

INFM 603: Information Technology and Organizational Context

Session I I: Cloud Computing and Big Data



Jimmy Lin
The iSchool
University of Maryland

Thursday, November 14, 2013

An aerial photograph showing a vast, dense layer of white, fluffy clouds stretching across the horizon. The sky above is a clear, deep blue. The clouds are illuminated from the side, creating soft shadows and highlights that give them a three-dimensional appearance. The overall scene is serene and expansive.

What is cloud computing?



The best thing since sliced bread?

- Before clouds...
 - Grids
 - Connection machines
 - Vector supercomputers
 - ...
- Cloud computing means many different things:
 - Large-data processing
 - Rebranding of web 2.0
 - Utility computing
 - Everything as a service

Rebranding of web 2.0

- Rich, interactive web applications
 - Clouds refer to the servers that run them
 - AJAX as the de facto standard (for better or worse)
 - Examples: Facebook, YouTube, Gmail, ...
- “The network is the computer”: take two
 - User data is stored “in the clouds”
 - Rise of the tablets, smartphones, etc.
 - Browser is the OS

GENERAL  ELECTRIC

Rr13⁸/₉



KILOWATTHOURS

CL 200

TYPE I-60-S
SINGLE STATOR



FM 2S
WATTHOUR METER

TA 30

240V

3W

CAT. NO.

720X1G1

K_h 7.2

60~

7
P
G
and
E

397128

•44 617 187•

MADE IN U.S.A.

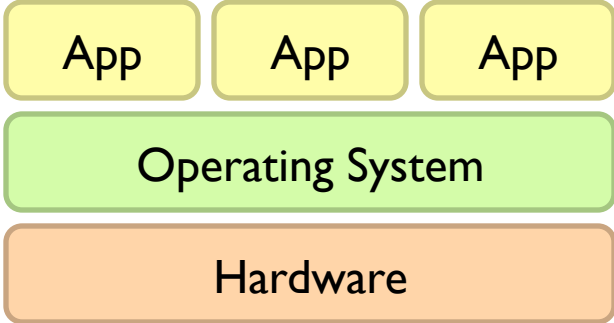
Utility Computing

- What?
 - Computing resources as a metered service (“pay as you go”)
 - Ability to dynamically provision virtual machines
- Why?
 - Cost: capital vs. operating expenses
 - Scalability: “infinite” capacity
 - Elasticity: scale up or down on demand
- Does it make sense?
 - Benefits to cloud users
 - Business case for cloud providers

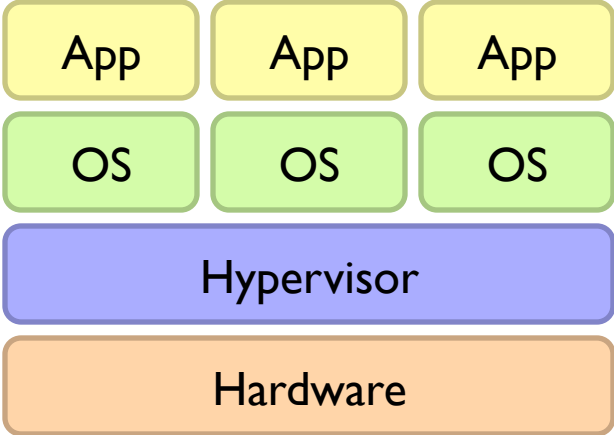
I think there is a world market for about five computers.



Enabling Technology: Virtualization



Traditional Stack



Virtualized Stack

Everything as a Service

- Utility computing = Infrastructure as a Service (IaaS)
 - Why buy machines when you can rent them?
 - Examples: Amazon's EC2, Rackspace
- Platform as a Service (PaaS)
 - Give me nice API and take care of the maintenance, upgrades, ...
 - Example: Google App Engine
- Software as a Service (SaaS)
 - Just run it for me!
 - Example: Gmail, Salesforce

Different Types of Clouds

- Public clouds
- Private clouds
- Hybrid clouds



Our World: Large Data

Google™ processes 20 PB a day (2008)

ebay®

10 PB data in Hadoop/Teradata
75B DB calls/day (6/2012)

100+ PB user data
+500 TB/day (8/2012)

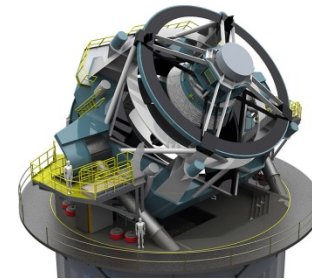
facebook®

640K ought to be
enough for anybody.



Wayback Machine:
10 PB web archive (10/2012)

LHC: 15 PB a year



LSST: 6-10 PB a year (~2015)

How much data?



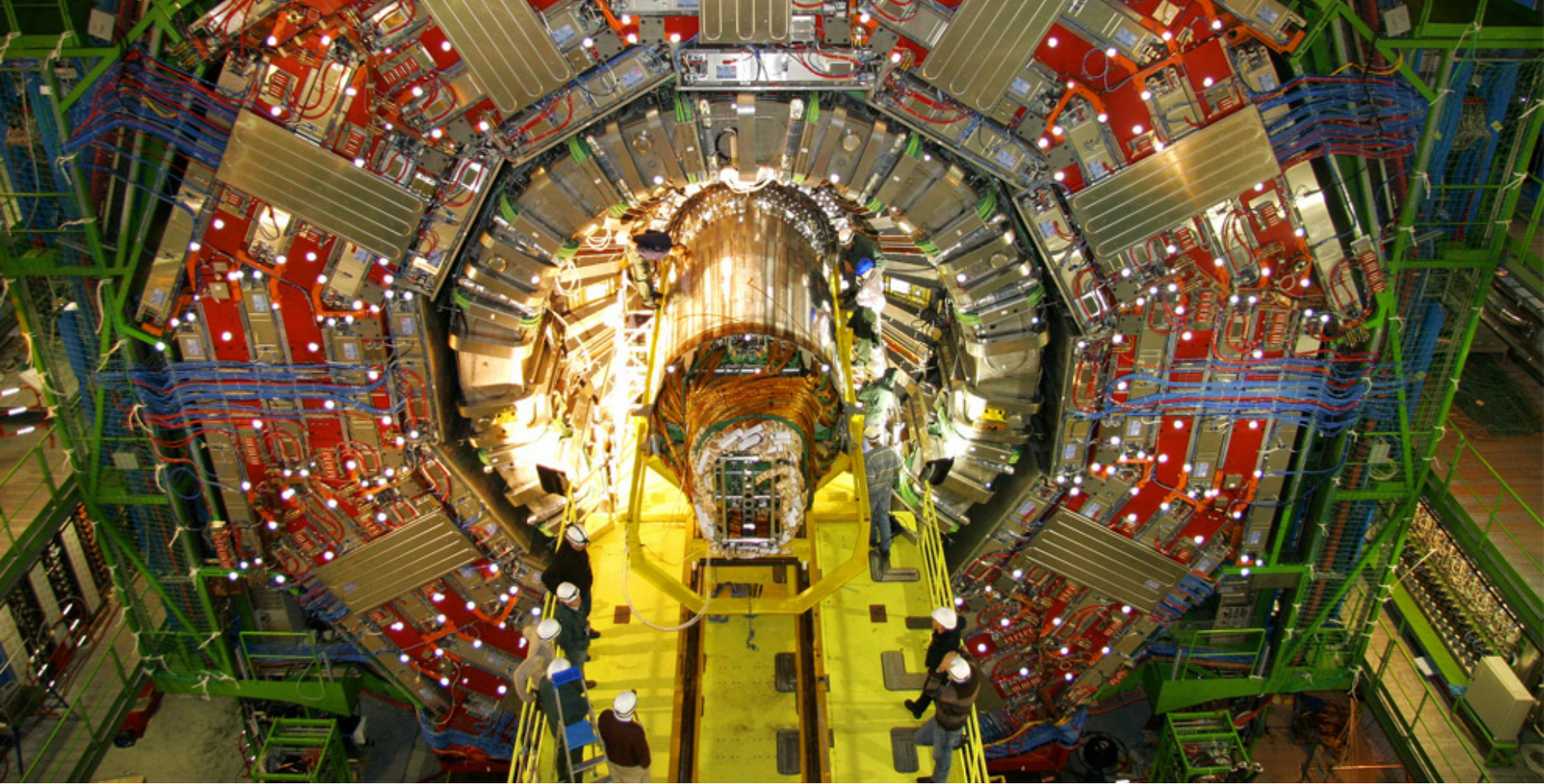
A photograph of a snow-capped mountain range under a clear blue sky. The mountains are rugged and covered in patches of snow and ice. The sky is a solid, clear blue.

Why large data?

Science

Engineering

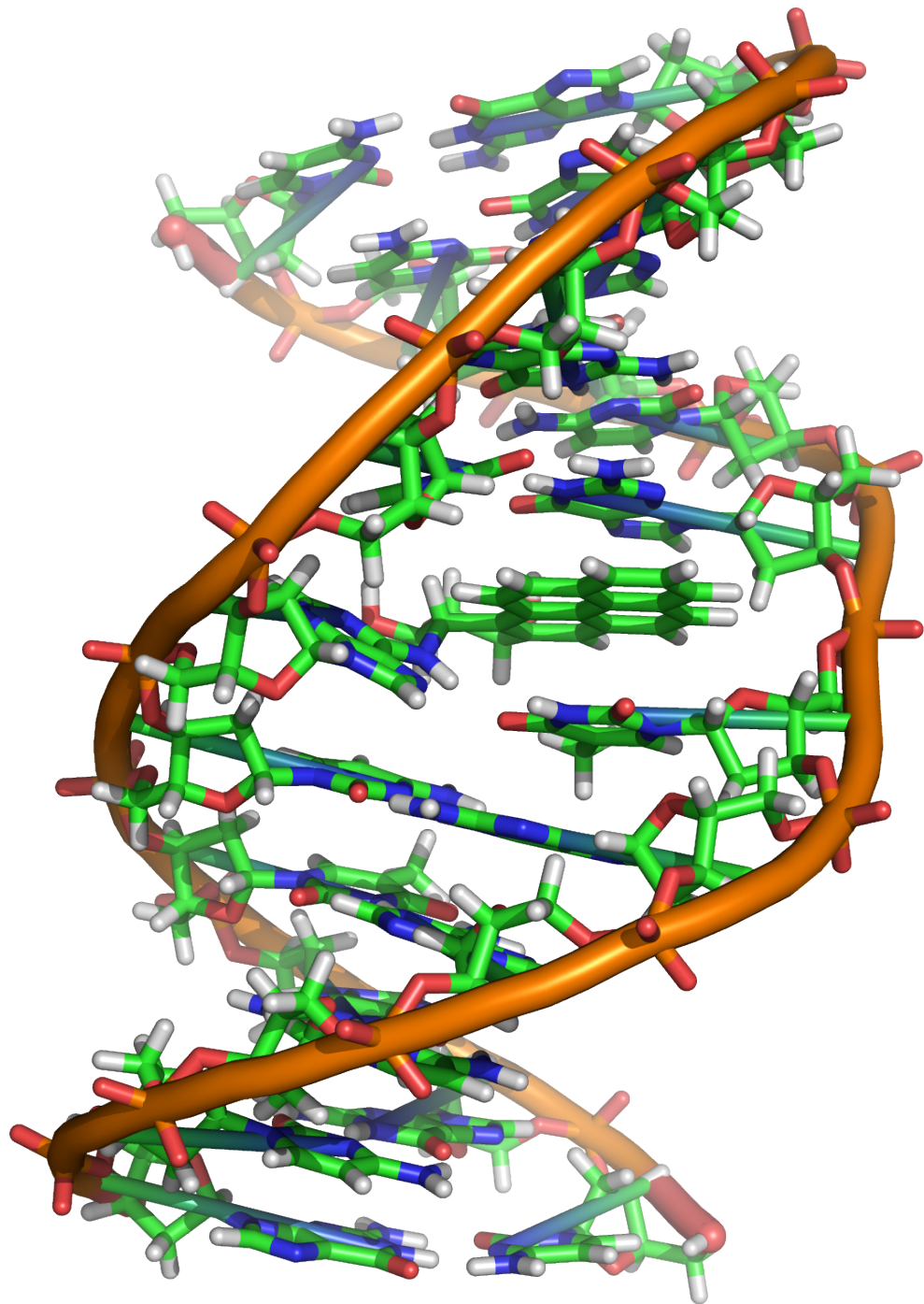
Commerce

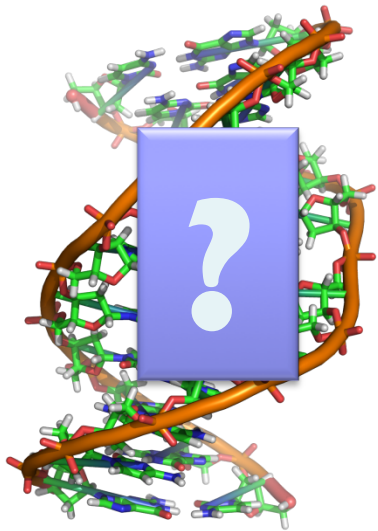


Science

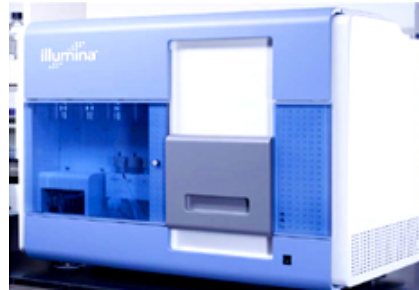
- Emergence of the 4th Paradigm
- Data-intensive e-Science







Subject genome



Sequencer

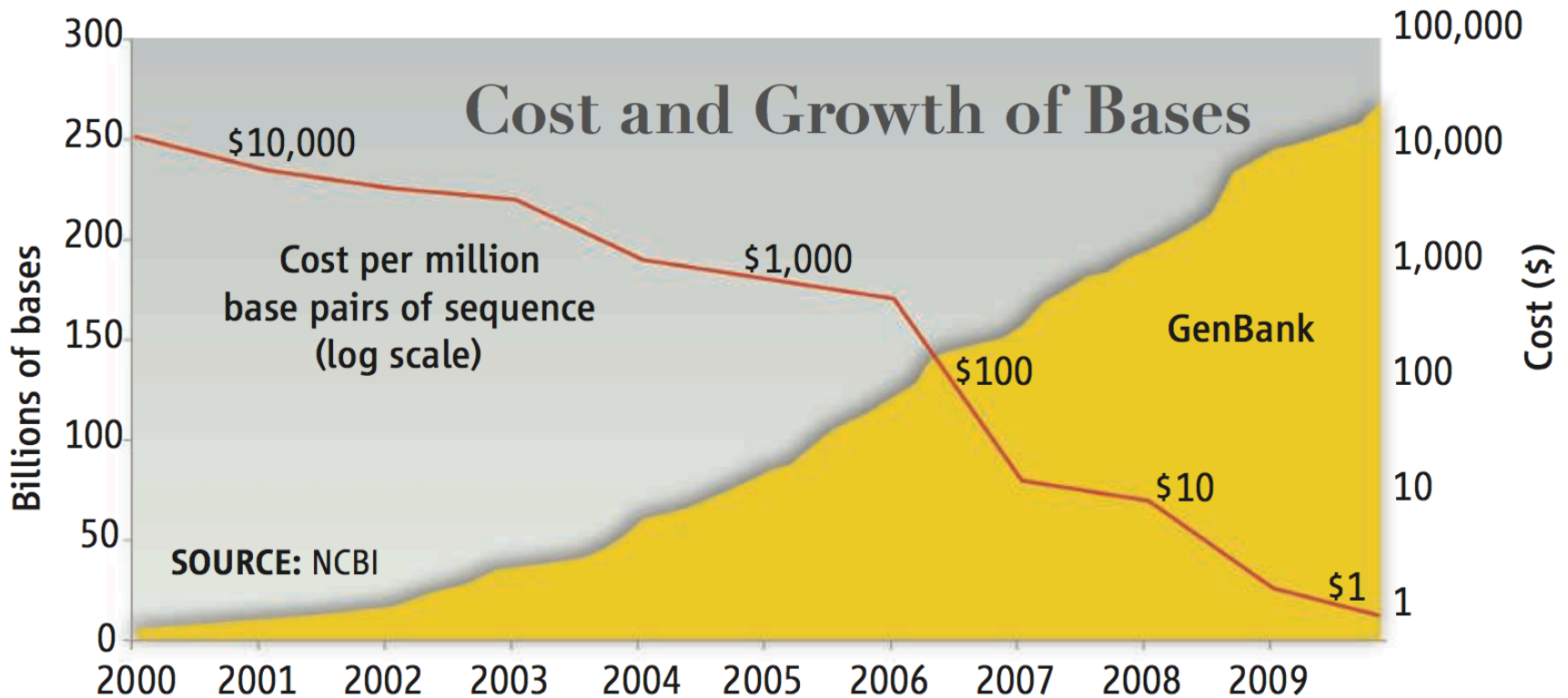
GATGCTTACTATGCGGGCCCC
CGGTCTAATGCTTACTATGC
GCTTACTATGCGGGCCCCCTT
AATGCTTACTATGCGGGCCCCCTT
TAATGCTTACTATGC
AATGCTTAGCTATGCGGGC
AATGCTTACTATGCGGGCCCCCTT
AATGCTTACTATGCGGGCCCCCTT
CGGTCTAGATGCTTACTATGC
AATGCTTACTATGCGGGCCCCCTT
CGGTCTAATGCTTAGCTATGC
ATGCTTACTATGCGGGCCCCCTT

Reads

Human genome: 3 gbp
A few billion short reads
(~100 GB compressed data)

DNA Data Tsunami

Current world-wide sequencing capacity exceeds 13 Pbp/year and is growing at 5x per year!



“Will Computers Crash Genomics?”

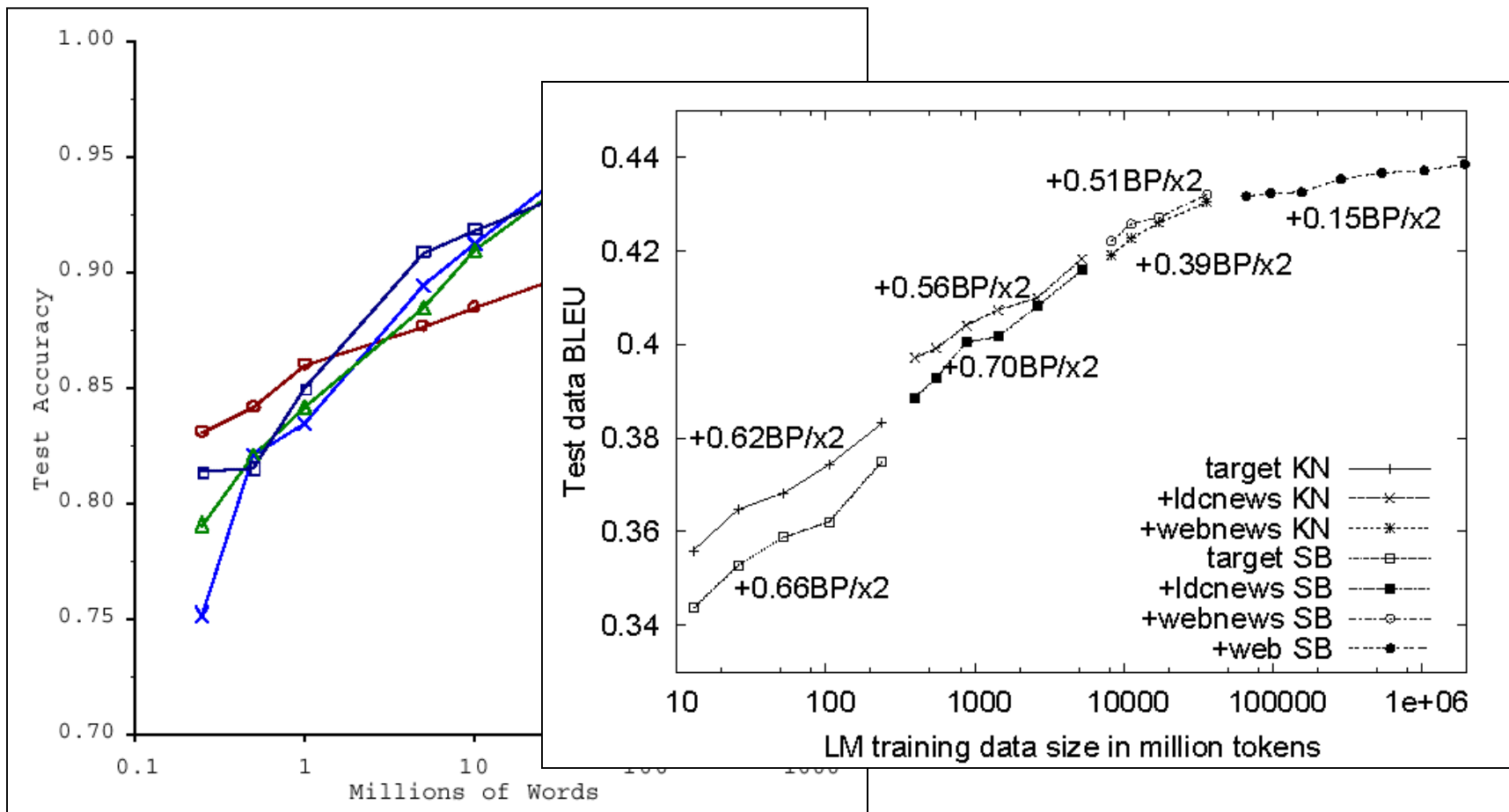
Elizabeth Pennisi (2011) *Science*. 331(6018): 666-668.



Engineering

- The unreasonable effectiveness of data
- Count and normalize!

No data like more data!



What to do with more data?

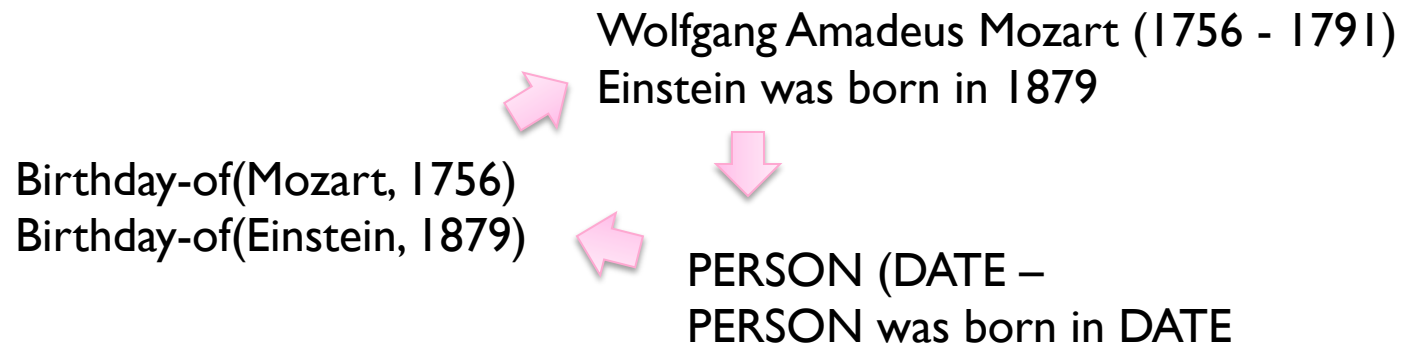
- Answering factoid questions

- Pattern matching on the Web
- Works amazingly well

Who shot Abraham Lincoln? → **X** shot Abraham Lincoln

- Learning relations

- Start with seed instances
- Search for patterns on the Web
- Using patterns to find more instances





Commerce

- Know thy customers
- Data → Insights → Competitive advantages

Business Intelligence

- Premise: more data leads to better business decisions
 - Periodic reporting as well as ad hoc queries
 - Rise of the data scientist
 - Listen to your customers, not the HiPPO
- Examples:
 - Slicing-and-dicing activity by different dimensions to better understand the marketplace
 - Analyzing log data to improve front-end experience
 - Analyzing log data to better optimize ad placement
 - Analyzing purchasing trends for better supply-chain management
 - Mining for correlations between otherwise unrelated activities

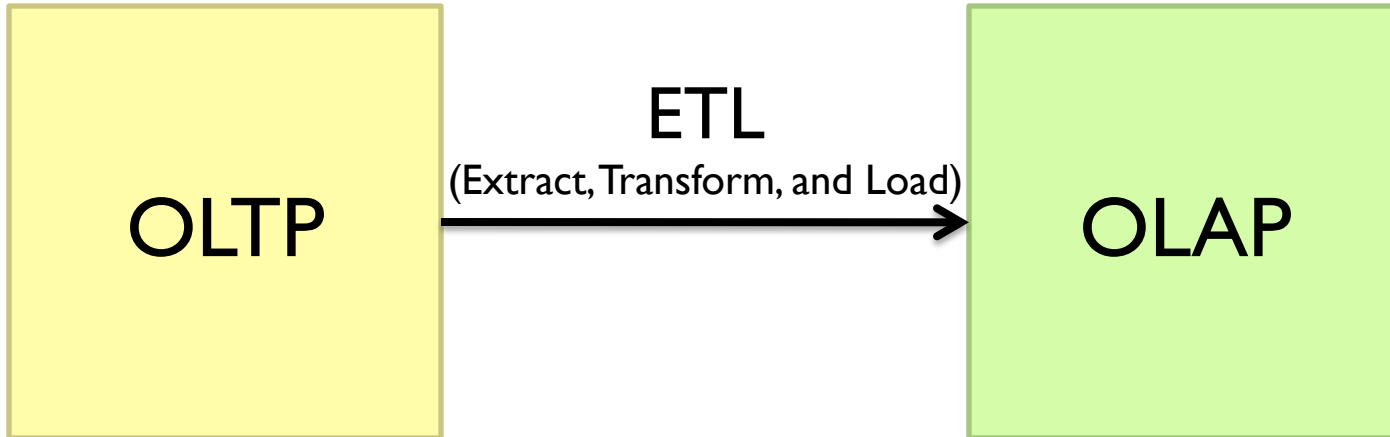
Database Workloads

- OLTP (online transaction processing)
 - Typical applications: e-commerce, banking, airline reservations
 - User facing: real-time, low latency, highly-concurrent
 - Tasks: relatively small set of “standard” transactional queries
 - Data access pattern: random reads, updates, writes (involving relatively small amounts of data)
- OLAP (online analytical processing)
 - Typical applications: business intelligence, data mining
 - Back-end processing: batch workloads, less concurrency
 - Tasks: complex analytical queries, often ad hoc
 - Data access pattern: table scans (involving large amounts of data)

One Database or Two?

- Downsides of co-existing OLTP and OLAP workloads
 - Poor memory management
 - Conflicting data access patterns
 - Variable latency
- Solution: separate databases
 - User-facing OLTP database for high-volume transactions
 - Data warehouse for OLAP workloads
 - How do we connect the two?

OLTP/OLAP Architecture



OLTP/OLAP Integration

- OLTP database for user-facing transactions
 - Retain records of all activity
 - Periodic ETL (e.g., nightly)
- Extract-Transform-Load (ETL)
 - Extract records from source
 - Transform: clean data, check integrity, aggregate, etc.
 - Load into OLAP database
- OLAP database for data warehousing
 - Business intelligence: reporting, ad hoc queries, data mining, etc.
 - Feedback to improve OLTP services

Challenge of Big Data

- Volume
- Cost
- ETL Latency

An aerial photograph showing a vast, dense layer of white, fluffy clouds stretching across the horizon. The sky above is a clear, deep blue. The clouds are illuminated from the side, creating soft shadows and highlights that give them a three-dimensional appearance. The overall scene is serene and expansive.

cloud computing meets big data

Cloud Computing Meets Big Data

- Rise of social media and user-generated content
 - Cloud services exacerbates big data problems
- Utility computing democratizes big data capabilities
 - Efficient dynamic allocation of large-scale computing resources

An aerial photograph showing a vast, dense layer of white, fluffy clouds stretching across the horizon. The clouds are illuminated from the side, creating soft shadows and highlights. The sky above is a clear, deep blue. The overall scene is serene and expansive.

What *really* is the cloud?



Source: Wikipedia (The Dalles, Oregon)





Source: Bonneville Power Administration



Source: Google



Source: Google



Source: Facebook

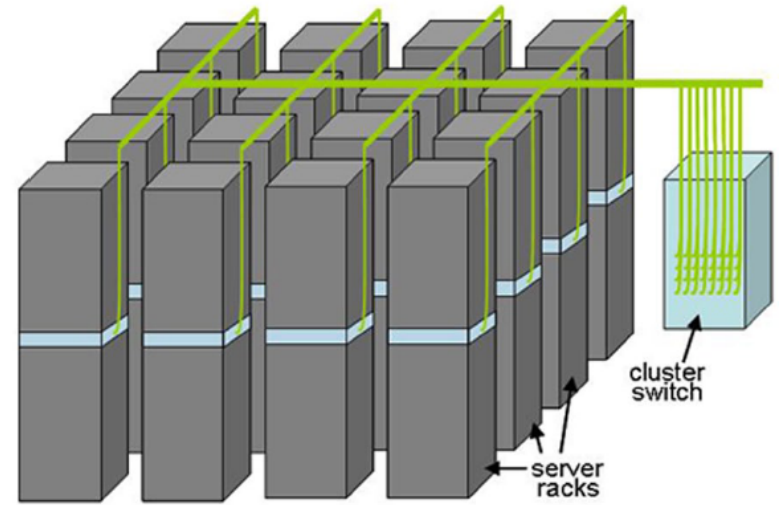
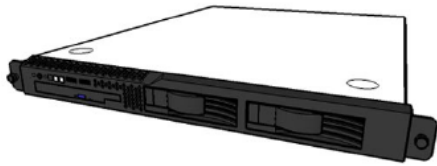


Source: Facebook

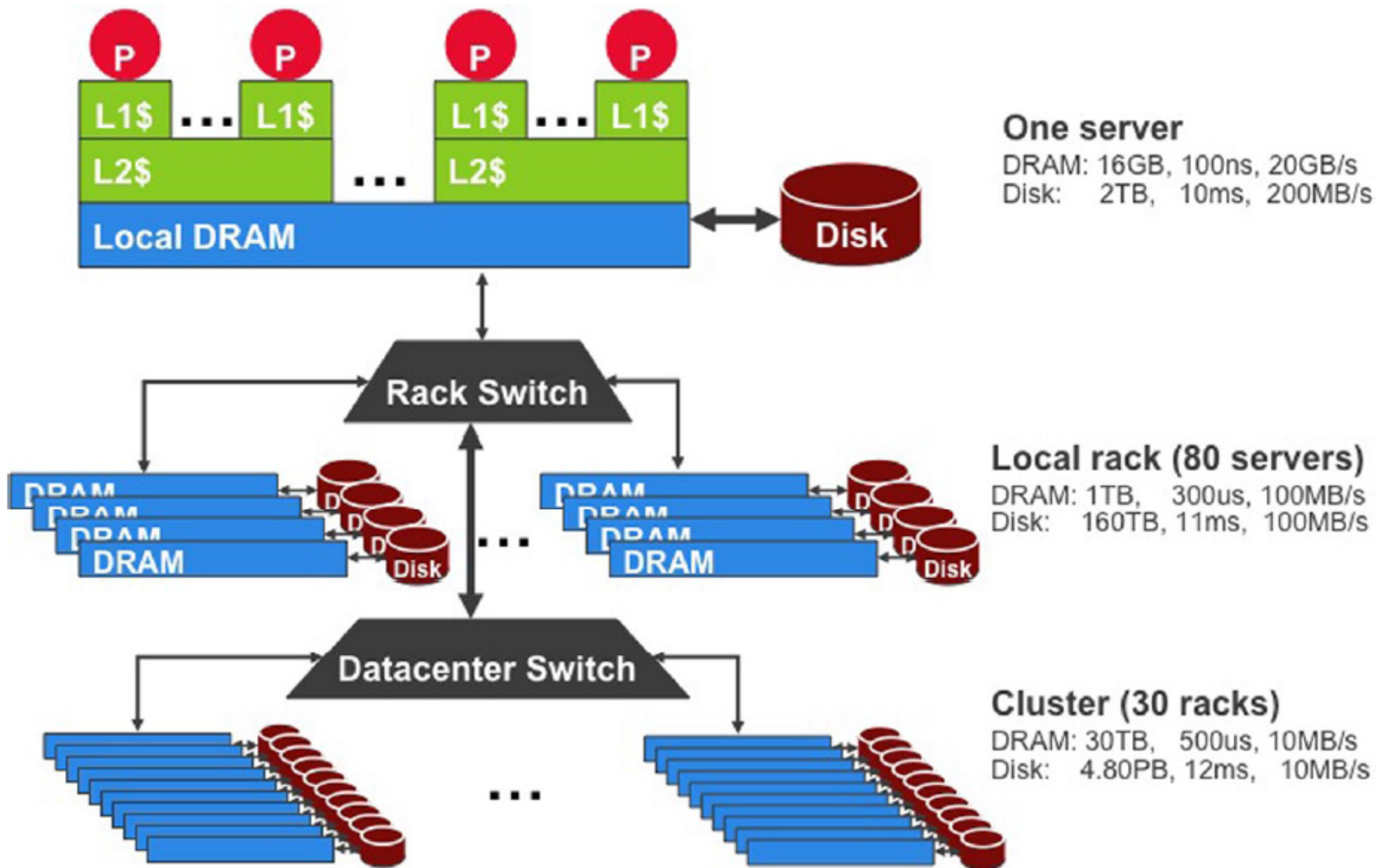


Source: Facebook

Building Blocks

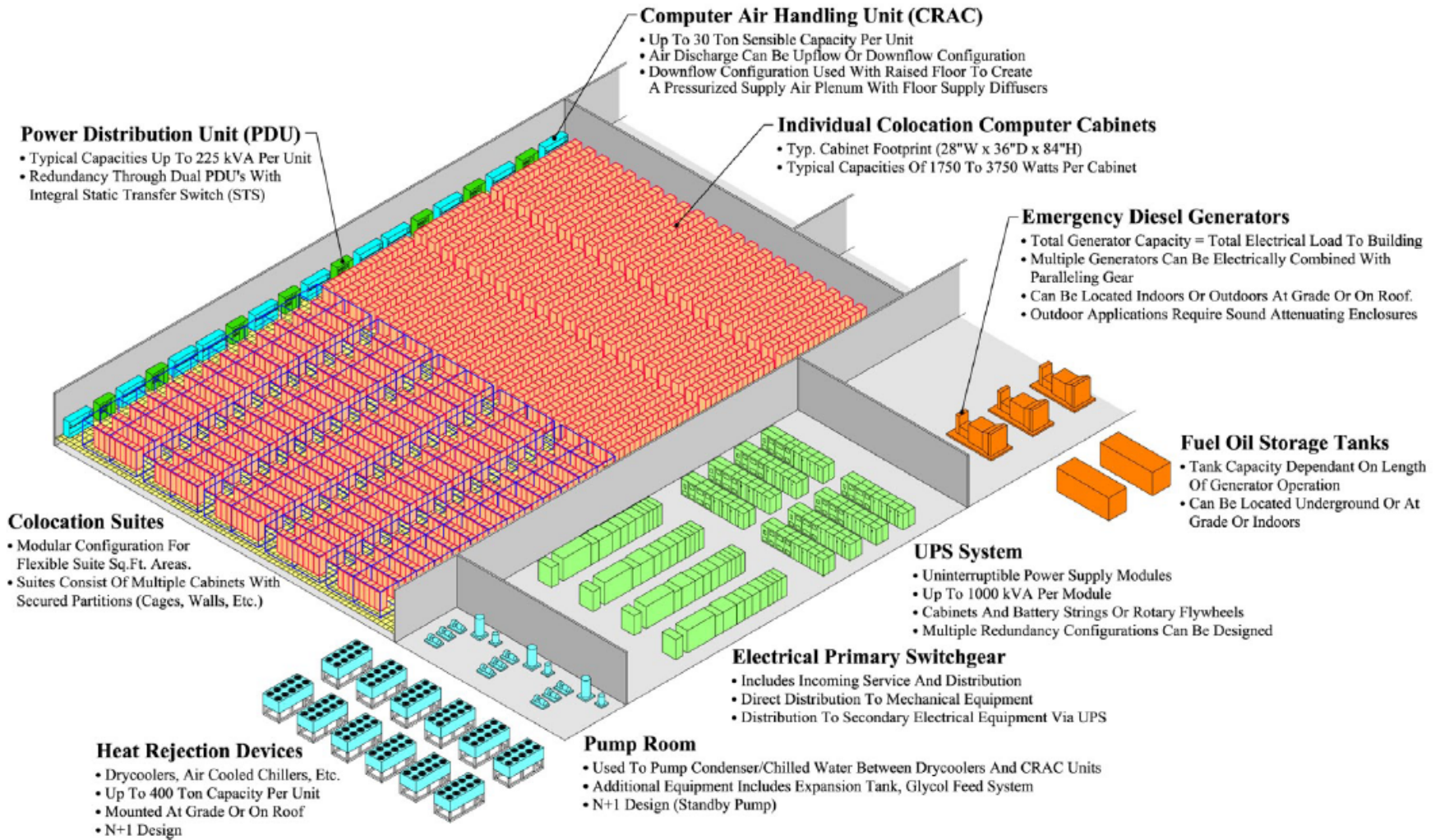


Storage Hierarchy

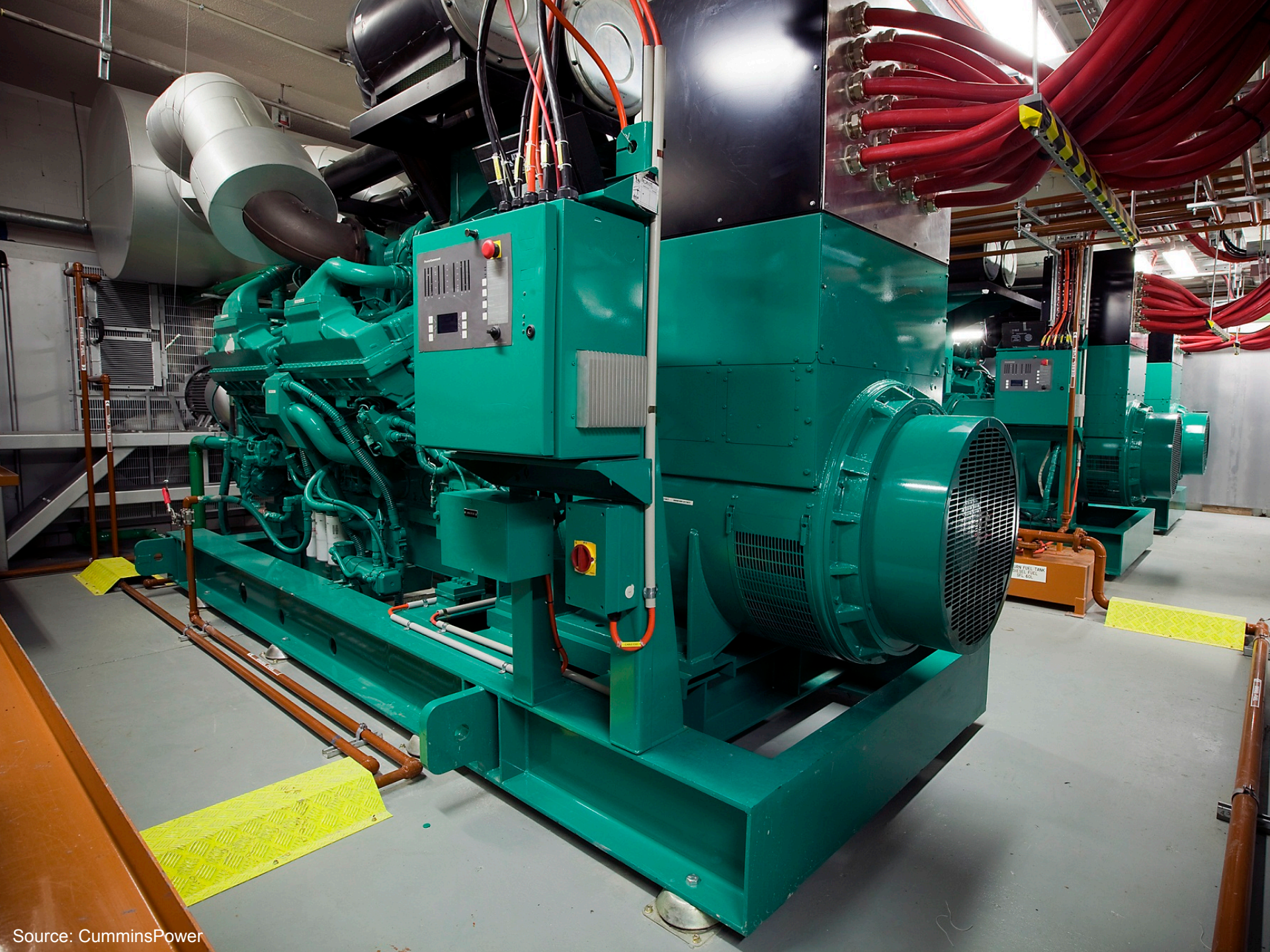


Funny story about sense of scale...

Anatomy of a Datacenter









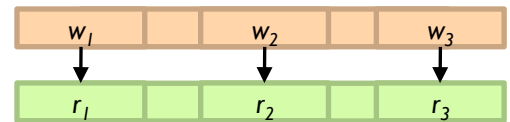
A photograph of a massive, snow-covered mountain range, likely the Himalayas, under a clear blue sky. The mountains are rugged and jagged, with significant snow cover. The text "How large data?" is overlaid in the center of the image in a bold, white, sans-serif font.

How large data?



Divide et impera

- Chop problem into smaller parts
- Combine partial results



Synchronization Challenges

- How to split large chunks up into smaller ones
- How to integrate results from each chunk
- How to distribute shared data
- How to update shared data
- How to coordinate access to shared resources
- How to schedule different processing chunks
- How to cope of machine failure



Source: Ricardo Guimarães Herrmann

Typical Large-Data Problem

- Iterate over a large number of records

Map Extract something of interest from each

- Shuffle and sort intermediate results

- Aggregate intermediate results

- Generate final output

Reduce

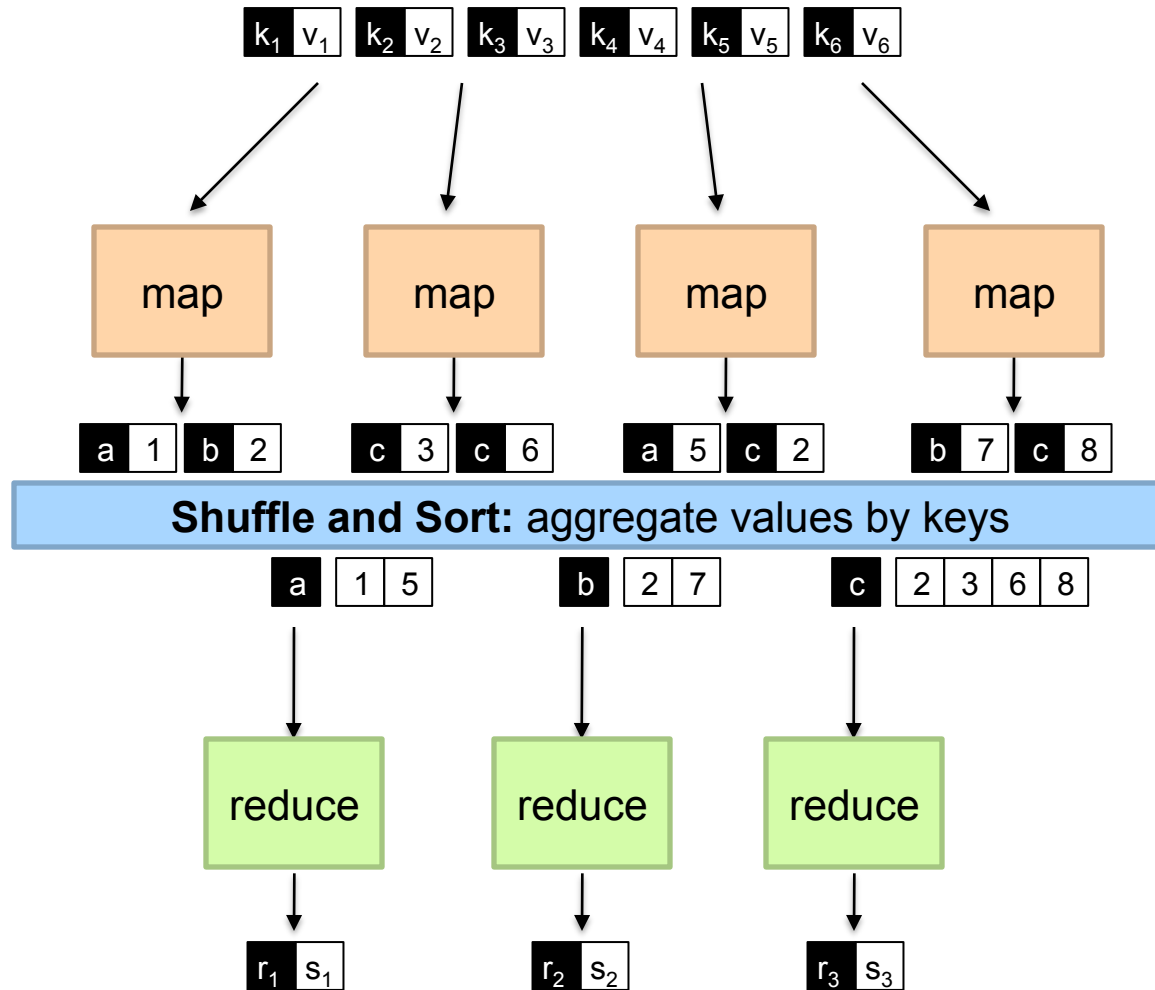
MapReduce

- Programmers specify two functions:

map $(k, v) \rightarrow \langle k', v' \rangle^*$

reduce $(k', v') \rightarrow \langle k', v' \rangle^*$

- All values with the same key are sent to the same reducer
- The execution framework handles everything else...



MapReduce

- Programmers specify two functions:

map $(k, v) \rightarrow \langle k', v' \rangle^*$

reduce $(k', v') \rightarrow \langle k', v' \rangle^*$

- All values with the same key are sent to the same reducer
- The execution framework handles everything else...

What's “everything else”?

MapReduce “Runtime”

- Handles scheduling
 - Assigns workers to map and reduce tasks
- Handles “data distribution”
 - Moves processes to data
- Handles synchronization
 - Gathers, sorts, and shuffles intermediate data
- Handles errors and faults
 - Detects worker failures and restarts

MapReduce Word Count

Map(String docid, String text):

for each word w in text:

Emit(w, 1);

Reduce(String term, Iterator<Int> values):

int sum = 0;

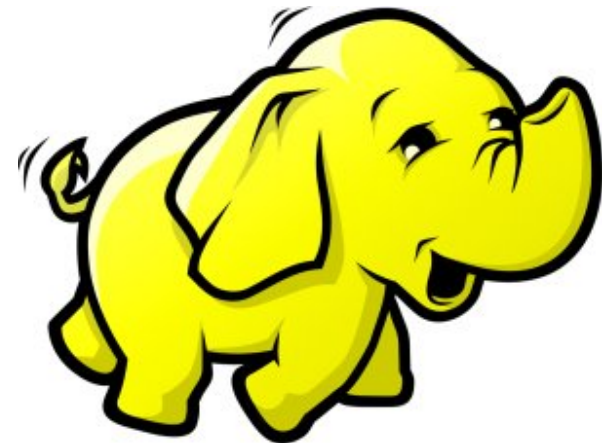
for each v in values:

sum += v;

Emit(term, value);

MapReduce Implementations

- Google has a proprietary implementation
- Hadoop is an open-source implementation in Java
 - Originally developed by Yahoo, now an Apache project
 - Center of a rapidly expanding software ecosystem



Now you know...

- Cloud computing
- Big data
- Relationship between the two
- Challenges with big data processing
- MapReduce/Hadoop

An aerial photograph showing a vast, dense layer of white, fluffy clouds stretching across the horizon. The clouds are illuminated from the side, creating soft shadows and highlights. The sky above is a clear, deep blue. The overall scene is serene and expansive.

Questions?