

DATA PROCESSING

Venkatesh Vinayakarao

venkateshv@cmi.ac.in

<http://vvtesh.co.in>

Chennai Mathematical Institute

Data is the new oil. - Clive Humby, 2006.

What Comes Next?

byte

kilobyte

megabyte

gigabyte

??

???

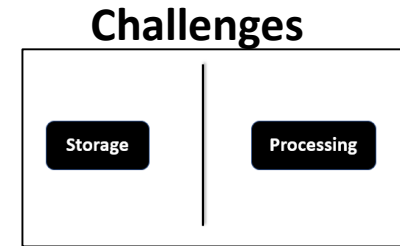
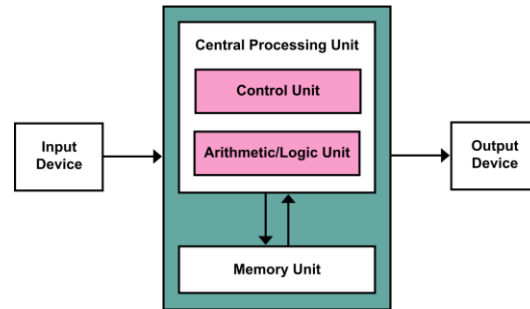
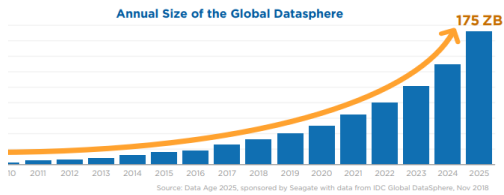
????

?????

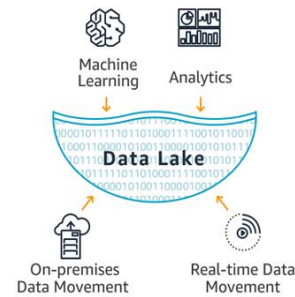
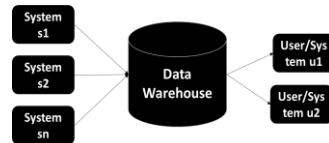
Sizes

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

Recap

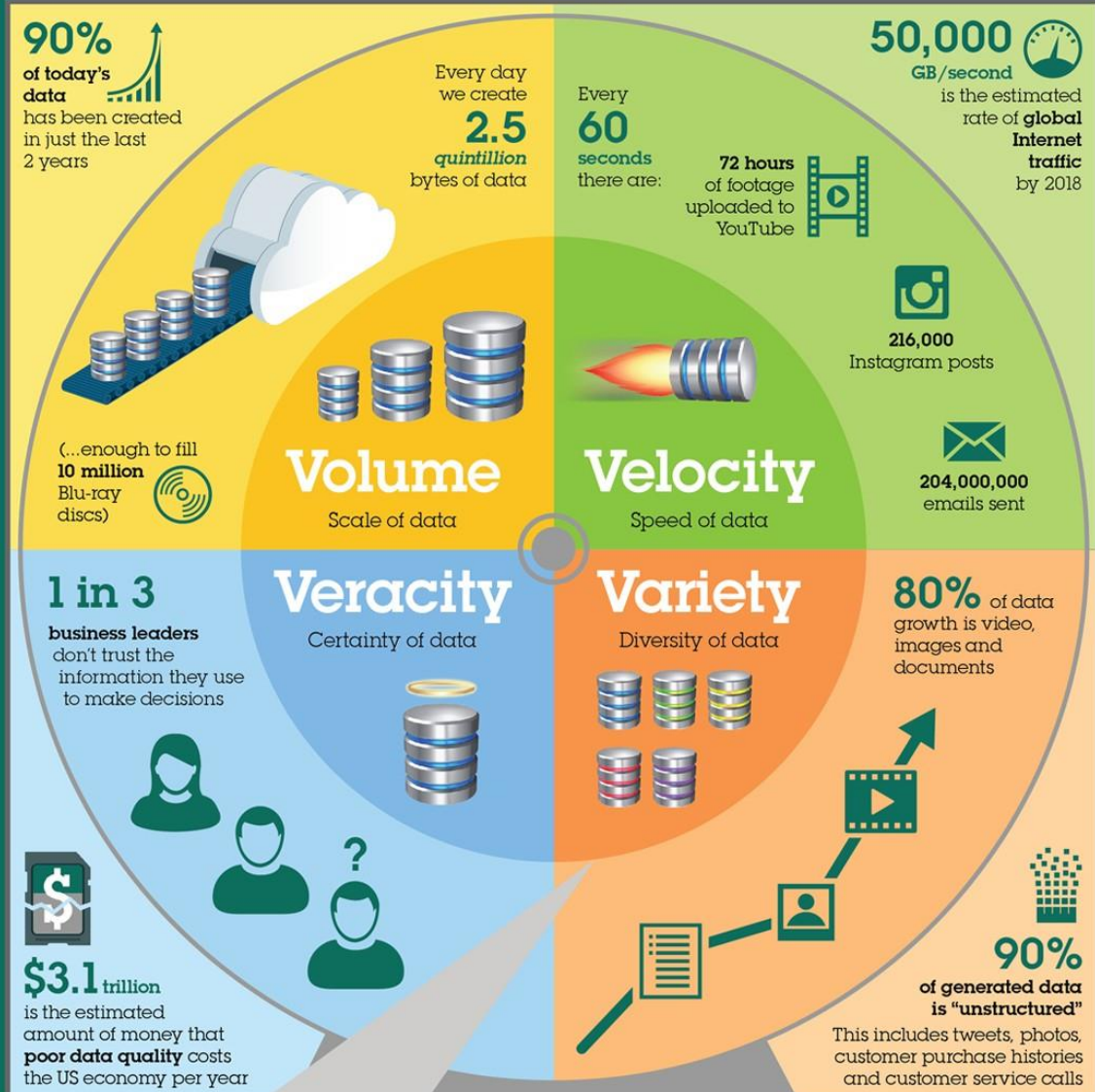


Data Storage



Amazon S3
STaaS

Extracting business value from the 4 V's of big data



Big Data Characteristics

- Volume
 - Petabytes, exabytes,...
- Variety
 - pdf, json, text, images, ...
- Velocity
 - real-time, near real-time, batch
- Veracity
 - Trustworthiness, correctness and consistency

“Where there is data smoke, there
is business fire.”

— Thomas Redman, Author.



Tuj mein rab diktha hai
Data mein kya karoon

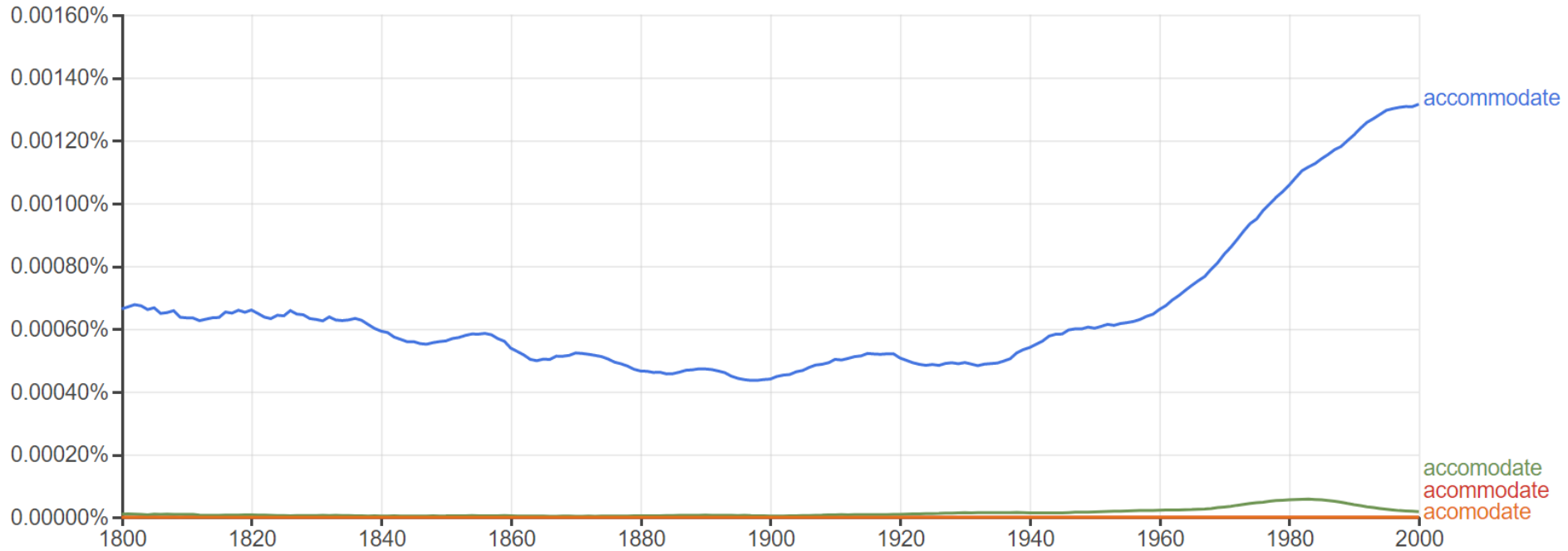
jab bhi koi **data** dekhun
mera dil deewana bole
ole ole ole...



Quiz

- Which is right?
 - accommodate
 - acommodate
 - accomodate
 - acomodate

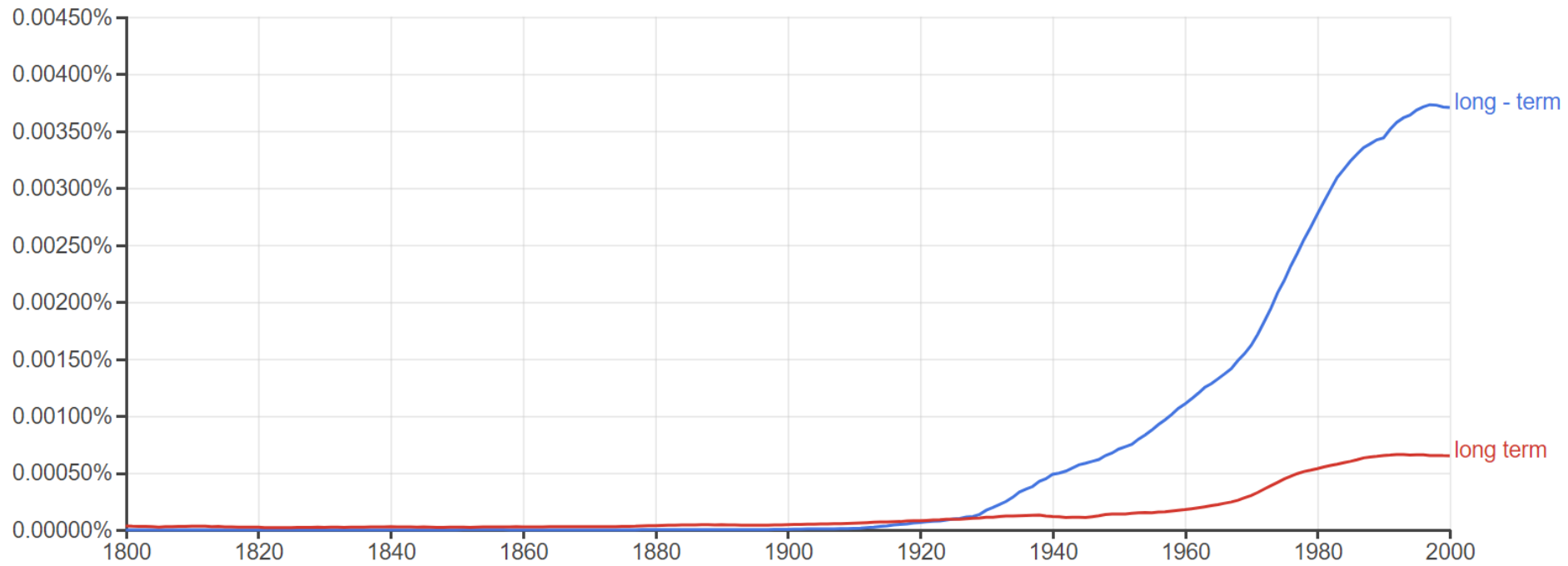
Google n-gram Viewer



<https://books.google.com/ngrams>

Quiz

- Long-term or long term



Data Processing

- Microprocessors
- Multi-core Processors
- Supercomputers
- ... all roads lead to ~~Rome~~ Cloud!



Microprocessors

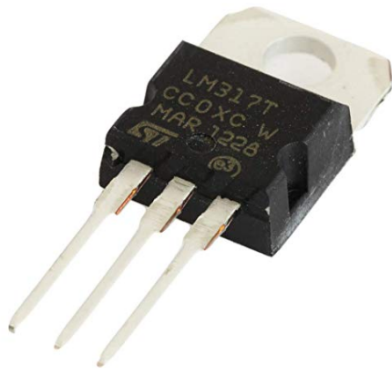
Processing unit on an integrated circuit

What are ICs made of?



Transistors

- Basic electronic component that alters the flow of current.
- Form the basic building block of an integrated circuit.
- Think of it as an electronic switch



SunRobotics LM317 Voltage Regulator IC TO-220 Adjustable Three-Terminal Regulators Field Effect Transistor Original Integrated circuit electronic Components (5 Pcs)

by sunrobotics

★★★★☆ 1 rating

M.R.P.: ₹ 199.00

Price: ₹ 165.00

You Save: ₹ 34.00 (17%)

Inclusive of all taxes

✓prime FREE Delivery by Monday



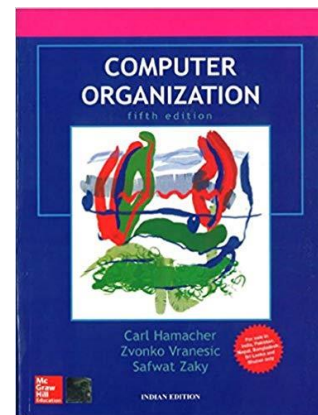
Pay on Delivery



10 Days Returnable



Amazon Delivered

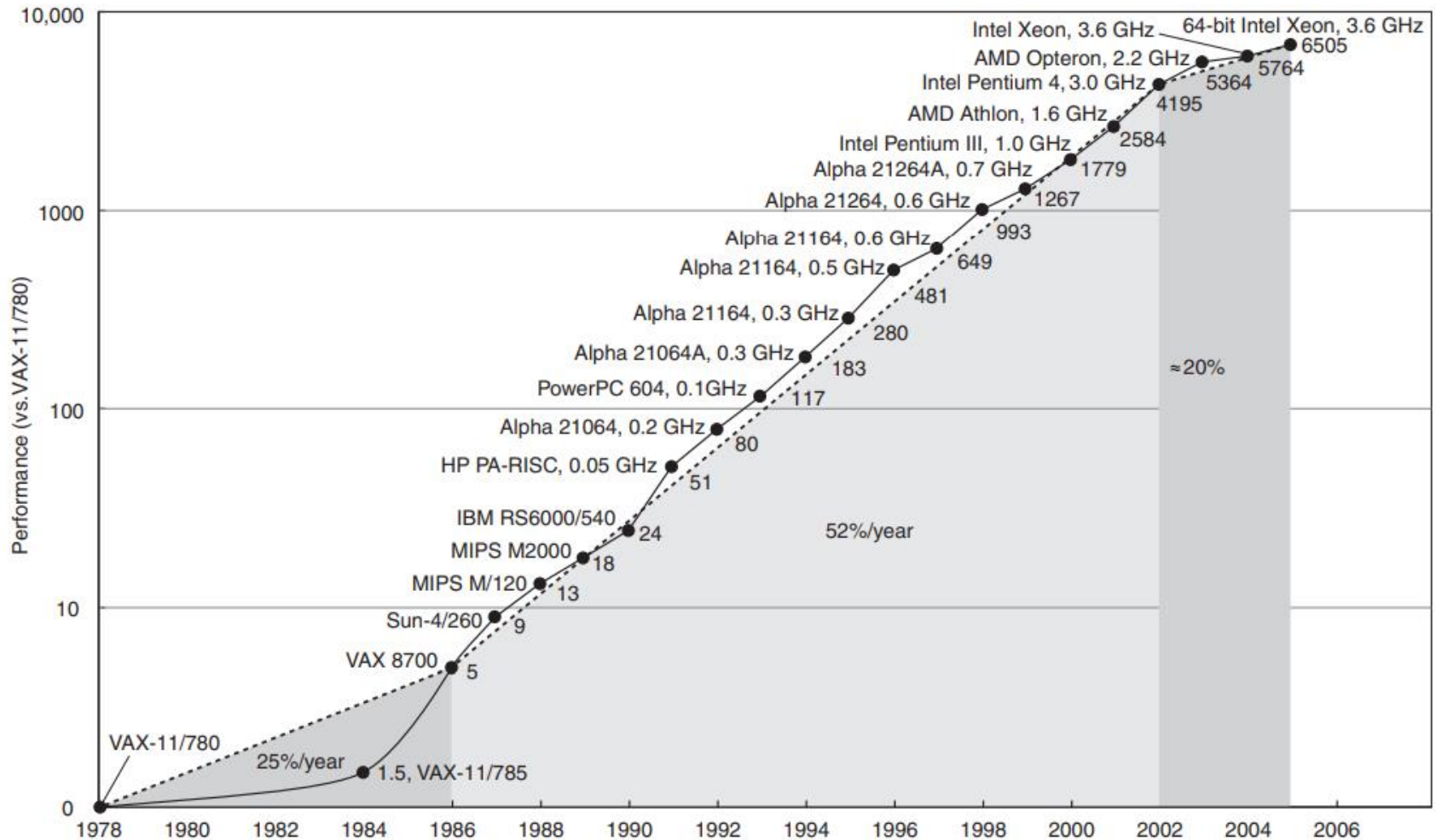


Logic Gates

- Implements Boolean functions (thus performs logical operations)
- Implemented using Transistors

Microprocessors contain millions of logic gates.

Processor Performance



Moore's Law

The number of transistors on a microchip doubles every two years, though the cost of computers is halved.

CES 2019: **Moore's Law is dead**, says Nvidia's CEO

CNET - 10-Jan-2019

Intel, for its part, doesn't think **Moore's Law is dead**. Companies are just finding new ways to keep it going, like Intel's new 3D chip stacking.

Moore's Law is far from death, according to Intel's Jim Keller

TweakTown - 10-Dec-2019

"The death of **Moore's Law** is coming and will limit how far we can go ... explanation of what **Moore's law** is, from Intel's co-founder, engineer, ...

Multi-Core Processors

- Two or more separate processing units (called cores)
- Enhances parallel processing



intel core duo
(has 2 cores, 2.66 GHz)



intel core i7
(has 4 cores, 4 GHz)

Quality Up

What do we achieve when we use p processors?

$$\text{Quality Up} = \frac{\textit{quality on } p \textit{ processors}}{\textit{quality on 1 processor}}$$

Read Section 1.1 of Jan Verschelde's book on ["Introduction to Supercomputing"](#).

Can we use multiple processors?

- Amdahl's law

- Let R be the fraction of the operations which cannot be parallelized. The speedup with p processors is bound by

$$\frac{1}{R + \frac{1-R}{p}}$$

- Example

- Say, 10% cannot be parallelized, and we have 8 processors. Best speedup = $\frac{1}{1/10 + \frac{1-1/10}{8}} \approx 4.7x$.

Multiple Processors for Speedup

- Amdahl's law

- Let R be the fraction of the operations which cannot be parallelized. The speedup with p processors is bound by

$$\frac{1}{R + \frac{1-R}{p}}$$

← Speed up in terms of problem size.

- Example

- Say, 10% cannot be parallelized, and we have 8 processors. Best speedup = $\frac{1}{1/10 + \frac{1-1/10}{8}} \approx 4.7x$.

- What happens if we had infinite processors?




Quiz

- Say, 10% of the operations cannot be parallelized. What happens if we had infinite processors?
- Answer: 10x.

Speedup

Speed up in terms of time.



- Gustafson's Law
 - If s is the fraction of serial operations in a parallel program run on p processors, then the scaled speedup is bounded by $p + (1 - p)s$.
- Example
 - Say, all other seven processors are kept idle while one processor completes 5% work, scaled speedup = $8 + (1 - 8) * 0.05 = 7.65$.

Our ability to parallelize determines the successful use of multi-core processors.

Supercomputer

- A computing system that provides close to the **best** currently **achievable sustained performance** on demanding computational problems.

How do supercomputers achieve such performance levels?

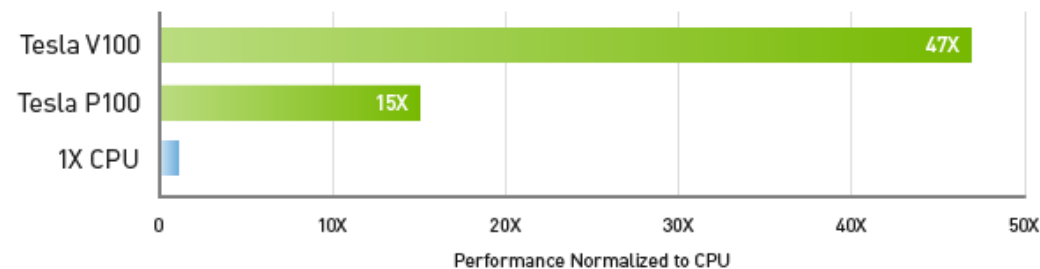
GPUs and GPGPUs

- Graphics Processing Unit (GPU)
 - Massive parallelization
 - Thousands of cores
 - Originally created for the gaming industry
- General Purpose GPU (GPGPU)
 - Architecture allows for programming (Example: Compute Unified Device Architecture (CUDA) on NVIDIA GPGPUs).
- Performance is measured in FLOP (Floating Point Operation)
 - sometimes, FLOPS (floating point operations per second)

CPU vs. GPU

- Say, two floating point operations could be performed in a clock cycle,
 - 3 GHz processor → 6 gigaflop per second.
- Top GPUs achieve petaflop per second.
 - Achieved through an array of cores (V100 has 5120 cores)

47X Higher Throughput Than CPU Server on Deep Learning Inference



Workload: ResNet-50 | CPU: 1X Xeon E5-2690v4 @ 2.6 GHz | GPU: Add 1X Tesla P100 or V100

My System



My Lenovo X390 uses
Intel® Core™ i7-8565U CPU @ 1.8 GHz

4 cores only! ☹️

Deep Blue

- Beat Chess World Champion Garry Kasparov in 1997



259th most powerful supercomputer.
Achieved 11.38 GFLOPS.

IBM Watson and Jeopardy! Game, 2011

- Cluster of **90 servers** each having **3.5GHz eight-core processor** and **16 TB of RAM**.
- Equivalent to 80 Teraflops (a slow supercomputer by today's standards).



Trivia

- Can you name the fastest supercomputer as of date?
 - How much data can it store?
 - How fast is it?

Trivia

- Can you name the fastest supercomputer as of date? **IBM SUMMIT**
 - How much data can it store? **250 PB**
 - How fast is it? **200 petaflops**

How fast is 200 petaflops?

Uses NVIDIA Tesla V100 GPU – How fast is its 200 petaflops?

"If every person on Earth completed one calculation per second, it would take the world population 305 days to do what Summit can do in 1 second" - Oak Ridge National Laboratory.

That is 200 quadrillion calculations in one second!

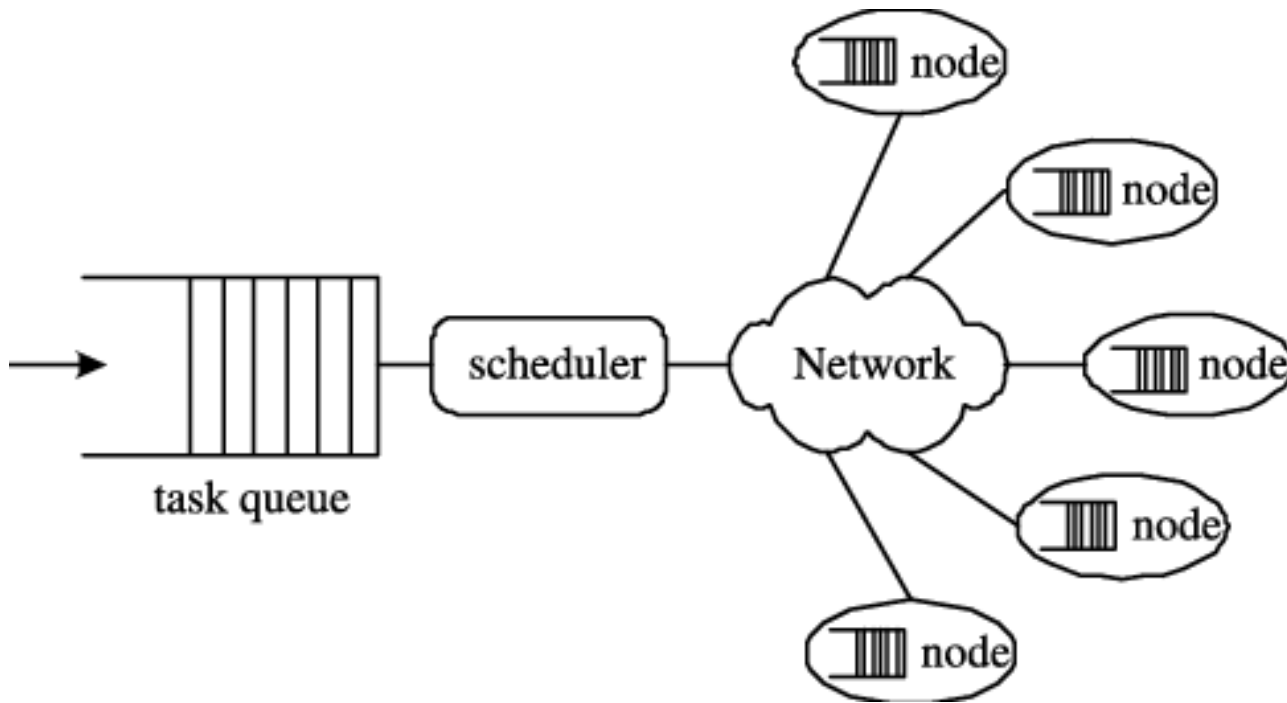
Limitations and Opportunities

- Supercomputers
 - are too expensive
 - still far away from achieving desirable speedups
 - need skilled programming (distributed computing algorithms, parallelizable code)
- But,
 - GPUs are becoming commonplace
 - High Performance Clusters are increasingly available

The Central Question!!!!

**Instead of using supercomputers,
can we put commodity hardware
into a cluster and achieve speedup?**

Computing with Commodity Hardware – Distributed Computing



Sun et al., Dynamic Task Flow Scheduling for Heterogeneous Distributed Computing, 2007.

Cluster Computing

- Multiple nodes acting as a single node
- High Performance Computing (HPC) clusters are becoming increasingly popular



Sun Microsystems, Solaris Cluster

Source: chrisdag, flickr.

Grid Vs. Cluster Computing

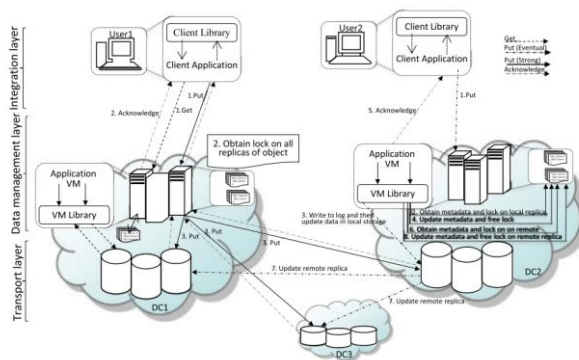
- Clusters have homogenous set of nodes.
- Grid refers to heterogenous systems.

All Roads Lead To... Cloud

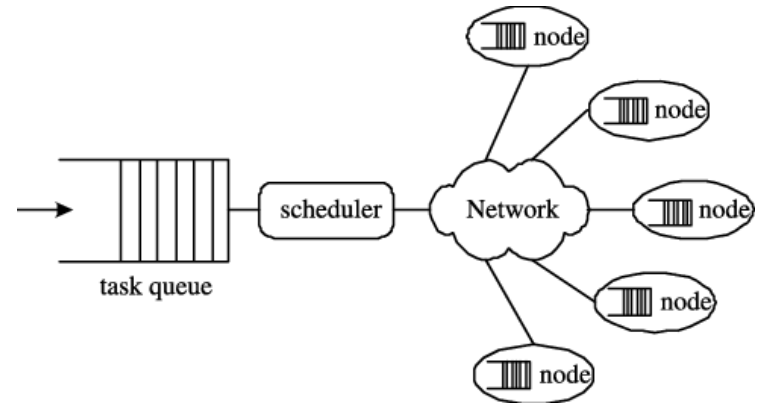
We are in the Big Data era!

Two kinds of Big Data Opportunities






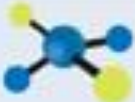
Storage



Processing



Imperatives for Big Data Platform

	Big Data Platform Imperatives		Technology Capability
1	Discover, explore and navigate big data sources		Federated Discovery, Search and Navigation
2	Extreme Performance – run analytics closer to data		Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data		Hadoop File System / MapReduce Text Analytics
4	Analyze data in real time		Stream Computing
5	Rich library of analytical functions and tools		In-Database Analytics Libraries Big Data visualization
6	Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM

Source: <https://www.ibmbigdatahub.com/blog/part-ii-big-data-platform-manifesto>

Key Questions

- How to setup and manage such clusters?
- How to achieve reliability, availability, scalability, ...?
- How to build services on cloud?

Apache Hadoop

Open source platform - reliable, scalable, - distributed processing of large data sets - built on clusters of commodity computers.

Remember...

- Presentation registration deadline is approaching.
- Register yourself on moodle.
- Do not ignore the readings.