X-Informatics
Cloud Computing Technology Part II

June 20 2013

Geoffrey Fox
gcf@indiana.edu
http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research, School of Informatics and Computing
Indiana University Bloomington
2013

Big Data Ecosystem in One Sentence

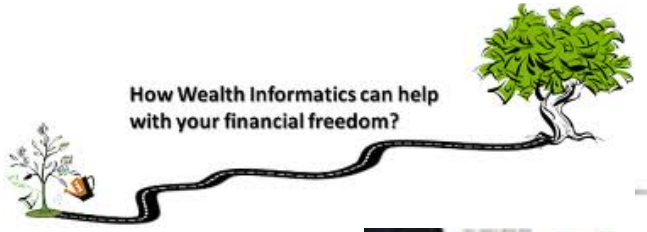
Use Clouds running Data Analytics processing Big Data to solve problems in X-Informatics (or e-X)

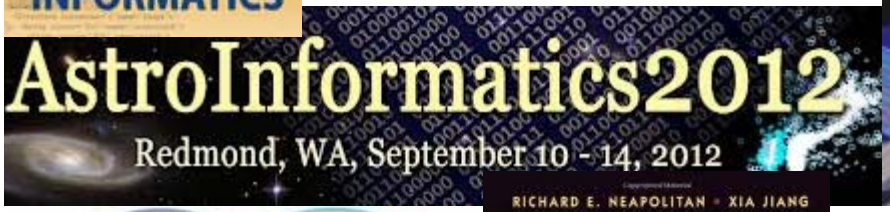
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

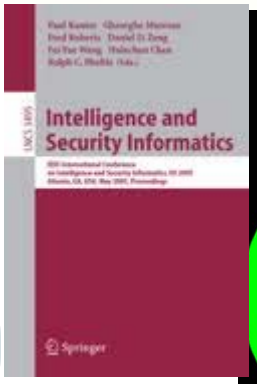
Spans Industry and Science (research)

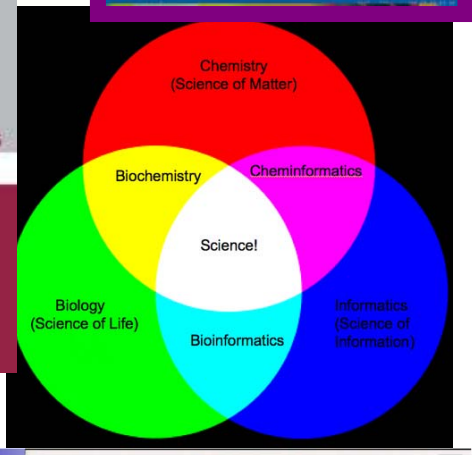
Education: Data Science see recent New York Times articles

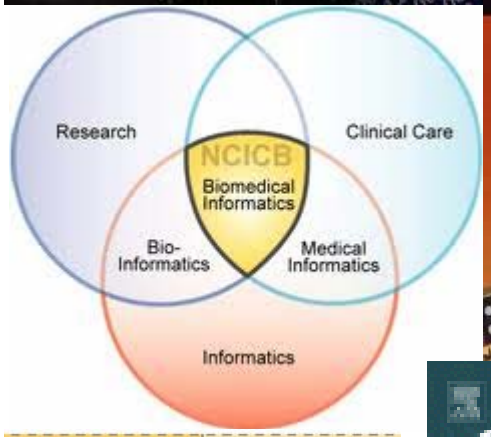
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

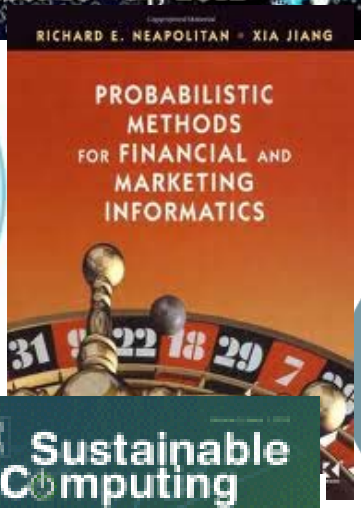


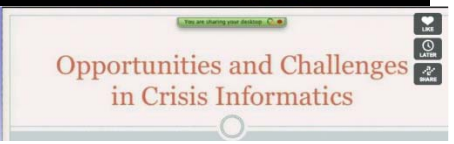


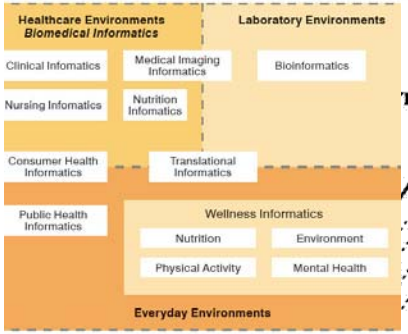


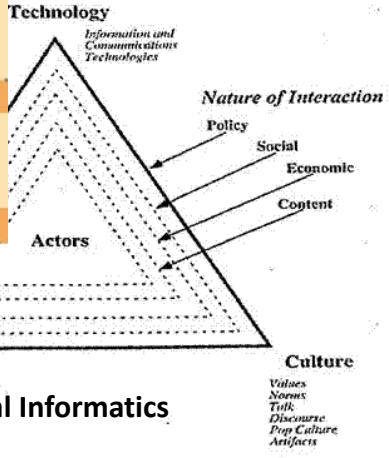




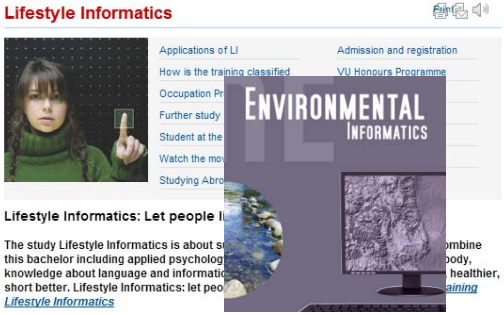








Social Informatics



What is Cloud Computing

In more detail

http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis

Cloud Computing provides solutions to a variety of challenges and opportunities

The classical problem

- Under-utilized server resources waste computing power (and energy)
- Over-utilized servers cause interruption or degradation of service levels

…today in an Internet setting

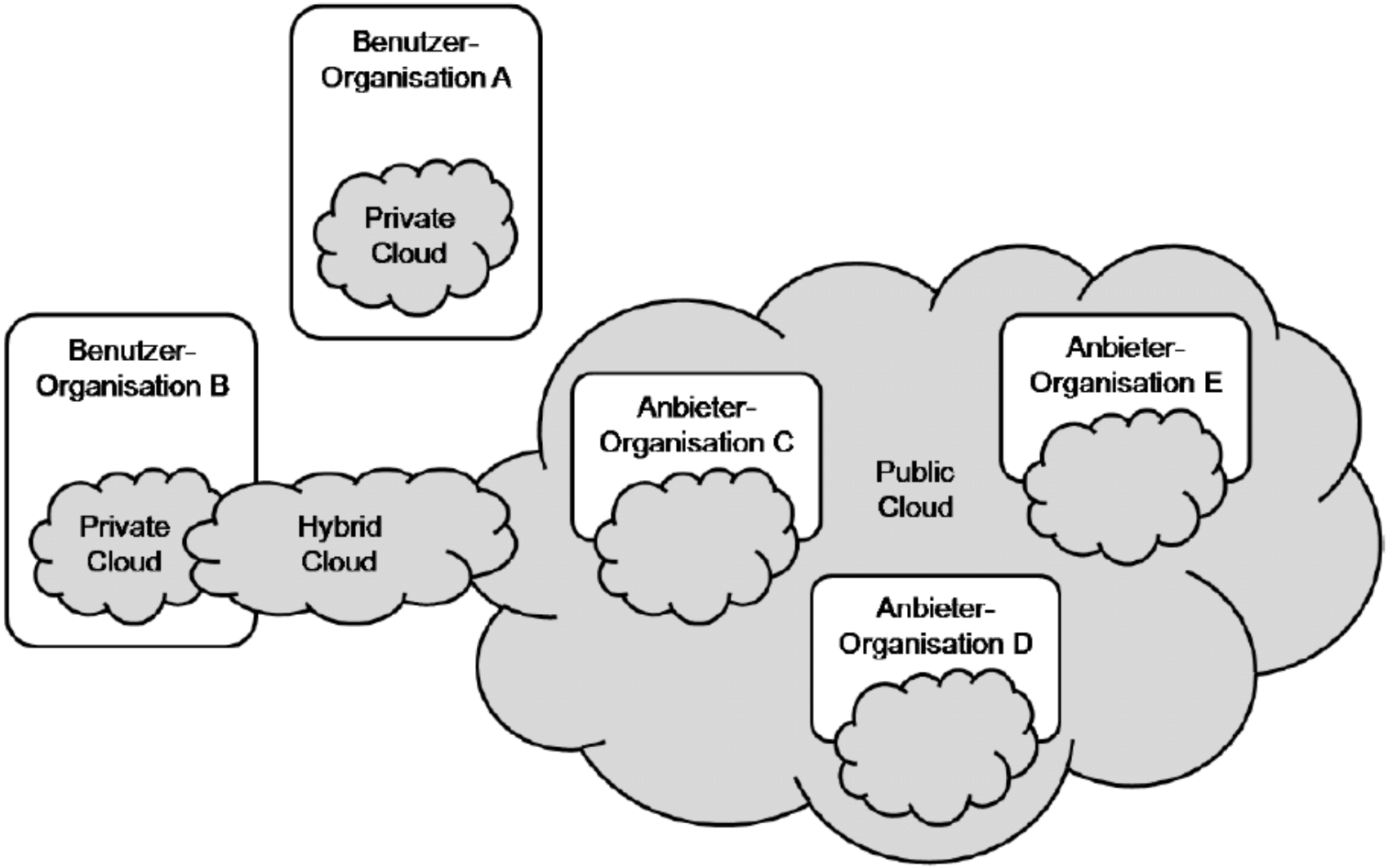
- Resource demands are increasingly of highly dynamic nature and Internet-scale
- On-demand resources are a means for faster time-to-market, and cost-effective innovation processes

…and tomorrow in the next-gen Web

- Leveraging the Web as a combined technology, business, and people collaboration platform:
  - Making effective use of sophisticated infrastructure which is increasingly available as (Web) services
  - Enabling dynamic (trans-)formation of open service and business networks

http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis

Organizational Cloud Architecture: Public-/Hybrid-/Private-Cloud



http://www.slideshare.net/JensNimis/cloud-computing-tutorial-jens-nimis

Players

Cloud infrastructure service providers – raw cloud resources
IaaS (infrastructure-as-a-service)

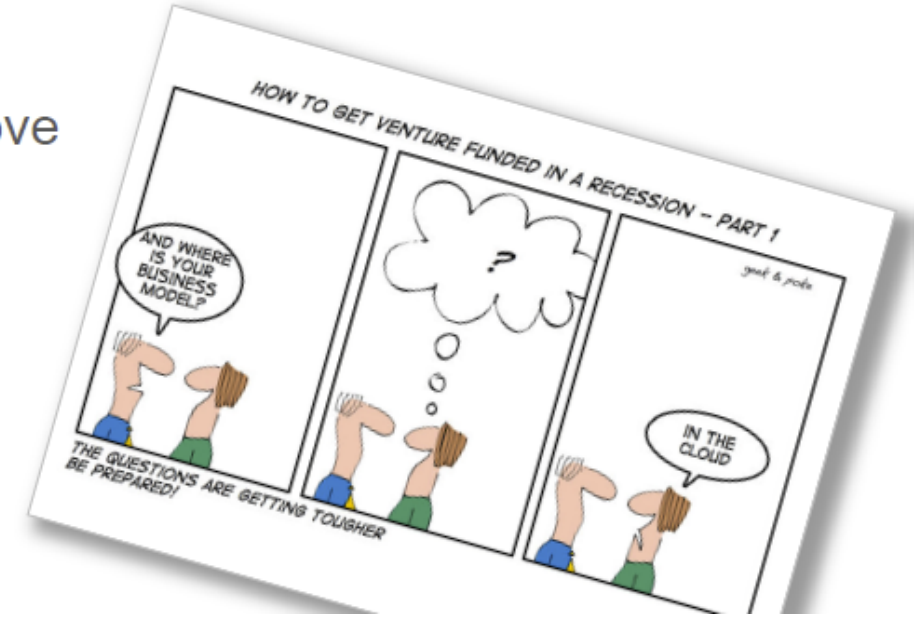
Cloud platform providers – resources + frameworks; PaaS (platform-as-a-service)

Cloud intermediares – help broker some aspect of raw resources and frameworks, e.g.,

server managers, application assemblers, application hosting

Cloud application providers (SaaS)

Cloud consumers – users of the above



[MM]

Cloud Talk

- ***Big "3" of Cloud Computing?***
- ***Cost***
  - ***Clouds are renowned for being dirt cheap for storage and burst-y processing.***
- ***Flexibility***
  - ***let someone else manage it for you.***
- ***Elasticity***
  - ***Growth and shrinkage***

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- ***More on why Clouds?***
- *On Demand Business*
  - *Unexpected loads, The <> effect*
- *Meeting Batch Load Demands*
  - *Batch*
- ***Parallelism***
  - *Large clusters of parallel jobs*
- ***Season Workloads***
  - *Retail, Travel, Financial*
- ***Backup Storage***

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- ***What is my definition of a "Holy-Grail" Cloud***
  - **Abstraction of the hardware infrastructure from the service.**
  - **Abstraction of the software infrastructure from the service.**

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- ***Cloud "Services" in Simple Terms***

- ***Applications***
  - ***Software provided as a service***
- ***Middle-ware***
  - ***Software stacks for developers LAMP, Java application servers, .Net***
- ***Servers***
  - ***Hardware and an operating system, perhaps a hyper-visor***

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- ***Processing Large Datasets***

- ***Map Reduce***
  - ***Jobs that run as hundreds or even thousands of separate parallel processes.***
  - ***Like counting the words in a book and break it up into multiple running parts (i.e., The Map)***
  - ***Then collect them all back into summary counts (i.e., The Reduce.)***

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- ***Processing Large Datasets***
- ***Hadoop***
  - ***Google invented GFS***
  - ***Yahoo***
  - ***AOL***
  - ***IBM***
  - ***Facebook***
  - ***Last.fm***

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- *Cloud Computing Challenges*
- *Retraining developers and operations people to deal with cloud computing*
- *Orchestration of multiple clouds*
- *24 by 7 by 365 operations in the cloud is usually more expensive*
- *Legacy applications might not port easily*
- *Virtualiztion project disruption*
- *Recent McKinsey Report $366 vs $150*

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- *More Challenges*

- *Workload Affinities*
- *Standards (Lock-in)*
- *Weak SLA's compared to Corp*
- *Service Management*
- *Security*
- *Compliance*
- *Image Sprawl*
- *Trojan Virtual Images*
- *Governance*

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

Cloud Talk

- *New Cloud Terms*
  - *Cloud Bursting*
    - *Analytics, Coding*
  - *Hybrid Clouds*
    - *VPN, Multiple Clouds*
  - *Cloud Spillage*
    - *An IBM Term*
  - *Cloud Orchestration*
    - *Managing multiple clouds*

johnmwillis.com

http://www.slideshare.net/botchagalupe/introduction-to-clouds-cloud-camp-columbus

The Cloud as an extension of your desktop and other client devices

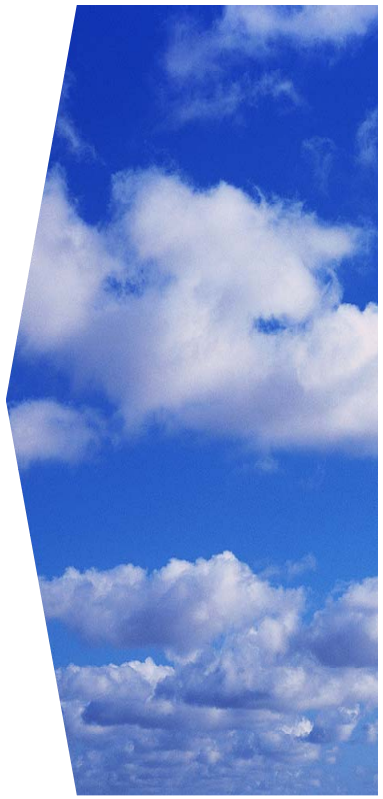
- Today
  - Cloud storage for your data files synchronized across all your machines (mobile me, live mesh, flicker, etc.)
  - Your collaboration space (Sakai, SharePoint)
  - Cloud-enabled apps (Google Apps, Office Live)
- Tomorrow (or even sooner)
  - The lens that magnifies the power of desktop
  - Operate on a table with a billion rows in excel
  - Matlab analysis of a thousand images in parallel

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

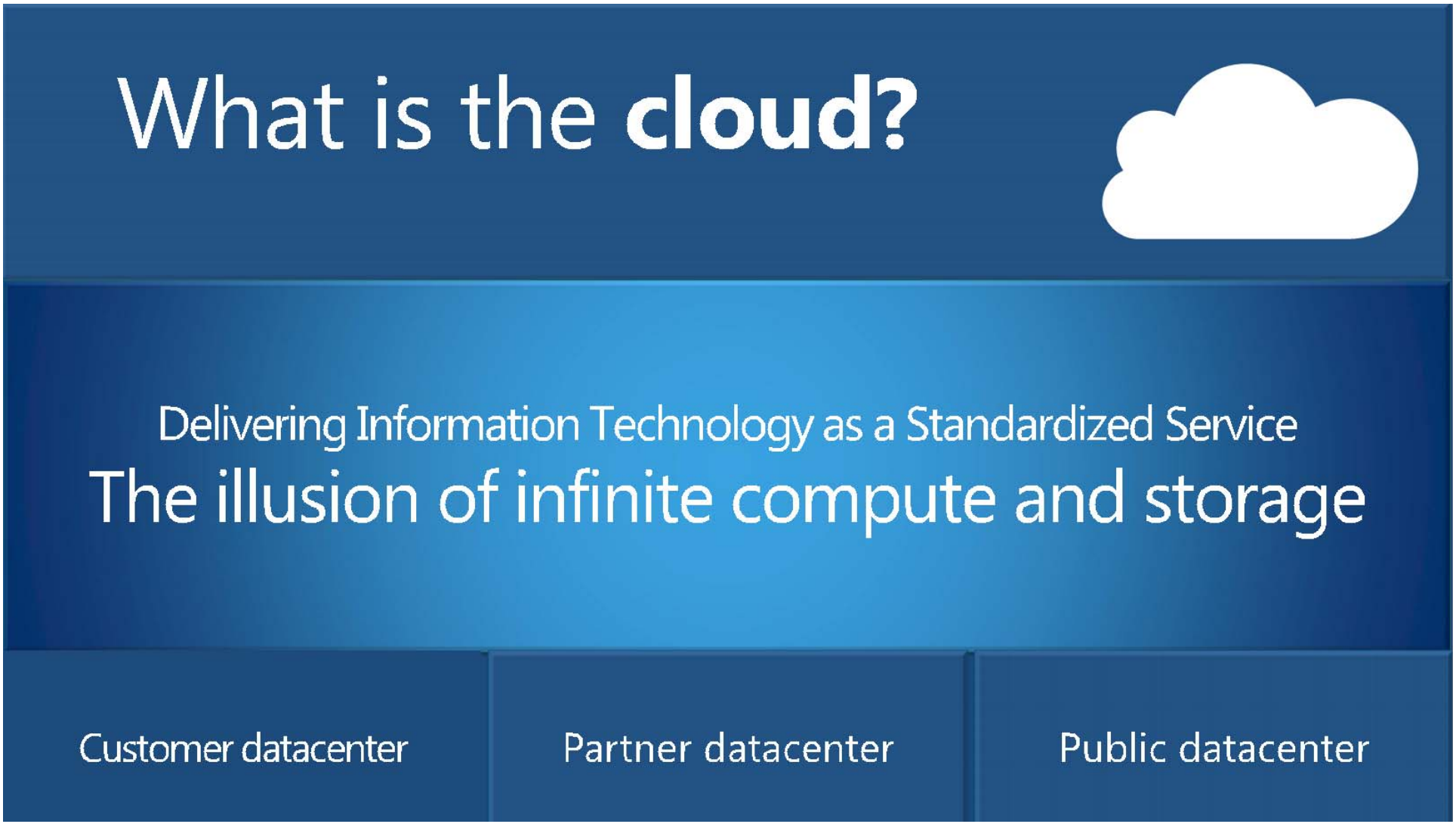
Microsoft

Cloud Properties

- **Designed to Provide Information and Computation to Many Users**
- Automatic Deployment and Management of Virtual Machine Instances
  - tens to thousands & dynamic scalability
- Dynamic Fault Recovery of Failed Resources
  - Cloud Services must run 24x7
- Automatic Data Replication
  - Geo-replication if needed
- Two levels of parallelism
  - Thousands of concurrent users
  - Thousands of servers for a single task.



http://www.oerc.ox.ac.uk/downloads/presentations-from-events/microsoftworkshop/gannon



http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAndChallenges_Yousef_Khalidi.pdf

Clouds Offer From different points of view

- **Features from NIST:**
  - On-demand service (elastic);
  - Broad network access;
  - Resource pooling;
  - Flexible resource allocation;
  - Measured service
- **Economies of scale** in performance and electrical power **(Green IT)**
- Powerful new **software models**
  - **Platform as a Service** is not an alternative to **Infrastructure as a Service –** it is instead an incredible valued added
  - Amazon is as much PaaS as Azure

What is Cloud?

- **Defining the Cloud**
  - A scalable, persistent outsourced infrastructure
  - A framework for massive data analysis
  - An amplifier of our desktop experience
- **The Origins**
  - Modern data center architecture

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

What is a cloud?

Eucalyptus

SLAs

The Or...s

...ata center architecture

Web Services

Virtualization

http://research.microsoft.com/en-us/people/barga/scog_cloudcomp_tutorial.pdf

Defining the Cloud

- A model of computation and data storage based on "pay as you go" access to unlimited remote data center capabilities.
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications.
- Examples:
  - Search, email, social networks
  - File storage (Live Mesh, Mobile Me, Flicker, …)
- A way for a start-up to build a scalable web presence without purchasing hardware.

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Cloud Mythologies

- **Cloud computing infrastructure is just a web service interface to operating system virtualization.**
  - –"I'm running Xen in my data center –I'm running a private cloud."
- **Clouds and Grids are equivalent**
  - –"In the mid 1990s, the term grid was coined to describe technologies that would allow consumers to obtain computing power on demand."
- **Cloud computing imposes a significant performance penalty over "bare metal" provisioning.**
  - –"I won't be able to run a private cloud because my users will not tolerate the performance hit."

Eucalyptus

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

The Clients+Cloud Platform

- At one time the “client” was a PC + browser.
- Now the cloud is an *integration point* for
  - The Phone
  - The laptop/tablet
  - The TV/Surface/Media wall
- And the future
  - The instrumented room
  - Aware and active surfaces
  - Voice and gesture recognition
  - Knowledge of where we are
  - Knowledge of our health

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

The History of the Cloud

- In the beginning …
  - There was search, email, messaging, web hosting
- The challenge: How do you
  - Support email for 375 million users?
  - Store and index 6.75 trillion photos?
  - Support 10 billion web search queries/month?
  - Build an index for the entire web? And do it over and over again…
- And
  - deliver deliver a quality response in 0.15 seconds to millions of simultaneous users?
  - never go down.
- Solution: build big data centers

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Microsoft

Public, Private, and Premise

Eucalyptus

- **Public Cloud**
  - Large scale infrastructure available on a rental basis
  - Virtualized compute, network and storage
  - Underlying infrastructure is shared but tenants are isolated
  - Interface is transactional
  - Accounting is e-commerce based
- **Private Cloud**
  - Dedicated resources either as a rental or on-premise
- **On-premise Cloud**
  - Like public clouds but
    - Isolation must be controllable
    - Accounting is organizational

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Public IaaS

Eucalyptus

- **Large scale infrastructure available on a rental basis**
  - Operating System virtualization (e.g. Xen, KVM) provides CPU isolation
  - "Roll-your-own" network provisioning provides network isolation
  - Locally specific storage abstractions
- **Fully customer self-service**
  - Customer-facing Service Level Agreements (SLAs) are advertized
  - Requests are accepted and resources granted via web services
  - Customers access resources remotely via the Internet
- **Accountability is e-commerce based**
  - Web-based transaction
  - "Pay-as-you-go" and flat-rate subscription
  - Customer service, refunds, etc.

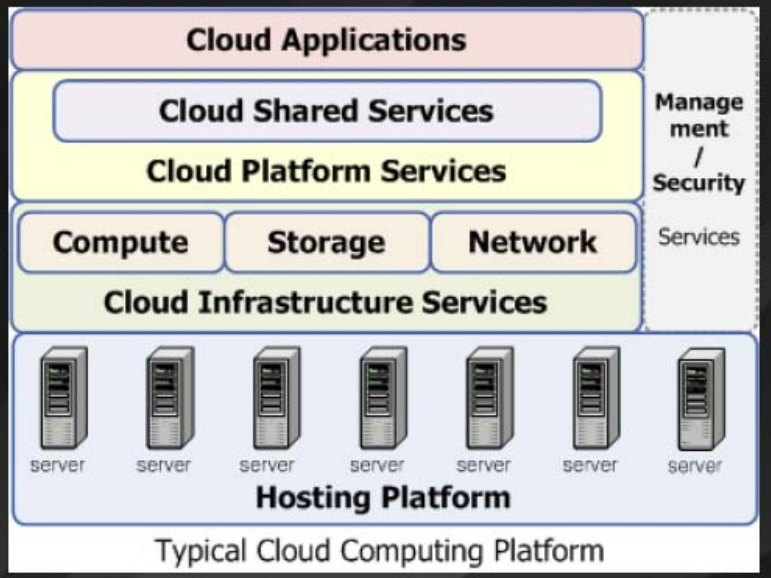
http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

As a Service and Platform Model

IaaS PaaS SaaS

Typical Cloud Computing Platform

- Hosting Platform
  - Provides the physical, virtual, and software assets which include physical machines, operating systems, network systems, storage systems, power management, and virtualization software
- Cloud Infrastructure Services(IaaS)
  - Abstract the hosting platform as a set of virtual resources(i.e. compute, storage, and network)
  - Manage those resources based on scalability and availability needs
- Cloud Platform Services(PaaS)
  - Provide a set of capabilities exposed as a services to help with integrating on-premise software with hosted services
- Cloud Applications(SaaS)
  - Houses applications that are built for cloud computing, which expose Web interfaces and Web Services for end users, enabling multitenant hosting models.



Typical Cloud Computing Platform

- Security Services
  - Ensure token provisioning, identity federation, and claims transformation
- Management Services
  - Provide a set of capabilities to automate scalability and availability administration such as deployment configurations, service usage analytics, and connection to enterprise management systems

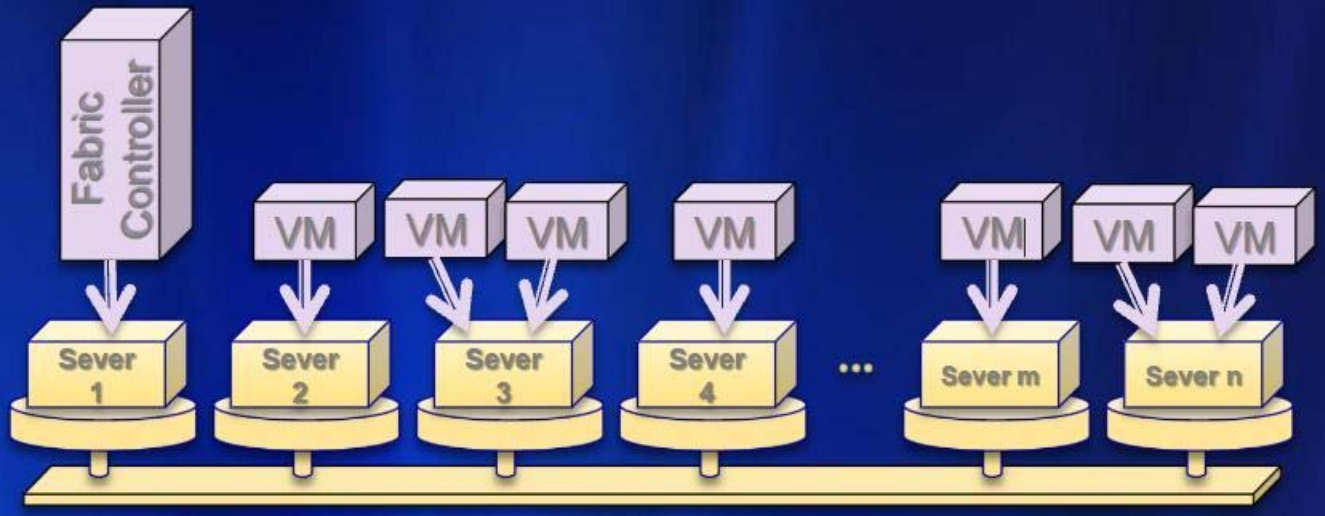
http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing

coord

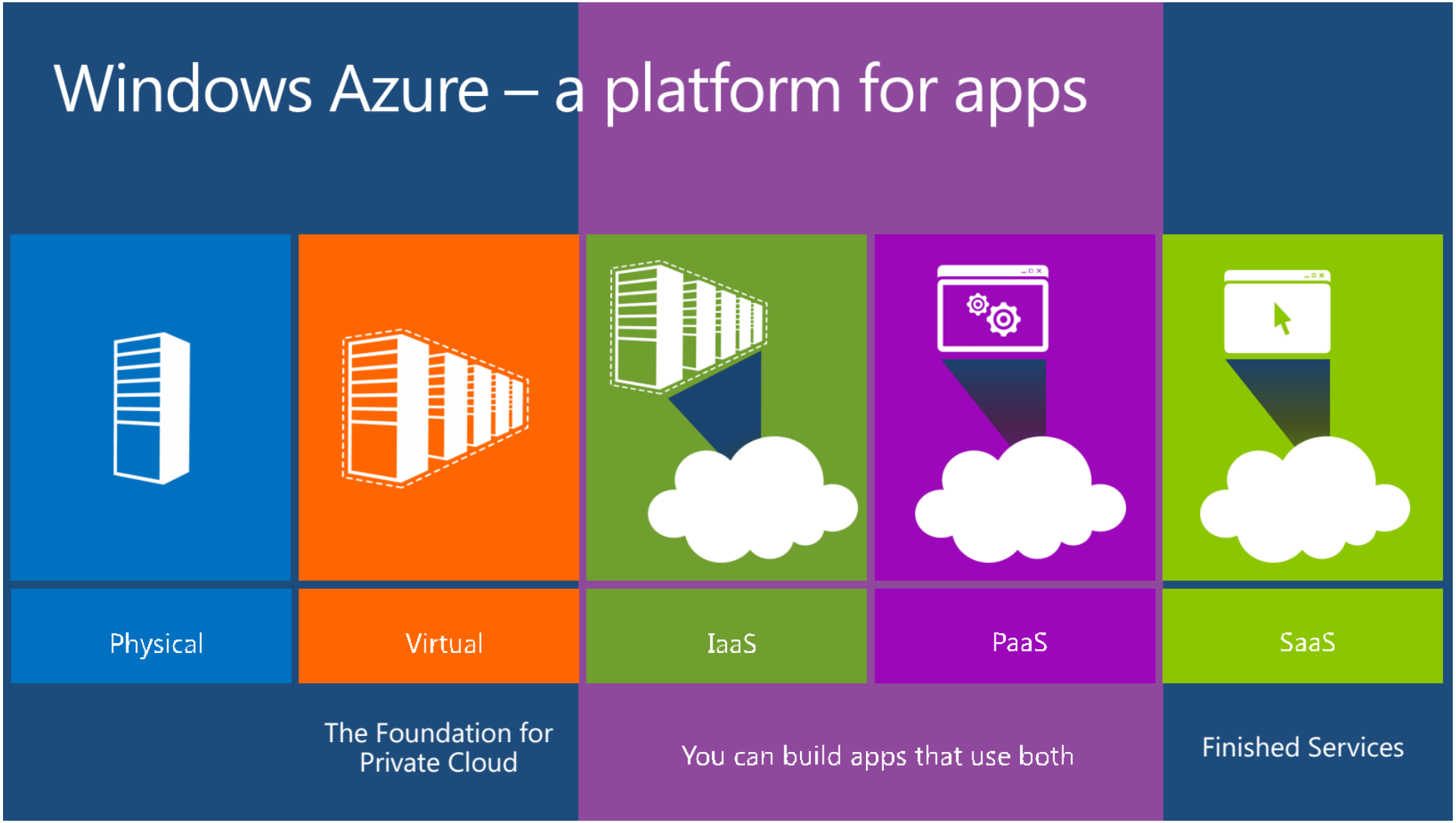
http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Three Levels of Cloud Arcitecture

- Infrastructure as a Service (IaaS)
  - Provide App builders a way to configure a Virtual Machine and deploy one or more instances on the data center
  - Each VM has access to local and shared data storage
  - The VM has an IP Address visible to the world
  - A Fabric controller manages VM instances
    - Failure and restart, dynamic scale out and scale back.



Microsoft



http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAndChallenges_Yousef_Khalidi.pdf

Clouds have highlighted SaaS PaaS IaaS

Software (Application Or Usage) SaaS
- Education
- Applications
- CS Research Use e.g. test new compiler or storage model

Platform PaaS
- Cloud e.g. MapReduce
- HPC e.g. PETSc, SAGA
- Computer Science e.g. Compiler tools, Sensor nets, Monitors

Infra structure IaaS
- Software Defined Computing (virtual Clusters)
- Hypervisor, Bare Metal
- Operating System

Network NaaS
- Software Defined Networks
- OpenFlow GENI

But equally valid for classic clusters

- Software Services are building blocks of applications
- The middleware or computing environment including **HPC, Grids** …
- Nimbus, Eucalyptus, OpenStack, OpenNebula CloudStack plus **Bare-metal**
- OpenFlow – *likely to grow in importance*

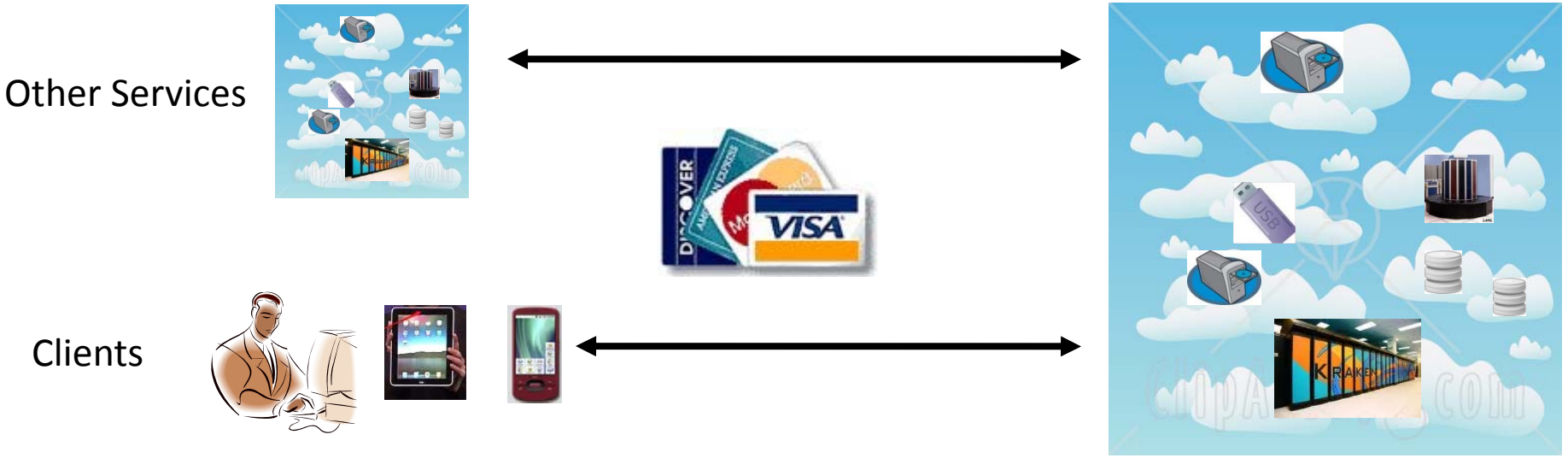
IaaS Open Source

- **Eucalyptus** (first major open source VM management environment)
- **Nimbus**
- **OpenStack** (currently dominant – 2800 at last OpenStack meeting)
- **OpenNebula** (Europe)
- **CloudStack** (break away from OpenStack)
- These manage basic computing and storage but these are separable ideas
- Management, Networking, Access Tools

- You install on your own clusters and do have control over operation
- Compare Public Clouds (Amazon Azure Google) where operation is opaque but you just need a credit card and a web browser
- Compare VMware commercial version which is similar to OpenSource but with better support and probably greater robustness/functionality
- Eucalyptus has Open Source and Commercial versions

X as a Service

- SaaS: Software as a Service imply software capabilities (programs) have a service (messaging) interface
  - Applying systematically reduces system complexity to being linear in number of components
  - Access via messaging rather than by installing in /usr/bin
- IaaS: Infrastructure as a Service or HaaS: Hardware as a Service – get your computer time with a credit card and with a Web interface
- PaaS: Platform as a Service is IaaS plus core software capabilities on which you build SaaS
- Cyberinfrastructure is "Research as a Service"



PaaS – What is a "cloud platform"?

"… data as a service…"

"cloud computing journal reports that…"

"… software as a service…"

"… everything as a service…"

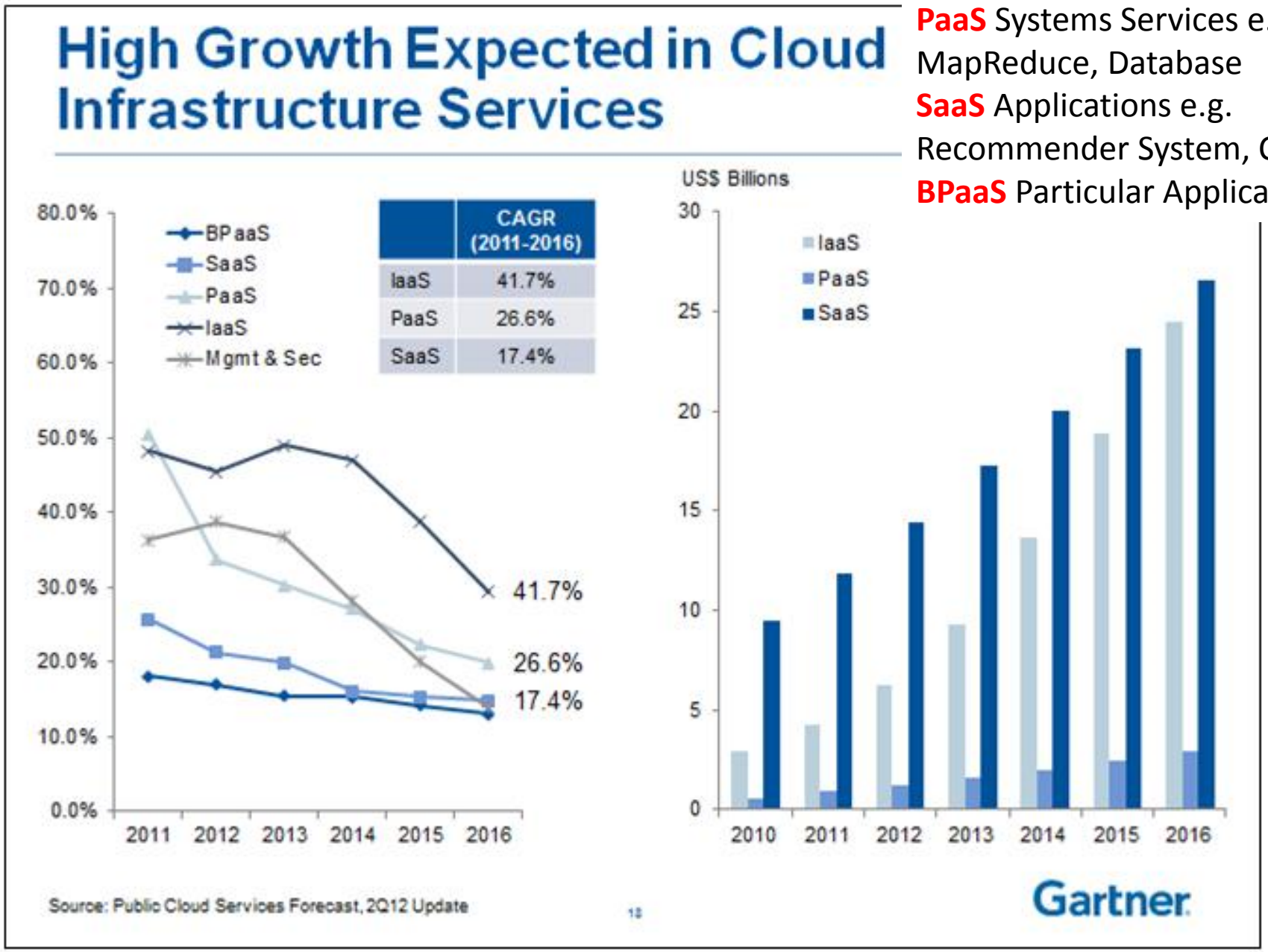
Platforms succeed when the platform helps others succeed

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Software as a Service

- Online delivery of applications
- Via Browser
  - Microsoft Office Live Workspace
  - Google Docs, etc.
  - File synchronization in the cloud – Live Mesh, Mobile Me
  - Social Networks, Photo sharing, Facebook, wikipedia etc.
- Via Rich Apps
  - Science tools with cloud back-ends
    - Matlab, Mathematica
  - Mapping
    - MS Virtual Earth, Google Earth
  - Much more to come.

IaaS Hardware e.g. Server
PaaS Systems Services e.g. MapReduce, Database
SaaS Applications e.g. Recommender System, Clustering
BPaaS Particular Application Set



BPM = Business Process management

Platform as a Service

http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Platform as a Service

- An application development, deployment and management fabric.
- User programs web service front end and computational & Data Services
- Framework manages deployment and scale out
- No need to manage VM images

App User

Internet

App Developer

PaaS Dev/Deploy Fabric

Web Access Layer

Fabric Controller

Data & Compute Layer

VM VM VM VM VM VM VM

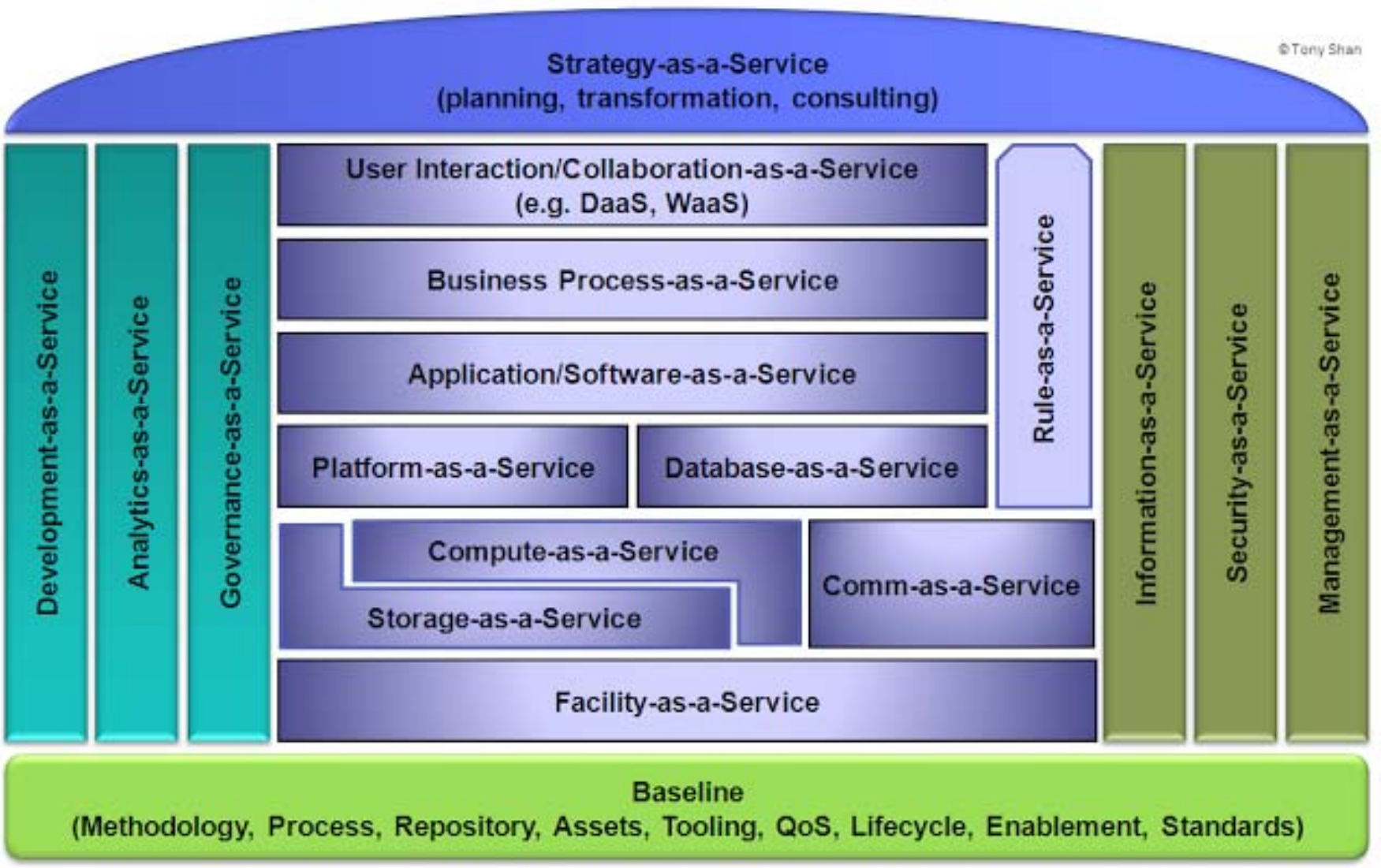
Sever 1 Sever 2 Sever 3 Sever 4 ... Sever m Sever n

Microsoft

Cloud Computing: Infrastructure and Runtimes

- Cloud Infrastructure: outsourcing of servers, computing, data, file space, utility computing, etc.
  - Handled through (Web) services that control virtual machine lifecycles.
- Cloud Runtimes or Platform: tools (for using clouds) to do data-parallel (and other) computations.
  - Apache Hadoop, Google MapReduce, Microsoft Dryad, Bigtable, Chubby (synchronization) and others
  - MapReduce designed for information retrieval but is excellent for a wide range of science data analysis applications
  - Can also do much traditional parallel computing for data-mining if extended to support iterative operations
  - MapReduce not usually done on Virtual Machines

Grid/Cloud	**Authentication and Authorization:** Provide single sign in to both FutureGrid and Commercial Clouds linked by workflow
	Workflow: Support workflows that link job components between FutureGrid and Commercial Clouds. Trident from Microsoft Research is initial candidate
	Data Transport: Transport data between job components on FutureGrid and Commercial Clouds respecting custom storage patterns
	Software as a Service: This concept is shared between Clouds and Grids and can be supported without special attention
	SQL: Relational Database
Cloud	**Program Library:** Store Images and other Program material (basic FutureGrid facility)
	Blob: Basic storage concept similar to Azure Blob or Amazon S3
	DPFS Data Parallel File System: Support of file systems like Google (MapReduce), HDFS (Hadoop) or Cosmos (Dryad) with compute-data affinity optimized for data processing
	Table: Support of Table Data structures modeled on Apache Hbase (Google Bigtable) or Amazon SimpleDB/Azure Table (eg. Scalable distributed “Excel”)
	Queues: Publish Subscribe based queuing system
	Worker Role: This concept is implicitly used in both Amazon and TeraGrid but was first introduced as a high level construct by Azure
	Web Role: This is used in Azure to describe important link to user and can be supported in FutureGrid with a Portal framework
	MapReduce: Support MapReduce Programming model including Hadoop on Linux, Dryad on Windows HPCS and Twister on Windows and Linux

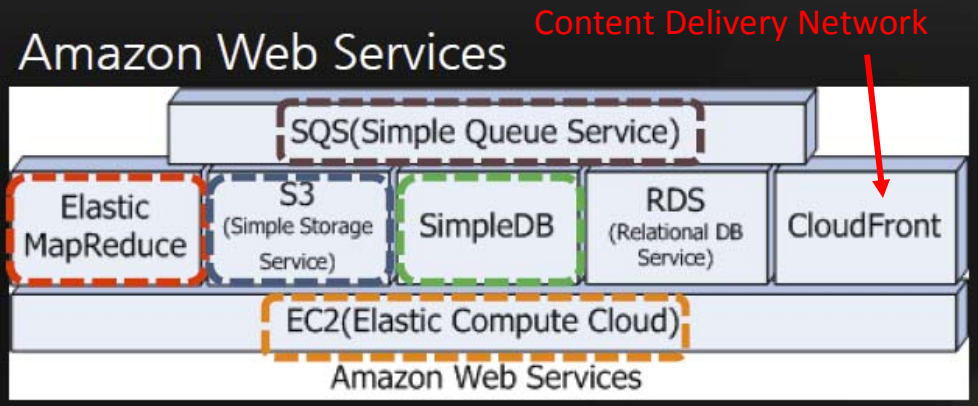


http://cloudonomic.blogspot.com/2009/02/cloud-taxonomy-and-ontology.html

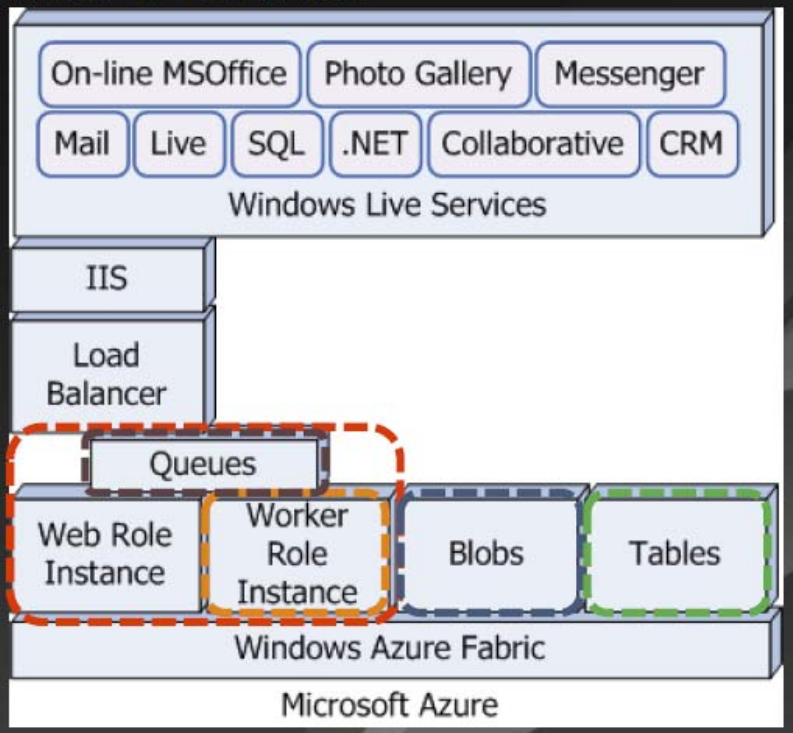
http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing

Architectures of Public Cloud Computing

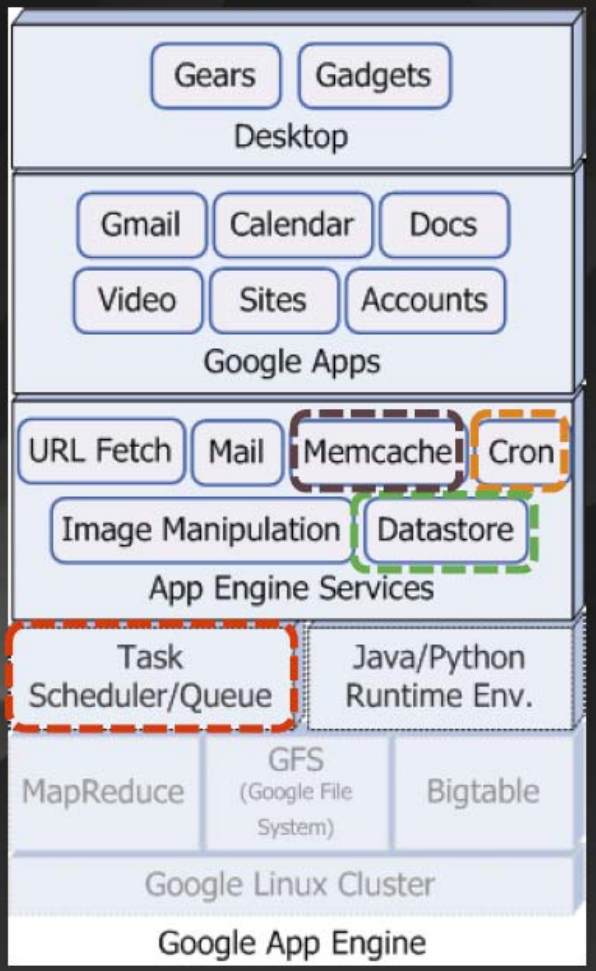
- Amazon Web Services



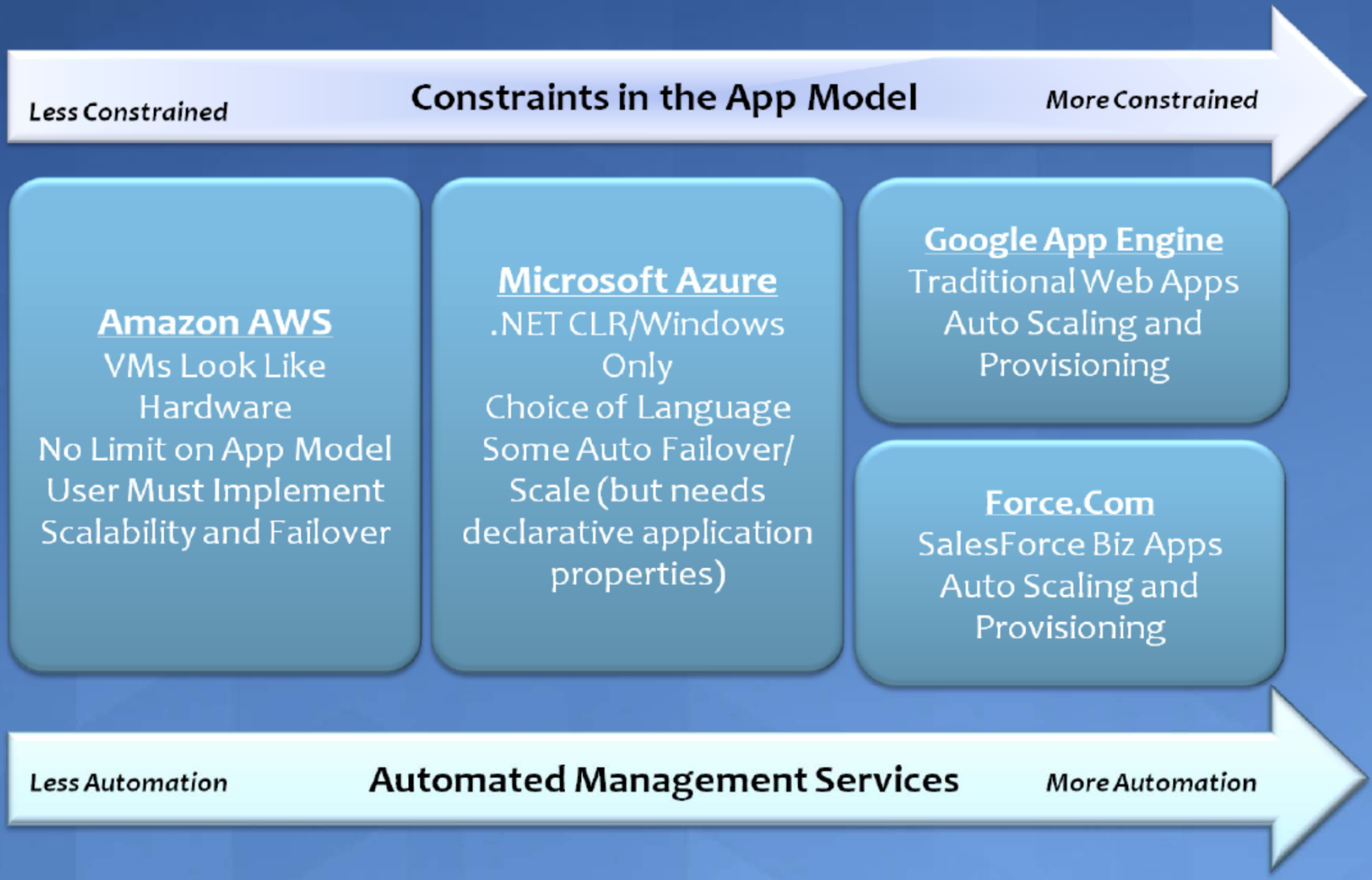
- Microsoft Azure



- Google App Engine



Platform Extension to Cloud is a Continuum



http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Amazon offers a lot! i.e. is PaaS

Data in the Cloud

Cloud Data Models

- Clouds have a unique processing model – MapReduce with Hadoop which we discussed
- They have also inspired a revolution in data models
  - Originally enterprises used databases but
  - Science did not as not efficient and “transactions” not typical use
  - Rather tended to access large chunks of data and/or did not do SQL queries. Rather used custom user software to access events of interest e.g. events like the Higgs
  - LHC explored use of “object oriented” databases (Objectivity) but abandoned
- “Web 2.0” applications had same issues as science
  - Huge data sizes (in fact larger than science) and non SQL queries
- This spurred NOSQL approaches to data and approaches to parallelism

Different formats for data

- Traditional Windows or **UNIX file systems** arranged in a hierarchy with directories
  - Generalize to share files across many computers
  - Access with standard UNIX I/O commands and in growing fashion via web (as in Oncourse or Dropbox)
- **Google File System** – realized in Hadoop File System HDFS – splits files up for parallelism
  - Traditional file systems also had parallel extensions but GFS very explicit
- **Object stores** like Amazon S3 have simpler interface than traditional file systems
- **Databases** will appear on a later page as a higher level management system

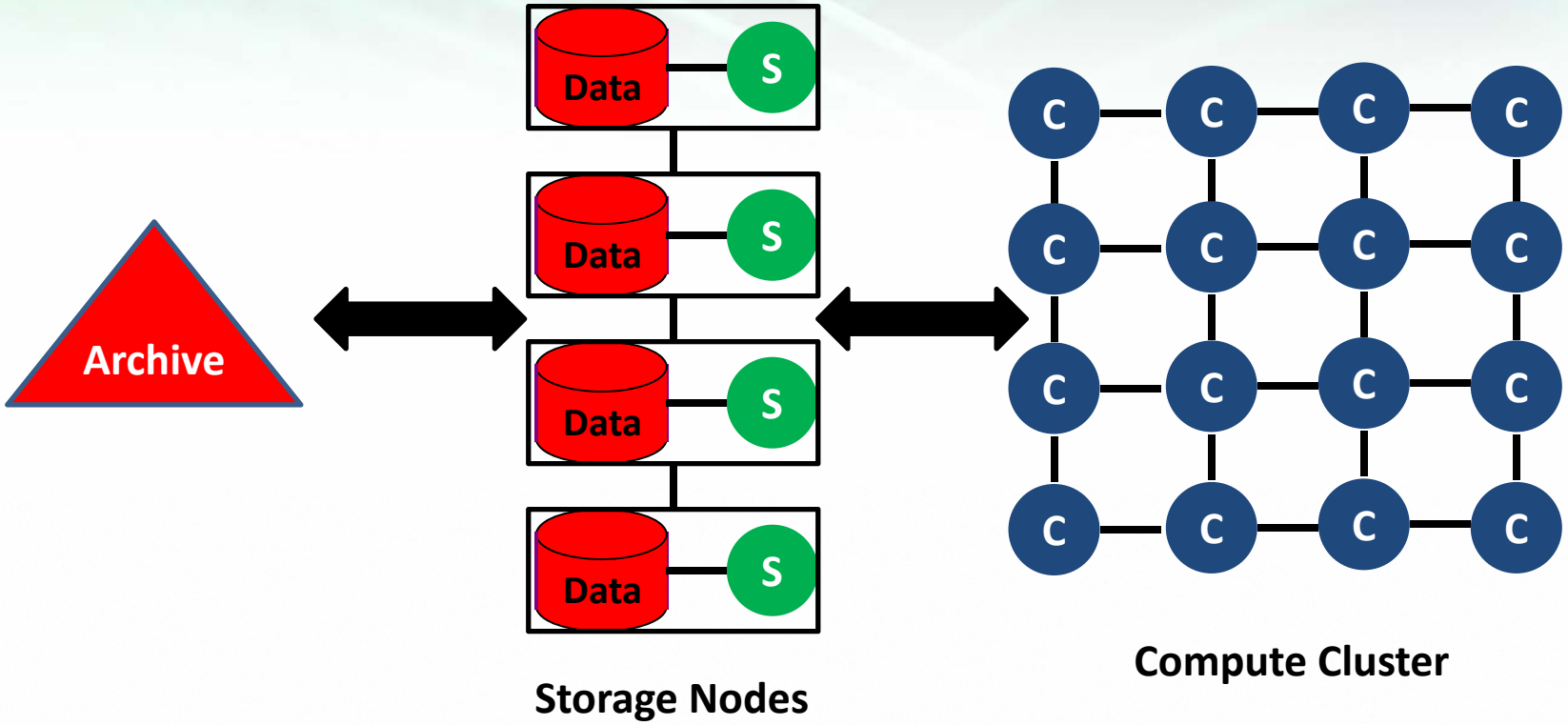
Clouds as Support for Data Repositories?

- The **data deluge** needs cost effective computing
  - Clouds are by definition cheapest
  - Need data and computing co-located
- **Shared resources** essential (to be cost effective and large)
  - Can't have every scientists downloading petabytes to personal cluster
- Need to reconcile **distributed** (initial source of) **data** with shared analysis
  - Can move data to (discipline specific) clouds
  - How do you deal with multi-disciplinary studies
- **Data repositories of future will have cheap data and elastic cloud analysis support?**
  - Hosted free if data can be used commercially?

Architecture of Data Repositories?

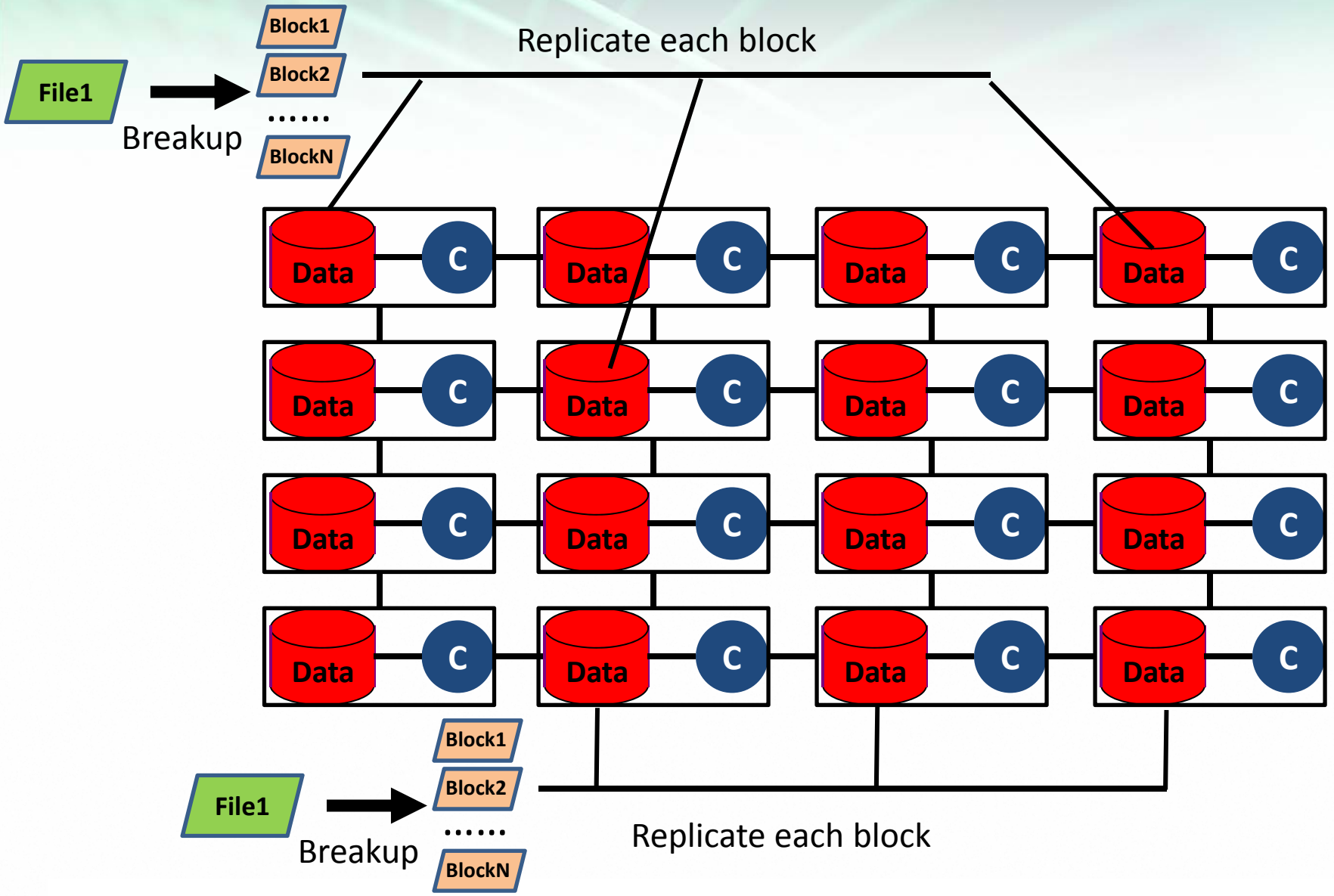
- Traditionally governments set up repositories for data associated with particular missions
  - For example EOSDIS (Earth Observation), GenBank (Genomics), NSIDC (Polar science), IPAC (Infrared astronomy)
  - LHC/OSG computing grids for particle physics
- This is complicated by volume of data deluge, distributed instruments as in gene sequencers (maybe centralize?) and need for intense computing like Blast
  - i.e. **repositories need lots of computing**?

Traditional File System?



- Typically a shared file system (Lustre, NFS …) used to support high performance computing
- Big advantages in flexible computing on shared data but doesn't "**bring computing to data**"
- Object stores similar structure (separate data and compute) to this

Data Parallel File System?



- No archival storage and computing brought to data

Different Approaches to managing Data

- Traditional is Directory structure and file names
  - With an add-on metadata store usually put in a database
  - Contents of file also can be self managed as with XML files
- Also traditional is databases which have tables and indices as core concept
- Newer NOSQL approaches are Hbase, Dynamo, Cassandra, MongoDB, CouchDB,Riak
  - Amazon SimpeDB and Azure table on public clouds
  - These often use an HDFS style storage and store entities in distributed scalable fashion
  - No fixed Schema
  - “Data center” or “web” scale storage
  - Document stores and Column stores depending on decomposition strategy
  - Typically linked to Hadoop for processing