Kai Hwang

Cloud Computing for Machine Learning and Cognitive Applications

The MIT Press, Cambridge, MA, 2017

Chapter 1: Cloud Computing Principles and Technologies

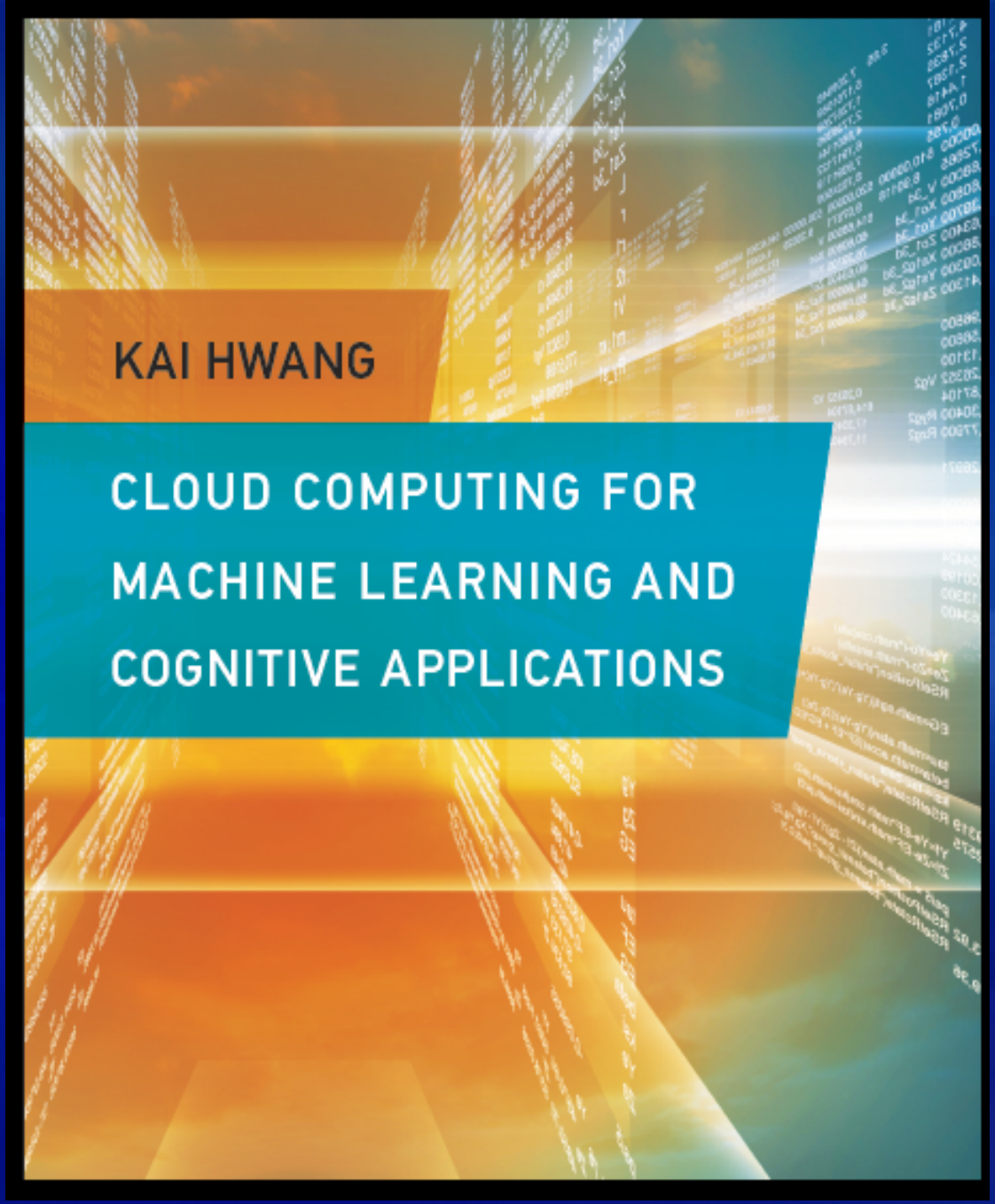
(52 slides for 3-hour lectures)

In total, there are 746 slides in 10 chapter files. These slides are suggested for use in 45 hours of lectures for senior undergraduate or graduate courses in one semester.

Published by The MIT Press, Cambridge, MA, on June 16, 2017. 601 pages.

Library of Congress ISBN 978-0-262-03641-2.

Available at the USC Bookstore, The MIT Press (mitpress.mit.edu), and Amazon (www.amazon.com).

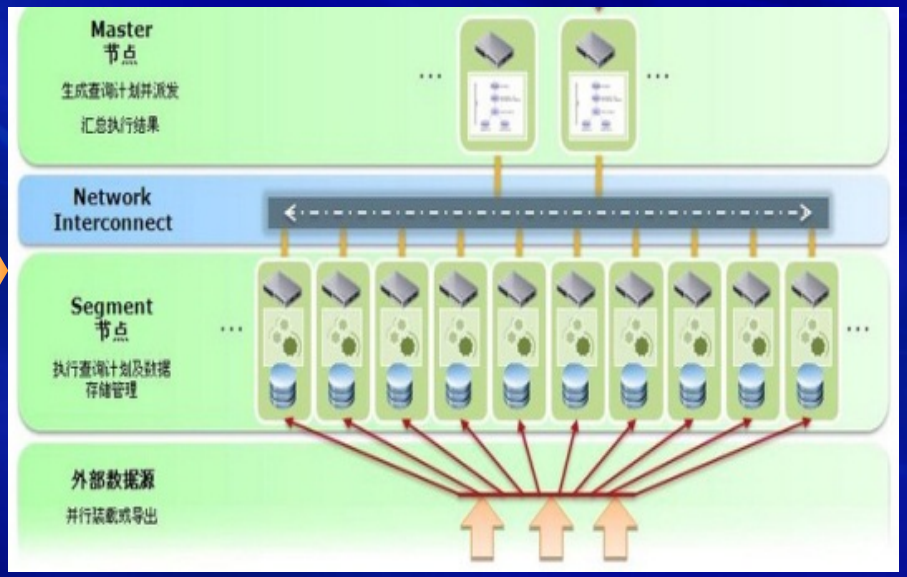


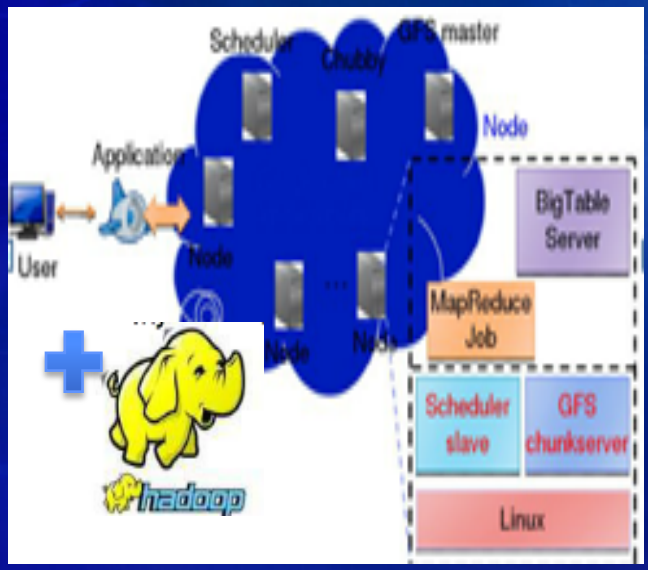
Cloud computing over big data

- Structured data; low processing efficiency of unstructured data
- Expensive hardware; Poor compatibility
- High scalability with Vendor lock-in of data resources

- Hybrid analysis capabilities of structured/unstructured data
- X86 servers; good compatibility
- High scalability, over 10,000 node-level deployment

Traditional Database/ Data Warehouse

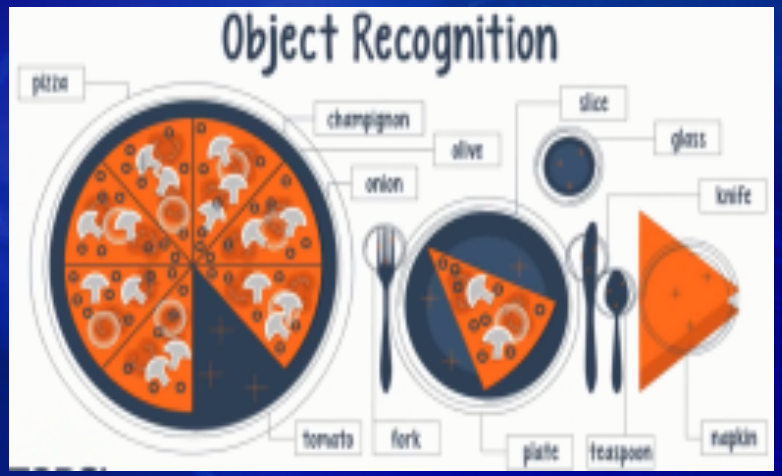




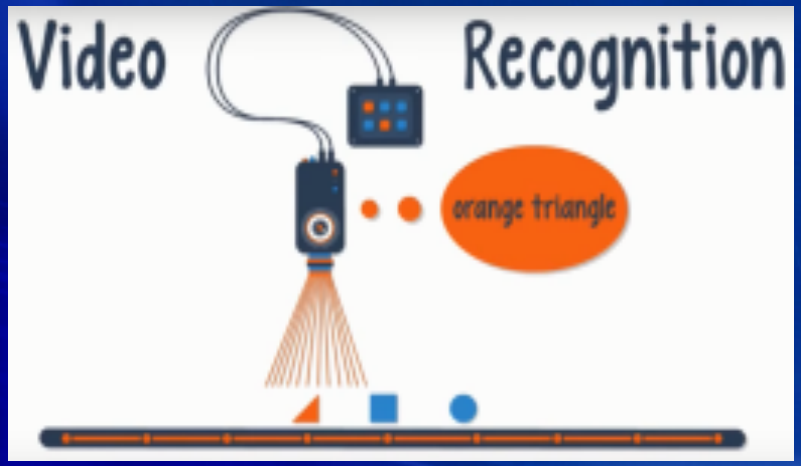
Supercluster (Data Warehouse) + Hadoop + Spark + Storm + TensorFlow + …..)

TB PB Distributed architecture EB → ZB

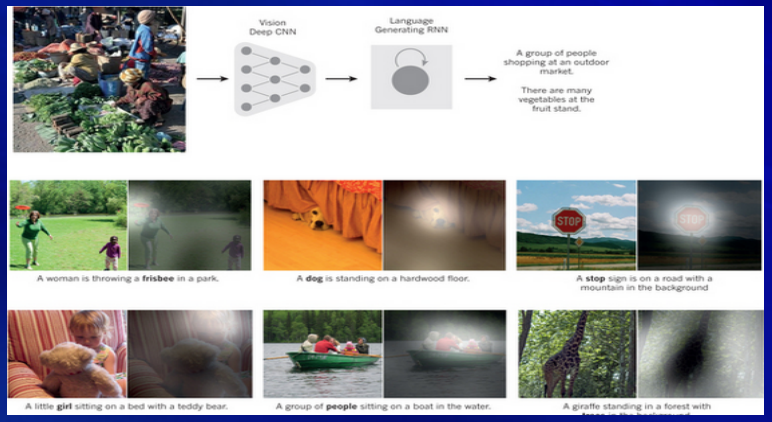
Machine learning and cognitive applications



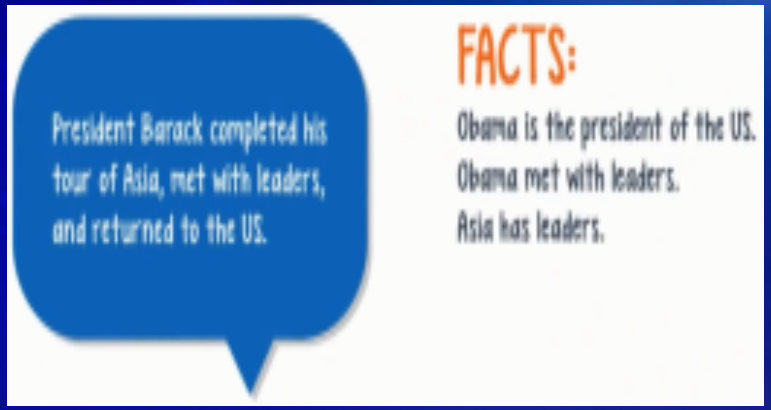
(a) Object recognition



(b) Video recognition



(c) Image retrieval

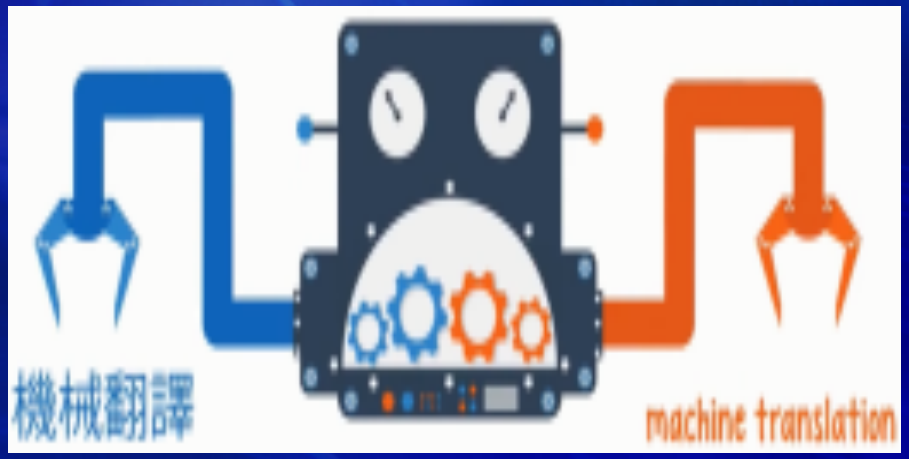


(d) Fact extraction

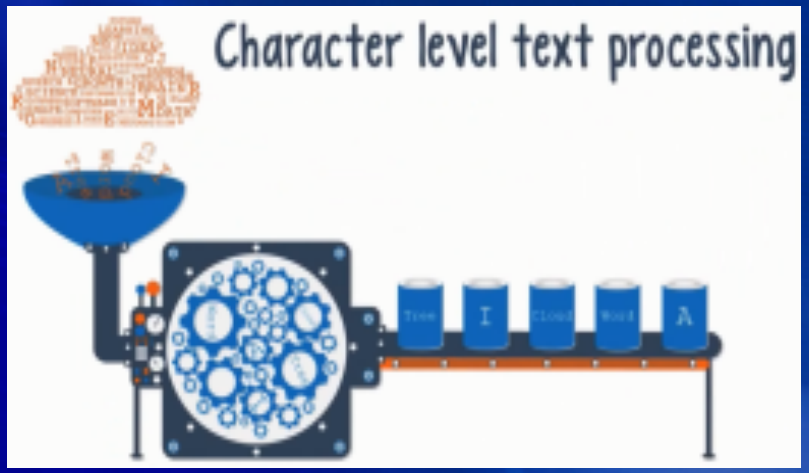
4

Machine and deep learning applications

Classified in 16 Categories (continued)



(e) Machine translation



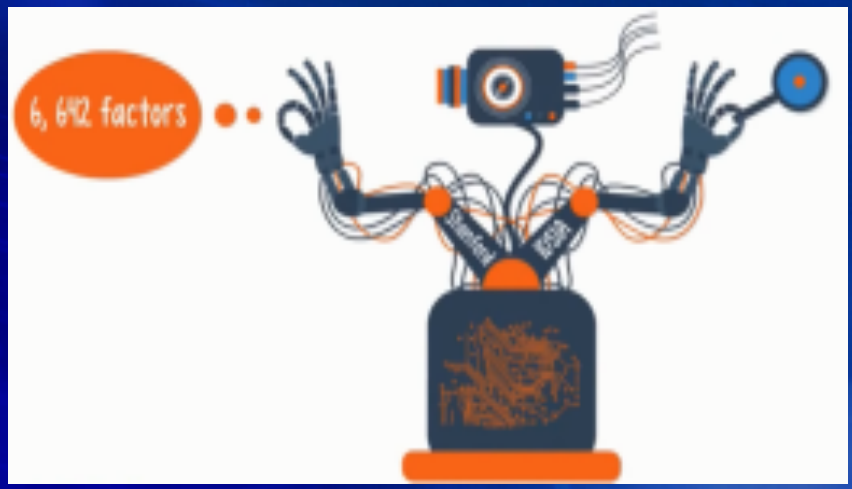
(f) Text processing

(g) Speech recognition

(h) Natural language processing

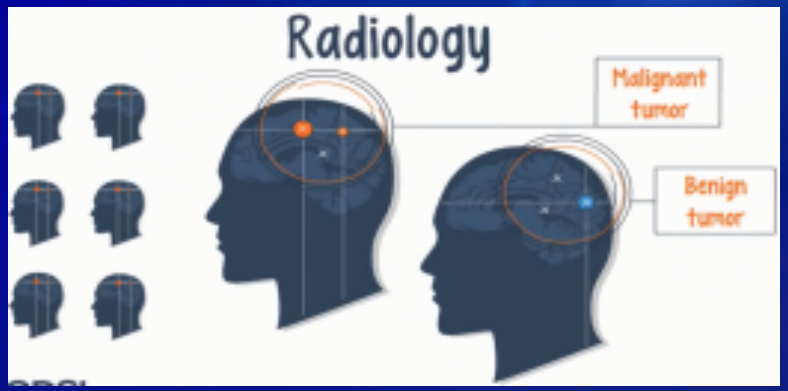
Machine and deep learning applications
Classified in 16 Categories (continued)

(i) Sentiment analysis



(j) Cancer detection

(k) Drug discovery/toxicology

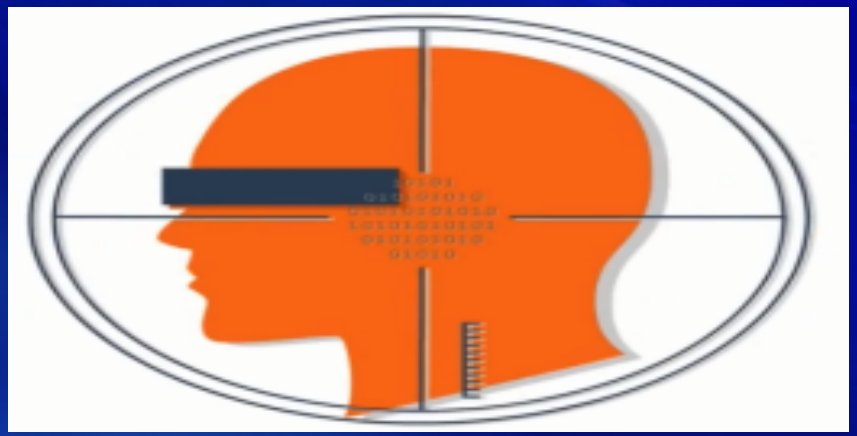


(l) Radiology

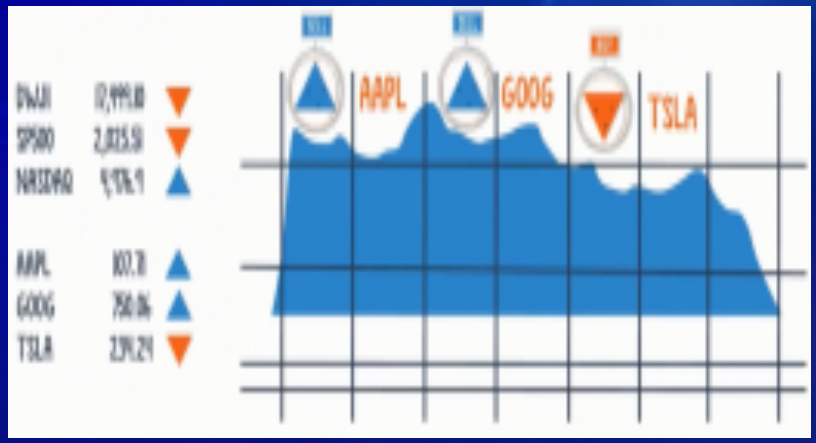
Machine and deep learning applications
Classified in 16 Categories (continued)

(m) Bioinformatics

(n) Digital advertising

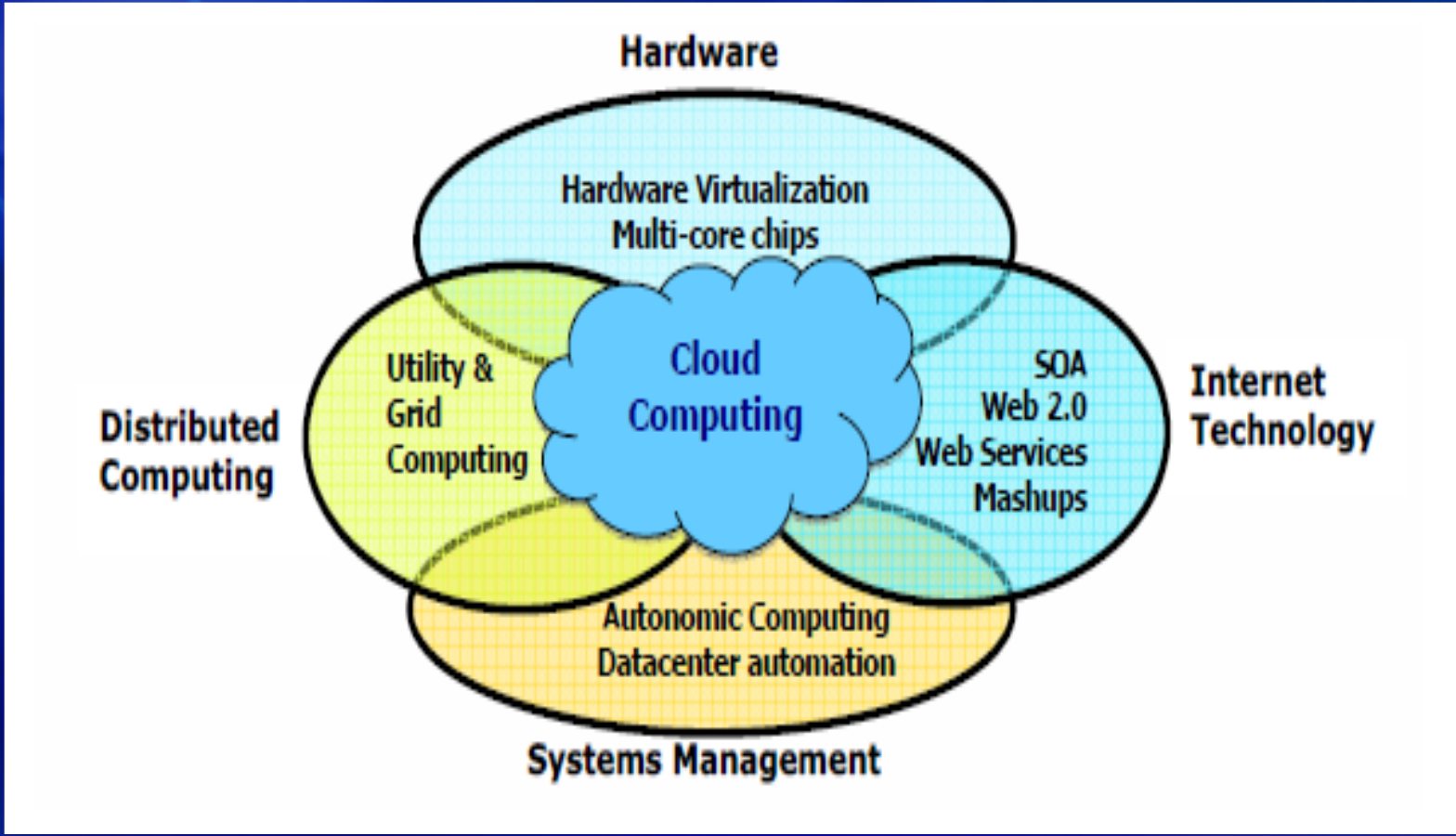


(o) Fraud detection

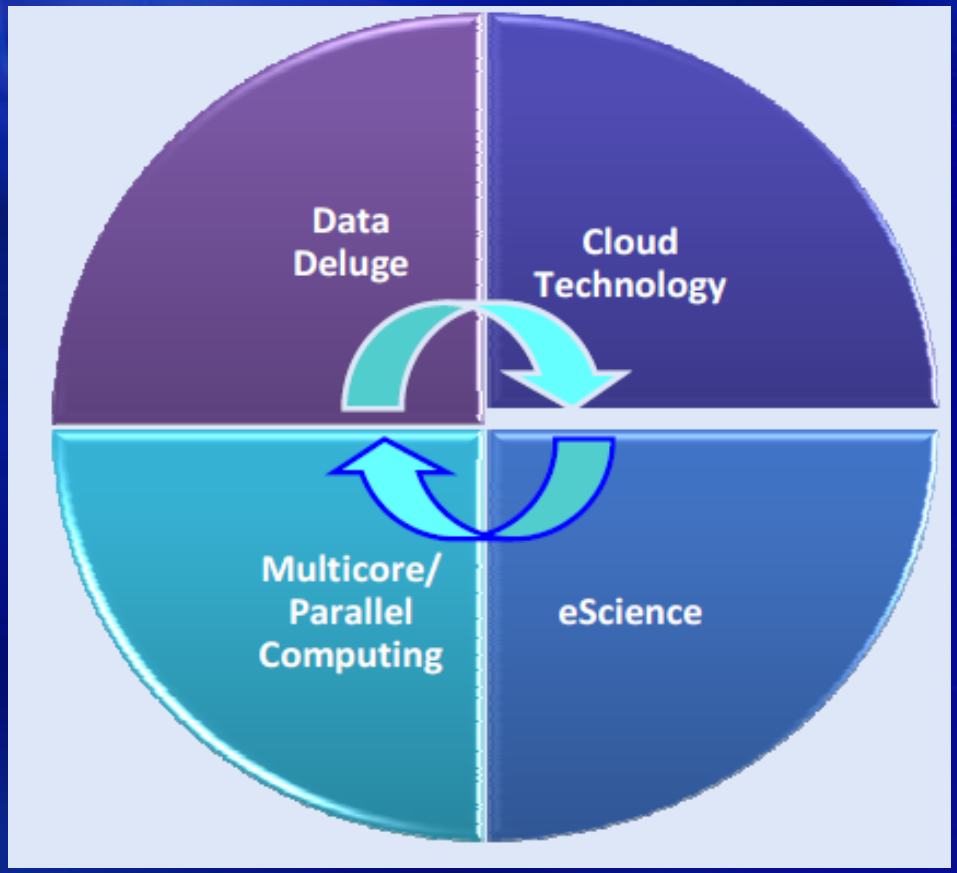


(p) Sell and buy prediction

Data deluge enabling new challenges



Interactions among 4 technical challenges: Data deluge, cloud technology, eScience, and multicore/parallel computing



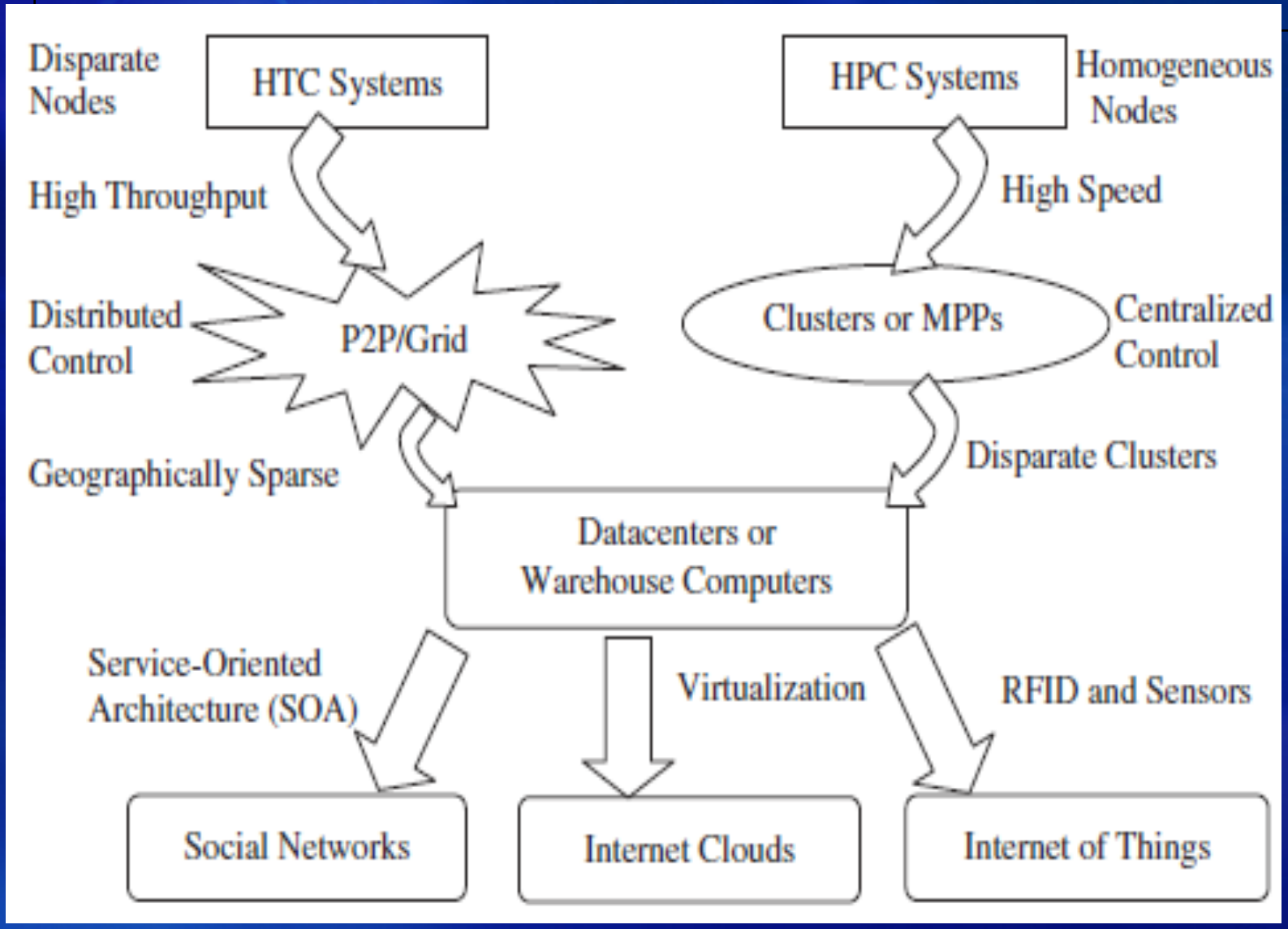
You have to attend all lectures, read books and some selected papers, do the homework, perform the cloud experiments, write the project report, and do well in quizzes, and mid-term and final exams.

Do not take this course if do not have the time to dedicate yourself to it.

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 have laid the foundation of cloud computing and have 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

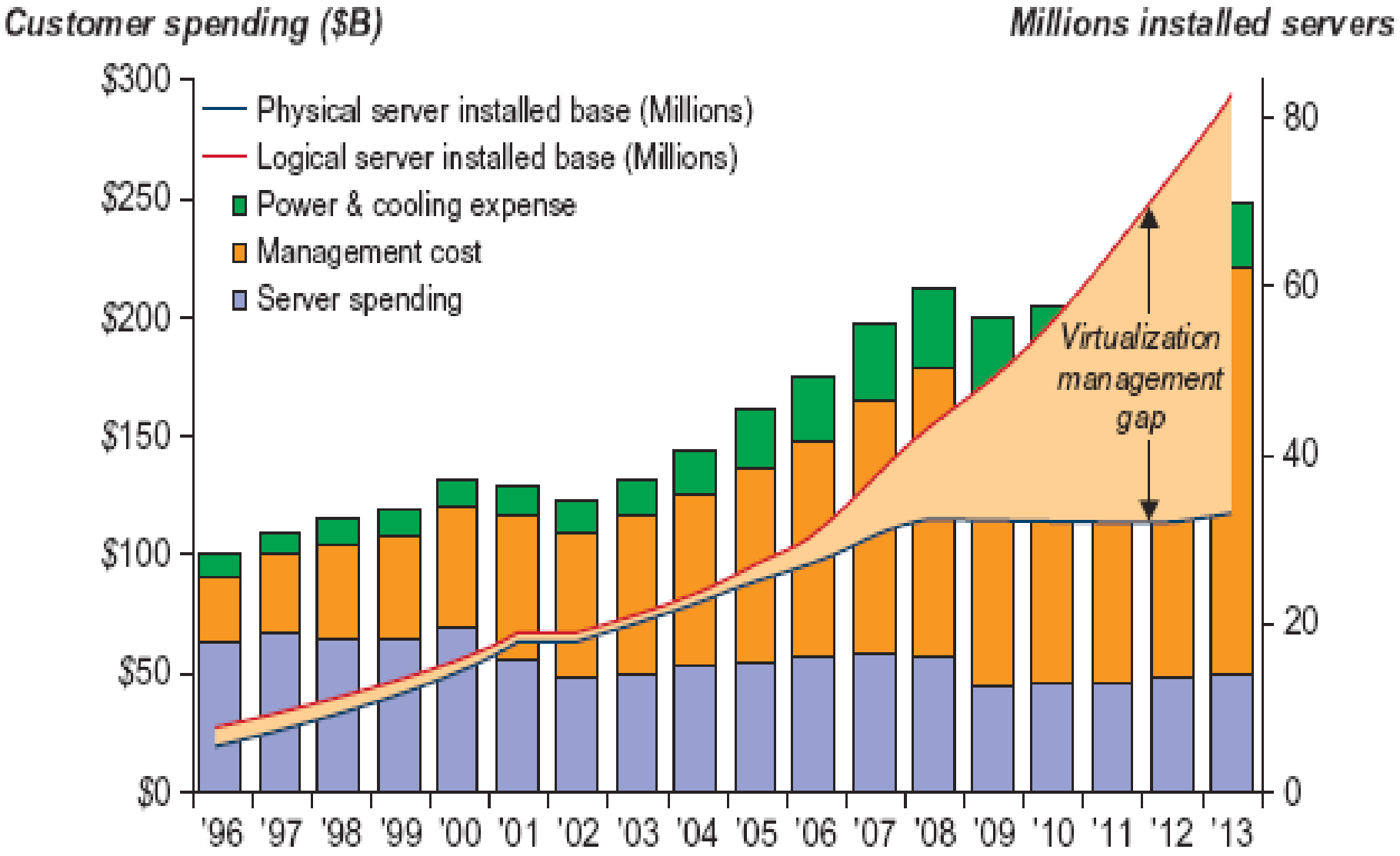


HPC: High-Performance Computing

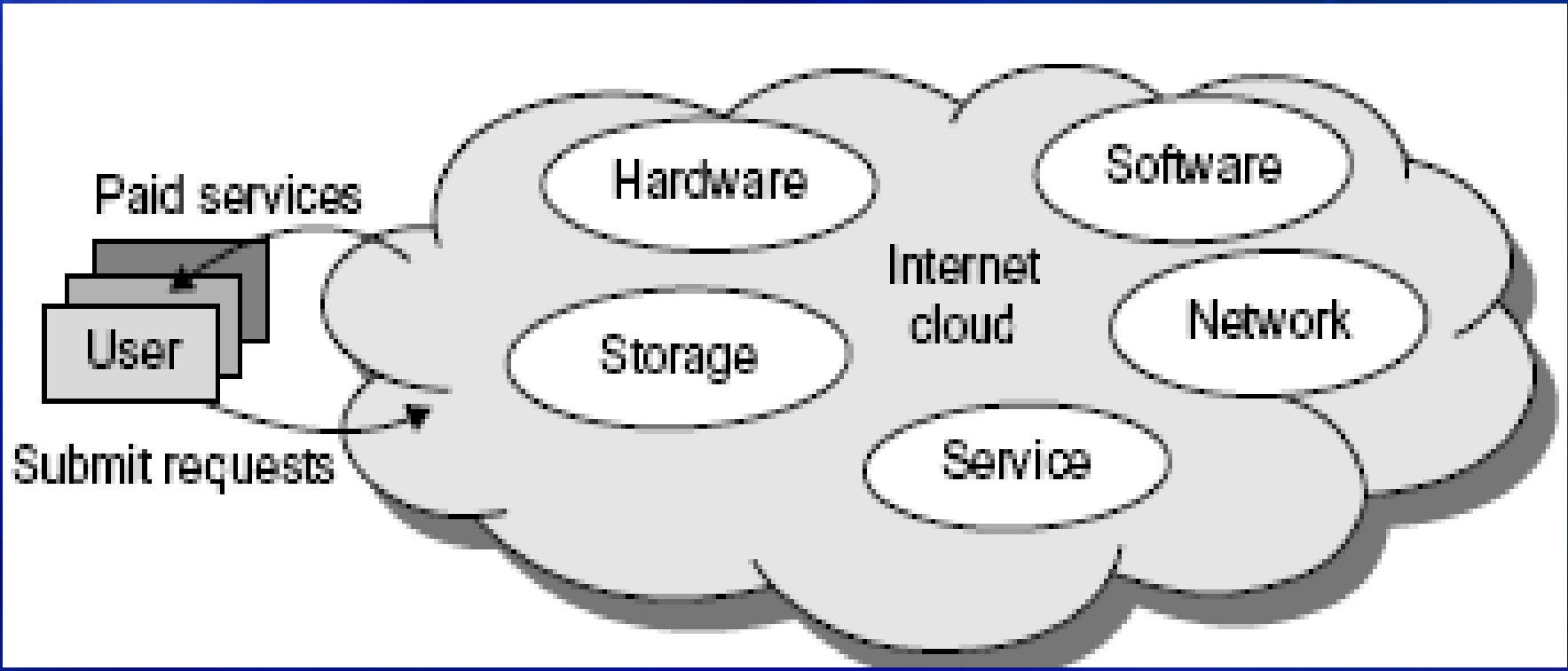
HTC: High-Throughput Computing

P2P: Peer-to-Peer

MPP: Massively Parallel Processors


Customer spending ($B)
Millions installed servers
$300
$250
$200
$150
$100
$50
$0
80
60
40
20
0
Physical server installed base (Millions)
Logical server installed base (Millions)
Power & cooling expense
Management cost
Server spending
Virtualization management gap
'96 '97 '98 '99 '00 '01 '02 '03 '04 '05 '06 '07 '08 '09 '10 '11 '12 '13

Basic concept of Internet clouds



From clusters, P2P networks, and grids, to clouds

Table 1.3
Classification of parallel and distributed computing systems.

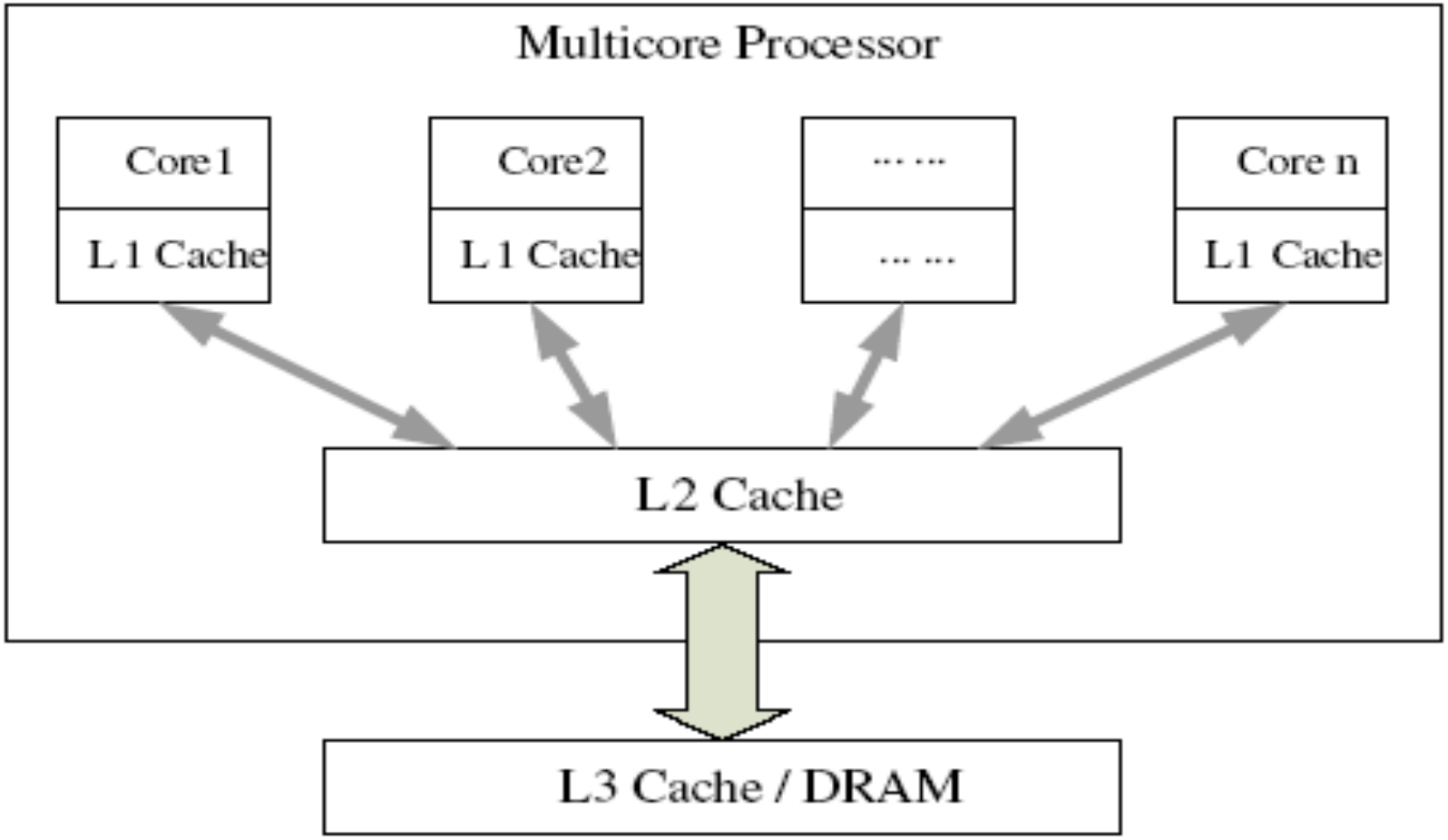
Functionality, Applications	Computer Clusters	Peer-to-Peer Networks	Computational Grids	Cloud Platforms
Architecture, Network Connectivity, and Size	Network of compute nodes interconnected by SAN, LAN, or WAN, hierarchically	Flexible network of client machines logically connected by an overlay network	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

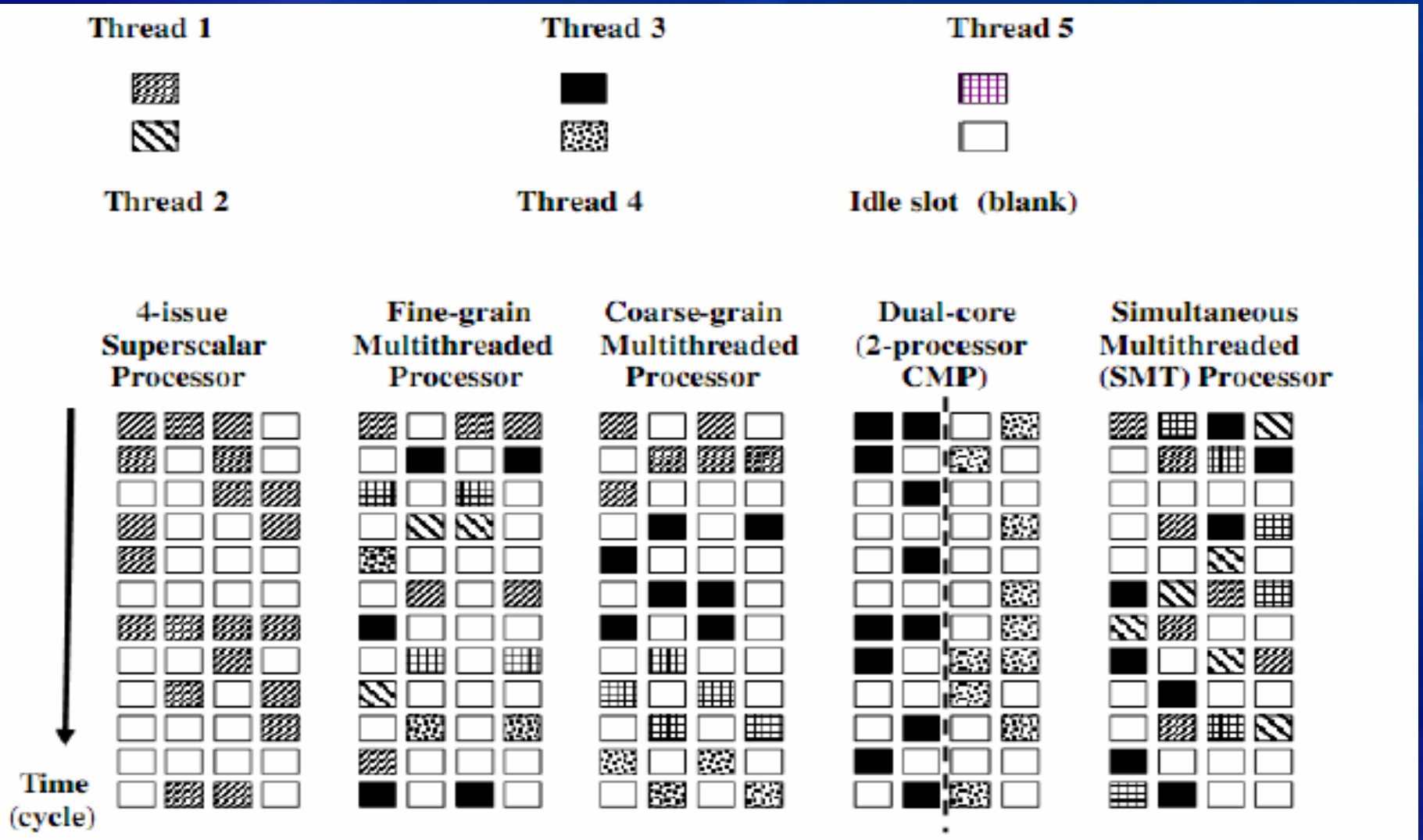
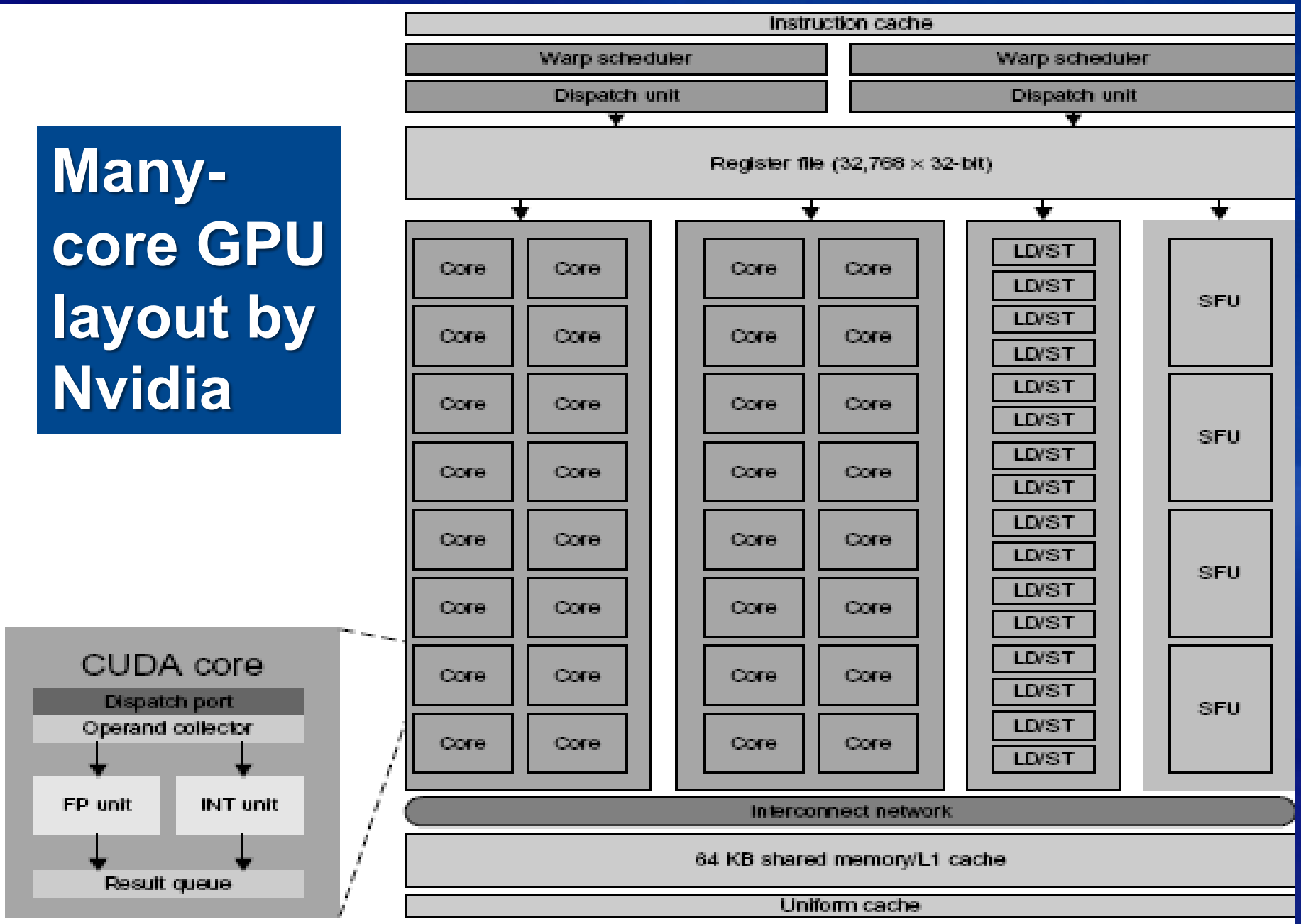


Figure 1. 8 Five micro-architectures that are current in use in modern processors that exploit both ILP and TLP supported by multicore and multithreading technologies

Many-core GPU layout by Nvidia



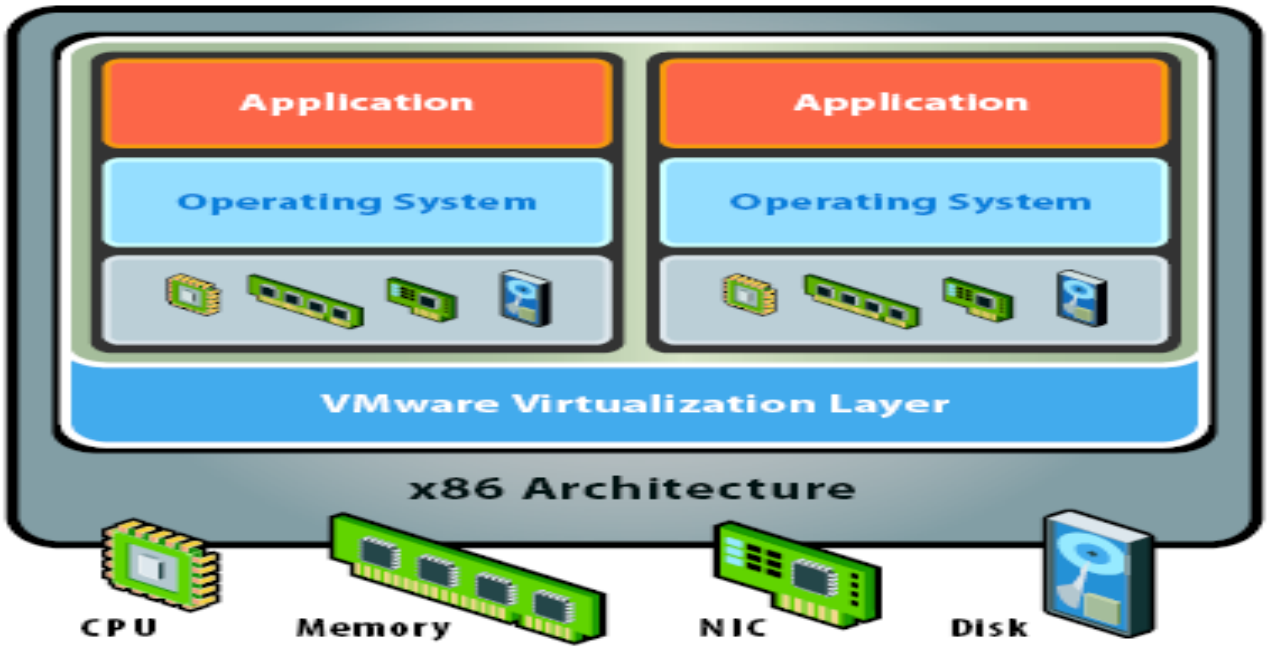
Intel building the eco-system of AI or DL chips

Environment	Brief Description
User-end	Human-machine interaction (Intel® Edison platform, Intel® Cedar Trail platform, Intel® RealSense™ technology)
Server-end	Xeon E5 v4 series CPU; Xeon Phi™ Product Family
Software Service	Intel® Math Kernel Library(Intel® MKL); Intel® Data Analytics Acceleration Library (Intel® DAAL)
Extending Performance	Purchased Altera Corporation; Integrated chip featuring Xeon core and FPGA
Business Acquisition	Purchased high-tech companies: Nervana, Movidius, Itseez, etc.

Nvidia GPUs for machine learning use

GPU Chip Model	Targeted applications in datacenters or cloud computing arena
Tesla P100	Deep learning training accelerator for the data center
Tesla P40/P4	Energy-efficient inference accelerator for deep learning
Jstson TKI1TX1	The embedded AI supercomputer for intelligent devices.
Drive PX2	Scalable in-vehicle AI supercomputer for autonomous driving

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

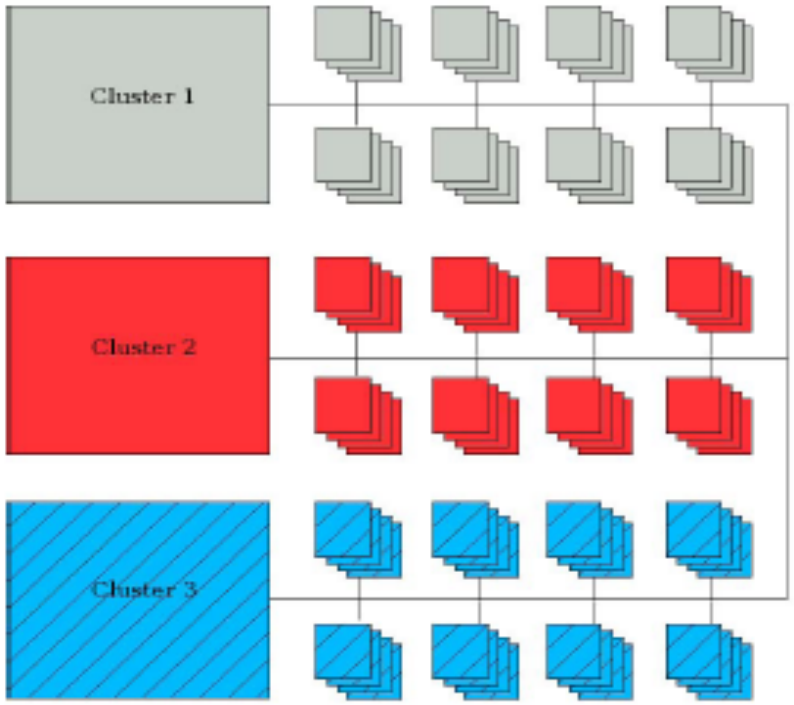


Fig. 1. A Campus Area Grid

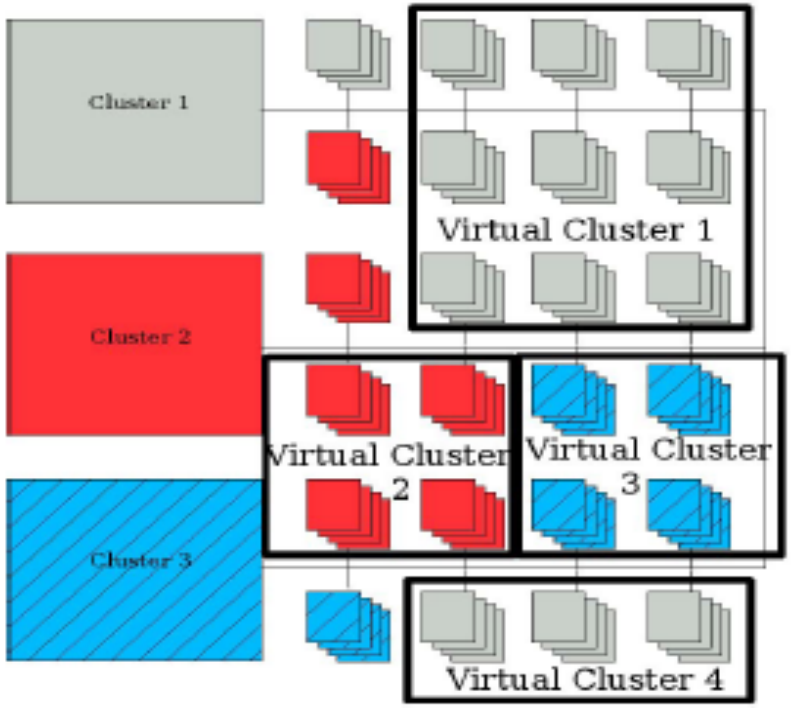


Fig. 2. Virtual machines in a cluster environment

The cloud

- Historical roots in today's Internet applications
  - Search, email, social networks
  - File storage (Live Mesh, Mobile Me, Flicker, etc.)
- 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)

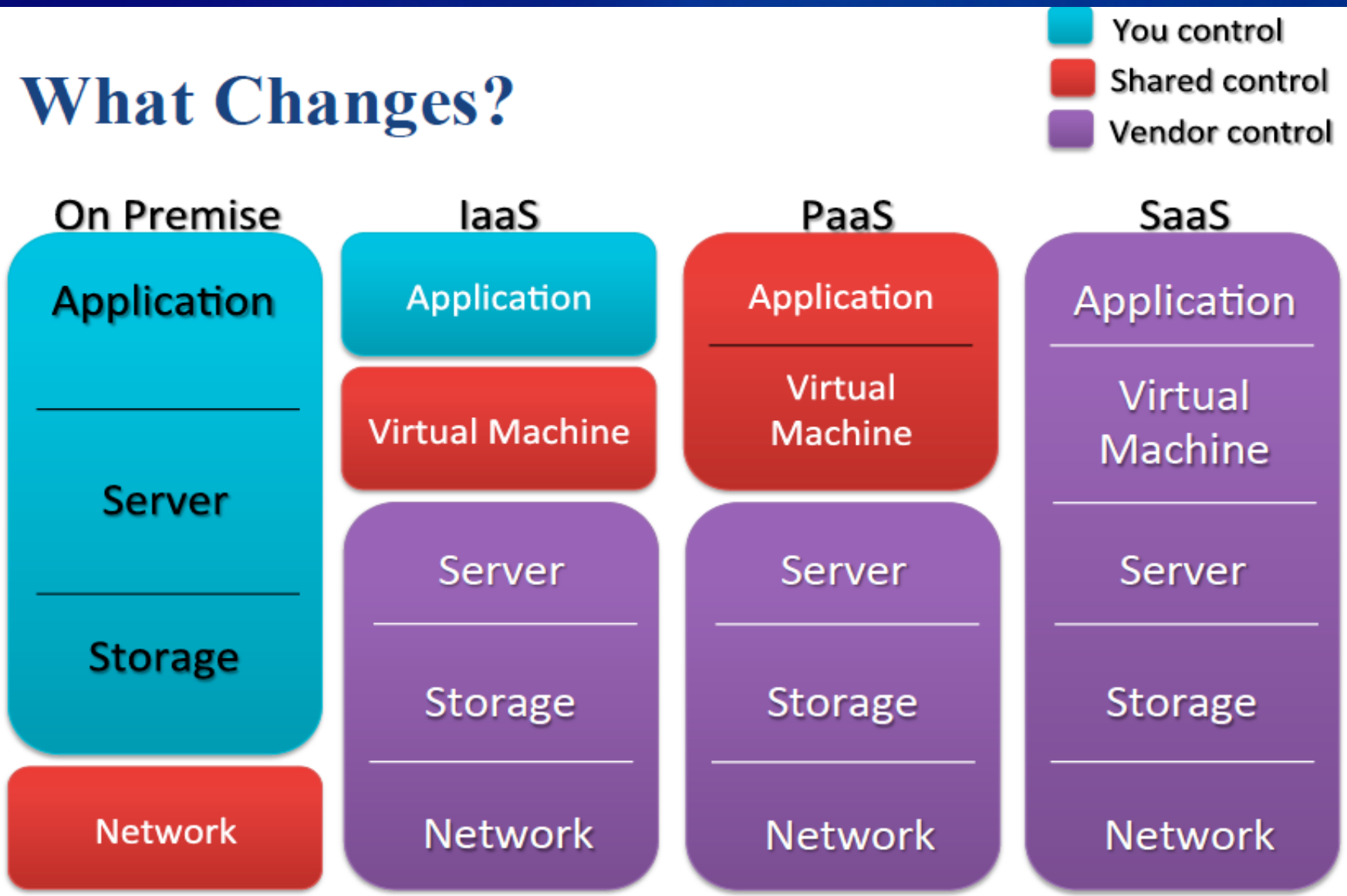
Every 18 months?

- **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

Lecture 2: Cloud architecture and service models, Aug. 23, 2017

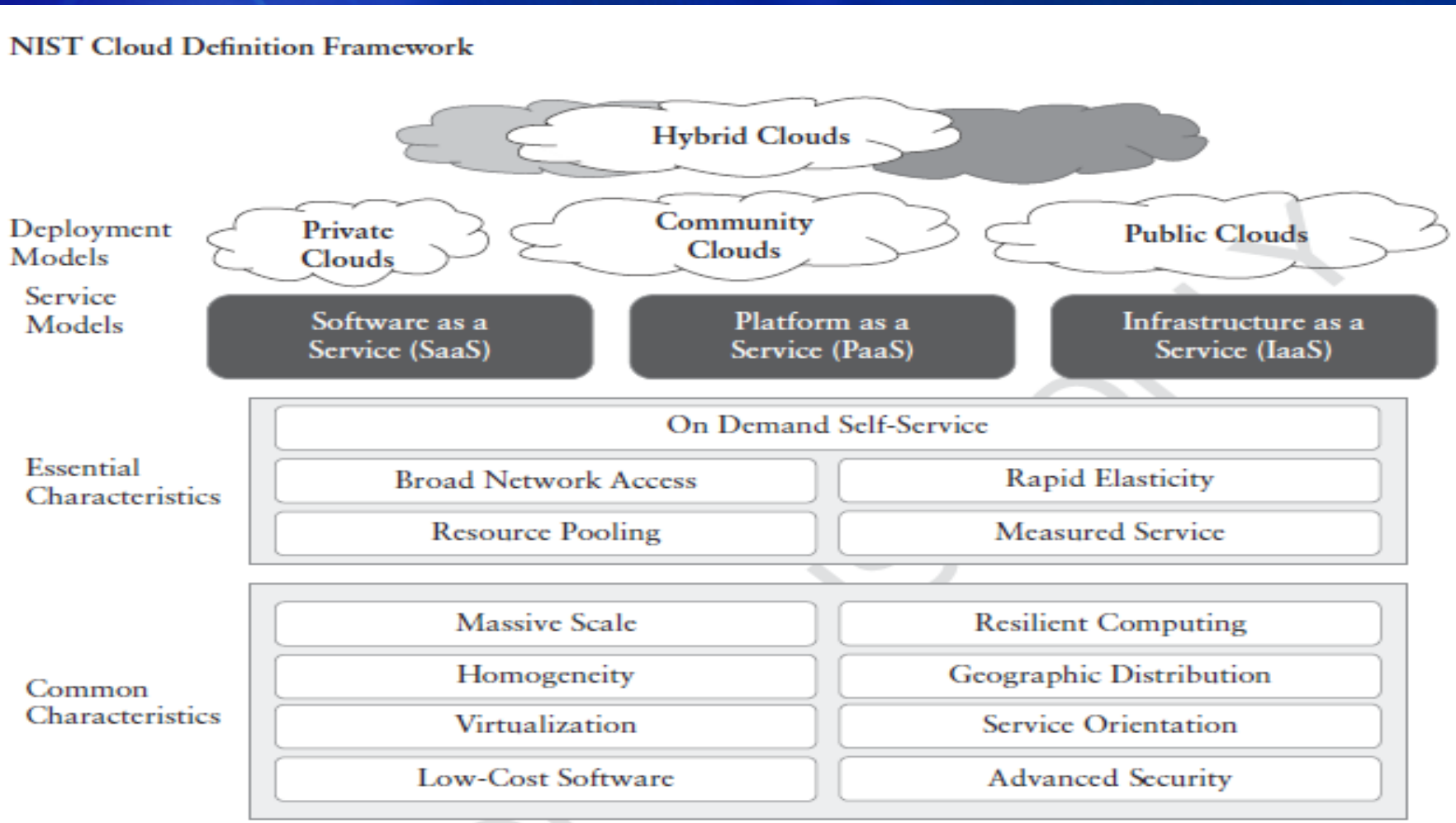


Figure 1.8
Public, private, community, and hybrid clouds. Courtesy of National Institute of Standards and Technology, 2013.

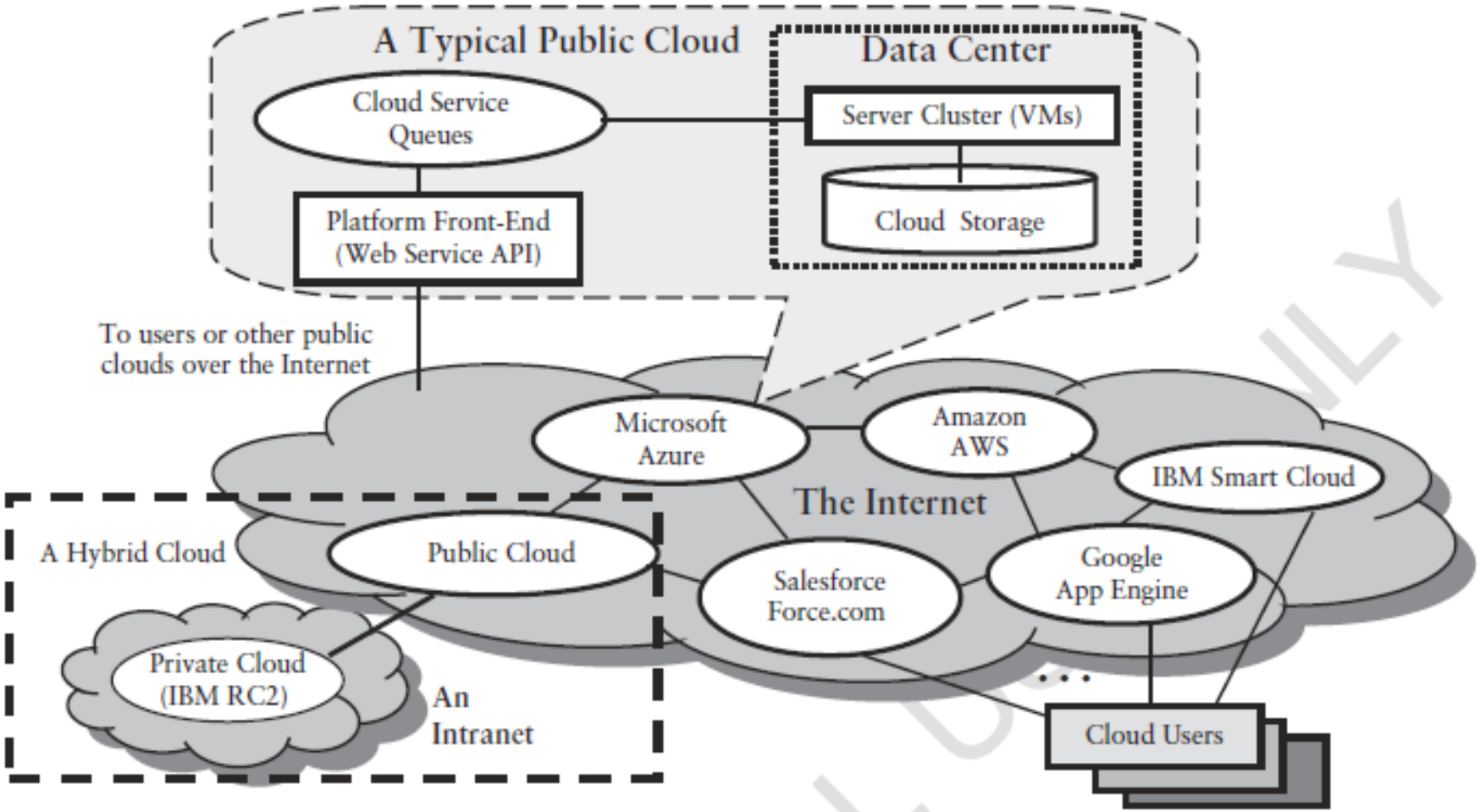
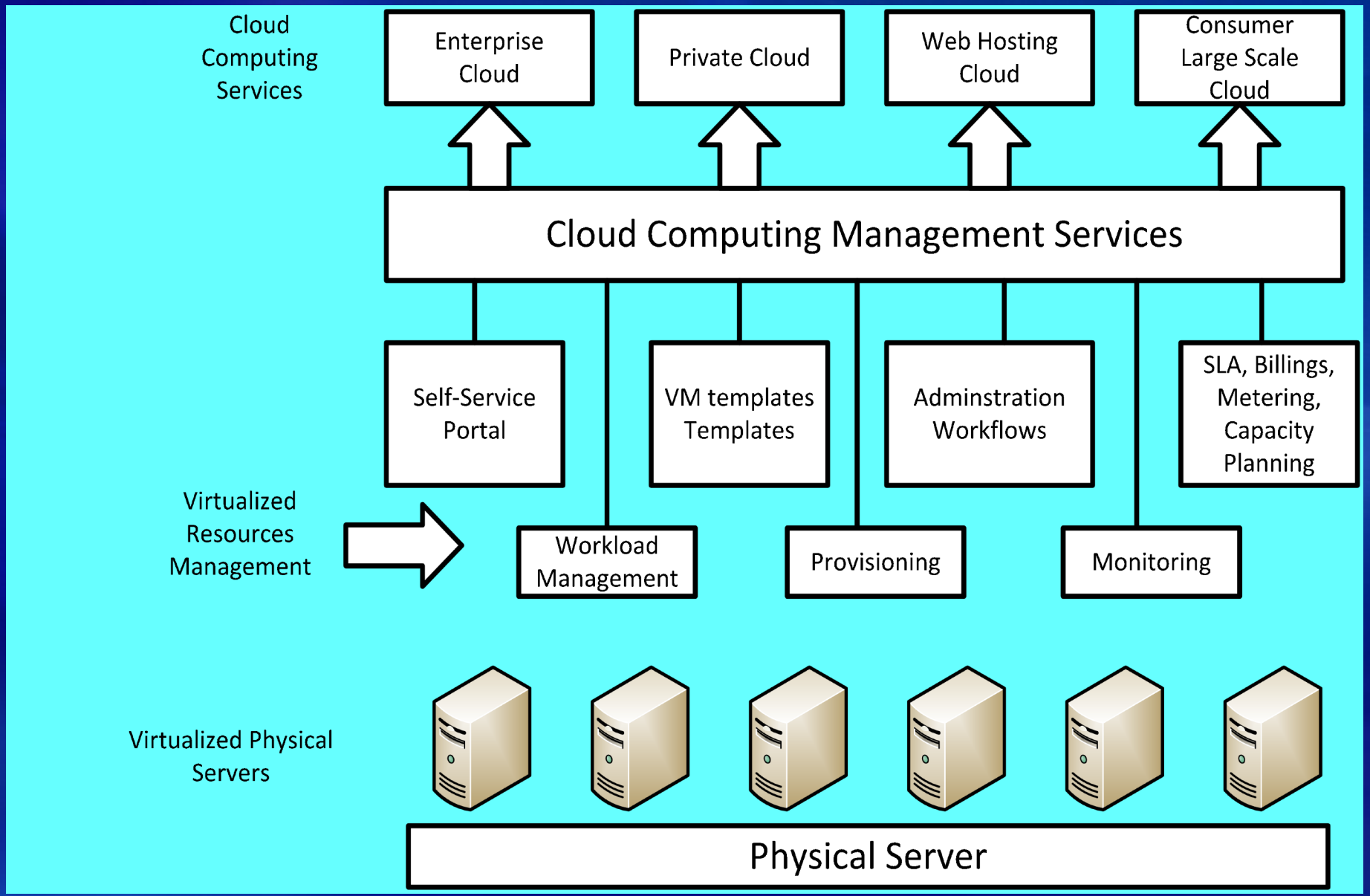
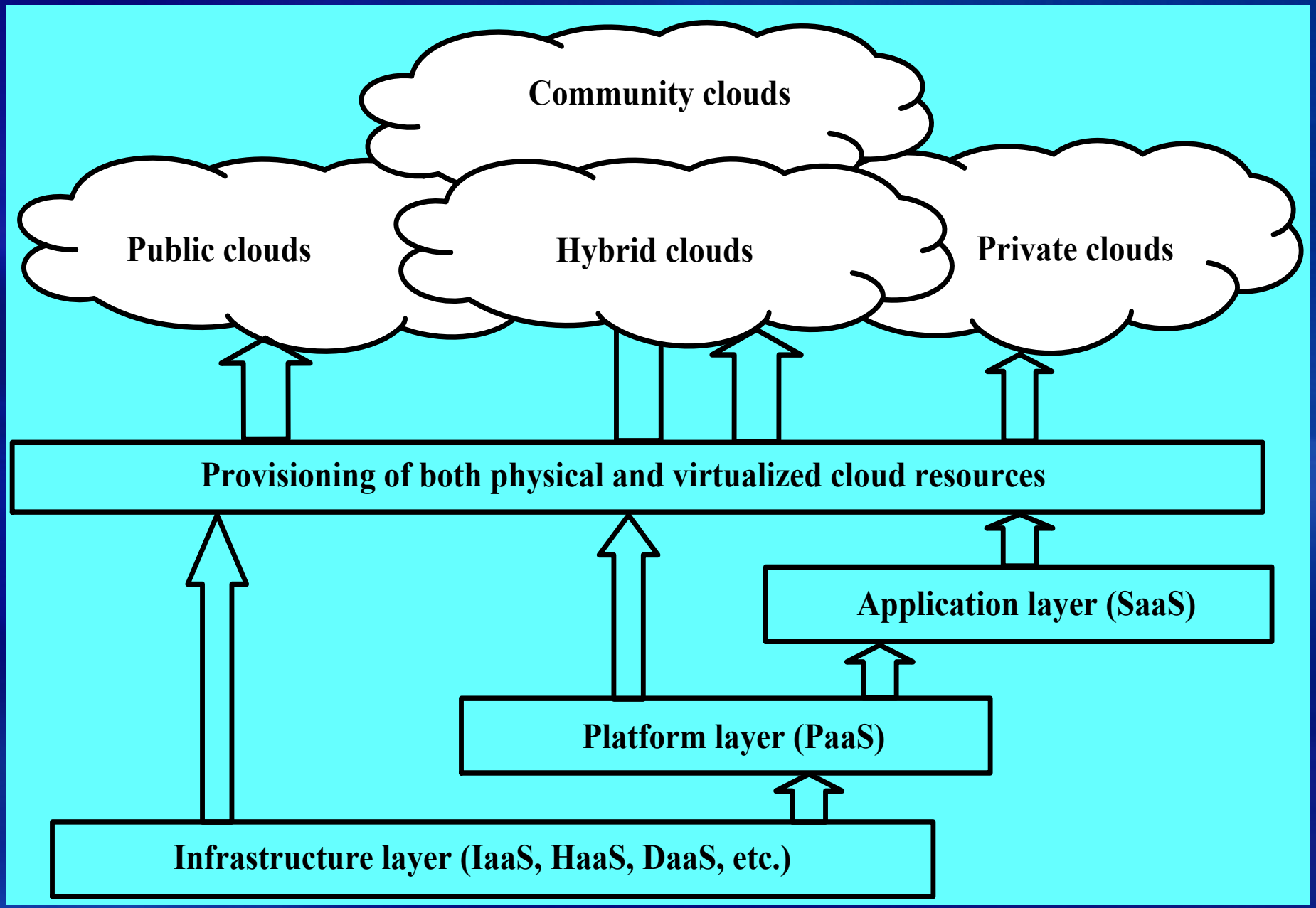
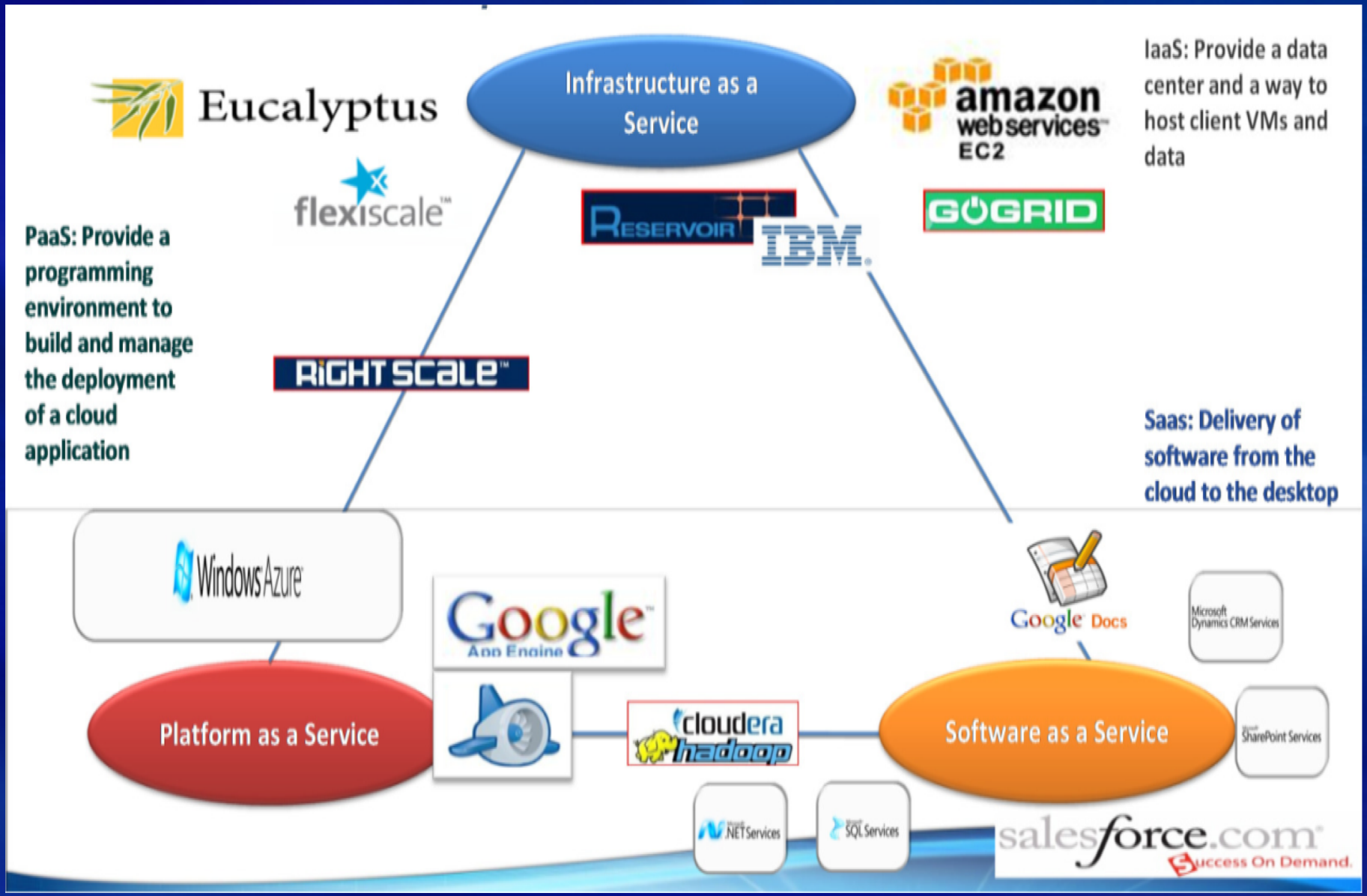
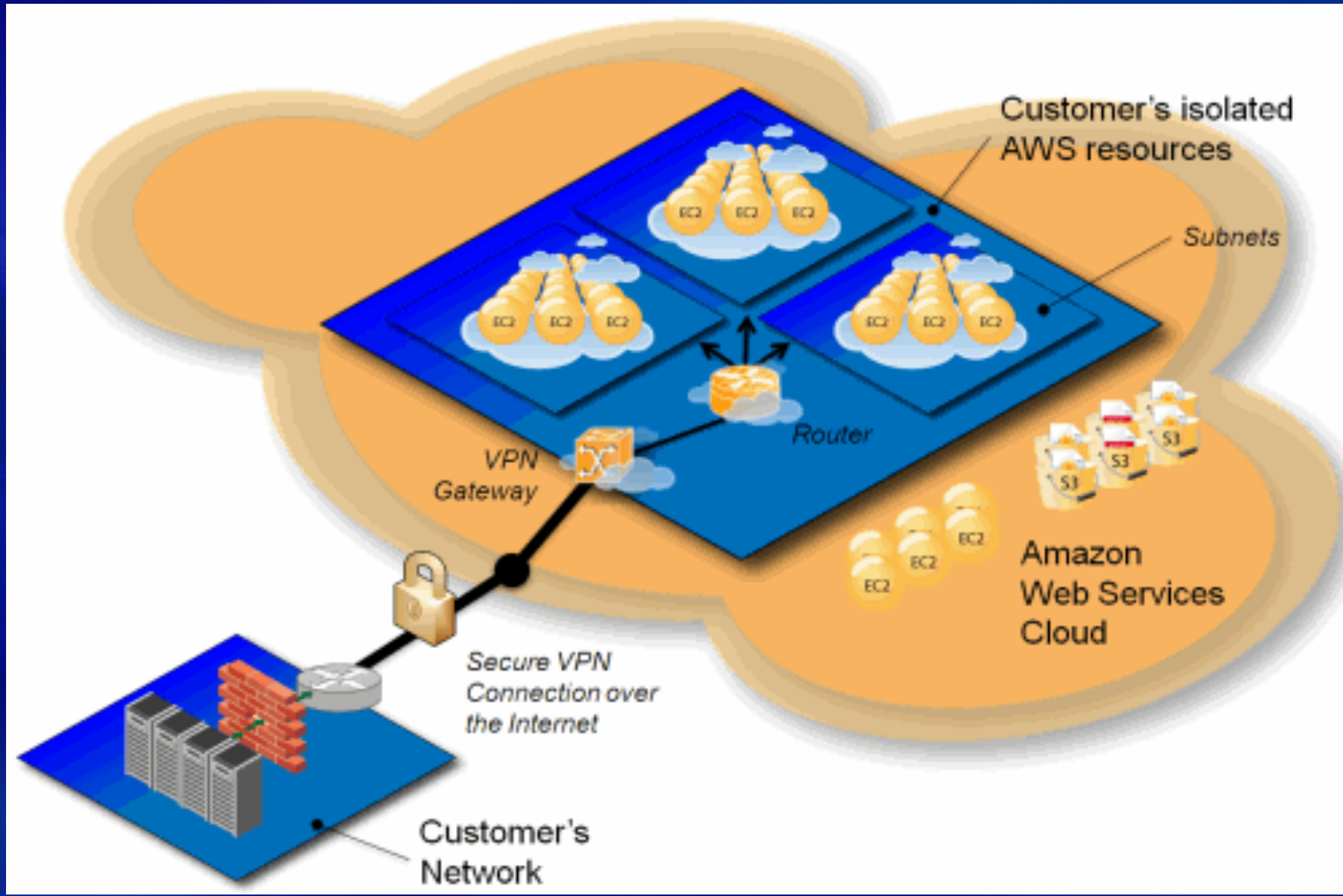


Figure 1.7
Public, private, and hybrid clouds. The callout box shows the architecture of a typical public cloud. A private cloud is built within an Intranet. A hybrid cloud involves both types in its operation range. Users access the clouds from a web browser or through a special API tool.


Cloud Computing Services
Enterprise Cloud
Private Cloud
Web Hosting Cloud
Consumer Large Scale Cloud
Cloud Computing Management Services
Self-Service Portal
VM templates Templates
Adminstration Workflows
SLA, Billings, Metering, Capacity Planning
Virtualized Resources Management
Workload Management
Provisioning
Monitoring
Virtualized Physical Servers
Physical Server


Community clouds
Public clouds
Hybrid clouds
Private clouds
Provisioning of both physical and virtualized cloud resources
Application layer (SaaS)
Platform layer (PaaS)
Infrastructure layer (IaaS, HaaS, DaaS, etc.)




Customer's isolated
AWS resources
Subnets
Router
VPN
Gateway
Amazon
Web Services
Cloud
Secure VPN
Connection over
the Internet
Customer's
Network

Market share of various cloud platforms

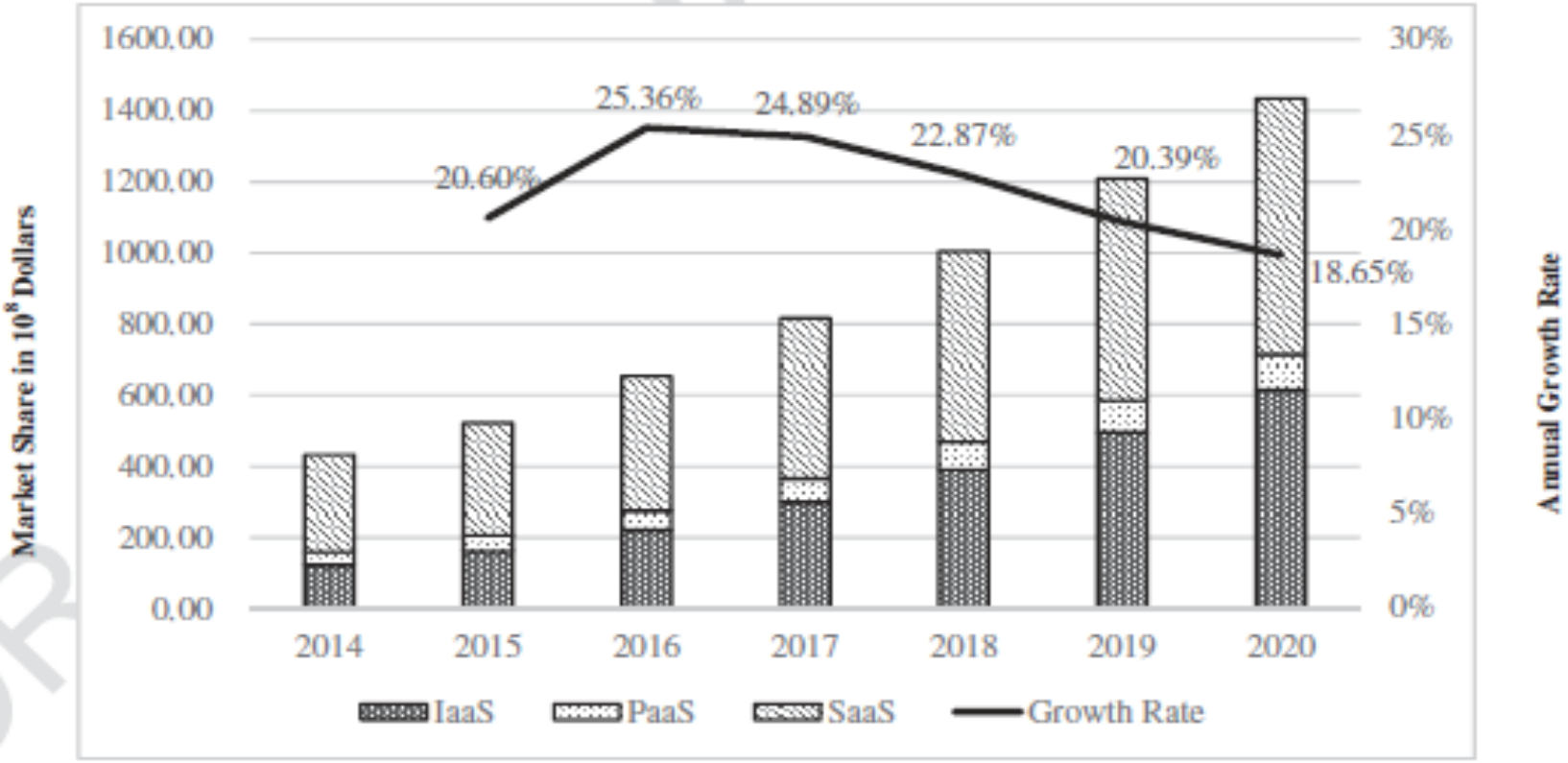


Figure 1.13
Worldwide distribution of cloud service models and the growth rate based on projections by Gartner Research from 2014–2020.

Cloud users vs. cloud providers

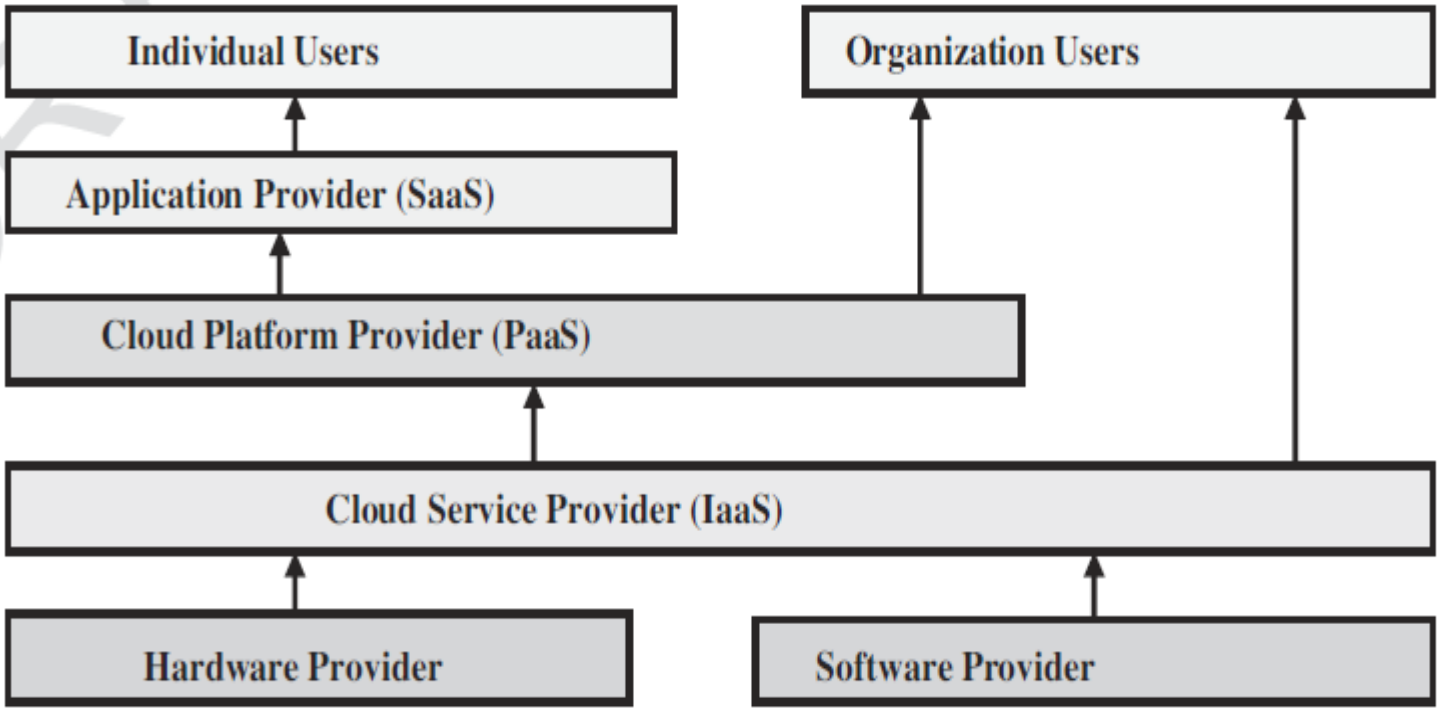


Figure 1.15
Individual versus organization users of cloud computing and their services, hardware, and software providers.

Table 1.4. Cloud providers and vendors

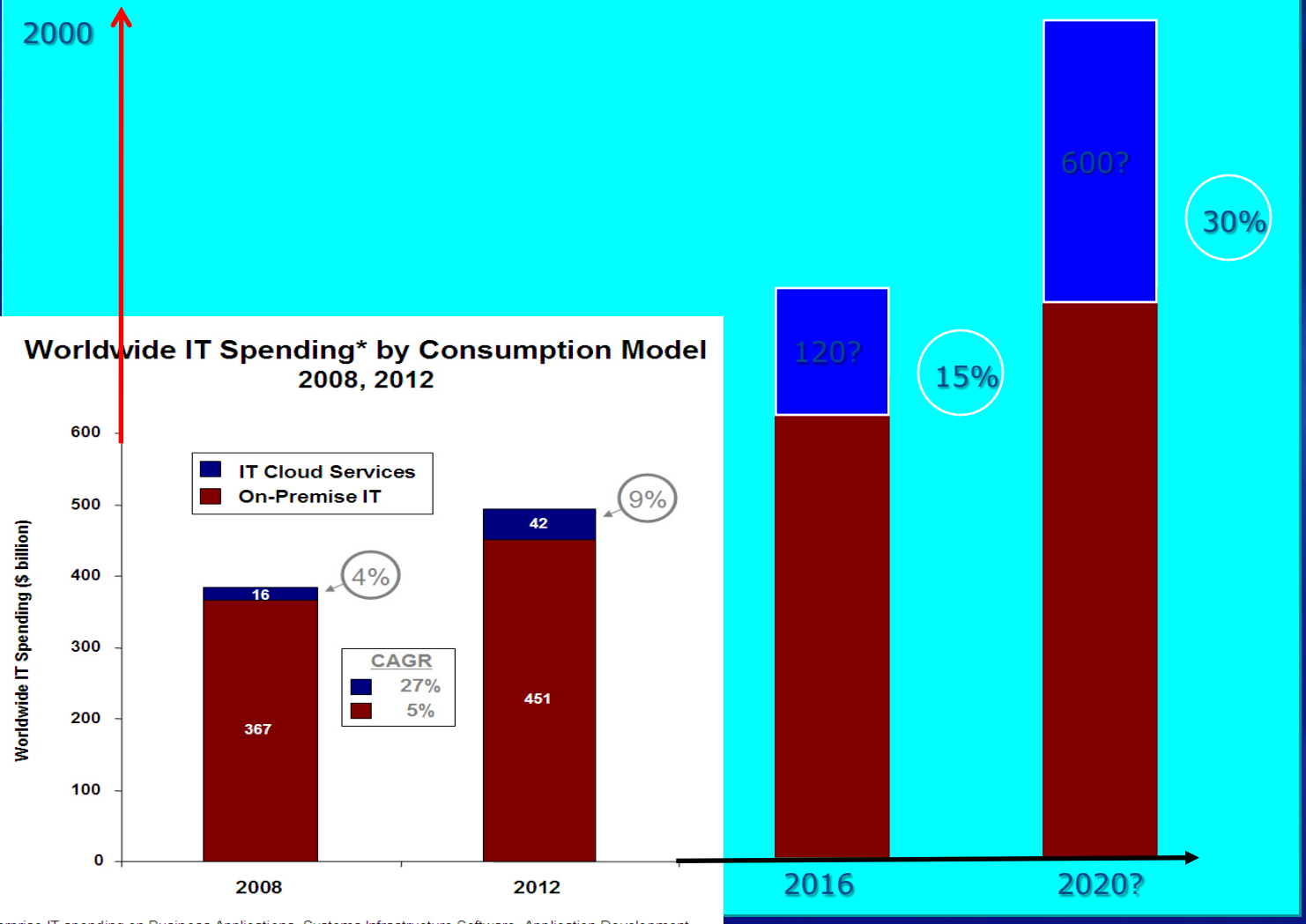
Cloud Players	IaaS	PaaS	SaaS
IT adminis-trators and cloud providers	Monitor SLAs	Monitor SLAs and enable Service Platforms	Monitor SLAs and deploy software
Software developers (vendors)	To deploy and store data	Enabling Platforms via configurator and APIs	Develop and Deploy Software

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 business potential:

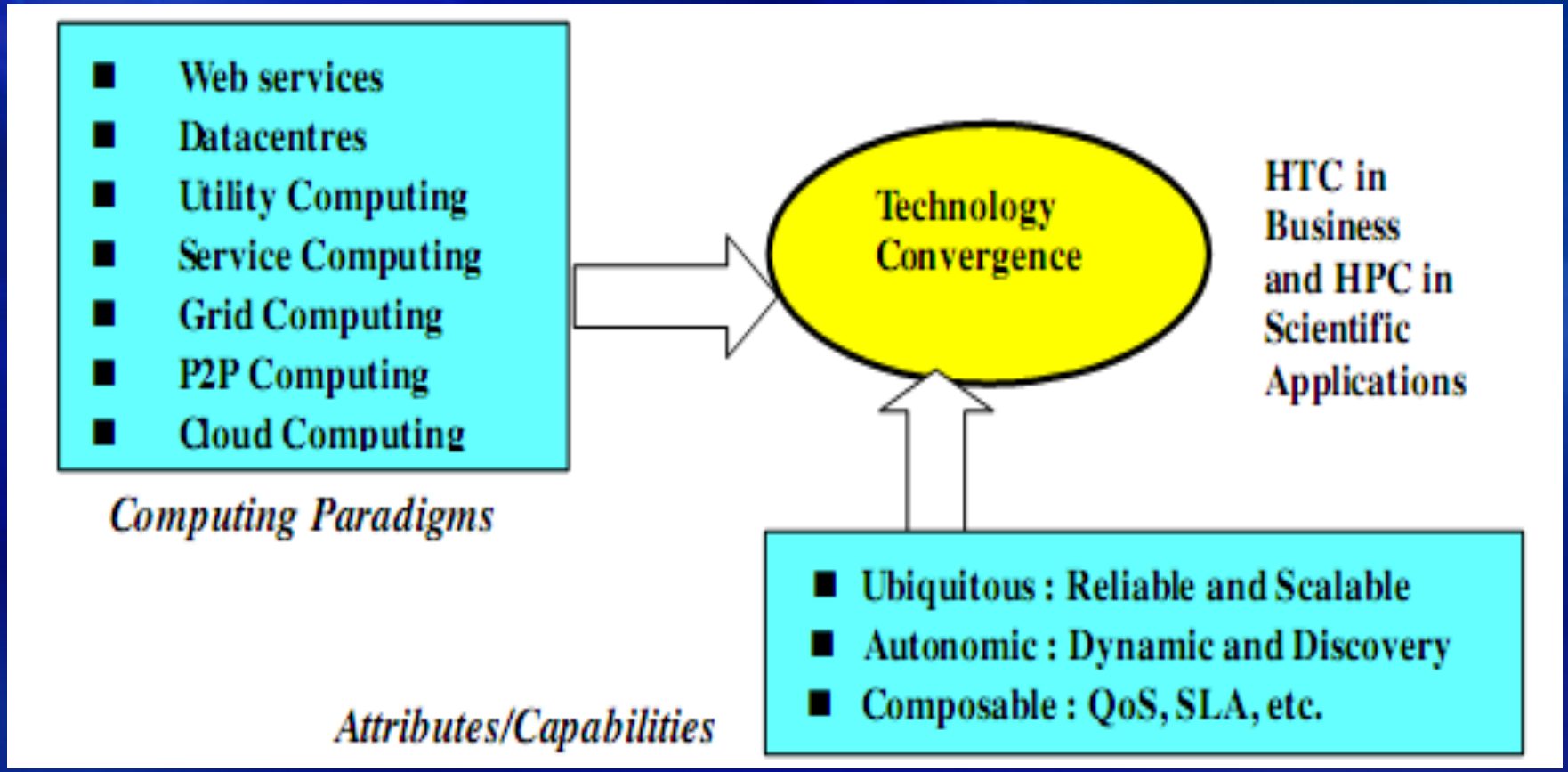
A trillion-dollar-business/year by 2020?



* 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

2015 cloud technologies

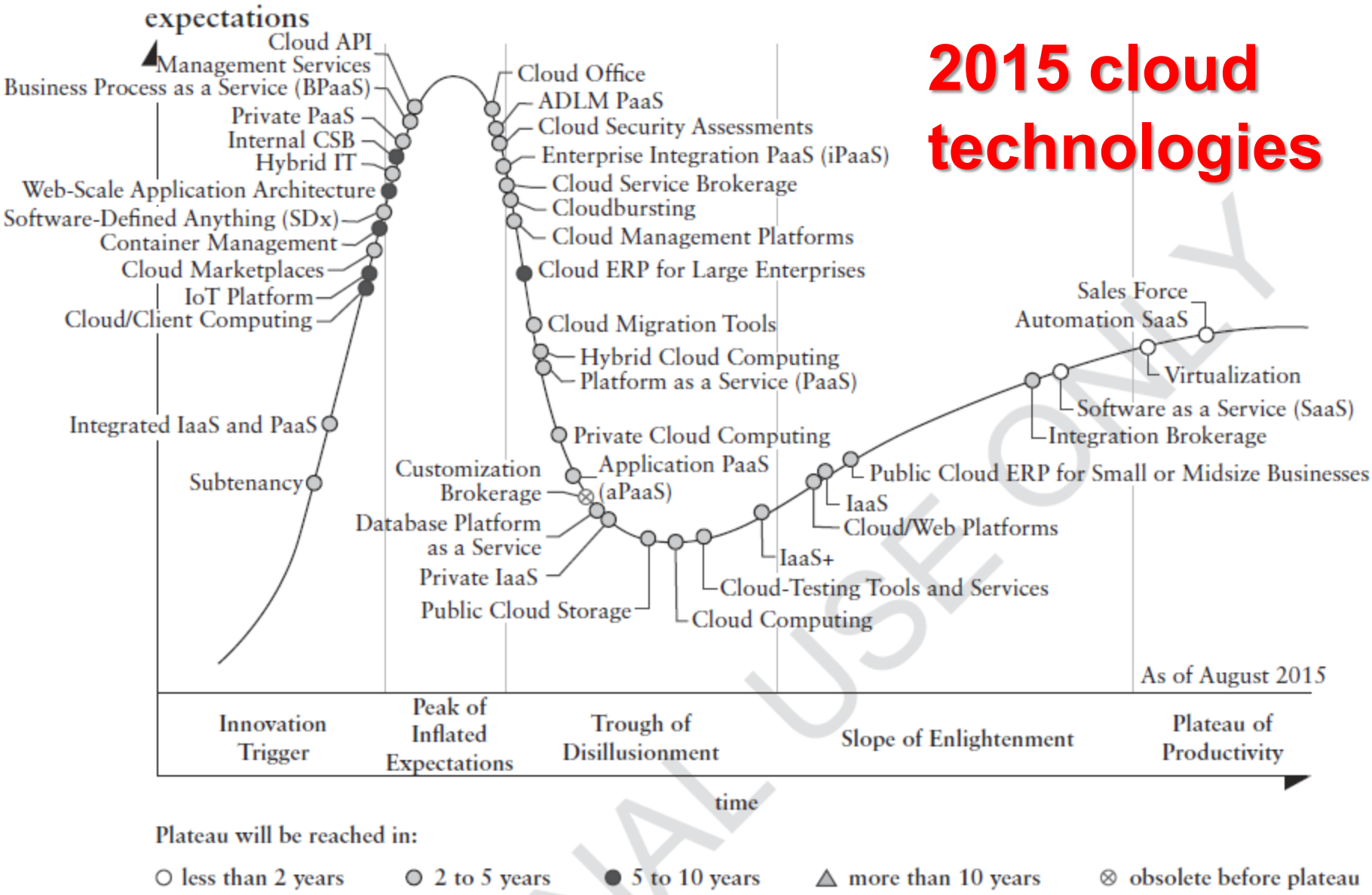


Figure 1.17
Gartner's Hype Cycle for cloud computing in August 2015. Reprinted with permission from Gartner Research, Inc.

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

Services-Oriented Architecture (SOA)

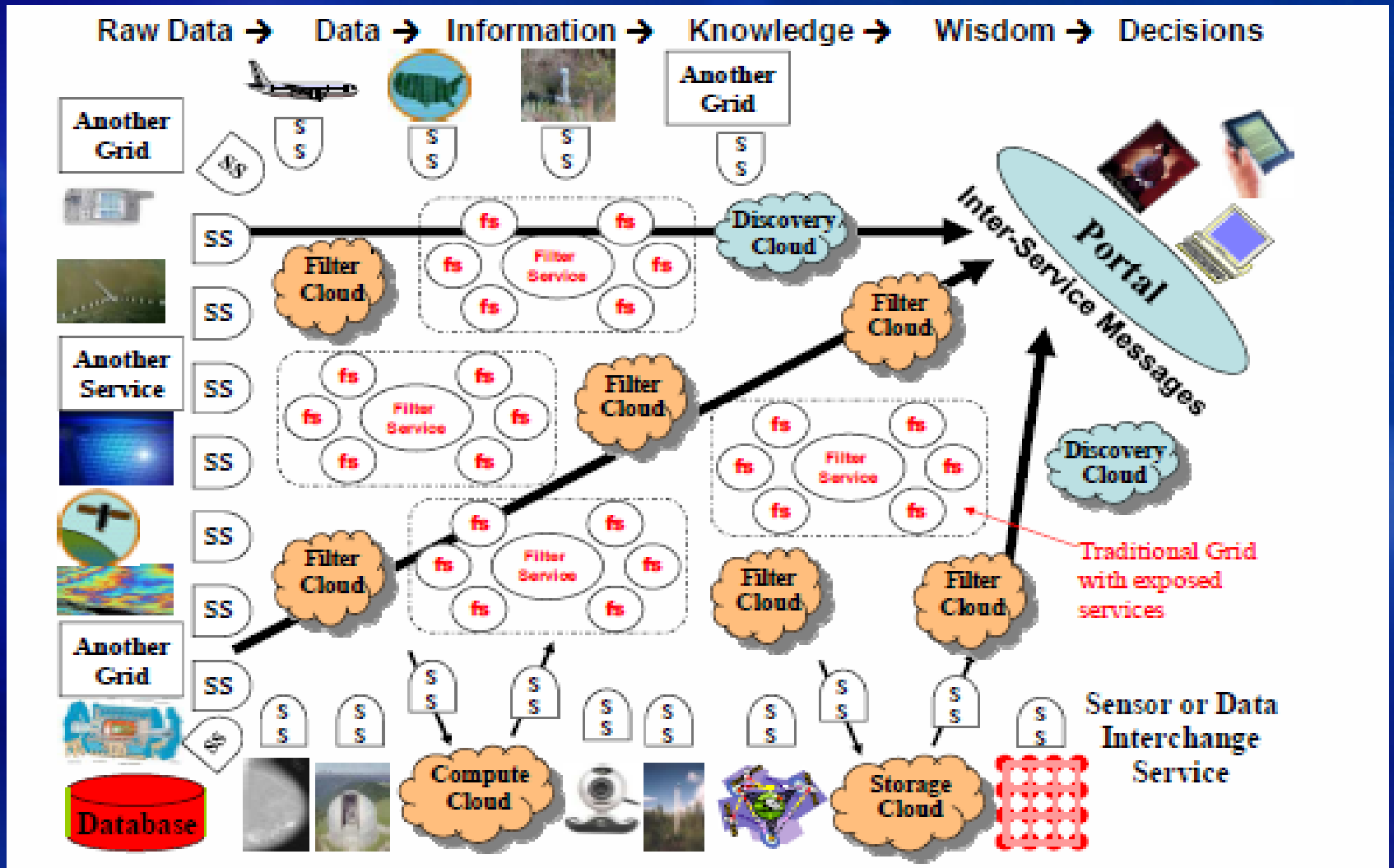


Figure 1.23 The evolution of service-oriented architecture: Grids of Clouds and Grids where SS refers to Sensor Service and fs to a filter or transforming service.

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

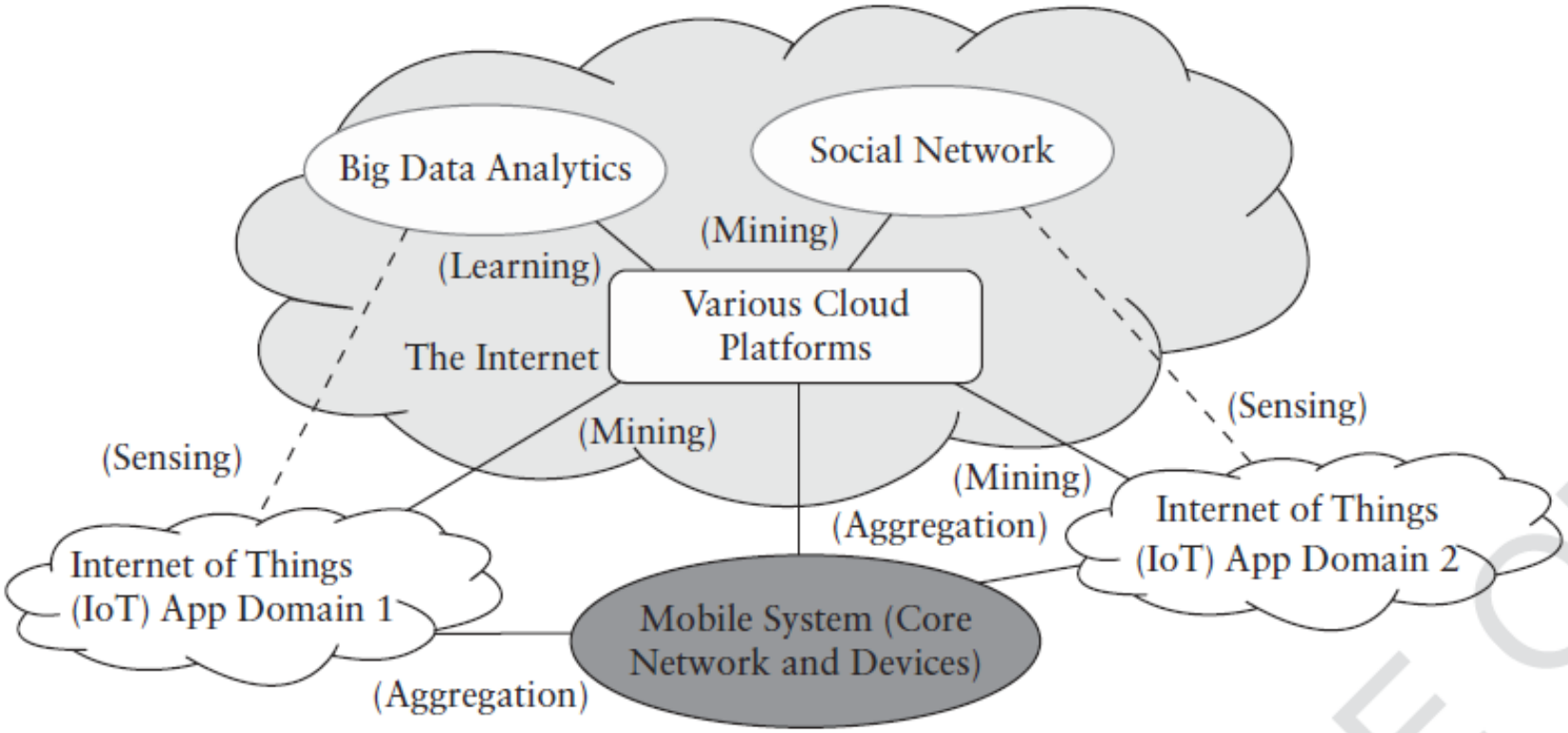


Figure 1.18
Interactions among social networks, mobile systems, big data analytics, and cloud platforms over various Internet of things (IoT) domains.

Transparent cloud computing environment

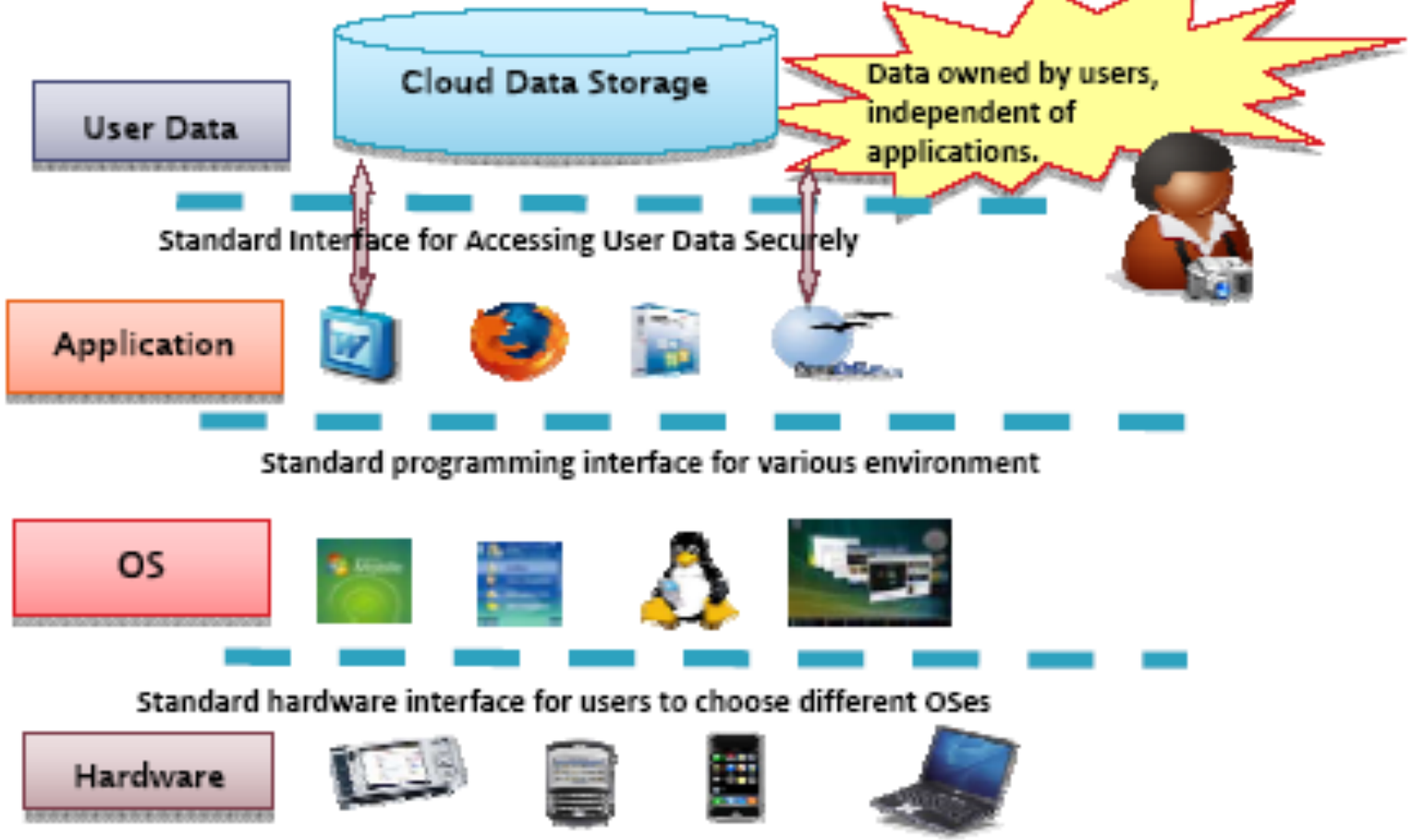


Figure 3 Transparent computing that separates the user data, application, OS, and hardware in time and space – an ideal model for future Cloud platform construction

System availability vs. configuration size

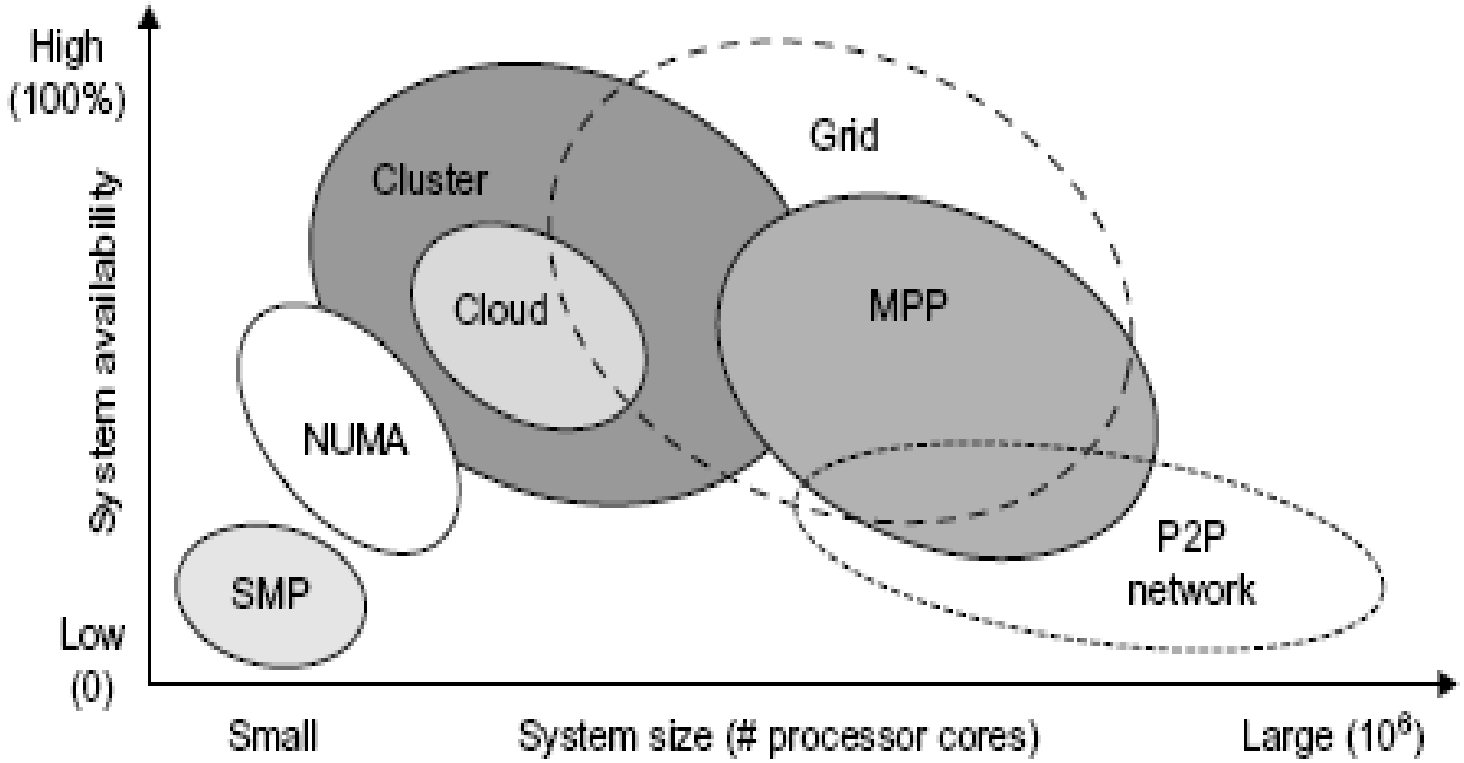
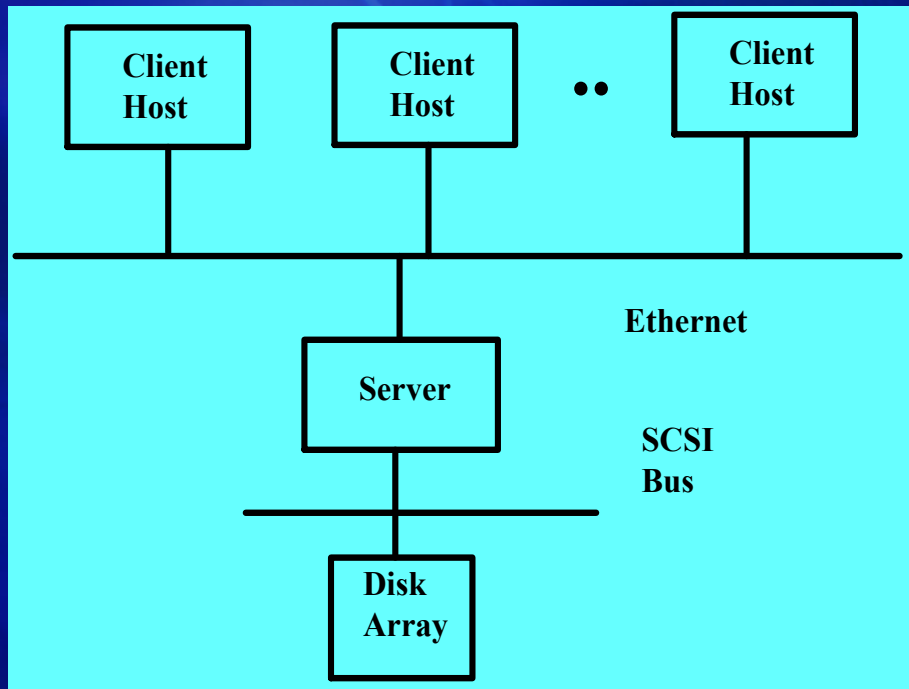
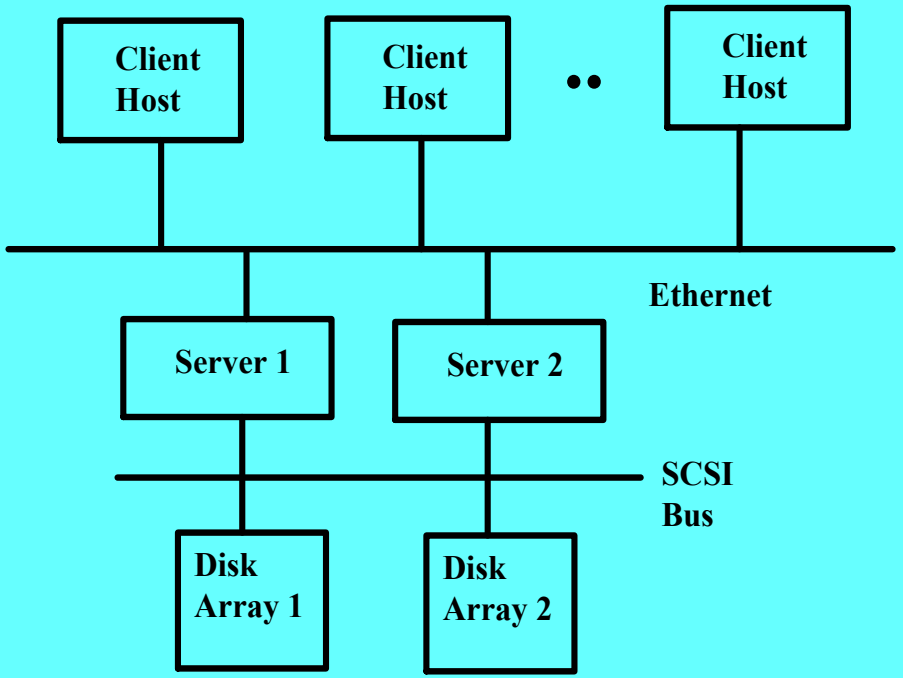


FIGURE 1.24

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

Single point of failure:
The server and the disk





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

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

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:

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

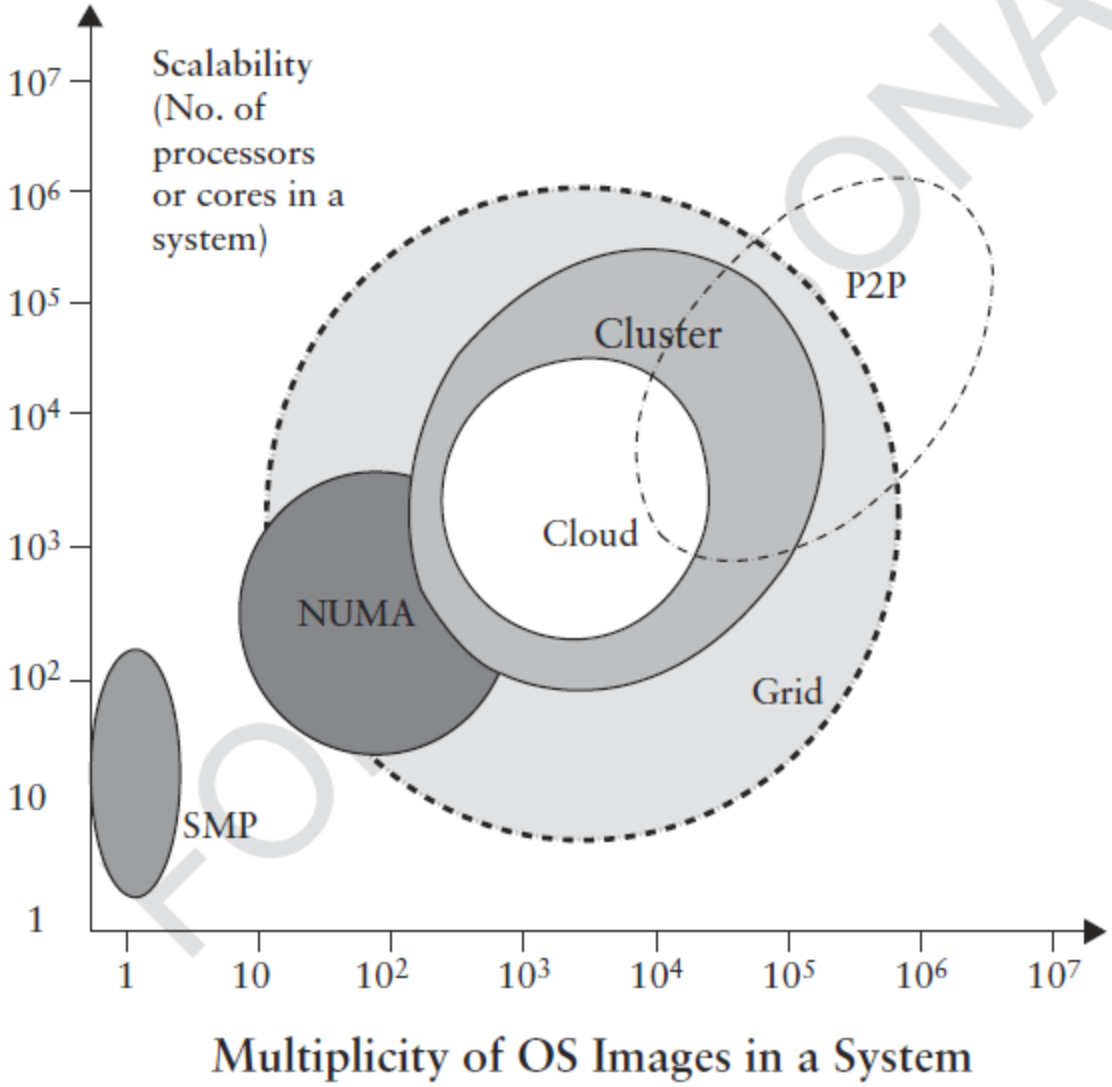


Figure 1.14
System scalability versus multiplicity of OS images based on 2010 technology. Reprinted with permission from K. Hwang and Z. Xu, *Scalable Parallel Computing*, McGraw-Hill, 1998.

System availability analysis

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

$$= \binom{n}{k} p^k (1-p)^{n-k} + \binom{n}{k+1} p^{k+1} (1-p)^{n-k-1}$$

$$+ \cdots + \binom{n}{n-1} p^{n-1} (1-p)^1 + \binom{n}{n} p^n (1-p)^0,$$

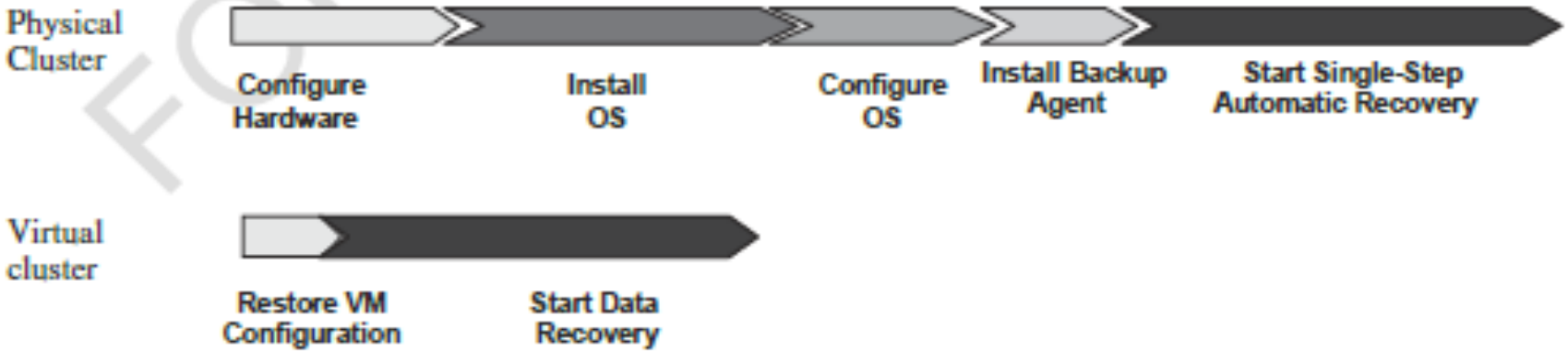
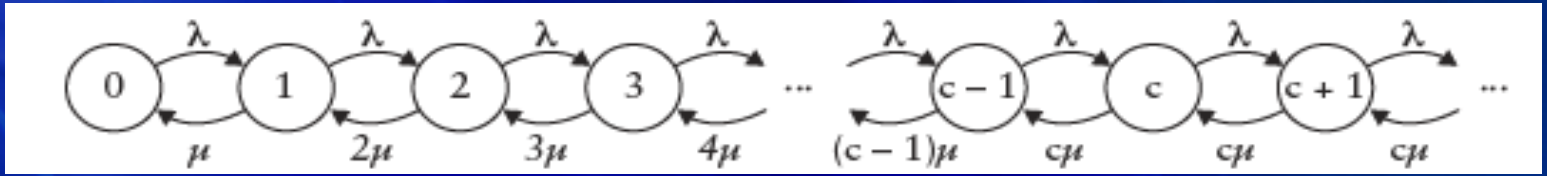


Figure 1.21
Recovery overhead on a physical cluster compared with that of a virtual cluster.

Multiserver cloud service model

The operation of an *m*-server cluster can be modeled by an *M/M/m* queueing system, where *λ* is service task arrival rate and *μ* is the average service rate of task requests. In what follows, we derive the formulas to estimate the average task response time. The parameter *ρ* = *λ* /*mμ* is the probability of having requests in the queue.



We calculate below the wait probability of a newly submitted service request that must wait in the queue, because all servers are busy.

$$P_q = \sum_{k=m}^{\infty} p_k = \frac{p_m}{1-\rho} = p_0 \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho}. \tag{1.8}$$

The average number of service requests in waiting or under execution in S is thus computed by:

$$\overline{N} = \sum_{k=0}^{\infty} k p_k = m\rho + \frac{\rho}{1-\rho} P_q. \tag{1.9}$$

Applying Little's Law, we get the average task response time as follows:

$$\overline{T} = \frac{\overline{N}}{\lambda} = \overline{x}\left(1 + \frac{P_q}{m(1-\rho)}\right) = \overline{x}\left(1 + \frac{p_m}{m(1-\rho)^2}\right). \tag{1.10}$$

The average waiting time of a service request is thus computed by:

$$\overline{W} = \overline{T} - \overline{x} = \frac{p_m}{m(1-\rho)^2}\overline{x}. \tag{1.11}$$

Multiserver cluster optimization

C is the expected charge to s service request, which is a very complex function of many parameters.

$G=\lambda C-(\beta m+\gamma(\lambda\bar{r}\xi s^{\alpha-1}+mP))$ is the expected business gain (i,e. net profit) of a service provider.

The following is profit gain function to be optimized with respect to the cluster size (m) and the server speed (s).

$$C \approx a\bar{r}\left(1-\frac{1}{\left(\sqrt{2\pi m}(1-\rho)\left(e^{\rho}/e\rho\right)^{m}+1\right)}\frac{1}{\left((ms-\lambda\bar{r})(c/s_0-1/s)+1\right)}\right.$$
$$\left.\times\frac{1}{\left((ms-\lambda\bar{r})(a/d+c/s_0-1/s)+1\right)}\right). \quad (1.15)$$

Our ultimate goal is to determine the optimal size m of the server cluster to be used. We optimize the cluster size m by making the derivative of the business gain G zero.

$$\frac{\partial G}{\partial m}=\lambda\frac{\partial C}{\partial m}-(\beta+\lambda P^{*})=0. \quad (1.16)$$

Furthermore, we need to find the *optimal server* speed that can maximize the business gain.

$$\frac{\partial G}{\partial s}=\lambda\frac{\partial C}{\partial s}-\gamma\lambda\bar{r}\xi(\alpha-1)s^{\alpha-2}=0. \quad (1.17)$$

Concluding remarks

- Big data and clouds demand a major overhaul of our educational programs in science and technology
- We must use clouds and big data analytics in storing, processing, and mining of big data, which changes rapidly in time and space
- Machine intelligence, clouds, IoT, and social networks are being integrated together to promote global economy, public healthcare, smart cities and environments
- On the negative side, 30% of jobs will be replaced by smart machines, autonomous vehicles, etc. Data privacy and cyber-space crimes will grow.