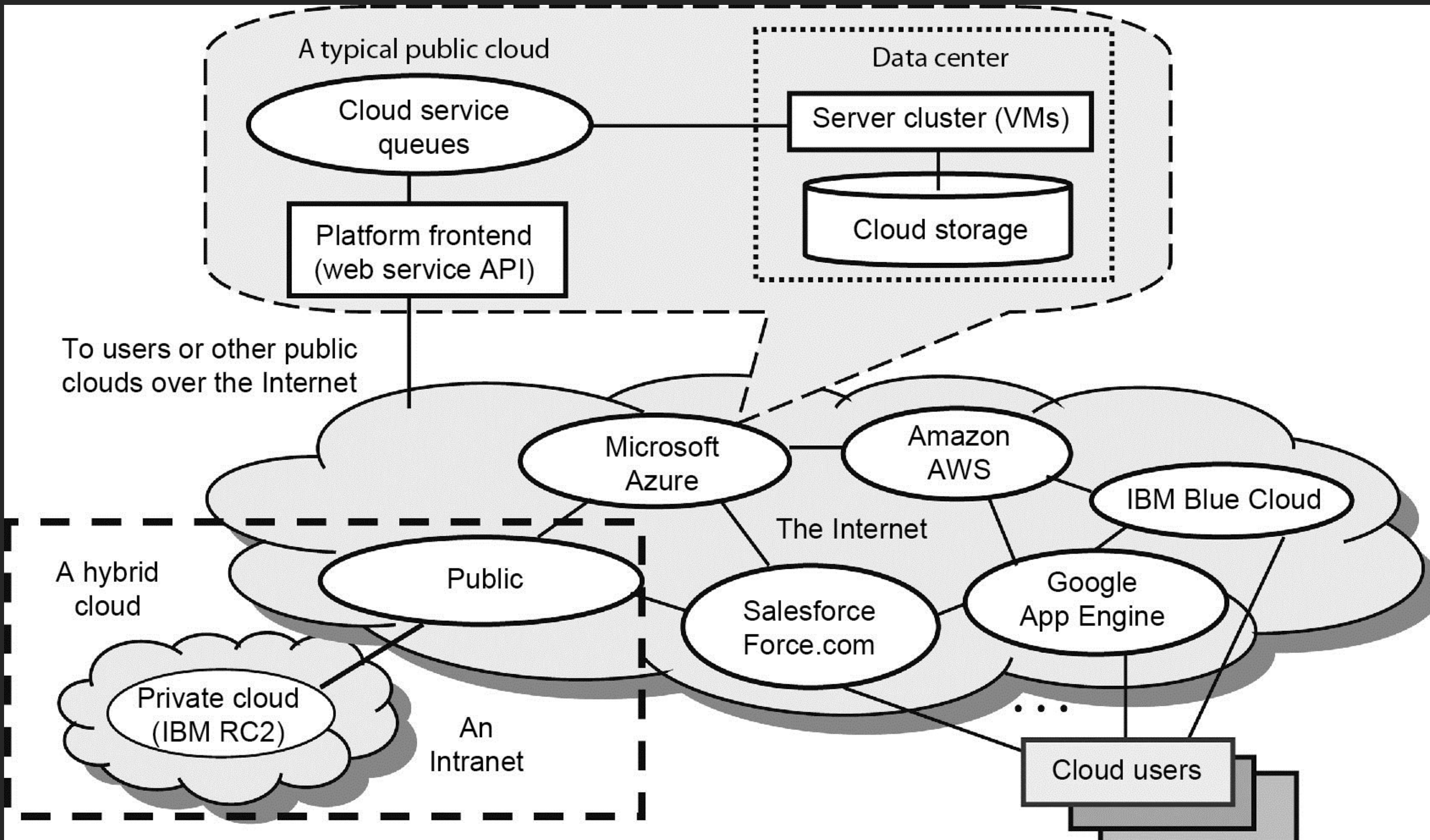
Distributed and Cloud Computing

K. Hwang, G. Fox and J. Dongarra

Chapter 4: Cloud Platform Architecture over Virtualized Datacenters

Adapted from Kai Hwang
University of Southern California

Public, Private & Hybrid Clouds



Public Clouds vs. Private Clouds :

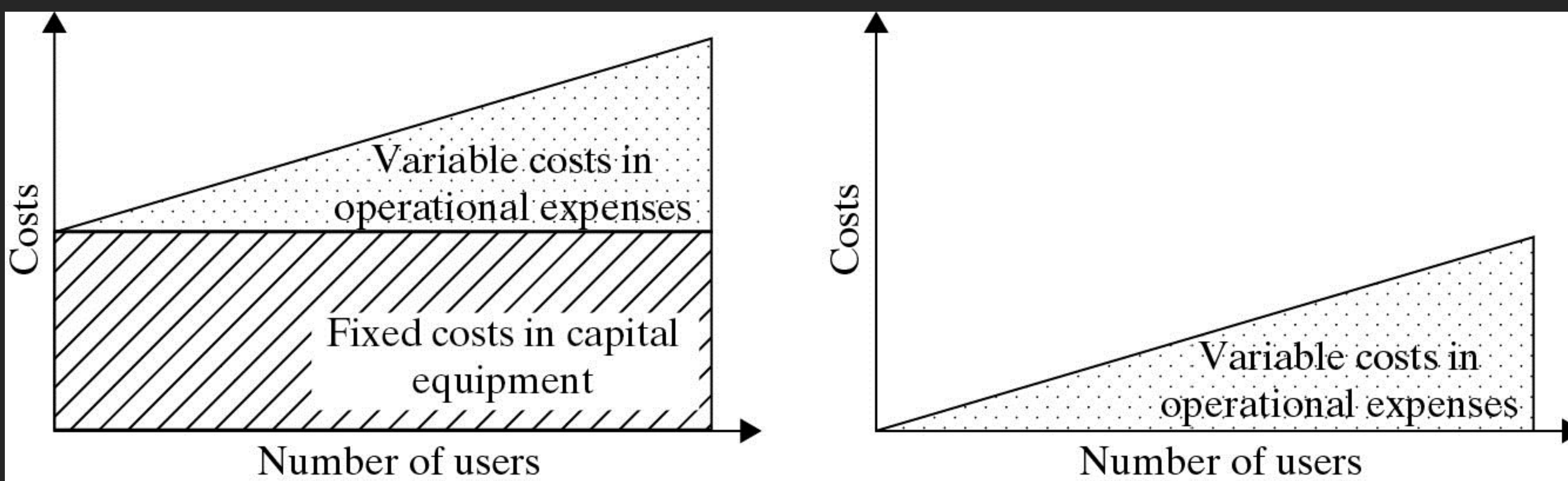
Characteristics	Public clouds	Private clouds
Technology leverage and ownership	Owned by service providers	Leverage existing IT infrastructure and personnel; owned by individual organization
Management of provisioned resources	Creating and managing VM instances within proprietary infrastructure; promote standardization, preserves capital investment, application flexibility	Client managed; achieve customization and offer higher efficiency
Workload distribution methods and loading policies	Handle workload without communication dependency; distribute data and VM resources; surge workload is off-loaded	Handle workload dynamically, but can better balance workloads; distribute data and VM resources
Security and data privacy enforcement	Publicly accessible through remote interface	Access is limited; provide pre-production testing and enforce data privacy and security policies
Example platforms	Google App Engine, Amazon AWS, Microsoft Azure	IBM RC2

Cost-Effectiveness in Cloud Computing vs. Datacenter Utilization

(Courtesy of M. Ambrust, et al 2009)

$$\text{UserHours}_{\text{cloud}} \times (\text{revenue} - \text{Cost}_{\text{cloud}}) \geq$$

$$\text{UserHours}_{\text{datacenter}} \times (\text{revenue} - \frac{\text{Cost}_{\text{datacenter}}}{\text{Utilization}})$$



(a) Traditional IT cost model

(b) Cloud computing cost model

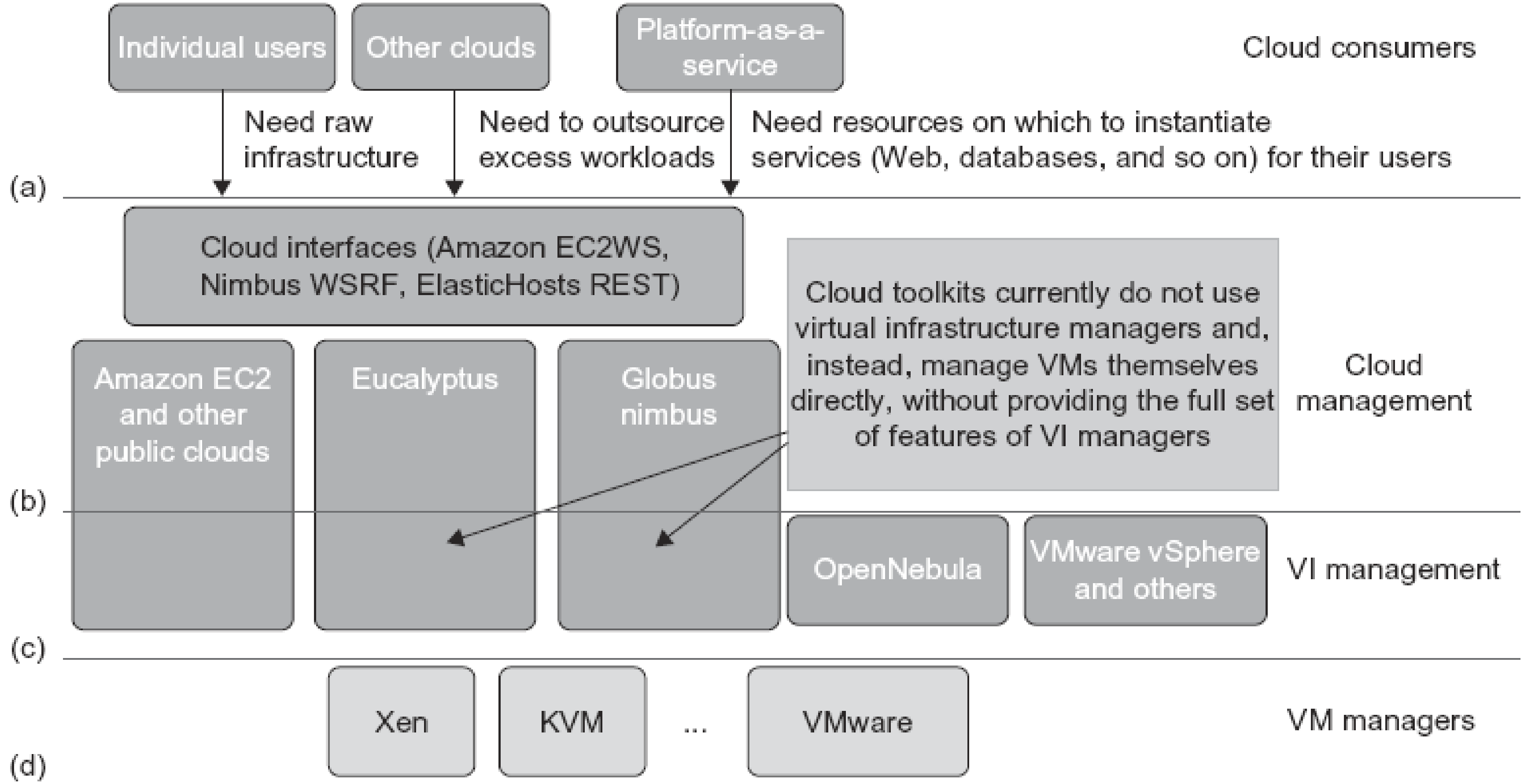


FIGURE 4.4

Cloud ecosystem for building private clouds: (a) Consumers demand a flexible platform; (b) Cloud manager provides virtualized resources over an IaaS platform; (c) VI manager allocates VMs; (d) VM managers handle VMs installed on servers.

(*Courtesy of Sotomayor, et al. [68]*)

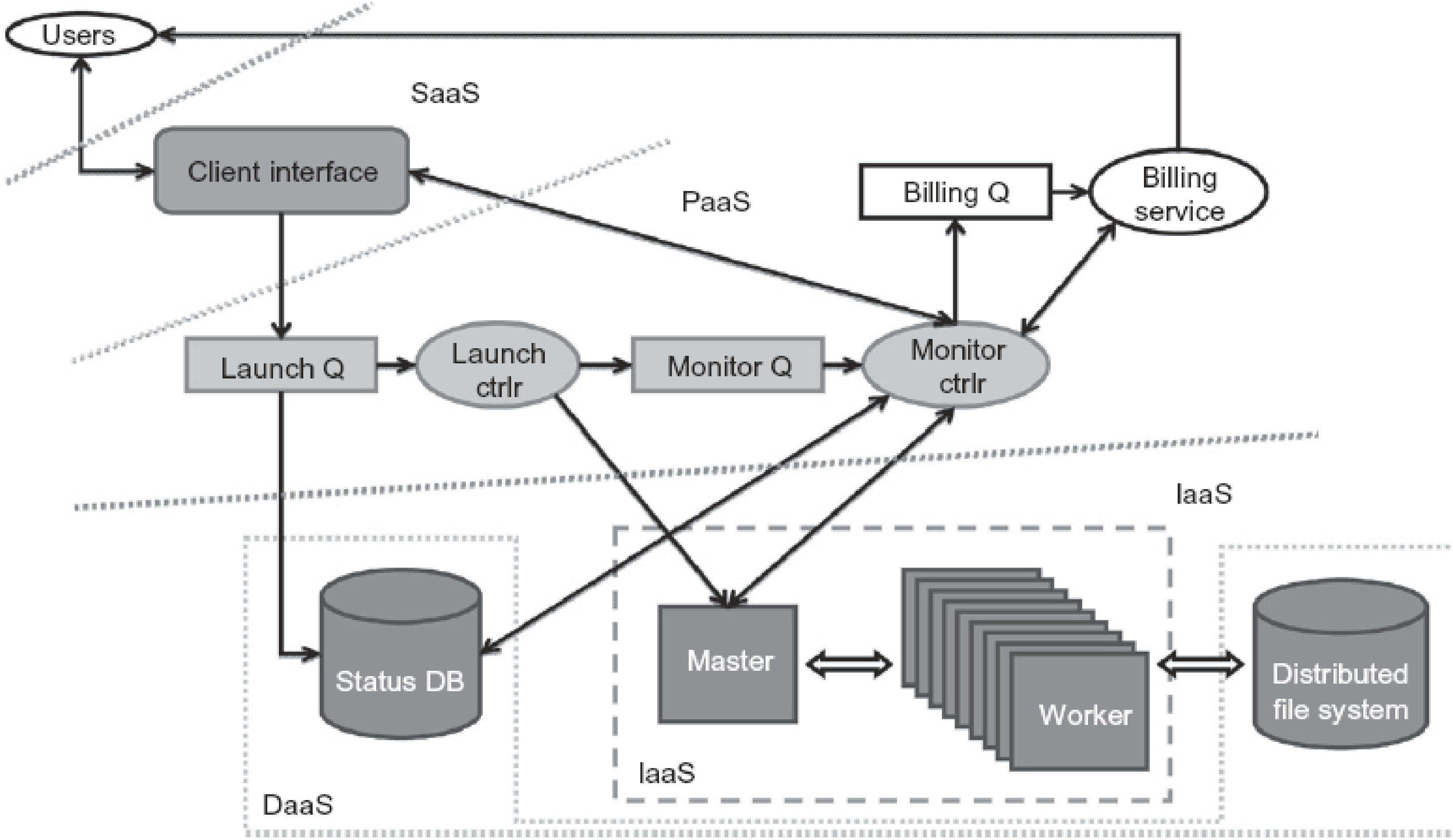


FIGURE 4.5

The IaaS, PaaS, and SaaS cloud service models at different service levels.

Infrastructure as a service (IaaS)

- Most basic cloud service model
- Cloud providers offer computers, as physical or more often as virtual machines, and other resources.
- Virtual machines are run as guests by a hypervisor, such as Xen or KVM.
- Cloud users deploy their applications by then installing operating system images on the machines as well as their application software.
- Cloud providers typically bill IaaS services on a utility computing basis, that is, cost will reflect the amount of resources allocated and consumed.
- Examples of IaaS include: Amazon CloudFormation (and underlying services such as Amazon EC2), Rackspace Cloud, Terremark, and Google Compute Engine.

Some IaaS Offerings from Public Clouds :

Table 4.1 Public Cloud Offerings of IaaS [10,18]

Cloud Name	VM Instance Capacity	API and Access Tools	Hypervisor, Guest OS
Amazon EC2	Each instance has 1–20 EC2 processors, 1.7–15 GB of memory, and 160–1.69 TB of storage.	CLI or Web Service (WS) portal	Xen, Linux, Windows
GoGrid	Each instance has 1–6 CPUs, 0.5–8 GB of memory, and 30–480 GB of storage.	REST, Java, PHP, Python, Ruby	Xen, Linux, Windows
Rackspace Cloud	Each instance has a four-core CPU, 0.25–16 GB of memory, and 10–620 GB of storage.	REST, Python, PHP, Java, C#, .NET	Xen, Linux
FlexiScale in the UK	Each instance has 1–4 CPUs, 0.5–16 GB of memory, and 20–270 GB of storage.	Web console	Xen, Linux, Windows
Joyent Cloud	Each instance has up to eight CPUs, 0.25–32 GB of memory, and 30–480 GB of storage.	No specific API, SSH, Virtual/Min	OS-level virtualization, OpenSolaris

Platform as a service (PaaS)

- Cloud providers deliver a computing platform typically including operating system, programming language execution environment, database, and web server.
- Application developers develop and run their software on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers.
- Examples of PaaS include: Amazon Elastic Beanstalk, Cloud Foundry, Heroku, Force.com, EngineYard, Mendix, Google App Engine, Microsoft Azure and OrangeScape.

PaaS Offerings from Public Clouds

Table 4.2 Five Public Cloud Offerings of PaaS [10,18]

Cloud Name	Languages and Developer Tools	Programming Models Supported by Provider	Target Applications and Storage Option
Google App Engine	Python, Java, and Eclipse-based IDE	MapReduce, Web programming on demand	Web applications and BigTable storage
Salesforce.com's Force.com	Apex, Eclipse-based IDE, Web-based Wizard	Workflow, Excel-like formula, Web programming on demand	Business applications such as CRM
Microsoft Azure	.NET, Azure tools for MS Visual Studio	Unrestricted model	Enterprise and Web applications
Amazon Elastic MapReduce	Hive, Pig, Cascading, Java, Ruby, Perl, Python, PHP, R, C++	MapReduce	Data processing and e-commerce
Aneka	.NET, stand-alone SDK	Threads, task, MapReduce	.NET enterprise applications, HPC

Software as a service (SaaS)

- Cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients.
- The pricing model for SaaS applications is typically a monthly or yearly flat fee per user, so price is scalable and adjustable if users are added or removed at any point.
- Examples of SaaS include: Google Apps, innkeypos, Quickbooks Online, Limelight Video Platform, Salesforce.com, and Microsoft Office 365.

Warehouse-Scale Computer (WSC)

- 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

(Courtesy of Hennessy and Patterson, 2012)

Design Considerations for WSC:

- Cost-performance
  - Small savings add up
- Energy efficiency
  - Affects power distribution and cooling
  - Work per joule
- Dependability via redundancy
- Network I/O
- Interactive and batch processing workloads
- Ample computational parallelism is not important
  - Most jobs are totally independent
  - "Request-level parallelism"
- Operational costs count
  - Power consumption is a primary constraint when designing system
- Scale and its opportunities and problems
  - Can afford customized systems since WSC require volume purchase

(Courtesy of Hennessy and Patterson, 2012)

Typical Datacenter Layout

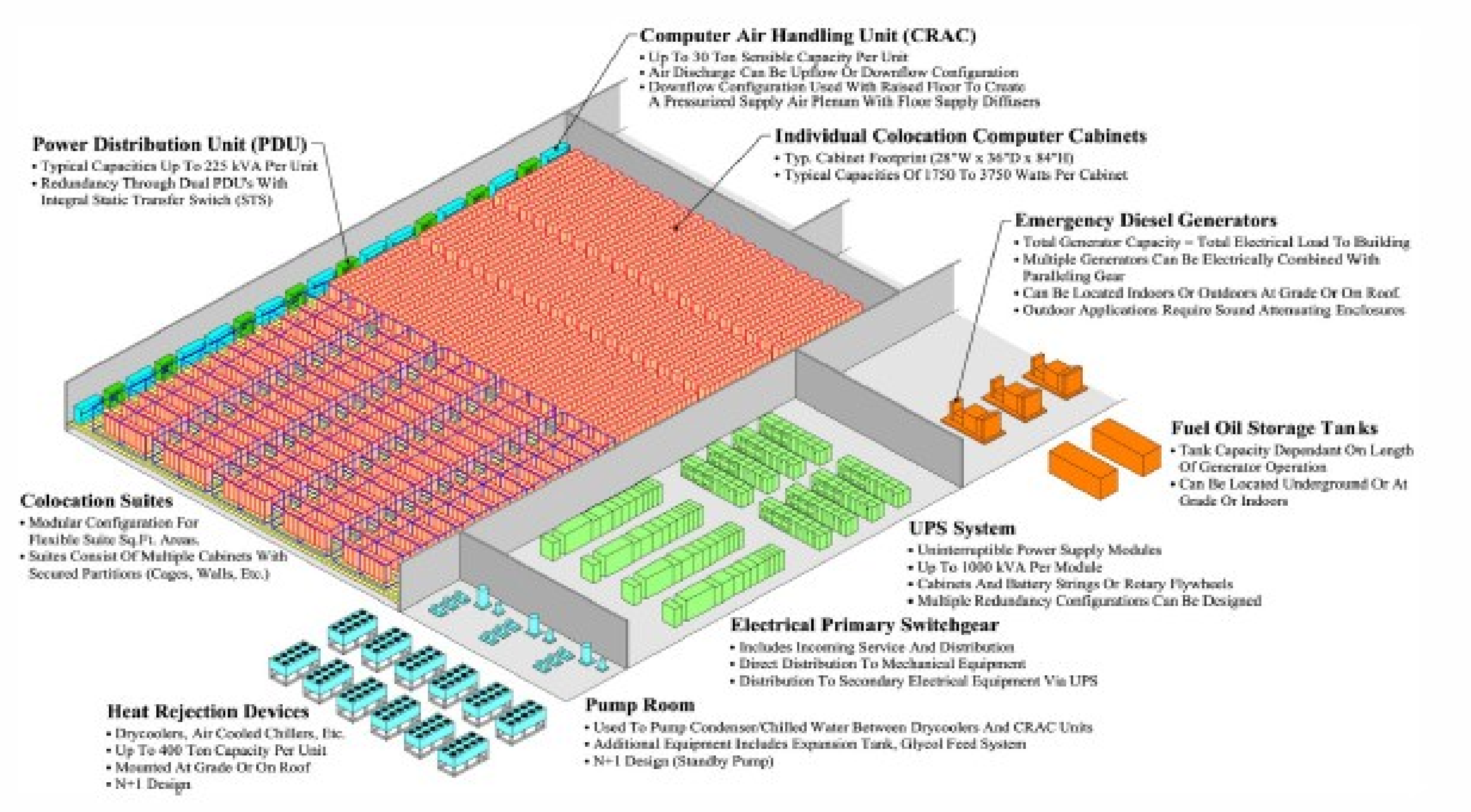


FIGURE 4.1: The main components of a typical datacenter (image courtesy of DLB Associates [23]).

Power and Cooling Requirements

- Cooling system also uses water (evaporation and spills)
  - E.g. 70,000 to 200,000 gallons per day for an 8 MW facility

- Power cost breakdown:
  - Chillers: 30-50% of the power used by the IT equipment
  - Air conditioning: 10-20% of the IT power, mostly due to fans

- How many servers can a WSC support?
  - Each server:
    - "Nameplate power rating" gives maximum power consumption
    - To get actual, measure power under actual workloads
  - Oversubscribe cumulative server power by 40%, but monitor power closely

(Courtesy of Hennessy and Patterson, 2012)

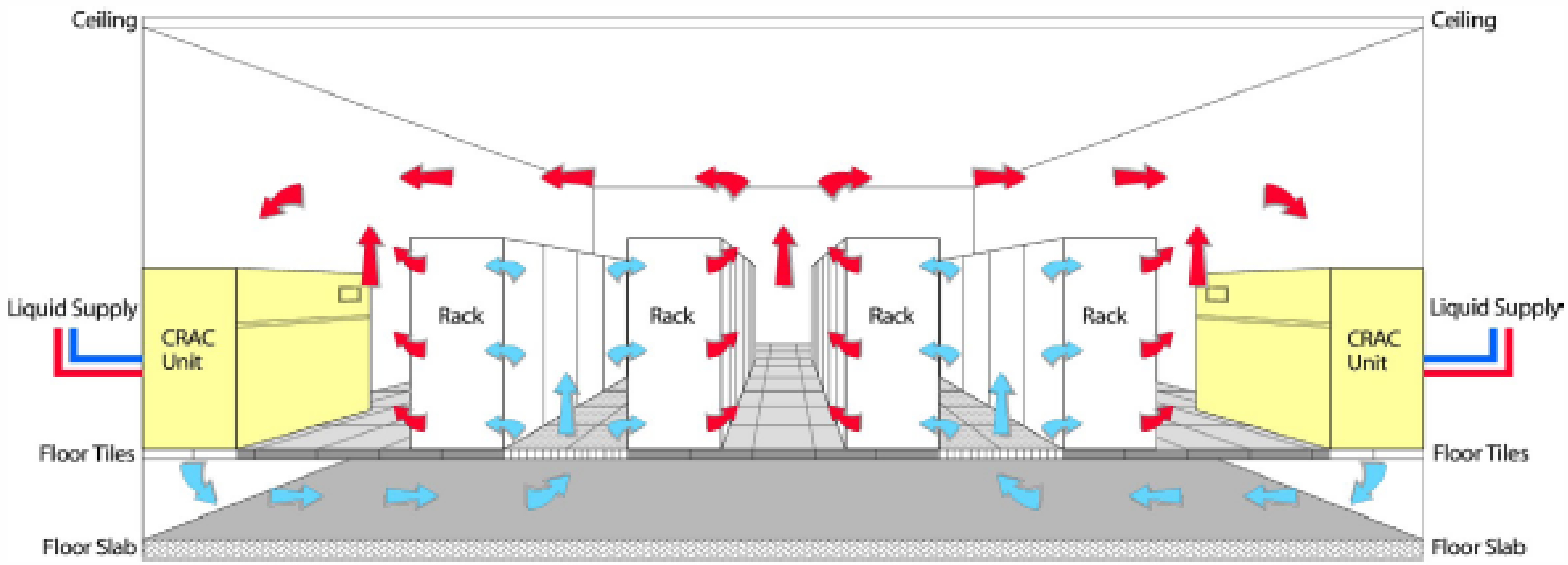


FIGURE 4.2: Datacenter raised floor with hot–cold aisle setup (image courtesy of DLB Associates [23]).

$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \underbrace{\left(\frac{1}{\text{PUE}}\right)}_{(a)} \times \underbrace{\left(\frac{1}{\text{SPUE}}\right)}_{(b)} \times \underbrace{\left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)}_{(c)}$$

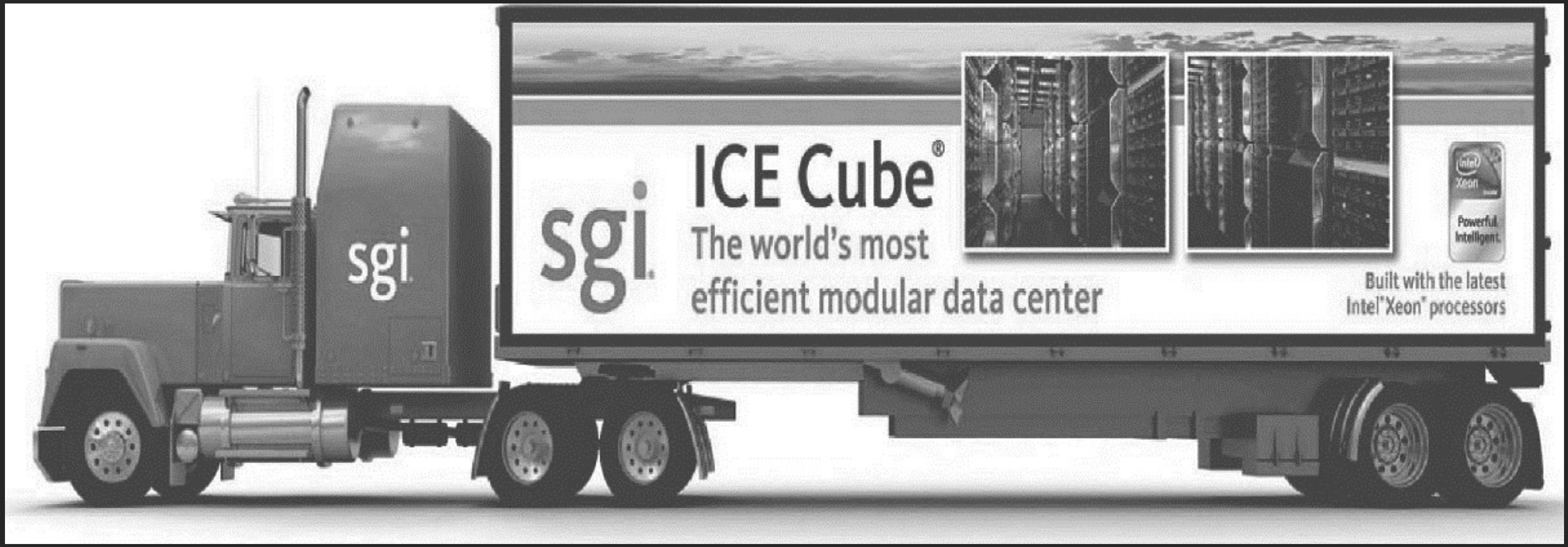
(Courtesy of Luiz Andre Barroso and Urs Holzle, Google Inc., 2009)

Measuring Efficiency of a WSC

- Power Utilization Effectiveness (PEU)
  - = Total facility power / IT equipment power
  - Median PUE on 2006 study was 1.69
- Performance
  - Latency is important metric because it is seen by users
  - Bing study: users will use search less as response time increases
  - Service Level Objectives (SLOs)/Service Level Agreements (SLAs)
    - E.g. 99% of requests be below 100 ms

(Courtesy of Hennessy and Patterson, 2012)

Modular Data Center



Cloud Computing

- WSCs offer economies of scale that cannot be achieved with a datacenter:
  - 5.7 times reduction in storage costs
  - 7.1 times reduction in administrative costs
  - 7.3 times reduction in networking costs
  - This has given rise to cloud services such as Amazon Web Services
    - "Utility Computing"
    - Based on using open source virtual machine and operating system software

(Courtesy of Hennessy and Patterson, 2012)

Enabling Technologies for The Clouds

Table 4.3 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 for distributing software to a large number of users for their simultaneous use and resource sharing if so desired
Massive data processing	Internet search and Web services which often require massive data processing, especially to support personalized services
Web-scale communication	Support for e-commerce, distance education, telemedicine, social networking, digital government, and digital entertainment applications
Distributed storage	Large-scale storage of personal records and public archive information which demands distributed storage over the clouds
Licensing and billing services	License management and billing services which greatly benefit all types of cloud services in utility computing

Cloud Computing as A Service

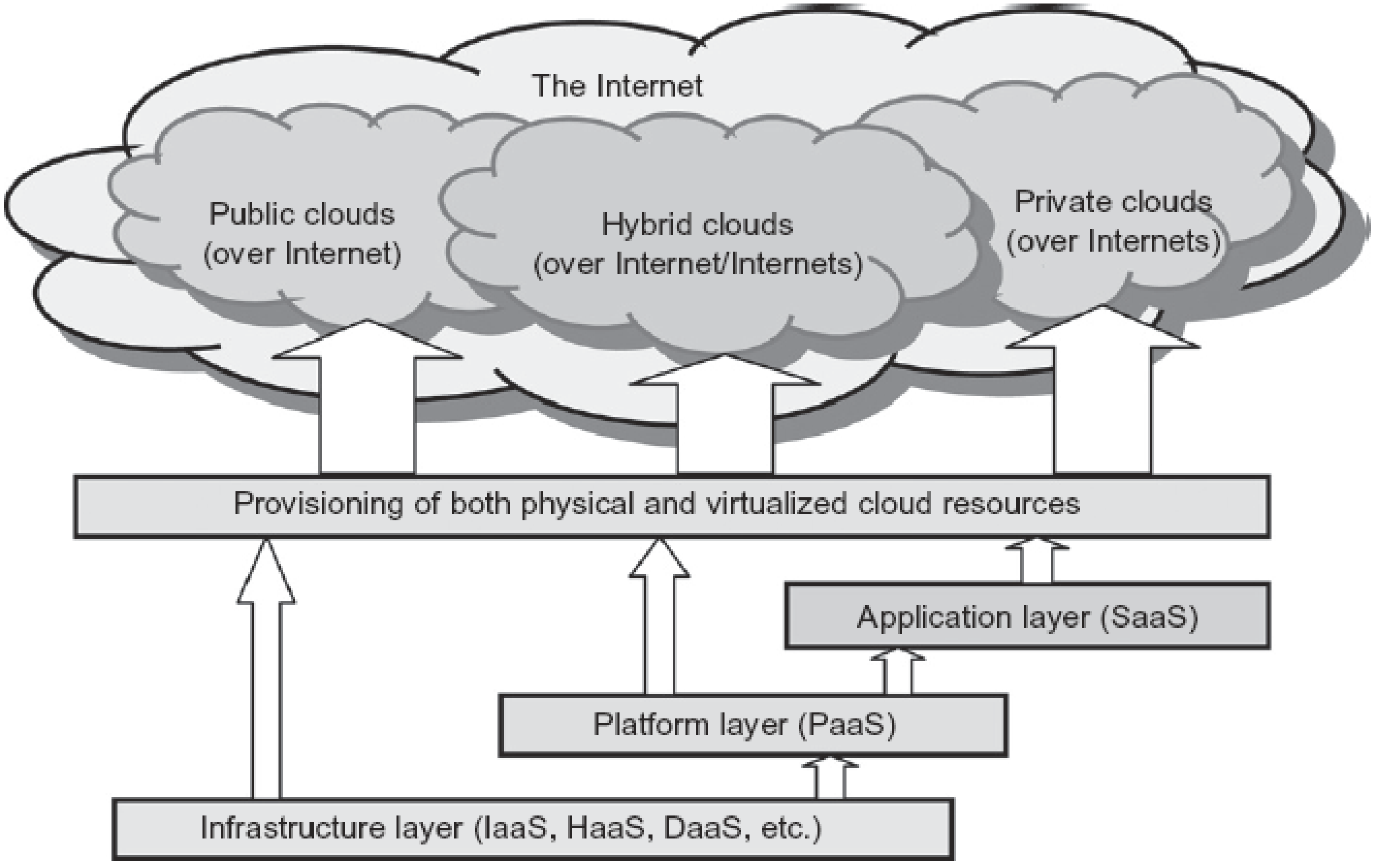


FIGURE 4.15

Layered architectural development of the cloud platform for IaaS, PaaS, and SaaS applications over the Internet.

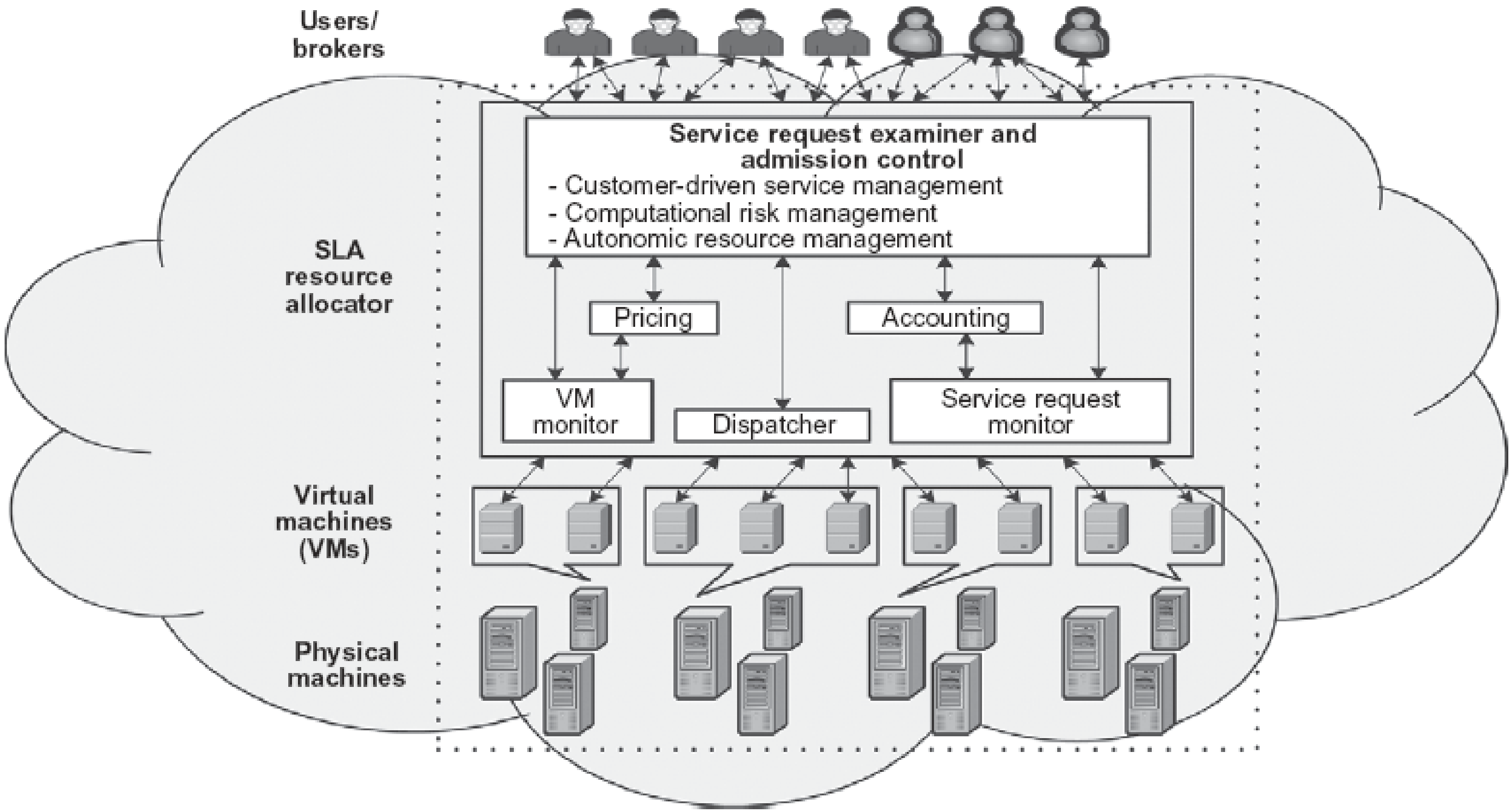
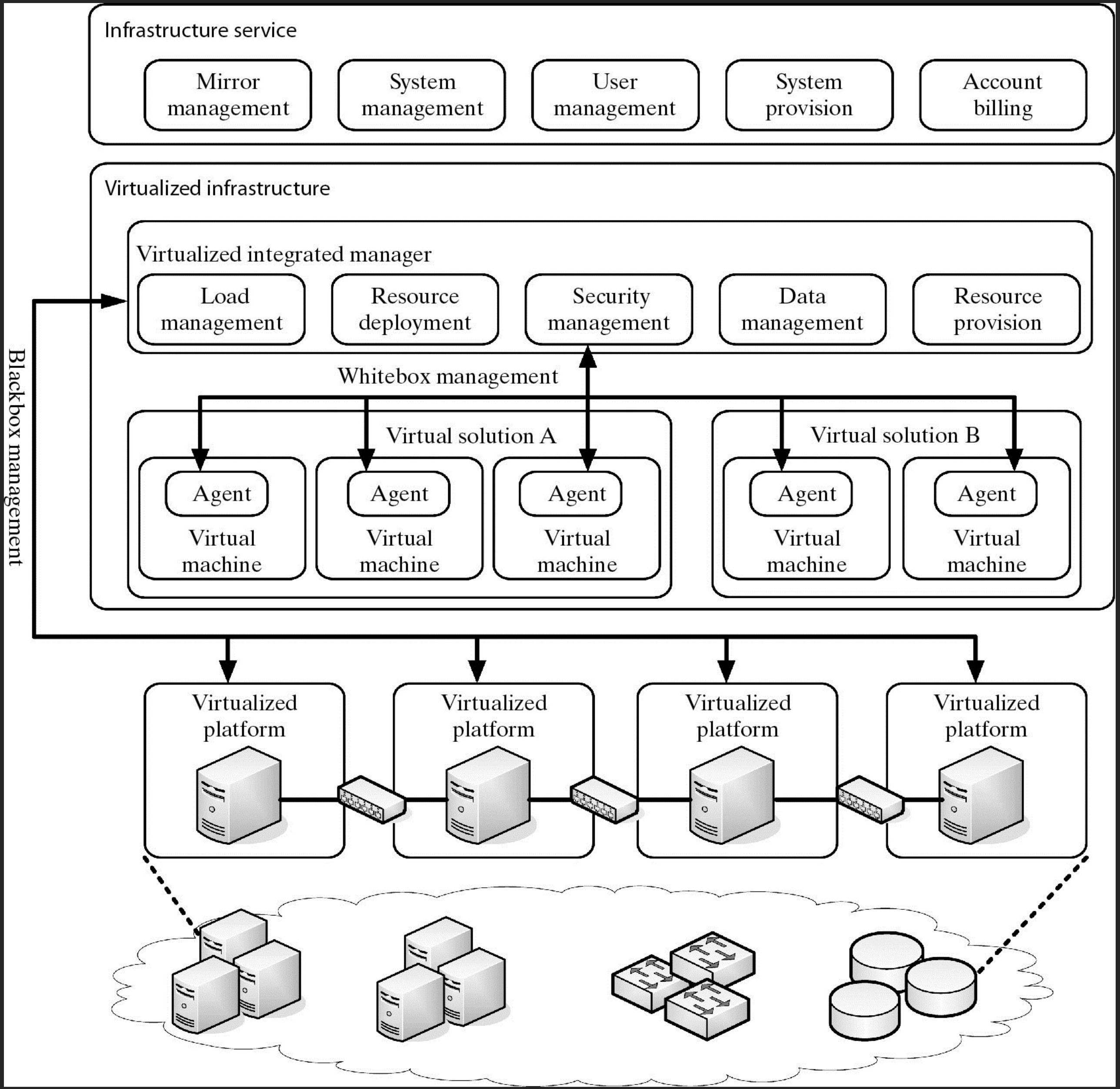


FIGURE 4.16

Market-oriented cloud architecture to expand/shrink leasing of resources with variation in QoS/demand from users.

(Courtesy of Raj Buyya, et al. [11])



Virtualized servers, storage , and network for cloud platform construction

Table 4.4 Virtualized Resources in Compute, Storage, and Network Clouds [4]

Provider	AWS	Microsoft Azure	GAE
Compute cloud with virtual cluster of servers	x86 instruction set, Xen VMs, resource elasticity allows scalability through virtual cluster, or a third party such as RightScale must provide the cluster	Common language runtime VMs provisioned by declarative descriptions	Predefined application framework handlers written in Python, automatic scaling up and down, server failover inconsistent with the Web applications
Storage cloud with virtual storage	Models for block store (EBS) and augmented key/blob store (SimpleDB), automatic scaling varies from EBS to fully automatic (SimpleDB, S3)	SQL Data Services (restricted view of SQL Server), Azure storage service	MegaStore/BigTable
Network cloud services	Declarative IP-level topology; placement details hidden, security groups restricting communication, availability zones isolate network failure, elastic IP applied	Automatic with user's declarative descriptions or roles of app. components	Fixed topology to accommodate three-tier Web app. structure, scaling up and down is automatic and programmer-invisible

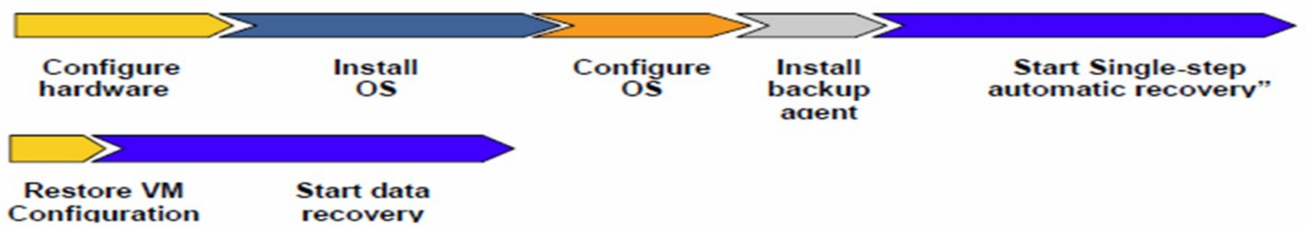
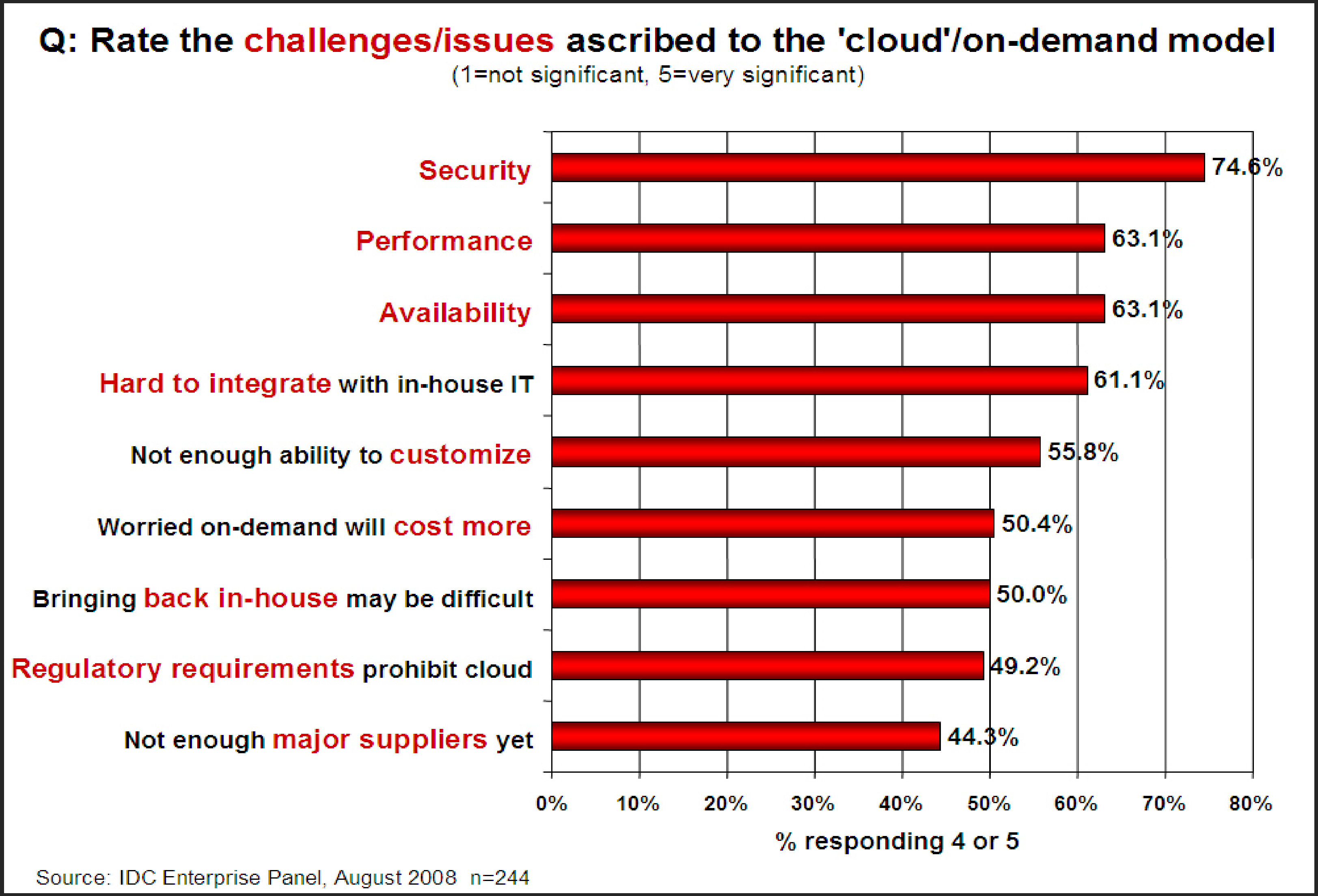


Figure 7.21 Recovery overhead of a conventional disaster recovery between physical machines, compared with that required to recover from live migration of virtual machines

Challenges/Issues in Cloud Computing



Challenges in Cloud Computing (1)

- **Concerns from The Industry (Providers)**
  - **Replacement Cost**
    - **Exponential increase in cost to maintain the infrastructure**
  - **Vendor Lock-in**
    - **No standard API or protocol can be very serious**
  - **Standardization**
    - **No standard metric for QoS is limiting the popularity**
  - **Security and Confidentiality**
    - **Trust model for cloud computing**
  - **Control Mechanism**
    - **Users do not have any control over infrastructures**

Challenges in Cloud Computing (2)

- **Concerns from Research Community :**
  - **Conflict to legacy programs**
    - **With difficulty in developing a new application due to lack of control**
  - **Provenance**
    - **How to reproduce results in different infrastructures**
  - **Reduction in Latency**
    - **No specially designed interconnect used**
    - **Very low controllability in layout of interconnect due to abstraction**
  - **Programming Model**
    - **Hard to debug where programming naturally error-prone**
    - **Details about infrastructure are hidden**
  - **QoS Measurement**
    - **Especially for ubiquitous computing where context changes**

Public Clouds and Service Offerings

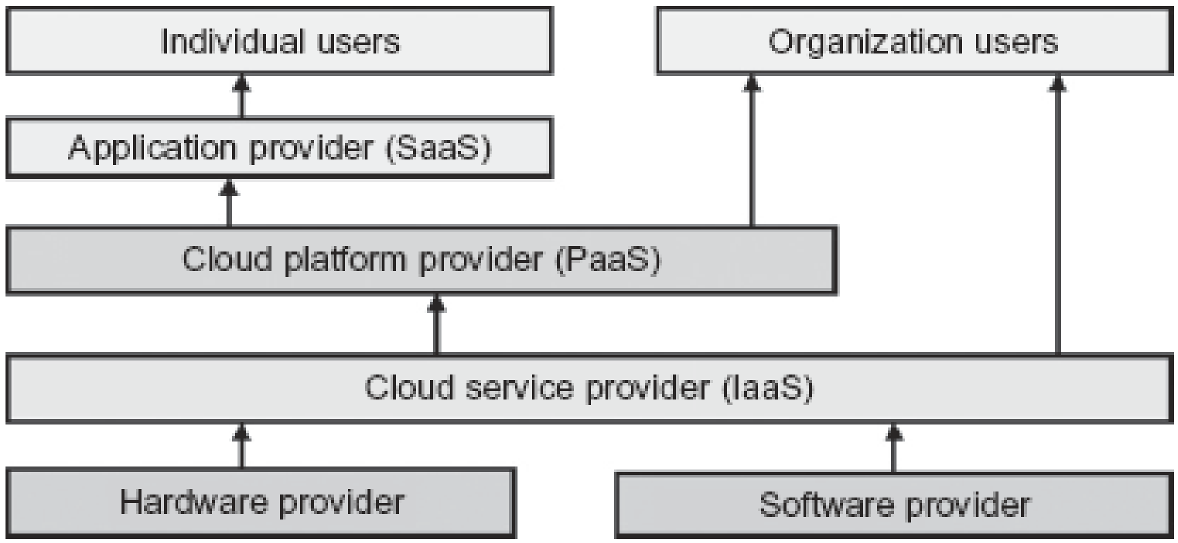


FIGURE 4.19

Roles of individual and organizational users and their Interaction with cloud providers under various cloud service models.

Table 4.5 Five Major Cloud Platforms and Their Service Offerings [30]

Model	IBM	Amazon	Google	Microsoft	Salesforce
PaaS	BlueCloud, WCA, RC2		App Engine (GAE)	Windows Azure	Force.com
IaaS	Ensembles	AWS		Windows Azure	
SaaS	Lotus Live		Gmail, Docs	.NET service, Dynamic CRM	Online CRM, Gifttag
Virtualization		OS and Xen	Application Container	OS level/ Hypel-V	
Service Offerings	SOA, B2, TSAM, RAD, Web 2.0	EC2, S3, SQS, SimpleDB	GFS, Chubby, BigTable, MapReduce	Live, SQL Hotmail	Apex, visual force, record security
Security Features	WebSphere2 and PowerVM tuned for protection	PKI, VPN, EBS to recover from failure	Chubby locks for security enforcement	Replicated data, rule-based access control	Admin./record security, uses metadata API
User Interfaces		EC2 command-line tools	Web-based admin. console	Windows Azure portal	
Web API	Yes	Yes	Yes	Yes	Yes
Programming Support	AMI		Python	.NET Framework	

Note: *WCA: WebSphere CloudBurst Appliance; RC2: Research Compute Cloud; RAD: Rational Application Developer; SOA: Service-Oriented Architecture; TSAM: Tivoli Service Automation Manager; EC2: Elastic Compute Cloud; S3: Simple Storage Service; SQS: Simple Queue Service; GAE: Google App Engine; AWS: Amazon Web Services; SQL: Structured Query Language; EBS: Elastic Block Store; CRM: Consumer Relationship Management.*

Platform as a Service (PaaS): Google App Engine

- This platform allows users to develop and host web application in Google datacenters with automatic scaling according to the demand.

- It is a free service for a certain limit and it only requires a Gmail account to access the services. After the free limit is exceeded the customers are charged for additional storage, bandwidth and instance hours.

- The current version supports Java, Python and Go as the programming languages and Google plans to add more languages in the future.

- All billed App Engine applications have a 99.95% uptime SLA. App Engine is designed to sustain multiple datacenter outages without any downtime.

- The app engine has a few restrictions - can only execute code called from an HTTP request, Java applications may only use a subset from the JRE standard edition and Java application cannot create new threads.

Google AppEgine (GAE)

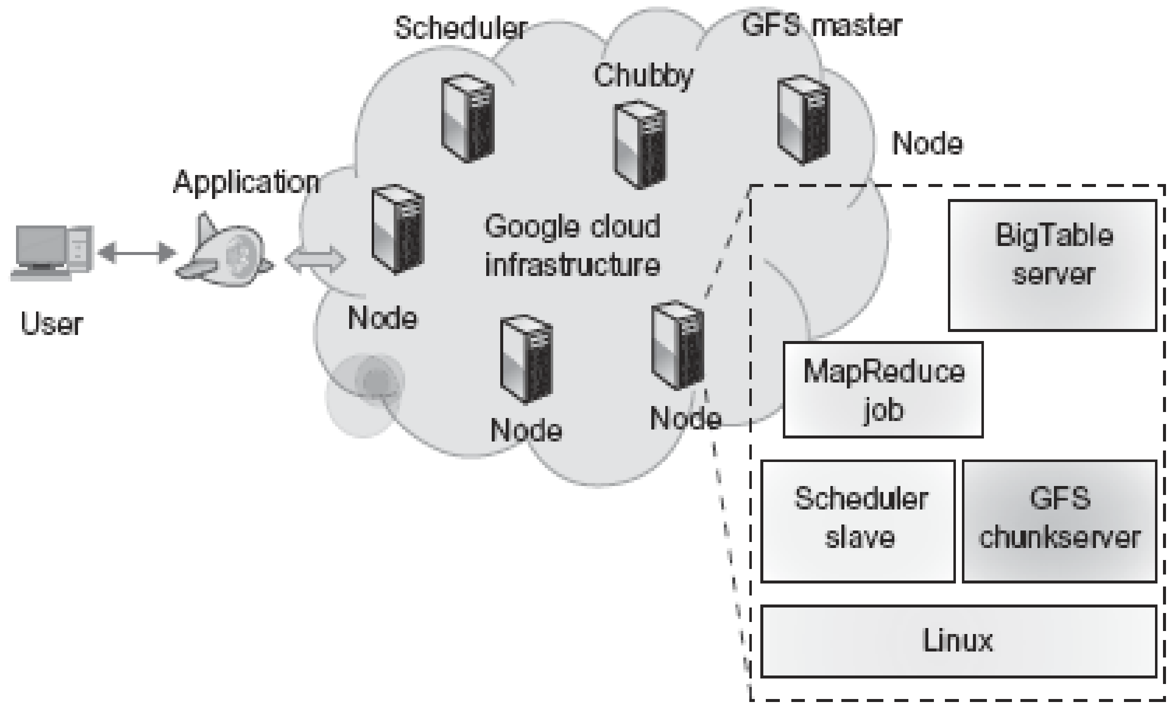


FIGURE 4.20

Google cloud platform and major building blocks, the blocks shown are large clusters of low-cost Servers.

(Courtesy of Kang Chen, Tsinghua University, China)

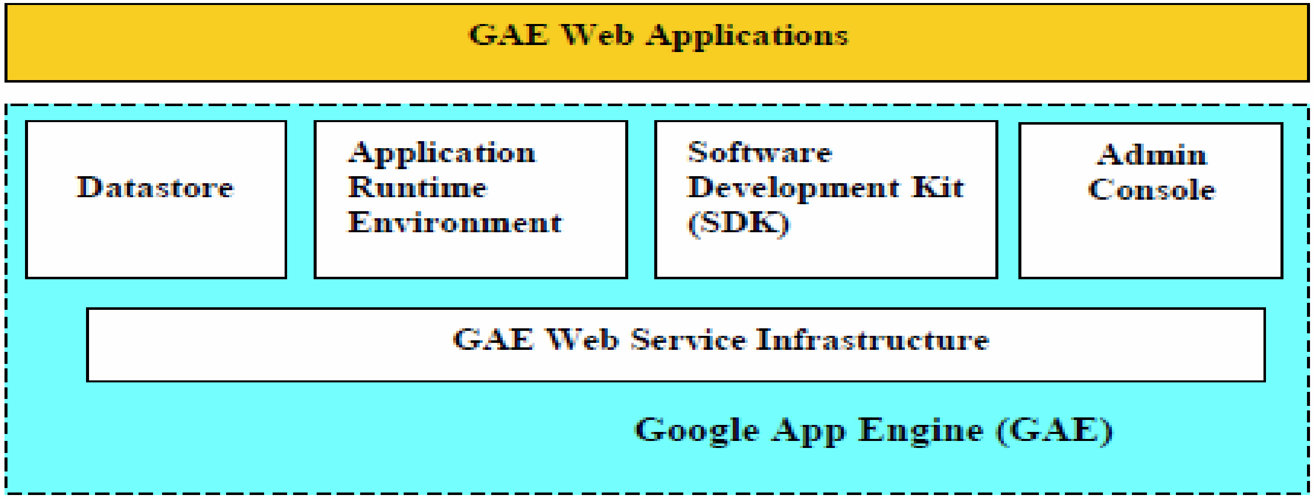


Figure 7.24 Functional components in the Google App Engine (GAE)
(Courtesy of Google, http://code.google.com/appengine/)

Google App Engine Front Page: http://code.google.com/appengine/

Signing up for an account or use your gmail account name : https://appengine.google.com/

Downloading GAE SDK : http://code.google.com/appengine/downloads.html

Python Getting Started Guide: http://code.google.com/appengine/docs/python/gettingstarted/

Java Getting Started Guide: http://code.google.com/appengine/docs/java/gettingstarted/

Quota page for free service: http://code.google.com/appengine/docs/quotas.html#Resources

Billing page if you go over the quota:
http://code.google.com/appengine/docs/billing.html#Billable_Quota_Unit_Cost

AWS – a leader in providing public IaaS services.

- EC2 (Elastic compute cloud allows users to rent virtual computers to run their own computer applications. It allows scalable deployment. A user can create, launch, and terminate server instances as needed, paying by the hour for active servers.
- S3 (simple storage service) provides the object-oriented storage service for users.
- EBS (Elastic block service) provides the block storage interface which can be used to support traditional applications.
- Amazon DevPay is a simple to use online billing and account management service that makes it easy for businesses
- MPI clusters uses hardware-assisted virtualization instead of para-virtualization and users are free to create a new AMIs
- AWS import/export allows one to ship large volumes of data to and from EC2 by shipping physical discs.
- Brokering systems offer a striking model for controlling sensors and providing office support of smartphones and tablets.
- Small-business companies can put their business on the Amazon cloud platform. Using AWS they can service a large number of internet users and make profits through those paid services.

Amazon Web Services (AWS)

Compute

- Amazon Elastic Compute Cloud (EC2)
- Amazon Elastic MapReduce
- Auto Scaling

Content Delivery

- Amazon CloudFront

Database

- Amazon SimpleDB
- Amazon Relational Database Service (RDS)

E-Commerce

- Amazon Fulfillment Web Service (FWS)

Messaging

- Amazon Simple Queue Service (SQS)
- Amazon Simple Notification Service (SNS)

Monitoring

- Amazon CloudWatch

Networking

- Amazon Virtual Private Cloud (VPC)
- Elastic Load Balancing

Payments & Billing

- Amazon Flexible Payments Service (FPS)
- Amazon DevPay

Storage

- Amazon Simple Storage Service (S3)
- Amazon Elastic Block Storage (EBS)
- AWS Import/Export

Support

- AWS Premium Support

Web Traffic

- Alexa Web Information Service
- Alexa Top Sites

Workforce

- Amazon Mechanical Turk

Amazon's Lesson

- Down for 3 days since 4/22/2011
- 1000x of businesses went offline. E.g. Pfizer, Netflix, Quora, Foursquare,Reddit
- SLA contract
  - 99.95% availability (<4.5hour down）
  - 10% penalty, otherwise

CNNMoney
A Service of CNN, Fortune & Money
FORTUNE M

Home Video Business News Markets Term Sheet Econo

Why Amazon's cloud Titanic went down

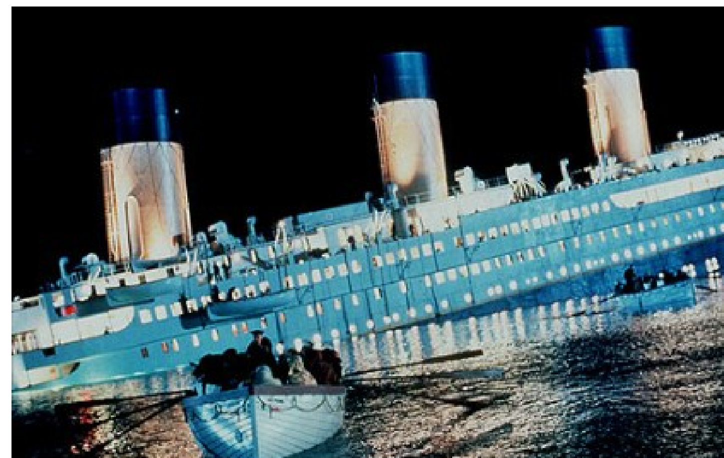


PHOTO: PARAMOUNT PICTURES/GETTY IMAGES

By David Goldman, staff writer April 22, 2011: 5:37 PM ET

NEW YORK (CNNMoney) -- This was never supposed to happen.

Amazon Web Services is the Titanic of cloud hosting, designed with backups to the backups' backups that prevent hosted websites and applications from failing.

Microsoft Azure Cloud :

This is essentially a PaaS Cloud.

- Windows Azure run its cluster hosted at Microsoft's datacenters that manages computing and storage resources.
  - One can download Azure development kit to run a local version of Azure. It allows Azure applications to be developed and debugged one the windows 7 hosts.
- All cloud services can interact with traditional MS software applications such as Windows Live, Office Live, Exchange Online, etc.
- If offers a Windows-based cloud platform using Microsoft virtualization technology.
  - Applications are built on VM’s deployed on the data-center services.
    - Azure manages all servers, storage and network resources of the data center.

Microsoft Windows Azure

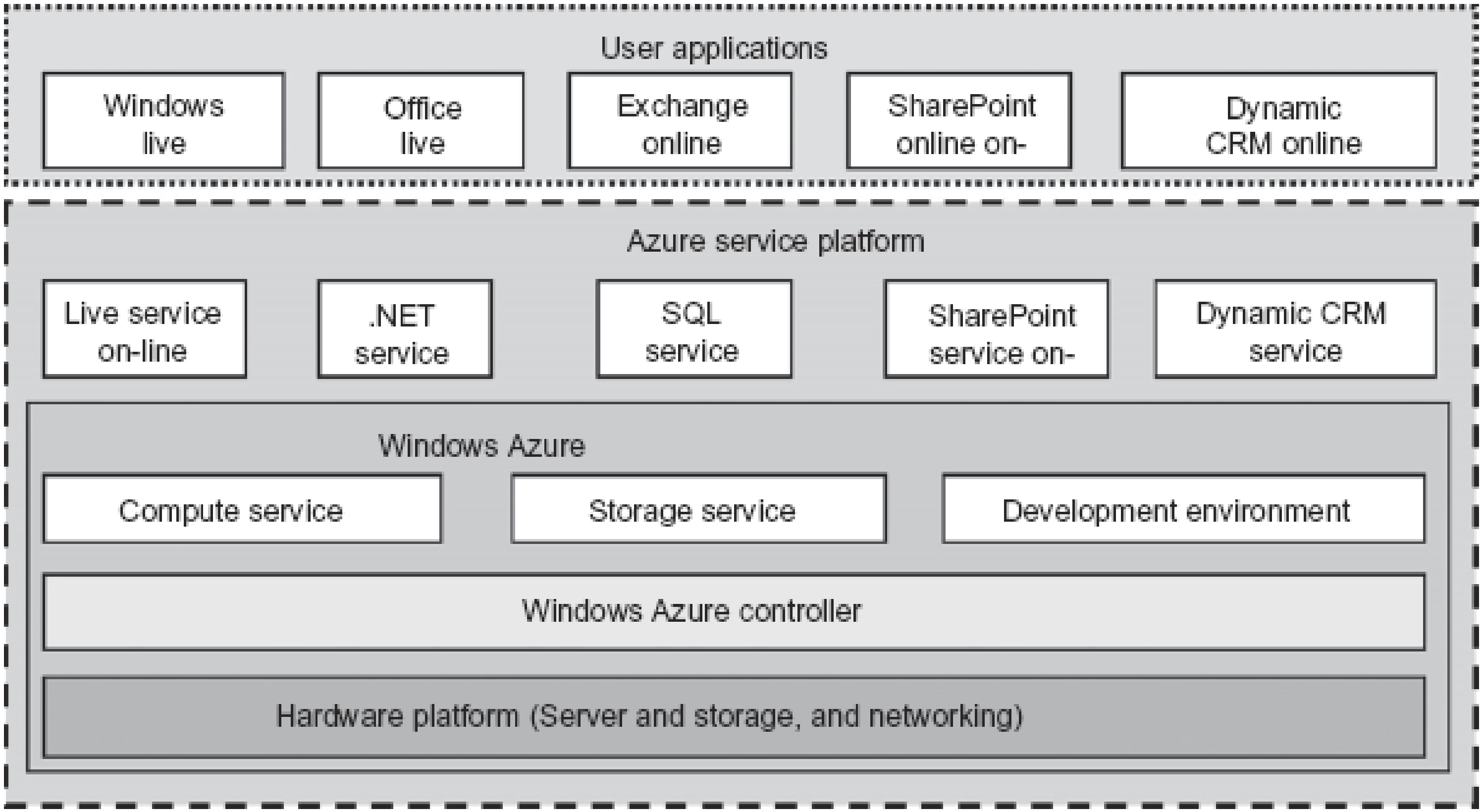


FIGURE 4.22

Microsoft Windows Azure platform for cloud computing.

(Courtesy of Microsoft, 2010, http://www.microsoft.com/windowsazure)

Cloud Services and Major Providers

Cloud application (SaaS)	Concur, RightNOW, Teleo, Kenexa, Webex, Blackbaud, salesforce.com, Netsuite, Kenexa, etc.
Cloud software environment (PaaS)	Force.com, App Engine, Facebook, MS Azure, NetSuite, IBM BlueCloud, SGI Cyclone, eBay
Cloud software infrastructure: Computational resources (IaaS); Storage (DaaS); Communications (Caas)	Amazon AWS, OpSource Cloud, IBM Ensembles, Rackspace cloud, Windows Azure, HP, Banknorth
Co-location cloud services (LaaS)	Savvis, Internap, NTTCommunications, Digital Realty Trust, 365 Main
Network cloud services (NaaS)	Owest, AT&T, AboveNet
Hardware/Virtualization cloud services (HaaS)	VMware, Intel, IBM, XenEnterprise

FIGURE 4.23

A stack of six layers of cloud services and their providers.

(Courtesy of T. Chou, Active Book Express, 2010 [16])

Table 4.7 Cloud Differences in Perspectives of Providers, Vendors, and Users

Cloud Players	IaaS	PaaS	SaaS
IT administrators/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 configurators and APIs	Develop and deploy software
End users or business users	To deploy and store data	To develop and test Web software	Use business software

Table 4.8 Storage Services in Three Cloud Computing Systems

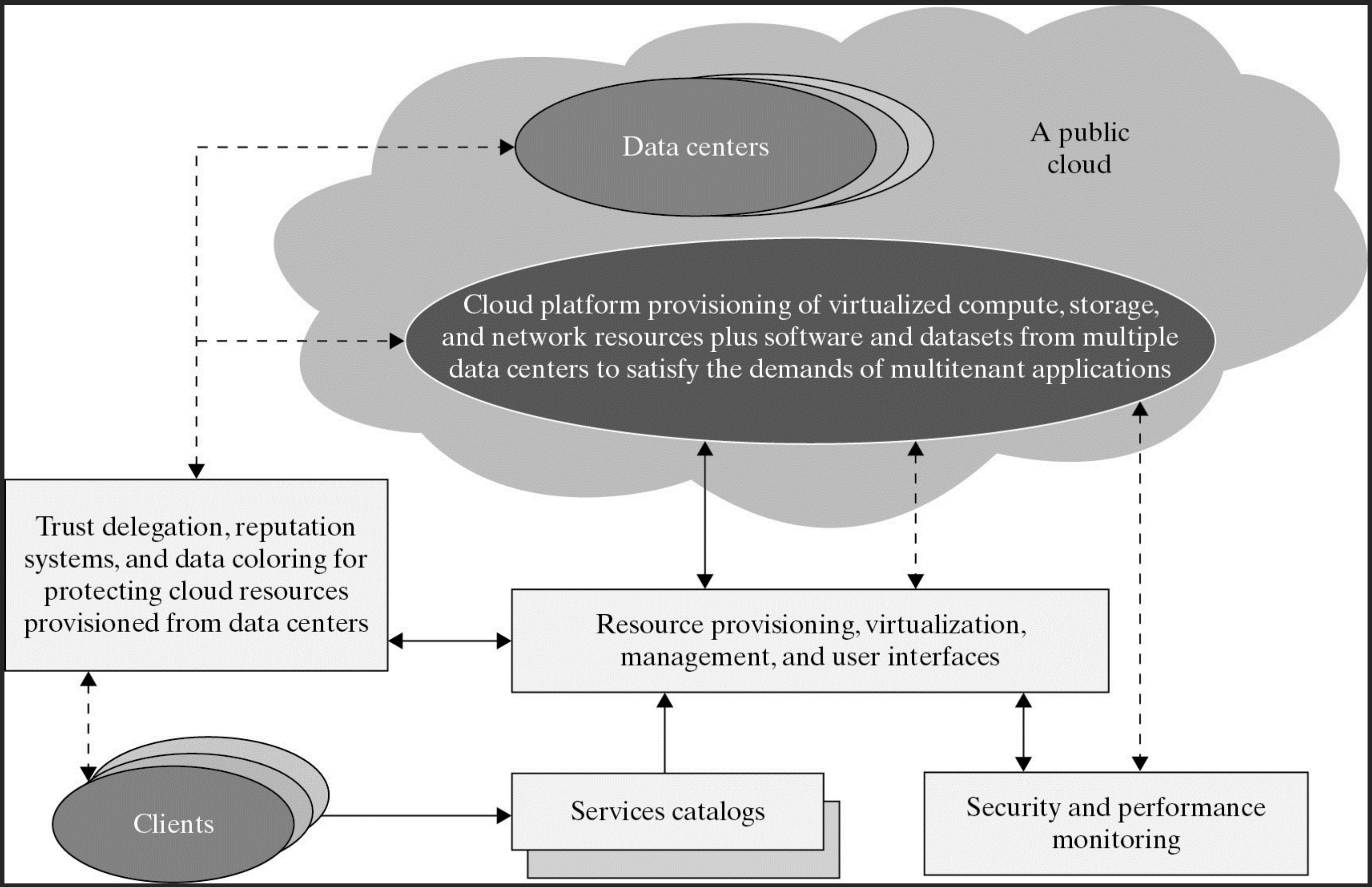
Storage System	Features
GFS: Google File System	Very large sustainable reading and writing bandwidth, mostly continuous accessing instead of random accessing. The programming interface is similar to that of the POSIX file system accessing interface.
HDFS: Hadoop Distributed File System	The open source clone of GFS. Written in Java. The programming interfaces are similar to POSIX but not identical.
Amazon S3 and EBS	S3 is used for retrieving and storing data from/to remote servers. EBS is built on top of S3 for using virtual disks in running EC2 instances.

$Ex' = \overline{X}$

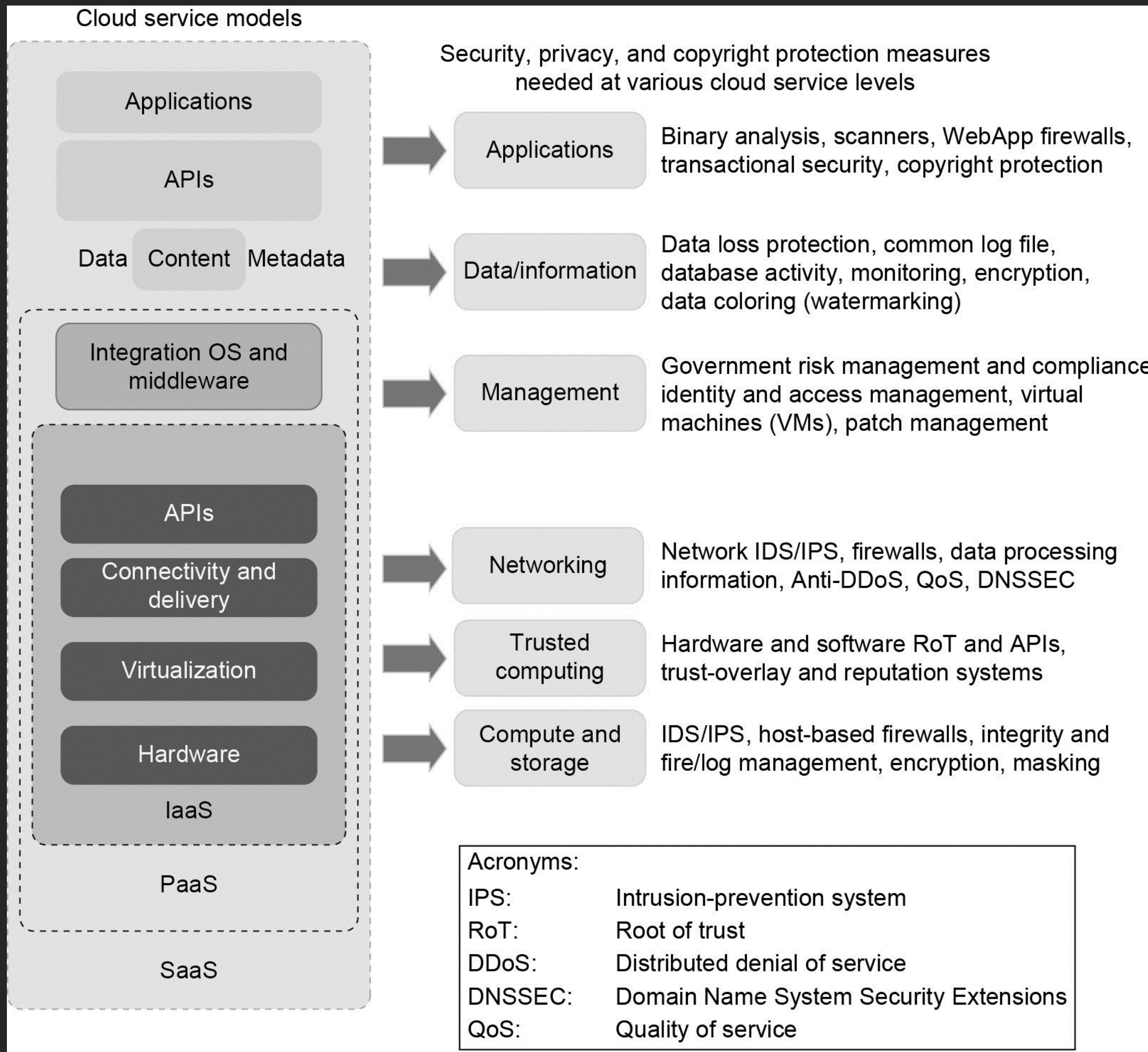
Security and Trust Barriers in Cloud Computing

- Protecting datacenters must first secure cloud resources and uphold user privacy and data integrity.
- Trust overlay networks could be applied to build reputation systems for establishing the trust among interactive datacenters.
- A watermarking technique is suggested to protect shared data objects and massively distributed software modules.
- These techniques safeguard user authentication and tighten the data access-control in public clouds.
- The new approach could be more cost-effective than using the traditional encryption and firewalls to secure the clouds.

Security Aware Cloud Platform



Cloud Service Models & Security Measures



(a) Cloud service models (b) Security, privacy, and copyright protection measures