EE 542
Lecture 9: Cloud Computing

Internet and Cloud Computing

Young Cho
Department of Electrical Engineering
University of Southern California

What is Cloud Computing?

- Cloud computing is a model for enabling *convenient*, *on-demand network access* to a *shared pool* of *configurable computing resources* (e.g., networks, servers, storage, applications, and services) [Mell_2009], [Berkely_2009].
- It can be *rapidly provisioned* and *released* with minimal management effort.
- It provides *high level abstraction* of computation and storage model.
- It has some essential **characteristics, service models**, and **deployment models**.

Essential Characteristics

- ***On-Demand Self Service:***
  - A consumer can unilaterally provision computing capabilities, automatically without requiring human interaction with each service's provider.
- ***Heterogeneous Access:***
  - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous ***thin*** or ***thick*** client platforms.

Essential Characteristics (cont.)

- ***Resource Pooling:***
  - The provider's computing resources are pooled to serve multiple consumers using a *multi-tenant model.*
  - Different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
- ***Measured Service:***
  - Cloud systems *automatically control* and *optimize* resources used by leveraging a metering capability at some level of abstraction appropriate to the type of service.
  - ***It will provide analyzable and predictable computing platform.***

Service Models

- ***Cloud Software as a Service (SaaS):***
  - The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure.
  - The applications are accessible from various client devices such as a web browser (e.g., web-based email).
  - The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage,…
  - ***Examples: Caspio, Google Apps, Salesforce, Nivio, Learn.com.***

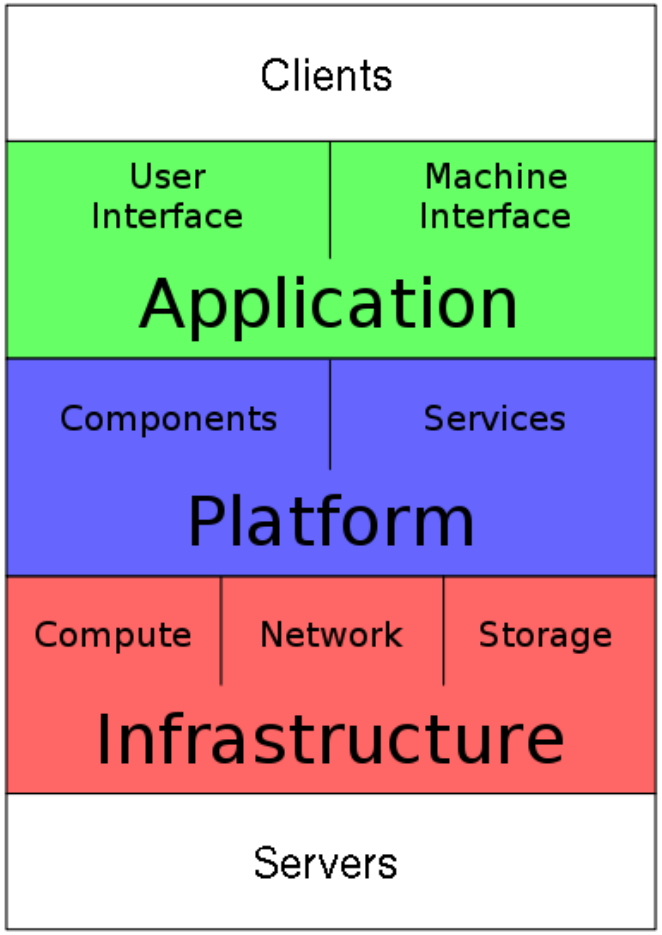
Service Models (cont.)

- ***Cloud Platform as a Service (PaaS):***
  - The capability provided to the consumer is to deploy onto the cloud infrastructure *consumer-created* or *acquired applications* created using *programming languages and tools* supported by the provider.
  - The consumer does not manage or control the underlying cloud infrastructure.
  - Consumer has control over the deployed applications and possibly application hosting environment configurations.
  - ***Examples: Windows Azure, Google App.***

Service Models (cont.)

- ***Cloud Infrastructure as a Service (IaaS):***
  - The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources.
  - The consumer is able to deploy and run arbitrary software, which can include operating systems and applications.
  - The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).
  - ***Examples: Amazon EC2, GoGrid, iland, Rackspace Cloud Servers, ReliaCloud.***

Service Models (cont.)



Cloud Computing Stack

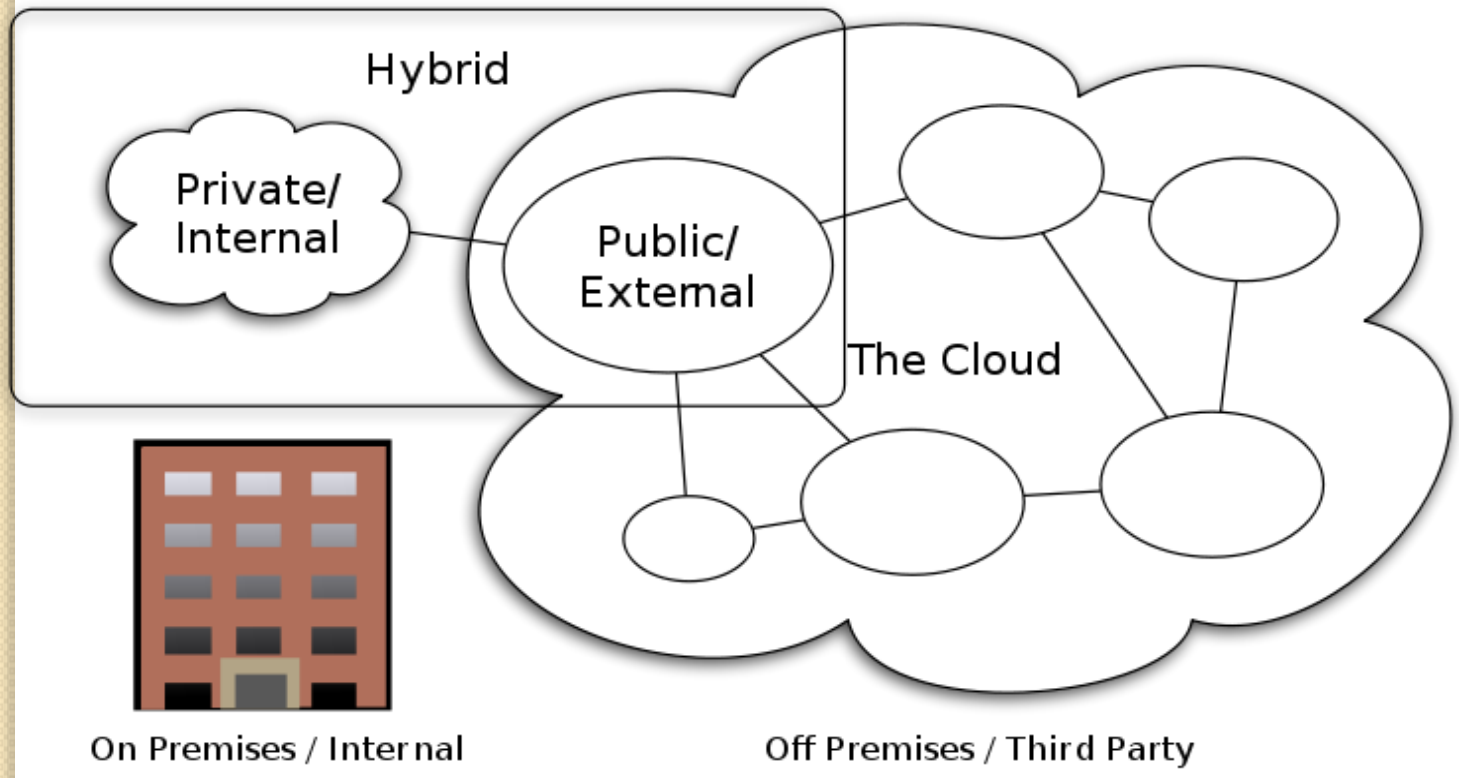
Service Model at a glance: Picture From http://en.wikipedia.org/wiki/File:Cloud_Computing_Stack.svg

Types of Cloud

- Private Cloud
  - The cloud is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.
- Community Cloud (Federated)
  - The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns.
  - It may be managed by the organizations or a third party and may exist on premise or off premise

Types of Cloud

- Public Cloud
  - The cloud infrastructure is made available to the general public or a large industry group and it is owned by an organization selling cloud services.
- Hybrid Cloud
  - The cloud infrastructure is a composition of two or more clouds (private, community, or public).



Cloud Computing Types

Advantages of Cloud Computing

- Cloud computing do not need high quality equipment for user, and it is very easy to use.
- Provides dependable and secure data storage center.
- Reduce run time and response time.
- Cloud is a large resource pool that you can buy on-demand service.
- Scale of cloud can extend dynamically providing nearly infinite possibility for users to use internet.

Infrastructure as a Service

- A category of cloud services which provides capability to provision processing, storage, intra-cloud network connectivity services, and other fundamental computing resources of the cloud infrastructure.

Source- [ITU –Cloud Focus Group]

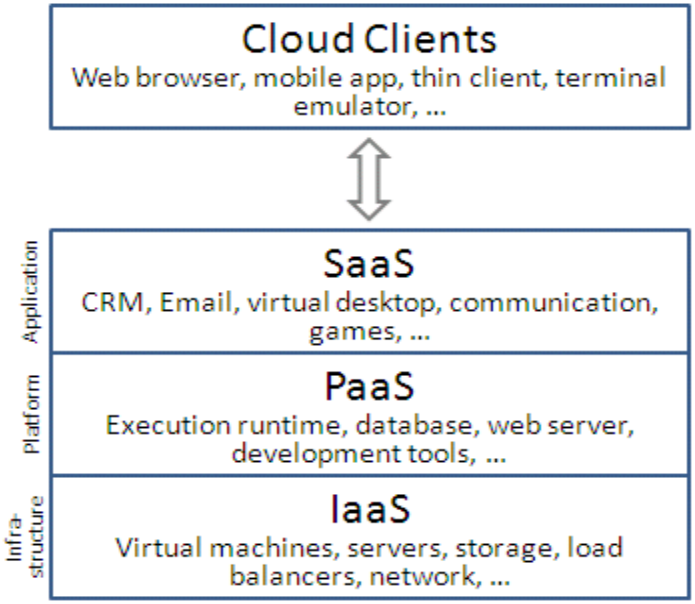


Diagram Source: Wikipedia

Highlights of IaaS

- On demand computing resources
  - Eliminate the need of far ahead planning
- No up-front commitment
  - Start small and grow as required
  - No contract, Only credit card!
- Pay for what you use
- No maintenance
- Measured service
- Scalability
- Reliability

Elastic Compute Cloud (EC2)

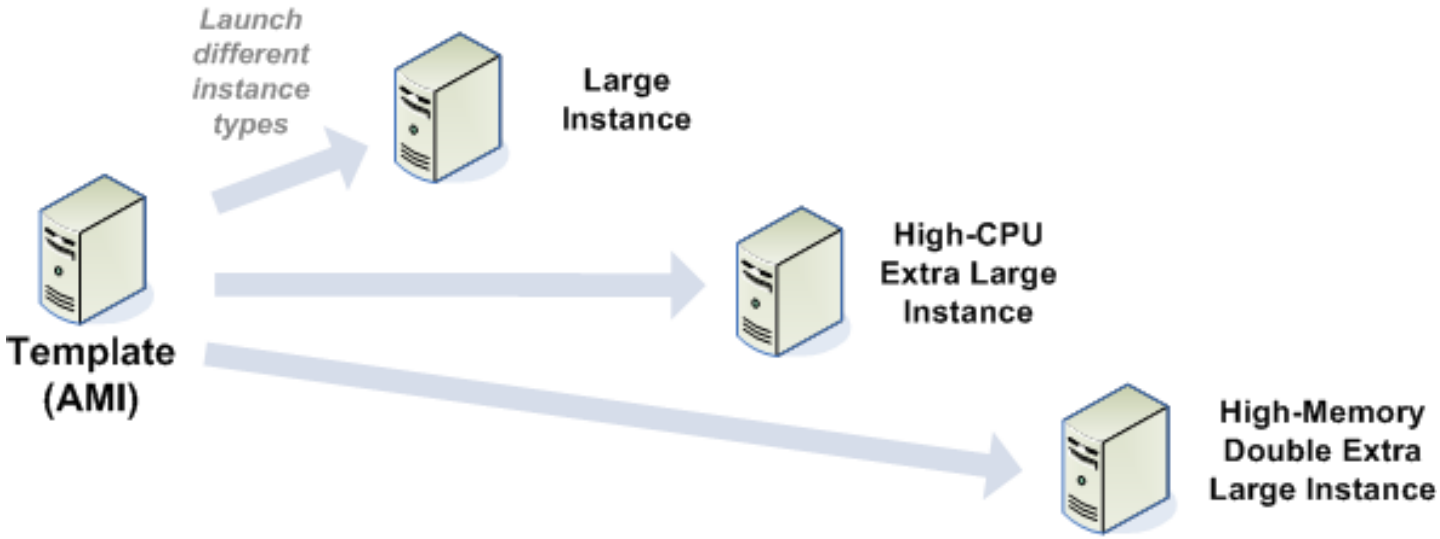
- Amazon Elastic Compute Cloud (EC2) is a web service that provides resizeable computing capacity that one uses to build and host different software systems.
- Designed to make web-scale computing easier for developers.
- A user can create, launch, and terminate server instances as needed, paying by the hour for active servers, hence the term "elastic".
  - Provides scalable, pay as-you-go compute capacity
  - Elastic - scales in both direction

EC2 Concepts

- AMI & Instance
- Region & Zones
- Storage
- Networking and Security
- Monitoring
- Auto Scaling
- Load Balancer

Amazon Machine Images (AMI)

- Is an immutable representation of a set of disks that contain an operating system, user applications and/or data.
- From an AMI, one can launch multiple instances, which are running copies of the AMI.



AMI and Instance

- Amazon Machine Image (AMI)
  - Software configuration templates
  - Operating System, Application Server, and Applications
- Instance is an AMI running on virtual servers
  - Type offers different compute and memory facilities

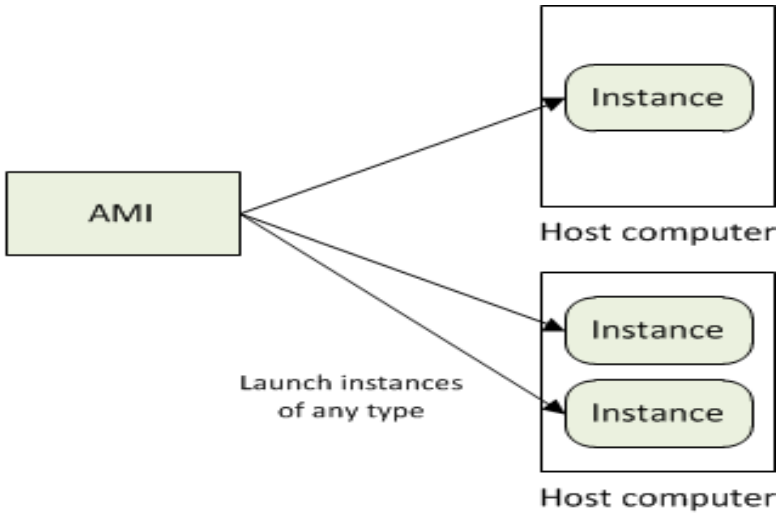


Diagram Source: http://docs.aws.amazon.com

Type	CPU	Memory	Local Storage	Platform	I/O	Name
Small	1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit)	1.7 GB	160 GB instance storage (150 GB plus 10 GB root partition)	32-bit	Moderate	m1.small
Large	4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each)	7.5 GB	850 GB instance storage (2 x 420 GB plus 10 GB root partition)	64-bit	High	m1.large
Extra Large	8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each)	15 GB	1690 GB instance storage (4 x 420 GB plus 10 GB root partition)	64-bit	High	m1.xlarge
Micro	Up to 2 EC2 Compute Units (for short periodic bursts)	613 MB	None (use Amazon EBS volumes for storage)	32-bit or 64-bit	Low	t1.micro
High-CPU Medium	5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each)	1.7 GB	350 GB instance storage (340 GB plus 10 GB root partition)	32-bit	Moderate	c1.medium

Region and Zones

- Data centers in different Regions across the globe
- An Instance can be launched in different Regions
  - Depending on the need
  - Closer to specific customer
  - Meet legal or other requirements
- Each Region has a set of Zones
  - Zones are isolated from failure in other zones
  - Low latency connectivity between zones in the same region

Storage

- Amazon EC2 provides three type of storage option
  - Amazon EBS
  - Amazon S3
  - Instance Storage

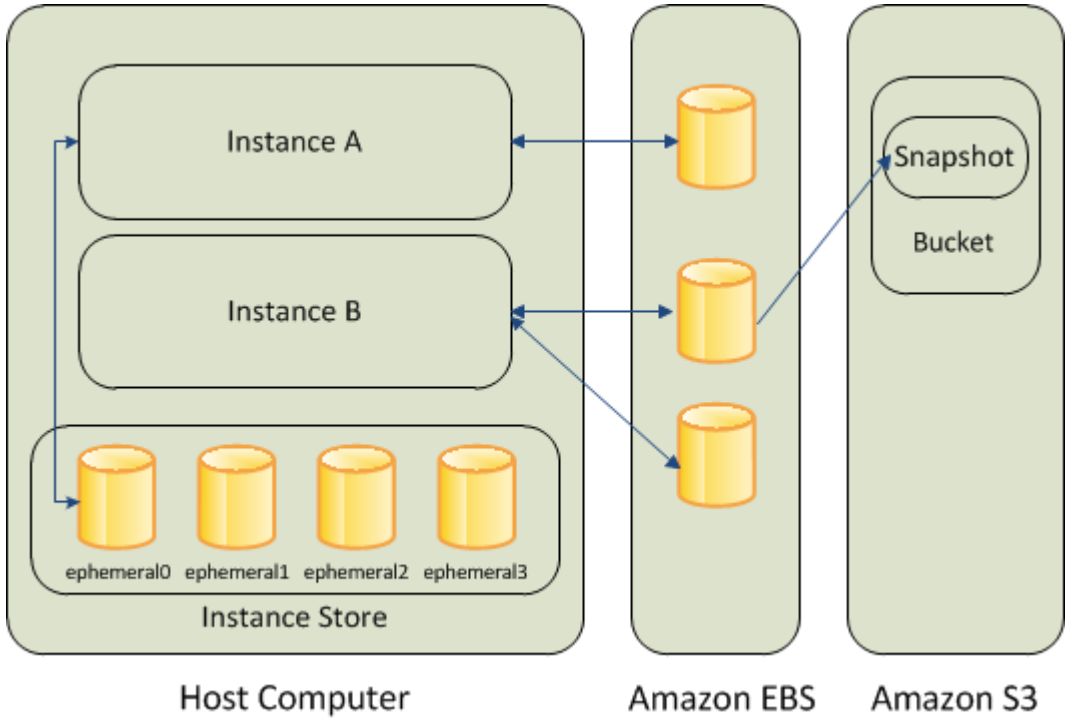
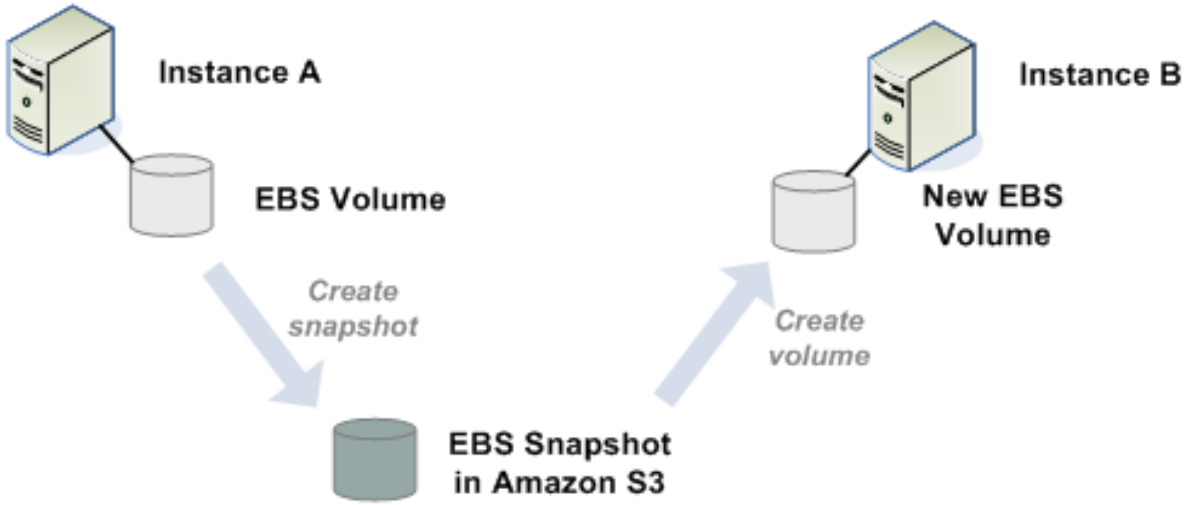


Diagram Source: http://docs.aws.amazon.com

Elastic Block Store(EBS) volume

- An EBS volume is a read/write disk that can be created by an AMI and mounted by an instance.
- Volumes are suited for applications that require a database, a file system, or access to raw block-level storage.



Amazon S3

- Simple Storage Service
  - Service Oriented Architecture
  - Provides online storage using web services.
  - Allows read, write and delete permissions on objects.
  - Specific messaging interface

Amazon SimpleDB

- Highly available
- Flexible, and scalable
- Non-relational data store
- Offloads the work of database administration.
- Creates and manages multiple geographically distributed replicas
- Fees based on resources consumed in data storage and query rates

Networking and Security

- Instances can be launched on one of the two platforms
  - EC2-Classic
  - EC2-VPC
- Each instance launched is assigned two addresses a private address and a public IP address.
  - A replacement instance has a different public IP address.
- Instance IP address is dynamic.
  - new IP address is assigned every time instance is launched
- Amazon EC2 offers Elastic IP addresses (static IP addresses) for dynamic cloud computing.
  - Remap the Elastic IP to new instance to mask failure
  - Separate pool for EC2-Classic and VPC
- Security Groups to access control to instance

Monitoring, Auto Scaling, and Load Balancing

- Monitor statistics of instances and EBS
  - CloudWatch
- Automatically scales amazon EC2 capacity up and down based on rules
  - Add and remove compute resource based on demand
  - Suitable for businesses experiencing variability in usage
- Distribute incoming traffic across multiple instances
  - Elastic Load Balancing

What is Hadoop?

- Map-Reduce Solution
- Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.
- It is a flexible and highly-available architecture for large scale computation and data processing on a network of commodity hardware.
- Designed to answer the question: "**How to process big data with reasonable cost and time?**"

Google Origins

2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*

2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.

2006

Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
{fay,jeff,sanjay,wilsonh,kerr,m3b,tushar,fikes,gruber}@google.com
Google, Inc.

Abstract

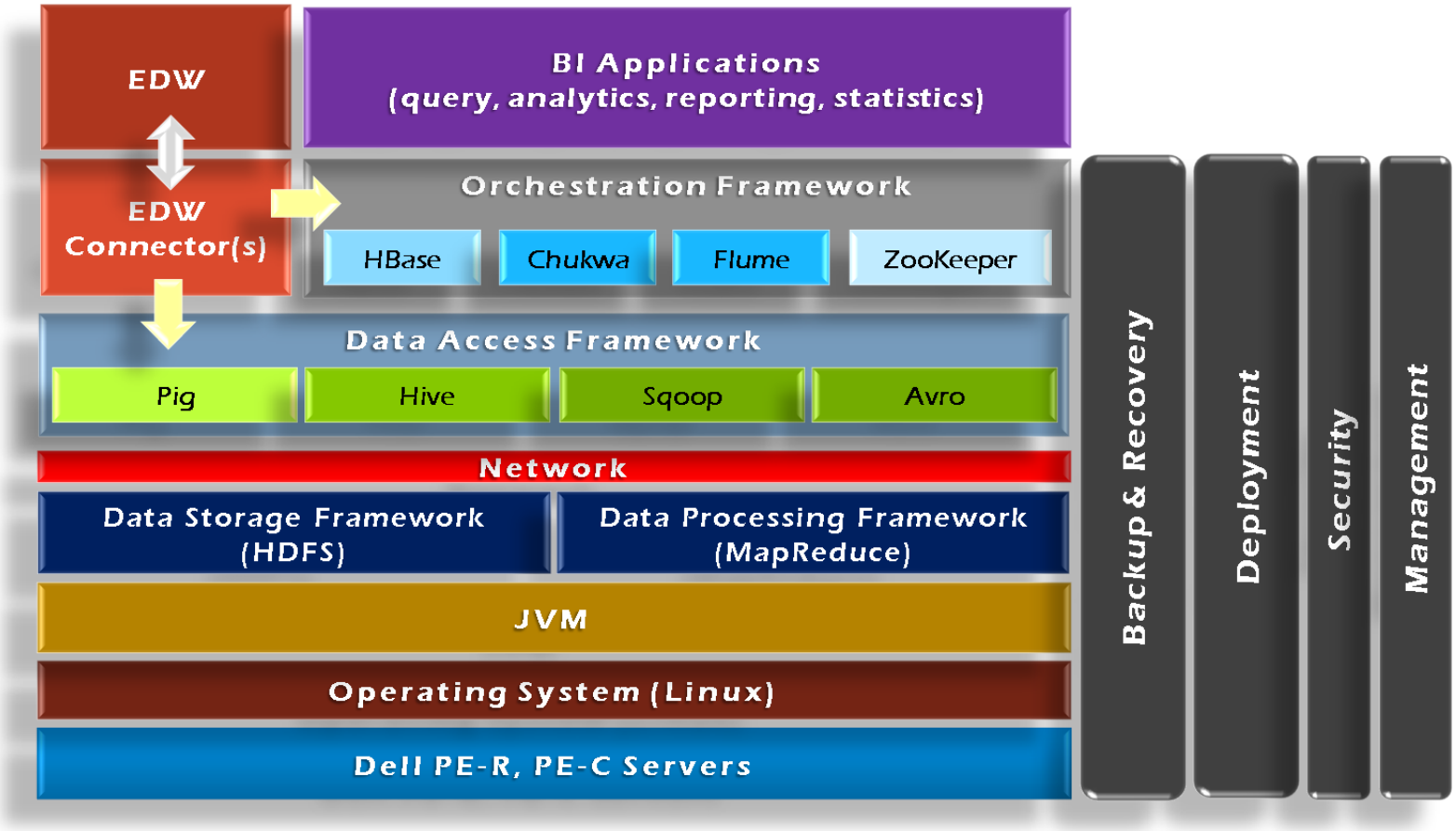
gtable is a distributed storage system for managing
tured data that is designed to scale to a very large
petabytes of data across thousands of commodity
rs. Many projects at Google store data in Bigtable,
ding web indexing, Google Earth, and Google Fi-
e. These applications place very different demands
igtable, both in terms of data size (from URLs to

achieved scalability and high performance, but Big
provides a different interface than such systems. Big
does not support a full relational data model; inste
provides clients with a simple data model that sup
dynamic control over data layout and format, an
lows clients to reason about the locality properties o
data represented in the underlying storage. Data i
dexed using row and column names that can be arbi
strings. Bigtable also treats data as uninterpreted str

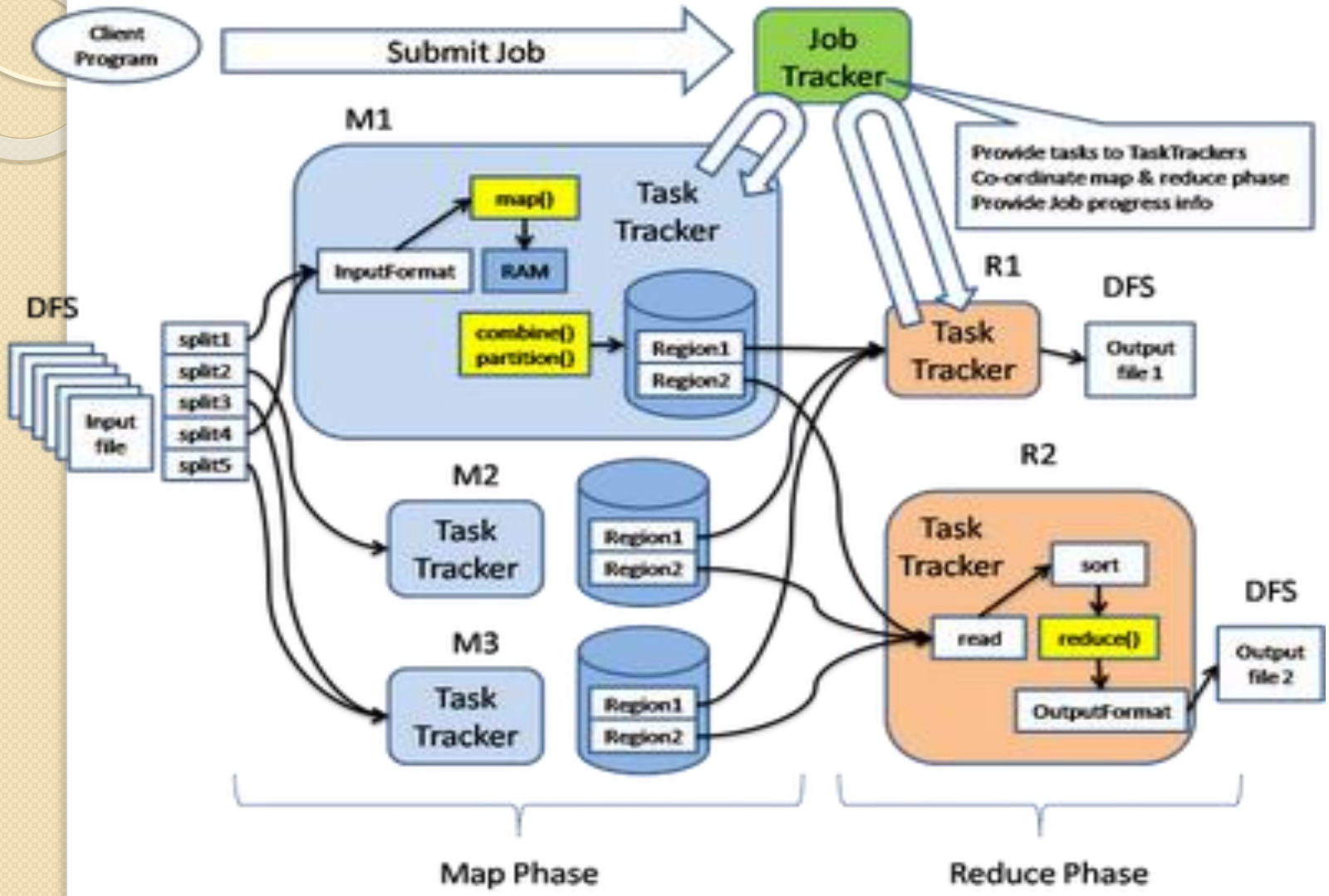
Hadoop Milestones

- **2008 - Hadoop Wins Terabyte Sort Benchmark (**sorted 1 terabyte of data in 209 seconds, compared to previous record of 297 seconds)
- 2009 - Avro and Chukwa became new members of Hadoop Framework family
- 2010 - Hadoop's Hbase, Hive and Pig subprojects completed, adding more computational power to Hadoop framework
- **2011 - ZooKeeper Completed**
- **2013 - Hadoop 1.1.2 and Hadoop 2.0.3 alpha.**

Hadoop Framework Tools



MapReduce Engine



Hadoop Use

- Hadoop is in use to handle big data:
  - Yahoo!'s Search Webmap runs on 10,000 core Linux cluster and powers Yahoo! Web search
  - FB's Hadoop cluster hosts 100+ PB of data (July, 2012) & growing at ½ PB/day (Nov, 2012)
  - Amazon and Netflix
  - NY Times was dynamically generating PDFs of articles from 1851-1922
    - Wanted to pre-generate & statically serve articles to improve performance
    - Using Hadoop + MapReduce running on EC2 / S3, converted 4TB of TIFFs into 11 million PDF articles in 24 hrs
- Key Applications
  - Advertisement (Mining user behavior to generate recommendations)
  - Searches (group related documents)
  - Security (search for uncommon patterns)

Requirements at Facebook

- Design requirements:
  - Integrate display of email, SMS and chat messages between users
  - Strong control over who users
  - Stringent latency & uptime
- System requirements
  - High write throughput
  - Cheap, elastic storage
  - Low latency
  - High consistency
  - Disk-efficient sequential and random read

Hadoop Use at Facebook

- Classic alternatives
  - These requirements typically met using large MySQL cluster & caching tiers using Memcache
  - Content on HDFS could be loaded into MySQL or Memcached if needed by web tier
- Problems with previous solutions
  - MySQL has low random write throughput… BIG problem for messaging!
  - Difficult to scale MySQL clusters rapidly while maintaining performance
  - MySQL clusters have high management overhead, require more expensive hardware

Hadoop Use at Facebook

- Hadoop + HBase as foundations
  - Improve & adapt HDFS and HBase to scale to FB's workload and operational considerations
  - NameNode is Single point of failure & failover times are at least 20 minutes
- AvatarNode
  - Eliminates single point of failure makes HDFS safe to deploy even with 24/7 uptime requirement
  - Performance improvements for realtime workload: RPC timeout.
  - Rather fail fast and try a different DataNode