CS 425 / ECE 428 Distributed Systems Fall 2026

Aishwarya Ganesan

w/ Ram Kesavan

(Thanks to Indranil Gupta)

Lecture 2: Introduction to Cloud Computing

For in-class activities:

- https://app.acadly.com/auth/signup/student
- Enter Illinois email address
- Use code: **ZHXVAN**
- Signup now if you haven't already!

How Big is the World’s Largest Datacenter?

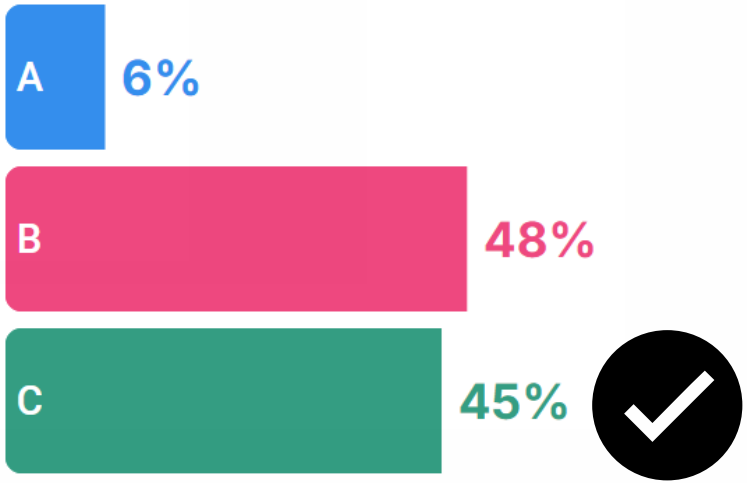
How big is the world's largest datacenter campus
(by reported floor area)?

A. ~500,000 sq. ft.

B. ~1 million sq. ft.

C. ~10 million sq. ft.

TOTAL RESPONSES: 238



World's Largest Datacenter?

- (2018-2025) China Telecom. 10.7 Million sq. ft.
- (2017) "The Citadel" Nevada. 7.2 Million sq. ft.
- (2015) In Chicago!
  - 350 East Cermak, Chicago, 1.1 MILLION sq. ft.
  - Shared by many different "carriers"
  - Critical to Chicago Mercantile Exchange
- See:
  - https://datacentremagazine.com/top10/top-10-biggest-data-centres

(There was and is) The Hype!

- Forrester in 2010 – Cloud computing will go from $40.7 billion in 2010 to $241 billion in 2020.
- Today [1]
  - Market size (2025): USD 943.6 Billion
  - Estimated (2026): USD 1.1 Trillion
  - CAGR (2026-2033): 16.0%
  - Projected (2033): USD 3.34 Trillion
- Companies and even Federal/state governments use cloud

[1] Source: Grand View Research

Many Cloud Providers

- AWS: Amazon Web Services (28%)
- Microsoft Azure (20%)
- Google Cloud (15%)
- Alibaba, Oracle, IBM, Tencent Cloud etc.
- And many many more!
- Edge computing
- Neo clouds

Two Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- Private clouds are accessible only to company employees
- Public clouds provide service to any paying customer:
  - Amazon S3 (Simple Storage Service): store arbitrary datasets, pay per GB-month stored
    - Recently: S3 Standard ≈ **2.3¢ per GB-month**; cheaper storage tiers available
  - Amazon EC2 (Elastic Compute Cloud): upload and run arbitrary OS images
    - from inexpensive CPU VMs to large **GPU/AI instances**;
    - pay per hour, with on-demand, spot, and committed-use options

How wide is the range of cloud compute prices?

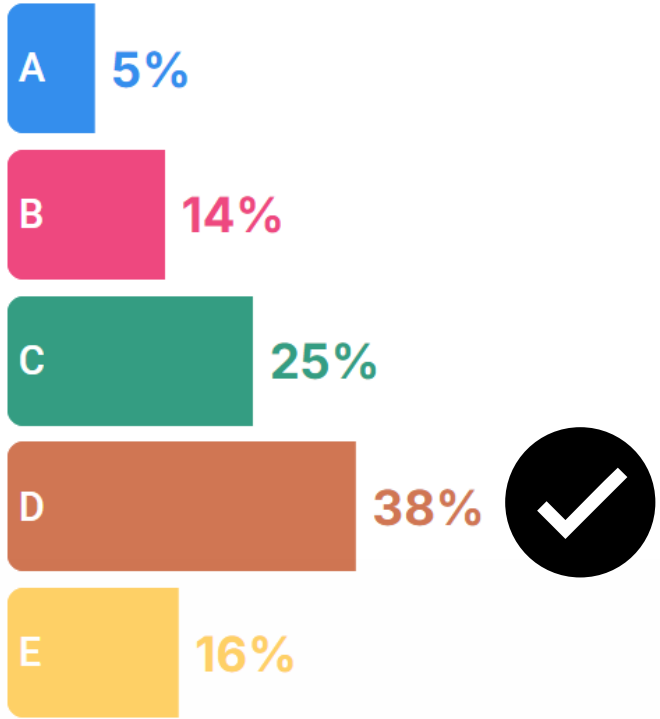
How much more expensive is AWS's **most expensive** EC2 compute option vs. its **cheapest**?

i.e., Let X= (Cost of most expensive compute per hour / cost of cheapest compute per hour)

What is X?

A. 150×
B. 1,500×
C. 15,000×
D. 150,000×
E. 1,500,000×

TOTAL RESPONSES: 254



Bar graph

Two Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- Private clouds are accessible only to company employees
- Public clouds provide service to any paying customer:
  - Amazon S3 (Simple Storage Service): store arbitrary datasets, pay per GB-month stored
  - Amazon EC2 (Elastic Compute Cloud): upload and run arbitrary OS images
    - **$0.005/hour** for a t3.nano (2 vCPU, 0.5 GiB RAM) –

      $761.904/hour for an ultra server (u-p6e) with 72 B200 GPUs.

  - Other clouds: similar pricing ranges as above

Customers Save Time and $$$

- (Stories from mid-2010s)
- Dave Power, Associate Information Consultant at Eli Lilly and Company: “With AWS, Powers said, a new server can be up and running in **three minutes** (it used to take Eli Lilly **seven and a half weeks** to deploy a server internally) and a **64-node Linux cluster** can be online in five minutes (compared with three months internally). … It's just shy of instantaneous.”

- Ingo Elfering, Vice President of Information Technology Strategy, GlaxoSmithKline: “With Online Services, we are able to reduce our IT **operational costs** by roughly **30%** of what we’re spending”

- Jim Swartz, CIO, Sybase: “At Sybase, a private cloud of virtual servers inside its datacenter has **saved nearly $US2 million annually** since 2006, Swartz says, because the company can share computing power and storage resources across servers.”

- 100s of startups in Silicon Valley can harness large computing resources without buying their own machines.

But what exactly IS a cloud?

What is a Cloud?

- It's a cluster!
- It's a supercomputer!
- It's a datastore!

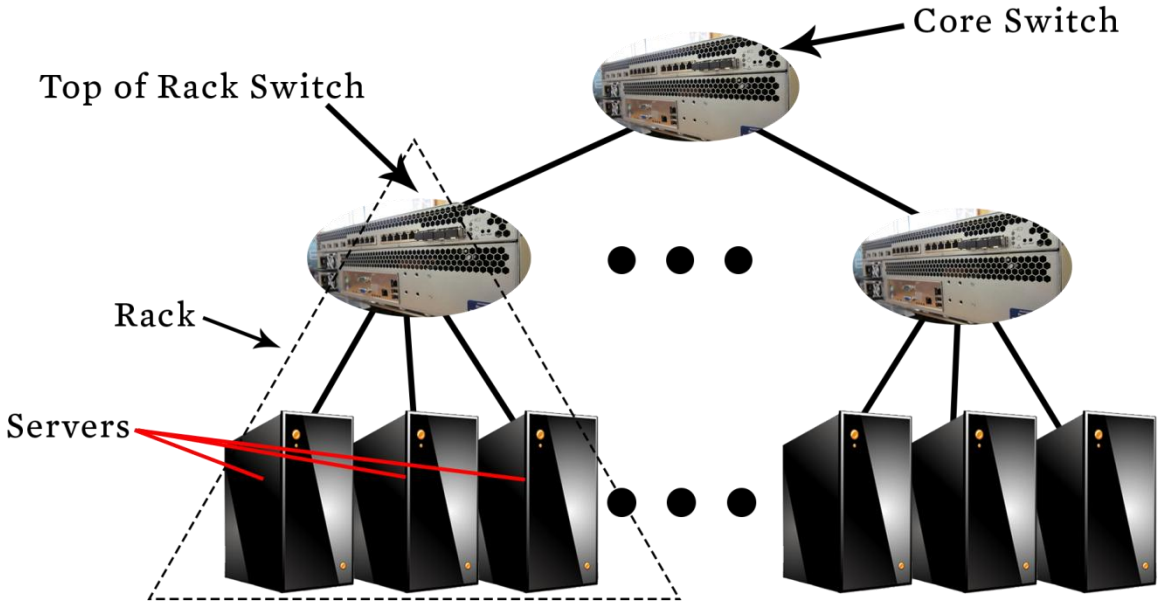
- All of the above

- Cloud = Lots of storage + compute cycles nearby

What is a Cloud?

- A single-site cloud (aka “Datacenter”) consists of
  - Compute nodes (grouped into racks)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical

A Sample Cloud Topology



Clos/fat-tree topology

What is a Cloud?

- A single-site cloud (aka “Datacenter”) consists of
  - Compute nodes (grouped into racks) (2)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical
  - Storage (backend) nodes connected to the network (3)
  - Front-end/load balancers for submitting jobs and receiving client requests (1)
  - (1,2,3: Often called “three-tier architecture”)
  - Cloud control plane — provisioning, scheduling, monitoring, authentication, etc.
  - Software Services

What is a Cloud? (2)

- A geographically distributed cloud consists of
  - Multiple such sites
  - Datacenters typically grouped into Availability Zones (AZs) (e.g., us-east-1a, us-east-1b)
    - AZs designed as separate failure domains
  - Multiple AZs form a regions (e.g., us-east-1)
  - Multiple such regions (e.g., us-east, us-west)
    - Connected by high-capacity WAN/backbone networks

Server → Rack → Datacenter → AZ → Region → Global

"A Cloudy History of Time"

e first datacenters!

ENIAC, ORDVAC, ILLIAC

Many used vacuum tubes and mechanical relays

1940

1950

1960

1970

1980

1990

2000

2012

IBM

xerox

Honeywell

Timesharing Companies

Data Processing Industry

Clouds and datacenters

Clusters

Grids

Data Processing Industry
- 1968: $70 M. 1978: $3.15 Billion

PCs
(not distributed!)

Peer to peer systems

Timesharing Industry (1975):
Honeywell 6000 & 635, IBM 370/168,
Xerox 940 & Sigma 9, DEC PDP-10, UNIVAC 1108

Trends: Technology

- Historical doubling Periods – storage: 12 mos, bandwidth: 9 mos, and cpu compute capacity: 18 mos
  - These trends did not continue indefinitely
  - E.g., 2005 is the end of Dennard/frequency scaling
- Then and Now
  - Bandwidth
    - 1985: mostly 56Kbps links nationwide
    - Today: Tbps backbone capacity widespread
  - Disk capacity
    - Today's PCs have TBs, far more than a 1990 supercomputer

Trends: Users

- Then and Now

  Biologists:

  - 1990: were running small single-molecule simulations
  - Today: CERN's Large Hadron Collider producing many PB/year

Four Features New in Today's Clouds

I. Massive scale.

II. On-demand access: Pay-as-you-go, no upfront commitment.
- And anyone can access it

III. Data-intensive Nature: What was MBs has now become TBs, PBs and XBs.
- Daily logs, forensics, Web data, etc.
- Humans have data numbness: Wikipedia (large) compressed is only about 10 GB!

IV. New Cloud Programming Paradigms: MapReduce/Hadoop, NoSQL/Cassandra/MongoDB and many others.
- High in accessibility and ease of programmability

Combination of one or more of these gives rise to novel and unsolved distributed computing problems in cloud computing.

Announcements

- **4 cr only: MP Groups Form DUE** this week **Mon Aug 31st** (see course webpage).
  - **Hard deadline, as Engr-IT will create and assign VMs the day after!**
  - **(Form does NOT automatically "enroll" you in 4cr section—you must do that yourself)**
- DO NOT
  - Change MP groups unless your partner has dropped
  - Leave your MP partner hanging: Both MP partners should contribute equally (we will ask!)

- **All: Please fill out Student Survey (see course webpage).**

- MP1 due Sep 13th
  - VMs being distributed soon (watch Piazza)
  - Demos will be Monday Sep 14th (schedule and details will be posted soon on Piazza)
- HW1 out today due Sep 20th: Solve problems right after lecture covers topic!

- **Check Piazza often! It's where all the announcements are at!**

I. Massive Scale

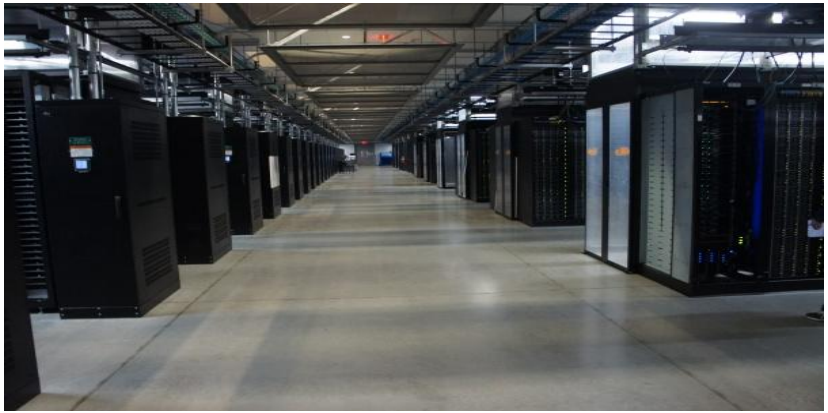
- Facebook [GigaOm, 2012]
  - 30K in 2009 -> 60K in 2010 -> 180K in 2012
- Microsoft [NYTimes, 2008]
  - 150K machines
- AWS EC2 [Randy Bias, 2009]
  - 40K machines
  - 8 cores/machine
- Google [2011, Data Center Knowledge] : 900K (reputed in 2016: 2.5M)

Today: exact server count not publicly disclosed

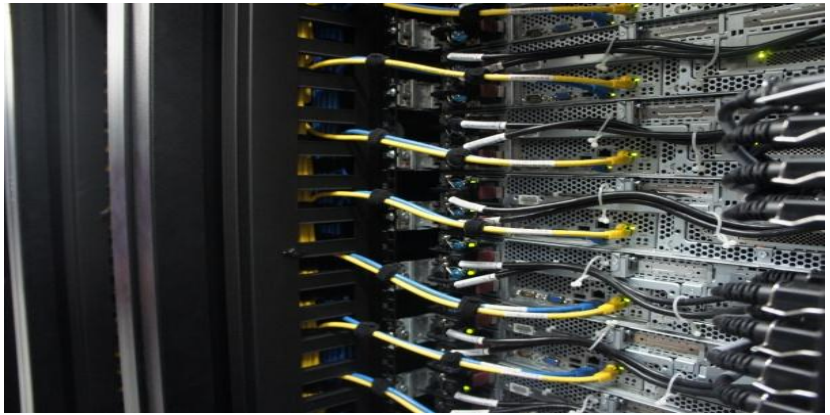
Estimated to have millions of machines globally

Meta: 125K-H100 GPUs

Servers

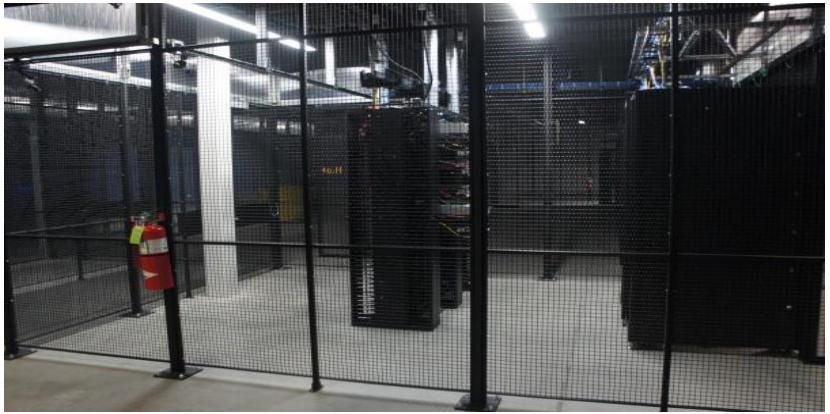


Front



Back

In



Some highly secure (e.g., financial info)

Cooling

Air sucked in from top (also, Bugzappers)

Water purified

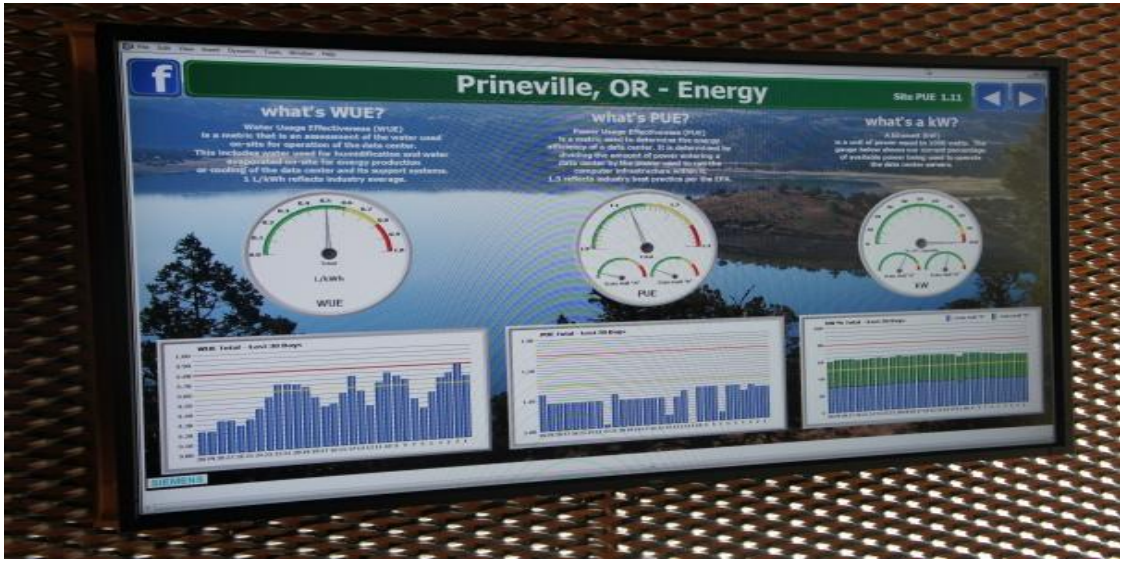
Water sprayed into air

15 motors per server bank

Power

Off-site

On-site



- WUE = Annual Water Usage / IT Equipment Energy (L/kWh) – low is good
- PUE = Total facility Power / IT Equipment Power – low is good (e.g., Google~1.1)

What does a datacenter look like from inside?

- A virtual walk through a datacenter (Facebook Prineville Datacenter, Mid 2010s)
  - https://www.youtube.com/watch?v=4A_A-CmrqpQ

- Timelapse of a Datacenter Construction on the Inside (Fortune 500 company)
  - http://www.youtube.com/watch?v=ujO-xNvXj3g

II. On-demand access: *aaS Classification

On-demand: renting a cab vs. (previously) renting a car, or buying one. E.g.:

- AWS Elastic Compute Cloud (EC2): a few cents to a few $ per CPU hour
- AWS Simple Storage Service (S3): a few cents per GB-month

- HaaS: Hardware as a Service
  - Barebones hardware machines
  - Not always a good idea because of security risks

- IaaS: Infrastructure as a Service
  - Flexible computing and storage infrastructure.
  - Using virtualization or containerization (cgroups, Kubernetes, Dockers, VMs,…).
  - Often said to subsume HaaS.
  - Ex: Amazon Web Services (AWS: EC2 and S3), Microsoft Azure, Google Cloud.

II. On-demand access: *aaS Classification

- PaaS: Platform as a Service
  - You get access to flexible computing and storage infrastructure, coupled with a software platform (often tightly coupled)
  - Ex: Google's AppEngine (Python, Java, Go)
- SaaS: Software as a Service
  - You get access to software services, when you need them.
  - Ex: Google docs, MS Office 365 Online
- FaaS = Function as a Service
  - you upload a function, and cloud provider runs it when triggered
  - e.g., AWS Lambda, Azure Functions, etc.

III. Data-intensive Computing

- Computation-Intensive Computing
  - Example areas: MPI-based, High-performance computing, Grids
  - Typically run on supercomputers (e.g., NCSA Blue Waters)

- Data-Intensive
  - Typically store data at datacenters
  - Use compute nodes nearby
  - Compute nodes run computation services

- In data-intensive computing, the focus shifts from computation to data: CPU utilization no longer the most important resource metric, instead I/O is (disk and/or network)

IV. New Cloud Programming Paradigms

- Easy to write and run highly parallel programs in new cloud programming paradigms:
  - Google: MapReduce and Sawzall
  - Amazon: Elastic MapReduce service (pay-as-you-go)
  - Google (MapReduce)
    - Indexing: a chain of 24 MapReduce jobs
    - ~200K jobs processing 50PB/month (in 2006)
  - Yahoo! (Hadoop + Pig)
    - WebMap: a chain of several MapReduce jobs
    - 300 TB of data, 10K cores, many tens of hours (~2008)
  - Facebook (Hadoop + Hive)
    - ~300TB total, adding 2TB/day (in 2008)
    - 3K jobs processing 55TB/day
  - Similar numbers from other companies, e.g., Yieldex, eharmony.com, etc.
  - NoSQL: MySQL is an industry standard, but Cassandra is 2400 times faster!

Two Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- Private clouds are accessible only to company employees
- Public clouds provide service to any paying customer

- You're starting a new service/company: should you use a public cloud or purchase your own private cloud?

Single site Cloud: to Outsource or Own?

- Medium-sized organization: wishes to run a service for *M* months
  - Service requires 128 servers (1024 cores) and 524 TB
  - Same as UIUC CCT (Cloud Computing Testbed) cloud site (bought in 2009, now decommissioned)
- Outsource (e.g., via AWS): *monthly* cost
  - S3 costs: $0.12 per GB month. EC2 costs: $0.10 per CPU hour (costs from 2009)
  - Storage = $ 0.12 X 524 X 1000 ~ $62 K
  - Total = Storage + CPUs = $62 K + $0.10 X 1024 X 24 X 30 ~ $136 K
- Own: monthly cost
  - Storage ~ $349 K / *M*
  - Total ~ $ 1555 K / *M* + 7.5 K (includes 1 sysadmin / 100 nodes)
    - using 0.45:0.4:0.15 split for hardware:power:network and 3 year lifetime of hardware

Single site Cloud: to Outsource or Own?

- Breakeven analysis: more preferable to own if:
  - $349 K / *M* < $62 K (storage)
  - $ 1555 K / *M* + 7.5 K < $136 K (overall)

  Breakeven points
  - *M* > 5.55 months (storage)
  - *M* > 12 months (overall)

As a result

Startups use clouds a lot

Academic Clouds: Emulab/CloudLab

- A community resource open to researchers in academia and industry. Very widely used by researchers everywhere today.
- https://www.emulab.net/
- A cluster, with currently ~500 servers
- Founded and owned by University of Utah (led by Late Prof. Jay Lepreau)

Accessible to researchers with a qualifying grant

- Chameleon Cloud: https://www.chameleoncloud.org/

CloudLab: https://www.cloudlab.us/

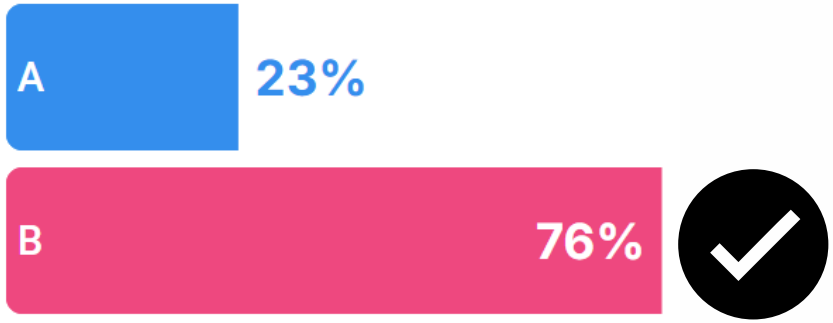
- Build your own cloud on their hardware

Writing to S3

TOTAL RESPONSES: 290

Writing 1 GB to Amazon S3 as 1 million small objects costs approximately the same as writing 1 GB as one large object?

A. True
B. False



Bar graph

Different costs

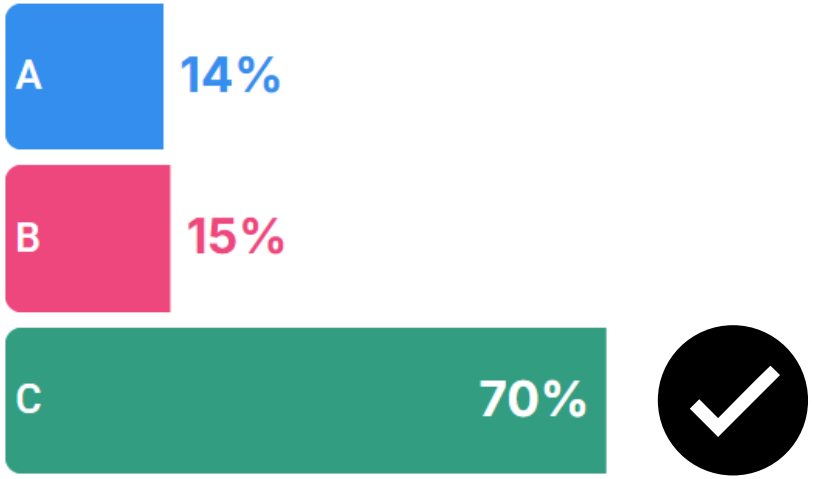
You run a replicated service across 3 Availability Zones (AZs) for fault tolerance. At high throughput, which is likely to be the biggest cloud infrastructure cost?

A. Compute — running the replicas

B. Storage — storing the data

C. Cross-AZ network traffic — replicating data between replicas

TOTAL RESPONSES: 308



Bar graph

Cloud economics

- Dictates system design
- Solutions:
  - Batch multiple small writes to write to S3
  - Use S3 for fault tolerance instead of replicating at the application level
- Leads to interesting designs…

Announcements

- **4 cr only: MP Groups Form DUE** this week **Mon Aug 31st** (see course webpage).
  - **Hard deadline, as Engr-IT will create and assign VMs the day after!**
  - **(Form does NOT automatically "enroll" you in 4cr section—you still have to do that yourself)**
- DO NOT
  - Change MP groups unless your partner has dropped
  - Leave your MP partner hanging: Both MP partners should contribute equally (we will ask!)
- **All: Please fill out Student Survey (see course webpage).**

- MP1 due Sep 13th
  - VMs being distributed soon (watch Piazza)
  - Demos will be Monday Sep 14th (schedule and details will be posted soon on Piazza)

- HW1 out today due Sep 20th: Solve problems right after lecture covers topic!
- **Check Piazza often! It's where all the announcements are at!**