EE 542 Review Session for Mid-Term Exam

Prof. Kai Hwang, October 11, 2017

Exam Period:

2 pm – 3:20 pm (80 minutes)
Wednesday, October 18, 2017

Exam Venue:

WPH Room 101 for 20 students with family names from Avinash to Kumar in alphabetical order.

WPH Room B27 for 45 students with family names from Kumaraswamy to Zhang in alphabetical order.

Format, Rules and Coverage:

- The mid-term is a close-book and close-note exam.
- No wireless phone or notebook computer are allowed.
- Bring your own calculator and erasers. No borrow from other students during the exam.
- There are 7 problems. 15% for mapping 30 key Terms, 13% for 13 multiple-choice questions.
- Remaining 5 problems (72%) cover Chapters 1 ~ 5, 8 and 10
- The coverage: Lectures 1 ~ 16, including the AWS project specification and HW Sets 1 ~ 2 solutions

Overview of 15 Lectures in 6 Chapters

- Chapter 1 on HW/SW background, distributed system models ranging from clusters to datacenters in Lectures 1 and 2.
- Chapter 2 is devoted to big data characteristics, software tools, and the Internet of Things (IoT) in Lectures 9 and 10
- Chapter 3 on virtualization, hypervisors, Docker engine, VMs, and Containers (Lectures 6, 7)
- Chapter 4 introduces cloud platforms and service models, such as AWS, GAE, MS Azures, Salesforce clouds, IBM SmarCloud, HP Helion, SGI Cyclone, etc. (Lectures 3, 4)
- Team Project was introduced in Lecture 5 along with benchmarks
- Chapter 8 covers MapReduce, Hadoop and Spark in Lectures 6 and 7.
- Chapter 10 on Cloud Performance in Lectures 14 and 15

Suggestions To Review:

- All lecture slides must be reviewed, some slides carry updated information from the book.
- Pay special attentions to key concepts, models, abbreviated key terms, and tabulations which summarizes clouds, services, HW/SW, and programming tools.
- A sample mid-term exam wil;l be handout to students who come to attend the review. The exam format and problem style are shown. Of course, the questions differ this year.
- You need to read 6 Chapters covered in the classes during the first 7 weeks in 15 lectures.
- Reading assignments were identified in the lectures. Those sections that you can skip are given in the next slide.

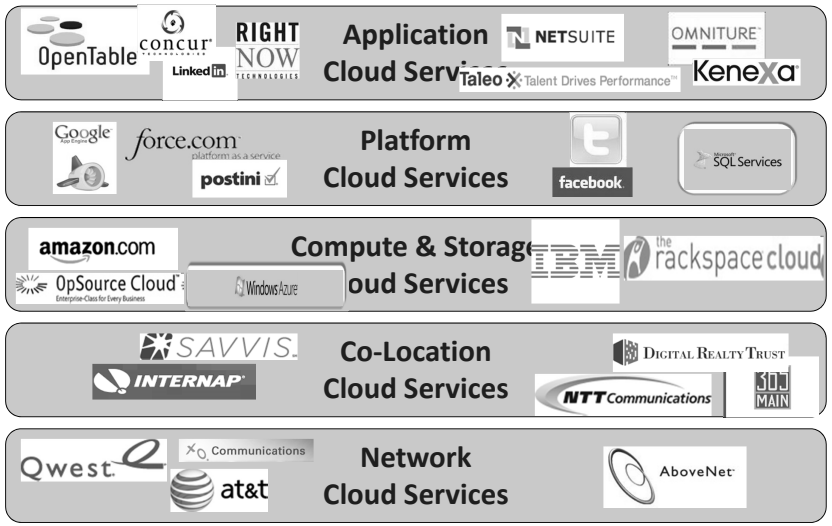
Preparation Suggestions:

- All 15 lecture slides must be reviewed including the project specification in Lecture 5.
- You need to read 7 Chapters (1~5, 8, 10) in the book, but sections I did not cover in lectures can be ignored
- You can skip reading the following sections: 1.4.3, 1.4.4, 2.4, 5.3, 5.4, 10.4, and 10.5
- Review Homework Solutions. You should correct errors or read alternate solutions.
- No make-up exam, if you miss the exam. Arrive at the exam room 10 minutes early to get seated properly.
- Graded exam papers will be returned in class on Oct. 23.

Important Key Concepts Related To Cloud Technology and Platforms

- **Basic Service Models (IaaS, PaaS, SaaS)**
- **Case Studies of many Public Cloud Platforms**
- **Virtualization Techniques (VM, Containers and Unikernel Approaches)**
- **MapReduce, Haddop and Spark Programming on Clouds**

Today's Cloud Services Stack



(Courtesy of T. Chou, 2010)

Market Share of Various Cloud Platforms

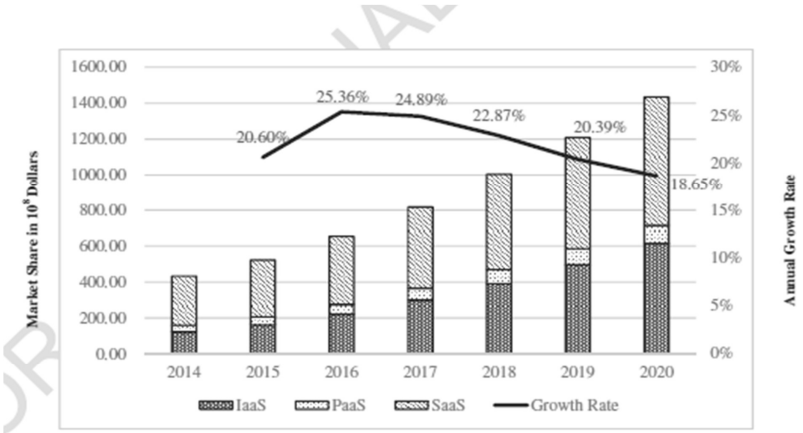


Figure 1.13
Worldwide distribution of cloud service models and the growth rate based on projections by Gartner Research from 2014–2020.

Table 4.6
Comparison of three cloud platform architectures

Cloud System Features	Amazon Web Service (AWS): Public Cloud	OpenStack Systems: Private Cloud	VMWare Systems: Hybrid Cloud
Service Model(s)	IaaS, PaaS	IaaS	IaaS, PaaS
Developer/Provider and Design	Amazon (Sec.4.3)	Rackspace/NASA and Apache (Section 2.3.4)	VMWare (Section 2.3.5) Proprietary
Architecture Packages and Scale	Data centers distributed as availability zones in many global regions (Figure 2.14)	Small cloud at owner sites, licensed thru Apache (Figure 2.15)	Private clouds interacting with public clouds (Figure 2.18)
Cloud OS/ Software Support	Supporting both Linux and Windows machine instances with autoscaling and billing	Open source, extending from Eucalyptus and OpenNebula	vSphere and vCenter, supporting x-86 servers with NSX and vSAN
User Spectrum	General public: enterprises and individual users	Research centers or small businesses	Enterprises and large organizations

The Amazon Web Service (AWS):

The Most popular Public Cloud in Use Today

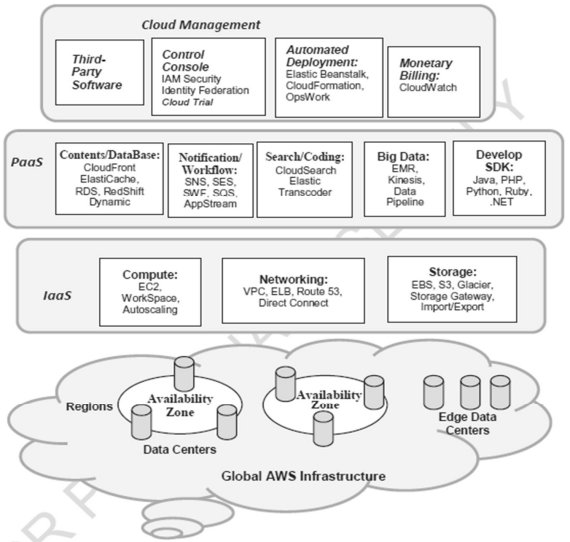


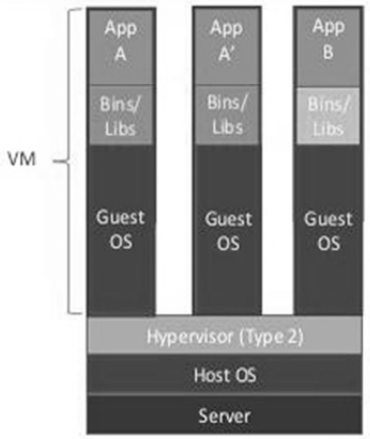
Figure 4.14
The AWS public cloud consisting of the top management layer, PaaS, and IaaS platforms, and the global infrastructure built over data centers in availability zones located in various regions.

Virtualization at Various Abstraction Levels

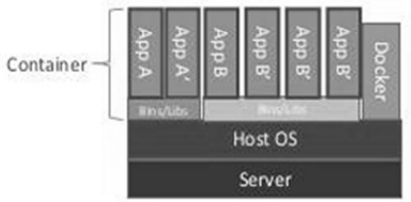
Table 1.7 Relative Merits of Virtualization at Five Abstraction Levels

Level of Virtualization	Functional Description	Example Pakages	Relative Merits, App Flexibility/Isolation, and Implementation Complexity
Instruction Set Architecture	*Emulation of a guest ISA by the host ISA*	*Dynamo, Bird, Bochs, Crusoe*	*Very Low performance, high app flexibility, and medium complexity and Isolation*
Hardware-level Virtualization	*Virtualization on top of bare metal hardware*	*XEN, VMWare, Virtual PC*	*High performance and complexity, medium app flexibility, and good app isolation*
Operating System Level	*Isolated containers as OS instances*	*Docker Engine, Jail, FVM,*	*Highest performance, Low App flexibility and Isolation, and average complexity*
Run-Time Library Level	*Creating VM via run-time library through API hooks*	*Wine, cCUDA, WABI, LxRun*	*Average performance, low app flexibility and isolation, and low complexity*
User Application Level	*Deploy HLL VMs at user app level*	*JVM, .NET CLR, Panot*	*Low performance and app flexibility, very high complexity and app isolation*

Containers vs. VMs



Containers are isolated, but share OS and, where appropriate, bins/libraries



Intel HiBench Benchmark for Testing Clouds:

- HiBench is a benchmark specifically tailored for running Hadoop programs based on MapReduce paradigm. The suite was developed at Intel for measuring the speed, throughput, HDFS bandwidth, and resources utilization in sort, word count, page ranking, Bayesian classifier, and distributed I/O workload. https://github.com/intel-hadoop/HiBench

- Hadoop programs in HiBench are :

1. Sort, 2. WordCount, 3. TeraSort, 4. Enhanced DFSIO,
5. Nutch indexing, 6. PageRank, 7. Bayesian classification,
8. K-means clustering, 9. Hive Query

Reference Paper: Huang, S., Huang, J., Dai, J., and Xie, T., and Hong, B., "The HiBench Benchmark Suite: Characterization of The MapReduce-based Data Analysis, *Int'l Conf. on Data Engineering Workshops,* March 2010.

Components in Apache Hadoop

- Hadoop Common: The common utilities that support the other Hadoop modules
- Hadoop Distributed File Syste, (HDFS): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management
- Hadoop MapReduce: A Yarn-based system for parallel processing of large data sets

Hadoop and Related Language and Programming Technologies

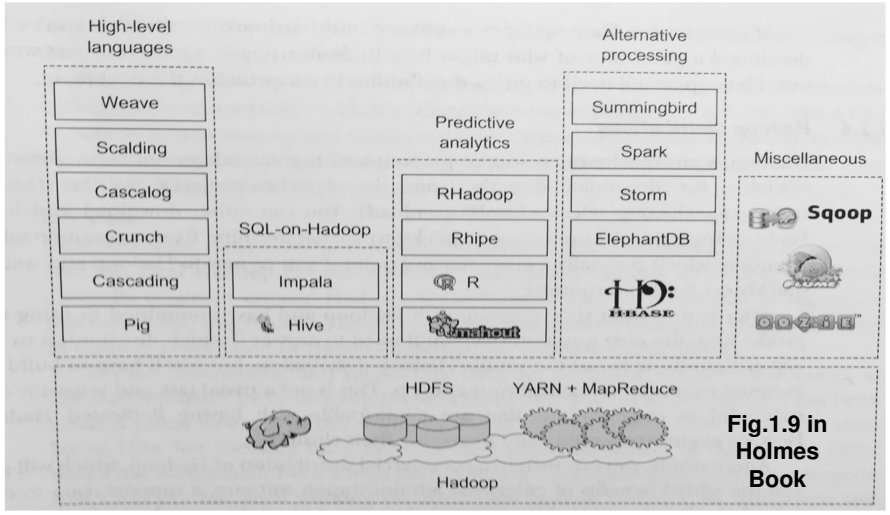


Fig.1.9 in Holmes Book

Core Concepts of Spark

- Spark's programming abstraction is enabled by RDD (Resilient Distributed Datasets), defined by many APIs for manipulating large collection of data items.
- Spark SQL deal with structured data.
- Spark Streaming handles live streams of data.
- Mllib library contains common machine learning functionality.
- GraphX for manuipulating social network graphs.
- Spark's Cluster Manager can run with
  - Hadoop YARN
  - Apache Mesos
  - Spark's own Standalone Scheduler.

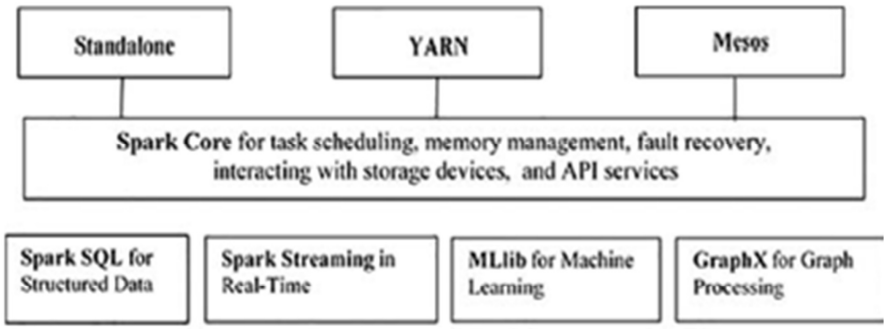
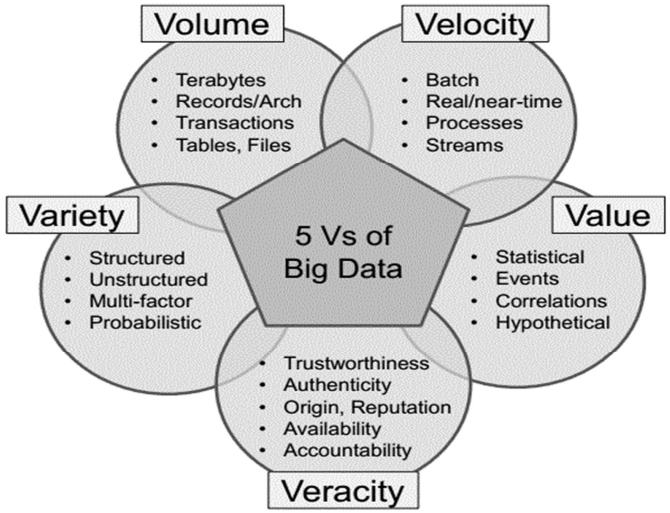
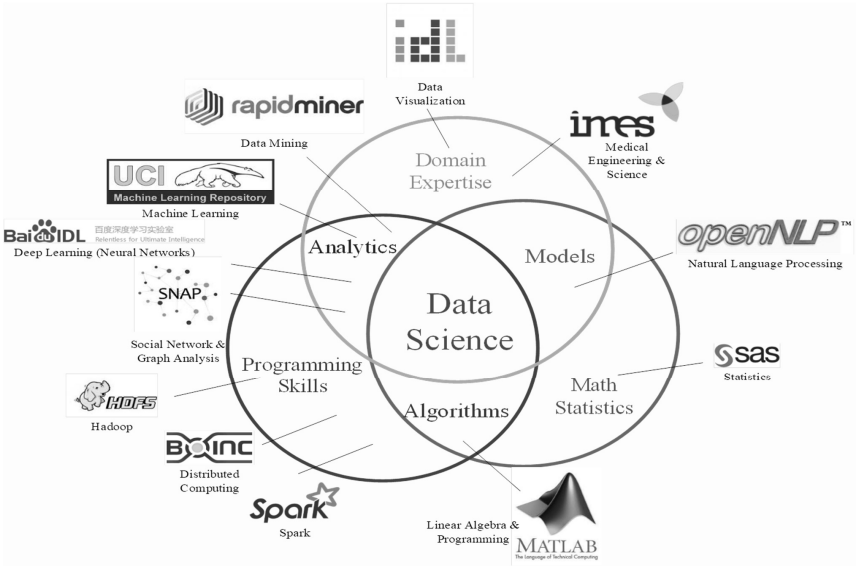


Figure 8.12
The core architecture of the Apache Spark software system.

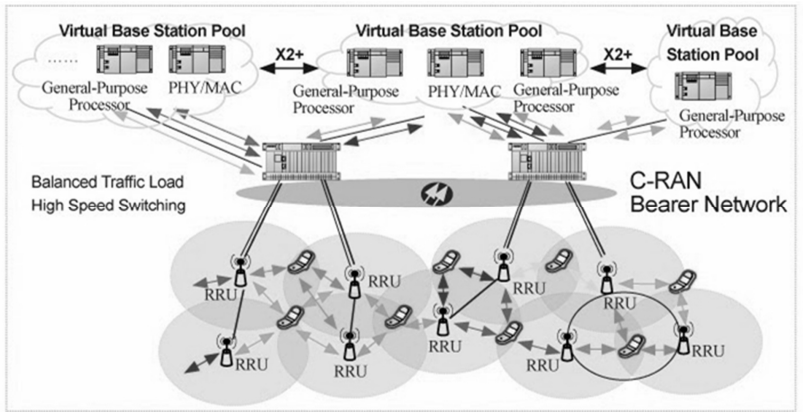
The Five V's of Big Data



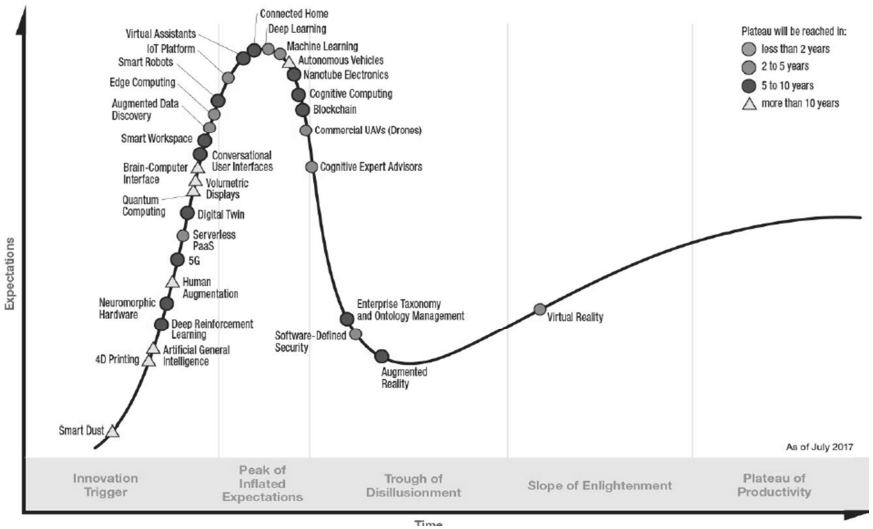
Today's Big Data Software Libraries

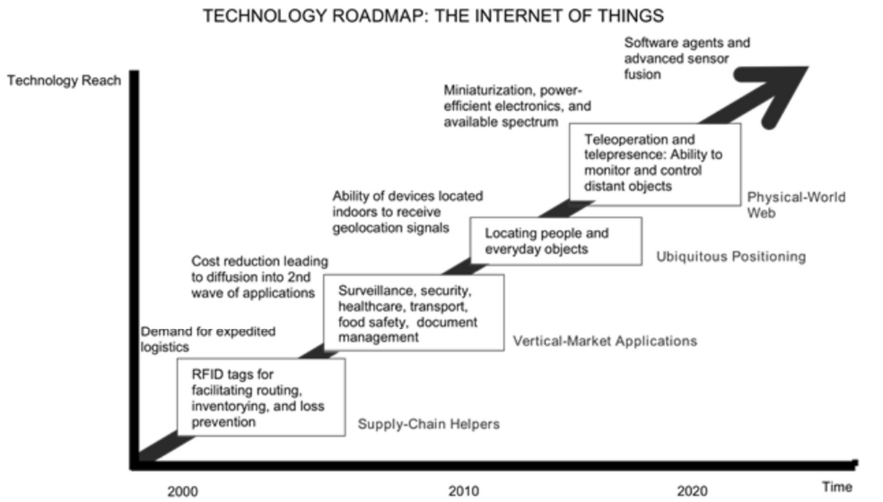


Virtual Base Station Pool and C-RAN Bear Network (1)



Gartner Hype Cycle for Emerging Technologies, 2017





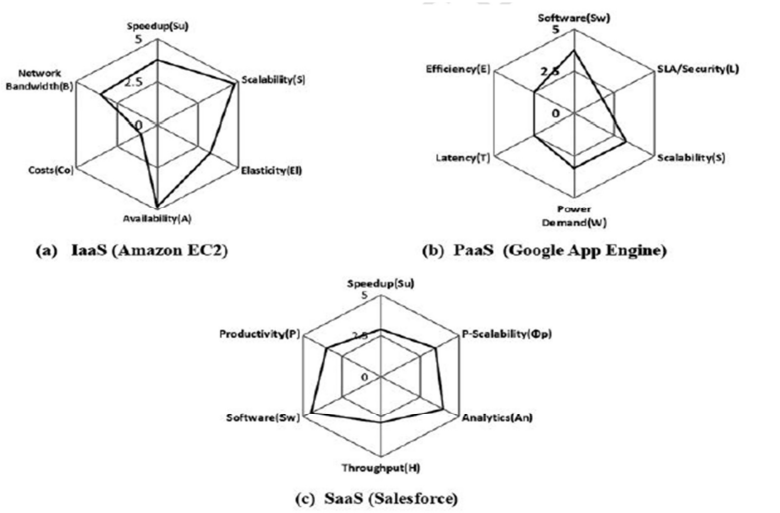


Figure 10.4
Performance maps of various clouds, where data points are extracted from reported Amazon EC2, Google App Engine, and Salesforce clouds. (Courtesy of Hwang et al.,"Cloud Performance Modeling with Benchmark Evaluation of Elastic Scaling Strategies," *IEEE Transactions on Parallel and Distributed Systems,* January 2016.)