Inspire…Educate…Transform.

Big Data

Introduction to Big Data

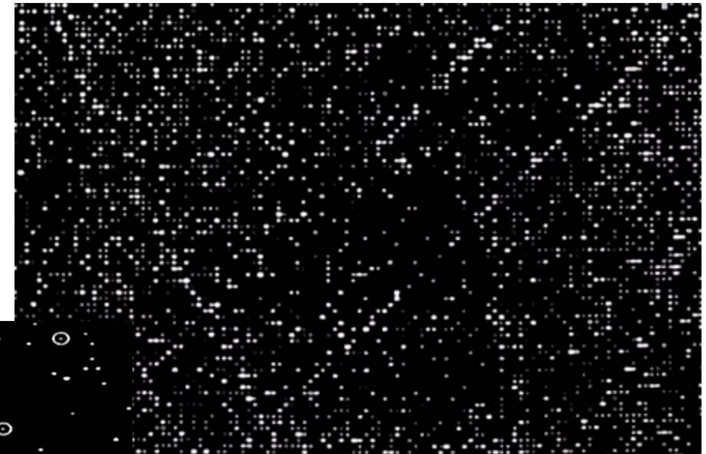
Suryaprakash Kompalli

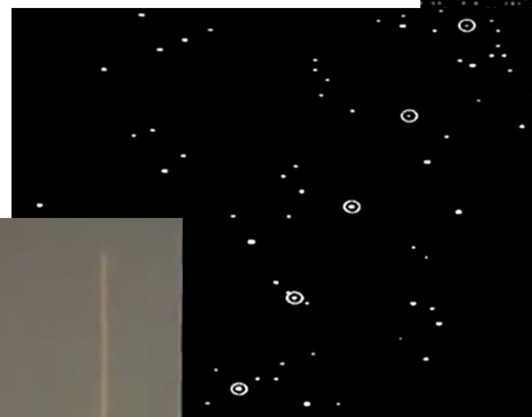
Senior Mentor, INSOFE

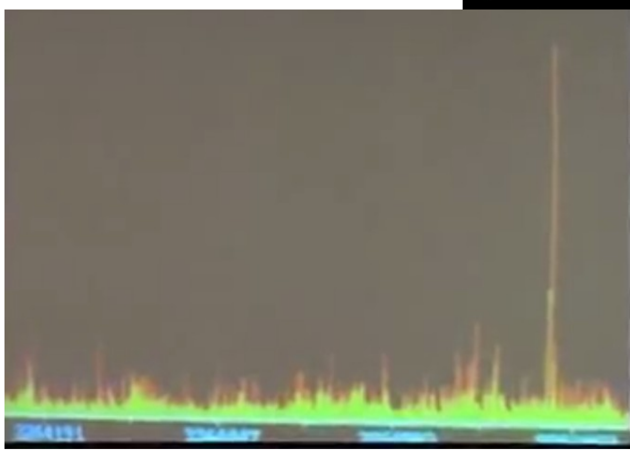
Agenda

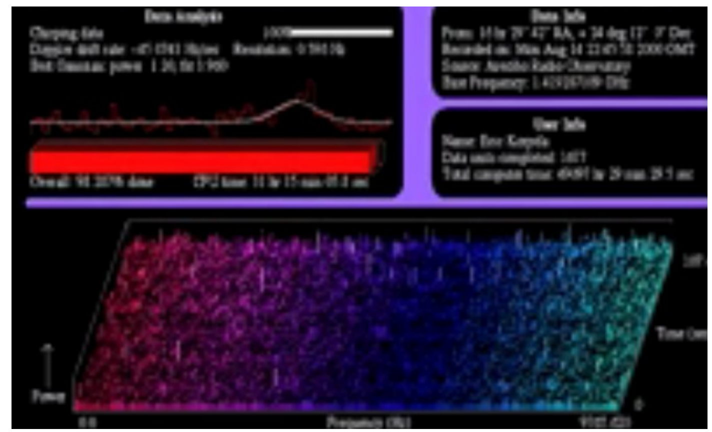
- Different architectures
- Transition from Databases to data warehouses and data lakes
- Thinking of Large Jobs as Task Decompositions
- How BigData is changing IT and business operations

Computing at a Glance: How huge is data, and how was it handled previously?









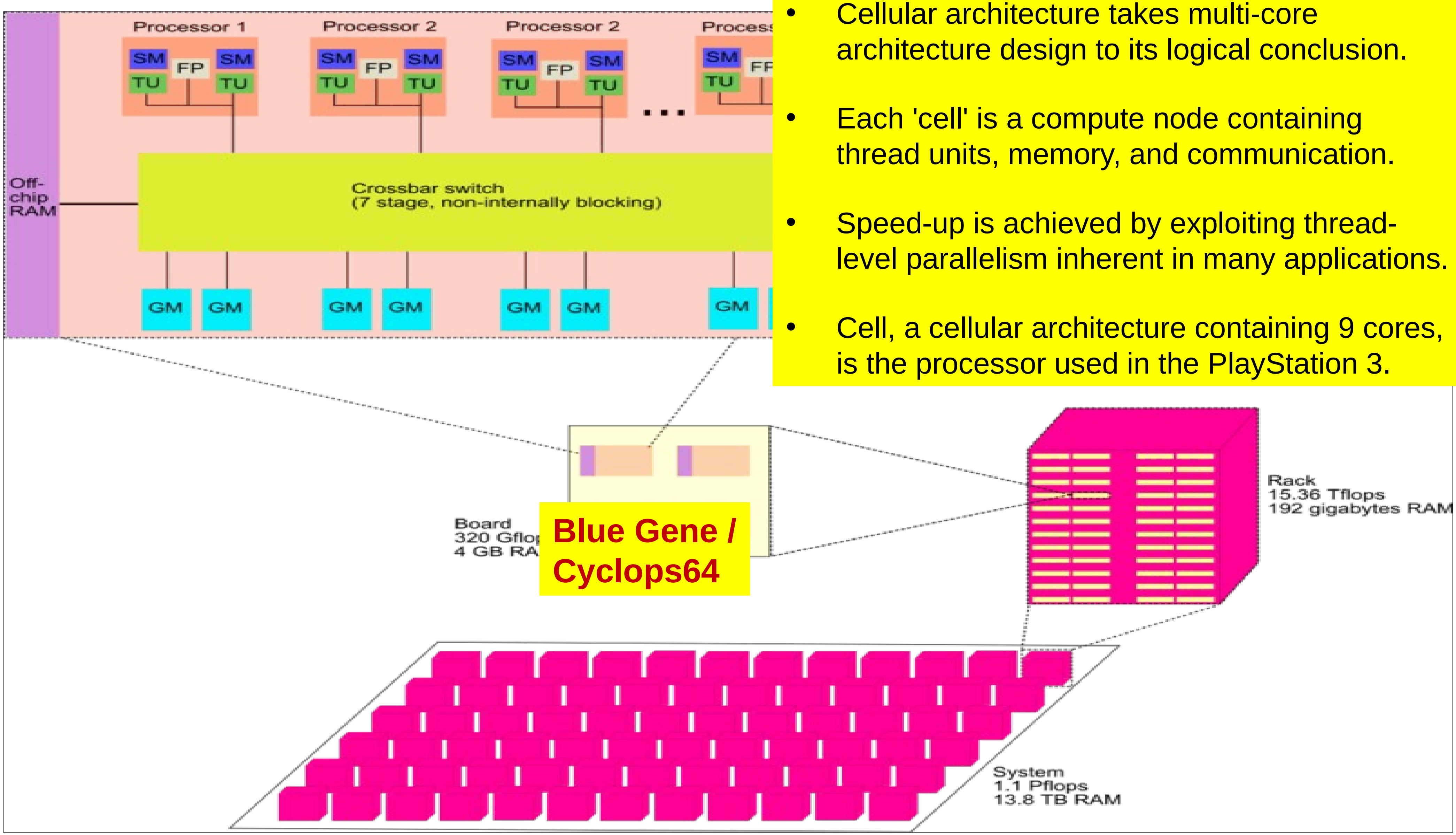
Link to SETI video

Large radio telescope in Arecibo, 100 million signals per second, multiple patterns, SETI@home screensaver on worldwide computers URL: https://www.youtube.com/watch?v=_alJV5aQR68

Connecting up computers is not a new idea.

- Cellular Computing
- Grid Computing
- Cluster Computing
- Cloud Computing

Cellular Computing

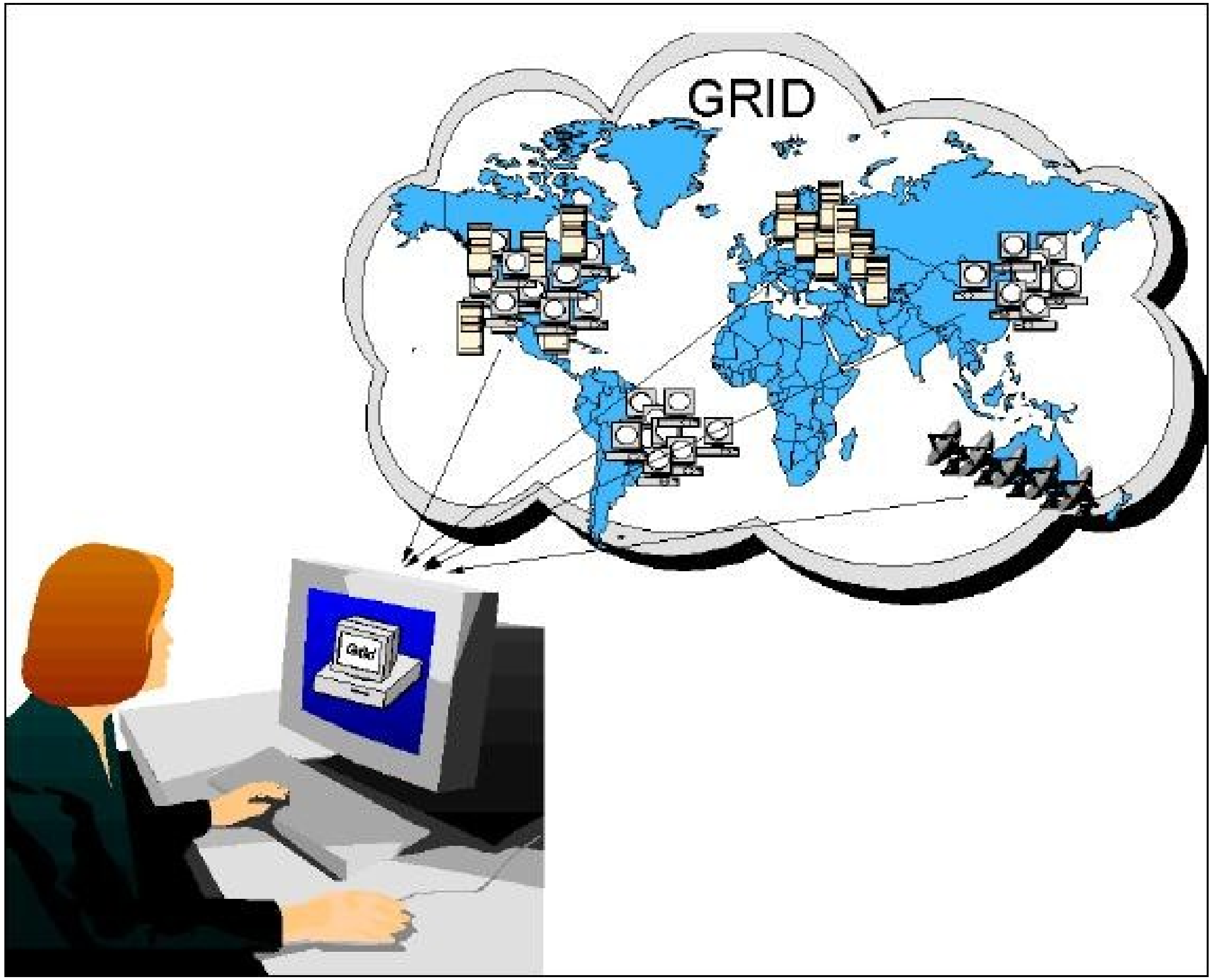


- Cellular architecture takes multi-core architecture design to its logical conclusion.
- Each 'cell' is a compute node containing thread units, memory, and communication.
- Speed-up is achieved by exploiting thread-level parallelism inherent in many applications.
- Cell, a cellular architecture containing 9 cores, is the processor used in the PlayStation 3.

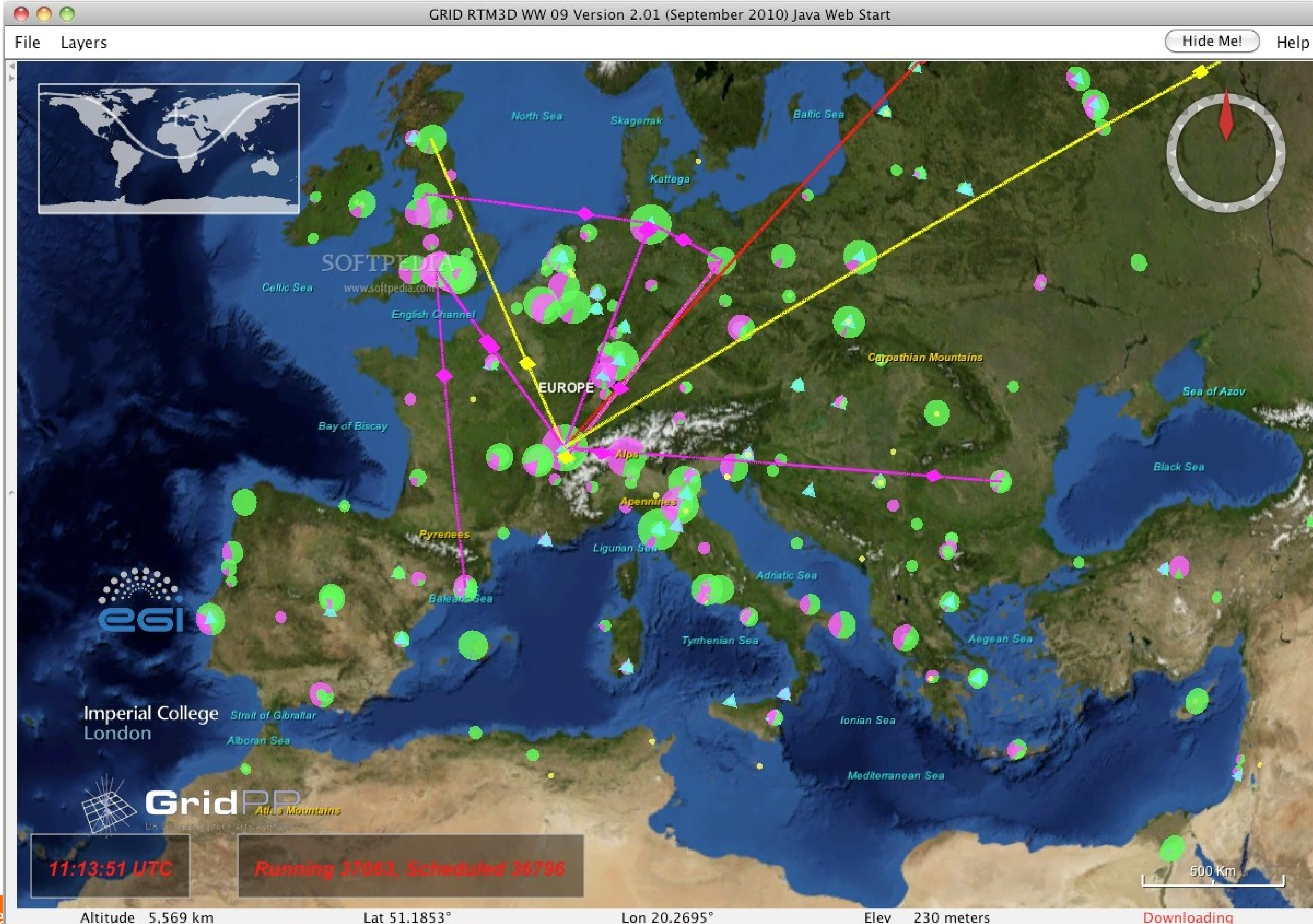
Blue Gene / Cyclops64

Grid Computing

Using *usually* geographically distributed and interconnected computers together for high performance computing *and/or* for resource sharing.



A Sample Grid



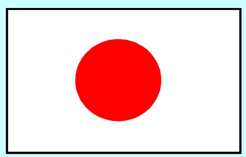
The be

Some Grid Projects & Initiatives

- Australia
  - Nimrod-G
  - Gridbus
  - GridSim
  - Virtual Lab
  - DISCWorld
  - GrangeNet.
  - ..etc

- Europe
  - UK eScience
  - EU Data Grid
  - Cactus
  - XtremeWeb
  - ..etc.

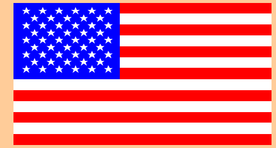
- India
  - I-Grid



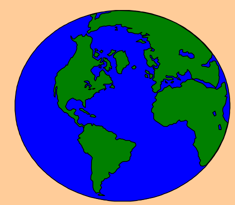
- Japan
  - Ninf
  - DataFarm
- Korea...

  N*Grid
- Singapore

  NGP



- USA
  - AppLeS
  - Globus
  - Legion
  - Sun Grid Engine
  - NASA IPG
  - Condor-G
  - Jxta
  - NetSolve
  - AccessGrid
  - and many more...
- Cycle Stealing & .com Initiatives
  - Distributed.net
  - SETI@Home, ….
  - Entropia, UD, SCS,….
- Public Forums
  - Global Grid Forum
  - Australian Grid Forum
  - IEEE TFCC
  - CCGrid conference
  - P2P conference



Figures due to Rajkumar Buyya, University of Melbourne, Australia, www.gridbus.org

Cluster Computing

- A cluster is a type of parallel or distributed processing system, which consists of a collection of interconnected stand-alone computers cooperatively working together as a single, integrated computing resource.

- A typical cluster:
  - Network: Faster, closer connection than a typical network (LAN)
  - Low latency communication protocols
  - Loose connections

Agenda

- Different architectures
- Transition from Databases to data warehouses and data lakes
- Thinking of Large Jobs as Task Decompositions
- How BigData is changing IT and business operations

What Changed?

1980-1990s

Mainframes
RDBMS
MPI
PCs

1990s-2000

Online applications:

Email, banking, retail, search

Distributed scientific computing

Data ware housing:
Online Analytical Processing (OLAP), BI

2000s-Till Date

Big Data and Cloud Computing
Significant explosion of data:

Desktop applications moved to web (Video/photo editing, office tools, productivity etc.)

SAAS, PAAS, Social and Mobile became pervasive

Data storage exploded

No SQL Dbs, MongoDB, Key-value, Column Store

Hadoop, MapReduce

Visualization: D3.js, Kibana

SOLR, Lucene

Transition from databases to data warehouses to data lakes

Then and Now

Then	Now
Measured on: Flops, or Floating Point Operations Per Second	Measured on: Flops + Data throughput
Longevity of hardware	Ability to use commodity hardware
Efficient batch processing	On-demand scaling of compute and storage, Uptime
Main load: Scientific compute	Main load: Anything !!! (But mostly analytics)
Programming tools: Proprietary, Highly technical	Programming tools: Open source + Proprietary, Retail/End user friendly

Big Data

*"Big Data" is **high-volume**, **high-velocity** and **high-variety** information assets that demand **cost-effective**, innovative forms of information processing for **enhanced insight** and decision making.*

Gartner

Agenda

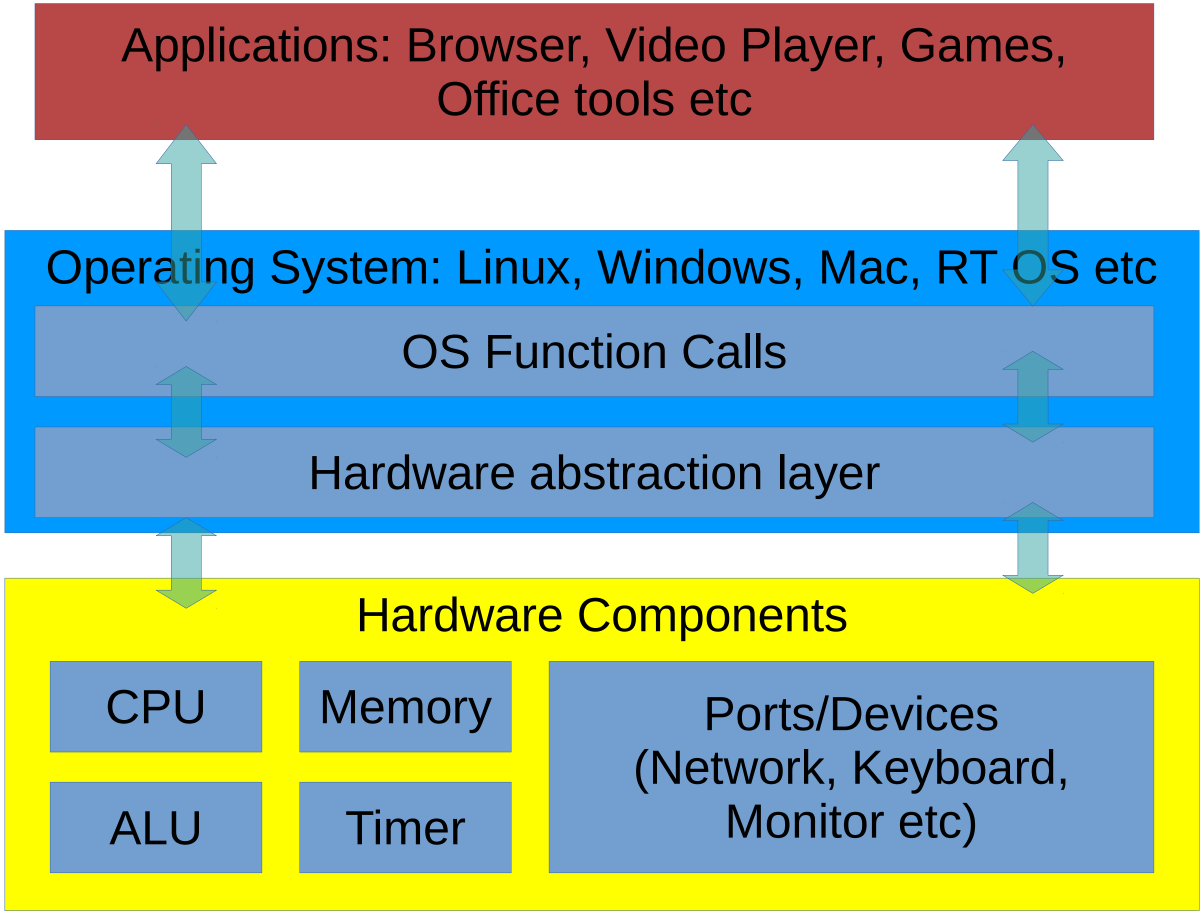
- Different architectures
- Transition from Databases to data warehouses and data lakes
- Thinking of Large Jobs as Task Decompositions
- How BigData is changing IT and business operations

Orders of Magnitude

Name (Symbol)	Value	Binaryusage
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

When individual applications need access to petabytes or more data, and need inexact answers you should explore Big Data solutions

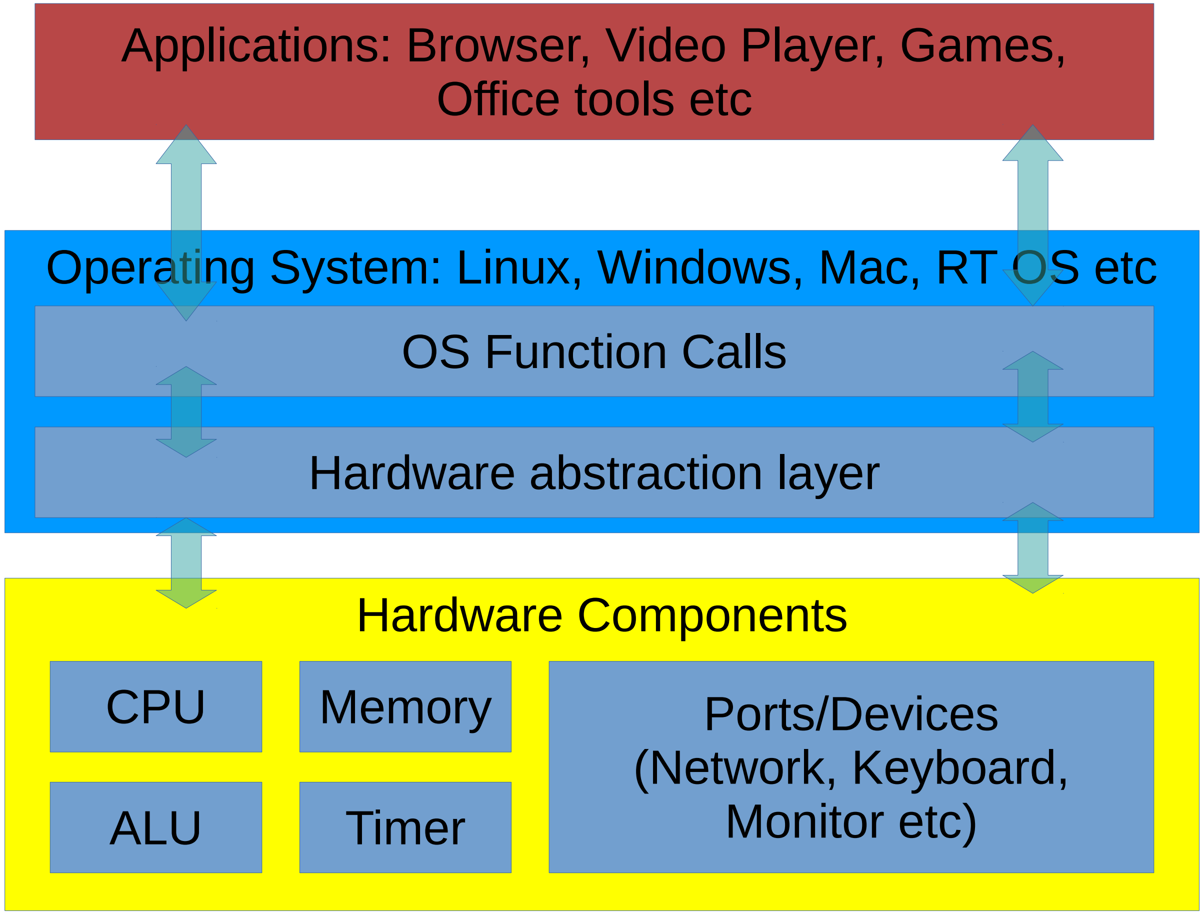
Big Data Tools: Why, really, Why???



Typical Operating System View

https://source.android.com/devices/

Big Data Tools: Why, really, Why???



Typical Operating System View

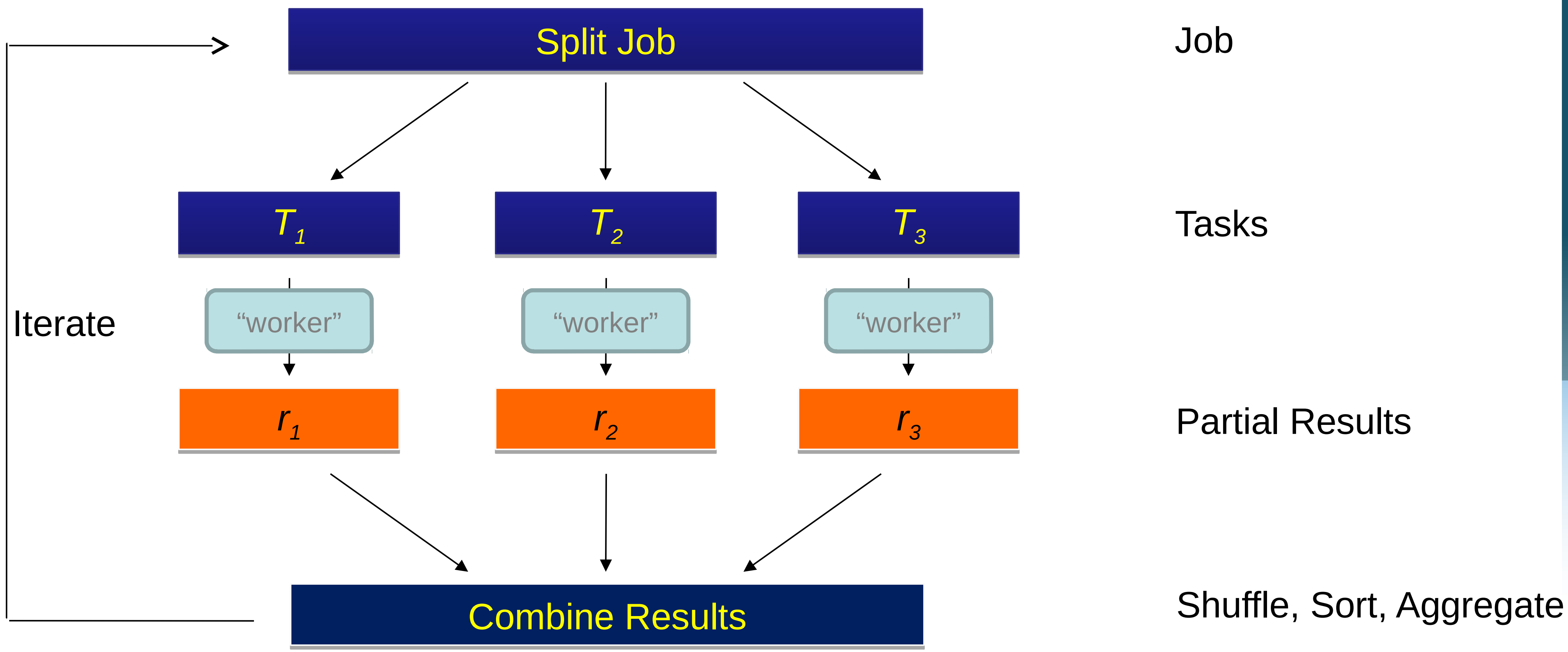
Role of an OS:

1) Application does not have to worry about hardware details. Ex: Different types or processors, Keyboards, Mice, or Graphics cards
2) Failure of hardware is handled gracefully, Ex: If the Graphics card fails or keyboard is not connected, give some beeps
3) Handle multiple applications

Role of a **Big Data** OS; All of the above plus:

1) Handle massively large amounts of data and applications
2) Keep backups
3) Provide system health
4) Support OS function calls that are err... ***complicated***

Map-Reduce


Split Job
Job
T_1
T_2
T_3
Tasks
Iterate
"worker"
"worker"
"worker"
r_1
r_2
r_3
Partial Results
Combine Results
Shuffle, Sort, Aggregate

Dean & Ghemawat, 2004
Map-Reduce: Solving large data problems

Map

- Iterate over a large number of records
- Extract something of interest from each
- Shuffle and sort intermediate results
- Aggregate intermediate results
- Generate final output

Reduce !!!

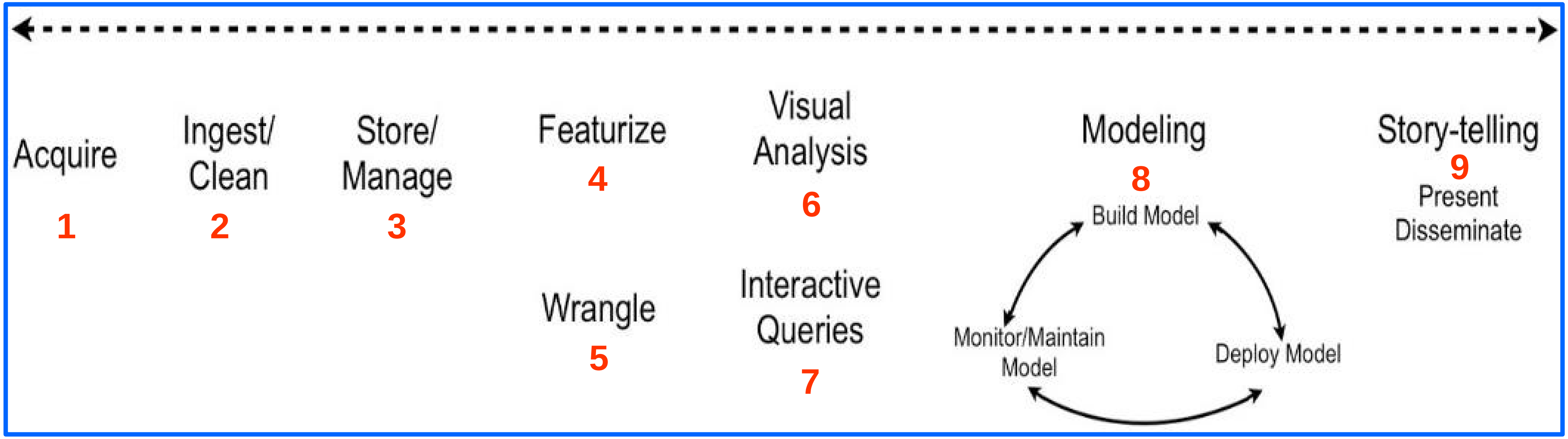
Orders of Magnitude: Sample Problems

- How many “likes” for an asset (photo/video/comment)
  - Approx 1.5 billion users
    - Let us assume 10 assets per user per day (photo/video/comment)
    - 15 billion assets per day.
    - Let us assume 3 likes per day per asset
  - Approx 45 billion counters updated per day and displayed
    - Let us further assume 50 friends may like these contents
  - For every counter you need to store ~50 addresses of “who” liked my asset
  - Assuming an address length of 32 bits, you are already accessing significantly more than a few peta bytes of data
- The feature was first implemented on MySQL, then Cassandra, then Hbase

Orders of Magnitude: Sample Problems

- What broke my system at 1:00 AM Pacific Time?
  - 10-20 million users of a system like office 365
  - A user can do say approx 500 actions: login, access documents, type, close documents, send email etc
  - Number of users performing an action at any given instant ~20K
    - Each action spawns a number of jobs and touches several back end services
  - How can you log these activities?
  - How can you search these logs?
  - Can you identify an outage just before it happens?

The Data Science Workflow



How is each of these stages affected by Big Data?

Large amounts of data

Inexact Queries

You will see these in multiple domains:
Retail, Banking, Manufacturing, Scientific computing

List out all the problems posed by customer

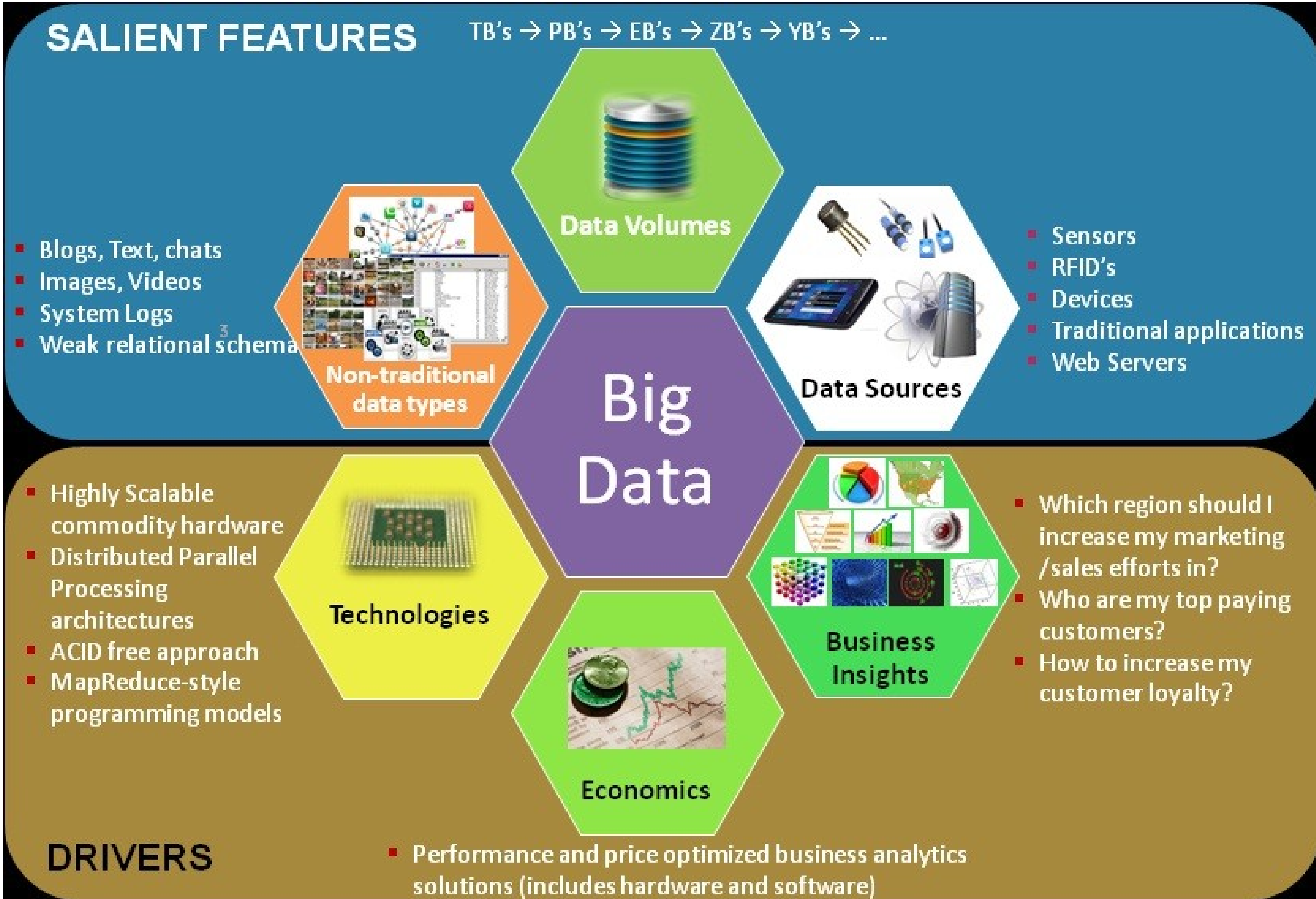
Create an abstract solution for the largest problem in your head using Big Data components

Then, designing solutions for smaller problems would be a cake-walk.

https://www.oreilly.com/ideas/data-analysis-just-one-component-of-the-data-science-workflow

Agenda

- Different architectures
- Transition from Databases to data warehouses and data lakes
- Thinking of Large Jobs as Task Decompositions
- How BigData is changing IT and business operations


SALIENT FEATURES
TB's → PB's → EB's → ZB's → YB's → ...
Data Volumes
Blogs, Text, chats
Images, Videos
System Logs
Weak relational schema
Non-traditional data types
Data Sources
Sensors
RFID's
Devices
Traditional applications
Web Servers
Big Data
Highly Scalable commodity hardware
Distributed Parallel Processing architectures
ACID free approach
MapReduce-style programming models
Technologies
Business Insights
Which region should I increase my marketing /sales efforts in?
Who are my top paying customers?
How to increase my customer loyalty?
Economics
DRIVERS
Performance and price optimized business analytics solutions (includes hardware and software)

Data: CERN

- Located at Meyrin,Switzerland
- 100K cores, 45 Petabytes of data
- Can process 1 PB of data per day
- Experimental data is mainly stored on tape. CERN uses Hadoop for storing the metadata of the experimental data
- Run 1: 30 PB per year. 100,000 processors with peaks of 20 GB/s writing.
- Tapes spread across 80 tape drives. 55,000 tape drives. Robot operated.
- Run 2: > 50 PB per year

Link to CERN video

CERN's Computer Center (1st floor)

https://www.youtube.com/watch?v=S0MgJFGL5jg
http://home.cern/about/computing

How huge is data?

Multiples of bytes		
SI decimal prefixes		Binary usage
Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

YouTube
>Billion users
72 hours of video per minute
Ads, TrueView, Paid channels
~200 – 350K servers

Microsoft Office 365
~10-20 Million users
Online documents/ppts/excel
Paid subscription service
~100K servers

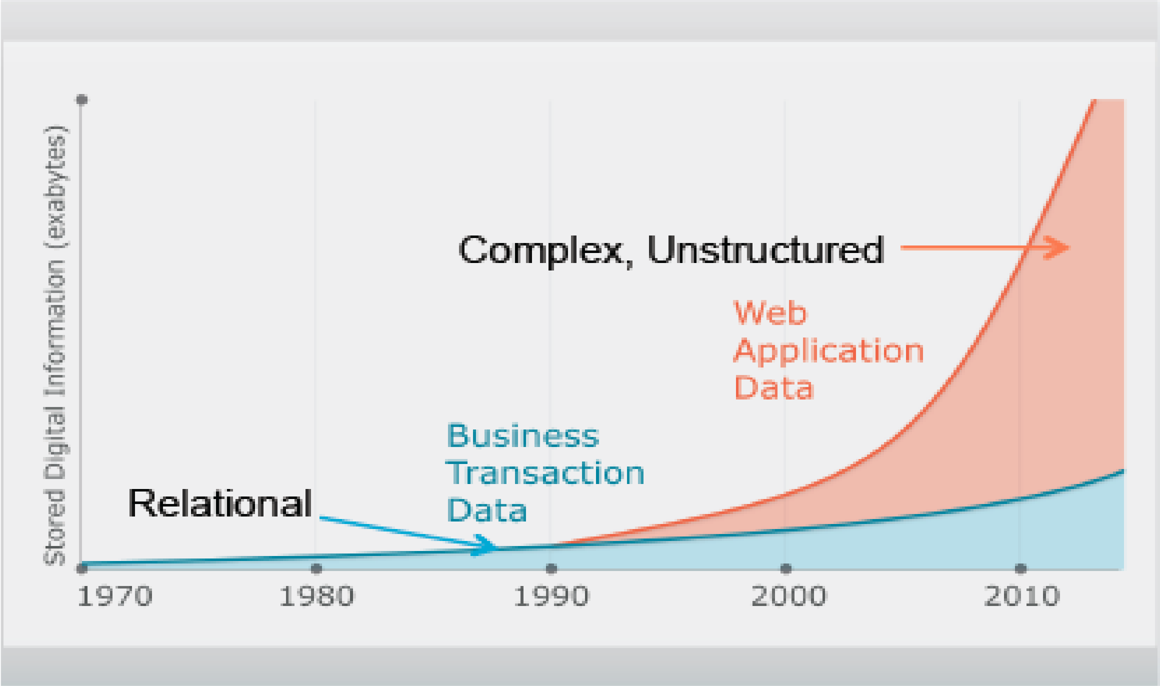
CERN
Worlds biggest machine
LHC has 27 km circumference
30 PetaBytes of data in 2012
~100K servers

http://home.cern/about/computing
https://www.youtube.com/yt/press/statistics.html
http://windowsitpro.com/blog/office-365-numbers-ever-increasing-trajectory

Explosion of Data

- Online
  - Web-ready devices
  - Social media
  - Digital content
  - Smart grids

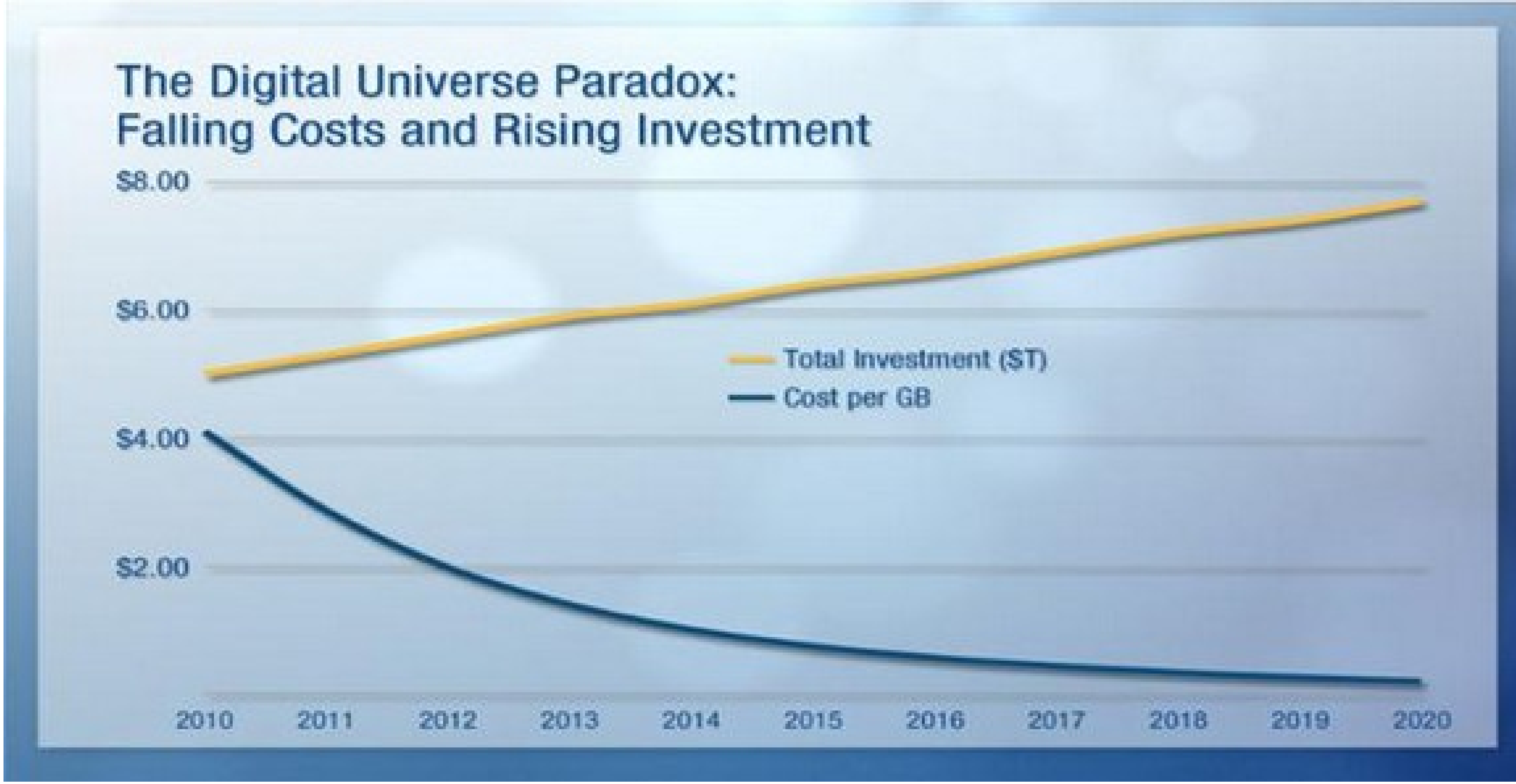
- Enterprise
  - Transactions
  - R&D data
  - Operational (control) data



2,500 exabytes of new information in 2012 with Internet as primary driver

Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 "zettabytes" this year

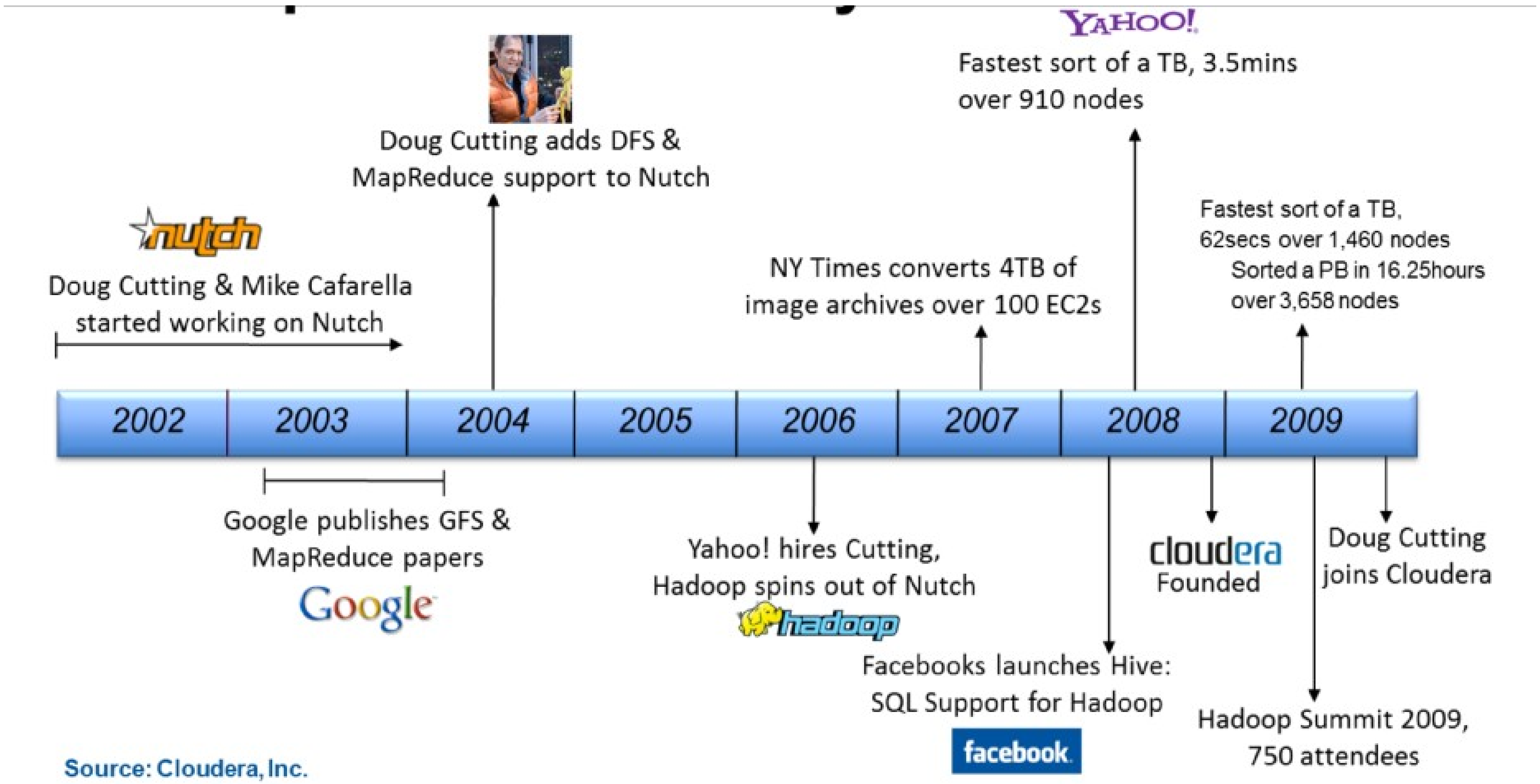
Source: An IDC White Paper - sponsored by EMC. As the Economy Contracts, the Digital Universe Expands. May 2009



Per gigabyte overall cost must come down exponentially for big data to be a reality.


Over the next
decade, the number
of "files,"
or containers
that encapsulate
the information in
the digital universe
.... will grow by
75x
01001
00101
01010
while the pool of
IT staff
available to
manage them
will grow only
1.5x
slightly.

Hadoop: How? The Timeline

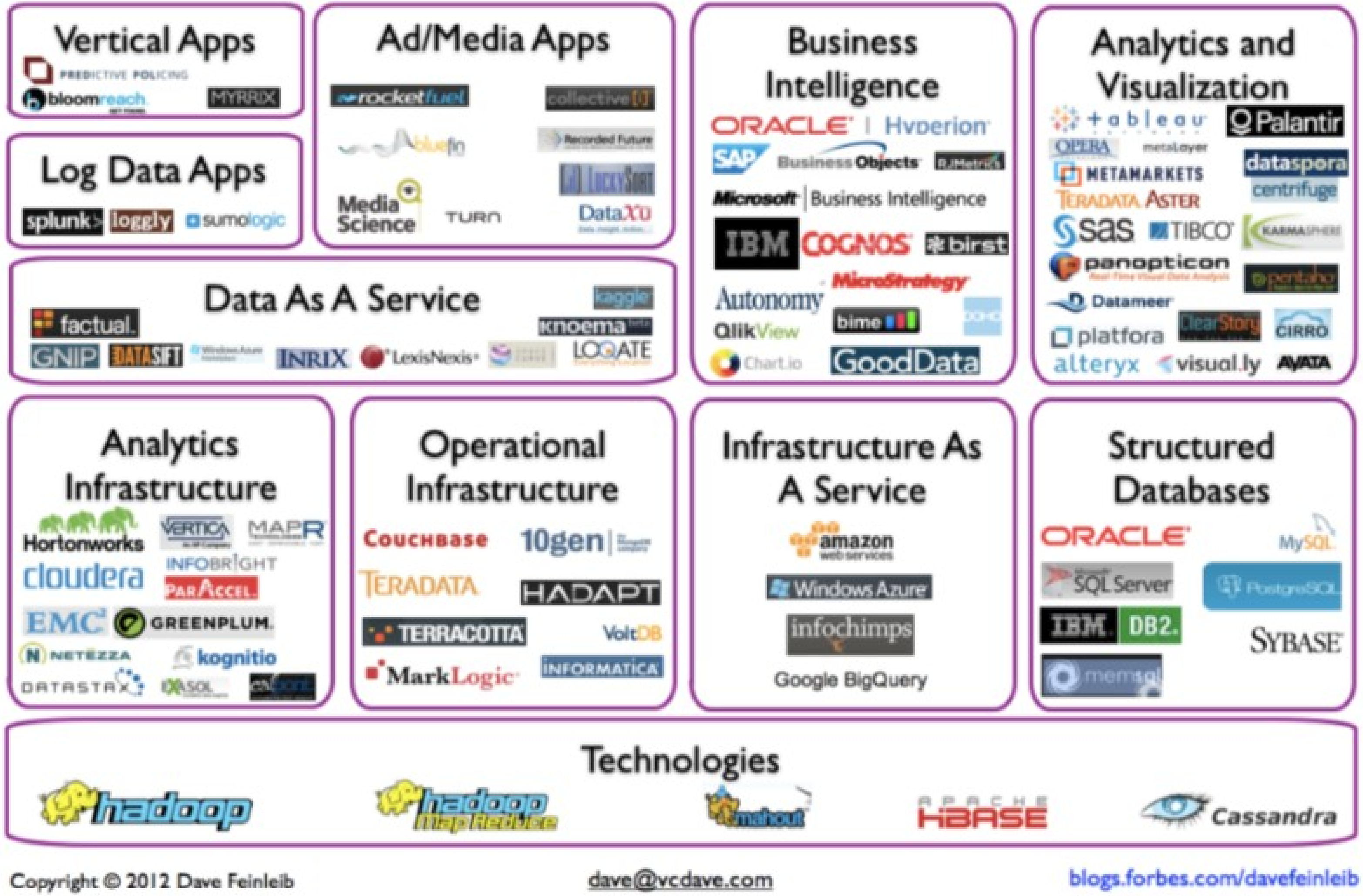


Hadoop: High Level Overview

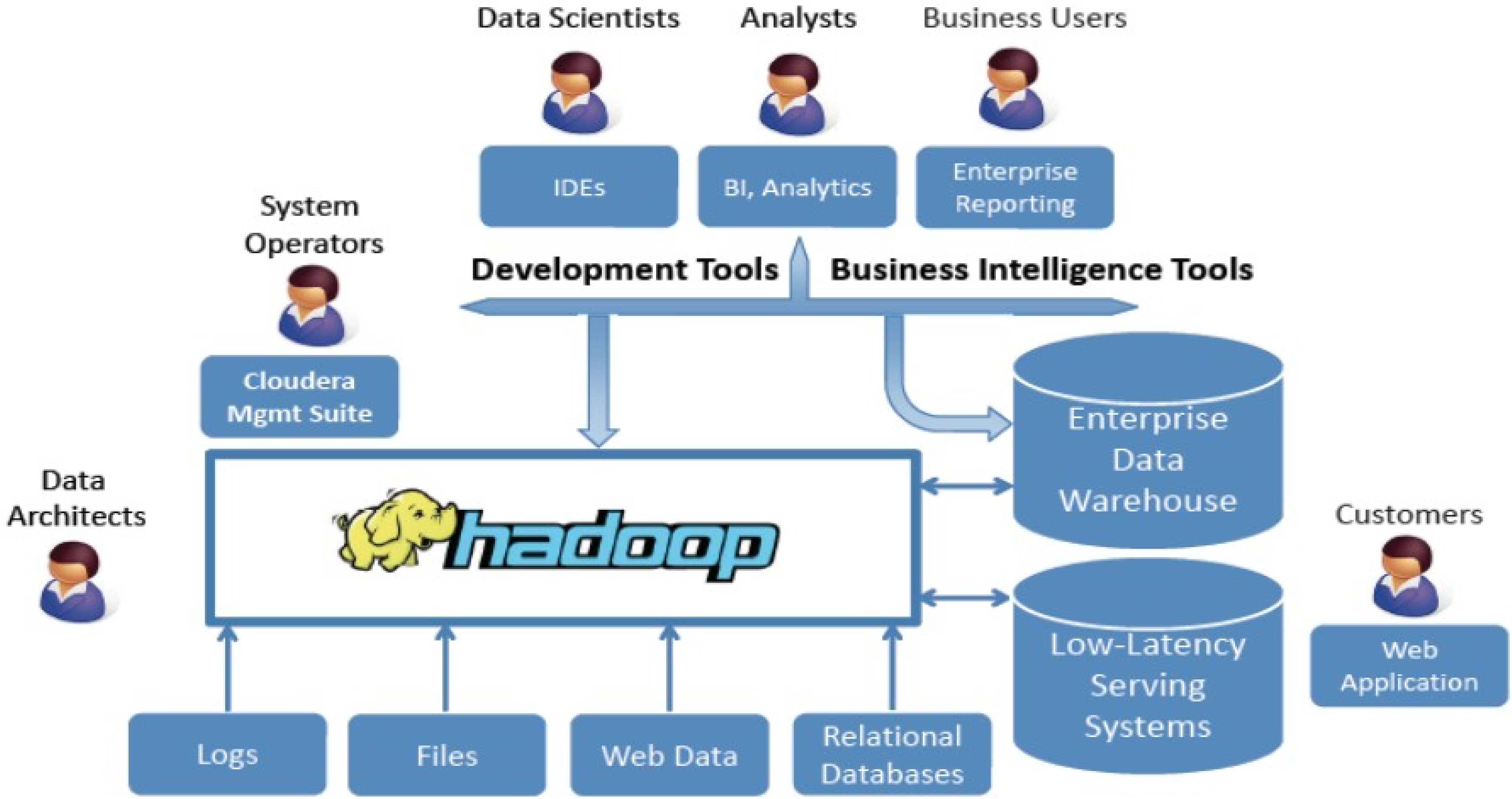
- Open Source Apache Project
  - http://hadoop.apache.org/
- Written in Java
  - Does work with other languages
- Runs on
  - Linux, Windows and more
  - Commodity hardware with high failure rate

The Hadoop Ecosystem

Big Data Landscape



Big Data Enterprise Roles



International School of Engineering

For Individuals:+91-9502334561/63 or 040-65743991

For Corporates:+91-9618483483

Web:http://www.insofe.edu.in

Facebook:https://www.facebook.com/insofe

Twitter:https://twitter.com/Insofeedu

YouTube:http://www.youtube.com/InsofeVideos

SlideShare:http://www.slideshare.net/INSOFE

LinkedIn:http://www.linkedin.com/company/international-school-of-engineering

Additional Material

Task Decomposition
Bulk Synchronous Processing

Task Decomposition

- Example query: select MODEL = ``CIVIC'' AND YEAR = 2001 AND (COLOR = ``GREEN'' OR COLOR = ``WHITE)

ID#	Model	Year	Color	Dealer	Price
4523	Civic	2002	Blue	MN	$18,000
3476	Corolla	1999	White	IL	$15,000
7623	Camry	2001	Green	NY	$21,000
9834	Prius	2001	Green	CA	$18,000
6734	Civic	2001	White	OR	$17,000
5342	Altima	2001	Green	FL	$19,000
3845	Maxima	2001	Blue	NY	$22,000
8354	Accord	2000	Green	VT	$18,000
4395	Civic	2001	Red	CA	$17,000
7352	Civic	2002	Red	WA	$18,000

Table 3.1 A database storing information about used vehicles.

COMP 422, Spring 2008 (V.Sarkar)

https://www.cs.rice.edu/~vs3/comp422/lecture-notes/comp422-lec4-s08-v1.pdf

Task Decomposition

The query can be divided into subtasks in various ways.

Each task generates an intermediate table of entries.

Combining these tables gives the final result

Task (1): 10

ID#	Model
4523	Civic
6734	Civic
4395	Civic
7352	Civic

Task (2): 10

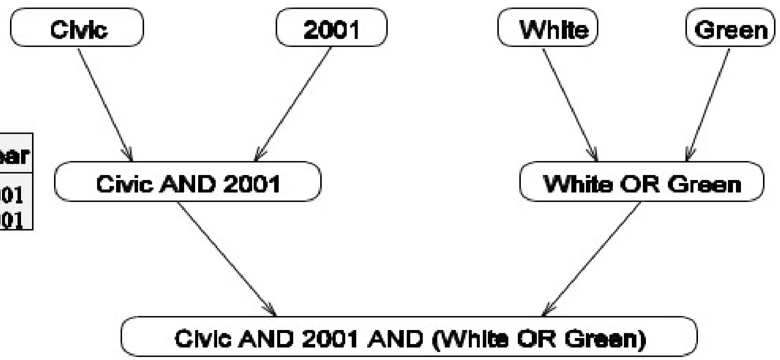
ID#	Year
7623	2001
6734	2001
5342	2001
3845	2001
4395	2001

Task (3): 10

ID#	Color
3476	White
6734	White

Task (4): 10

ID#	Color
7623	Green
9834	Green
5342	Green
8354	Green



Task (5): 6

ID#	Color
3476	White
7623	Green
9834	Green
6734	White
5342	Green
8354	Green

Task (6): 9

ID#	Model	Year
6734	Civic	2001
4395	Civic	2001

Task (7): 8

ID#	Model	Year	Color
6734	Civic	2001	White

Given a data-set and sub-tasks, one can identify the minimum number of data elements you need to visit to complete the task.

This is indicated on the graph in bold.

For example, sub task 1 requires you to find out which car is a civic. This task will require visiting all 10 records in our sample table

Subtask 6 requires you to access data from task 1 and 2, that means 9 records

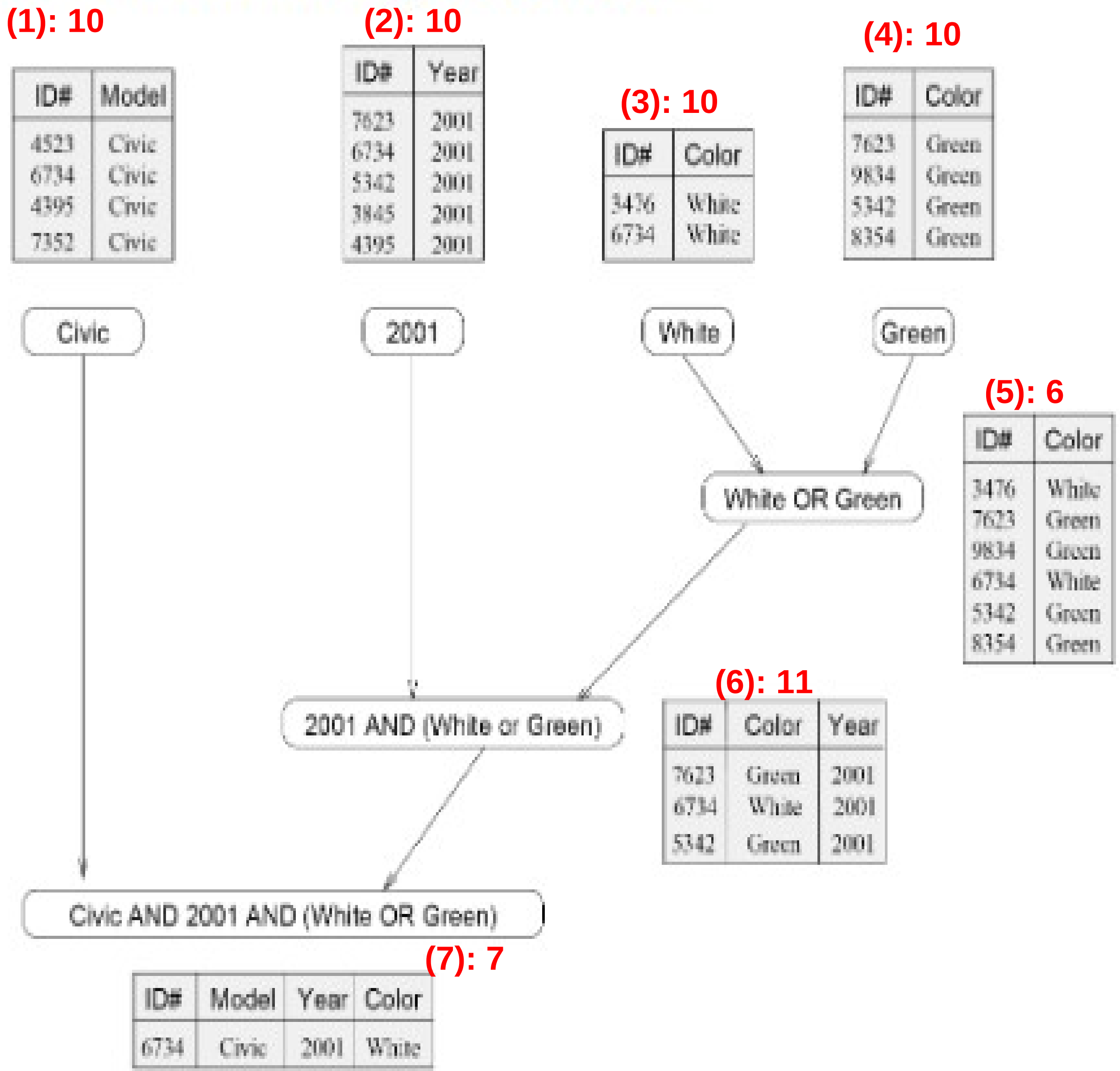
Task Decomposition

This sequence of tasks that must be processed one after the other can be visually shown as a directed graph, called task dependency graph

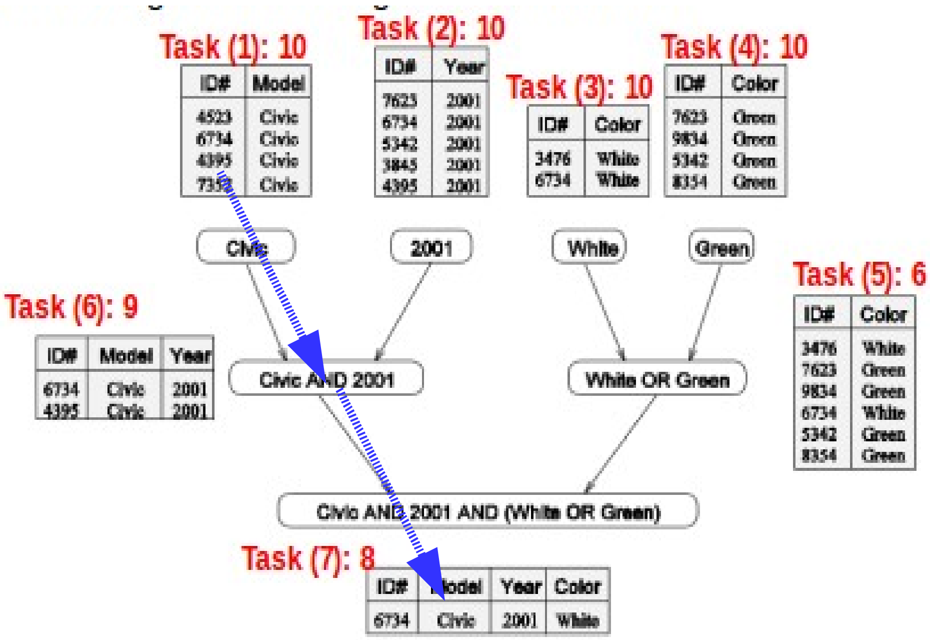
The longest path in this graph determines the shortest time in which the program can be executed in parallel.

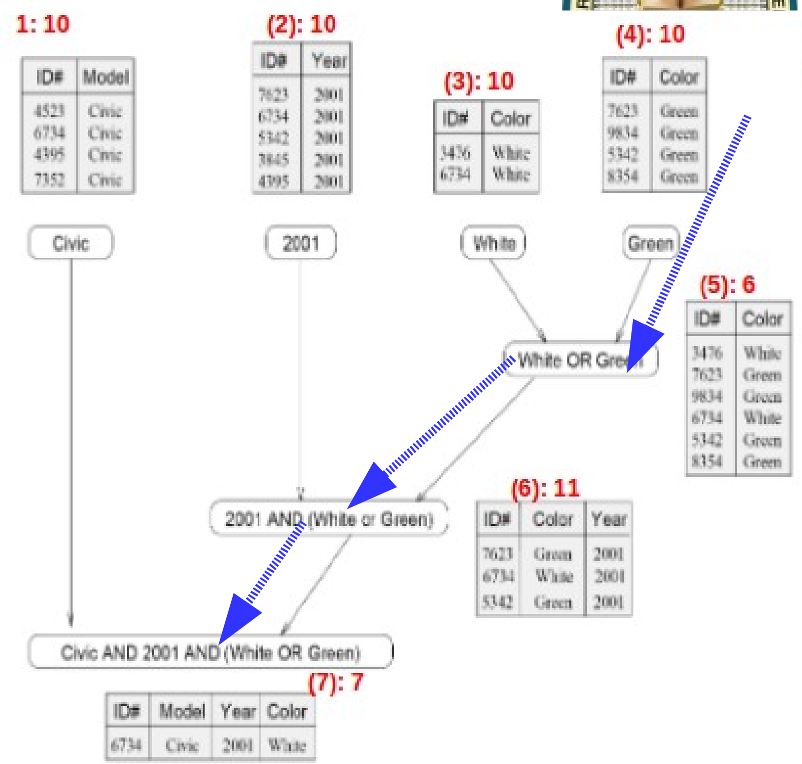
The length of the longest path in a task dependency graph is called the critical path length

The ratio of the total amount of work to the critical path length is the average degree of concurrency



Task Decomposition





Total work (T): 63
Critical path (Blue arrows) (Tc): 27
Avg concurrency: 63/27 = ~2.3

Total work: 64
Critical path (Blue arrows): 34
Avg concurrency: ~1.9

If task is to be run on “p” processors we can show the max and min time needed for execution to be:
Upper bound time complexity: Tparallel <= (T/p) + Tc
Lower bound: Tparallel >= (T/p), Tc

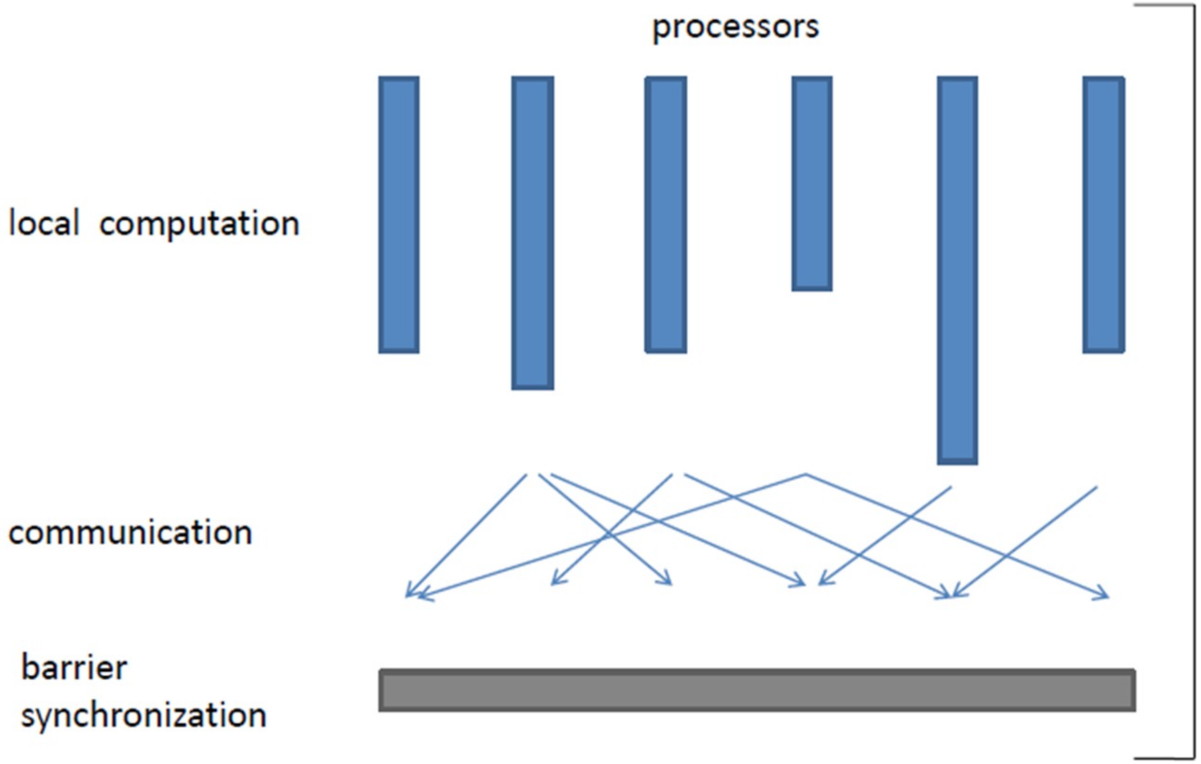
Assumes that additional tasks not required to distribute tasks on “p” processors

Task Decomposition

- How to perform task decomposition?
  - recursive decomposition: Algorithms that use divide-conquer like merge sort or quick sort
  - data decomposition: Item-set operations (induction), matrix operations
  - exploratory decomposition: Multi-option search (Chess)
  - speculative decomposition: Used in branch prediction
  - Hybrid?

http://parallelcomp.uw.hu/ch03lev1sec2.html
http://suif.stanford.edu/papers/lam92/subsection3_2_1.html

Bulk Synchronous Processing (BSP)



In BSP, a task is broken down into supersteps that can run on a node

Barrier ensures that messages are sent to everyone only when all nodes finish the super steps

- Developed in the 1990s
  - Parallel local computation
  - Synchronized peer to peer communication

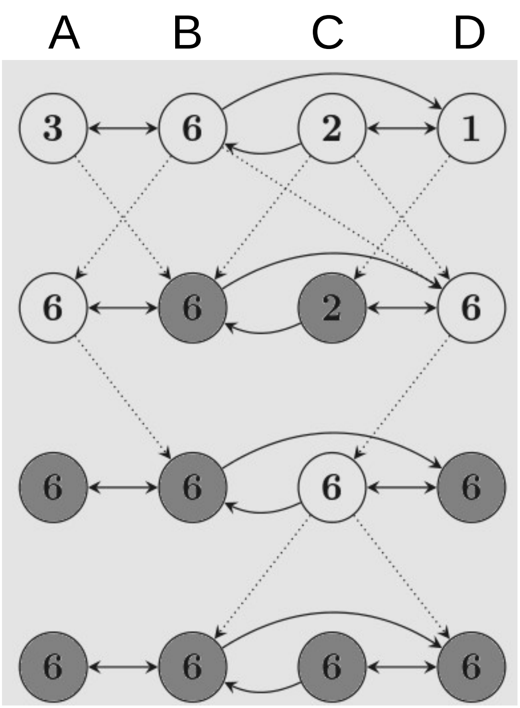
Example (max of numbers, matrix multiplication): http://sbrinz.di.unipi.it/~peppe/FilesPaginaWeb/BSP.pdf

https://en.wikipedia.org/wiki/Bulk_synchronous_parallel

Bulk Synchronous Processing (BSP)

- Input:
  - Directed edge
  - Each vertex associated with id and value
  - Edges may also have values
- Edges have no computation
  - Vertices may modify its value, active/halt state or edges
- Computation ends when all vertices reach halt state

Active vertex   Halted vertex



Superstep 0
D gets message from B, C
C from D
B from C, A
A from B
B and C decide they are large A and D change their values

Superstep 1
B, D, A decide they are max, C changes value

Superstep 2
B, D, A, C decide they are max

Superstep 3
Everyone is in halt stage, read out max value

Maximum value example, dotted lines indicate messages, dark lines indicate edges

https://kowshik.github.io/JPregel/pregel_paper.pdf
https://www.cs.duke.edu/courses/spring13/compsci590.2/slides/lec14.pdf

Bulk Synchronous Processing (BSP)

- Example: Page rank
- Note: In map-reduce Task Trackers cannot talk to each other
- BSP (HAMA) allows you to do that

```
class PageRankVertex
    : public Vertex<double, void, double> {
 public:
  virtual void Compute(MessageIterator* msgs) {
    if (superstep() >= 1) {
      double sum = 0;
      for (; !msgs->Done(); msgs->Next())
        sum += msgs->Value();
      *MutableValue() =
          0.15 / NumVertices() + 0.85 * sum;
    }

    if (superstep() < 30) {
      const int64 n = GetOutEdgeIterator().size();
      SendMessageToAllNeighbors(GetValue() / n);
    } else {
      VoteToHalt();
    }
  }
};
```

https://prezi.com/tabqzlvzohii/apache-hama-introduction/
http://arasan-blog.blogspot.in/

Bulk Synchronous Processing (BSP)

HDFS

BSP

HAMA

Suitable for iterative tasks

Map-Reduce

BSP

Giraph

Primarily for graph processing

https://prezi.com/tabqzlvzohii/apache-hama-introduction/
http://www.hadoopsphere.com/2015/06/large-scale-graph-processing-with.html
http://arasan-blog.blogspot.in/

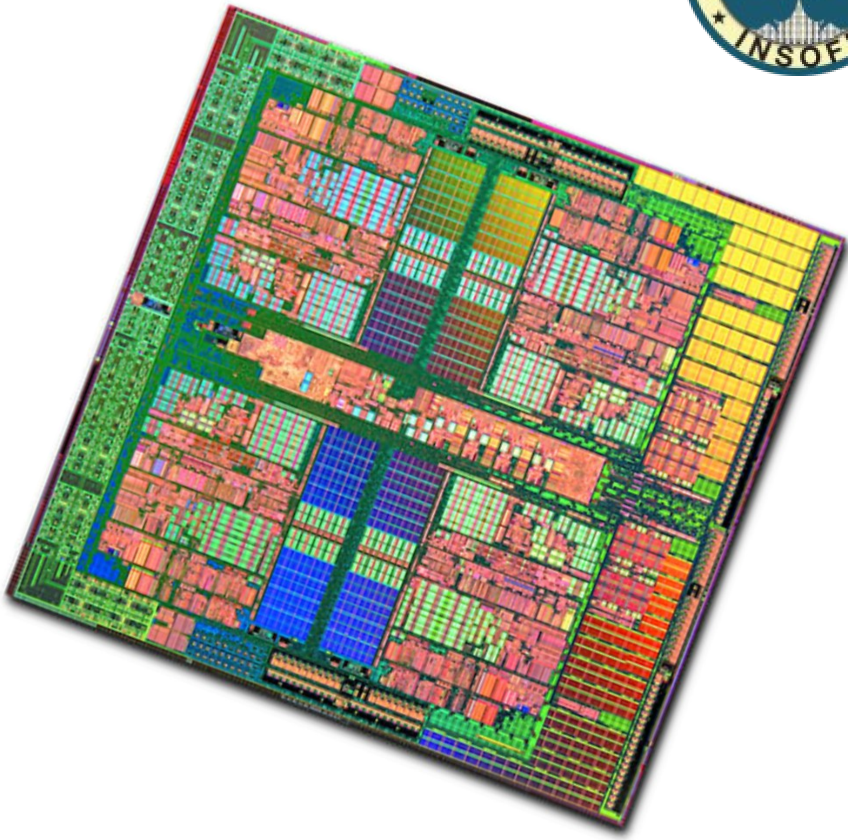
Google Data Center Video

Inside a Google data centre:

https://www.youtube.com/watch?v=XZmGGAbHqa0

Inside a Google data center-XZmGGAbHqa0

A Server Farm



Left: Server Farm, Right: Opteron 4-core processor

https://upload.wikimedia.org/wikipedia/commons/thumb/3/34/Quad-Core_AMD_Opteron_processor.jpg/330px-Quad-Core_AMD_Opteron_processor.jpg

Doug Cutting Basics of Hadoop Video

https://www.youtube.com/watch?v=0GOxDBR6VAU

Time is Important

- Common crashes and lessons:
  - http://highscalability.com/blog/2012/3/14/the-azure-outage-time-is-a-spof-leap-day-doubly-so.html
  - https://azure.microsoft.com/en-us/blog/summary-of-windows-azure-service-disruption-on-feb-29th-2012/
  - https://www.groovehq.com/blog/downtime
  - http://www.evolven.com/blog/downtime-outages-and-failures-understanding-their-true-costs.html
  - http://www.wired.com/2012/07/leap-second-glitch-explained/