Cloud Computing

José Fortes
Center for Cloud and Autonomic Computing
Advanced Computing and Information Systems Lab

Kasetsart University, Bangkok, March 20, 2013

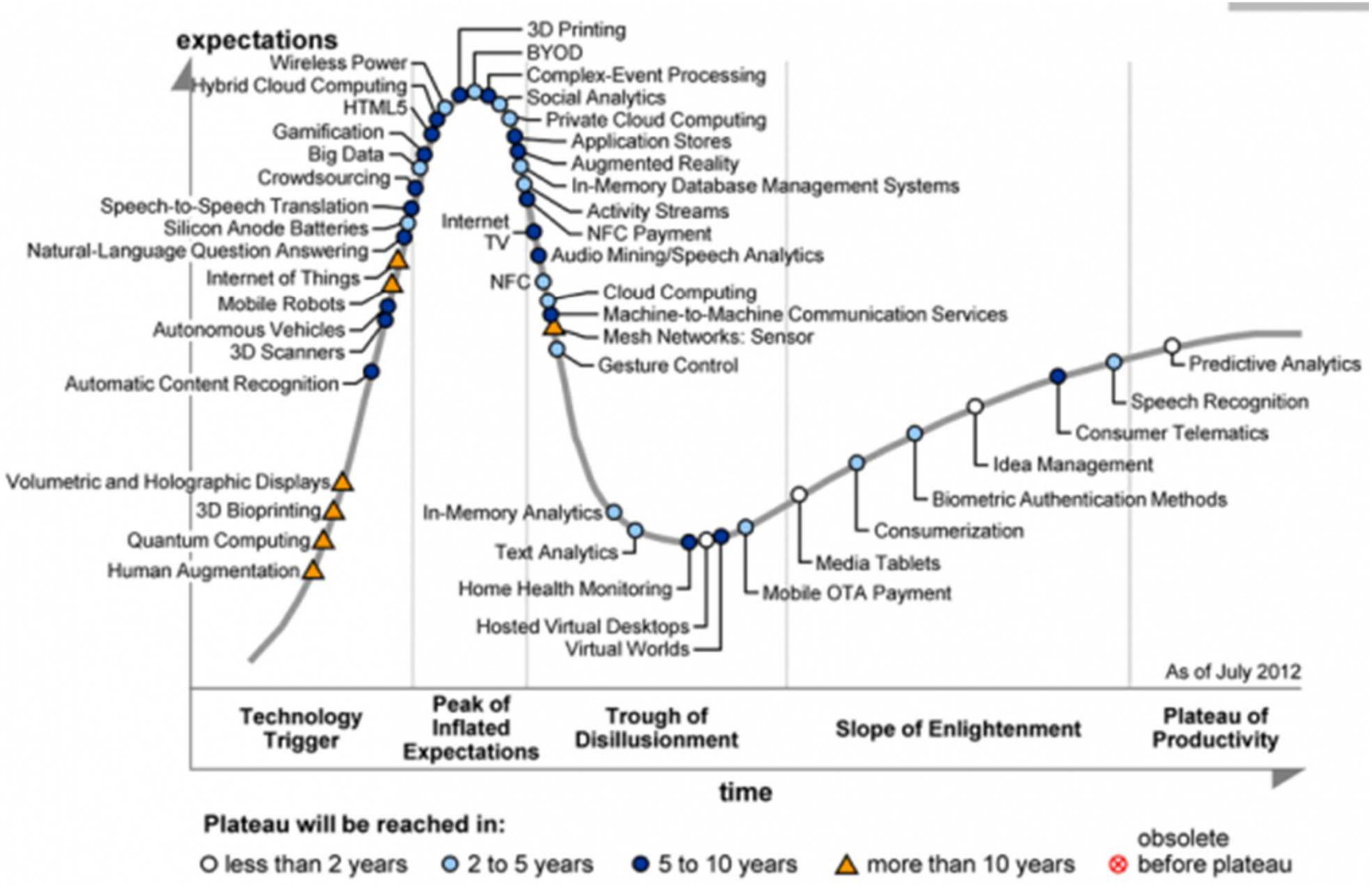
Agenda

- **09:00-10:30** Tutorial on Cloud Computing: Trends and Technology (Jose Fortes)
- **10:30-10:45** Coffee Break
- **10:45-12:15** Tutorial on Software-Defined Networking (Mauricio Tsugawa)
- **12:15-13:30** Lunch
- **13:30 -14:15** Research Example 1: Building the PRAGMA International Cloud (Cindy Zhang)
- **14:15-15:00** Research Example 2: VM Migration across International Clouds (Kohei Ishikawa)
- **15:00-15:15** Coffee Break
- **15:15-16:00** General Discussion on Research Directions
- **16:00-16:30** Questions and Discussion

Outline

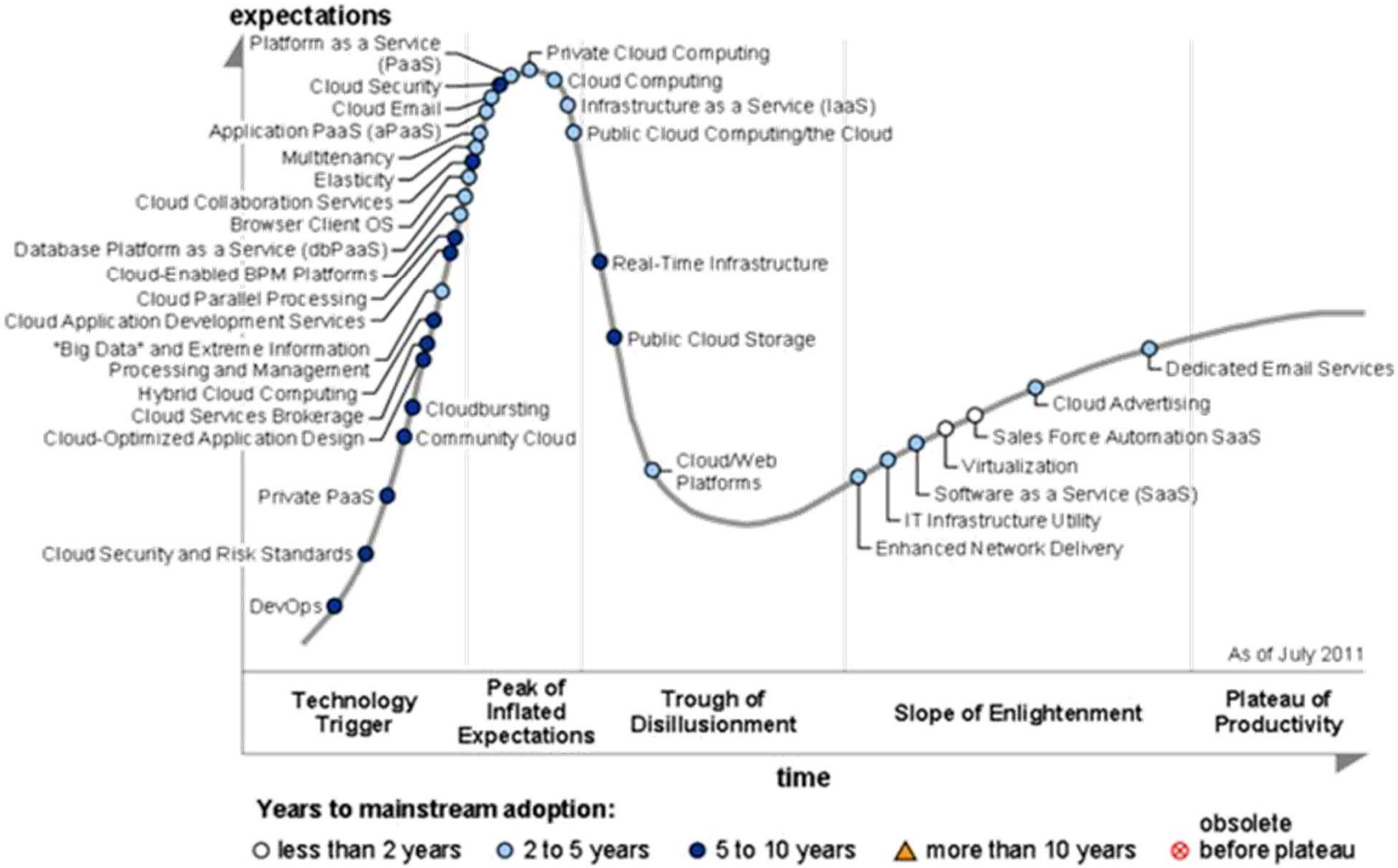
- Informal introduction to cloud computing
- A more precise introduction
- Examples
- Some technologies
- Demo
- Conclusions

Emerging technologies hype cycle (2012)



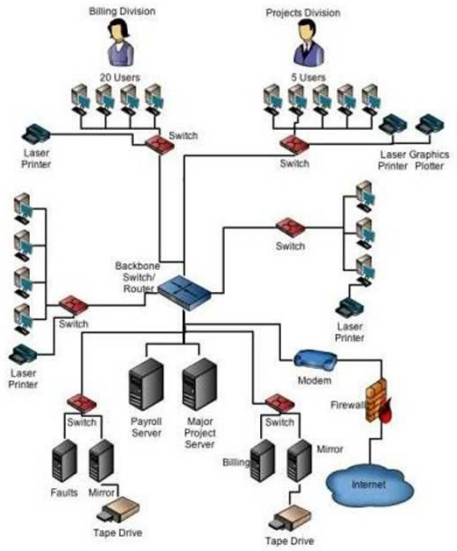
Source: Gartner 2011

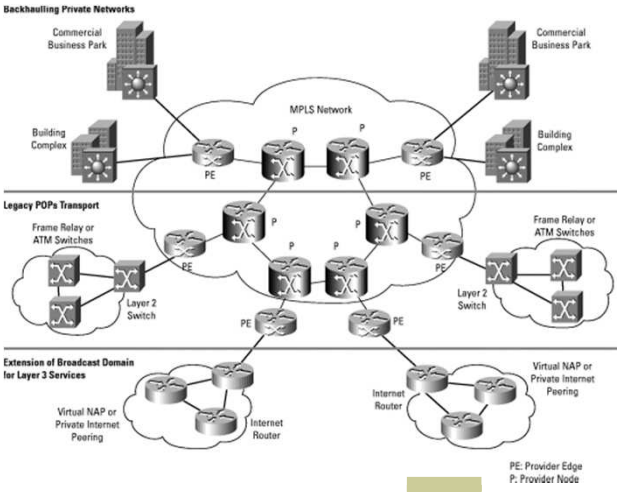
Cloud computing hype cycle (2011)

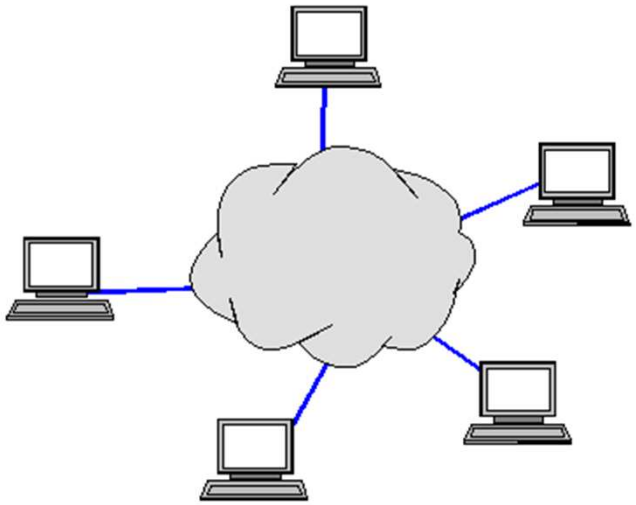


Source: Gartner 2012

“Cloud” name origins







One wonders

- What enables a cloud to do what it does?

and

- What is it that a cloud does?

Virtualization + pay-per-use

Main points

- Using clouds you can
  - Get the IT system/software/applications you need,
  - when you need it
  - for as long as you need it

and

  - pay only (very little!) for what you use

How little for a virtual computer?

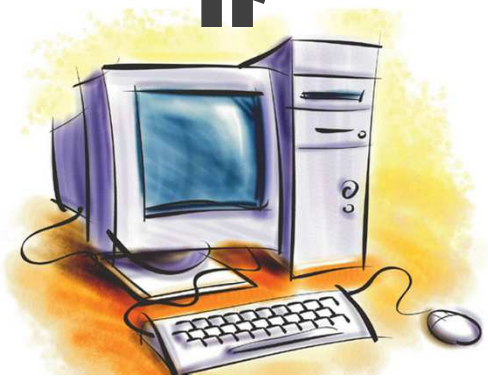
	Pricing (Small Instance)
amazon web services™	$0.085/VM-HR (Linux) $0.120/VM-HR (Windows) 1.7GB/160GB
GOGRID	$0.19/VM-HR
Azure	0.120/VM-HR
rackspace HOSTING	0.120/VM-HR (Windows) 2.048GB/80GB

How little for a Google app engine?

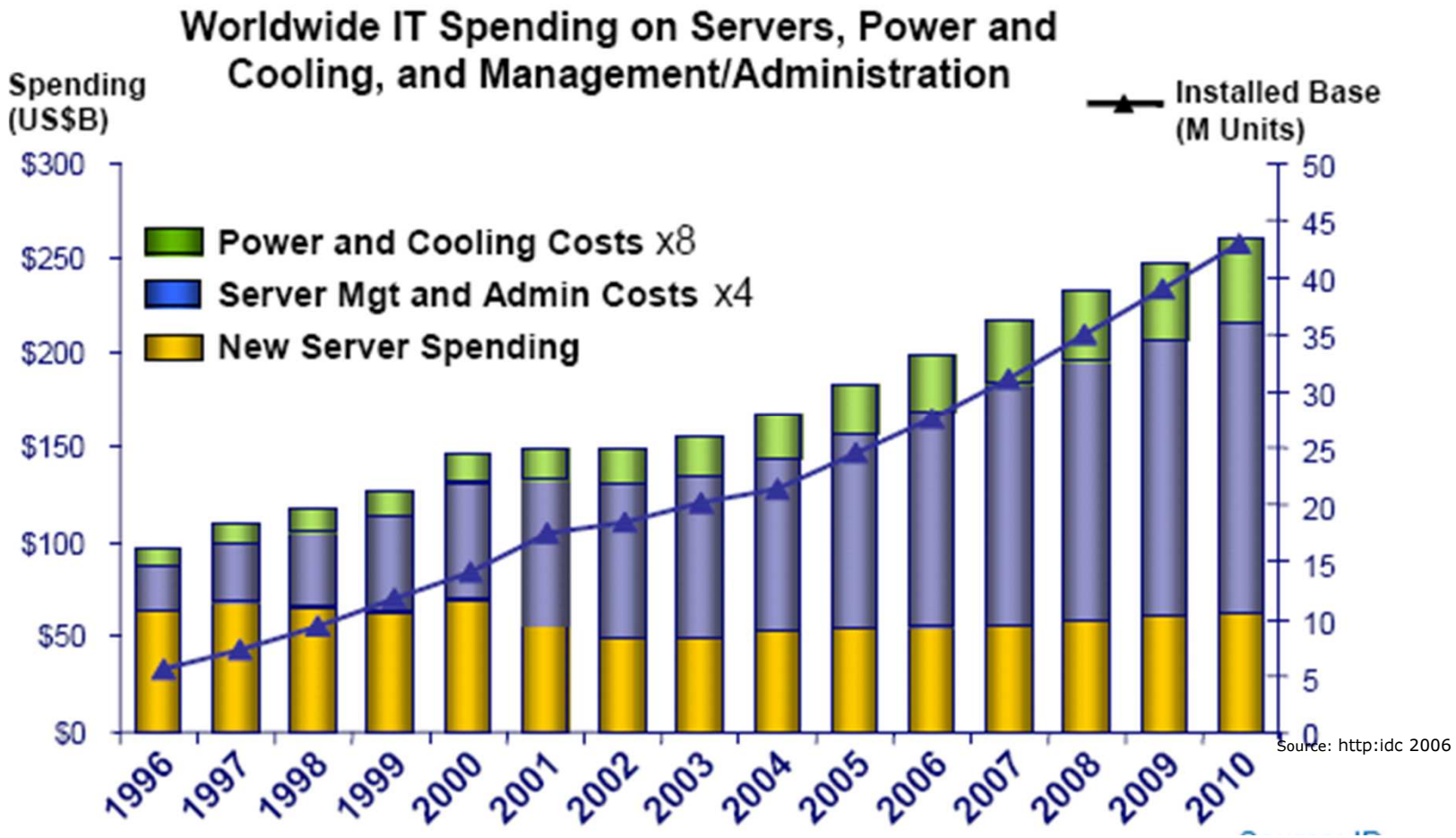
	Free quota per app per day	Pricing if you exceed your free quota
Hosting	**Free quota per app per day**	**Price**
On-demand Frontend Instances	28 free instance hours	$0.08 / hour
Reserved Frontend Instances		$0.05 / hour
High Replication Datastore	1G	$0.24 / G / month
Outgoing Bandwidth	1G	$0.12 / G
Incoming Bandwidth	1G	Free
APIs		
Datastore API	50k free read/write/small	$0.10/100k write ops $0.07/100k read ops $0.01/100k small ops
Blobstore API	5G	$0.13 / G / month

On-demand IT: Up-in-the-air decisions

Hilton



Where are the real savings?



Also,

No procurement, delivery and deployment delays

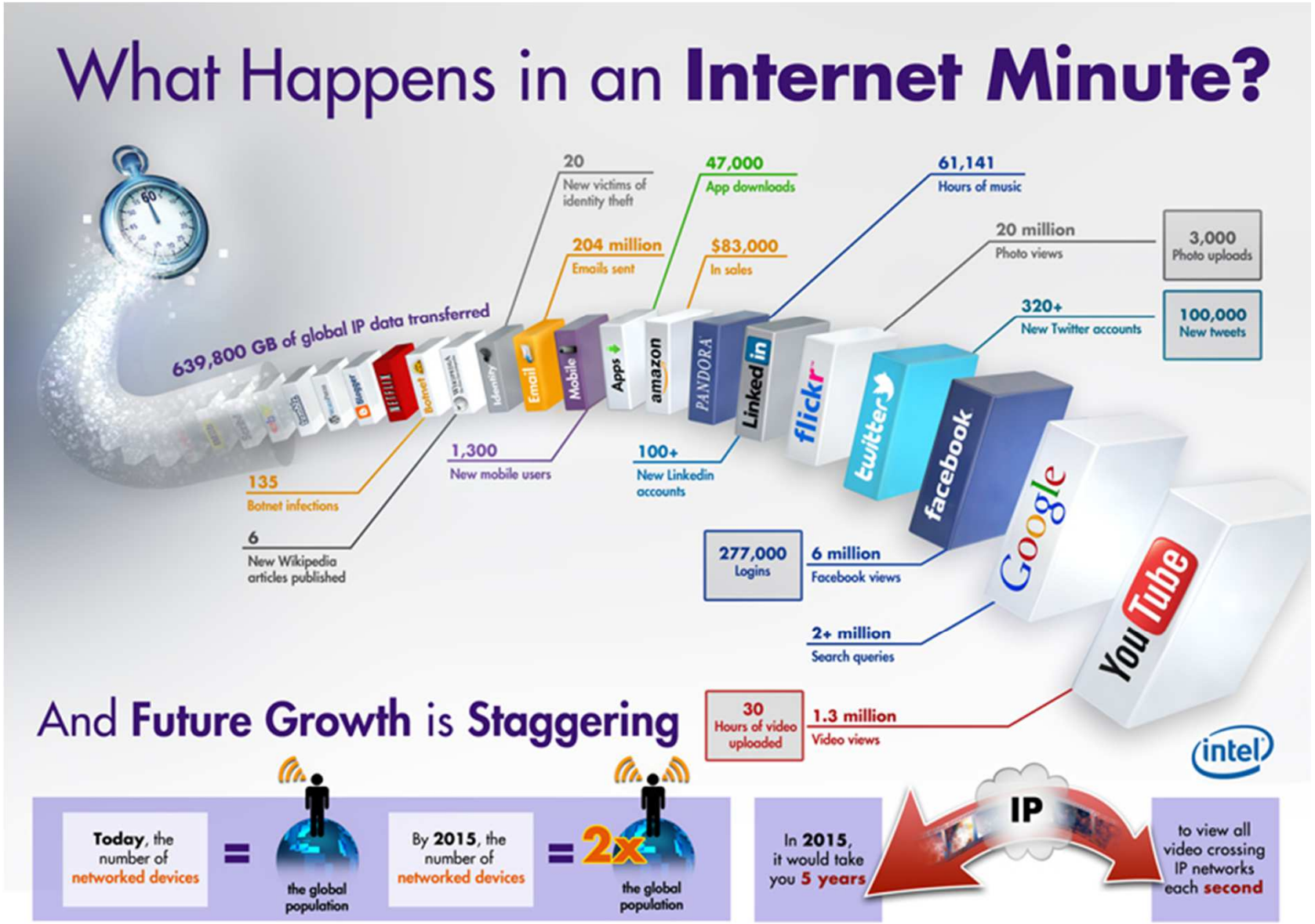
No amortization, updates, de-comissioning costs

Radu Sion's view

sion@cs.stonybrook.edu

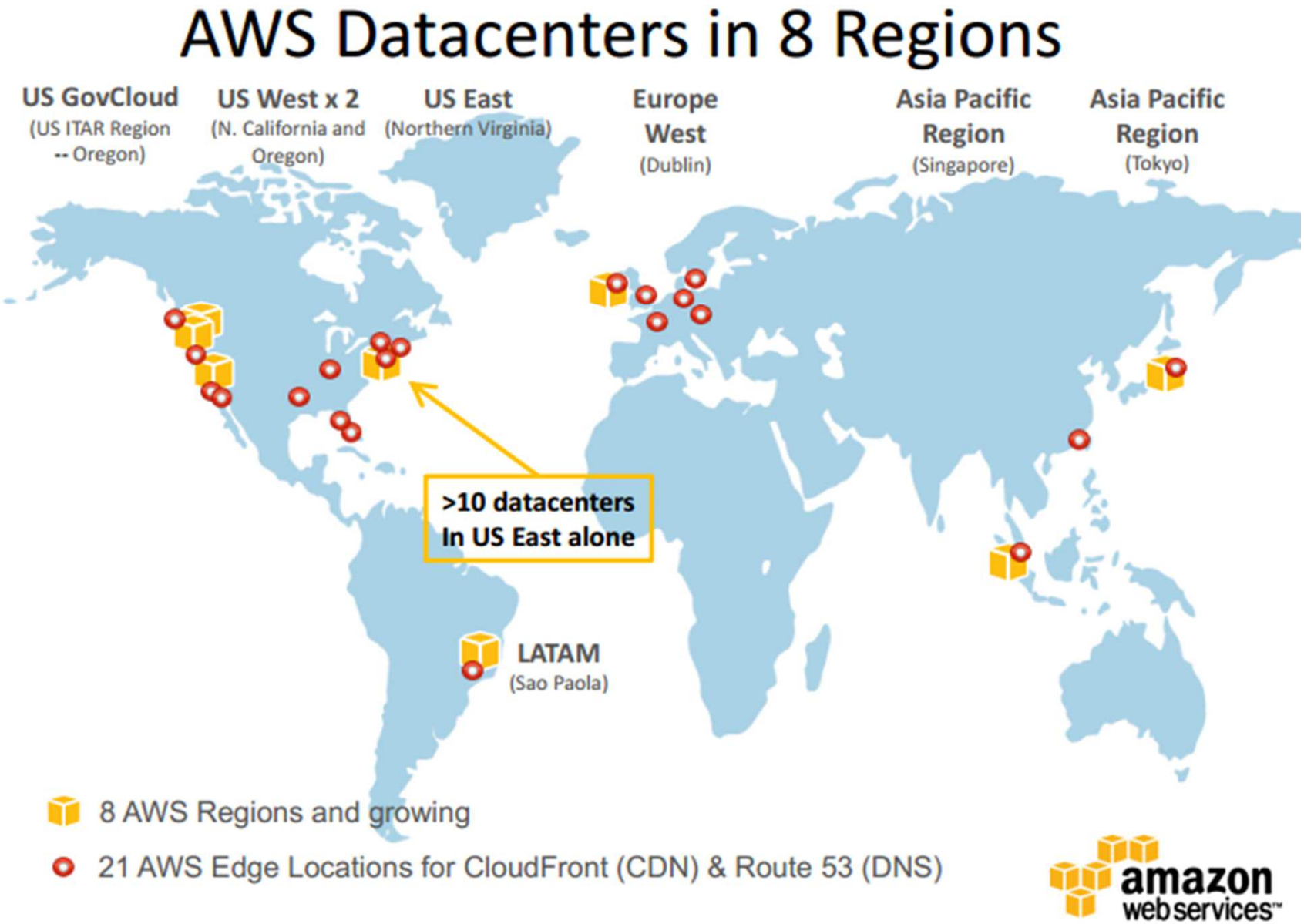
- Costs for consumer resources (in picocents)
  - CPU cycle: 6 -27; 1 bit-storage/year: 6
- Costs of cloud-provided resources
  - CPU cycle: 0.58; 1 bit-storage/year: 5.3 – 6
- 1-bit network transfer: 800-6000
  - not worth outsourcing any task of less than 4000 CPU cycles per transferred 32-bit input (single client)
- Multi-client cost per access
  - 500 if outsourced vs. 5000 otherwise
- E.g. CPU cycle: 25 Network transfer: 3000 (consumer)
  - cloud deployment saves >4500 picocents per client-to-app traffic bit + tens of picocents per CPU cycle.

Mostly content distribution/access



http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

Cloud as a CDN



NIST Cloud Computing Definition

- 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) that can be rapidly provisioned and released with minimal management effort or service provider interaction. NIST SP 800-145, SP 500-292

- a model where resources … are abstracted and provided as services on the Internet…

docs.openstack.org

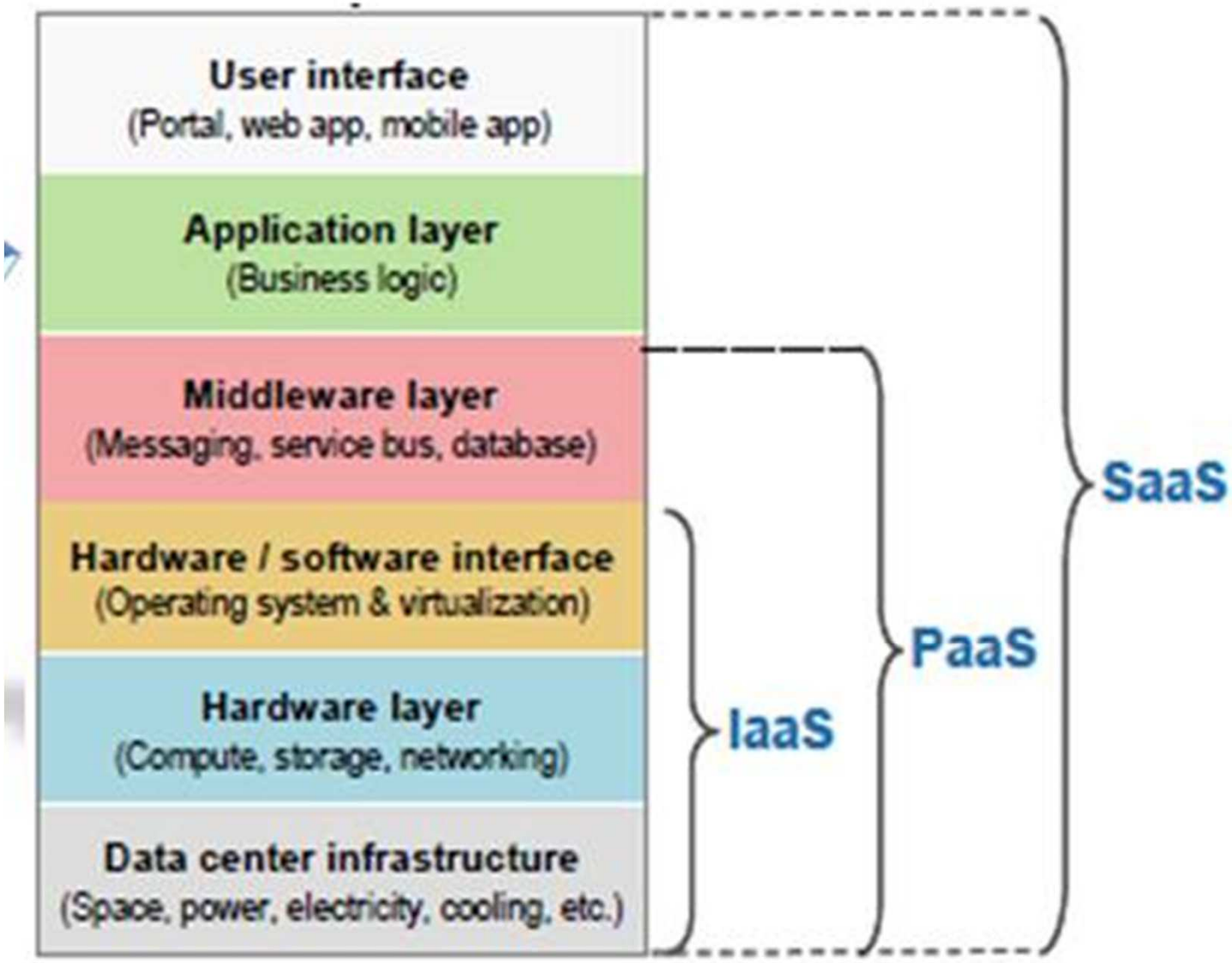
NIST's 5 essential characteristics

- On-demand self-service
  - When needed at customer's initiative
- Broad network access
  - Access through standard networks
- Resource pooling
  - Transparent sharing and location of resources
- Rapid elasticity
  - So one can use as many resources as needed
- Measured Service
  - So one can pay/be paid per use

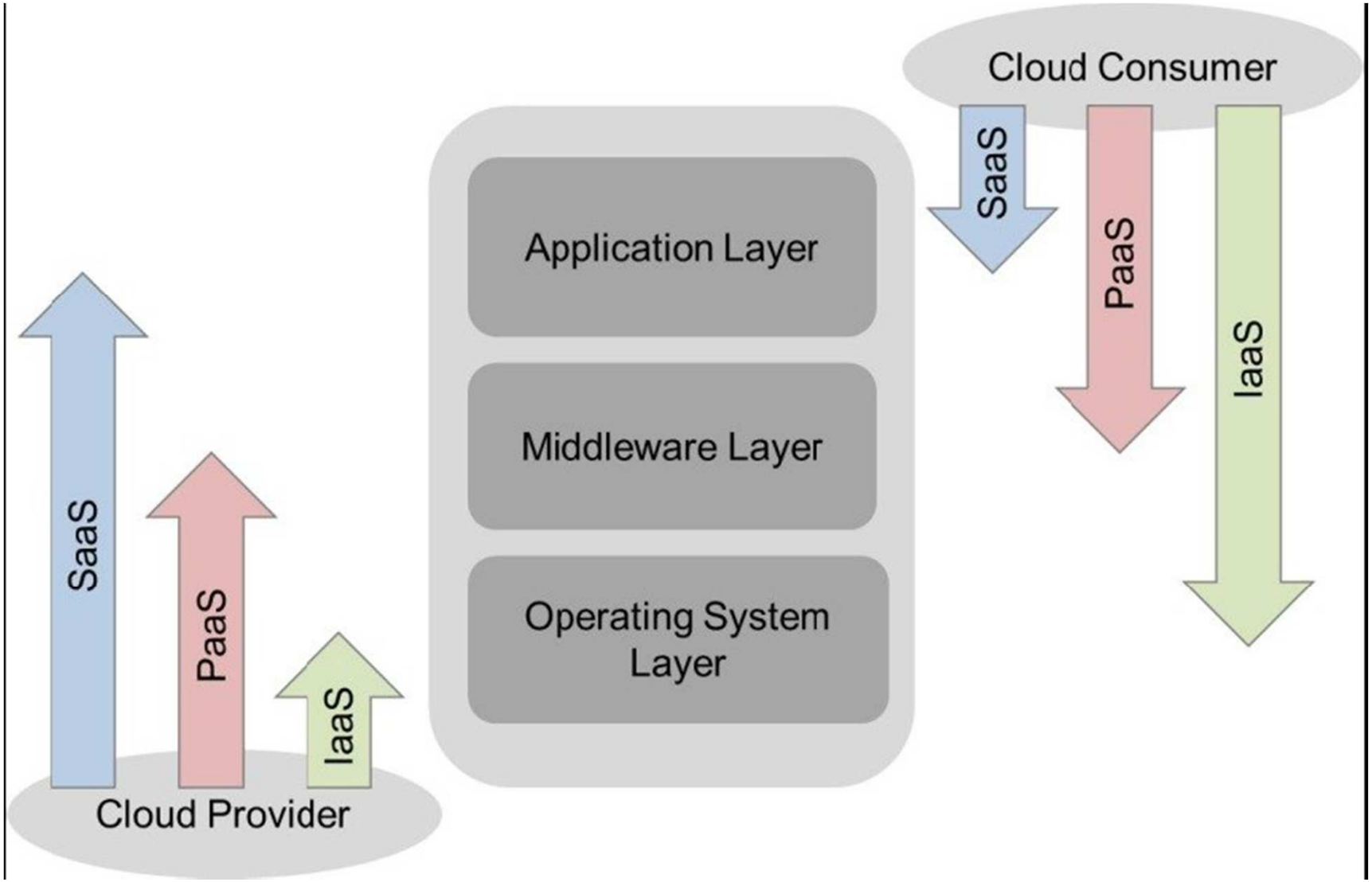
NIST Cloud Service Models

- Cloud Software as a Service (SaaS)
  consumer uses the provider's applications running on a cloud infrastructure.
- Cloud Platform as a Service (PaaS)
  consumer deploys or develops applications created using programming languages and tools supported by the provider.
- Cloud Infrastructure as a Service (IaaS)
  consumer deploys/runs arbitrary software on processing, storage, networks, and other managed computing resources

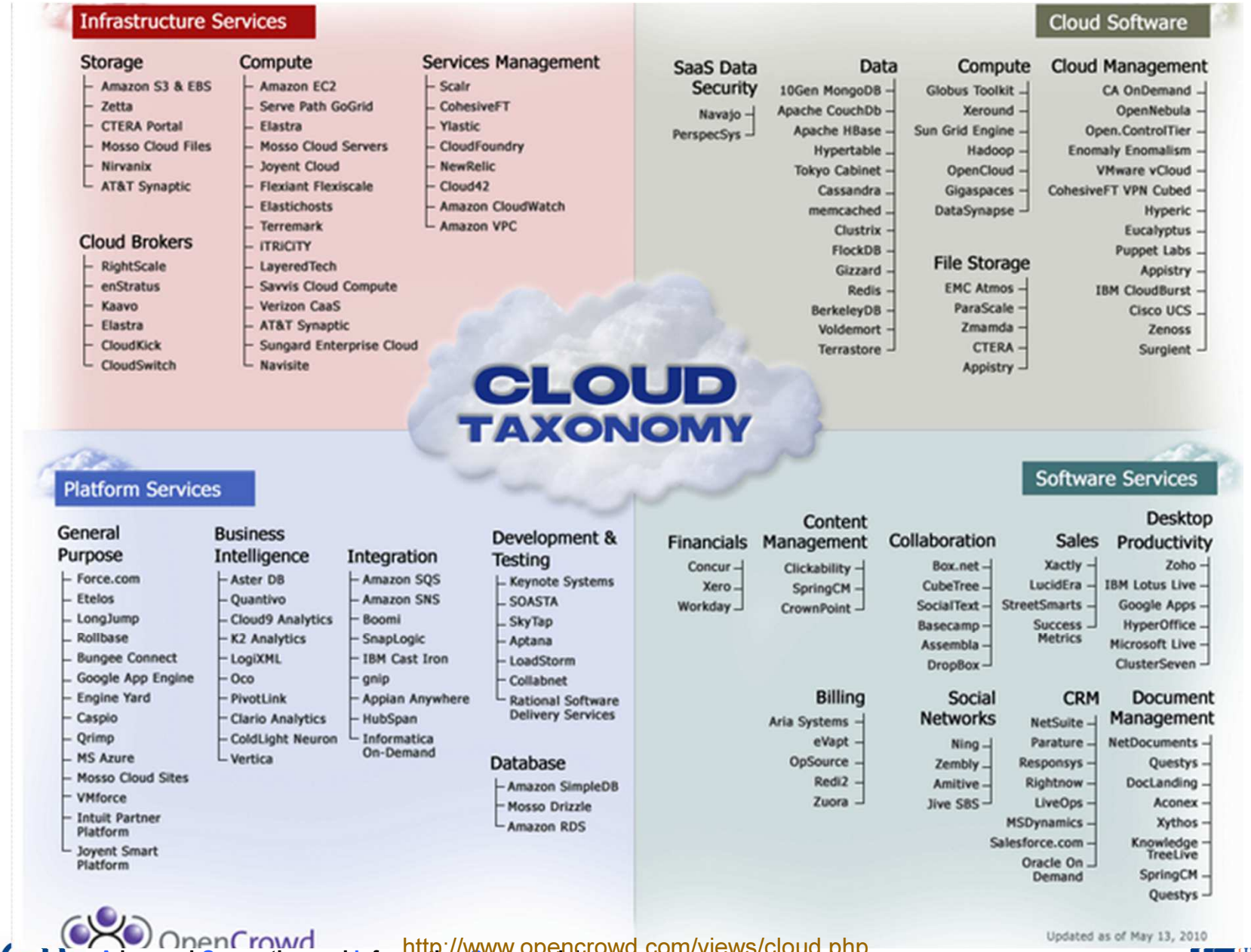
IaaS vs. PaaS vs. SaaS



Consumer vs. provider concerns



Much can be done

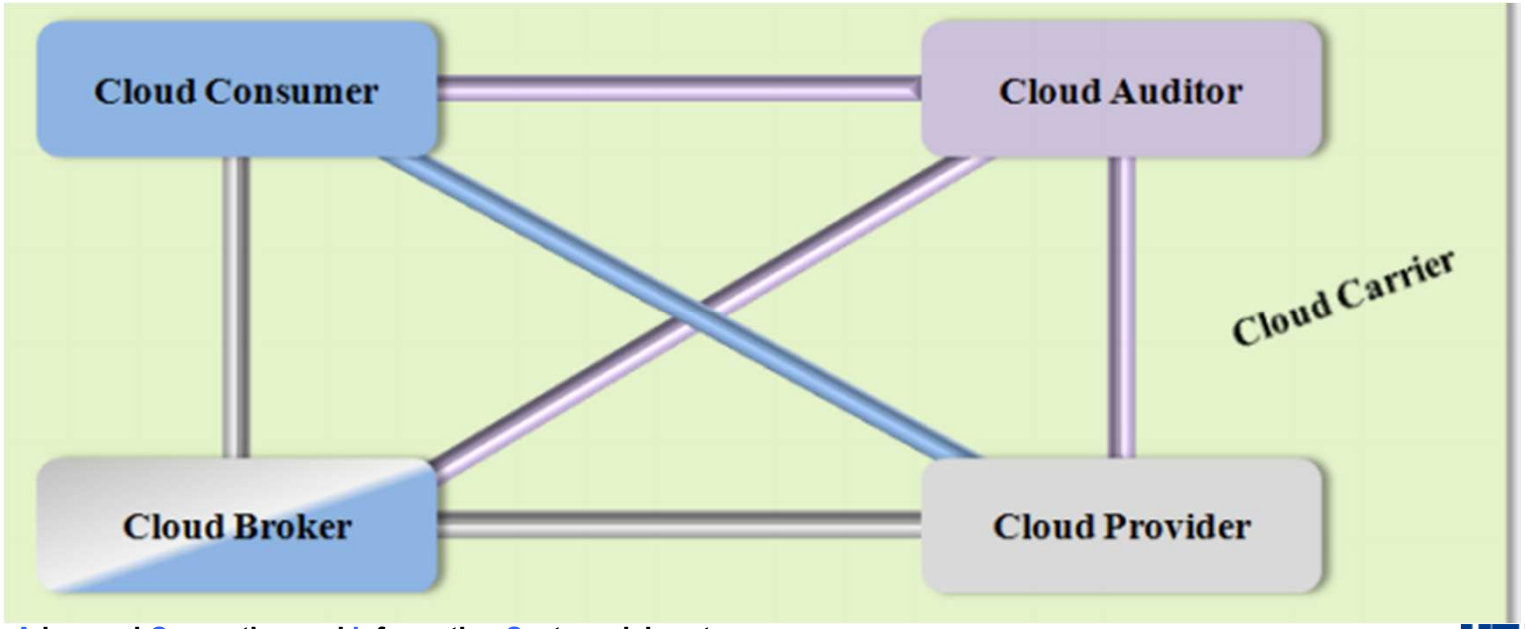


NIST Deployment Models

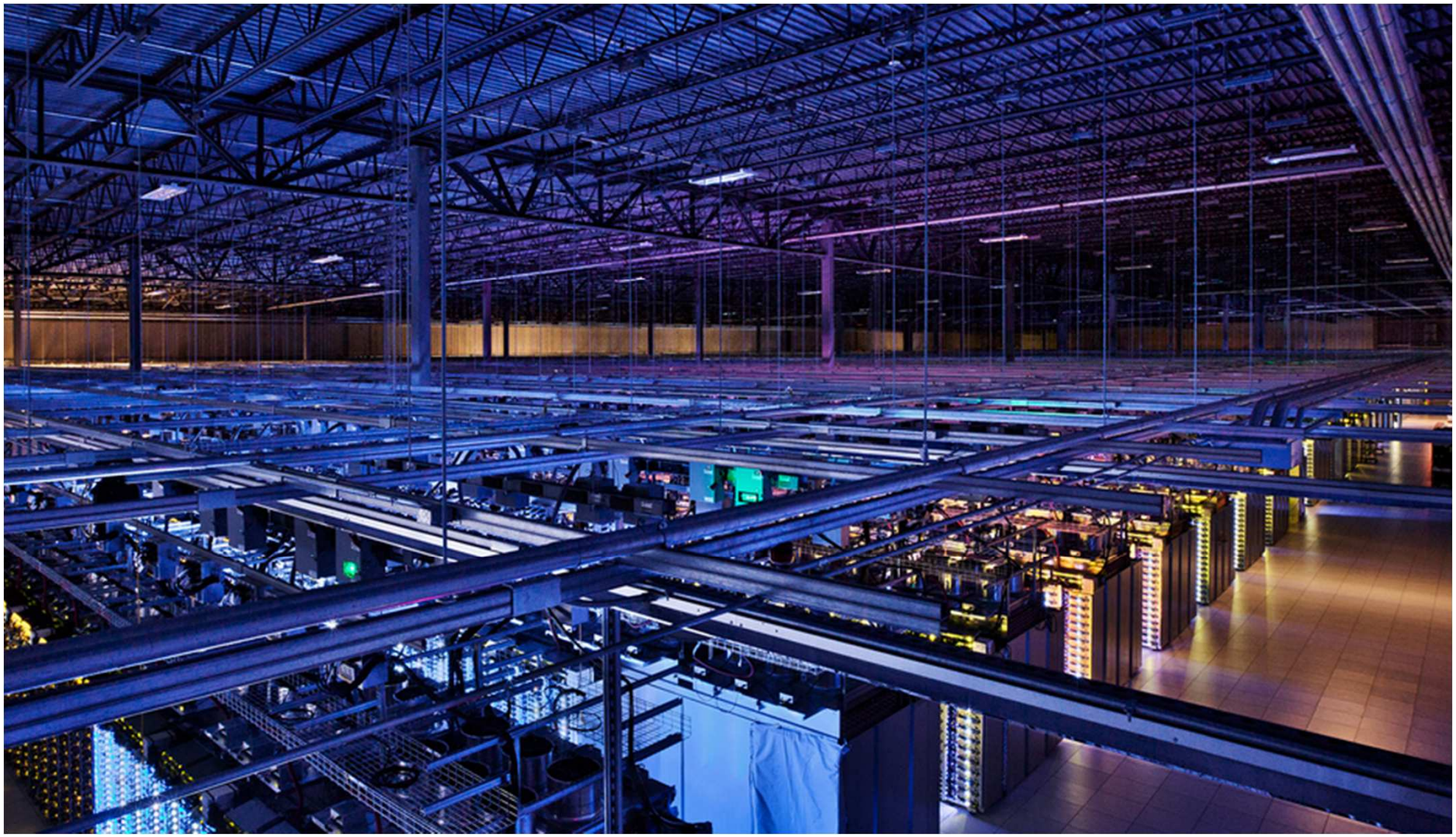
- Private cloud: operated solely for an organization (managed by the organization or a third party, on-premise or off-premise)
- Community cloud: shared by several organizations, for a specific community with shared concerns
- Public cloud: generally available from seller
- Hybrid cloud: composition of clouds (private, community, or public) so to enable data and application portability (e.g., cloud bursting for load balancing between clouds).

Cloud Actors (NIST view)

- Auditor: independent assessment services, information system operations, performance and security of the cloud
- Broker: manages the use, performance and delivery of cloud services, and negotiates relationships between Providers and Consumers.
- Carrier: intermediary that provides connectivity and transport of cloud services

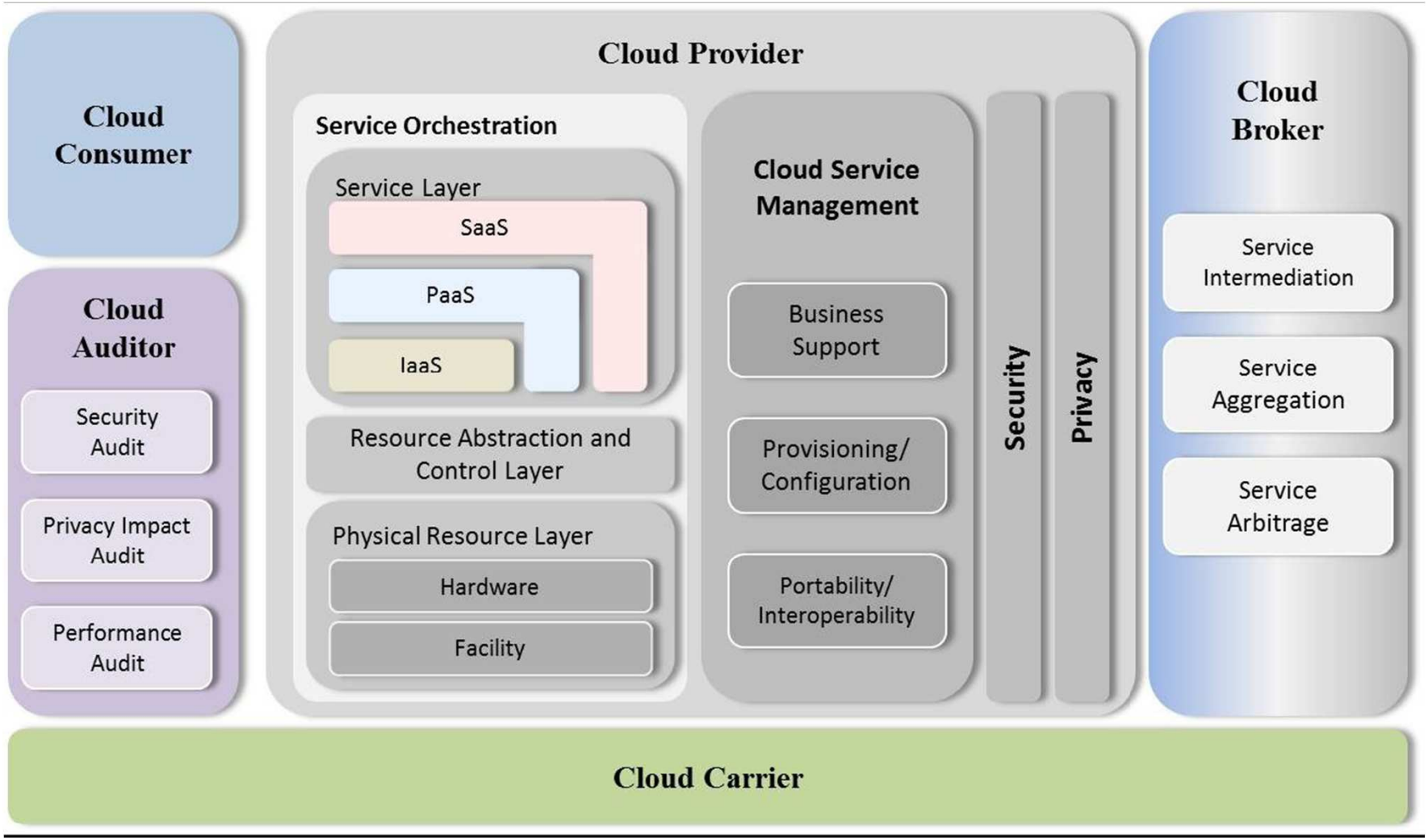


www.google.com/about/datacenters



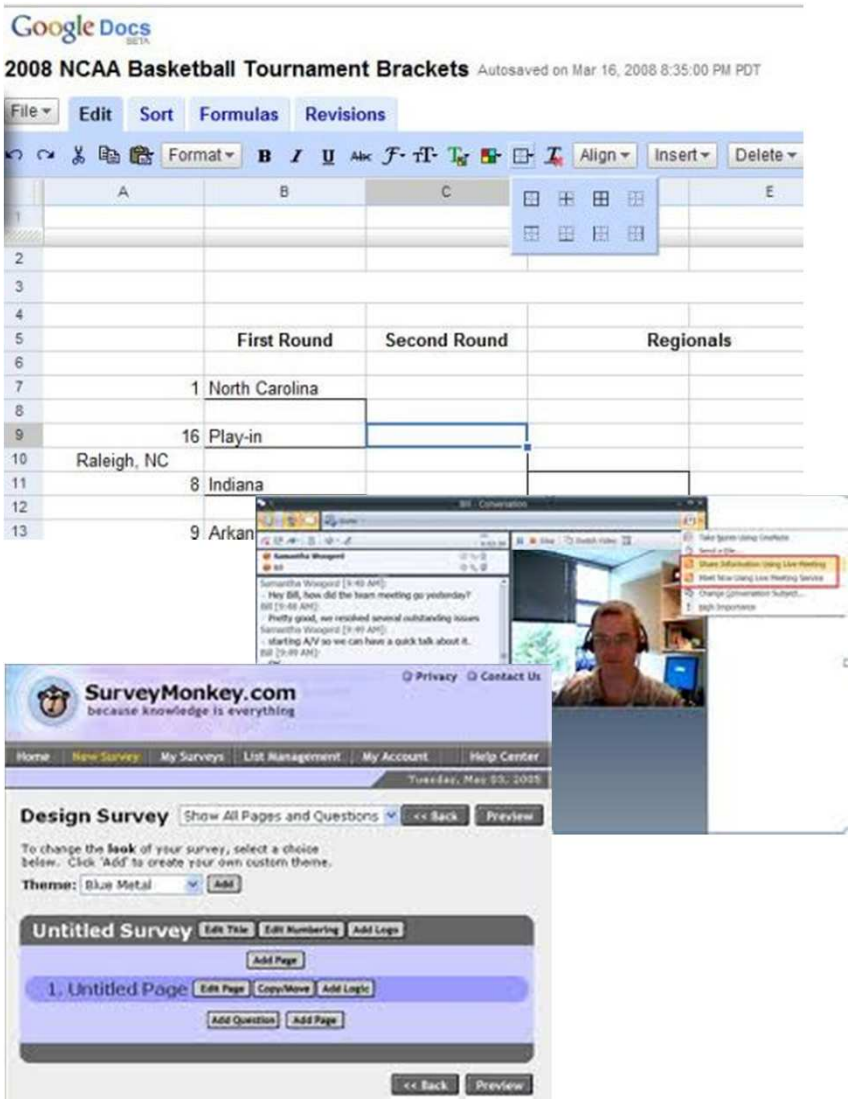
www.google.com/about/datacenters

NIST Reference Architecture

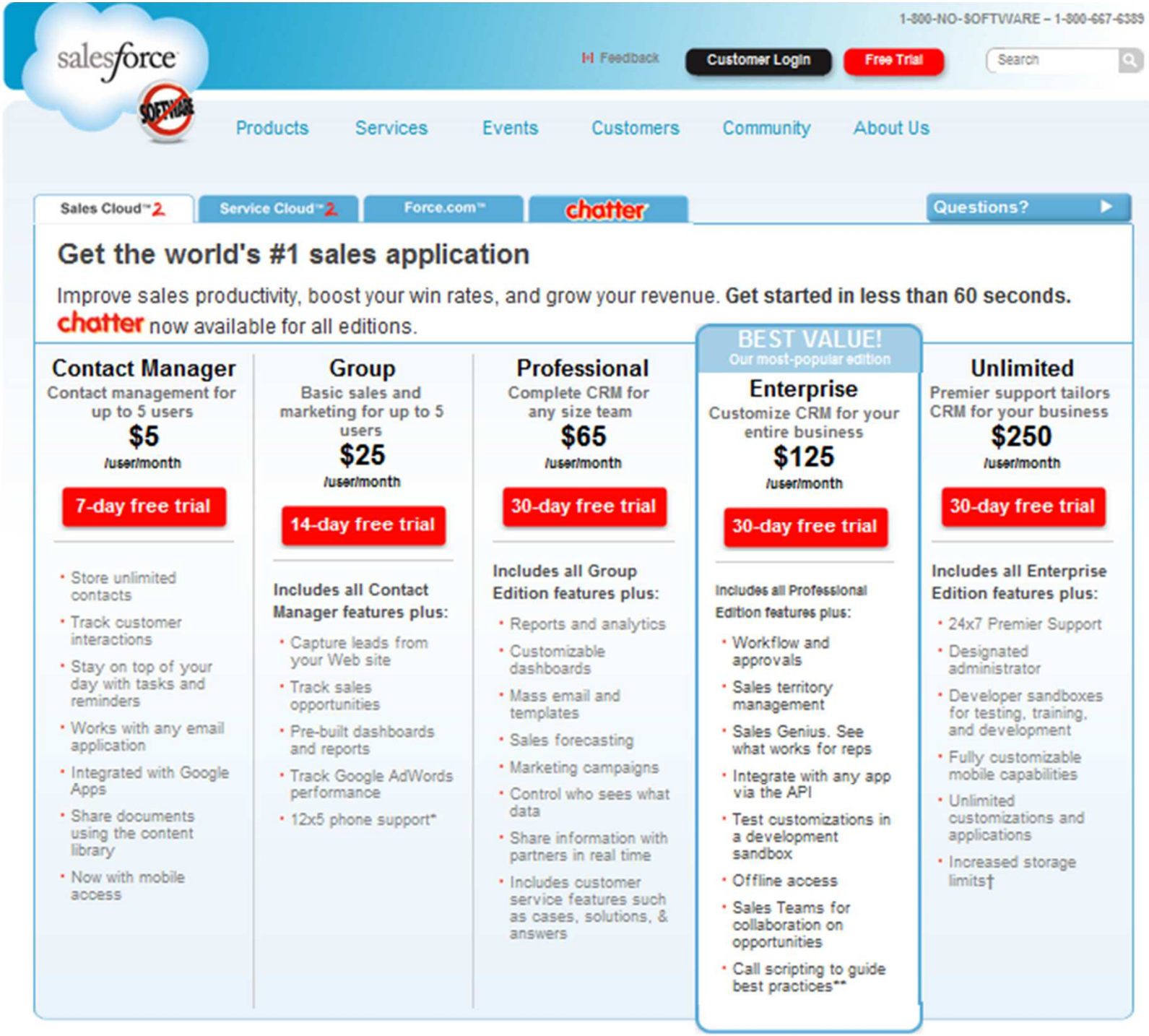


SaaS

- No software to purchase, install, update or maintain
- Everything handled by providers
  - Sign up + use cloud apps in minutes
  - Apps and data are accessible from any connected computer
  - No data is lost computer breaks
- Examples: Google's gmail, Google Docs, Dropbox, etc.
- One code base for all customers, yet configurable for their needs
  - Data aggregation, instant access to application updates and an optimized environment



For the software vendors: Reduce startup costs and faster time to market

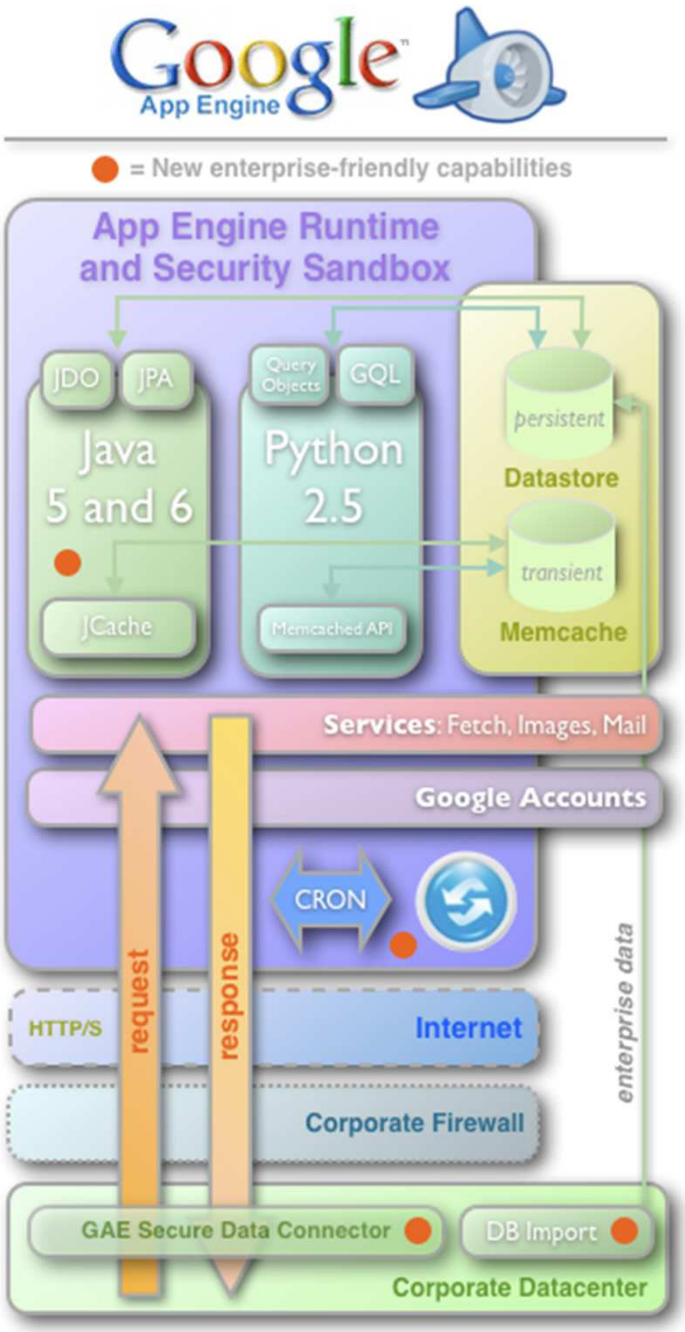

salesforce
1-800-NO-SOFTWARE – 1-800-667-6389
Feedback
Customer Login
Free Trial
Search
Products
Services
Events
Customers
Community
About Us
Sales Cloud™ 2
Service Cloud™ 2
Force.com™
chatter
Questions?
Get the world's #1 sales application
Improve sales productivity, boost your win rates, and grow your revenue. Get started in less than 60 seconds.
chatter now available for all editions.
Contact Manager
Contact management for up to 5 users
$5
/user/month
7-day free trial
• Store unlimited contacts
• Track customer interactions
• Stay on top of your day with tasks and reminders
• Works with any email application
• Integrated with Google Apps
• Share documents using the content library
• Now with mobile access
Group
Basic sales and marketing for up to 5 users
$25
/user/month
14-day free trial
Includes all Contact Manager features plus:
• Capture leads from your Web site
• Track sales opportunities
• Pre-built dashboards and reports
• Track Google AdWords performance
• 12x5 phone support*
Professional
Complete CRM for any size team
$65
/user/month
30-day free trial
Includes all Group Edition features plus:
• Reports and analytics
• Customizable dashboards
• Mass email and templates
• Sales forecasting
• Marketing campaigns
• Control who sees what data
• Share information with partners in real time
• Includes customer service features such as cases, solutions, & answers
BEST VALUE!
Our most-popular edition
Enterprise
Customize CRM for your entire business
$125
/user/month
30-day free trial
Includes all Professional Edition features plus:
• Workflow and approvals
• Sales territory management
• Sales Genius. See what works for reps
• Integrate with any app via the API
• Test customizations in a development sandbox
• Offline access
• Sales Teams for collaboration on opportunities
• Call scripting to guide best practices**
Unlimited
Premier support tailors CRM for your business
$250
/user/month
30-day free trial
Includes all Enterprise Edition features plus:
• 24x7 Premier Support
• Designated administrator
• Developer sandboxes for testing, training, and development
• Fully customizable mobile capabilities
• Unlimited customizations and applications
• Increased storage limits†

PaaS

- Like SaaS but, rather than being software delivered over the web, it is a platform for the creation of software, delivered over the web
- No cost and complexity of buying and managing the underlying HW, SW, provisioning and hosting
- The service providers offer:
  - Develop applications and get to market faster
  - Deploy new web applications to the cloud in minutes
  - Reduce complexity with middleware as a service
- Examples: Google App Engine, Windows Azure Platform, Amazon Web Services (Elastic Beanstalk), Vmware Cloud Foundry, IBM SmartCloud Application Services, etc.

Google App Engine

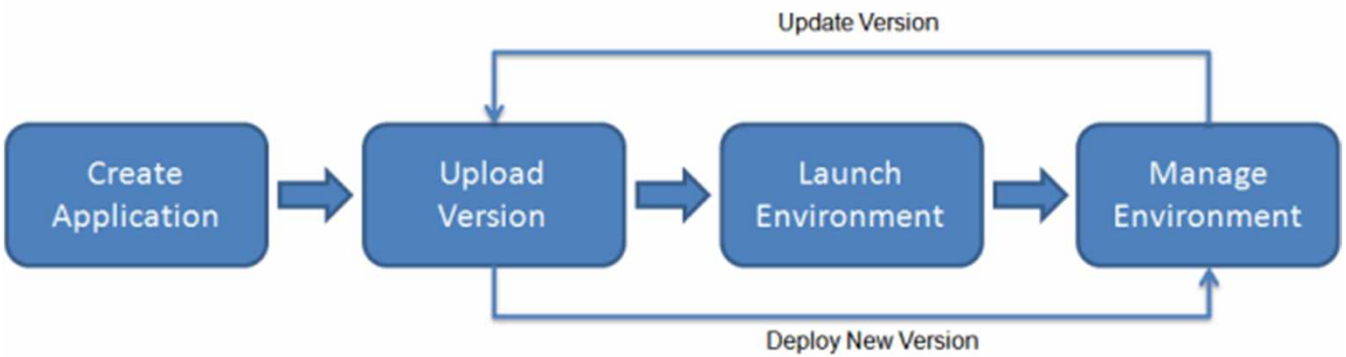
- Support apps written in several programming languages (e.g., Java, Python, Ruby, Go)
- No set-up costs and recurring fees
- Offer limited secure sandbox environment
- Provide a range of options for storing your data
  - App Engine Datastore
  - Google Cloud SQL
  - Google Cloud Storage



From http://blogs.zdnet.com/Hinchcliffe

Amazon Elastic Beanstalk

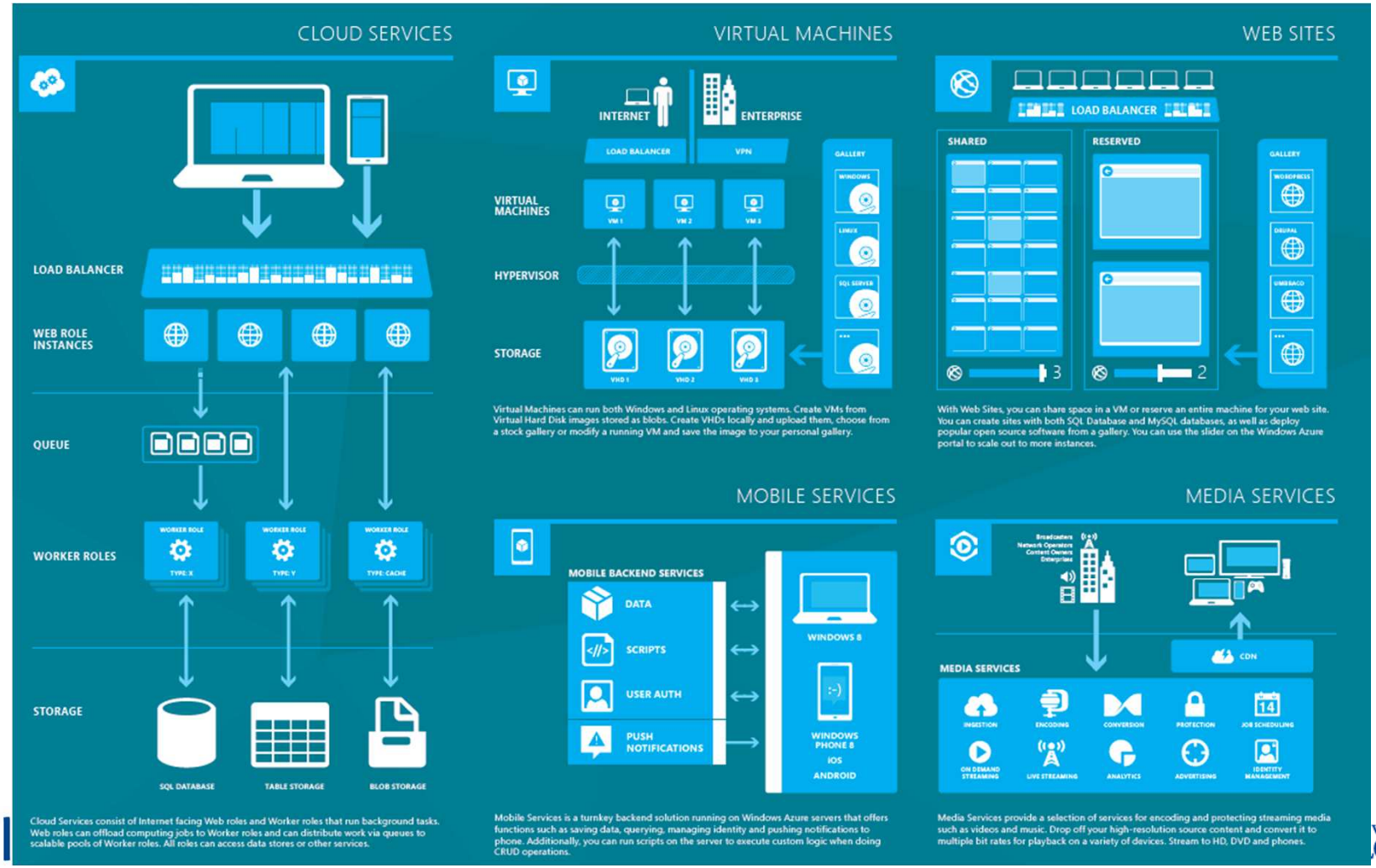
- Quickly deploy and manage applications with the AWS cloud
- Automatically handle the details of capacity provisioning, load balancing, scaling and application health monitoring
- Workflow



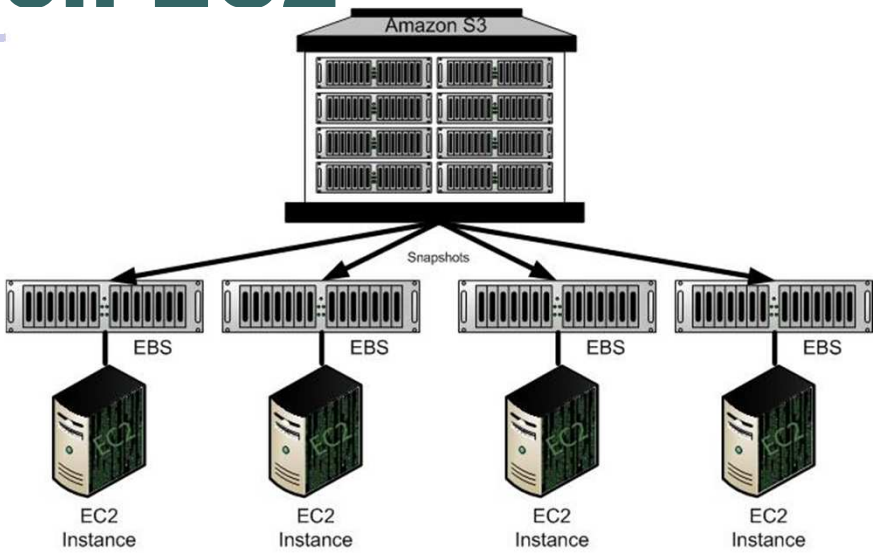
- Supported platforms: Java, PHP, Windows .NET, Python and Ruby

Windows Azure

- enables you to quickly build, deploy, and manage applications across Microsoft-managed datacenters
  - E.g. Web sites: Build with ASP.NET, PHP or Node.js and deploy in seconds with FTP, Git or Team Foundation Server (TFS)



IaaS example: Amazon EC2



- Instance types
  - **On-demand**: pay for compute capacity by hour with no long-term commitment
  - **Reserved**: make a low one-time payment for each instance to reserve and in turn receive discount on the hourly charge for that instance.
  - **Spot**: bid on unused Amazon EC2 capacity and run those instances for as long as their bid exceeds the current spot price
- Amazon Elastic Block Store (EBS) and Amazon Simple Storage Service (S3)
  - Offer persistent storage and independent from the life of an instance

Amazon Web Services (AWS)

Compute

- **Amazon Elastic Compute Cloud** (EC2)
  Amazon Elastic Compute Cloud delivers scalable, pay-as-you-go compute capacity in the cloud.

- **Amazon Elastic MapReduce**
  Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

- **Auto Scaling**
  Auto Scaling allows you to automatically scale your Amazon EC2 capacity up or down according to conditions you define.

- **Elastic Load Balancing**
  Elastic Load Balancing automatically distributes incoming application traffic across multiple Amazon EC2 instances.

Content Delivery

- **Amazon CloudFront**
  Amazon CloudFront is a web service that makes it easy to distribute content with low latency via a global network of edge locations.

Software

- **AWS Marketplace**
  AWS Marketplace is an online store that helps customers find, buy, and immediately start using software that runs on the AWS cloud. It includes software from trusted vendors like SAP, Zend, Microsoft, IBM, Canonical, and 10gen as well as many widely used open source offerings including Wordpress, Drupal, and MediaWiki.

Networking

- **Amazon Route 53**
  Amazon Route 53 is a highly available and scalable Domain Name System (DNS) web service.

- **Amazon Virtual Private Cloud** (VPC)
  Amazon Virtual Private Cloud (Amazon VPC) lets you provision a private, isolated section of the Amazon Web Services (AWS) Cloud where you can launch AWS resources in a virtual network that you define. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter.

- **AWS Direct Connect**
  AWS Direct Connect makes it easy to establish a dedicated network connection from your premise to AWS, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

http://aws.amazon.com/products/

Amazon Web Services (AWS)(2)

Database

- **Amazon Relational Database Service** (RDS)
  Amazon Relational Database Service is a web service that makes it easy to set up, operate, and scale a relational database in the cloud.
- **Amazon DynamoDB**
  Amazon DynamoDB is a fully-managed, high performance, NoSQL database service that is easy to set up, operate, and scale.
- **Amazon ElastiCache**
  Amazon ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud.
- **Amazon Redshift**
  Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Amazon Redshift offers you fast query performance when analyzing virtually any size data set using the same SQL-based tools and business intelligence applications you use today.

Deployment & Management

- **AWS Identity and Access Management (IAM)**
  IAM enables you to securely control access to AWS services and resources for your users. IAM enables you to create and manage users in AWS, and enables you to grant access to AWS resources for users managed outside of AWS in your corporate directory.
- **Amazon CloudWatch**
  Amazon CloudWatch is a web service that provides monitoring for AWS cloud resources, starting with Amazon EC2.
- **AWS Elastic Beanstalk**
  AWS Elastic Beanstalk is an even easier way to quickly deploy and manage applications in the AWS cloud. You simply upload your application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.

Payments & Billing

- **Amazon Flexible Payments Service** (FPS)
  Amazon Flexible Payments Service facilitates the digital transfer of money between any two entities, humans or computers.
- **Amazon DevPay**
  Amazon DevPay is a billing and account management service which enables developers to collect payment for their AWS applications.

Storage

- **Amazon Simple Storage Service** (S3)
  Amazon Simple Storage Service provides a fully redundant data storage infrastructure for storing and retrieving any amount of data, at any time, from anywhere on the Web.
- **Amazon Glacier**
  Amazon Glacier is an extremely low-cost storage service that provides secure and durable storage for data archiving and backup.
- **Amazon Elastic Block Store** (EBS)
  Amazon Elastic Block Store provides block level storage volumes for use with Amazon EC2 instances. Amazon EBS volumes are off-instance storage that persists independently from the life of an instance.
- **AWS Import/Export**
  AWS Import/Export accelerates moving large amounts of data into and out of AWS using portable storage devices for transport.
- **AWS Storage Gateway**
  AWS Storage Gateway is a service connecting an on-premises software appliance with cloud-based storage to provide seamless and secure integration between an organization's on-premises IT environment and AWS's storage infrastructure.

Amazon Web Services (AWS)(3)

- **AWS CloudFormation**
  AWS CloudFormation is a service that gives developers and businesses an easy way to create a collection of related AWS resources and provision them in an orderly and predictable fashion.

- **AWS Data Pipeline**
  AWS Data Pipeline is a service that helps you reliably process and move data between different AWS compute and storage services as well as on-premise data sources at specified intervals.

- **AWS OpsWorks**
  AWS OpsWorks is a DevOps platform for managing applications of any scale or complexity on the AWS cloud.

Application Services

- **Amazon CloudSearch**
  Amazon CloudSearch is a fully-managed search service in the cloud that allows customers to easily integrate fast and highly scalable search functionality into their applications.

- **Amazon Simple Workflow Service** (SWF)
  Amazon Simple Workflow Service (Amazon SWF) helps you coordinate the processing steps in your applications and manage distributed execution state.

- **Amazon Simple Queue Service** (SQS)
  Amazon Simple Queue Service provides a hosted queue for storing messages as they travel between computers, making it easy to build automated workflow between Web services.

- **Amazon Simple Notification Service** (SNS)
  Amazon Simple Notification Service is a web service that makes it easy to set up, operate, and send notifications from the cloud.

- **Amazon Simple Email Service** (SES)
  Amazon Simple Email Service is a highly scalable and cost-effective bulk and transactional email-sending service for the cloud.

- **Amazon Elastic Transcoder**
  Amazon Elastic Transcoder is a fully managed service that makes it easy to convert media files in the cloud with scalability and at a low cost.

Support

- **AWS Support**
  AWS Support is a one-on-one, fast-response support channel to help you build and run applications on AWS Infrastructure Services.

Web Traffic

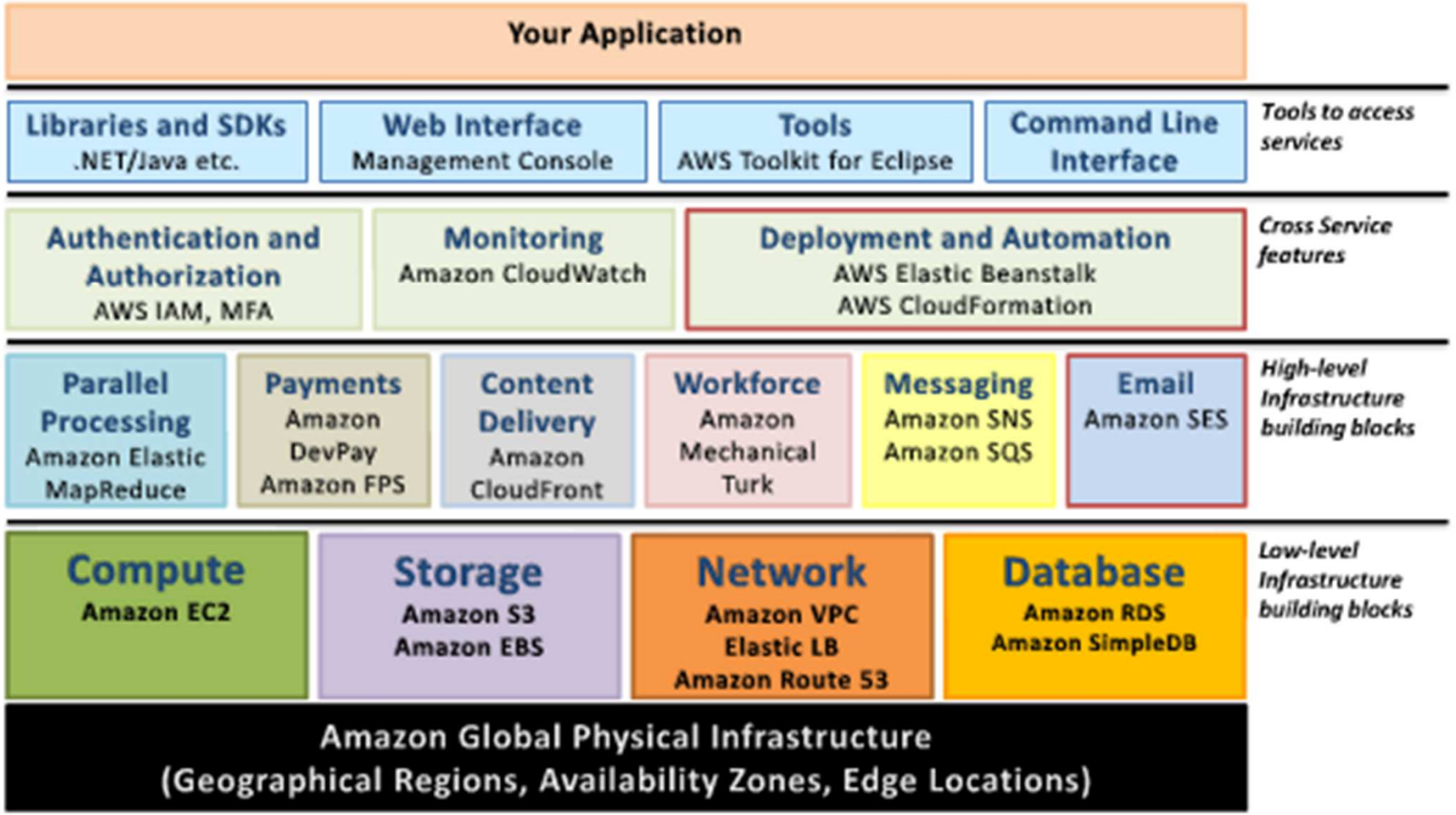
- **Alexa Web Information Service**
  Alexa Web Information Service makes Alexa's huge repository of data about structure and traffic patterns on the Web available to developers.

- **Alexa Top Sites**
  Alexa Top Sites exposes global website traffic data as it is continuously collected and updated by Alexa Traffic Rank.

Workforce

- **Amazon Mechanical Turk**
  Amazon Mechanical Turk enables companies to access thousands of global workers on demand and programmatically integrate their work into various business processes.

AWS: How they fit together



Storage

- Simple Storage Service (S3)
- Elastic Block Store (EBS)
- AWS Import/Export
  - Send Disk to Amazon
- AWS Storage Gateway

Storage: Simple Storage Service (S3)

- "Objects Storage for the Internet"
- Write, read, and delete unlimited number of objects, containing from 1 byte to 5 terabytes of data each
- Each object stored in a bucket and retrieved via a unique, developer-assigned key
  - E.g: An object named **photos/puppy.jpg** and stored in the **johnsmith** bucket, is addressable using the URL http://johnsmith.s3.amazonaws.com/photos/puppy.jpg
- Public or Private objects, access control
- Relaxed Eventual Consistency Model
  - If you PUT to an existing key, a subsequent read might return the old data or the updated data, but it will never write corrupted or partial data.
  - -C +A +P (CAP model)
- Simple Interfaces
  - REST
    - HTTP PUT, GET
  - SOAP
  - Bittorrent
- Pricing a combination of:
  - Per Storage GB, Per # of requests, Per Transfer GB

Region: US Standard

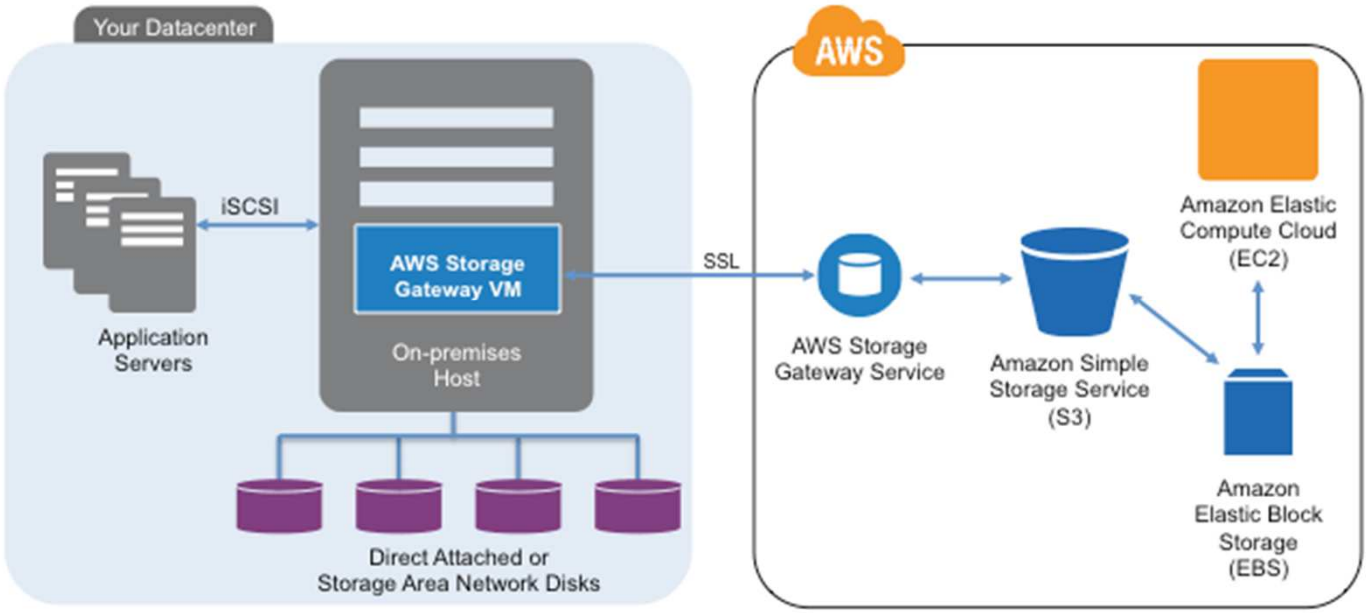
	Standard Storage	Reduced Redundancy Storage
First 1 TB / month	$0.140 per GB	$0.093 per GB
Next 49 TB / month	$0.125 per GB	$0.083 per GB
Next 450 TB / month	$0.110 per GB	$0.073 per GB
Next 500 TB / month	$0.095 per GB	$0.063 per GB
Next 4000 TB / month	$0.080 per GB	$0.053 per GB
Over 5000 TB / month	$0.055 per GB	$0.037 per GB

Storage: Elastic Block Store (EBS)

- “Cloud-based virtual hard drives”
- Block level storage volumes for use with Amazon EC2 instances
- Off-instance storage, persists independently from the life of an instance
- Can be attached to a running Amazon EC2 instance and exposed as a device within the instance
  - 1 GB to 1 TB
- Amazon CloudWatch exposes performance metrics for EBS volumes, giving insight into bandwidth, throughput, latency, …
- EBS can be (incrementally) backed up on S3
- Higher throughput than Amazon EC2 instance stores for applications performing a lot of random accesses
- Can attach multiple volumes to an instance and stripe across the volumes (RAID0) to achieve further increases in throughput.

Storage: AWS Storage Gateway

- Service for hybrid cloud storage
- Provides for “cloud-bursting”
- Designed for Enterprise storage and backup
- Use a gateway VM to connect to the cloud



Elastic Compute Cloud: EC2

- VMs on-demand (per-hour) or Reserved (Annual + discounted pay-per-hour
- Managed via

CLI API or

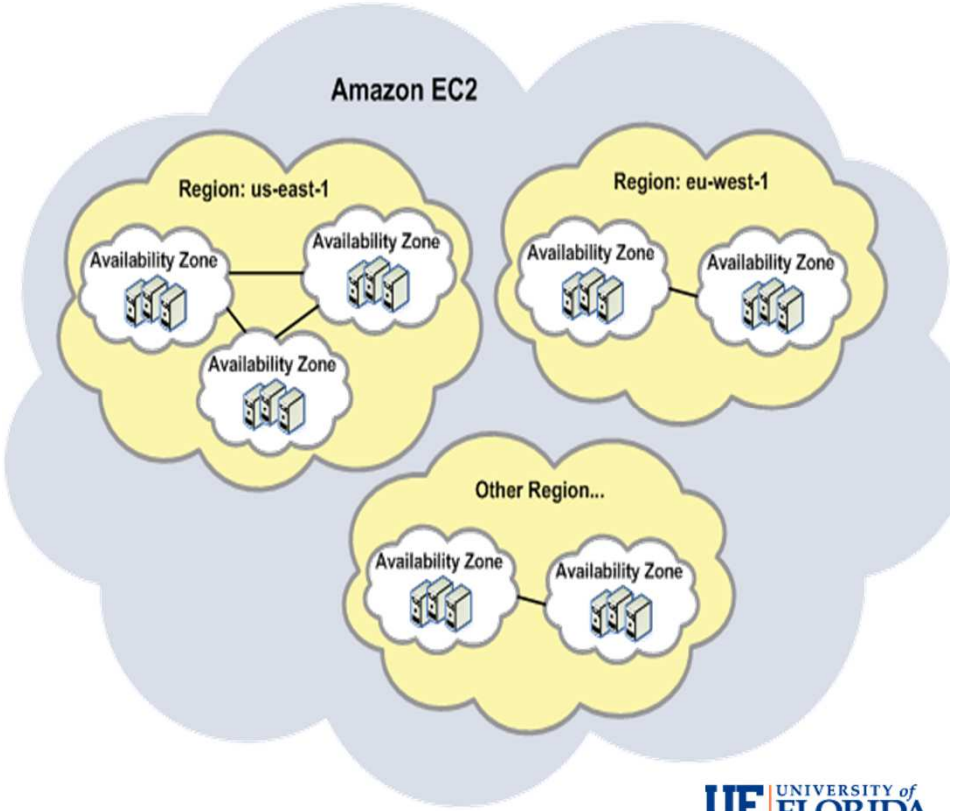
web tools

Region: US East (Virginia)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.12 per hour
Large	$0.34 per hour	$0.48 per hour
Extra Large	$0.68 per hour	$0.96 per hour
Micro On-Demand Instances		
Micro	$0.02 per hour	$0.03 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.62 per hour
Double Extra Large	$1.00 per hour	$1.24 per hour
Quadruple Extra Large	$2.00 per hour	$2.48 per hour
Hi-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.29 per hour
Extra Large	$0.68 per hour	$1.16 per hour
Cluster Compute Instances		
Quadruple Extra Large	$1.30 per hour	$1.61 per hour
Eight Extra Large	$2.40 per hour	$2.97 per hour
Cluster GPU Instances		
Quadruple Extra Large	$2.10 per hour	$2.60 per hour

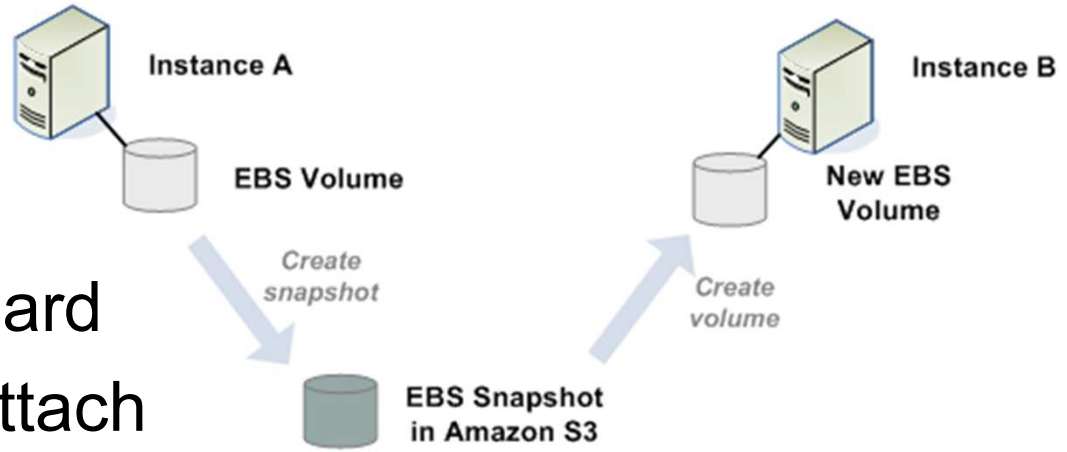
Elastic Compute Cloud: EC2

- Regions
  - Data centers in different parts of the world
  - Each has locations called Availability Zones which are isolated from failures in others and have Inexpensive, low-latency network connectivity to other zones in the same Region
  - Launching instances in separate Zones protects applications from failure in a single location
  - Applications closer to specific customers or to meet legal or other requirements



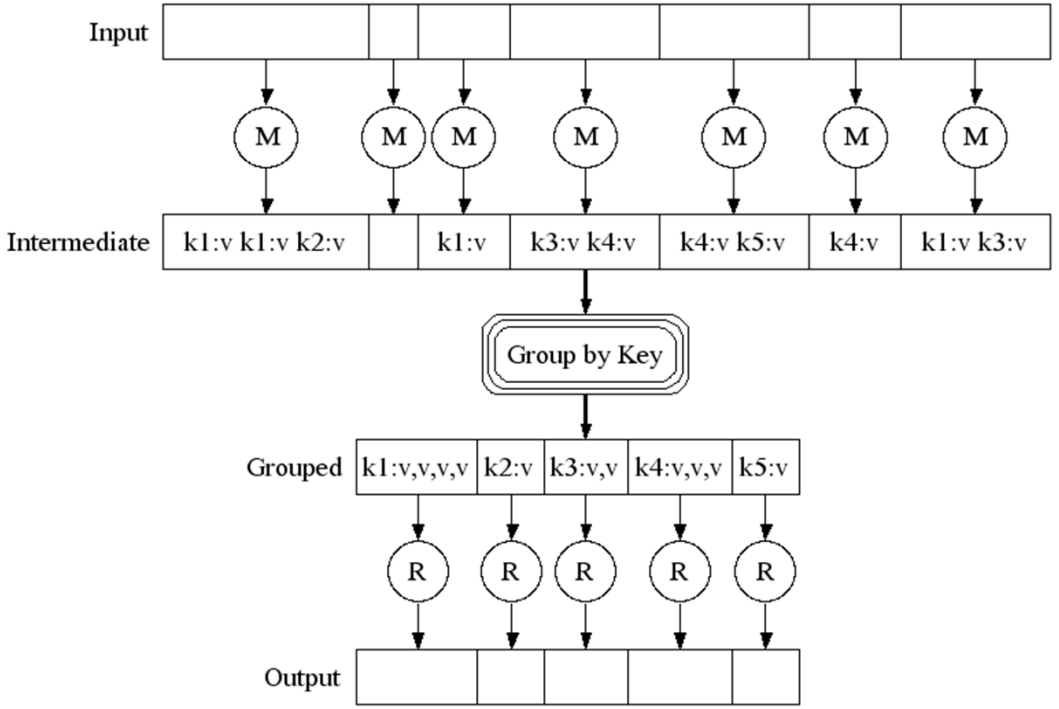
Elastic Compute Cloud: EC2

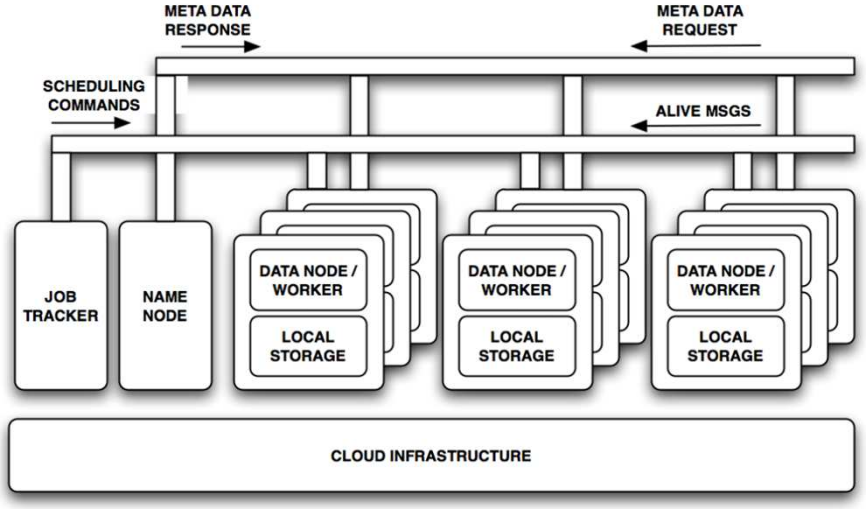
- Amazon Machine Images
  - Basically a Xen VM image: operating system, application server, and applications
  - Launch instances: run copies of the AMI
  - Runs until you stop or terminate them or if it fails
- Storage
  - Store the AMI images in S3
  - EBS: essentially hard disks that you can attach to a running instance



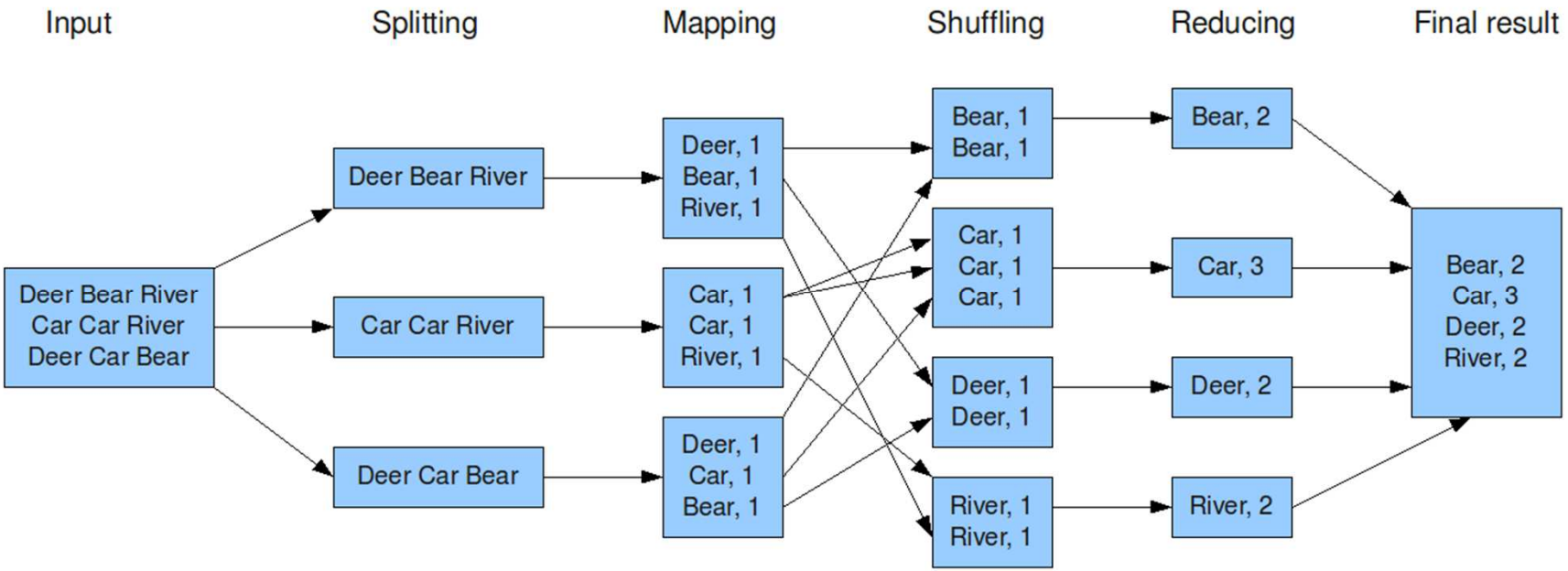
Background: Map-Reduce

- Programming model + software framework
- Parallelization + data distribution
- Hadoop:
  - Each job is split into map and reduce tasks
  - Master & Slave roles
  - Computation Layer: "MapReduce"
  - Storage Layer: "Hadoop Distributed File System (HDFS)"





Word count with Map-Reduce



http://www.rabidgremlin.com/data20/#(3)

Elastic MapReduce (EMR)

- Cheaper and more convenient than building it on EC2
- Uses hosted Hadoop framework on EC2 and S3

1. Write Hadoop program in Java
2. Submit the jar to EMR
3. Store the input in S3
4. Invoke EMR to run it
5. EMR runs it and stores the results back in S3

- EMR looks for unused resources to minimize the costs

Region: US East (Virginia)

	Amazon EC2 Price	Amazon Elastic MapReduce Price
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.015 per hour
Large	$0.34 per hour	$0.06 per hour
Extra Large	$0.68 per hour	$0.12 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.09 per hour
Double Extra Large	$1.00 per hour	$0.21 per hour
Quadruple Extra Large	$2.00 per hour	$0.42 per hour
Hi-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.03 per hour
Extra Large	$0.68 per hour	$0.12 per hour
Cluster Compute On-Demand Instances		
Quadruple Extra Large	$1.30 per hour	$0.27 per hour
Cluster Compute Eight Extra Large	$2.40 per hour	$0.50 per hour
Cluster GPU On-Demand Instances		
Quadruple Extra Large	$2.10 per hour	$0.42 per hour

Auto Scaling & Elastic Load Balance

- Auto Scaling
  - Monitor the load on EC2 instances using CloudWatch
  - Define Conditions
  - Spawn new instances when there is too much load or remove instances when not enough load
- Elastic Load Balance
  - Automatically distributes incoming application traffic across multiple EC2 instances
  - Detects EC2 instance health and divert traffic from bad ones
  - Support different protocols
    - HTTP, HTTPS, TCP, SSL, or Custom
- They can work together

Relational Database Service (RDS)

- Preconfigured EC2 instances with MySQL or Oracle installed
  1. Create an RDS instance
  2. Dump your database into it
     - mysqldump acme | mysql --host=hostname --user=username --password acme
  3. Update SQL connection strings in your application (which might be running anywhere, including EC2 VMs)
- Features
  - Pre-configured
  - Monitoring and Metrics (CloudWatch)
  - Automatic Software Patching
  - Automated Backups
  - DB Snapshots
  - Changing the instance type (= increase computer power)
    - Through EBS snapshots
  - Multi-AZ Deployments
  - Read Replicas
    - Scaling for read-heavy database workloads
  - Isolation and Security

SimpleDB

- A NoSQL database, non-relational
- Eventual consistency, no ACID compliance
- Data model is comprised of domains, items, attributes and values
  - Large collections of items organized into domains
  - Items are little hash tables containing attributes of key, value pairs
- Use Put, Batch Put, & Delete to create and manage the data set
- Use GetAttributes to retrieve a specific item
- Attributes can be searched with various lexicographical queries
- The service manages infrastructure provisioning, hardware and software maintenance, replication, indexing of data items, and performance tuning
- Tables limited to 10 GB, typically under 25 writes/second
- User manages partitioning and re-partitioning of data over additional SimpleDB tables

SimpleDB	S3
Indexes all the attributes	Stores raw data
Uses less dense drives	Uses dense storage drives
Better optimized for random access	Optimized for storing large objects

DynamoDB

- Amazon Dynamo paper (2007) → Open-source Apache Cassandra project → DynamoDB (1/2012) *
  - Dynamo is a highly available, key-value structured storage system
- Fully managed NoSQL non-relational Database
- Data model is comprised of domains, items, attributes and values (similar to SimpleDB)
  - Domains are collections of items that are described by attribute-value pairs
- Pay by throughput, not storage
- Integrates with Hadoop MapReduce using Elastic MapReduce
- Run on solid state disks (SSDs)
- There are no limits on the request capacity or storage size for a given table.
  - DynamoDB automatically partitions data and workload over a sufficient number of servers to meet the scale requirements

*http://www.datastax.com/dev/blog/amazon-dynamodb

ElasticCache

- a web service that makes it easy to set up, manage, and scale distributed in-memory cache environments in the cloud.
  - Protocol-compliant with Memcached
- Memcached is
  - a free & open source, high-performance, distributed memory object caching system, generic in nature, but intended for use in speeding up dynamic web applications by alleviating database load.
  - Memcached is an in-memory key-value store for small chunks of arbitrary data (strings, objects) from results of database calls, API calls, or page rendering.

Content delivery: CloudFront

- Delivers static and streaming content using a global network of edge locations
- Store the original versions of your files on an origin server.
  - Amazon S3 bucket, Amazon EC2 instance, or your own server
- Register the origin server with CloudFront through a simple API call
- When users request an object using the original domain name, they are automatically routed to the nearest edge location
- Similar to Akamai services, but in cloud
  - To use akamai you have to contact them and get individual pricing

Deployment & Management

- Identity and Access Management (IAM)
- IAM enables you to securely control access to AWS services and resources for your users. IAM enables you to create and manage users in AWS, and enables you to grant access to AWS resources for users managed outside of AWS in your corporate directory.

D&M: Amazon CloudWatch

- Monitor AWS resources automatically
  - Monitoring for Amazon EC2 instances: seven pre-selected metrics at five-minute frequency
  - Amazon EBS volumes: eight pre-selected metrics at five-minute frequency
  - Elastic Load Balancers: four pre-selected metrics at one-minute frequency
  - Amazon RDS DB instances: thirteen pre-selected metrics at one-minute frequency
  - Amazon SQS queues: seven pre-selected metrics at five-minute frequency
  - Amazon SNS topics: four pre-selected metrics at five-minute frequency
- Custom Metrics generation and monitoring
- Set alarms on any of the metrics to receive notifications or take other automated actions
- Use Auto Scaling to add or remove EC2 instances dynamically based on CloudWatch metrics

D&M: Elastic Beanstalk

- Solution for Enterprise server-side java application deployment
- Create your application (e.g. Eclipse).
- Package deployable code into a standard Java Web Application Archive (WAR file).
- Upload the WAR file to Elastic Beanstalk using the AWS Management Console, …
- Deploy the application
  - Elastic Beanstalk handles the provisioning of a load balancer and the deployment of the WAR file to one or more EC2 instances running the Apache Tomcat application server
- Access the application at a customized URL (e.g. http://myapp.elasticbeanstalk.com/).

Deployment & Management

- AWS CloudFormation is a service that gives developers and businesses an easy way to create a collection of related AWS resources and provision them in an orderly and predictable fashion.
  - Amazon's version of chef, opscode, puppet,…
- AWS Data Pipeline is a service that helps you reliably process and move data between different AWS compute and storage services as well as on-premise data sources at specified intervals.
- AWS OpsWorks is a DevOps platform for managing applications of any scale or complexity on the AWS cloud.
  - DevOps is a software development method that stresses communication, collaboration and integration between software developers and information technology (IT) professionals.

Application services

- CloudSearch is a fully-managed search service in the cloud that allows customers to integrate fast and highly scalable search functionality into their applications.
- Simple Workflow Service helps you coordinate the processing steps in your applications and manage distributed execution state.
- Amazon Elastic Transcoder is a fully managed service to convert media files in the cloud with scalability and at a low cost.

Messaging

- Simple Queue Service provides a hosted queue for storing messages as they travel between computers, to build automated workflow between Web services.
- Simple Notification Service is a web service to set up, operate, and send notifications from the cloud.
- Simple Email Service is a scalable and cost-effective bulk and transactional email-sending service for the cloud.

Network

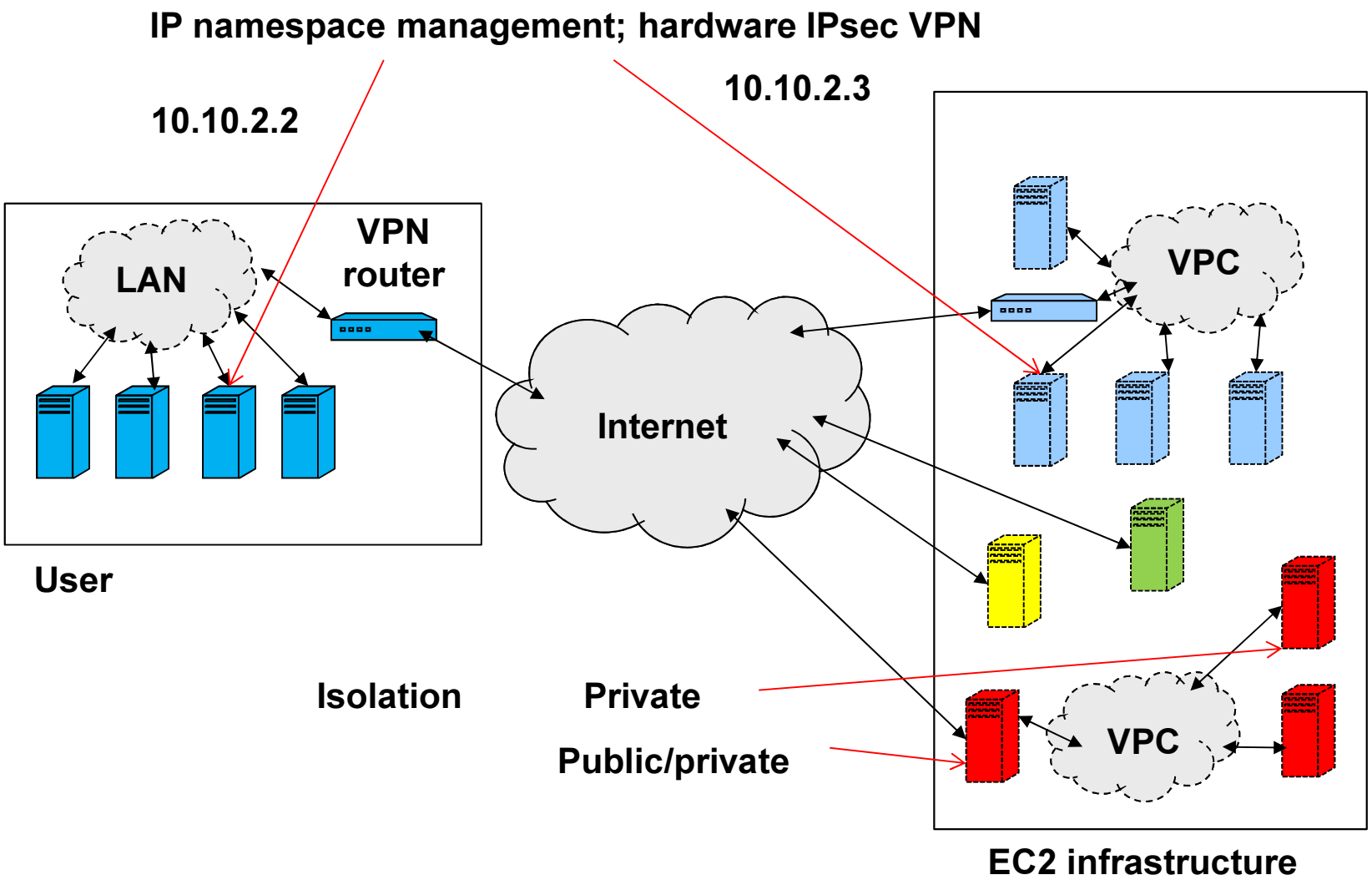
- Amazon Route 53 is a highly available and scalable Domain Name System (DNS) web service.
- AWS Direct Connect makes it easy to establish a dedicated network connection from your premise to AWS, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

Amazon Virtual Private Cloud (VPC)

- “lets you provision a private, isolated section of the Amazon Web Services (AWS) Cloud where you can launch AWS resources in a virtual network that you define.”
- “You have complete control over your virtual networking environment, including your IP address range, creation of subnets, and configuration of route tables and network gateways.”
- “you can create a Hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC and leverage the AWS cloud as an extension of your datacenter.”

From http://aws.amazon.com/vpc

Amazon Virtual Private Cloud



Web Traffic

- Alexa Web Information Service makes Alexa's huge repository of data about structure and traffic patterns on the Web available to developers.
  - Alexa is the leading provider of free, global web metrics. Search Alexa to discover the most successful sites on the web by keyword, category, or country.
- Alexa Top Sites exposes global website traffic data as it is continuously collected and updated by Alexa Traffic Rank.

Workforce

- Amazon Mechanical Turk enables companies to access thousands of global workers on demand and programmatically integrate their work into various business processes.

Payment and Billing

- Flexible Payments Service facilitates the digital transfer of money between any two entities, humans or computers.

- DevPay is a billing and account management service which enables developers to collect payment for their AWS applications.

Open-Source Cloud Software Solutions

- Motivation: replace a mix of different APIs - most proprietary - making it difficult or infeasible to deploy and to evaluate security by a standard API that is open and freely available.
  - "the Linux of Cloud Computing Platforms"
  - *open source software that enterprises/service providers can use to setup and run their cloud compute and storage infrastructure*
- Competing initiatives
  - Openstack
    - More than 100 consortium members
  - Cloudstack
    - Currently a project of the Apache Software Foundation

CloudStack

Compute

Hypervisor

XenServer | VMware | Oracle VM | KVM | Bare metal

Storage

Block & Object

Local Disk | iSCSI | Fiber Channel | NFS | Swift

Primary Storage | Secondary Storage

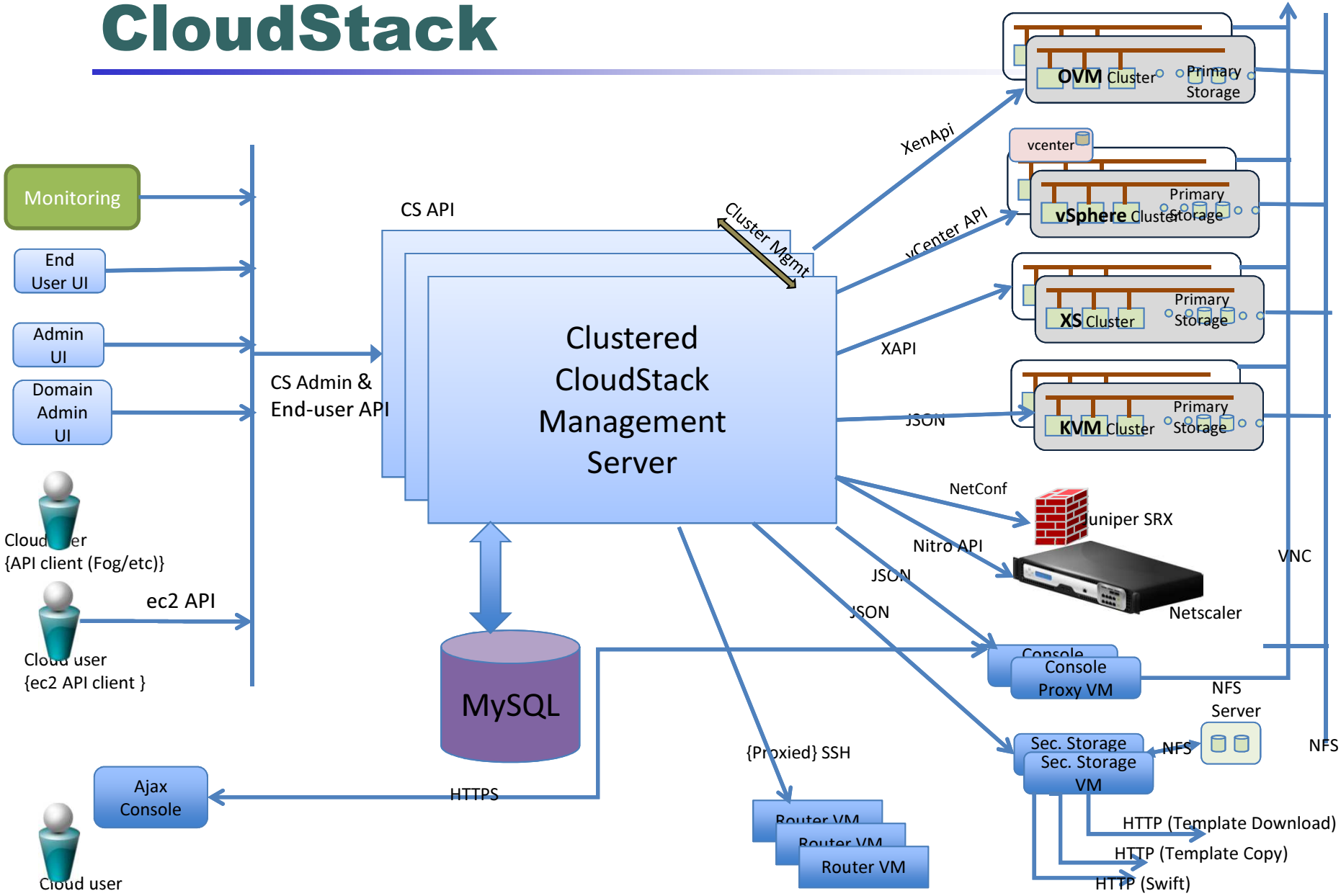
Network

Network & Network Services

Network Type | Isolation | Firewall | Load balancer | VPN

http://www.slideshare.net/cloudstack/cloudstack-architecture

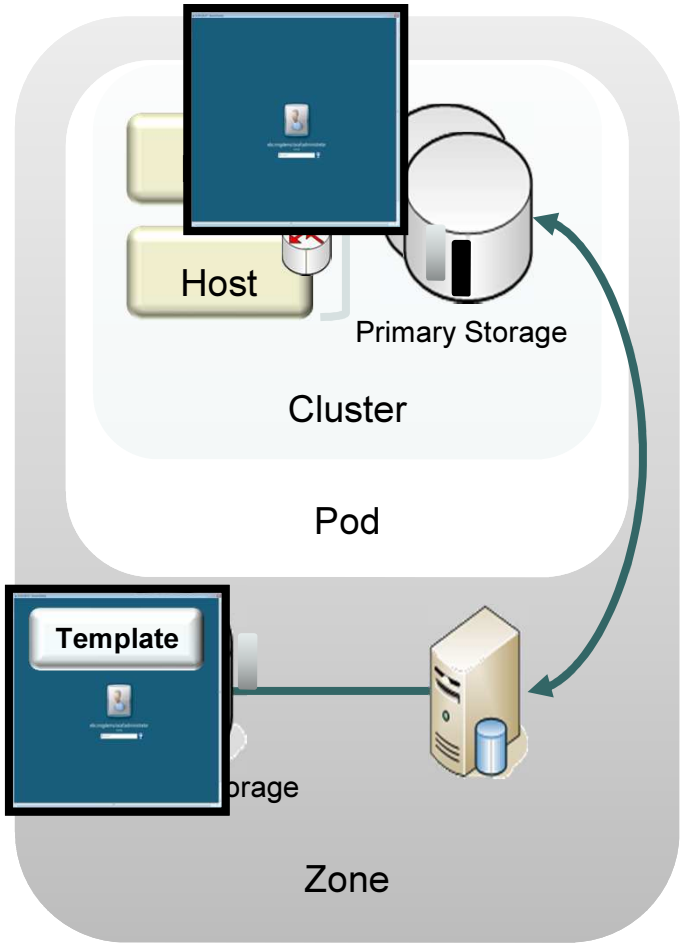
CloudStack



http://www.slideshare.net/cloudstack/cloudstack-architecture

Provisioning Process

1. User Requests Instance
2. Provision Optional Network Services
3. Copy instance template from secondary storage to primary storage on appropriate cluster
4. Create any requested data volumes on primary storage for the cluster
5. Create instance
6. Start instance



http://www.slideshare.net/cloudstack/cloudstack-architecture

DevCloud

- Pre-installed/configured CloudStack management
  - Provided as VirtualBox VM image
  - Nested virtualization (Xen hypervisor runs in a VirtualBox environment)

- https://cwiki.apache.org/confluence/display/CLOUDSTACK/DevCloud

DevCloud demo

OpenStack's Core Components

- Compute ("Nova")

  Orchestrates large networks of Virtual Machines.

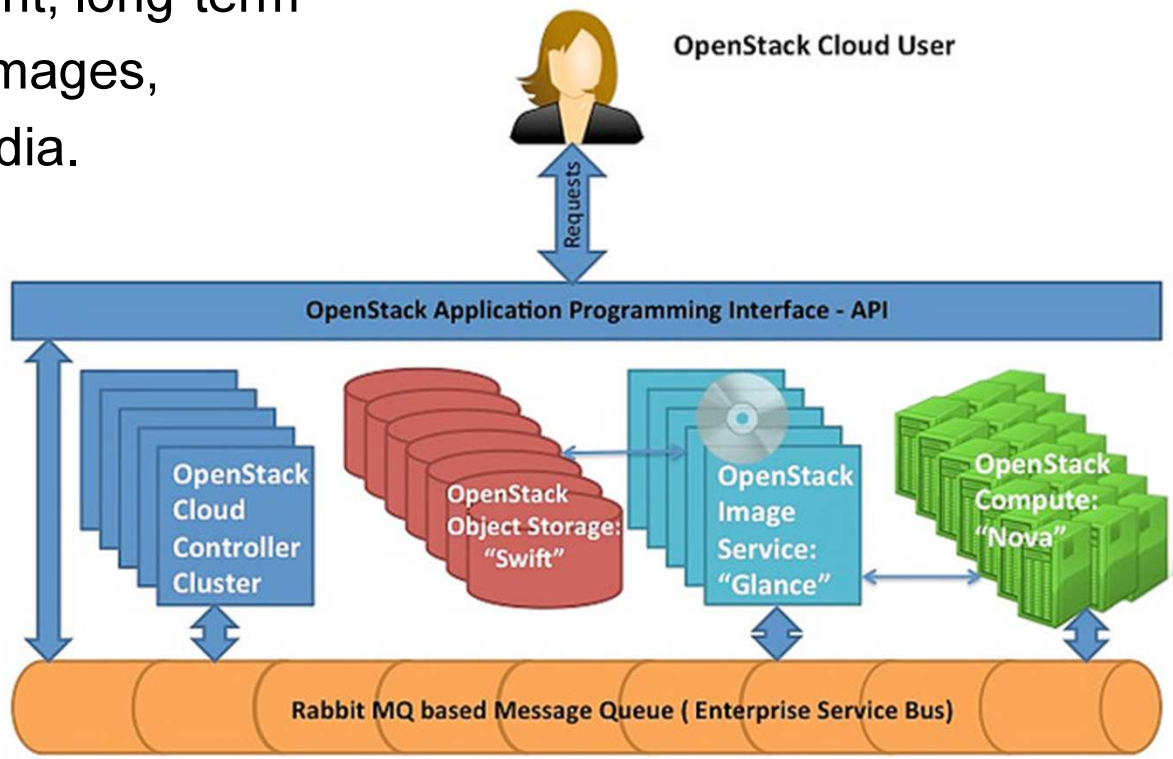
  Responsible for VM instance lifecycle, network management, and user access control.

- Object Storage ("Swift")

  Provides scalable, redundant, long-term storage for things like VM images, data archives, and multimedia.

- Image Service ("Glance")

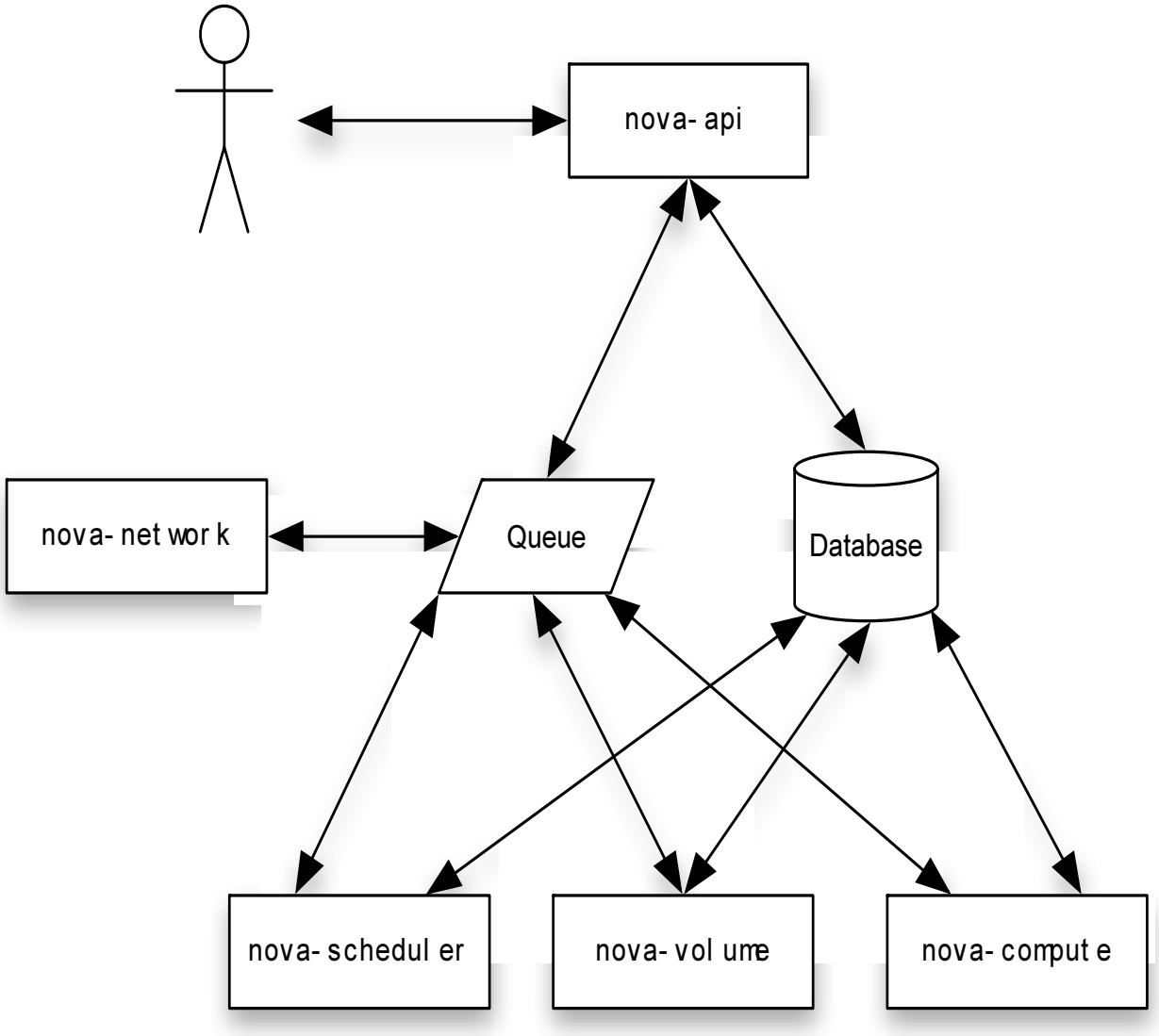
  Manages VM disk images.



OpenStack Nova

- Contributed by NASA from the Nebula platform.
  - allows users to create, destroy, and manage virtual machines using user-supplied images.
- Corresponds to Amazon's EC2.
- Users can use OpenStack API or Amazon's EC2 API.
- Uses Python and Web Server Gateway Interface (WSGI).

OpenStack Nova: Architecture



OpenStack Nova: nova-api

A daemon that is the workhorse of Nova.

- Handles API requests.
- Manages most orchestration.
- Enforces some policies.

If it can, it will handle the request on its own with help from the database.

Otherwise, it will delegate to the other nova daemons using the message queue as well as the database.

OpenStack Nova: nova-compute

Worker that does the actual work of starting and stopping virtual machine instances.

Takes its orders from the message queue, and executes the appropriate VM API calls to accomplish the task.

Commonly uses “libvirt” (RedHat), but can use Xen, vSphere (VMware), or Windows Management Interface.

OpenStack Nova: nova-network

Worker that does the actual work of configuring the network.

Network is specified as one of three types:

- Flat
- FlatDHCP
- VLAN
- Can use Quantum to take advantage of software-defined network components

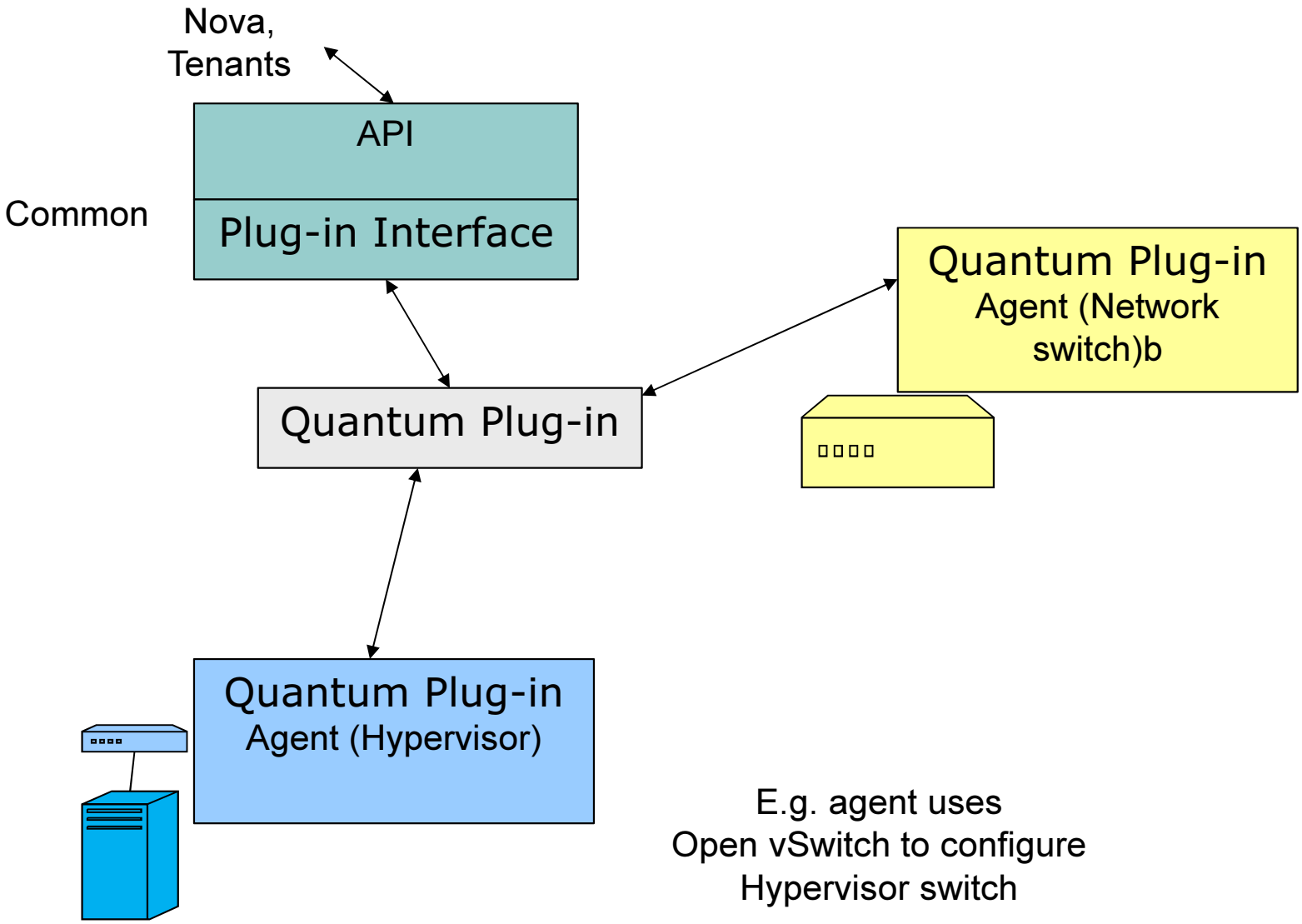
Openstack Quantum

- Service to establish connectivity among virtual NICs managed by Openstack cloud
- From Wiki: http://wiki.openstack.org/Quantum
  - “Give cloud tenants an API to build rich networking topologies, and configure advanced network policies in the cloud.”
  - “Let anyone build advanced network services (open and closed source) that plug into Openstack tenant networks.”
- Quantum plugin - manage configuration of virtual switches (at VMM) and physical switches
  - Plug-in may use OpenFlow to manage switches

Openstack Quantum

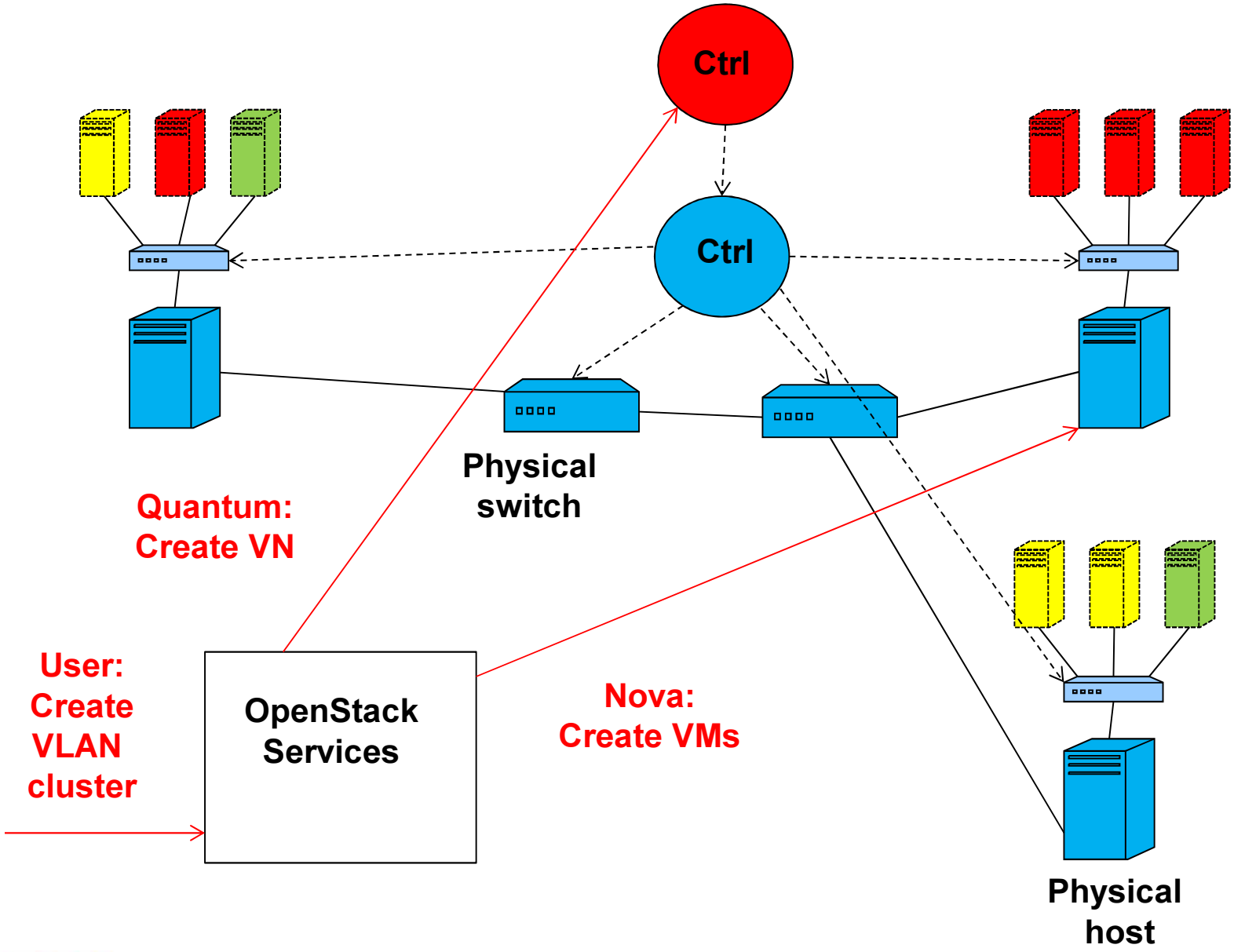
- Quantum specifies service APIs
  - Simple APIs for creating, managing virtual networks
    - Add, update, remove networks and ports
    - Plug/unplug attachments
  - Technology-agnostic
- Layer-2 networking
- Expose tenant-facing APIs
- Enable rich network topologies
- Leverage emerging network virtualization technologies

Openstack Quantum



E.g. agent uses
Open vSwitch to configure
Hypervisor switch

Quantum and OpenFlow



Opportunities ahead

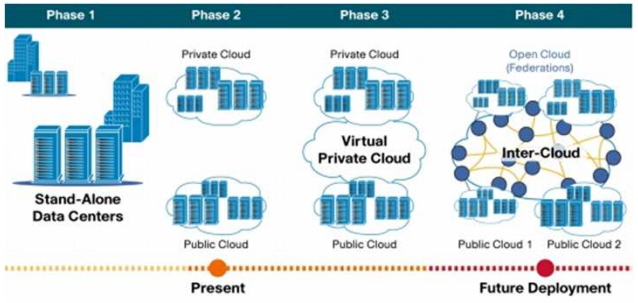
- Challenges remain
  - Security
  - Intercloud
  - Software-defined networking

Cloud security

- More secure (than consumer resources)
  - Better design, configuration, support, software, etc.
- Less secure
  - Multi-tenancy, exposure, expensive crypto, etc.
  - Existing “secure” data outsourcing mechanisms are 2-5 orders of magnitude more expensive than local execution.
- Depends on meaning of security
  - Applications behave as expected
  - Data Integrity, Query Correctness, Data Confidentiality, Query Privacy, Access Privacy

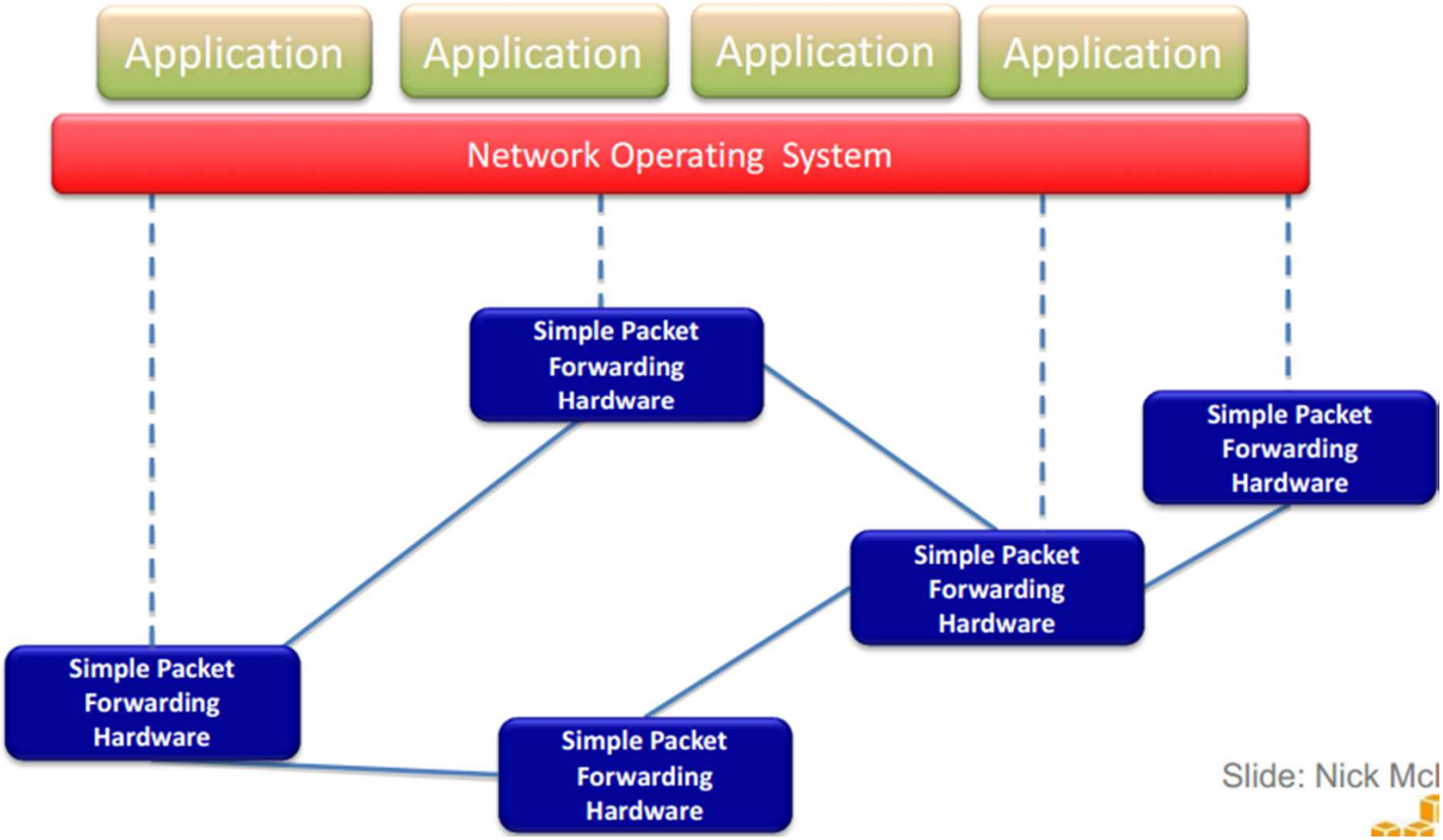
Why federated/multi/inter clouds?

- Different clouds for different
  - Applications and/or Data
  - Components of an application
  - Stages of application lifecycle
  - Customer locations
  - Levels of demand
  - Security/confidentiality/policy/reliability requirements
  - Service models (centralized data distributed computing or vice-versa)
  - Costs
- Common example: cloud bursting



Software-defined networks

- Next talk



Conclusions

- Cloud computing is here to stay
  - Not pure hype
- Changes many things about computing
  - Economics
  - How we program
  - How we design and build systems
  - How we build services and businesses
- A rich set of technologies
- New applications (social nets, big data, …)
- New set of problems
  - Security, intercloud, scalability,