HPC meets Big Data

Dr. Putchong Uthayopas
Department of Computer Engineering,
Faculty of Engineering, Kasetsart University
Email: putchong@ku.th
Brief History of HPC Platform

- HPC on single processor
- HPC using Vector machine
- HPC with SMP, SIMD
- MPP and Cluster Computing
- GPU computing
- Heteronomous computing

Mostly build for compute-intensive application!
Growth of Supercomputing

Doubling Time = 1.46 years
Emergence of a Fourth Research Paradigm

1. Thousand years ago – Experimental Science
   – Description of natural phenomena
2. Last few hundred years – Theoretical Science
   – Newton’s Laws, Maxwell’s Equations...
3. Last few decades – Computational Science
   – Simulation of complex phenomena
4. Today – Data-Intensive Science
   – Scientists overwhelmed with data sets from many different sources
   - Data captured by instruments
   - Data generated by simulations
   - Data generated by sensor networks

   eScience is the set of tools and technologies to support data federation and collaboration
   - For analysis and data mining
   - For data visualization and exploration
   - For scholarly communication and dissemination

(With thanks to Jim Gray)
Data-intensive Research

Data

- Acquisition & modelling
- Archiving and preserving
- Collaboration and visualisation
- Dissemination & sharing
- Analysis & data mining

fourthparadigm.org

Science@Microsoft
The Fourth Paradigm in Practice
We are using Big Data All the time

• How can google map know about the traffic condition?
Facebook Usage Statistics (March 2016)

• 1.09 billion daily active users
• 989 million mobile daily active users
• 1.65 billion monthly active users
• 1.51 billion mobile monthly active users

### Data processed per day

<table>
<thead>
<tr>
<th>Organization</th>
<th>Est. amount of data processed per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>100 pb</td>
</tr>
<tr>
<td>Baidu</td>
<td>10-100 pb</td>
</tr>
<tr>
<td>NSA</td>
<td>29 pb</td>
</tr>
<tr>
<td>Facebook</td>
<td>600 Tb</td>
</tr>
</tbody>
</table>

4 more rows

[Data size estimates | Follow the Data - WordPress.com](https://followthedata.wordpress.com/2014/06/24/data-size-estimates/)
Gartner: 21 Billion IoT Devices To Invade By 2020

Gartner indicates the market for Internet of Things devices is poised to explode and will reach nearly 21 billion connected devices by 2020.

Aside from connected vehicles, a number of different consumer uses will continue to account for the greatest number of Internet-connected devices, fueling a booming market for the Internet of Things. At the same time, enterprises will account for the largest spending on these devices, according to a Nov. 10 Gartner report.

Google, Tesla, Nissan: 6 Self-Driving Vehicles
DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.

Sequencers begin giving flurries of data

Terabases

200
160
120
80
40
0
2004
2006
2008
2010
2012

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

Extremely powerful computers are needed to help biologists to handle big data traffic jams.

BY YVONNE MARES

Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and interpreting information that were once the domain of computer scientists.

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why cancer cells thrive, what microbes dwell where in human body cavities and how the genetic makeup of different cancer influences how cancer patients fare. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world’s largest biology-data repositories, currently stores 16 petabytes (1 petabyte is 1015 bytes) of data and looks ahead about genes, proteins and small molecules that scientists believe contribute to diseases.

This data set is just one-tenth the size of the data stored at CERN, Europe’s particle-physics laboratory near Geneva, Switzerland. Every year, particle-collision events at CERN’s Large Hadron Collider generate around 7 petabytes of data — the equivalent of about 4 million high-definition feature-length films. But the EBI and institutes like it face similar data-storing challenges at a fraction of CERN’s scale.

The EBI and its colleagues are regularly meet with computer scientists at the European Space Agency (ESA) in Paris to exchange lessons about data storage, analysis and sharing. All sides need to manipulate data to yield research answers. As prices drop for high-throughput instruments such as next-generation sequencing, the challenge is to store, analyze and share large volumes of data.
Google Open Image Dataset

- The Open Images Dataset
- YouTube-8M Dataset
- Google Books Ngrams
- Google Trends Datastore

New big data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time (*schema-on-read*).
- Apps and users interpret the data as they see fit
What is Data Science?

• Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured.

K-Mean on iris data

Iris setosa

Iris versicolor

Iris virginica

Ref: https://en.wikipedia.org/wiki/Iris_flower_data_set

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td><em>l. setosa</em></td>
</tr>
</tbody>
</table>

Iris data: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

SparkML kmean app: https://github.com/apache/spark/tree/branch-1.5/examples/src/main/python/mllib
Ex. Iris K-mean

Iris Dataset, KMeans clustering with K=3

Kaggle

• Titanic competition
  – What is the factor involve in surviving Titanic

• Data set of the passengers has been provided

https://www.kaggle.com/c/titanic
Deep Learning

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, partially supervised or unsupervised.

Deep learning - Wikipedia

What is Deep Learning?

- The modern reincarnation of Artificial Neural Networks from the 1980s and 90s.
- A collection of simple trainable mathematical units, which collaborate to compute a complicated function.
- Compatible with supervised, unsupervised, and reinforcement learning.
New Technology to Handle Big Data and Machine learning

• Hadoop/Spark Ecosystem
• GPU system
• GPU Cluster
• AI supercomputer using dense GPUs
A DECADE OF SCIENTIFIC COMPUTING WITH GPUs

- **2006**: Stream Processing @ Stanford
- **2008**: CUDA Launched
- **2010**: Oak Ridge Deploys World’s Fastest Supercomputer w/ GPUs
- **2012**: AlexNet beats expert code by huge margin using GPUs
- **2014**: Stanford Builds AI Machine using GPUs
- **2016**: World’s First 3-D Mapping of Human Genome
- **2016**: Google Outperform Humans in ImageNet
- **2016**: GPU-Trained AI Machine Beats World Champion in Go
- **2016**: World’s First Atomic Model of HIV Capsid

---

**IMAGENET**

- Discovered How H1N1 Mutates to Resist Drugs
HEAD IN THE CLOUDS
In cloud computing, large data sets are processed on remote Internet servers, rather than on researchers’ local computers.
Summary

• Scientific Research is rapidly changing to Data Intensive Research
  – Driving by Big data analytics and Machine learning

• Innovative Platform is needed that put data storage, and very high computing power in one place
  – Hadoop move computing to data not traditional data move to computing

• Everything going to the CLOUD