USC EE 542: Fall 2017
Internet and Cloud Computing

Lectures 1 and 2 : August 21, 23, 2017
Introducing the Course : Cloud Architecture, Service Models and Basic Principles
(Chapter 1 in Hwang's Text, 2017)

Prof. Kai Hwang, EEB 212
kaihwang@usc.edu

EE 542 : Internet and Cloud Computing

Class Period : 15 weeks from Aug. 21 to Nov.29, 2017

Class Website: http://blackboard.usc.edu (Campus only)

Sec. 30533R : M.W. 2 pm – 3:15 am,

Class Room: WPH B27

Instructor: Kai Hwang, Professor of Electrical Engineering and Computer Science,

Office Hours: M.W. 9 am – 12 noon in EEB 212

Email: kaihwang@usc.edu

Teaching Assistant: Yue Shi, Yueshi@usc.edu

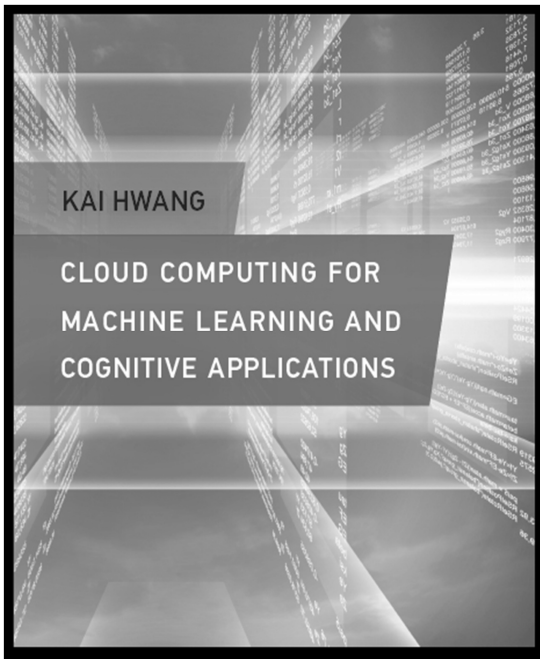
Course Description:

This course is designed for graduate students in electrical engineering and computer science. Students will learn the theory, architecture, hardware/software, and programming of computing clouds, Internet of Things (IoT). machine learning, big data analytics, cognitive computing and brain-inspired future computers.. Students will have the opportunity to gain hands-on experience in using Amazon cloud (AWS), where real-life cloud, big data or IoT applications will be developed and executed on Amazon EC2 and S3, etc. We will cover various clouds, namely AWS, GAE, Salesforce, Azure, Hadoop, Spark, Eucalyptus, vSphere, XEN, Docker containers, VMWare Tools, etc..

Required Textbook: Kai Hwang : *Cloud Computing for Machine Learning and Cognitive Applications* MIT Press, June 2017. (order immediately from www.mitpress.mit.edu or www.amazon.com). You must get hold of the book asap. Read all covered sections within 7 weeks in order to do well in homework sets, cloud programming project and in both exams.

Published by the MIT Press, Cambridge, MA. June 16, 2017 , (601 pages). Library of Congress ISBN 978-0-262-03641-2

Available from USC Bookstore, or order from the MIT Press via the web site:
mitpress.mit.edu
or order from the Amazon web site:
www.amazon.com



Visit EE 542 Web Site : blackboard.usc.edu
frequently for Syllabus update, Lecture Slides, Homework, Project Spec, Grade Book, and Handout Paper, etc.

Lectures and Dates in 2017	Topics Covered, Source, Due Dates and Exams
Lectures 1, 2 , Aug. 21, 23	Course Introduction, Principles of Cloud Computing Chapter 1
Lectures 3, 4, Aug. 28, 30,	Cloud Architecture, and Service Models, Chapter 4
Lectures 5, 6, Sept. 6, 11	Virtual Machines, Containers, Chapter 3 (No class on Labor Day , Sept.4)
Lecture 7, Sept. 13, 2017	Cloud Project Specification, (Team Proposal due Sept.20). HW#1 due Sept. 13
Lectures 8, 9, Sept. 18, 20	MapReduce, Hadoop and Spark Programming, Chapter 8
Lectures 10, 11, Sept. 25, 27	Big Data, IoT and Cognitive Computing, Chapter 2, ,
Lectures 12, 13, Oct. 2, 4,	Mobile Clouds, Cloud Mashup and Cloud OS, Chapters 3 and 5 ,
Lectures 14, 15, Oct. 9, 11	Cloud Performance and Scaling Techniques, Chapter 10, HW#2 due Oct. 9
Lecture 16, Oct. 16, 2017	Review Session of the first 15 lectures, Chapters 1 ~ 5, 8 , 10
Mid-Term Exam, Oct. 18	2 pm to 3:20 pm (80 minutes), Class Room WPH B27 plus overflow room

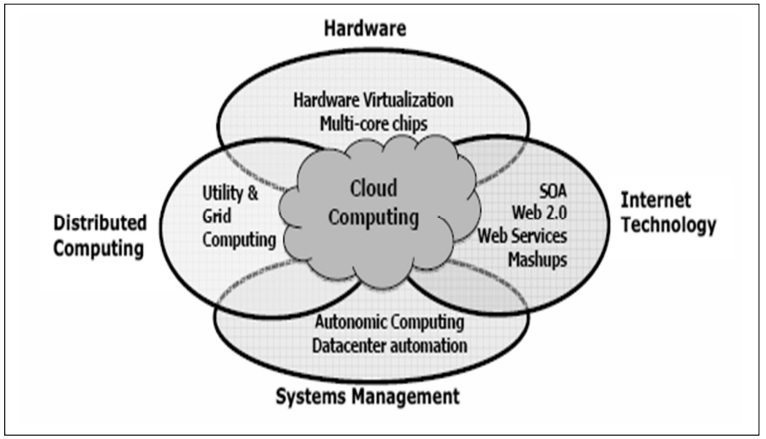
Lectures.17, 18, Oct. 23, 25	Machine Learning Algorithms Chapter 6
Lectures 19, 20 , Oct.30, Nov.1	AI Machines, and Deep Learning Tools Chapter 7
Lectures 21, 22, Nov. 6, 8	TensorFlow and Cognitive Systems, Chapter 9, HW#3 due Nov.13.
Lectures 23, 24, Nov.13, 15	Social Media, Health-Care Apps, and Security Issues, Chapters 8, 9, 10
Lecture 25 , Nov. 20, 2017	SMACT Technologies and Brain-inspired Future Computers, Chapters 1, 2, 7 Project Report due Nov. 20, (No class on Nov.22, Thanksgiving)
Lecture 26, Nov.27, 2017	Review of the entire course and return of graded Project Reports
Final Exam, Nov. 29	2 – 4 pm, Exam Rooms to be announced, covering the entire course .

- **The Syllabus will be updated in the lecture series. Check the class web site frequently for syllabus updates.**
- **You should not miss live lectures. According to our past experience, those students who raise the habit of cutting classes regularly perform the worst and end up with C's or D's or fail the course.**
- **To earn a decent grade, you must work hard all the way to the end. The Term Project (16%) and Final Exam (38%) are heavily weighted.**
- **Getting a good score in the mid-term exam and home works cannot guarantee your grade, if you intend to cut classes towards the end.**

Grading Policy and Class Rules:

- **Three Homework Sets (18%), Mid-Term Exam (28%), Project Report (16%), and Final Exam (38%).**
- **Typed homework solutions are submitted at the beginning of class on the due days. No late home work will be accepted due to handout of solutions promptly.**
- **All exams are close-book/close-notes. No make-up exam for any excuses. No use of computers or wireless phones during exams. Calculator is fine to use without WiFi or mobile connections.**
- **The Term Project requires you to experiment on existing public cloud like Amazon AWS, It is done by 3 students per team. The TA will help you form the Team, once the Project Spec is given in Lecture 7.**
- **The Project Report is due on Nov.20 before Thanksgiving recess. You have to submit a professionally prepared Project Report using IEEE Conference paper format (10 pages).**
- **The final exam is scheduled Nov.29, the last day of lecture. The whole class must agree if we decide to do so. We can shift your exam time slot slightly, If you have conflicts with other course on that day.**

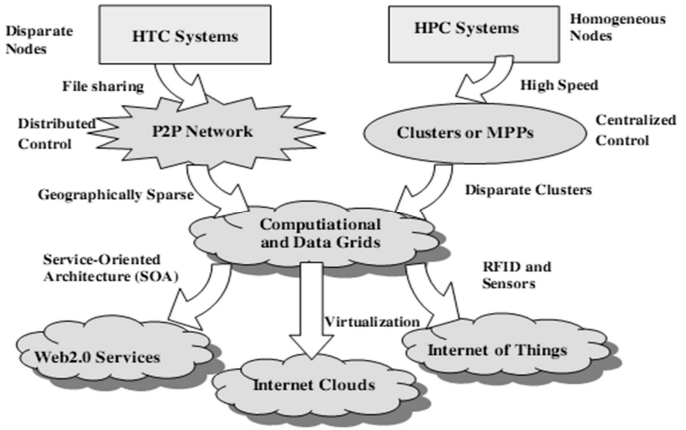
Data Deluge Enabling New Challenges



From Desktop/HPC/Grids to Datacenters and Clouds in 30 Years

- HPC moving from centralized supercomputers to geographically distributed desktops, desksides, clusters, and grids to clouds over last 30 years
- R/D efforts on HPC, clusters, Grids, P2P, and virtual machines has laid the foundation of cloud computing that has been greatly advocated since 2007
- Location of computing infrastructure in areas with lower costs in hardware, software, datasets, space, and power requirements – moving from desktop computing to datacenter-based clouds

From HPC Systems and Clusters to Grids, P2P Networks, Clouds, and the Internet of Things

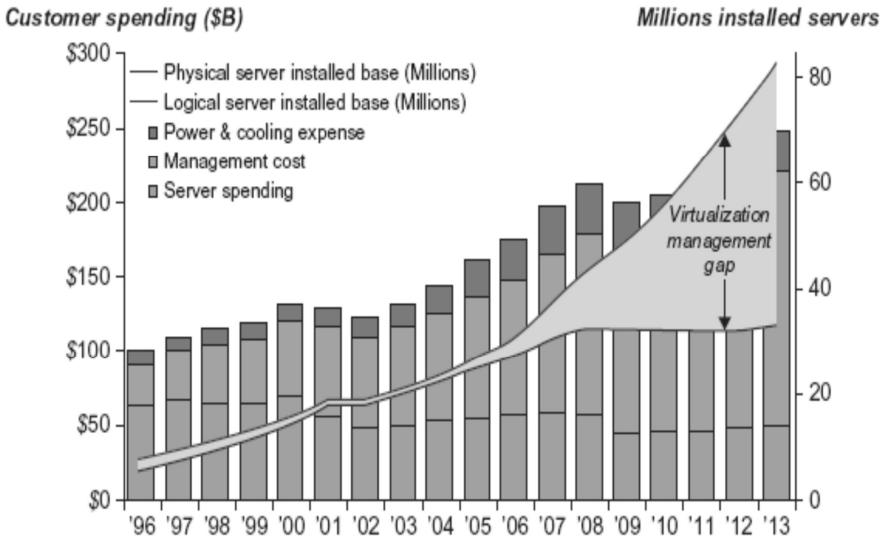


HPC: High-Performance Computing

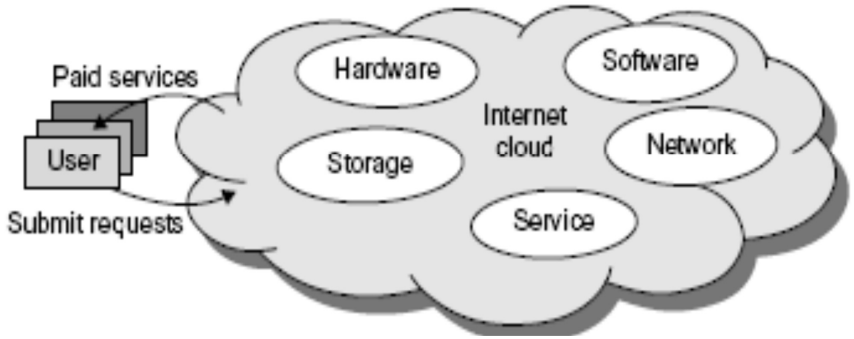
HTC: High-Throughput Computing

P2P: Peer to Peer

MPP: Massively Paralle Processors



Basic Concept of Internet Clouds



From Clusters, P2P Networks, Grids to Clouds

Table 1.3
Classification of parallel and distributed computing systems.

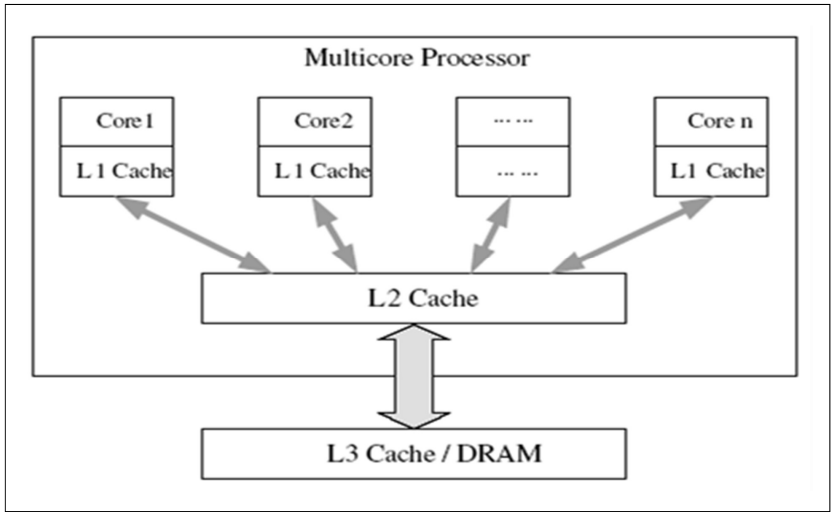
Functionality, Applications	Computer Clusters	Peer-to-Peer Networks	Computational Grids	Cloud Platforms
Architecture, Network Connectivity, and Size	Network of compute nodes interconnected by SAN, LAN, or WAN, hierarchically	Flexible network of client machines logically connected by an overlay network	Heterogeneous clusters interconnected by high-speed network links over selected resource sites	Virtualized cluster of servers over data centers via service-level agreement
Control and Resources Management	Homogeneous nodes with distributed control, running Unix or Linux	Autonomous client nodes, free in and out, with self-organization	Centralized control, server oriented with authenticated security	Dynamic resource provisioning of servers, storage, and networks
Applications and Network-Centric Services	High-performance computing, search engines, web services, etc.	Most appealing to business file sharing, content delivery, and social networking	Distributed super-computing, global problem solving, and data center services	Upgraded web search, utility computing, and outsourced computing services
Representative Operational Systems	Google search engine, Sun Blade, IBM Road-Runner, Cray XT4, etc.	Gnutella, eMule, BitTorrent, Napster, KaZaA, Skype, JXTA	TeraGrid, GriPhyN, UK EGEE, D-Grid, ChinaGrid, etc.	Google App Engine, IBM Smart Cloud, AWS, and Microsoft Azure

Enabling Technologies for Clouds

Table 1.1
Cloud-enabling technologies in hardware, software, and networking.

Technology	Requirements and Benefits
Fast Platform Deployment	Fast, efficient, and flexible deployment of cloud resources to provide dynamic computing environment to users.
Virtual Clusters on Demand	Virtualized cluster of VMs provisioned to satisfy user demand and virtual cluster reconfigured as workload changes.
Multitenant Techniques	SaaS distributes software to a large number of users for their simultaneous uses and resource sharing if so desired.
Massive Data Processing	Internet search and web services often require massive data processing, especially to support personalized services.
Web-Scale Communication	Support e-commerce, distance education, telemedicine, social networking, digital government, digital entertainment, etc.
Distributed Storage	Large-scale storage of personal records and public archive information demand distributed storage over the clouds.
Licensing and Billing Services	License management and billing services greatly benefit all types of cloud services in utility computing.

A Typical Multi-Core Processor



Multi-Core and Multithreaded Processors

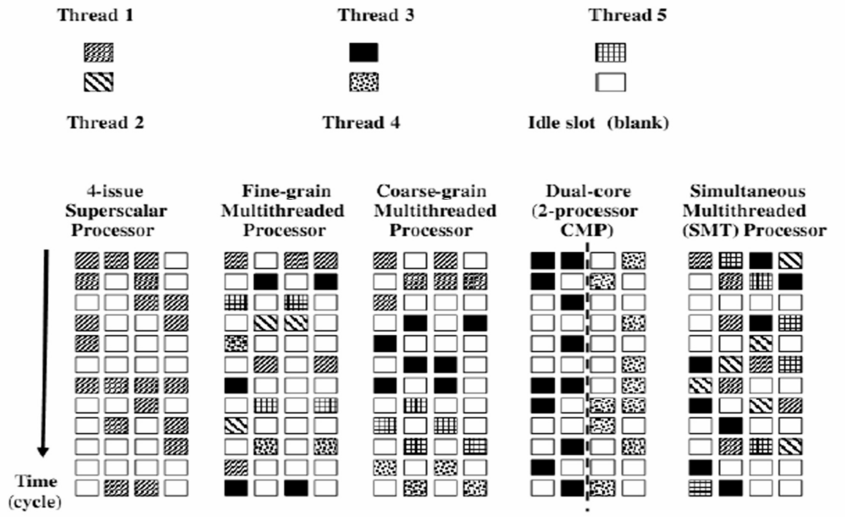
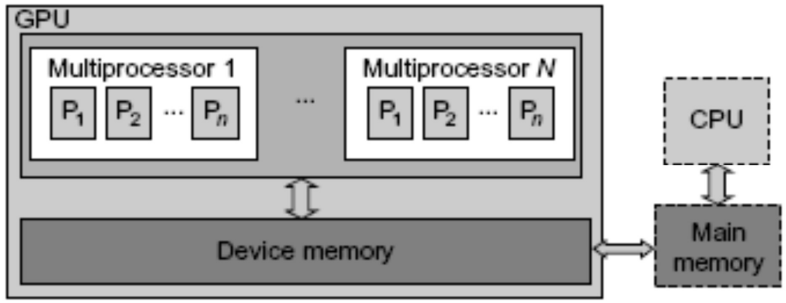
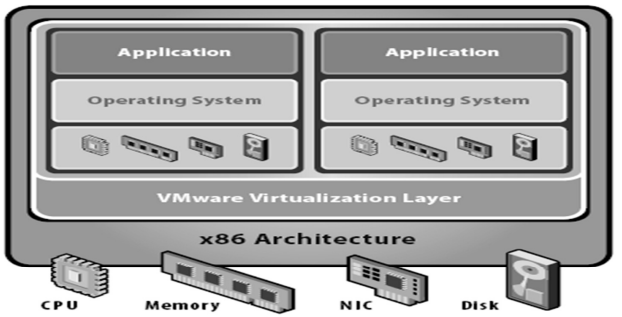


Figure 1. 8 Five micro-architectures that are current in use in modern processors that exploit both ILP and TLP supported by multicore and multithreading technologies

Architecture of A Many-Core Multiprocessor GPU interacting with a multi-core CPU Processor



Virtual Computer Architecture



After Virtualization:

- Hardware-independence of operating system and applications
- Virtual machines can be provisioned to any system
- Can manage OS and application as a single unit by encapsulating them into virtual machines

Concept of Virtual Clusters

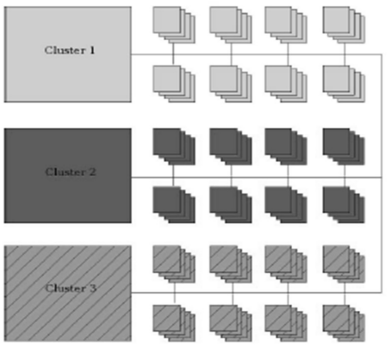


Fig. 1. A Campus Area Grid

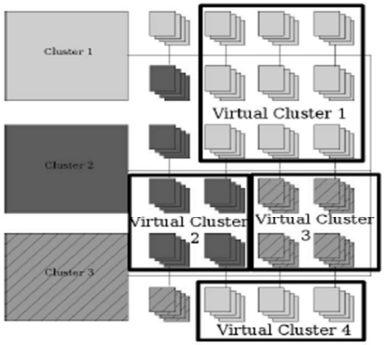


Fig. 2. Virtual machines in a cluster environment

The Cloud

- Historical roots in today's Internet applications
  - Search, email, social networks
  - File storage (Live Mesh, Mobile Me, Flicker, …)
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications
- A cloud is the "invisible" backend to many of our mobile applications
- A model of computation and data storage based on "pay as you go" access to "unlimited" remote data center capabilities

The Next Revolution in IT
Cloud Computing

- **Classical Computing**
  - Buy & Own
    - Hardware, System Software, Applications often to meet peak needs.
  - Install, Configure, Test, Verify, Evaluate
  - Manage
  - ..
  - Finally, use it
  - $$$$....$(High CapEx)

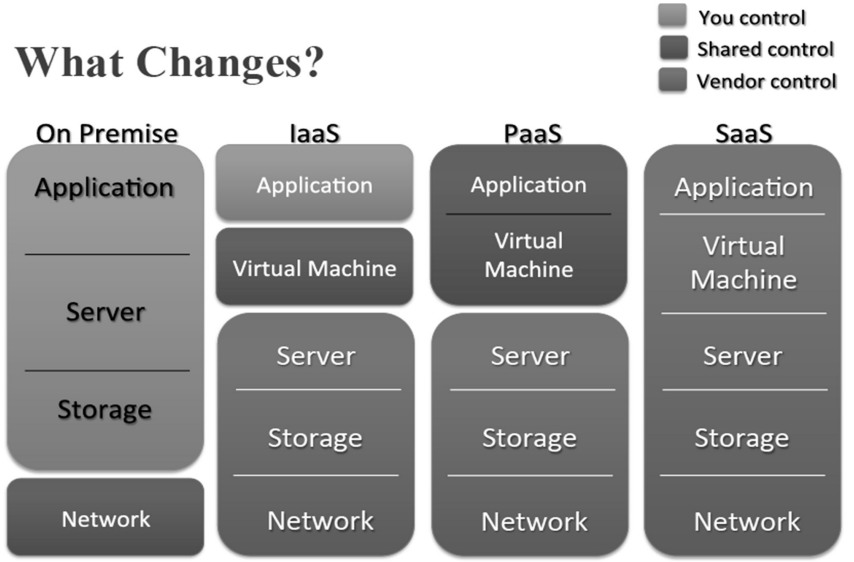
Every 18 months?

- **Cloud Computing**
  - Subscribe
  - Use

  - $ - pay for what you use, based on QoS

(Courtesy of Raj Buyya, 2012)

What Changes?



Mather, Kumaraswamy and Latif, "Cloud Security and Privacy," O'Reilly 2009

2015 Cloud Technologies

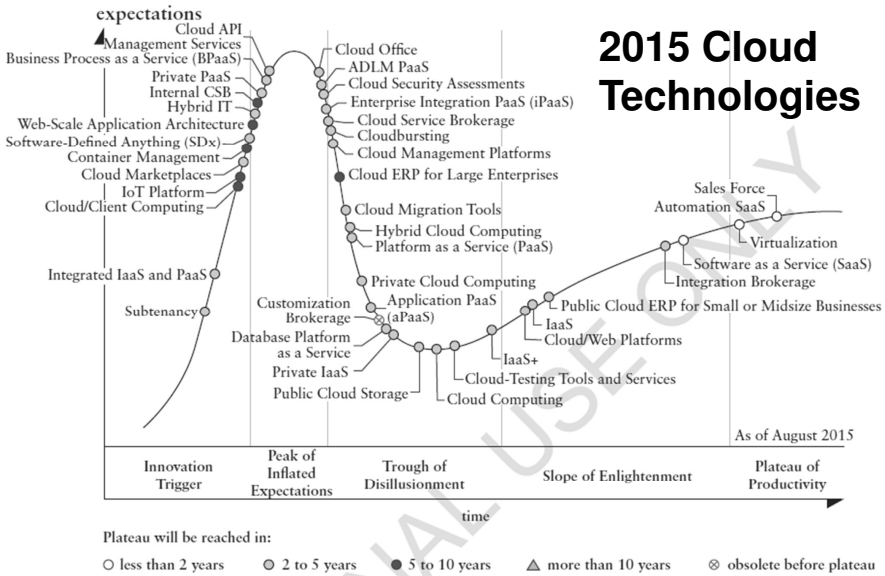


Figure 1.17
Gartner's Hype Cycle for cloud computing in August 2015. Reprinted with permission from Gartner Research, Inc.

Table 1.6
Top 10 strategic technology trends for cloud computing in 2015.

Merging the Real World and the Virtual World	1	Computing everywhere
	2	The Internet of things
	3	3D printing
Intelligence Everywhere	4	Advanced, pervasive, and invisible analytics
	5	Context-rich systems
	6	Smart machines
The New IT Reality Emerges	7	Cloud/client computing
	8	Software-defined application and infrastructure
	9	Web-scale IT
	10	Risk-based security and self-protection

Lecture 2: Cloud Architecture and Service Models Aug.23, 2017

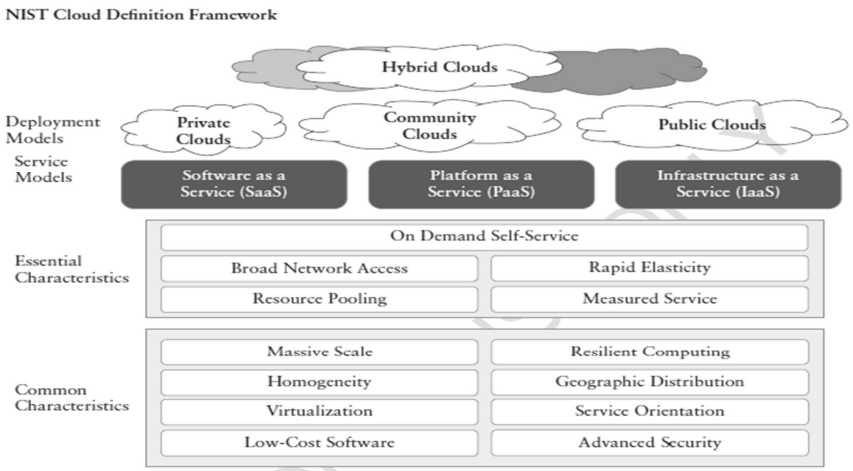


Figure 1.8
Public, private, community, and hybrid clouds. Courtesy of National Institute of Standards and Technology, 2013.

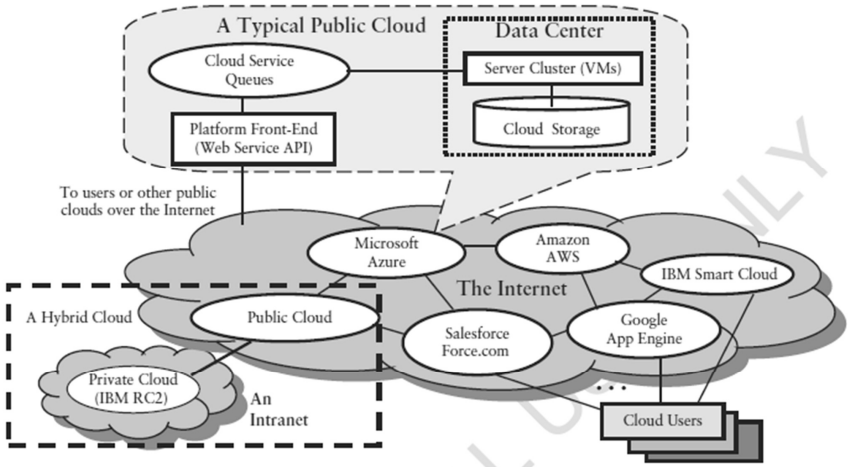
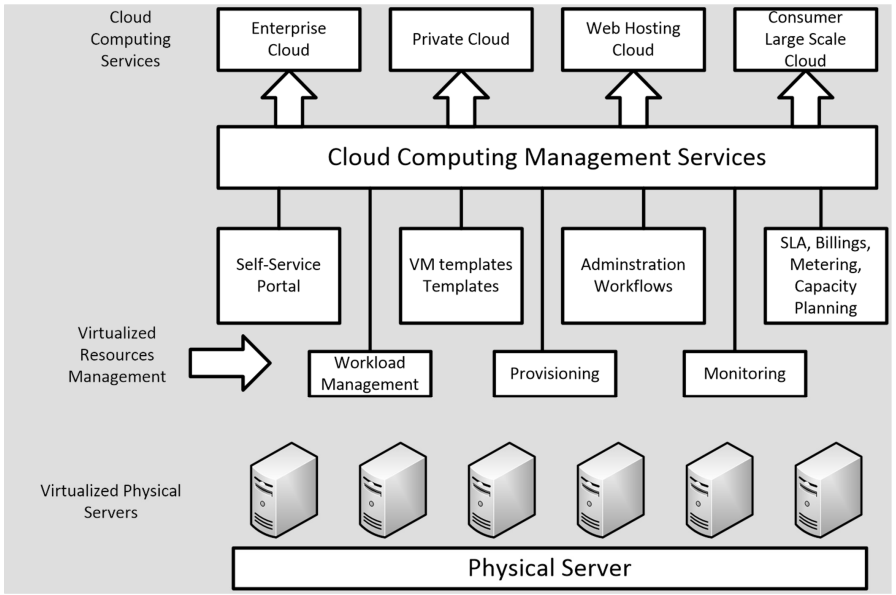
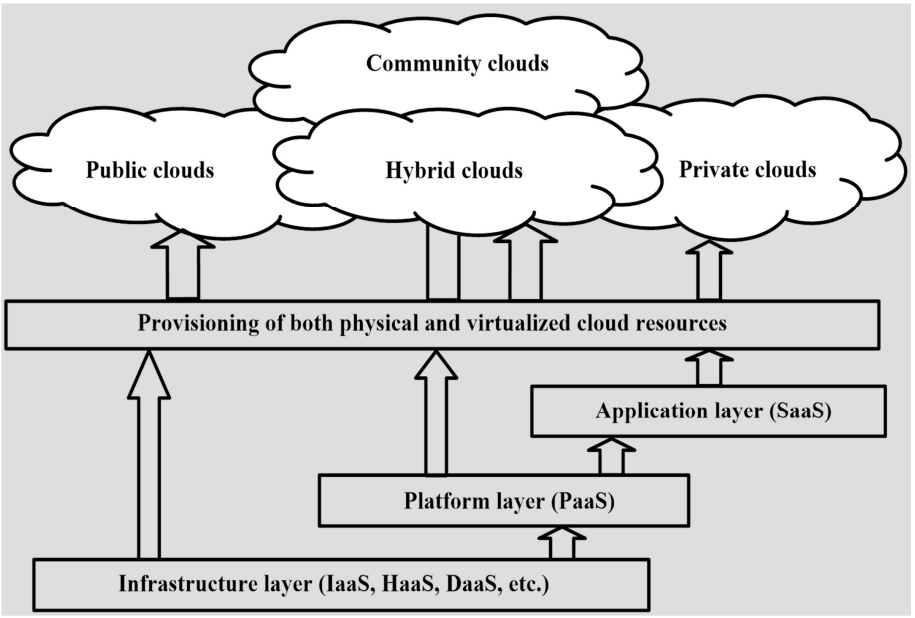
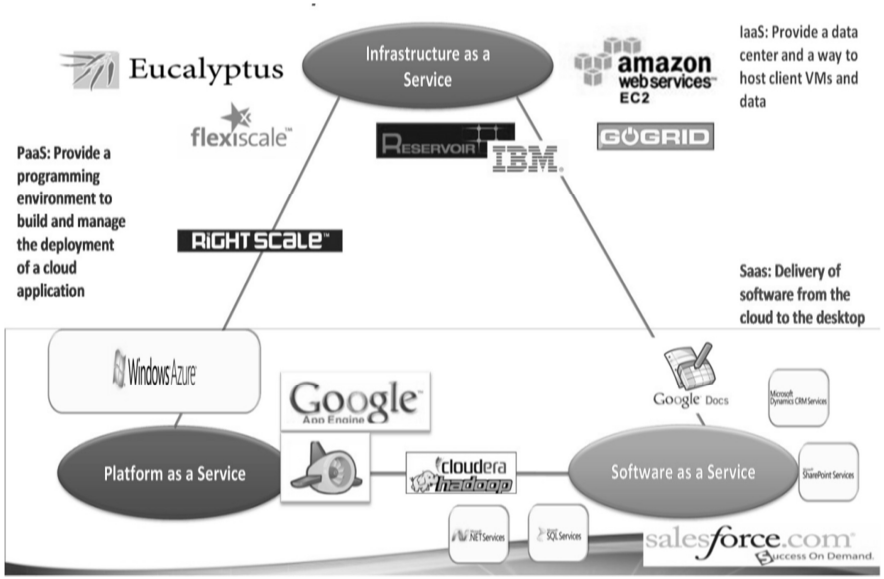
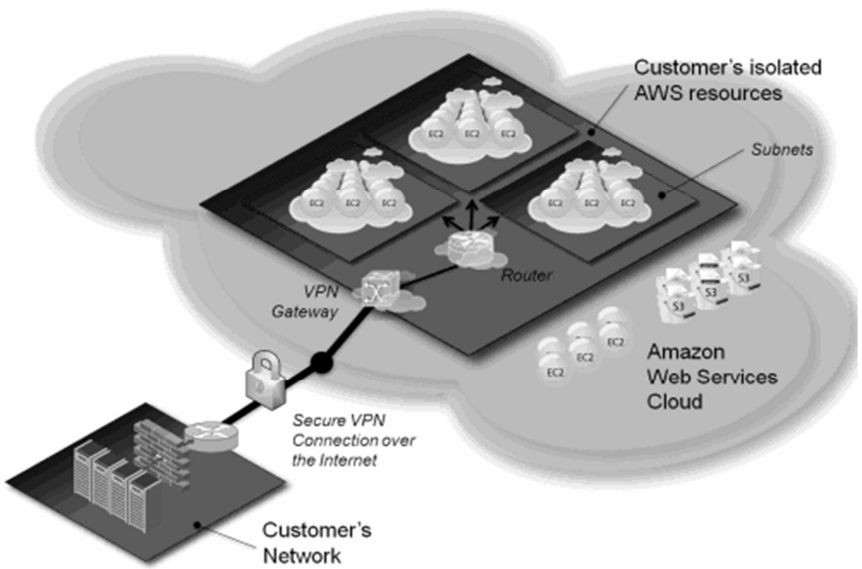


Figure 1.7
Public, private, and hybrid clouds. The callout box shows the architecture of a typical public cloud. A private cloud is built within an Intranet. A hybrid cloud involves both types in its operation range. Users access the clouds from a web browser or through a special API tool.









Market Share of Various Cloud Platforms

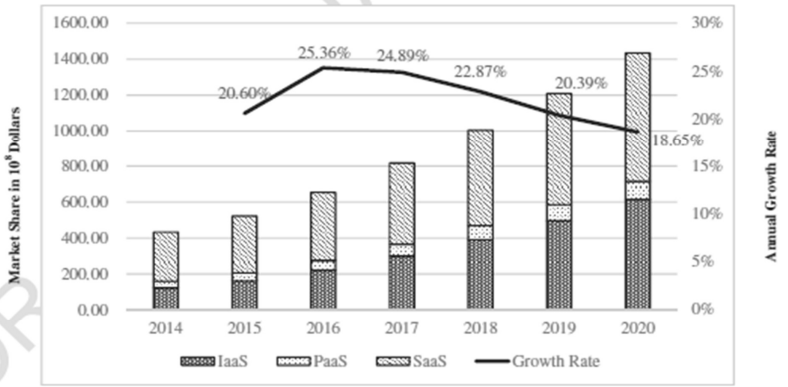


Figure 1.13
Worldwide distribution of cloud service models and the growth rate based on projections by Gartner Research from 2014–2020.

Cloud Users vs. Cloud Providers

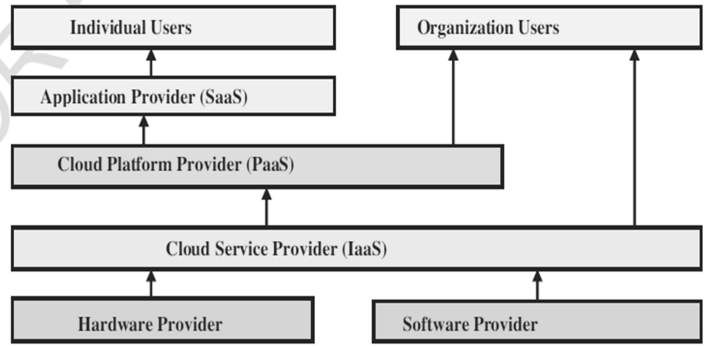


Figure 1.15
Individual versus organization users of cloud computing and their services, hardware, and software providers.

Table 1.4 Cloud Providers and Vendors

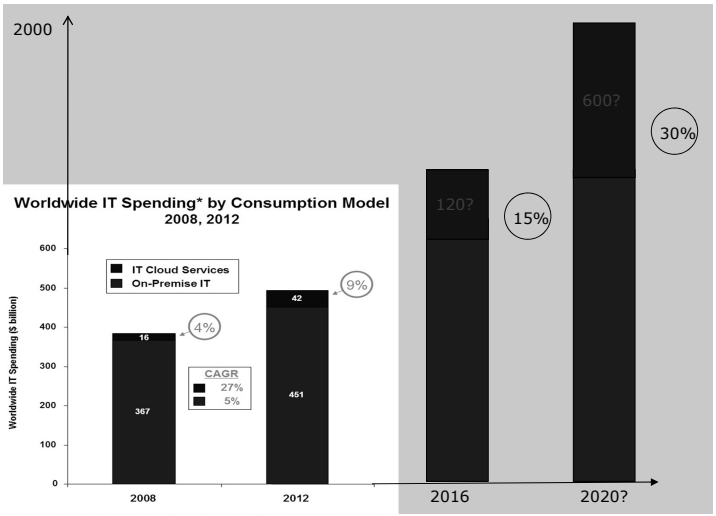
Cloud Players	IaaS	PaaS	SaaS
IT Administrators and Cloud Providers	Monitor SLAs	Monitor SLAs and enable Service Platforms	Monitor SLAs and deploy software
Software Developers (Vendors)	To deploy and store data	Enabling Platforms via configurator and APIs	Develop and Deploy Software

Table 1.5
Cloud application trends beyond web services and Internet computing.

Categories	Some Cloud Service Examples
Document and Databases	Collaborative word processing using docs.google.com, joint co-authorship using Dropbox for synchronization
Community/ Communications	Group exchanges, community services, security watch, social welfare, alert, and alarming systems
Storage and Data Sharing	Backup storage on Dropbox, records on iCloud, photo sharing on Facebook, and professional profiling and job hunting on LinkedIn
Activity/Event Management	Calendar, contacts, event planning, family budgeting, school events, exercise team, and scheduling
Project/Mission Management	Joint design, collaborative project, virtual organizations, mission coordination, strategic defense, battlefield management, crisis handling, etc.
e-Commerce and Business Analytics	Online shopping on Amazon, Taobao, Jingdong, eBay, Salesforce CRM, and sales clouds
Healthcare and Environment	Big data for healthcare through hospitals and public clinics, pollution control, environmental protection, emotion control, caring for the elderly
Social Media and Entertainment	Centralized e-mail services like the Outlook Web App (OWA) through MS Office 365, Facebook, Twitter, Gmail, QQ, LinkedIn, cloud gaming, etc.

Cloud Bisiness Potential:

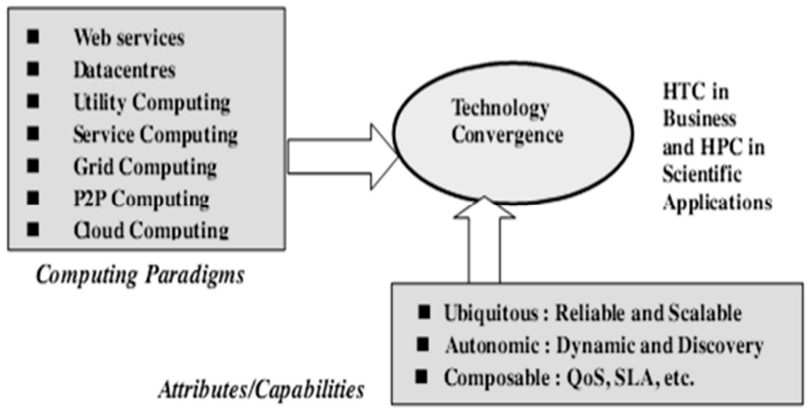
A trillion $ business/year by 2020?



* Includes enterprise IT spending on Business Applications, Systems Infrastructure Software, Application Development & Deployment Software, Servers and Storage

Source: IDC, October 2008

Technology Convergence toward HPC for Science and HTC for Business



Warehouse-Scale Computer :

- **Provides Internet services**
  - **Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.**
- **Differences with HPC "clusters":**
  - **Clusters have higher performance processors and network**
  - **Clusters emphasize thread-level parallelism, WSCs emphasize request-level parallelism**
- **Differences with datacenters:**
  - **Datacenters consolidate different machines and software into one location**
  - **Datacenters emphasize virtual machines and hardware heterogeneity in order to serve varied customers**

Services-Oriented Architecture (SOA)

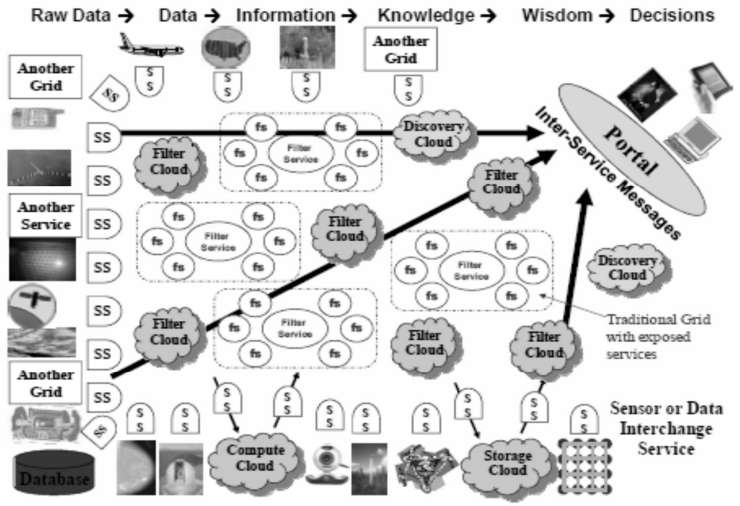


Figure 1.23 The evolution of service-oriented architecture: Grids of Clouds and Grids where SS refers to Sensor Service and fs to a filter or transforming service.

System Availability vs. Configuration Size :

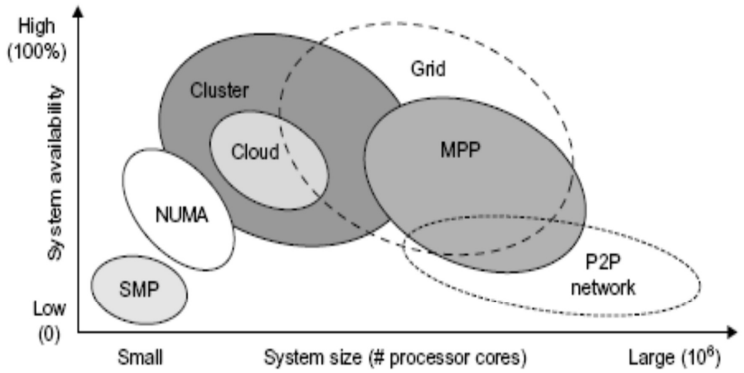
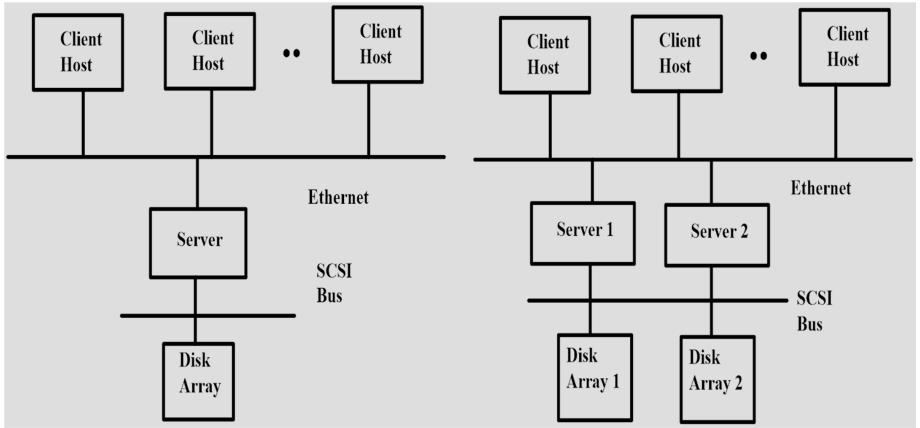


FIGURE 1.24

Estimated system availability by system size of common configurations in 2010.

Single Point of Failure :
The Server and the Disk



$$Cluster\ Availability = MTTF / (MTTF + MTTR).$$

System Availability Analysis

$$A = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$
$$= \binom{n}{k} p^k (1-p)^{n-k} + \binom{n}{k+1} p^{k+1} (1-p)^{n-k-1}$$
$$+ \cdots + \binom{n}{n-1} p^{n-1} (1-p)^1 + \binom{n}{n} p^n (1-p)^0,$$

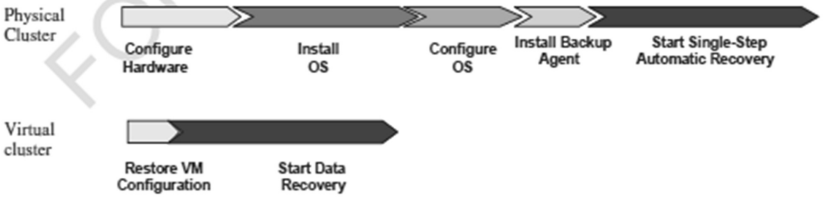


Figure 1.21
Recovery overhead on a physical cluster compared with that of a virtual cluster.

$$Speedup = S = T / [\alpha T + (1-\alpha)T/n] = 1 / [\alpha + (1-\alpha)/n]. \quad (1.1)$$

The maximum speedup of n is achieved only if the *sequential bottleneck* α is reduced to 0 or the code is fully parallelizable with $\alpha = 0$. As the cluster becomes sufficiently large, i.e., $n \to \infty$, S approaches $1/\alpha$, which is the upper bound on the speedup S. Surprisingly, this upper bound is independent of the cluster size n.

Sequential bottleneck is the portion of the code that cannot be parallelized. For example, the maximum speedup achievable is 4, if $\alpha = 0.25$ or $1-\alpha = 0.75$, even if one uses hundreds of processors. Amdahl's Law implies that one should make the sequential bottleneck of all programs as small as possible. Increasing the cluster size alone may not give a good speedup as the program structure is essentially sequential in nature.

Amdahl's Law assumes the workload (or problem size) is fixed regardless how large a cluster is used. Hwang [18] refer to this as *fixed-workload speedup*. To execute a fixed workload on n servers, parallel processing may lead to a *cluster efficiency* defined by:

$$E = S/n = 1/[\alpha n + 1 - \alpha]. \quad (1.2)$$

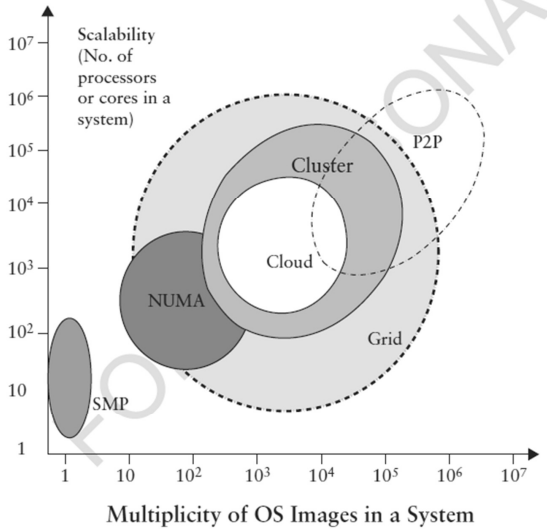


Figure 1.14
System scalability versus multiplicity of OS images based on 2010 technology. Reprinted with permission from K. Hwang and Z. Xu, *Scalable Parallel Computing*, McGraw-Hill, 1998.

Integrated SMACT Technologies

Table 1.7
SMACT technologies characterized by basic theories, typical hardware, software tooling, networking, and service providers needed.

SMACT Technology	Theoretical Foundations	Hardware Advances	Software Tools and Libraries	Networking Enablers	Representative Service Providers
Mobile Systems	Telecommunication, radio access theory, mobile computing	Smart devices, wireless, mobility infrastructures	Android, iOS, Uber, WeChat, NFC, iCloud, Google Play	4G LTE, WiFi, Bluetooth, radio access networks	AT&T Wireless, T-Mobile, Verizon, Apple, Samsung
Social Networks	Social science, graph theory, statistics, social computing	Data centers, search engines, and www. infrastructure	Browsers, APIs, Web 2.0, YouTube, WhatsApp, WeChat	Broadband Internet, sofware-defined networks	Facebook, Twitter, QQ, LinkedIn, Baidu, Amazon, Taobao
Big Data Analytics	Data mining, machine learning, artificial intelligence	Data centers, clouds, search engines, big data lakes, data storage	Spark, Hama, BitTorrent, MLlib, Impala, GraphX, KFS, Hive, HBase	Co-location clouds, mashups, P2P networks, etc.	AMPLab, Apache, Cloudera, FICO, Databricks, eBay, Oracle
Cloud Computing	Virtualization, parallel/distributed computing	Server clusters, clouds, VMs, interconnection networks	OpenStack, GFS, HDFS, MapReduce, Hadoop, Spark, Storm, Cassandra	Virtual networks, OpenFlow networks, software-defined networks	AWS, GAE, IBM, Salesforce, GoGrid Apache, Azure, Rackspace, DropBox
Internet of Things (IoT)	Sensing theory, cyber physics, pervasive computing	Sensors, RFID, GPS, robotics, satellites, ZigBee, gyroscope	TyneOS, WAP, WTCP,IPv6, Mobile IP, Android, iOS, WPKI, UPnP, JVM	Wireless LAN, PAN, MANET, WMN Mesh, VANET, Bluetooth	IoT Council, IBM, social media, Smart Earth, Google, Samsung

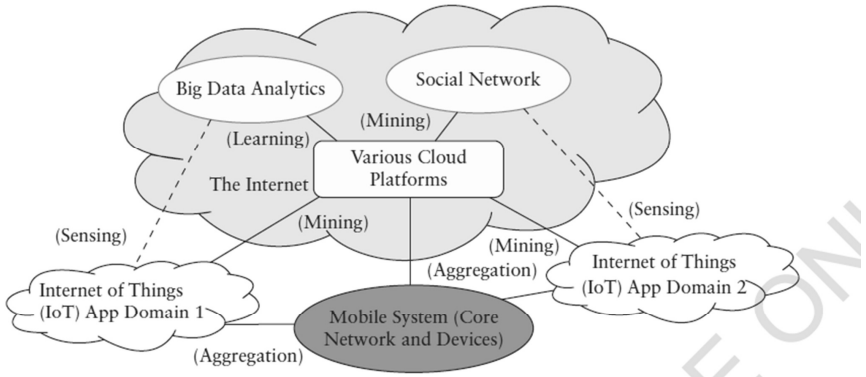


Figure 1.18
Interactions among social networks, mobile systems, big data analytics, and cloud platforms over various Internet of things (IoT) domains.

Home Work Problems assigned to H.W. Set #1 Due Sept. 13, 2017 (4%)

Note: Homework must be done independently. You should work on the assigned problems immediately after each lecture. Copying homework from fellow classmate or from solution handouts in previous years will result in zero credit for both loaner and borrower.

Chapter 1: Prob. 1.4, 1.9, 1.12, (week 1)

Chapter 3: Prob. 3.14, 3.16, 3.18 (week 3)

Chapter 4: Prob. 4.1, 4.2, 4.11, 4.12 (week 2)

Lectures 1~6 are relevant to HW #1 in the first 3 weeks . Several problems require open research by digging out updated information from Wikipedia or the Internet.

Textbook and 3 Reference Books:

1. K. Hwang, *Cloud Computing for Machine Learning and Cognitive Applications, MIT Press, Cambridge, MA. 2017*
2. R. Buyya, C. Vecchiola, and S. Selvi, *Mastering Cloud Computing:* , *Morgan Kauffmann Publishers*, July 2013
3. T. Chou, *Introduction to Cloud Computing: Business and Technology,* Lecture Notes at Stanford University, Active Book Press, 2010.
4. K. Hwang, G. Fox and J. Dongarra, *Distributed and Cloud Computing: From Parallel Processing to The IoT,* Morgan Kaufmann, 2011

Reading Assignments in Chapter 1:

All sections except the last two subsections 1.4.3 and 1.4.4. You should finish reading the relevant sections before solving the assigned problems.