa 2007: Google file system (GFS)
  - Use local disks; local access patterns
  - At least three replicas for disk reliability
    - replicas used for several failure modes
  - Eventual consistency for lower cost

51

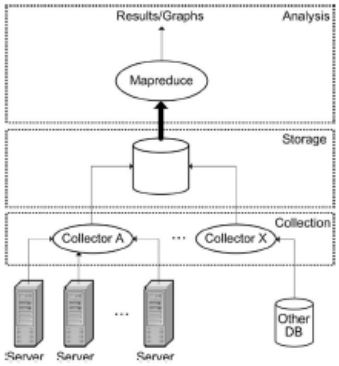
Reliability & Availability

- Common goal for services: 99.99% availability
  - 1 hour of down-time per year
- Graceful degradation under faults
  - E.g., search quality results with loss of systems
  - E.g., delay in access to email
  - E.g., corruption of data of web
  - Internet itself has only two nines of availability
- More appropriate availability metrics
  - Yield = fraction of requests satisfied by service/total number of requests made by users

52

Manageability

- Monitoring and repair in Google WSC
  - 1 operator per 1000 servers
  - Google System Health Monitoring software to track health of all servers/network
  - Diagnostics and simple automated solutions for failures; Reboot/Reimage/Replace; Batched failure processing
  - Minimal human-in-the-loop
    - Goal< 1% of all nodes in manual repair queue
    - MTTR: 1 week; depends on state
  - Power management
    - DVFS not used
      - Feasible only in very low activity modes; System-wide savings small, complexity of management control loop too high
    - Emphasis on energy proportionality and resource scheduling
    - PowerCapping



53

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

54

WSC Applications

- 3-tier applications
  - E.g., web-mail, maps, webdocs, social networks, …
- Search
  - Similar to 3rd tier but more latency critical
  - E.g., most data are kept in memory
- Analytics
  - E.g., create Netflix recommendations, search index, …
  - They are typically off-line (not customer facing)
- Virtualized computations
  - E.g. Amazon EC2 or Microsoft Azure

55

Example: Amazon Web Services (AWS)

- Amazon started offering utility computing via the Amazon Simple Storage Service (Amazon S3) and then Amazon Elastic Computer Cloud (Amazon EC2) in 2006 with several innovations:
  - Virtual Machines: WSC used x86-commodity computers running the Linux operating system and Xen virtual machine
  - Very low cost: When AWS announced a rate of $0.10 per hour per instance in 2006, it was a startlingly low amount. An instance is one VM on a 1.0 to 1.2 GHz AMD Opteron or Intel Xeon of that era
  - (Initial) reliance on open source software: recently, AWS started offering instances including commercial third-party software at higher prices
  - No (initial) guarantee of service: Amazon originally promised only best effort. The low cost was so attractive that many could live without a service guarantee. Today, AWS provides availability SLAs of up to 99.95% on services such as Amazon EC2 and Amazon S3
  - No contract required. In part because the costs are so low, all that is necessary to start using EC2 is a credit card

56

Why Virtual Machines?

- Allowed Amazon to protect users from each other
- Simplified software distribution within a WSC
  - only need install an image and then AWS will automatically distribute it to all the instances being used
- Ability to kill a VM reliably makes it easy for Amazon and customers to control resource usage
- VMs can limit rate at which they use the physical processors, disks, network as well as main memory,
  - gave AWS multiple price points: the lowest price option by packing multiple virtual cores on a single server, the highest price option of exclusive access to all the machine resources, as well as several intermediary points
- VMs hide the identity of older hardware
  - allowing AWS to continue to sell time on older machines that might otherwise be unattractive to customers if they knew their age.

57

Cloud Computing with AWS

- Low cost and a pay-for-use model of utility computing
- Cloud computing provider (Amazon) take on the risks of over-provisioning or under-provisioning
  - Very attractive for startups who want to minimize risks
- E.g.. Farmville from Zynga
  - had 1 million players 4 days after launch; 10 million players after 60 days after 270 days, it had 28 million daily players and 75 million monthly players
  - Because they were deployed on AWS, they were able to grow seamlessly with the number of users
- E.g. Netflix
  - migrated its Web site and streaming video service from a conventional datacenter to AWS in 2011
  - Streaming to mobile devices, computers, HDTVs requires batch processing on AWS to convert new movies to the myriad formats
  - Accounts for ~30% of all Internet traffic today; powered by AWS

58

July 2018 AWS Instances & Prices

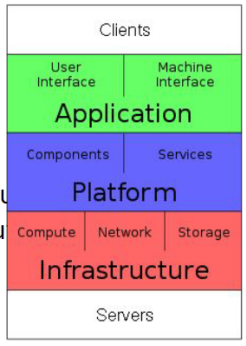
	vCPU	ECU	Memory (GiB)	Storage (GB)	Usage Cost
t2.nano	1	Variable	0.5	EBS Only	$0.0077 per Hour
t2.micro	1	Variable	1	EBS Only	$0.015 per Hour
t2.small	1	Variable	2	EBS Only	$0.031 per Hour
t2.medium	2	Variable	4	EBS Only	$0.061 per Hour
t2.large	2	Variable	8	EBS Only	$0.122 per Hour
m4.large	2	6.5	8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

- For latest see: https://aws.amazon.com/ec2/pricing/on-demand/

59

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 clou infrastructure to implement apps. Deploy apps withou 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



Cloud Computing Stack

60

What Else is Running in a WSC?

- Platform-level software
  - Firmware, operating system, key libraries
- Cluster-level infrastructure
  - Distributed file system, cluster schedulers, distributed programming models (e.g., MapReduce), software for monitoring and deployment management (e.g., Autopilot)…
- Application services
  - Easier to develop given cluster-level infrastructure…

61

Outline

- Part 1: Warehouse Scale Computing (WSC)
  - Introduction
  - Warehouse Scale Computing
  - WSC vs. Datacenters
  - Programming Models (MapReduce)
  - Building a WSC; Considerations
  - Applications
  - Containers
  - Summary
- Part 2: Exascale Computing

62

Containers in WSCs

Inside WSC

Inside Container

63

Server, Rack, Array

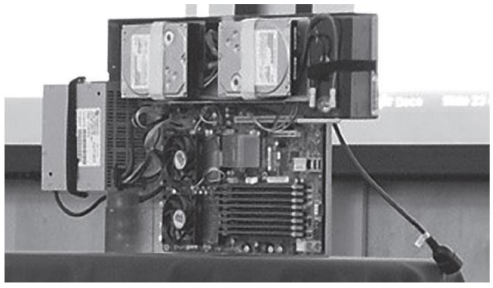
64

Google's Container Based WSC

- Both Google and Microsoft have built WSCs using shipping containers
- Each container is independent
  - Only external connections are: networking, power, and water
  - The containers in turn supply networking, power, and cooling to the servers placed inside them
- Google mini WSC (Oregon): 45 40-foot-long containers (standard 1AAA container 40 x 8 x 9.5 feet) in a 300-foot by 250-foot space, or 75,000 square feet
- To fit in the warehouse, 30 of the containers are stacked two high, or 15 pairs of stacked containers
- WSC offers 10 megawatts with a PUE of 1.23

65

Google WSC: Server



- power supply is on the left and two disks are on top
- two fans below the left disk cover two sockets of the AMD Barcelona processor, each with two cores, running at 2.2 GHz
- eight DIMMs in the lower right each hold 1 GB, giving a total of 8 GB
- single network interface card (NIC) for a 1 Gbit/sec Ethernet link
- peak power of baseline is about 160 watts, idle power is 85 watts
- Alternative to baseline compute node: Storage node
  - 12 SATA disks, 2 Ethernet NICs
  - Peak power is about 300 watts, and it idles at 198 watts
  - takes up two slots in the rack
  - ratio was about two compute nodes for every storage node (but different Google WSCs have different ratios)

66

WSC Summary

- WSC not the same as traditional datacenters
  - Scale, datacenter-is-computer, RLP, costs
- Total cost of ownership
  - Capex-amortized (facility, P&C, server, networking), Opex (P&C, people, bandwidth)
- Architecture and key building blocks
  - Warehouse, container, computer, storage, network, power delivery design, cooling design
  - Balance: HW/SW co-design
- Reliability, availability and serviceability (RAS), Energy
  - Scale changes everything: non-traditional models for RAS, more aggressive energy efficiency focus
- WSC Software
  - Cluster-level infrastructure software
    - Resource management (e.g., cluster scheduler), Hardware abstraction and other basic services (e.g., GFS), programming frameworks (e.g., MapReduce/Spark)
  - Deployment and maintenance
    - Service-level dashboards, performance debugging tools, platform-level monitoring (Google Health Infrastructure)
  - Application-level software

67

Fallacies and Pitfalls

- Cloud computing providers are losing money
  - AWS has a margin of 25%, Amazon retail 3%
- Focusing on average performance instead of 99th percentile performance
- Using too wimpy a processor when trying to improve WSC cost-performance
- Inconsistent Measure of PUE by different companies
- Capital costs of the WSC facility are higher than for the servers that it houses

68

68

Fallacies and Pitfalls

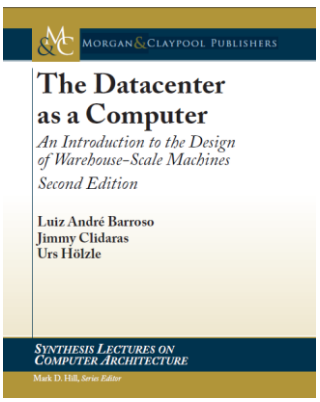
- Trying to save power with inactive low power modes versus active low power modes
- Given improvements in DRAM dependability and the fault tolerance of WSC systems software, there is no need to spend extra for ECC memory in a WSC
- Coping effectively with microsecond (e.g. Flash and 100 GbE) delays as opposed to nansecond or millisecond delays
- Turning off hardware during periods of low activity improves the cost-performance of a WSC

69

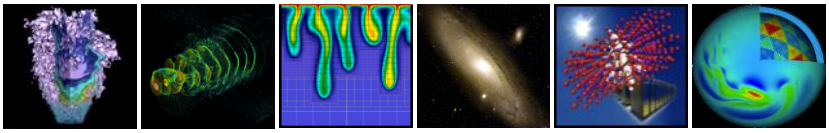
69

Resources

- Textbook
- Luiz André Barroso, Jimmy Clidaras, Urs Hölzle, The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines, Second Edition
  - http://web.eecs.umich.edu/~mosharaf/Readings/DC-Computer.pdf
- Open Compute Project (OCP)
  - A collaborative community focused on re-designing hardware technology to efficiently support the growing demands on compute infrastructure.
  - https://www.opencompute.org/



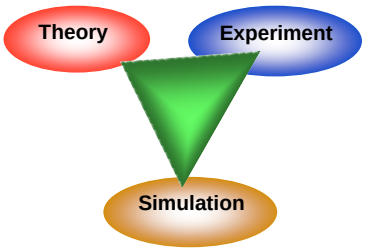
70



Exascale Computing: The Future of Supercomputing

71

SIMULATION: Third Pillar of Science

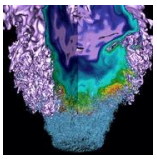


- Traditional scientific and engineering paradigm
  1) Do Theory or paper design
  2) Perform Experiments or build systems
- Limitations:
  - Too difficult (build large wind tunnels)
  - Too expensive (build throw-away passenger jet)
  - Too slow (wait for climate or galactic evolution)
  - Too dangerous (weapons, drug design, climate experimentation)
- Computational science paradigm:
  3) Use high performance computing systems to Simulate the phenomenon
     - Based on known physical laws and numerical methods

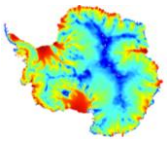
72

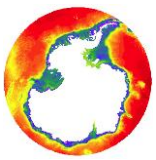
Science at Scale

- Combustion simulations improve future designs
  - Model fluid flow, burning and chemistry
  - Uses advanced math algorithms
  - Requires petascale systems today
  - Need exascale (10^18 operations per second) computing to design for alternative fuels, new devices
- Impacts of Climate Change
  - Warming ocean and Antarctic ice sheet key to sea level rise
    - Previous climate models inadequate
  - Adaptive Mesh Refinement (AMR) to resolve ice-ocean interface
    - Dynamics very fine resolution (AMR)
    - Antarctica still very large (scalability)
  - Exascale machines needed to improve detail in models, including ice and clouds



Simulations reveal features not visible in lab experiments

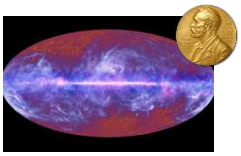


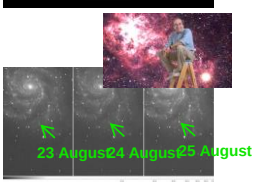


73

Science in Data

- From Simulation to Image Analysis
  - Computing on Data - key in 4 of 10 Breakthroughs of the decade
    - 3 Genomics problems (better DNA, microbe, ancestry analysis) + CMB (cosmic microwave background; to understand origin of universe)
  - Data rates from experimental devices will require exascale volume computing
- Image Analysis in Astronomy
  - Data Analysis in 2006 Nobel Prize
    - Measurement of temperature patterns
  - Simulations used in 2011 Prize
    - Discovery of the accelerating expansion of the universe through observations of distant supernovae
  - More recently: astrophysics discover early nearby supernova.
    - Rare glimpse of a supernova within hours of explosion, 20M light years away
    - Telescopes world-wide redirected to catch images

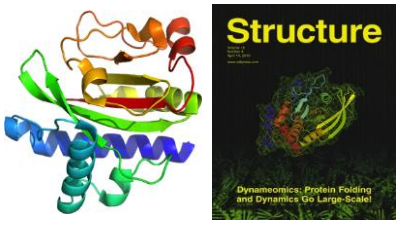


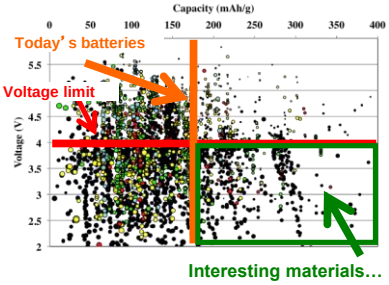


74

Science through Volume: Screening Drugs to Batteries

- Large number of simulations covering a variety of related materials, chemicals, proteins,…





Dynameomics Database

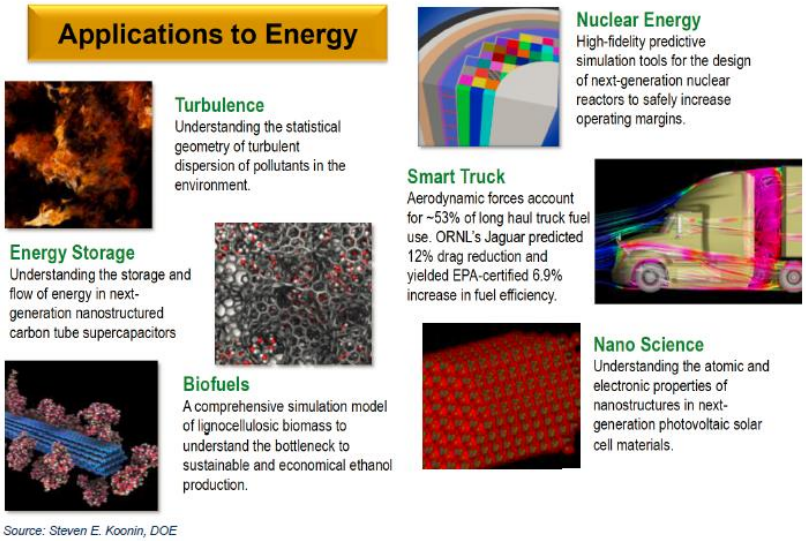
Improve understanding of disease and drug design, e.g., 11,000 protein unfolding simulations stored in a public database.

Materials Genome

Cut in half the 18 years from design to manufacturing, e.g., 20,000 potential battery materials stored in a database

75

Many Other Domains Need Exascale



76

Summary: Need for Supercomputing

- Strategic importance for supercomputing
  - Essential for scientific research
  - Critical for national security
  - Fundamental contributor to the economy and competitiveness through use in engineering and manufacturing
- Supercomputers are the tools for solving the most challenging problems through SIMULATIONS

77

Top500 List – June 2021: http://www.top500.org

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM** DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, **IBM / NVIDIA / Mellanox** DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, **NRCPC** National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, **HPE** DOE/SC/LBNL/NERSC United States	706,304	64.59	89.79	2,528

78

Top500 List – June 2022: http://www.top500.org

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942
4	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
5	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438

79

79

Top500 List – June 2023: http://www.top500.org

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096

80

80

Challenges of Exascale Computing

- 10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket?
- Energy costs?
  - At ~$1M per MW, energy costs are substantial
  - 1 petaflop in 2010 used 3 MW
  - 1 exaflop in 2018 possible in 200 MW with "usual" scaling
- 3D packaging?
- Large-scale optics/photonics based interconnects?
- 10-100 PB of aggregate memory?
- Hardware and software-based fault management?
- Heterogeneous cores?
- Performance per watt?
- Power, area and capital costs will be significantly higher?

81

Resources

- Exascale Computing Project
  - https://www.exascaleproject.org/about/
- The Energy Exascale Earth System Model (E3SM) Project
  - https://e3sm.org/about/
- Modeling and Simulation at the Exascale for Energy and the Environment
  - https://science.osti.gov/-/media/ascr/pdf/program-documents/docs/Townhall.pdf
- CrossCut Report – Exascale Requirements Reviews
  - https://science.osti.gov/-/media/ascr/pdf/programdocuments/docs/2018/DOE-ExascaleReport-CrossCut.pdf
- NVIDIA at Supercomputing
  - https://www.nvidia.com/en-us/events/supercomputing/?ncid=pa-srch-goog-32907&gclid=EAIaIQobChMI4pyVp_-E7QIVTfDACh0n2QIeEAAYAiAAEgIzyPD_BwE#cid=hpc06_pa-srch-goog_en-us
- The Convergence of AI and HPC
  - https://www.intel.com/content/www/us/en/high-performance-computing/supercomputing/exascale-computing.html
- …

82