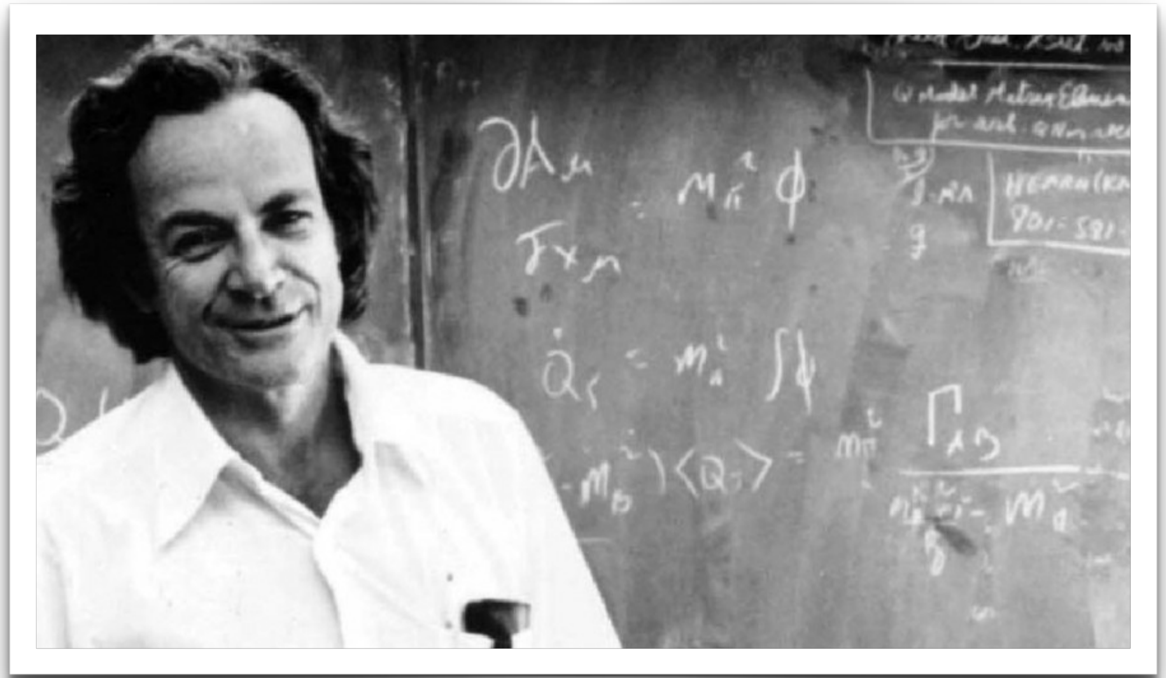
CS3640

Research (3): Cloud Computing

Prof. Supreeth Shastri
Computer Science
The University of Iowa

Reflections on our semester



❝ Students don't need a perfect teacher. Students need a happy teacher, who's going to make them excited to come to school and grow a love for learning ❞

– Richard Feynman

I am here to help YOU succeed (this course & beyond)

Technical Foundation

equip you with the what, why, how of the Internet, so you can create your own network systems

Recommendations

for your job application, grad school application, TA application etc.

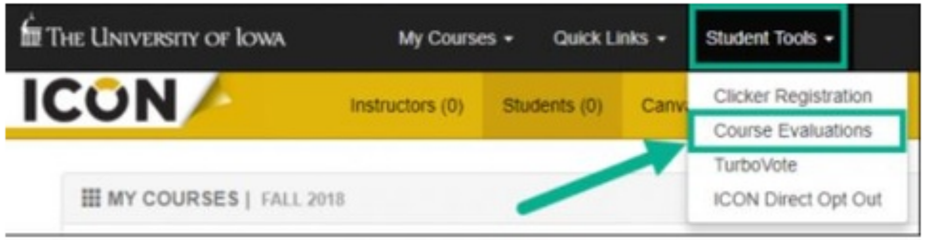
Mentorship/Introductions

if you are interested in working with me; or seeking collaborations w/ other researchers or practitioners

Now, it is your turn!

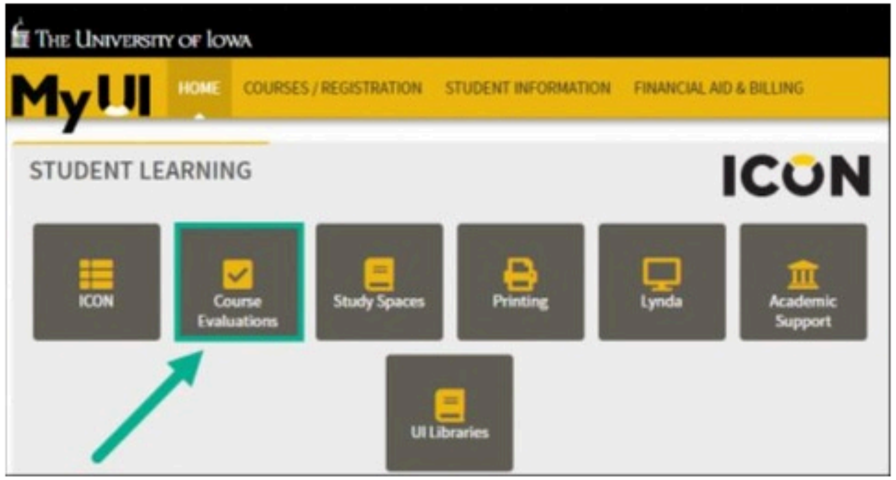
Access ACE Online from ICON:

1. In a browser (Chrome or Firefox preferred), go to icon.uiowa.
2. Drop down "Student Tools."
3. Click on "Course Evaluations."
4. Enter your Hawk ID and password.



Access ACE Online from MyUI:

1. In a browser (Chrome or Firefox preferred), go to myui.uiowa.edu.
2. Click on the "Course Evaluations" button.
3. Enter your Hawk ID and password.



Finals: Format and Logistics

Category	Example questions and topics	Weight
Networking **Principles**	*Internet's hour-glass model; Middleboxes*	25%
Networking **Protocols**	*A day in the life of a packet*	25%
Networking **Problems**	*Construct Dijkstra's LS table*	25%
Networking **Practice**	*SDN; Cloud computing; Solar superstorms*	25%

May 2

Written 12:30 - 1:30pm (in-class); alternate slots for SDS

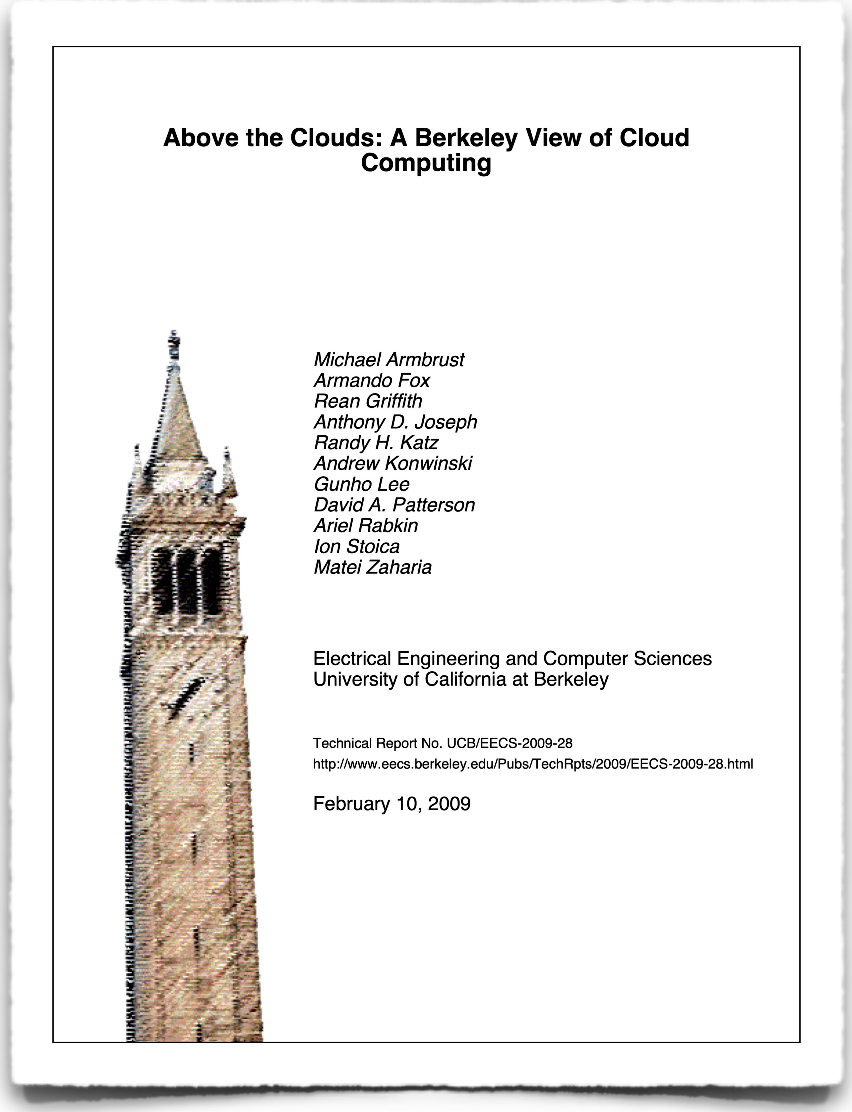
May 1 - 5

Interview location: MLH 201J
Arrive 5min prior to your slot

Lecture goals

A brief overview of cloud computing

- *What, why, and how of the cloud*
- *Datacenters (and networking within)*
- *Challenges and opportunities*



Above the Clouds: A Berkeley View of Cloud Computing

Michael Armbrust
Armando Fox
Rean Griffith
Anthony D. Joseph
Randy H. Katz
Andrew Konwinski
Gunho Lee
David A. Patterson
Ariel Rabkin
Ion Stoica
Matei Zaharia

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-28
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

February 10, 2009

*Cloud computing refers to both **the computing resources delivered as services over the Internet** and **the hardware and software systems in datacenters** that provide those services*

(circa 2006)

Elastic Cloud Compute (EC2)

Standard Instances	Linux/UNIX	Windows
Small (Default)	$0.10 per hour	$0.125 per hour
Large	$0.40 per hour	$0.50 per hour
Extra Large	$0.80 per hour	$1.00 per hour

- Small Instance (Default) 1.7 GB of memory, 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), 160 GB of instance storage, 32-bit platform
- Large Instance 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of instance storage, 64-bit platform
- Extra Large Instance 15 GB of memory, 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), 1690 GB of instance storage, 64-bit platform

Simple Storage Service (S3)

Storage

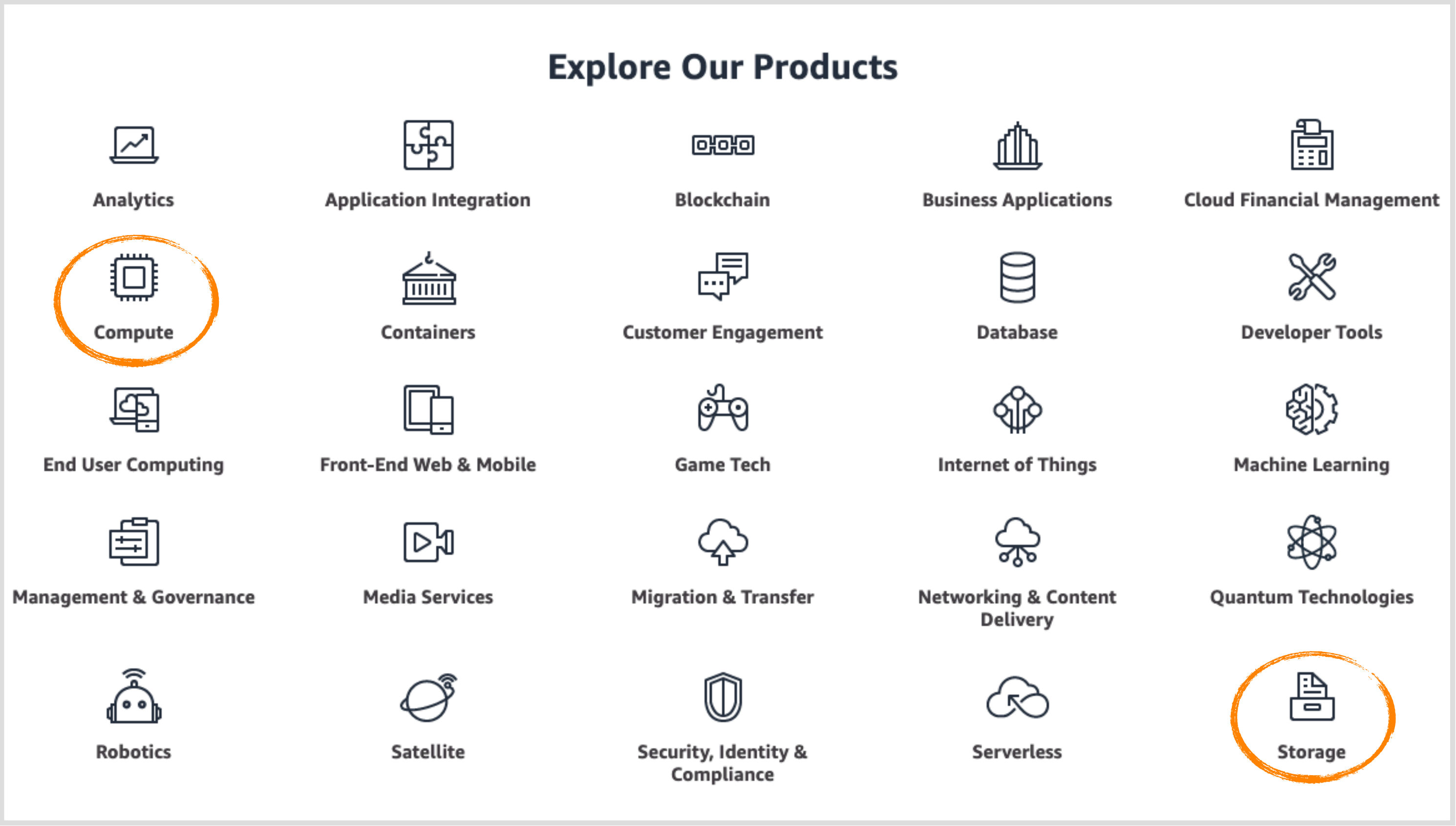
- $0.150 per GB – first 50 TB / month of storage used
- $0.140 per GB – next 50 TB / month of storage used
- $0.130 per GB – next 400 TB /month of storage used
- $0.120 per GB – storage used / month over 500 TB

Data Transfer

- $0.100 per GB – all data transfer in

- $0.170 per GB – first 10 TB / month data transfer out
- $0.130 per GB – next 40 TB / month data transfer out
- $0.110 per GB – next 100 TB / month data transfer out
- $0.100 per GB – data transfer out / month over 150 TB

aws (circa 2023)



A collection of 200+ services

Three models of computing:
Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS)

Four different access methods:
Web-based management console; Command line tools; Software development kits, or RESTful APIs

Characterizing Amazon’s public cloud

On-demand
users can procure resources if/ when needed; no need for making commitments a priori

Scalability
offers an illusion of infinite scalability; allows users to scale their resources up/down in real-time

Billing model
allows users to trade capital expense for operating expense; fine-grained billing proportional to the time and size of resources used

Strong guarantees
services come with high levels of availability and reliability (three to four nines)

Ease of administration
hardware (and low-level system software) are virtualized, so users don’t have to maintain any infrastructure

Global deployment
users have the ability to select the geographical regions in which their data/compute will reside

Why now? Key enablers of cloud computing

Inevitable rise of distributed systems/infrastructure

- ➡ In early 2000, companies realized that vertical scaling of servers has hit a ceiling (i.e., one cannot buy powerful enough servers to keep up with increasing load)
- ➡ This resulted in an increased focus on horizontal scaling a.k.a. building a distributed infrastructure using 1000s of commodity servers
- ➡ Companies such as Google and Amazon developed expertise in building and operating massive datacenters; designed software systems to make them work reliably

Q: who is Google cloud's first customer?

- ➡ Google operates eight global-scale applications each with billion+ user base
- ➡ Google applications run with average uptime of 99.99%

Why now? Key enablers of cloud computing

Advances in resource virtualization

- ➡ First commercial virtualization (IBM Mainframes in **1970**)
- ➡ First virtualization of x86 architecture (VMware ESX in **1998**)
- ➡ Significant reduction in virtualization overheads (Linux containers in **2008**)

Increasing broadband speeds

- ➡ Faster and better quality access to the Internet over the last two decades; average broadband speed in the US ~100 Mbps

Emergence of applications that benefited from the cloud model

- ➡ Large-scale data analytics. For e.g., election campaigns
- ➡ Interactive mobile applications. For e.g., Pokemon Go
- ➡ Video streaming. For e.g., YouTube

❝ One reason you should not use cloud is that you lose control. You're putty in the hands of whoever developed that software. ❞

– Richard Stallman

In The Guardian (11/29/2008)

Datacenters

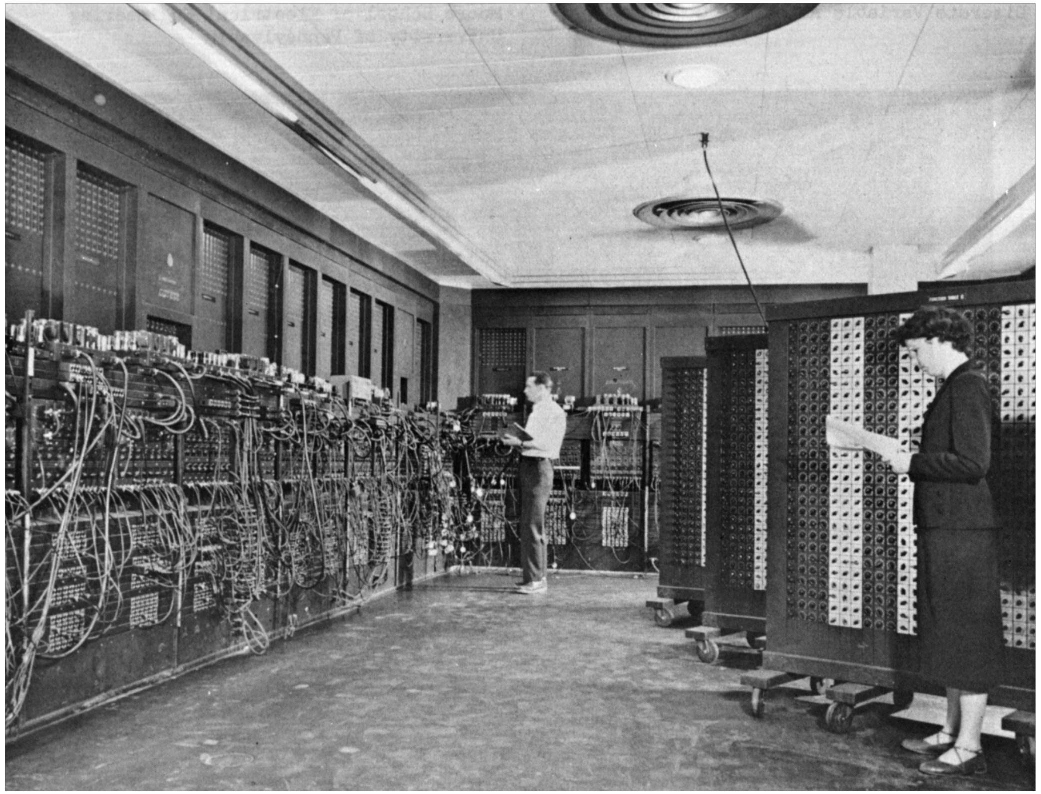
Datacenter

A building that houses computer systems, storage, and telecommunication equipment

Why a dedicated building?

- Complex, heavy, and elaborate machinery
- Component connectivity needed special accommodations
- Consumed significant amount of power and needed cooling
- Expensive; thus, needed security and isolation

Q: Is this a datacenter?



ENIAC (1946) at UPenn and Army Research Lab

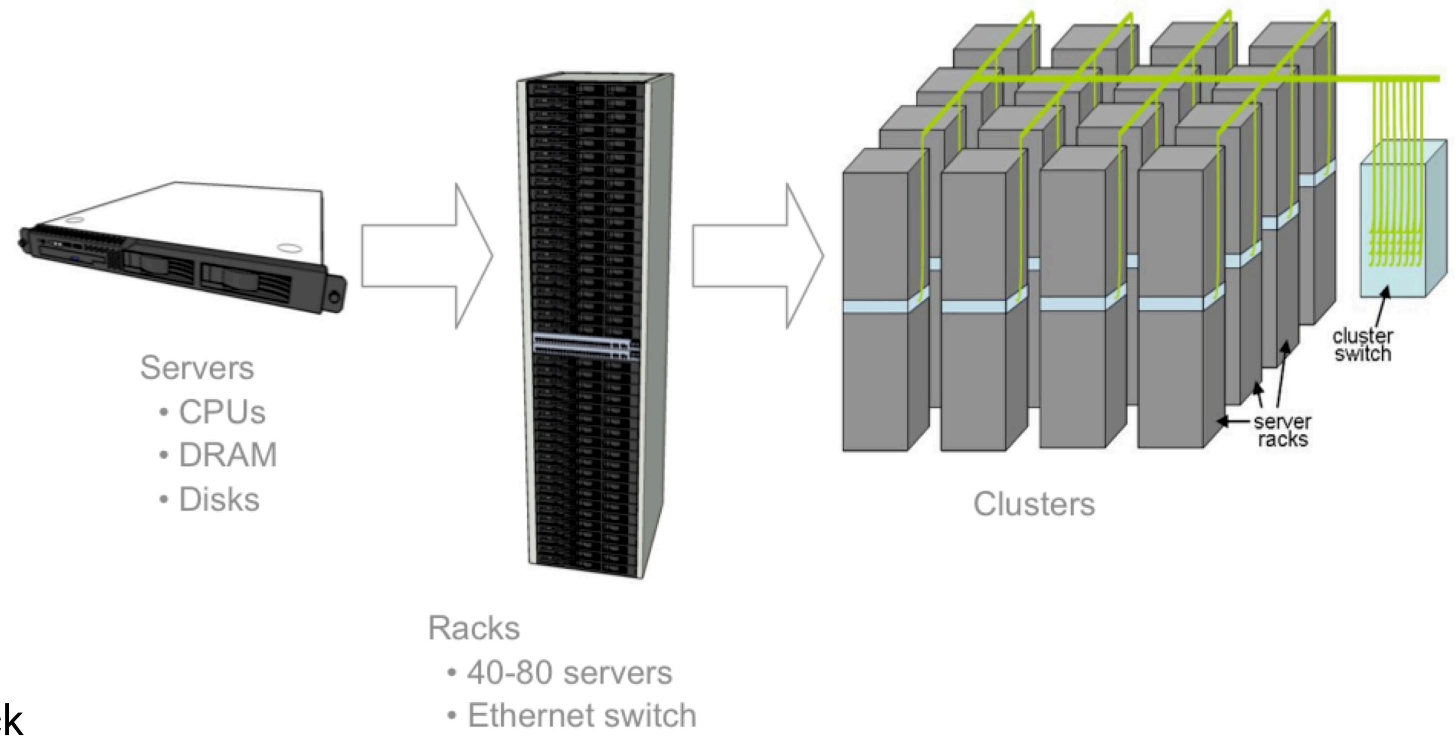
Modern Datacenters

A building that houses computer systems, storage, and telecommunication equipment

- ➡ Built with commodity components
- ➡ O(10K - 100K) servers; O(100K) hard disks
- ➡ High-bandwidth commodity networking
  (1—100Gbps Ethernet switches)
- ➡ Dedicated power generators
- ➡ Heavily secured and guarded

E.g., Google's datacenter in Amsterdam, Netherlands (CapEx: $1.1B)

Hardware Organization



Hierarchical structure

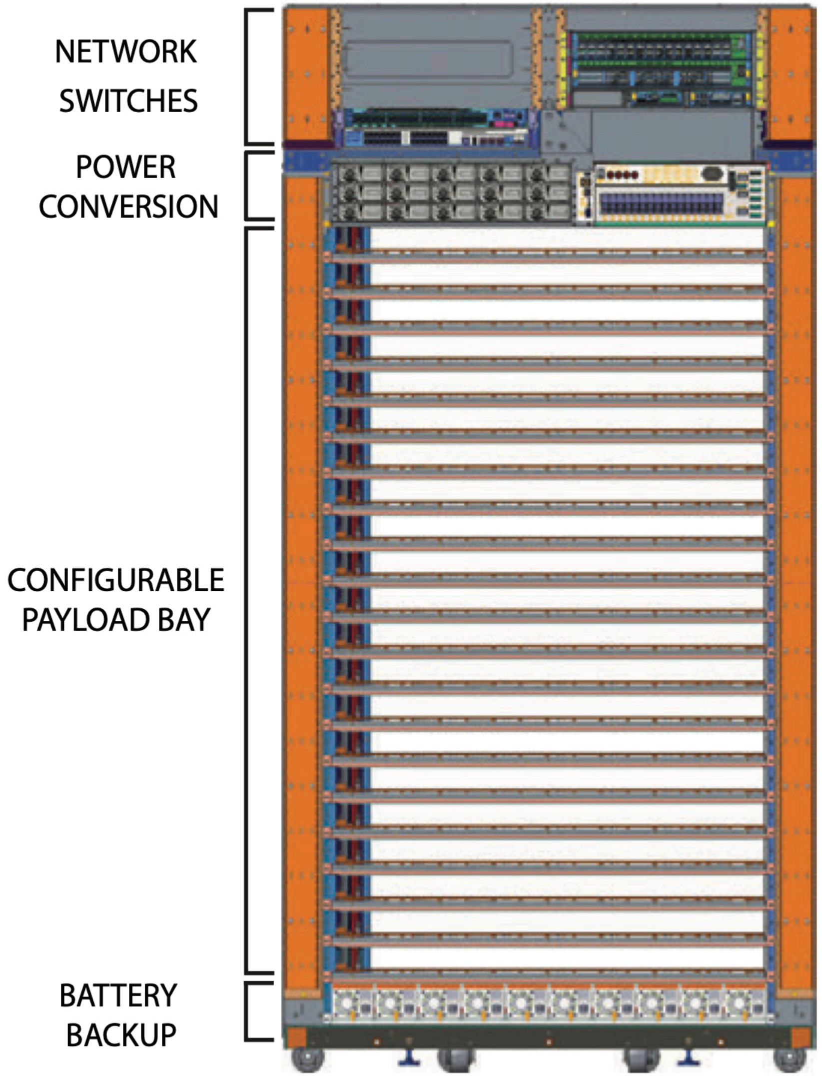
- ➡ low-end, commodity components
- ➡ Blade servers mounted within a rack
- ➡ Racks organized as clusters
- ➡ Ethernet switches (lower capacity at rack level, and denser interconnects at cluster level)

How do these schematics look in the real datacenters?

Hardware Organization

Racks up close!

- ➡ Network switches and power management on the top
- ➡ Compute and storage blades in the middle (10-40 Rack Units)
- ➡ Made of reinforced metal; open in the front and back; wheels for ease of movement



Hardware Organization

Compute + Networking

Server racks have 2-4 switches to which servers connect using different colored cables.

Fiber optic cables (running in yellow cable trays near the ceiling) provide connectivity to the Internet and other Google datacenters.

Datacenter networking

Difficulty of horizontal scaling

- Compute and storage scale well horizontally i.e., *if you need extra aggregate capacity, simply add more boxes*
- In networking, simply increasing the bandwidth of the leaf node is not enough (why?)
- **Solution**: increase the bisection bandwidth of the network i.e., *bandwidth across the narrowest line that divides the cluster into two parts*
- **Challenge**: we cannot increase bisection bandwidth by simply making or buying arbitrarily large switches and routers (limitations in physics and manufacturing)
- **Solution**: build novel network topologies; for e.g., fat-tree, or CLOS

Why bisection bandwidth?
Assuming uniformly random communication between servers, we expect every node to send about half its bits across the "middle" of the network

Hardware Organization

Putting it all together

Datacenter in Council Bluffs, Iowa provides 115,000 sq. feet of rack space.

It supports services including Search and YouTube

Spot Quiz (ICON)