Introduction to Cloud Computing and Technical Issues

Eom, Hyeonsang
School of Computer Science & Engineering
Seoul National University

Future Internet Summer Camp 2009

Outline

- What is Cloud Computing?
- Clouds vs. Grids
- Cloud Computing Architecture
- Cloud Services
- Technical Issues on Cloud Computing

Cloud Computing
- Buzz and Fuzz Word

"…computation may someday be organized as a public utility …"

John McCarthy,
1960

"cloud computing can take on different shapes depending on the viewer, and often seems a little fuzzy at the edges"

James O'brien

"a cloud is a pool of virtualized resources that can host a variety of different workloads, allow workloads to be deployed and scaled-out quickly, allocate resources when needed, and support redundancy

Greg Boss et al.,
IBM

Chaisiri's presentation

What Is Cloud Computing?

- Internet computing
  - Computation done through the Internet
  - No concern about any maintenance or management of actual resources
- Shared computing resources
  - As opposed to local servers and devices
- Comparable to Grid Infrastructure
- Web applications
- Specialized raw computing services

Cloud Computing Resources

- Large pool of easily usable and accessible virtualized resources
  - Dynamic reconfiguration for adjusting to different loads (scales)
  - Optimization of resource utilization
- Pay-per-use model
  - Guarantees offered by the Infrastructure Provider by means of customized SLAs(Service Level Agreements)
- Set of features
  - Scalability, pay-per-use utility model and virtualization

Technical Key Points

- User interaction interface: how users of cloud interface with the cloud
- Services catalog: services a user can request
- System management: management of available resources
- Provisioning tool: carving out the systems from the cloud to deliver on the requested service
- Monitoring and metering: tracking the usage of the cloud
- Servers: virtual or physical servers managed by system administrators

Key Characteristics (1/2)

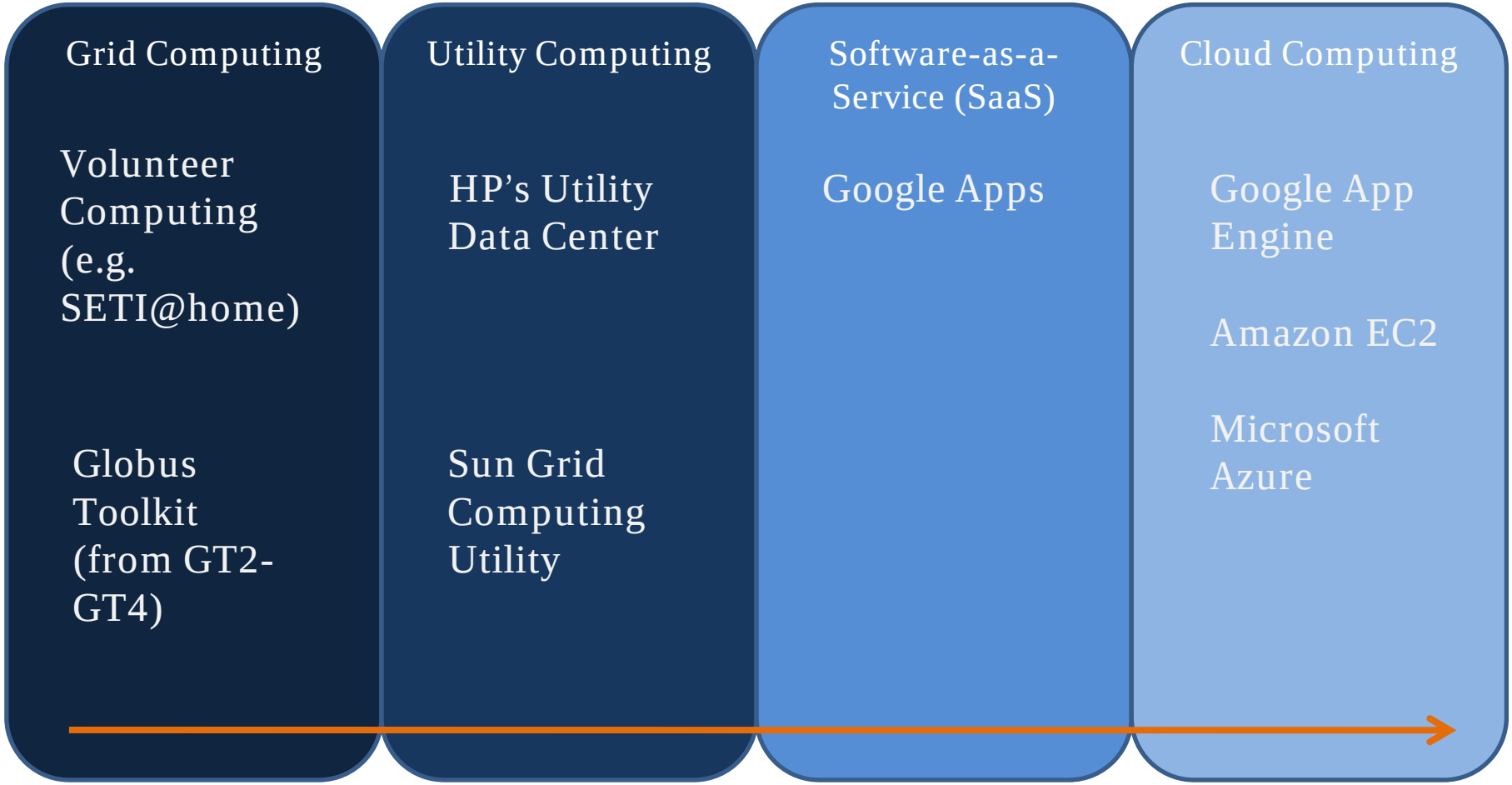
- Cost savings for resources
  - Cost is greatly reduced as initial expense and recurring expenses are much lower than traditional computing
  - Maintenance cost is reduced as a third party maintains everything from running the cloud to storing data
- Platform, Location and Device independency
  - Adoptable for all sizes of businesses, in particular small and mid-sized ones

Key Characteristics (2/2)

- Scalable services and applications
  - Achieved through server virtualization technology
- Redundancy and disaster recovery

Cloud computing means getting the best performing system with the best monetary value

Timeline



Wikipedia.com

Clouds vs. Grids

- Distinctions: not clear maybe because Clouds and Grids share similar visions
  - Reducing computing costs
  - Increasing flexibility and reliability by using third-party operated hardware
- Grid
  - System that coordinates resources which are not subject to centralized control, using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service
  - Ability to combine resources from different organizations for a common goal

Feature Comparison (1/3)

- Resource Sharing
  - Grid: collaboration (among Virtual Organizations, for fair share)
  - Cloud: assigned resources not shared
- Virtualization
  - Grid: virtualization of data and computing resources
  - Cloud: virtualization of hardware and software platforms
- Security
  - Grid: security through credential delegations
  - Cloud: security through isolation

Feature Comparison (2/3)

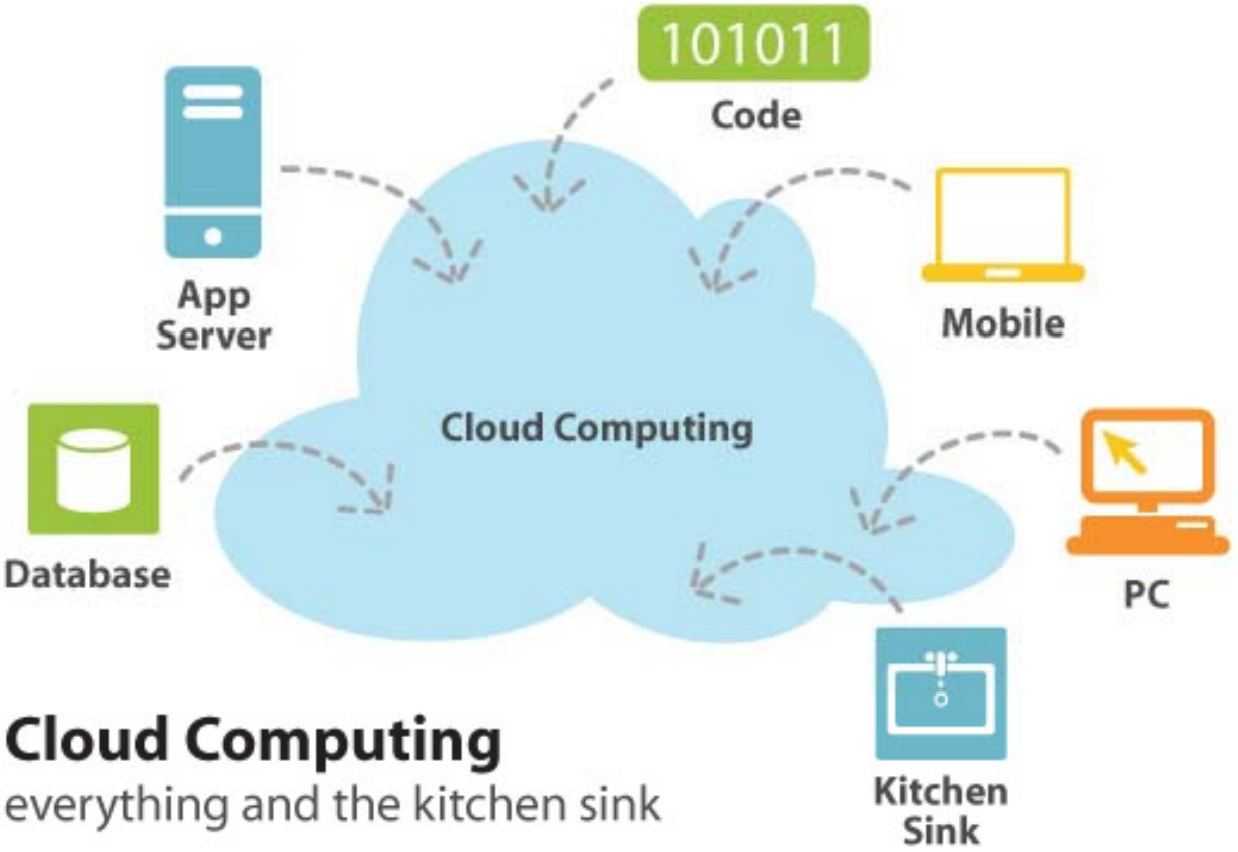
- High Level Service
  - Grid: Plenty of high level services
  - Cloud: No high level services defined yet.
- Architecture
  - Grid: Service oriented
  - Cloud: User chosen architecture
- Platform Awareness
  - Grid: need for Grid-enabled client software
  - Cloud: SP (Service Provider) software working in a customized environment

Feature Comparison (3/3)

- Scalability
  - Grid: node and site scalability
  - Cloud: node, site and hardware scalability
- Self Management
  - Grid: reconfigurability
  - Cloud: reconfigurability and self-healing
- Payment Model
  - Grid: rigid
  - Cloud: flexible

Cloud Computing Architecture

- Front End
  - End user, client or any application (i.e., web browser, etc.)
- Back End (Cloud services)
  - Network of servers with any computer program and data storage system
    - It is usually assumed that clouds have infinite storage capacity for any software available in market

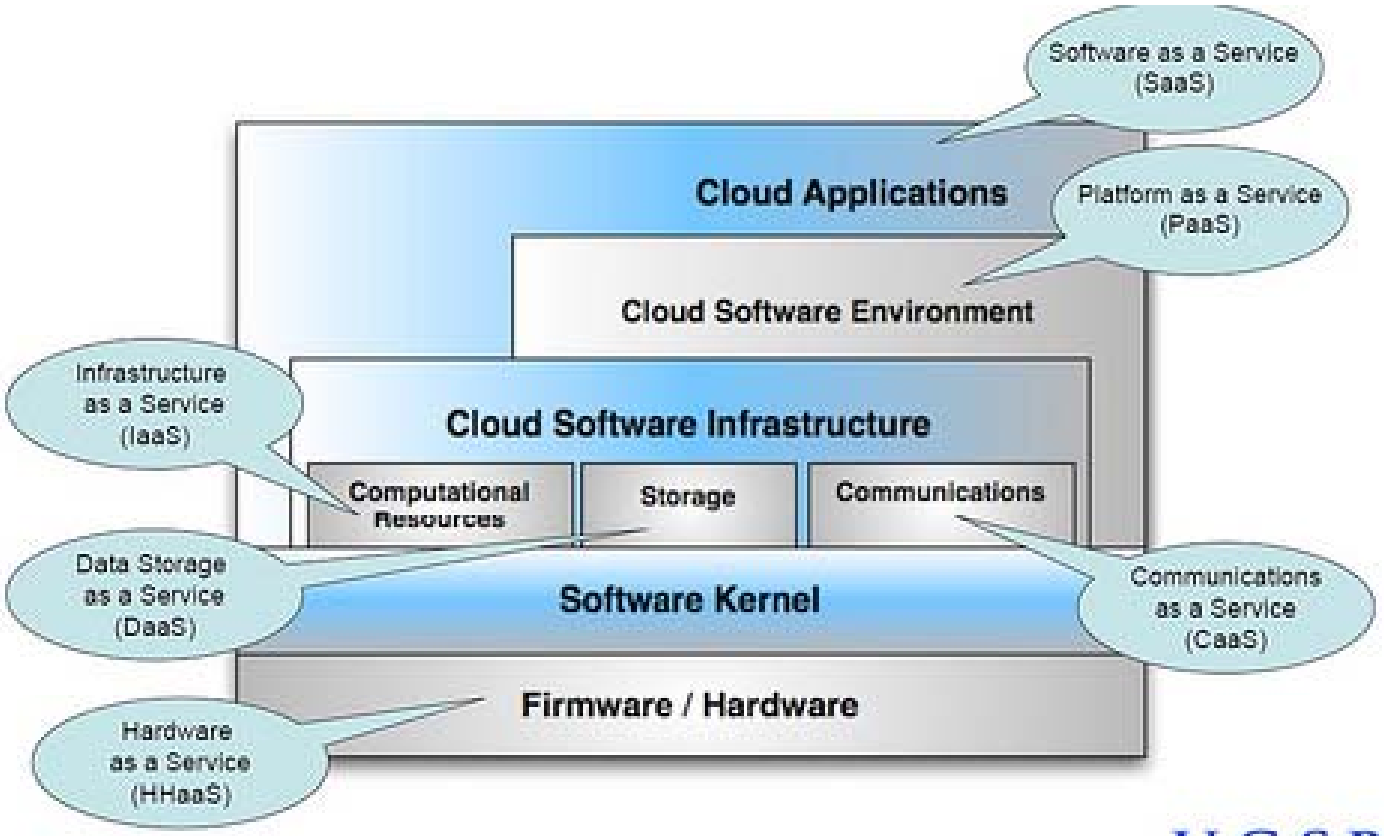


<cloudcomputingarchitect.com>

Cloud Service Taxonomy

- Layer
  - **Software-as-a-Service (SaaS)**
  - **Platform-as-a-Service (PaaS)**
  - **Infrastructure-as-a-Service (IaaS)**
  - **Data Storage-as-a-Service (DaaS)**
  - Communication-as-a-Service (CaaS)
  - Hardware-as-a-Service(HaaS)
- Type
  - Public cloud
  - Private cloud
  - Inter-cloud

Cloud Computing Service Layer



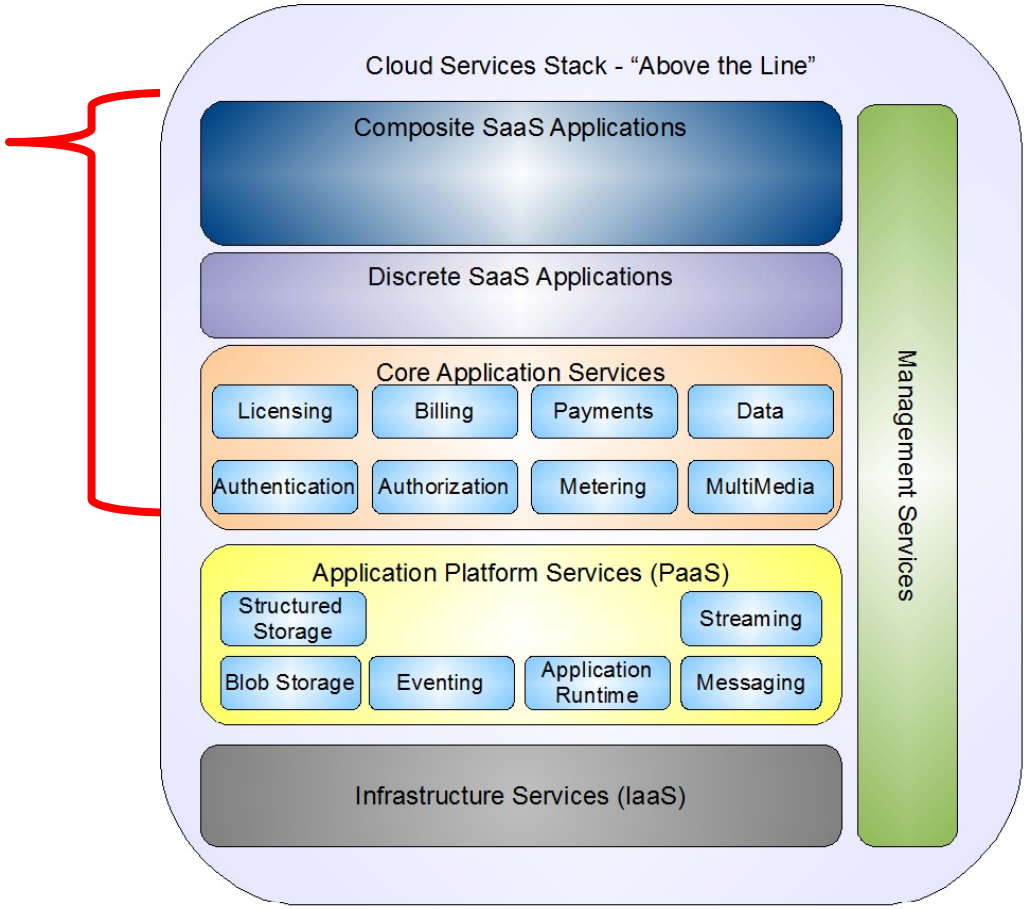
UCSB

Software-as-a-Service (SaaS)

- Definition
  - *Software deployed as a hosted service and accessed over the Internet*
- Features
  - *Open, Flexible*
  - *Easy to Use*
  - *Easy to Upgrade*
  - *Easy to Deploy*

Overall Technology Stack

- SaaS layer
  - Most of SaaS Applications are implemented based on available modules in the way that new business logics are realized;
    e.g., salesforce.com



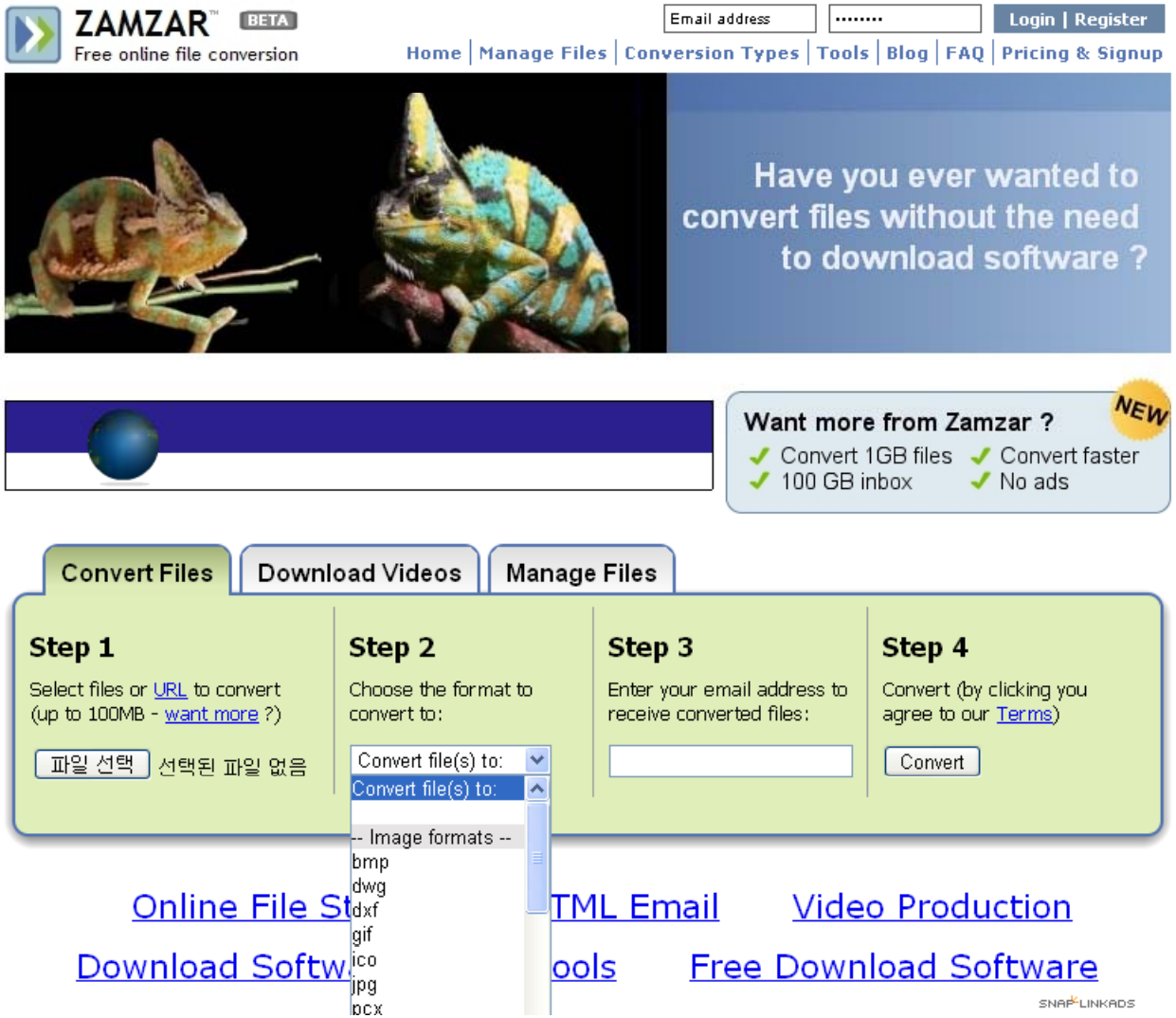
memia

SaaS Business Example 1

그림 3-4-3 AppStore의 Checkout 프로그램

자료 : 'Announcing AppStore', 세일즈포스닷컴 홈페이지, 2006.12.

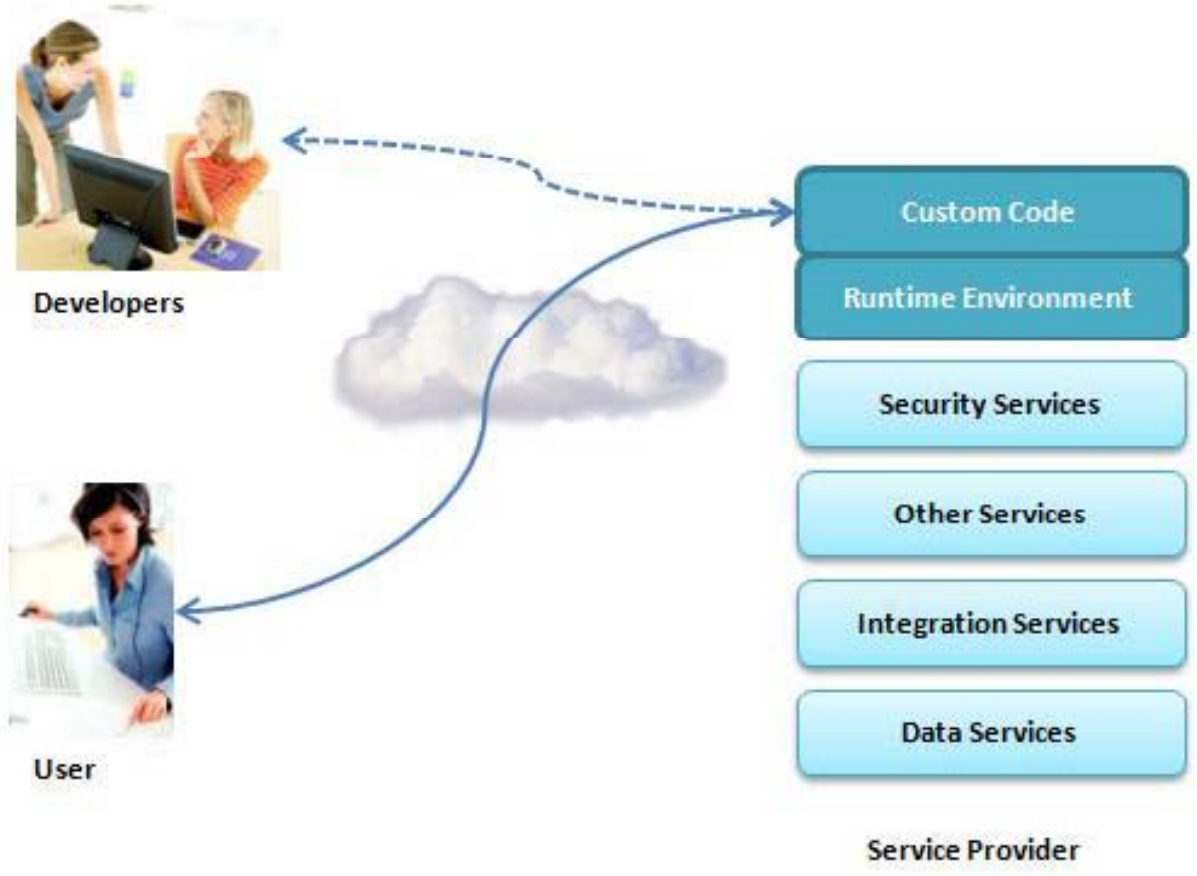
SaaS Business Example 2



Platform-as-a-Service (PaaS)

- Definition
  - Platform providing all the facilities necessary to support the complete process of building and delivering web applications and services, all available over the Internet
  - Entirely virtualized platform that includes one or more servers, operating systems and specific applications

What Is PaaS ?



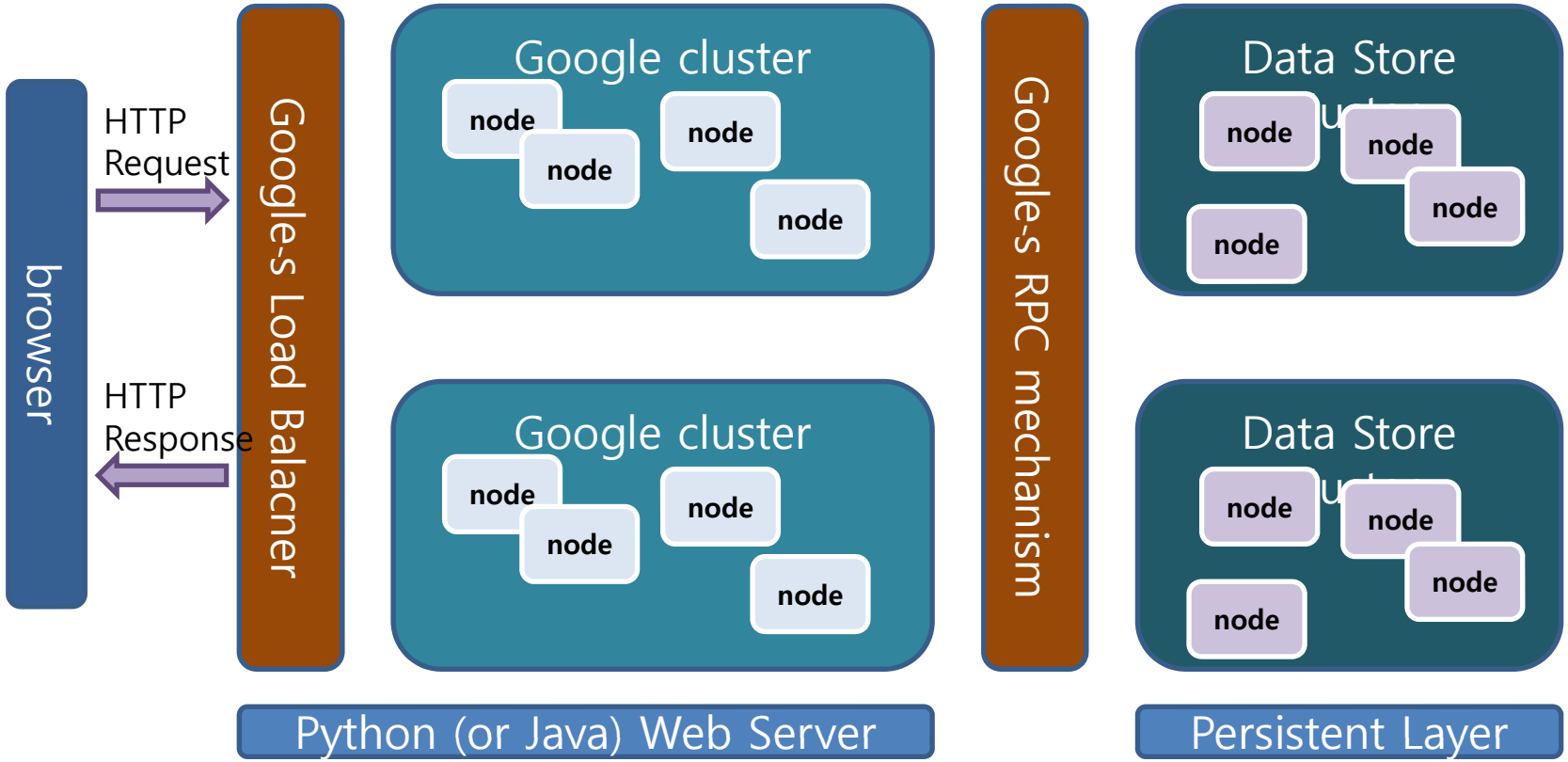
PssS Example: Google App Engine

- Service that allows user to deploy user's Web applications on Google's very scalable architecture
  - Providing user with a sandbox for user's Java and Python application that can be referenced over the Internet
  - Providing Java and Python APIs for persistently storing and managing data (using the Google Query Language or GQL)

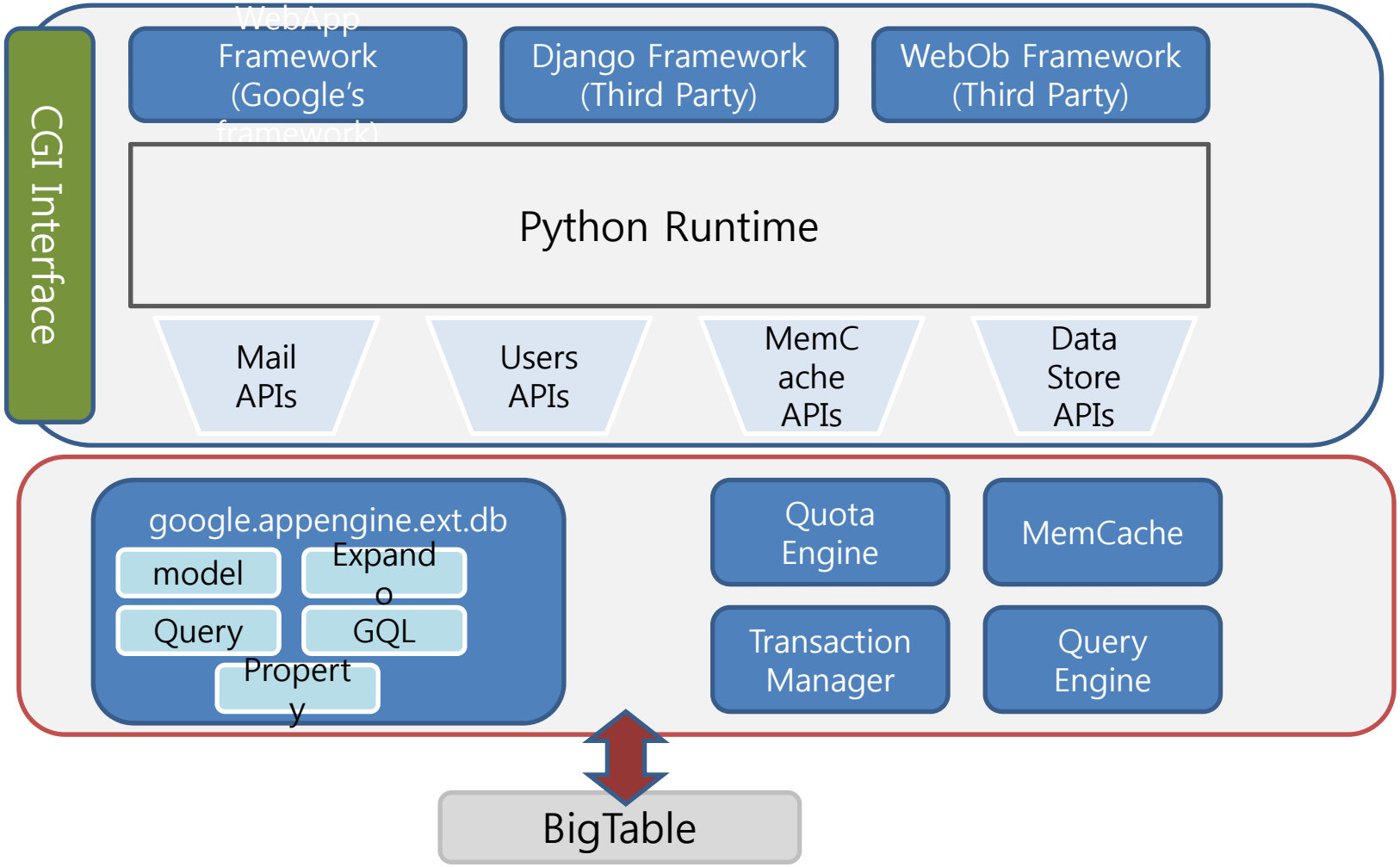
Google App Engine

Google App Engine lets you run your web applications on Google's infrastructure

Google App Engine supports apps written in several programming languages including Python and Java



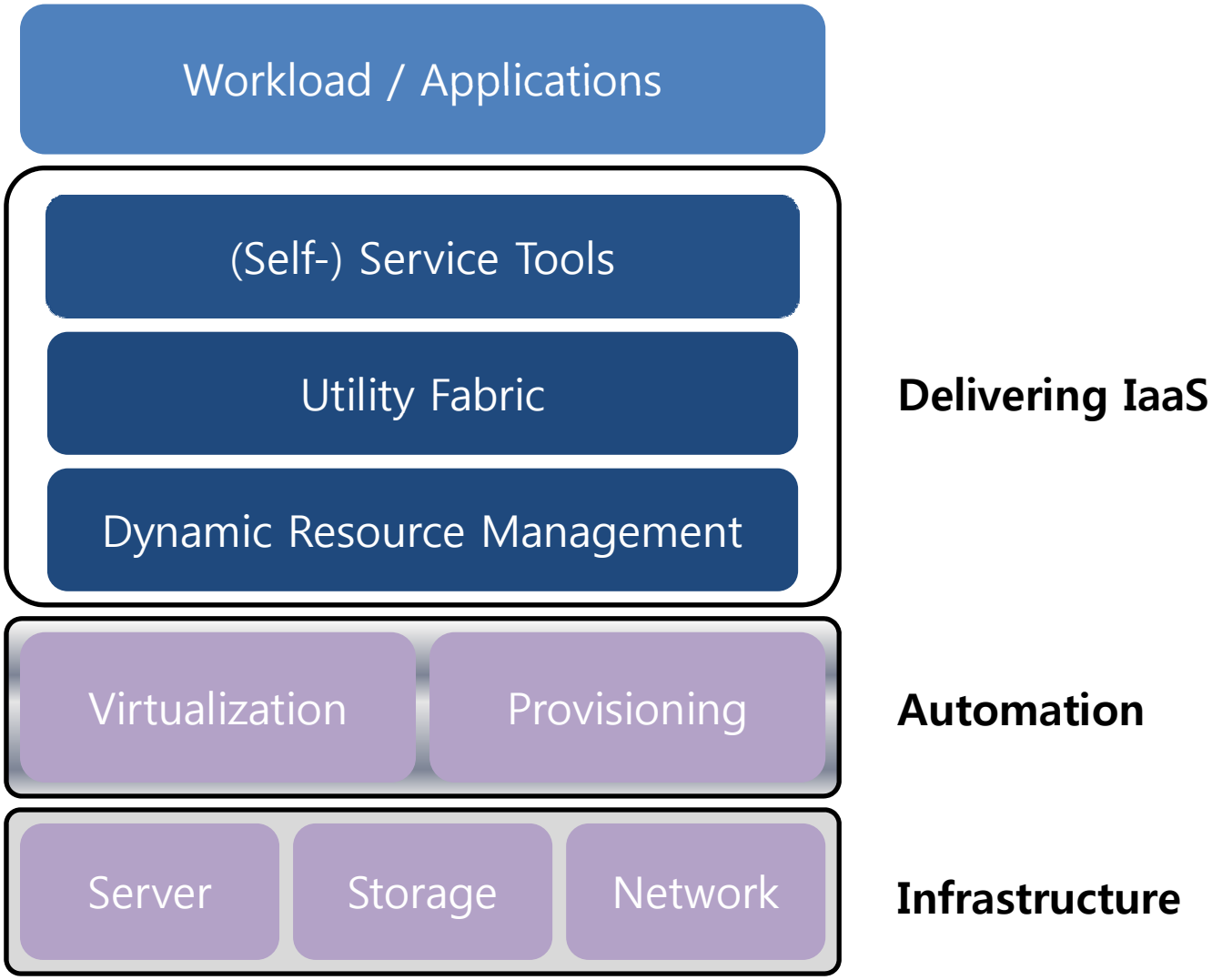
Google App Engine (Python Example)



Infrastructure-as-a-Service (IaaS)

- Definition
  - Provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers and networking components
    - Infrastructure as a Service is sometimes referred to as Hardware as a Service (HaaS).
    - The service provider owns the equipment and is responsible for housing, running and maintaining it
    - The client typically pays on a per-use basis

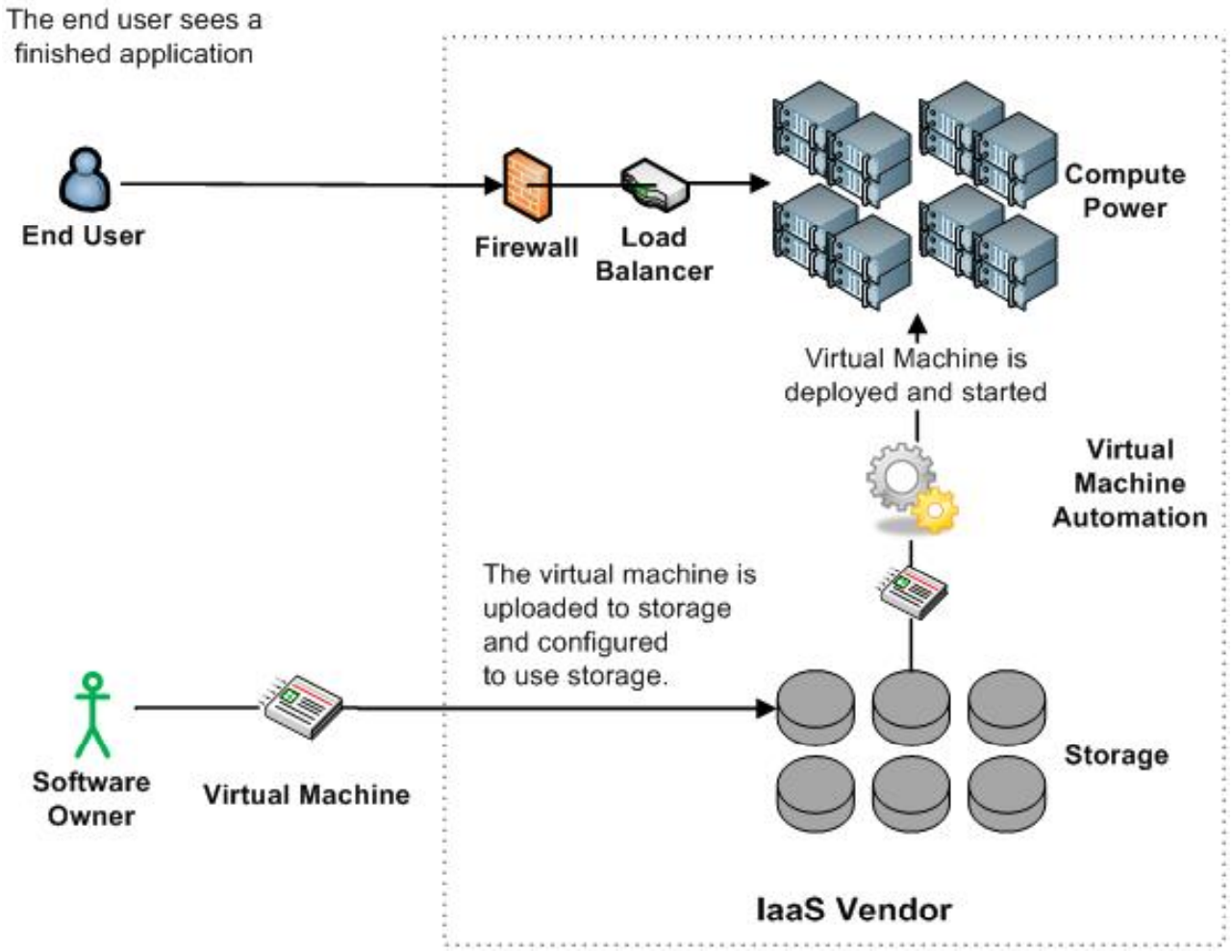
IaaS Platform



What Is Infrastructure-as-a-Service (IaaS)

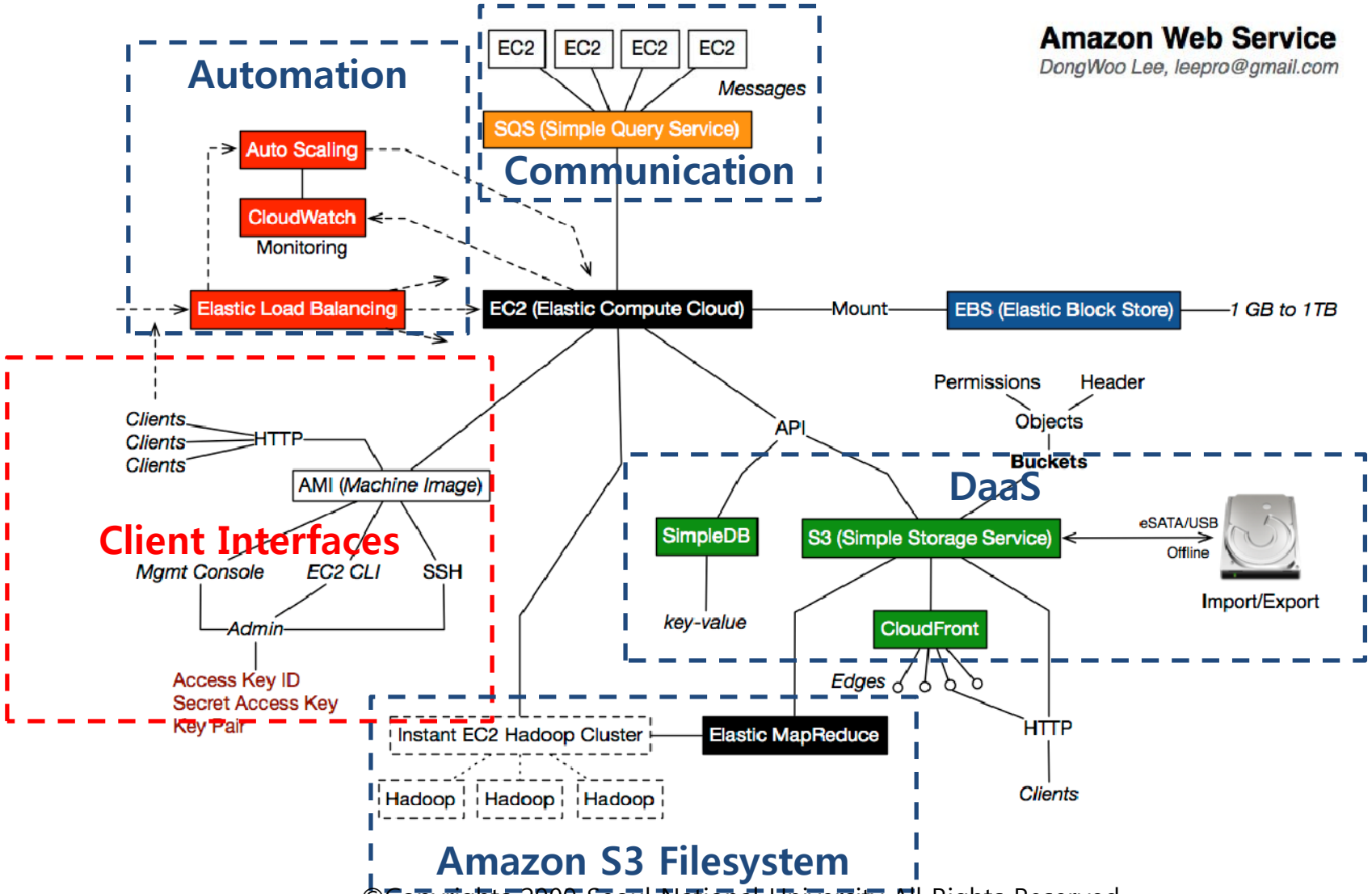
- Characteristics
  - Utility computing and billing model
  - Automation of administrative tasks
  - Dynamic scaling
  - Desktop virtualization
  - Policy-based services
  - Internet connectivity

Use Scenario for IaaS

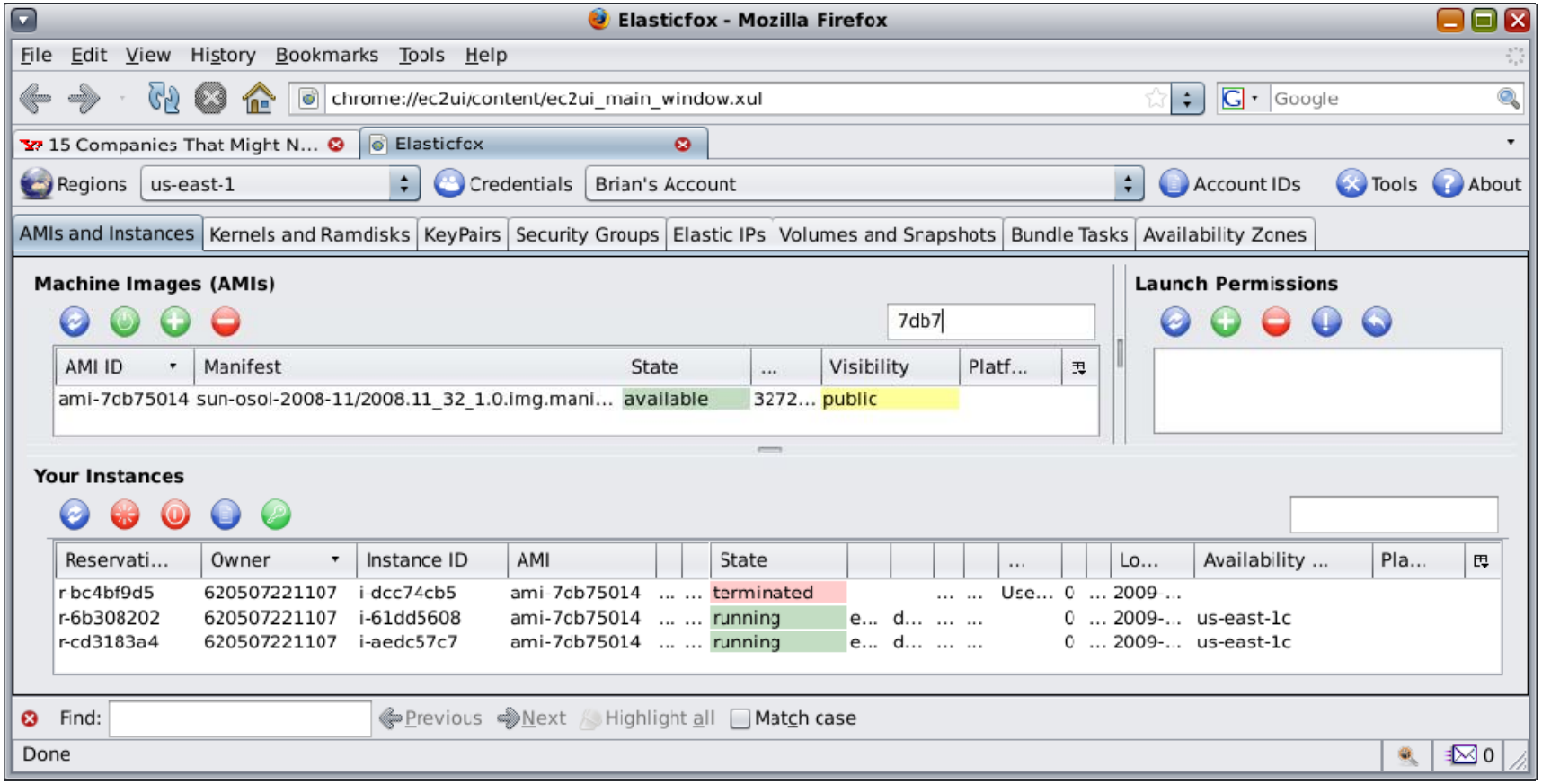


keithpij.com.dnnmax.com

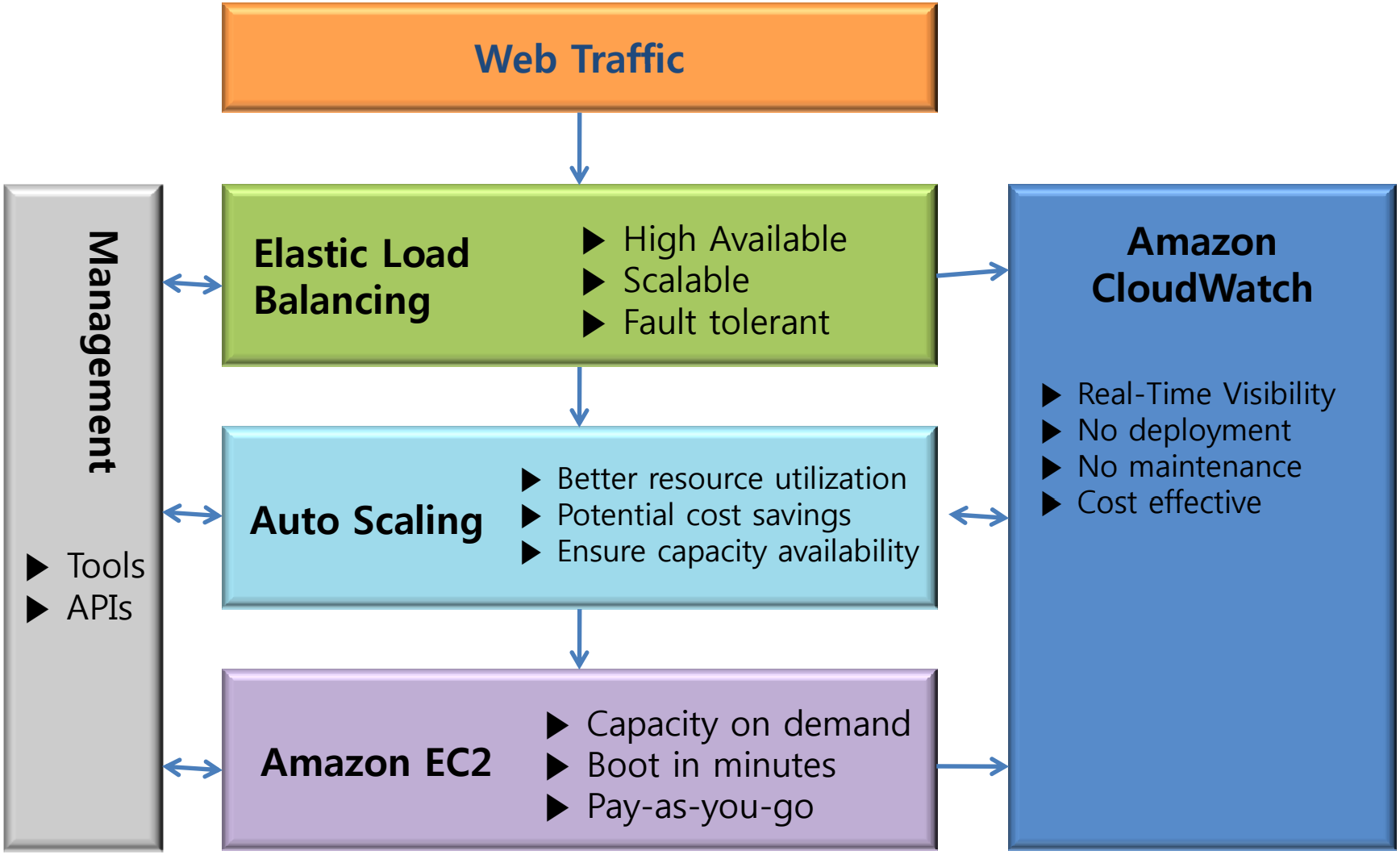
Business Example: Amazon EC2



Amazon EC2 CLI
(Client Level Interface)



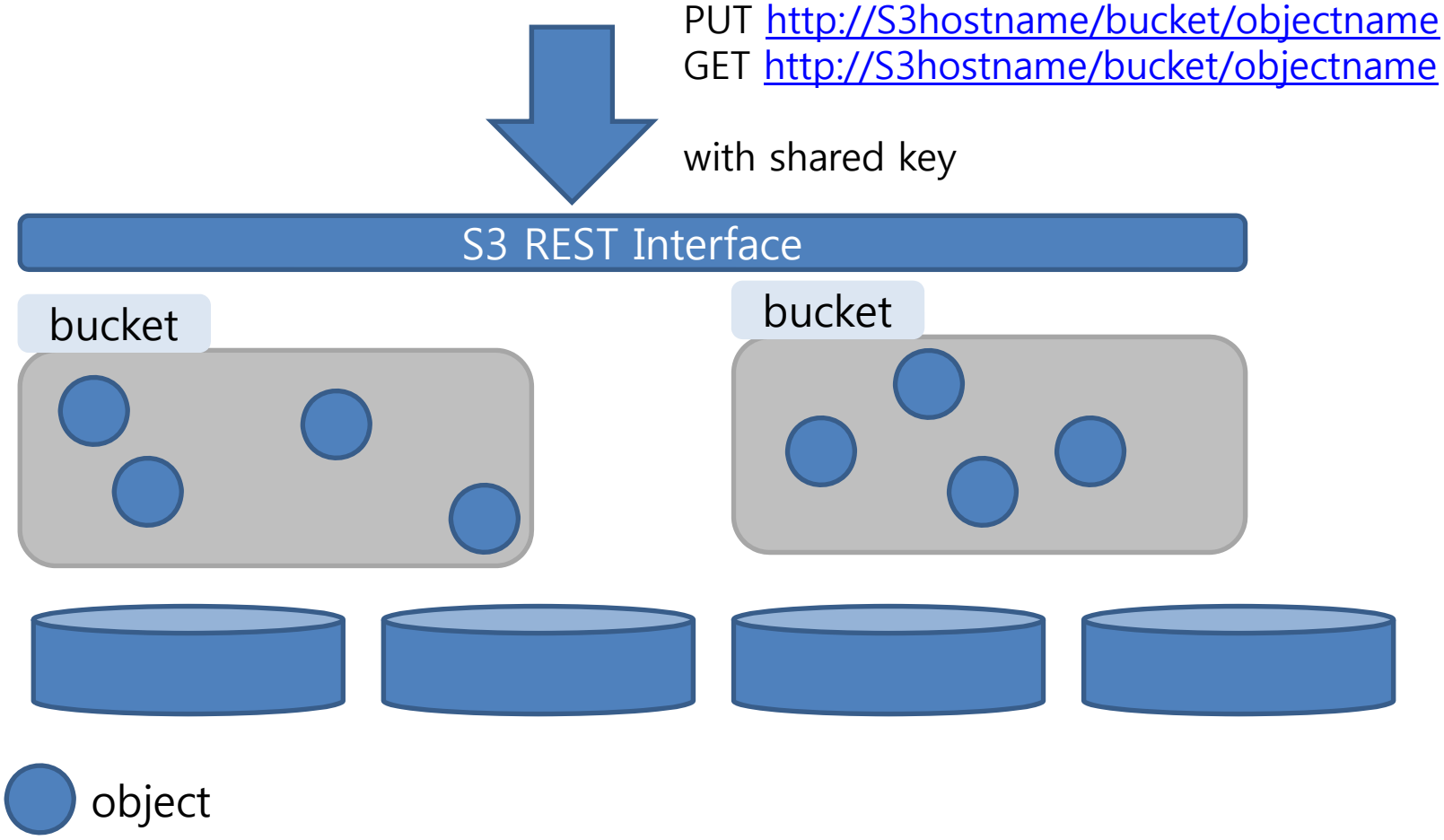
Amazon EC2 Automated Management



Data Storage as a Service (DaaS)

- Definition
  - Delivery of data storage as a service, including database-like services, often billed on a utility computing basis
    - Database (Amazon SimpleDB & Google App Engine's BigTable datastore)
    - Network attached storage (MobileMe iDisk & Nirvanix CloudNAS)
    - Synchronization (Live Mesh Live Desktop component & MobileMe push functions)
    - Web service (Amazon Simple Storage Service & Nirvanix SDN)

Amazon S3



Putting It All Together – Case Study of Alexa Web

- Alexa Web Search web service
  - To build customized search engines against the massive data that Alexa crawls every night
  - To query the Alexa search index and get Million Search Results (MSR) back as output

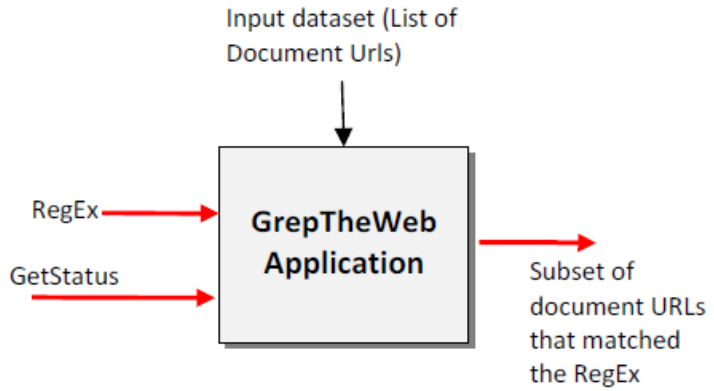


Figure 1 : GrepTheWeb Architecture - Zoom Level 1

GrepTheWeb Architecture

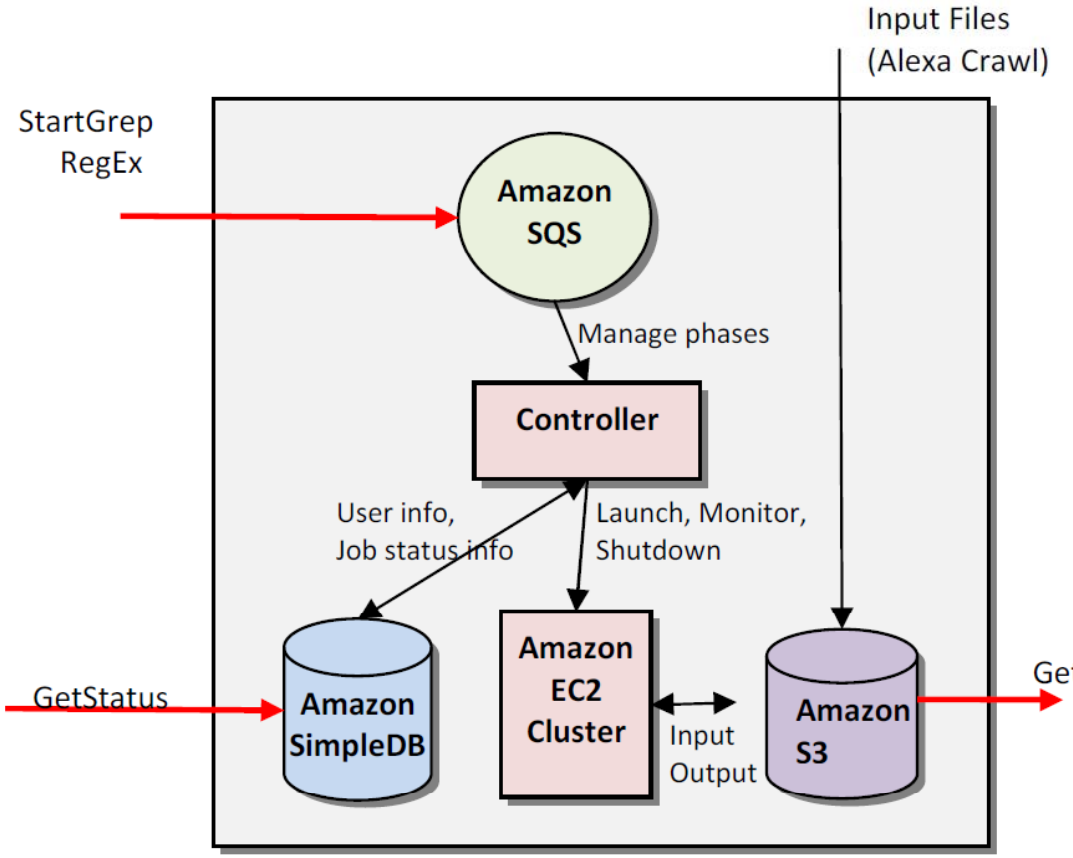


Figure 2: GrepTheWeb Architecture - Zoom Level 2

Amazon S3

◆For retrieving input datasets and for storing the output data set

Amazon SQS

◆For durably buffering requests acting as a "glue" between controllers

Amazon SimpleDB

◆For storing intermediate status, log, and for user data about tasks

Amazon EC2

◆For running a large distributed processing Hadoop cluster on-demand

Hadoop

◆For distributed processing, automatic parallelization, and job scheduling

Data and Control Flow

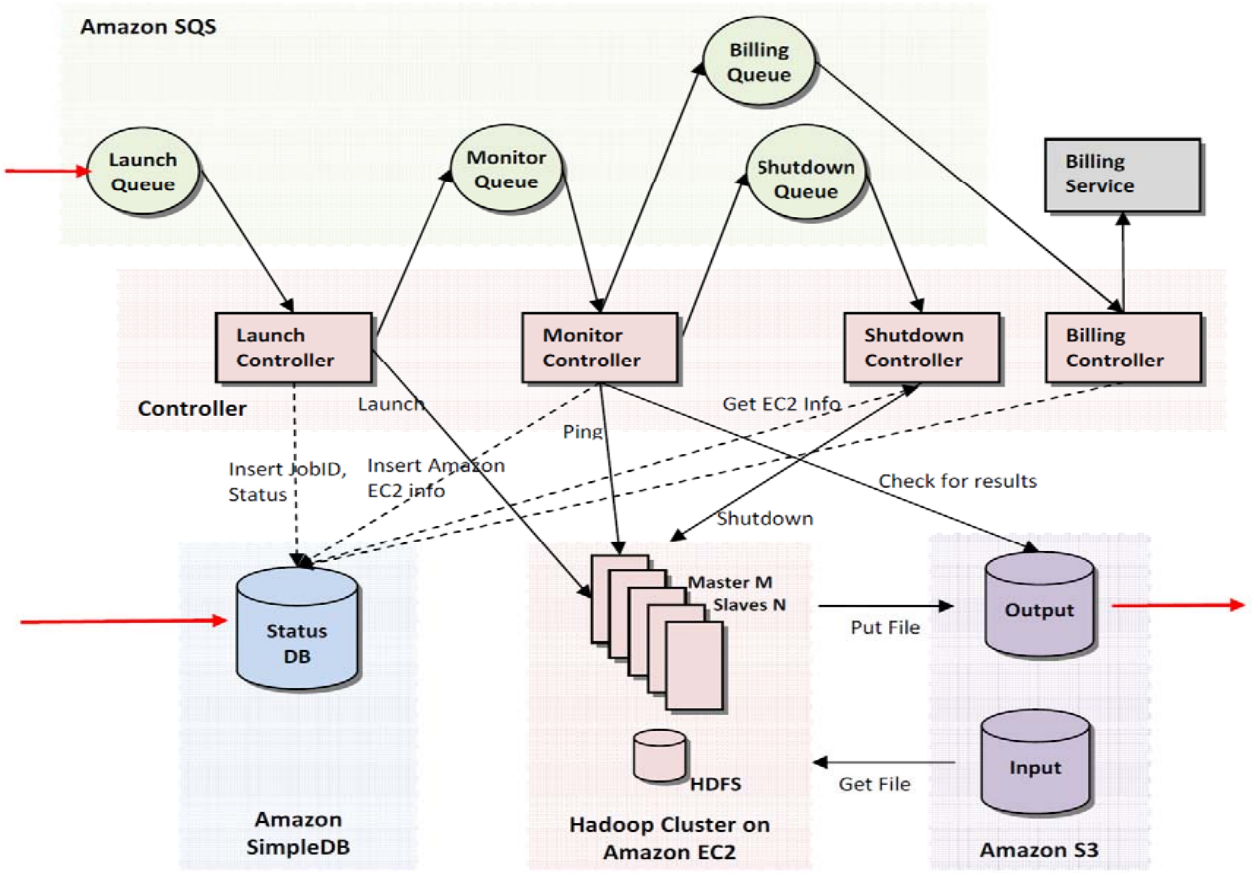


Figure 4: GrepTheWeb Architecture - Zoom Level 3

MapReduce

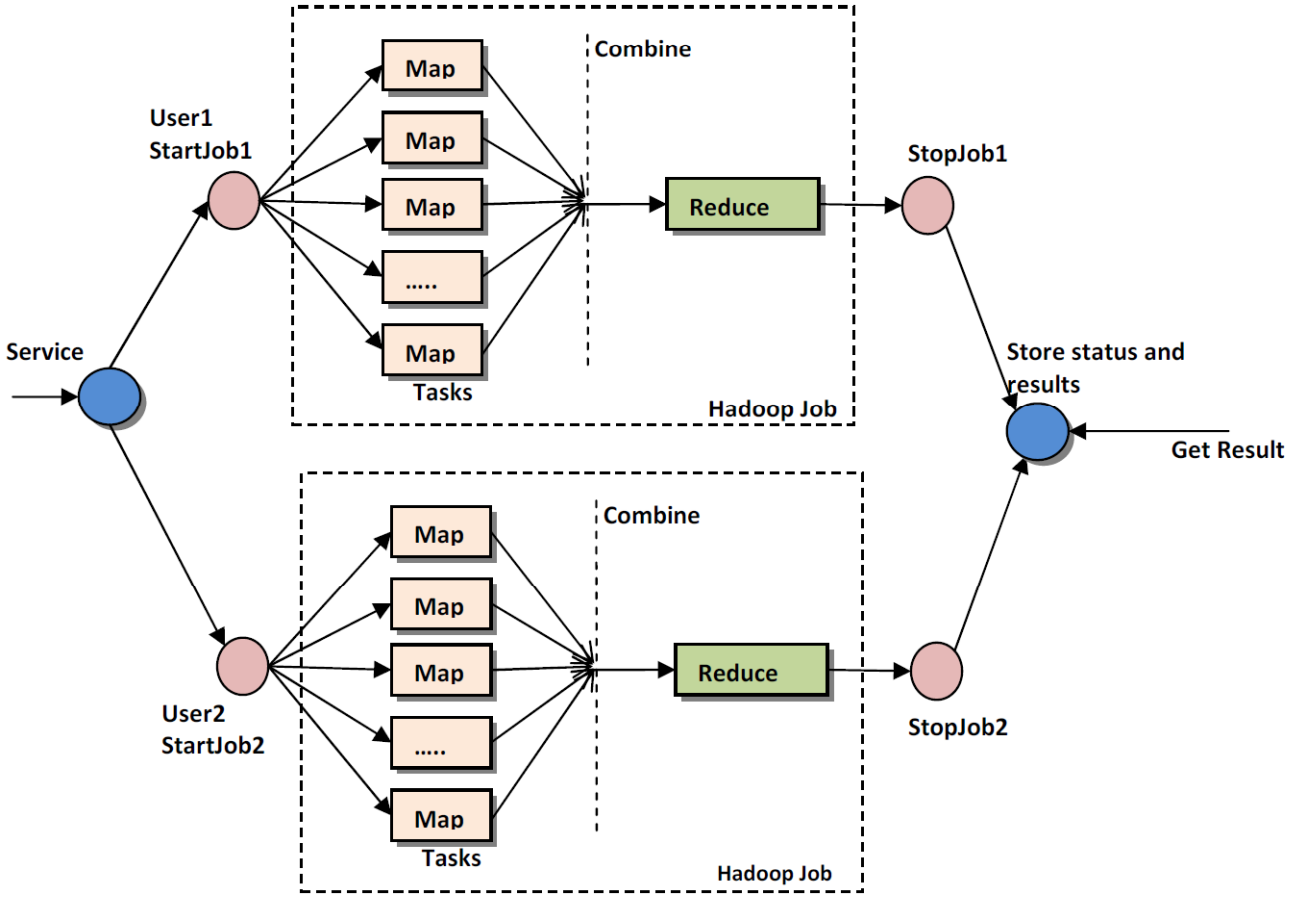


Figure 5: Map Reduce Operation (in GrepTheWeb)

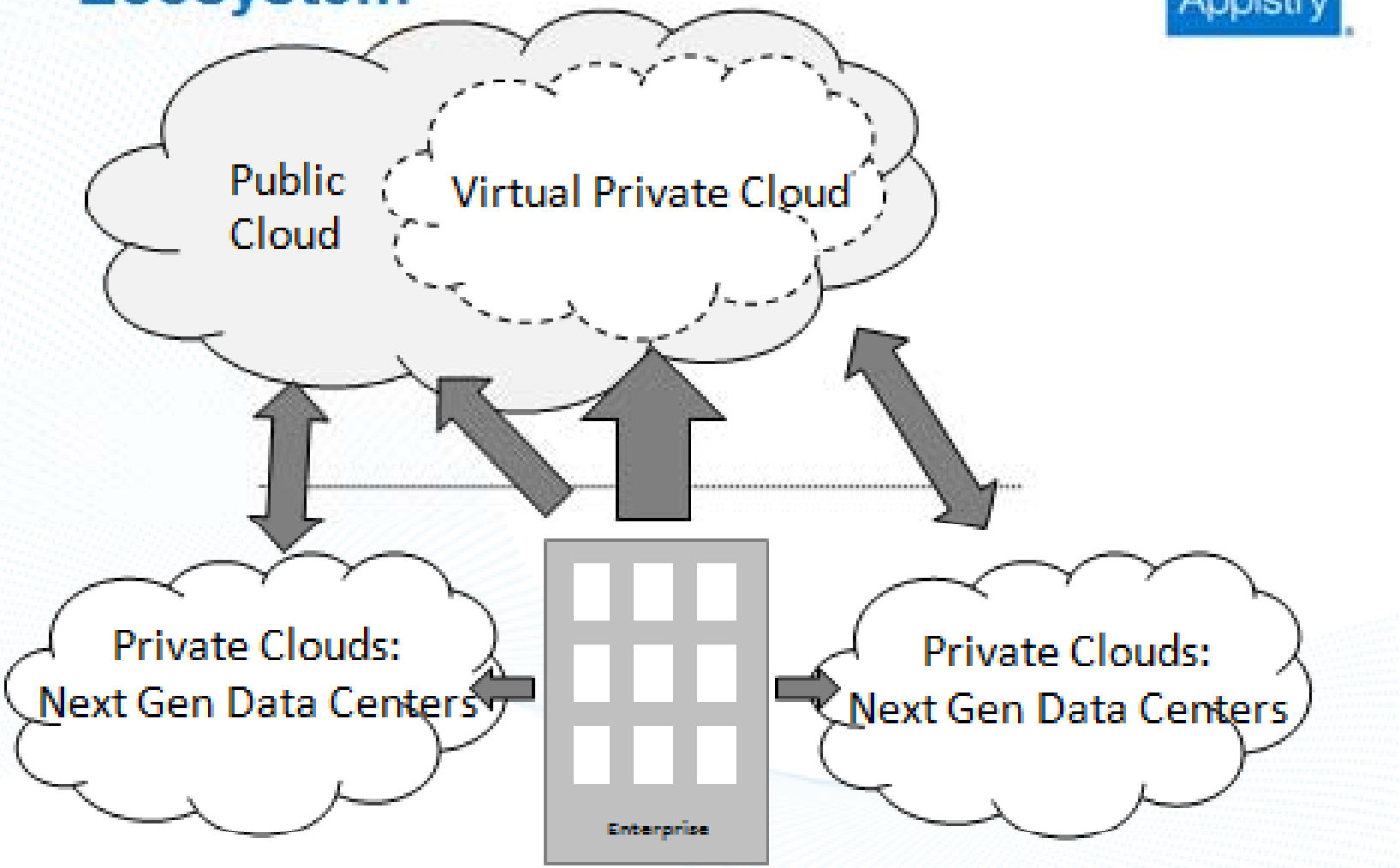
Public Cloud Services

- Definition
  - The standard cloud computing model
    - The SP makes resources, such as applications and storage, available to the general public over the Internet
  - Free or offered on a pay-per-use model
- Examples of public clouds
  - Amazon Elastic Compute Cloud (EC2), IBM's Blue Cloud, Sun Cloud, Google AppEngine and Windows Azure Services Platform.

Private Cloud Services

- Internal cloud or corporate cloud
- Definition
  - Proprietary computing architecture that provides hosted services to a limited number of people behind a firewall.
    - Designed to appeal to an organization that needs or wants more control over their data than they can get by using a third-party hosted service

Emerging Cloud Computing Ecosystem



Inter-Cloud

- Definition
  - Federation of clouds based on open standards
  - Concept primarily promoted by Cisco
- Similar Concepts
  - Cloud of Clouds, Cloud Interoperability, etc.

Why “Inter-Cloud”?

- InterGrid
  - Acquiring resources from peering Grids to serve occasional peak demands
  - Solving the problem of resource provisioning in environments with multiple Grids
- Evolvable and sustainable system

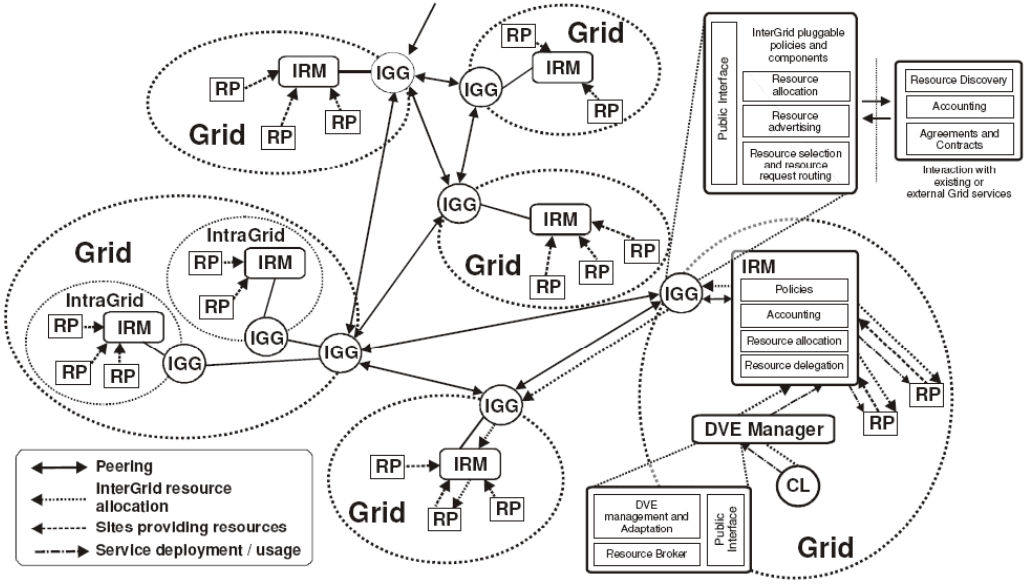
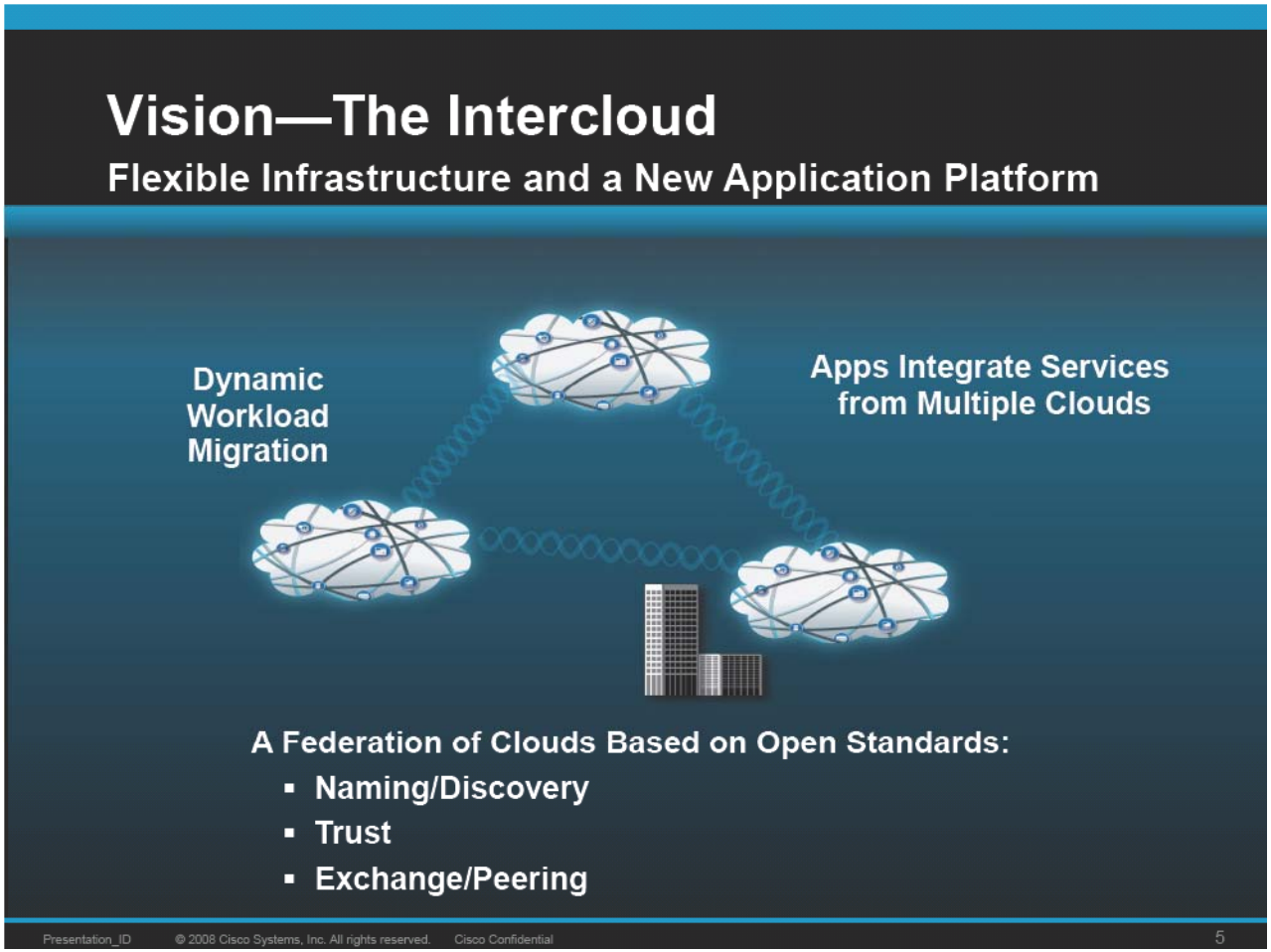


Figure 3. Architecture for internetworking of Grids.

Cisco's Vision



Architecture of Inter-Cloud Standards (by Cisco)

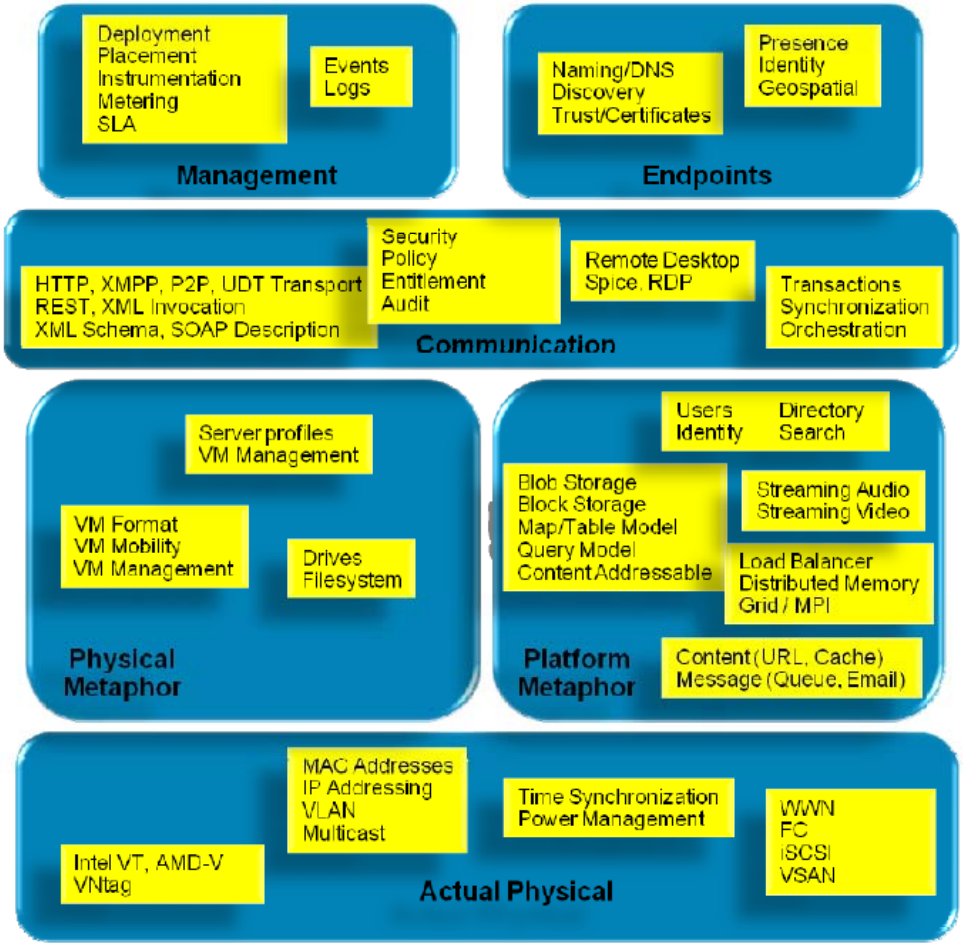


Figure 4. An Architecture for Intercloud Standards

Technical Issues of Cloud Computing

- Virtualization Security
- Reliability
- Monitoring
- Manageability

Virtualization Security (1/2)

- Virtualization
  - Abstracting the underlying resources so that multiple operating systems can be run on a single physical simultaneously
  - Improving resource utilization by sharing available resources to multiple on demand needs

Virtualization Security (2/2)

- Virtualization security
  - Including the standard enterprise security policies on access control, activity monitoring and patch management
  - Most enterprises just starting to grasp but not fully understanding
    - Many IT people still believe that the hypervisor and virtual machines are safe
  - Becomeing one of the factors when virtualization technologies move into the cloud
    - Access control and monitoring of the virtual infrastructure will be on top of providers' mind

Reliability

- One of the biggest factors in enterprise adoption
  - Almost no SLAs provided by the cloud providers today
    - Enterprises cannot reasonably rely on the cloud infrastructures/platforms to run their business
    - Amazon said that “AWS (Amazon Web Service) only provides SLA for their S3 service”
  - Hard to imagine enterprises signing up cloud computing contracts without SLAs clearly defined
- Some startup coming up with clever idea to provide SLA as a third party vendor
- Cloud providers to grow/wake up and actually do something to encourage the enterprise adoption

Monitoring (1/2)

- Critical to any IT for performance, availability and payment
- Monitoring in Cloud Computing
  - How much CPU or memory the machines are using
    - CPU and memory usage are misleading most of the time in virtual environments
  - Only real measurements: how long your transactions are taking and how much latencies there are
- High Availability's article on latency
  - Amazon finding every 100ms of latency cost them 1% in sales
  - Google finding an extra .5 seconds in search page generation time dropped traffic by 20%

Monitoring (2/2)

- High Availability's article on latency (cont'd)
  - Broker possibly losing $4 million in revenues per millisecond if their electronic trading platform is 5 milliseconds behind the competition
- Hypernic's CloudStatus: one of the first to recognize this issue and develop a solution
  - Monitoring Amazon's web services
  - Recently added monitoring for Google App Engine
- RightScale's solution possibly providing monitoring for the virtual machines under their management

Manageability (1/2)

- Most of IaaS/PaaS providers being raw infrastructures and platforms that do not have great management capabilities
- Auto-scaling: example of missing management capabilities for cloud infrastructures
  - Amazon EC2 claiming to be elastic; however, it really means that it has the *potential* to be elastic
  - Amazon EC2 not automatically scaling your application as your server becomes heavily loaded

Manageability (2/2)

- Many startups having recognized the need for management early on and built management capabilities on top of the existing cloud infrastructure/platforms
  - RightScale being one of the early pioneers
  - Solving many of the management issues such as auto-scaling and load balancing

Q & A

Thank you!

Eom, Hyeonsang
School of Computer Science & Engineering
Seoul National University