X-Informatics
Cloud Computing Technology Part III

July 5 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

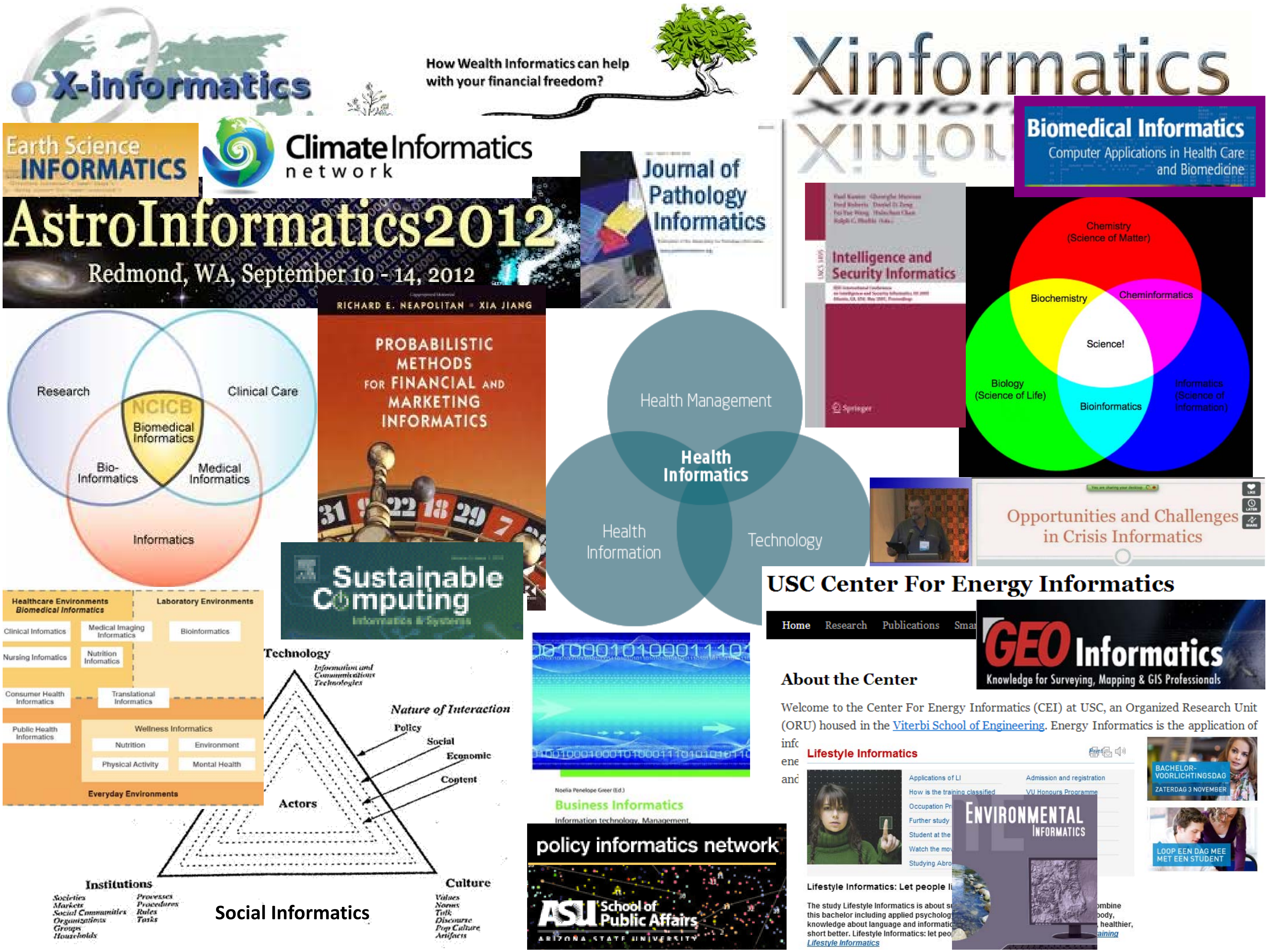
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, 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

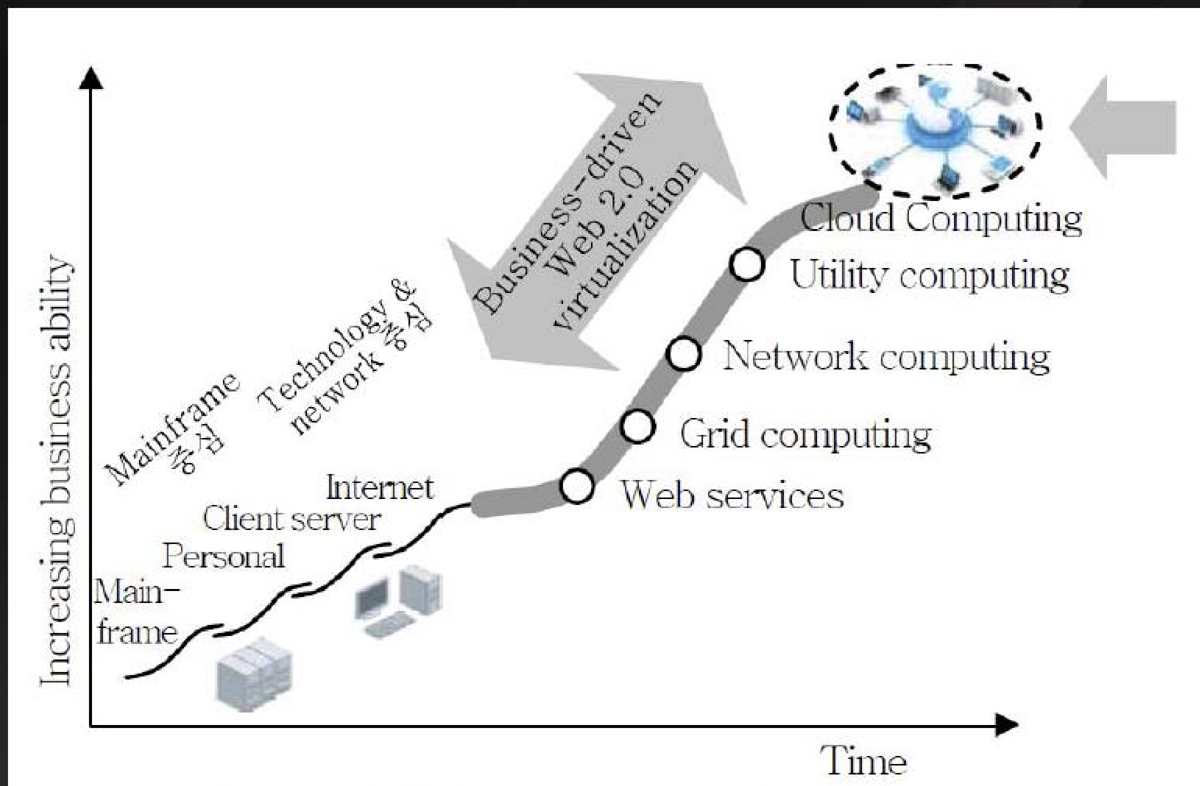
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/


X-informatics
How Wealth Informatics can help with your financial freedom?
Xinformatics
Earth Science INFORMATICS
Climate Informatics network
Biomedical Informatics
Computer Applications in Health Care and Biomedicine
AstroInformatics2012
Redmond, WA, September 10 - 14, 2012
Journal of Pathology Informatics
Intelligence and Security Informatics
Springer
Chemistry (Science of Matter)
Biochemistry
Cheminformatics
Science!
Biology (Science of Life)
Bioinformatics
Informatics (Science of Information)
Research
Clinical Care
NCICB
Biomedical Informatics
Bio-Informatics
Medical Informatics
Informatics
RICHARD E. NEAPOLITAN · XIA JIANG
PROBABILISTIC METHODS FOR FINANCIAL AND MARKETING INFORMATICS
Health Management
Health Informatics
Health Information
Technology
Opportunities and Challenges in Crisis Informatics
Sustainable Computing
Informatics & Systems
USC Center For Energy Informatics
Home
Research
Publications
GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals
About the Center
Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the Viterbi School of Engineering. Energy Informatics is the application of
Healthcare Environments
Biomedical Informatics
Laboratory Environments
Clinical Infomatics
Medical Imaging Informatics
Bioinformatics
Nursing Infomatics
Nutrition Infomatics
Consumer Health Informatics
Translational Informatics
Public Health Informatics
Wellness Informatics
Nutrition
Environment
Physical Activity
Mental Health
Everyday Environments
Technology
Information and Communications Technologies
Nature of Interaction
Policy
Social
Economic
Content
Actors
Institutions
Societies
Markets
Social Communities
Organizations
Groups
Households
Processes
Procedures
Rules
Tasks
Culture
Values
Norms
Talk
Discourse
Pop Culture
Artifacts
Social Informatics
Business Informatics
policy informatics network
ASU School of Public Affairs
ARIZONA STATE UNIVERSITY
Lifestyle Informatics
Applications of LI
Admission and registration
Lifestyle Informatics: Let people li
ENVIRONMENTAL INFORMATICS
BACHELOR-VOORLICHTINGSDAG
ZATERDAG 3 NOVEMBER
LOOP EEN DAG MEE MET EEN STUDENT

Cloud (Data Center) Architectures

Evolution of Computing Environments

Cloud Computing is NOT a brand-new revolution



Stolen from Trends in Technology of Cloud Computing, ETRI 2009.08

But why now? then not?

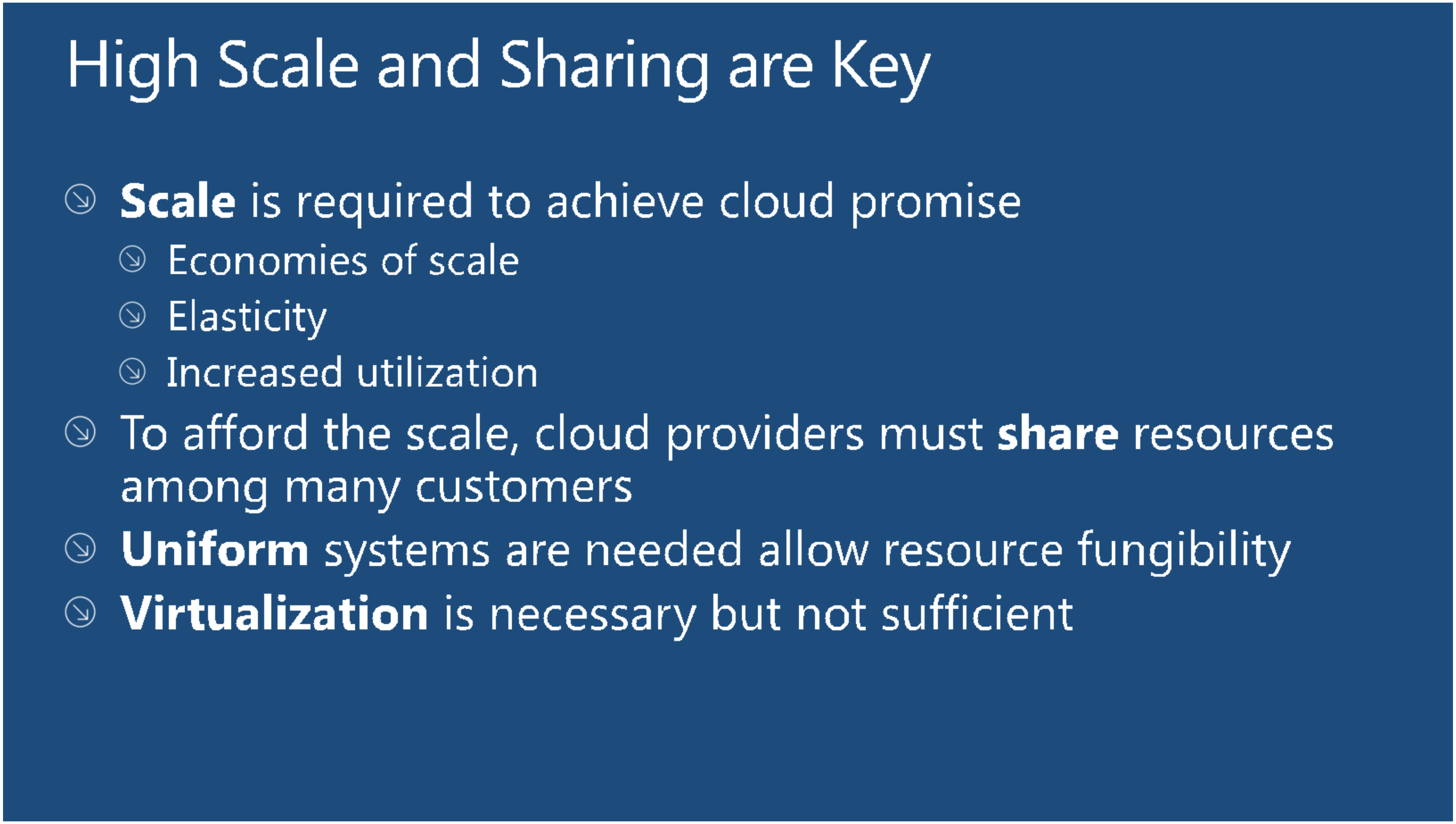
http://www.slideshare.net/woorung/trend-and-future-of-cloud-computing

Amazon making money

- It took Amazon Web Services (AWS) eight years to hit $650 million in revenue, according to Citigroup in 2010.
- Just three years later, Macquarie Capital analyst Ben Schachter estimates that AWS will top $3.8 billion in 2013 revenue, up from $2.1 billion in 2012 (estimated), valuing the AWS business at $19 billion.
- It's a lot of money, and it underlines Amazon's increasingly dominant role in cloud computing, and the rising risks associated with enterprises putting all their eggs in the AWS basket.

Over time, the cloud will replace company-owned data centers

- That is what **Adam Selipsky of Amazon feels**. He says it may not happen overnight, it may take 5, 10 or even 20 years, but it will happen over time.
- According to Amazon, clouds enable 7 transformation of how applications are designed, built and used.
  - Cloud makes distributed architectures easy
  - Cloud enables users to embrace the security advantages of shared systems
  - Cloud enables enterprises to move from scaling by architecture to scaling by command
  - Cloud puts a supercomputer into the hands of every developer
  - Cloud enables users to experiment often and fail quickly
  - Cloud enables big data without big servers
  - Cloud enables a mobile ecosystem for a mobile-first world
- http://www.eweek.com/c/a/Cloud-Computing/AWS-Innovation-Means-Cloud-Domination-307831/



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

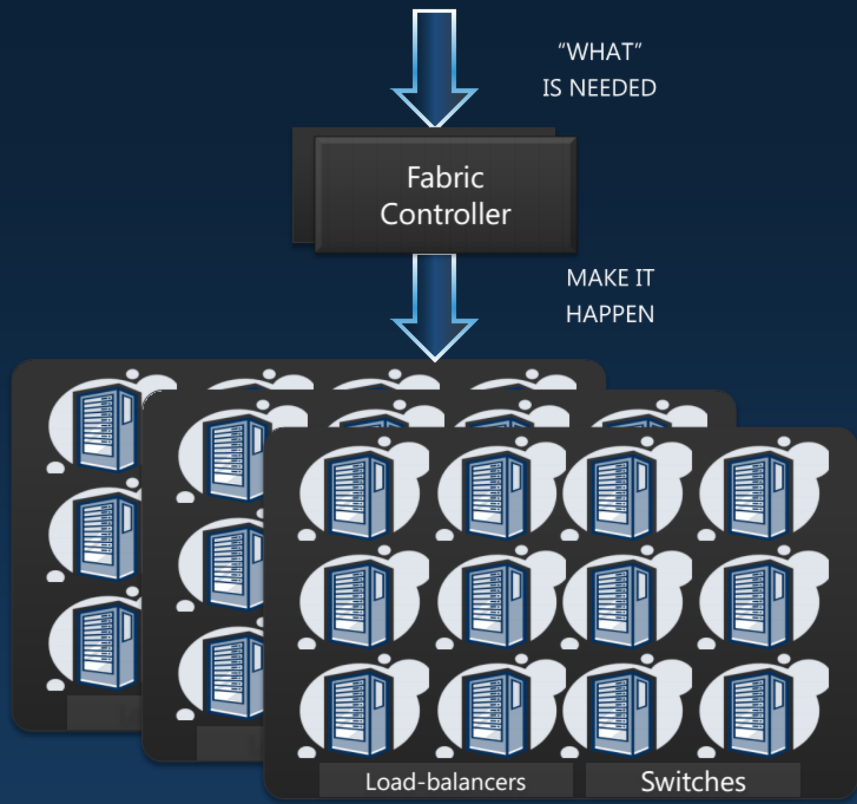
Computing At Scale: Windows Azure

- Redundancy at multiple levels
- Virtualized compute and network

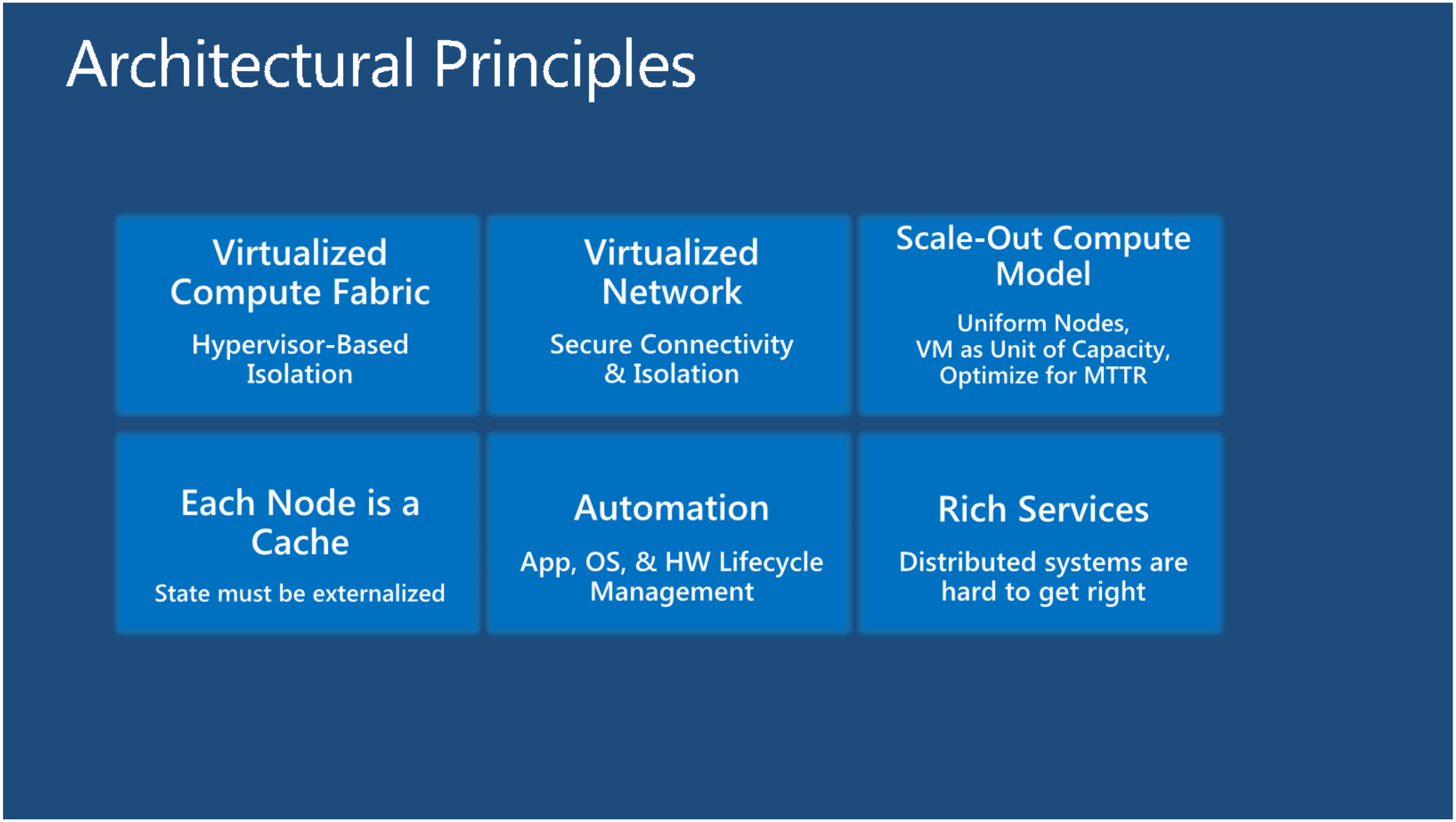
- Model-driven automation
  - Logical application graphs mapped over physical topology

- Development and Management tools
  - REST APIs, multiple language and tool support
  - Web and enterprise management tools

- Rich services - built as Azure apps
  - Storage, SQL Azure, etc.
  - Secure cloud federation services



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



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

Applying Architectural Principles in Azure

Virtualized Compute Fabric
HyperV based virtualization

Virtualized Network
Logical networks over physical network

Scale-Out Compute Model
Fixed set of VM/mem/bw sizes, all components optimized for MTTR

Each Node is a Cache
Durable network drives, local drives used as a cache

Automation
Model-driven automation of sw + hw, provisioning, configuration, and health

Rich Services
Blob, table, & database services, queues, caching, identity, …

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

The Challenge of Data Centers & Apps

- The impact on the environment
  - In 2006 data centers used 61 *Tera*watt-hours of power
    - 1.5 to 3% of US electrical energy consumption today
  - Great advances are underway in power reduction
- With 100K+ servers and apps that must run 24x7 constant failure must be an axiom of hardware and software design.
  - Huge implication for the application design model.
  - How can hardware be designed to degrade gracefully?
- Two dimensions of parallelism
  - Scaling apps from 1 to 1,000,000 simultaneous users
  - Some apps require massive parallelism to satisfy a single request in less than a second.

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

Microsoft

The Microsoft Cloud is Built on Data Centers

~100 Globally Distributed Data Centers

Range in size from "edge" facilities to megascale (100K to 1M servers)

Quincy, WA

eneration 4 DCs

Data Centers Clouds & Economies of Scale I

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

2 Google warehouses of computers on the banks of the Columbia River, in The Dalles, Oregon

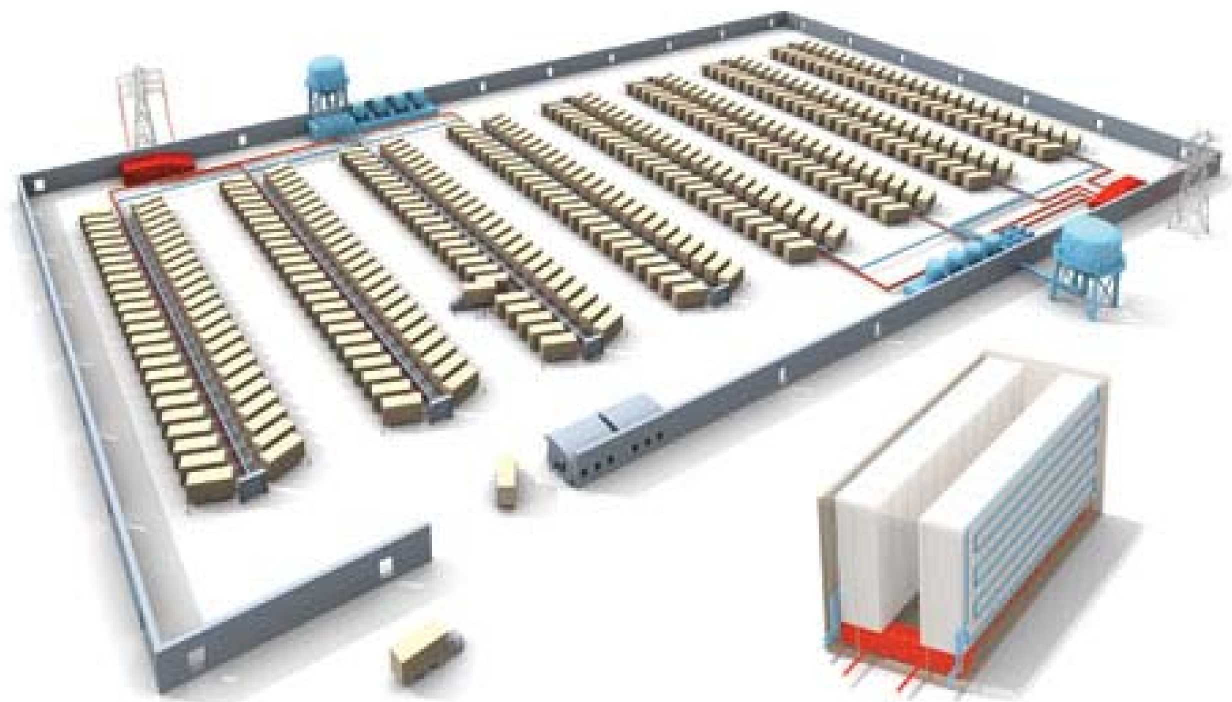
Such centers use 20MW-200MW each with 150 watts per CPU

Save money from large size, positioning with cheap power and access with Internet

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

Data Centers, Clouds & Economies of Scale II

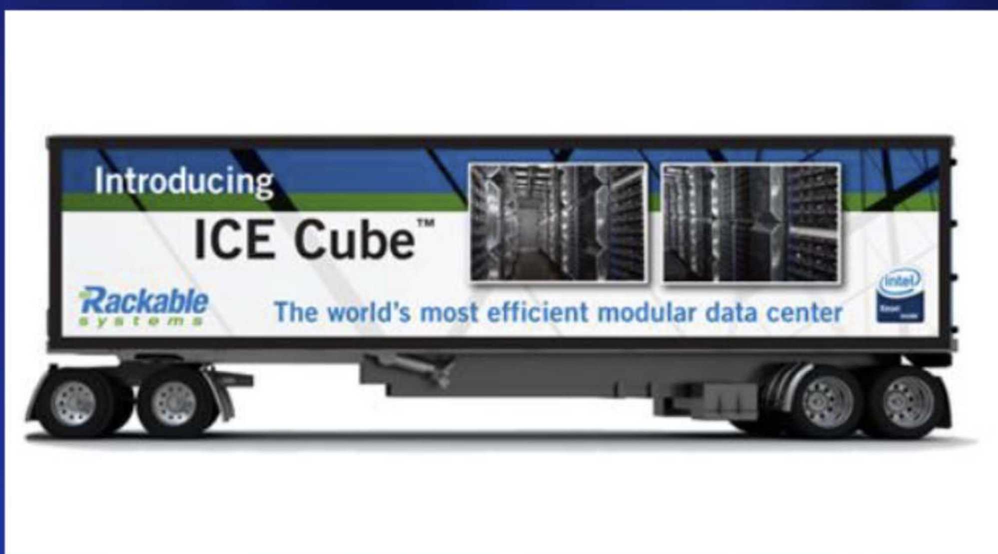
- Builds giant data centers with 100,000's of computers; ~ 200-1000 to a shipping container with Internet access
- "Microsoft will cram between 150 and 220 shipping containers filled with data center gear into a new 500,000 square foot Chicago facility. This move marks the most significant, public use of the shipping container systems popularized by the likes of Sun Microsystems and Rackable Systems to date."



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

Advances in DC deployment

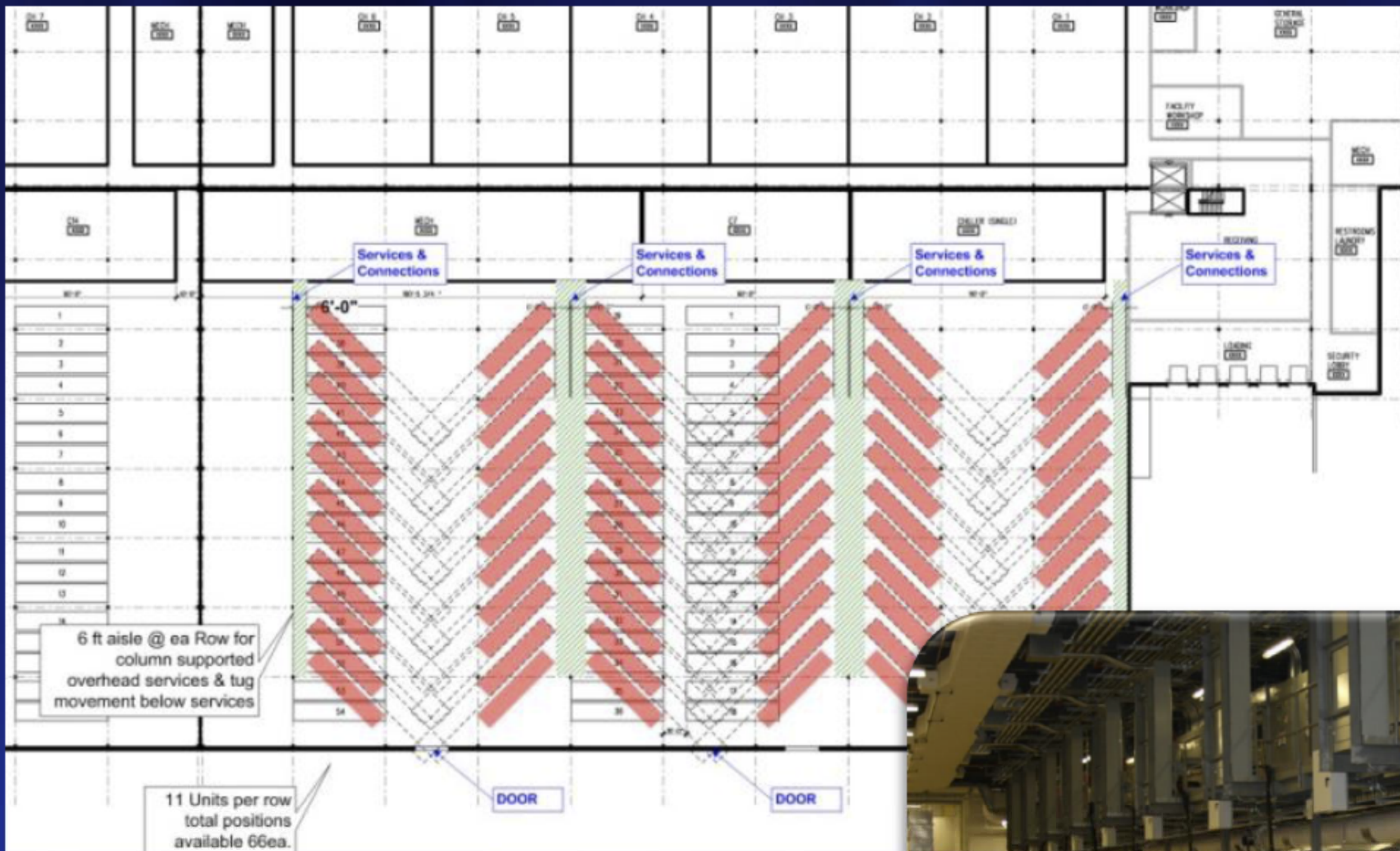
- Conquering complexity.
  - Building racks of servers & complex cooling systems all separately is not efficient.
  - Package and deploy into bigger units:



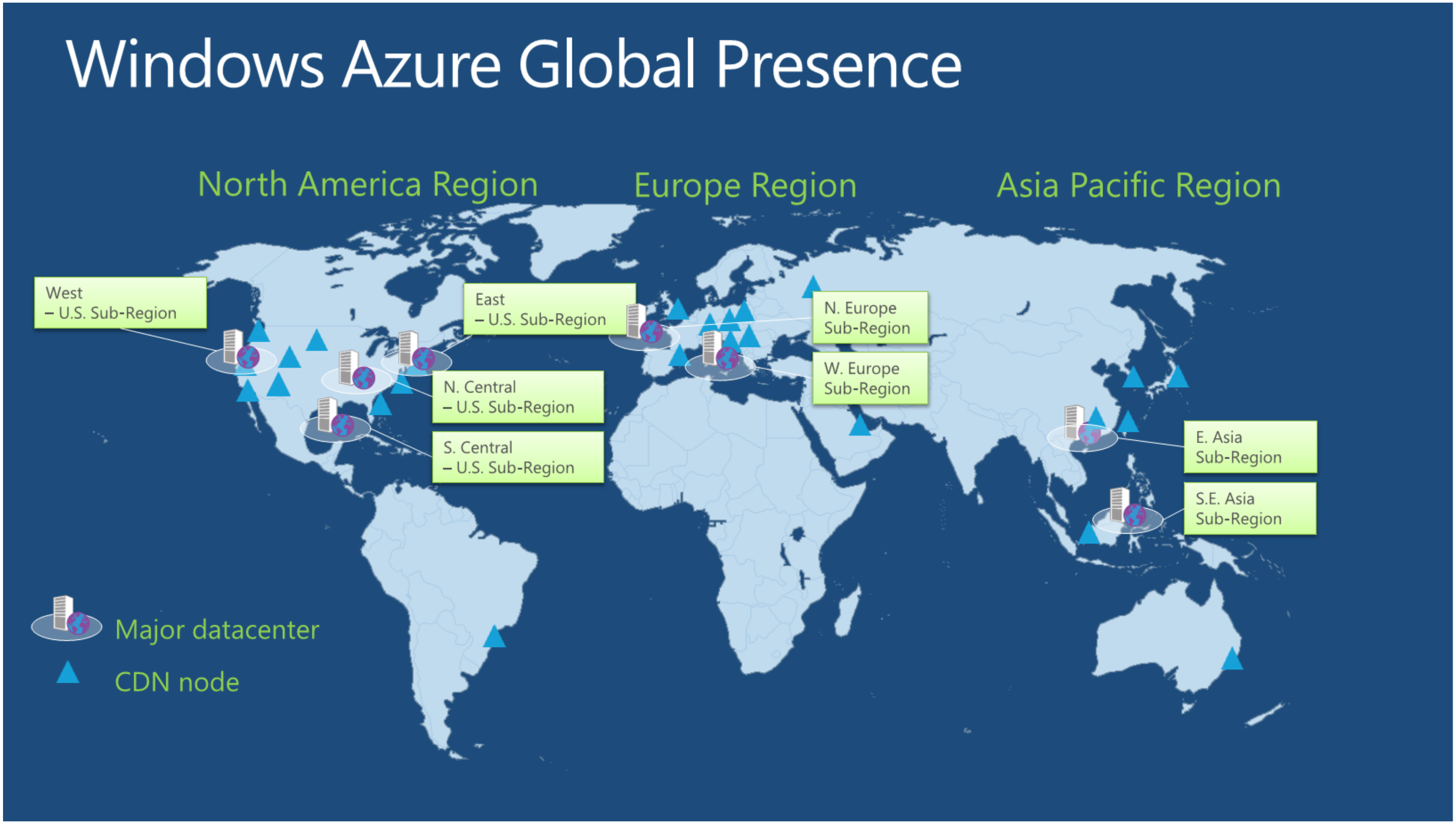
http://research.microsoft.com/en-us/people/barga/sc09_cloudcomp_tutorial.pdf

Microsoft

Containers: Separating Concers



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



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

Green Clouds

- Cloud Centers optimize life cycle costs and power use

$$\text{PUE} = \frac{\text{Total facility power}}{\text{IT equipment power}}$$

- http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
- Average PUE = 1.8 (was nearer 3) ; Good Clouds are 1.1-1.2
- 4th generation data centers (from Microsoft) make everything modular so data centers can be built incrementally as in modern manufacturing
- http://loosebolts.wordpress.com/2008/12/02/our-vision-for-generation-4-modular-data-centers-one-way-of-getting-it-just-right/
- Extends container based third generation

Some Sizes in 2010

- http://www.mediafire.com/file/zzqna34282frr2f/koomeydatacenterelectuse2011finalversion.pdf
- 30 million servers worldwide
- Google had 900,000 servers (3% total world wide)
- Google total power ~200 Megawatts
  - < 1% of total power used in data centers (Google more efficient than average – **Clouds are Green**!)
  - ~ 0.01% of total power used on anything world wide
- Maybe total clouds are 20% total world server count (a growing fraction)

Some Sizes Cloud v HPC

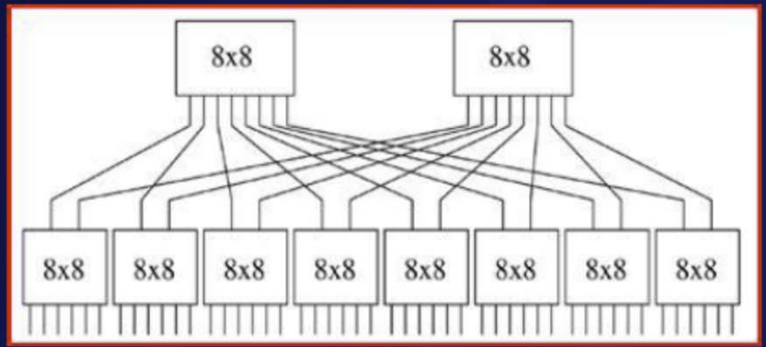
- **Top Supercomputer** Sequoia Blue Gene Q at LLNL
  - 16.32 Petaflop/s on the Linpack benchmark using 98,304 CPU compute chips with 1.6 million processor cores and 1.6 Petabyte of memory in 96 racks covering an area of about 3,000 square feet
  - 7.9 Megawatts power
- **Largest (cloud)** computing data centers
  - 100,000 servers at ~200 watts per CPU chip
  - Up to 30 Megawatts power
  - Microsoft says upto million servers
- So **largest supercomputer** is around **1-2% performance of total cloud computing systems** with Google ~20% total

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

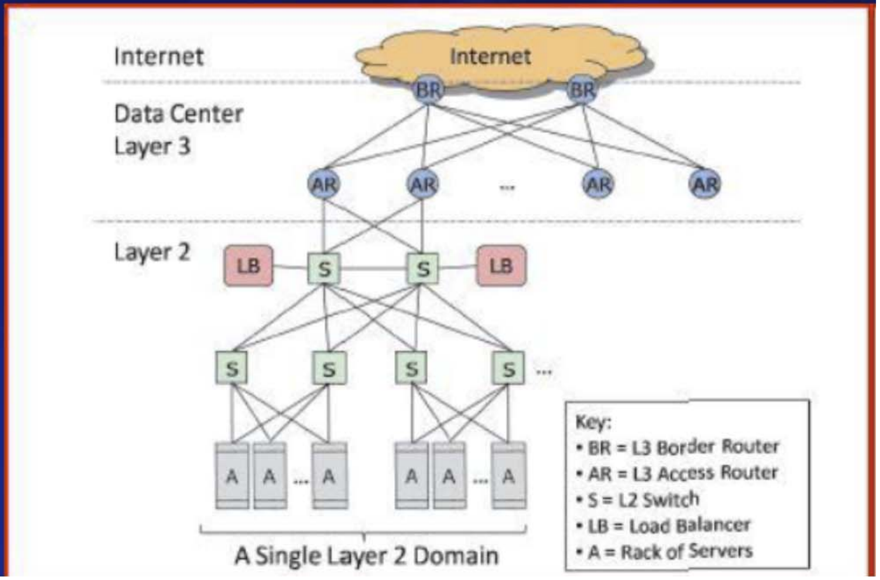
Data Center vs Supercomputers

- Scale
  - Blue Waters = 40K 8-core "servers"
  - Road Runner = 13K cell + 6K AMD servers
  - MS Chicago Data Center = 50 containers = 100K 8-core servers.
- Network Architecture
  - Supercomputers: CLOS "Fat Tree" infiniband
    - Low latency – high bandwidth
    - protocols
  - Data Center: IP based
    - Optimized for Internet Access
- Data Storage
  - Supers: separate data farm
    - GPFS or other parallel file system
  - DCs: use disk on node + memcache

Fat tree network



Standard Data Center Network



Cloud Industry Players

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

Players: Providers

FZI

Programmatic access via Web Services and/or Web APIs

"Pure" virtualized resources

- CPU, memory, storage, and bandwidth
- Data store

amazon webservices™

Joyent Cloud Computing Brought to you with the power of the Joyent Accelerator™

versus

Virtualized resources plus application framework (e.g., RoR, Python, .NET)

- Imposes an application and data architecture
- Constrains how application is built

Google App Engine

Live Mesh TECH PREVIEW

[MM]

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

Players: Cloud Intermediaires

FZI

- Resells (aspects of) raw cloud resources, with added value propositions
  - Packaging resources as bundles
  - Facilitating cloud resource management,
    - e.g., setup, updates, backup, load balancing, etc.
  - Providing tools and dashboards
- Enabler of the cloud ecosystem

zimory

RightScale™

morph™ Simplify Innovation

enomaly

elastra

Engine Yard™

BETA heroku

cohesiveFT assemble. deploy. manage

[MM]

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

Players: Application Providers

- Software as a Service (SaaS):
  - Applications provided and consumed over the Web
- Infrastructure usage (mostly) hidden

[MM]

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

Cloud Talk

- *The Private Cloud "Big Four"*
  - ***IBM***
  - ***VMware***
  - ***Sun/Oracle***
  - ***3Tera***

johnmwillis.com

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

Cloud Applications

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.

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

Cloud Talk

Success Stories

- *Animoto*
  - ***25k customers to 750k in one week***
  - ***40 servers to 500 servers in one week***
  - ***No system administrators***
- *NY Times*
  - ***Convert 11 million files in one night***
  - ***4 TB's of data***
  - ***Total cost $240***
- *Eli Lilly*
  - ***10 weeks to get a server now 5 minutes***
  - ***Amazon has redefined "Time" at Eli Lilly***

johnmwillis.com

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

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

Success Cases in Cloud Computing

SmugMug(http://www.smugmug.com/)

- an online photo storage application that stores more than half a petabyte of data on S3
- estimates cost savings on service and storage to be close to $1 million

New York Times(http://www.nytimes.com)

- use EC2 to process terabytes of archival data using hundreds of EC2 instances within 36 hours

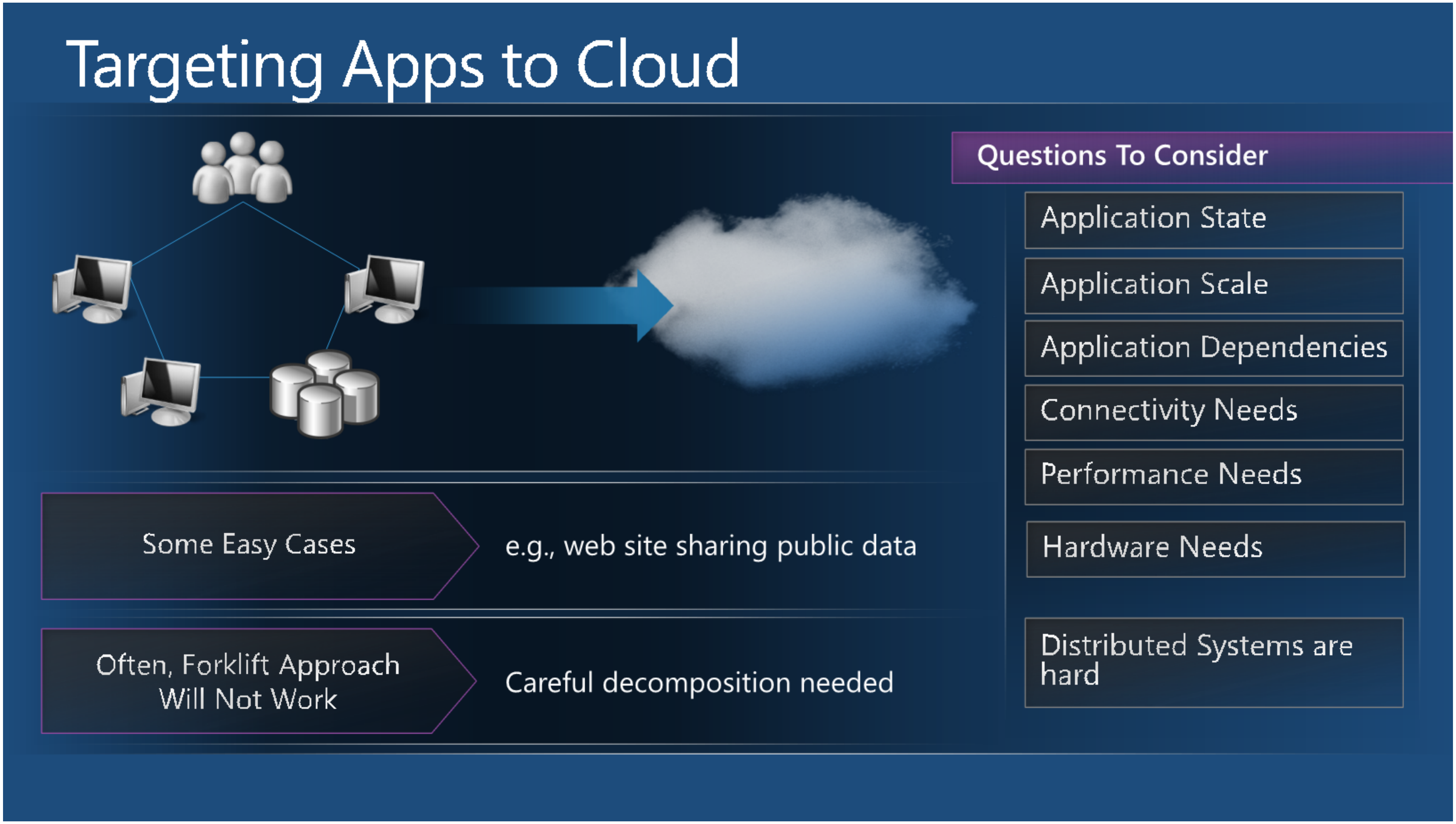
37Signals(http://37signals.com/)

- maker of popular online project-management software Basecamp, uses S3 for storage needs.

Animoto(http://animoto.com/)

- an online presentation video generator that needs gobs of computing power for video processing
- recently successfully withstood a surge in Web traffic that would kill most companies' systems by scaling up their processing power quickly using EC2 with RightScale
  - Animoto ramped from 25,000 users to 250,000 users in three days, signing up 20,000 new users per hour at peak
  - Using RightScale, EC2 instances automatically scaled out 40 to 4000 at that time
  - For more detail, refer to http://blog.rightscale.com/2008/04/23/animoto-facebook-scale-up/

coOrd

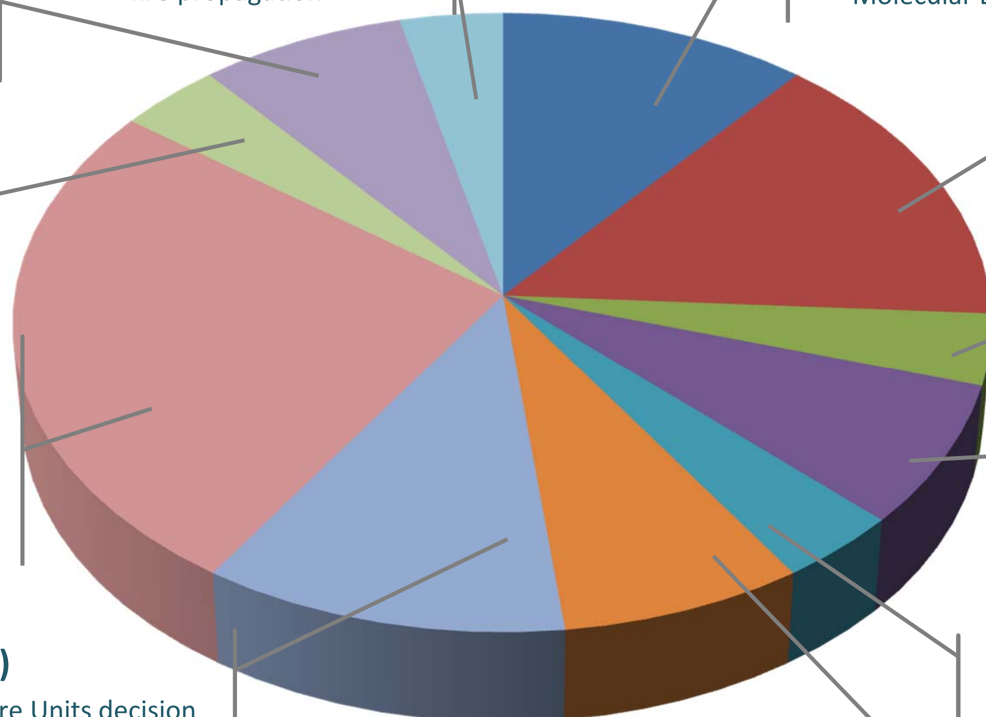


http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/tuesday/Keynote_OpportunitiesAndChallenges_Yousef_Khalidi.pdf

What Applications work in Clouds

- **Pleasingly (moving to modestly) parallel** applications of all sorts with roughly independent data or spawning independent simulations
  - **Long tail** of science and integration of distributed sensors
- **Commercial and Science Data analytics** that can use MapReduce **(**some of such apps) or its **iterative** variants (most other data analytics apps)
- **Which science applications are using clouds**?
  - **Venus-C** (Azure in Europe): 27 applications **not using** Scheduler, Workflow or MapReduce (except roll your own)
  - 50% of applications on **FutureGrid** are from Life Science
  - Locally **Lilly** corporation is commercial cloud user (for drug discovery) but not IU Biology
- **But overall very little science use of clouds**

27 Venus-C Azure Applications



Chemistry (3)
- Lead Optimization in Drug Discovery
- Molecular Docking

Civil Protection (1)
- Fire Risk estimation and fire propagation

Biodiversity & Biology (2)
- Biodiversity maps in marine species
- Gait simulation

Civil Eng. and Arch. (4)
- Structural Analysis
- Building information Management
- Energy Efficiency in Buildings
- Soil structure simulation

Physics (1)
- Simulation of Galaxies configuration

Earth Sciences (1)
- Seismic propagation

Mol, Cell. & Gen. Bio. (7)
- Genomic sequence analysis
- RNA prediction and analysis
- System Biology
- Loci Mapping
- Micro-arrays quality.

ICT (2)
- Logistics and vehicle routing
- Social networks analysis

Medicine (3)
- Intensive Care Units decision support.
- IM Radiotherapy planning.
- Brain Imaging

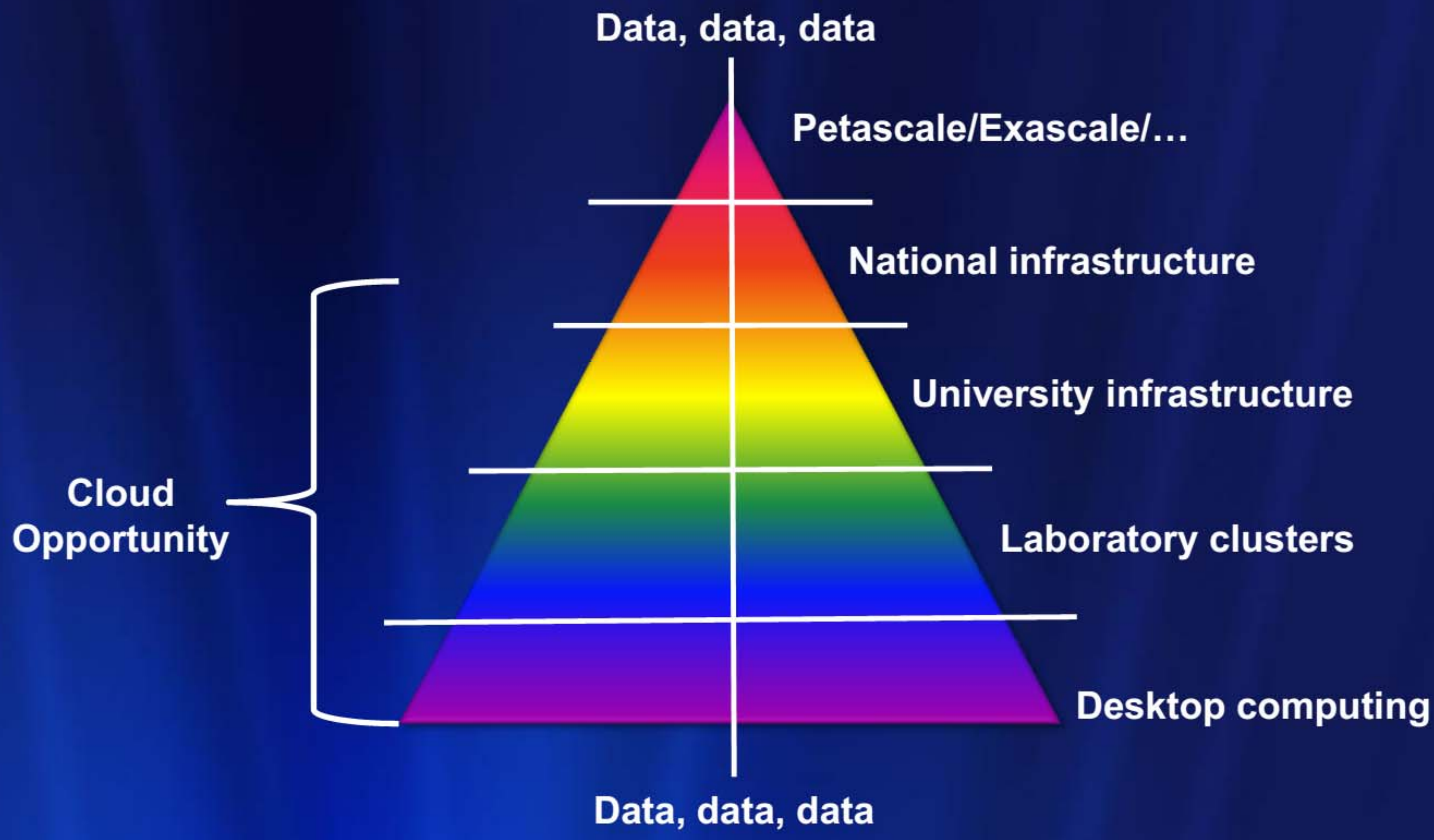
Mathematics (1)
- Computational Algebra

Mech, Naval & Aero. Eng. (2)
- Vessels monitoring
- Bevel gear manufacturing simulation

Parallelism over Users and Usages

- "**Long tail of science**" can be an important usage mode of clouds.
- In some areas like particle physics and astronomy, i.e. "**big science**", there are just a few major instruments generating now petascale data driving discovery in a coordinated fashion.
- In other areas such as genomics and environmental science, there are many **"individual" researchers** with distributed collection and analysis of data whose total data and processing needs can match the size of big science.
- **Clouds** can provide scaling convenient resources for this important aspect of science.
- Can be **map only** use of MapReduce if different usages naturally linked e.g. exploring docking of multiple chemicals or alignment of multiple DNA sequences
  - Collecting together or summarizing multiple "maps" is a **simple Reduction**

The Computing Research Pyramid



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

Classic Parallel Computing

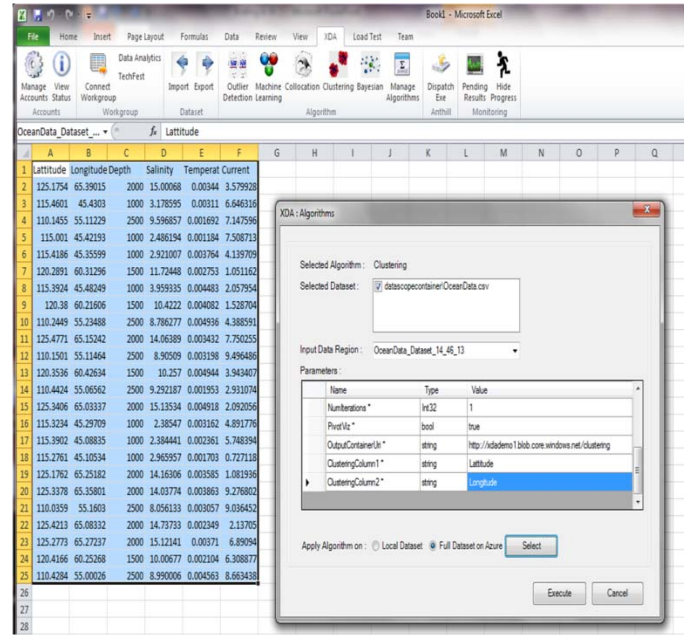
- **HPC:** Typically SPMD (Single Program Multiple Data) “maps” typically processing particles or mesh points interspersed with multitude of low latency messages supported by specialized networks such as Infiniband and technologies like MPI
  - Often run large capability jobs with 100K (going to 1.5M) cores on same job
  - National DoE/NSF/NASA facilities run 100% utilization
  - Fault fragile and cannot tolerate “outlier maps” taking longer than others
- **Clouds:** MapReduce has asynchronous maps typically processing data points with results saved to disk. Final reduce phase integrates results from different maps
  - Fault tolerant and does not require map synchronization
  - **Map only** useful special case
- **HPC + Clouds:** Iterative MapReduce caches results between “MapReduce” steps and supports SPMD parallel computing with large messages as seen in parallel kernels (linear algebra) in clustering and other data mining

Data Intensive Applications

- **Applications** tend to be **new** and so can consider **emerging technologies** such as clouds
- Do not have lots of small *messages* but rather **large reduction** (aka Collective) operations
  - **New optimizations** e.g. for huge messages
- **EM (expectation maximization**) tends to be **good** for **clouds** and **Iterative MapReduce**
  - Quite **complicated computations** (so compute largish compared to communicate)
  - Communication is **Reduction** operations (global sums or linear algebra in our case)
- We looked at **Clustering** and **Multidimensional Scaling using deterministic annealing** which are both EM
  - See also **Latent Dirichlet Allocation** and related Information Retrieval algorithms with similar EM structure

Excel DataScope

Cloud Scale Data Analytics from Excel



Bringing the power of the cloud to the laptop

- **Data sharing in the cloud**, with annotations to facilitate discovery and reuse;
- **Sample and manipulate** extremely large data collections in the cloud;
- **Top 25 data analytics algorithms**, through Excel ribbon running on Azure;
- **Invoke models**, perform analytics and visualization to gain insight from data;
- **Machine learning** over large data sets to discover correlations;
- **Publish** data collections and visualizations to the cloud to share insights;

Researchers use familiar tools, **familiar but differentiated**.

CSTI Meeting. October 2012 Dennis Gannon

Security

Security, Trust, Regulations

- Application isolation and data privacy
  - Over shared network and compute fabrics
- Identity and access management
- Policy controls for data & applications
  - Do you know where your data reside?

- Independently verified compliance
  - Audit certifications, e.g., ISO/IEC 27001:2005

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

Security

- Compute isolation
  - Hypervisor, Root OS, Guest VMs
  - Guest VMs are untrusted
- Network isolation
  - Packet filtering at network edge
- Control software
  - Fabric Controllers isolation
  - SSL mutual auth for internal traffic
  - Encrypted store for platform secrets

Authentication mechanisms

Subjects	Objects	Auth mechanism
Customers	Subscription (account)	Windows Live ID
Developers & Operators	Web portal REST APIs	Live ID (portal) Self-signed certificate
Applications	Storage	Storage account key
External applications	Azure apps	Customer-defined Platform support for single-sign on and federated identity scenarios

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

Comments on Fault Tolerance and Synchronicity Constraints

Service Level Guarantees

- Availability
  - Typically expressed as percentage of total operations to succeed
  - Calculated over some fixed period
  - Example
    - "We guarantee that at least 99.9% of the time we will successfully process correctly formatted requests that we receive to add, update, read and delete data"
- Performance
  - Typically expressed as minimum (or a range) of memory, network bandwidth, or VM computing power
  - Applications want end-to-end guarantees

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

Monthly Service Level Agreement

Compute connectivity	Instance monitoring and restart	Storage availability	Database availability	Service bus and access control availability
• Your service is connected and reachable via web • Internternet facing roles will have external connectivity	• All running roles will be continuously monitored • If role is not running, we will detect and initiate corrective state	• Storage service will be available/reachable (connectivity) • Your storage requests will be processed successfully	• Database is connected to the internet gateway • All databases will be continuously monitored	• Service bus and access control endpoints will have external connectivity • Message operation requests processed successfully
>99.95%	>99.9%	>99.9%	>99.9%	>99.9%

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

Fault Tolerance

- The basic simple approach to fault tolerance is replication
- Amazon basic S3 designed for 99.999999999% durability and 99.99% availability of objects over a given year.
  - Done by replicating each bit of storage ~6 times
- Amazon Reduced Redundancy Storage (RRS) has 99.99% durability and same availability as basic
  - Implies that average expected loss of 0.01% of objects in a year
  - 33% cheaper
  - Perhaps stored ~4 times
  - Use if can regenerate data

Fault Tolerance of Programs

- Data is relatively easy to make static but programs are harder
  - One can of course save the images but what's hard is saving running programs
- Programs can stop but you need to be certain that no messages in system from other programs (processes)
- Safest is save "state" (data defining program execution) on disk in such a way that program can read data to restart
  - This is MapReduce strategy where processes communicate via disk
  - Unfortunately doing messaging via disk is slow
- In parallel processing different processes are correlated; when one breaks the others have to wait
  - Thus cost of a failure multiplied by number of parallel processes (up to a million today)
  - So fault tolerance strategy for loosely coupled processes (typical cloud scenario where recovery approach clear so can allow faults) different from that for tightly coupled case (HPC case where more effort made to avoid faults)

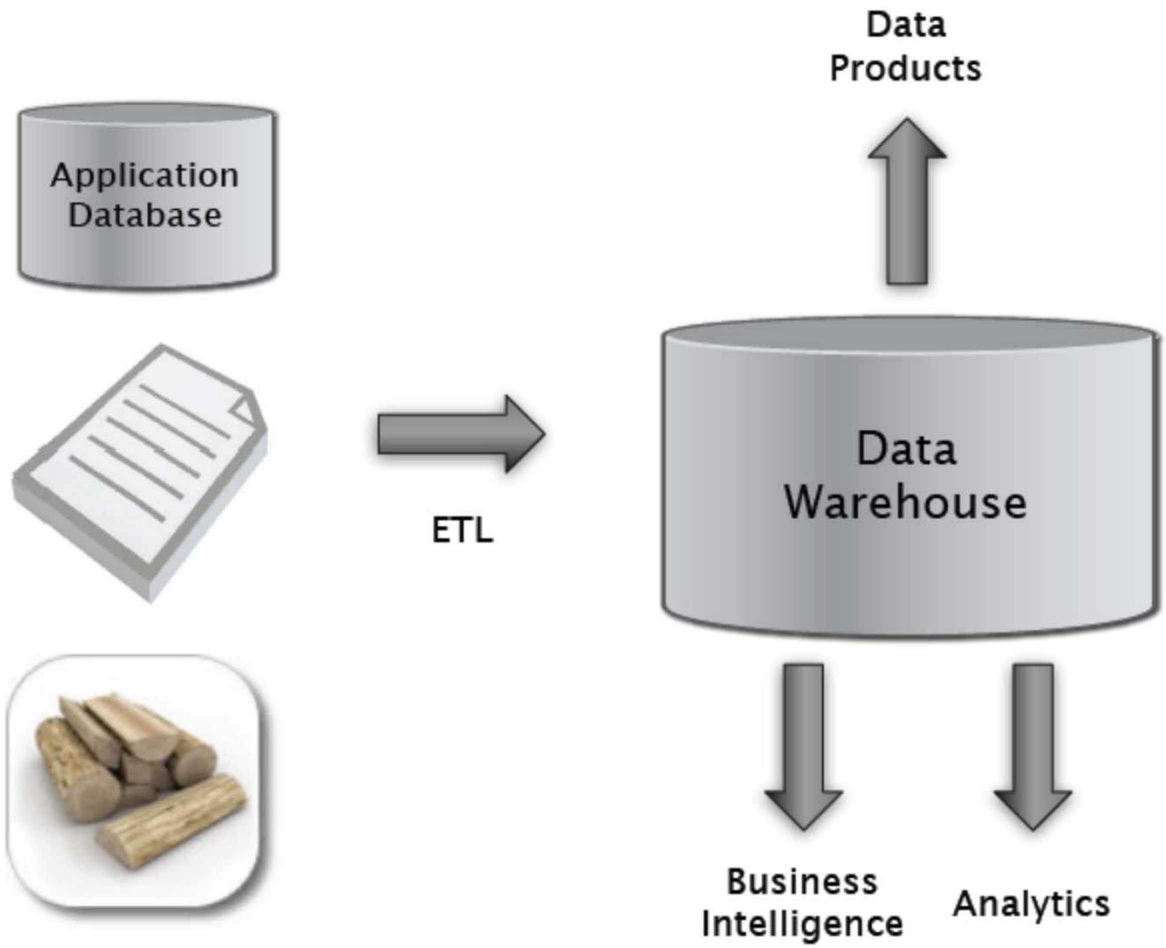
Application Fault Tolerance

- Typical commercial applications can afford to lose data
  - If you do a recommendation and miss out 0.1% of ratings, it will make essentially no difference as you are averaging over all the data (in a clever way)
- On the other hand if you are doing a parallel weather simulation decomposed geometrically and the process in charge of Martinsville breaks then that will eventually cause process to break
  - Weather information propagates out from Martinsville and will reach Bloomington after an hour or so.
  - Another reason why classic HPC is more sensitive to faults

Big Data Processing

From Application Perspective

Technology discussed earlier



http://berkeleydatascience.files.wordpress.com/2012/01/20120119berkeley.pdf
Jeff Hammerbacher

"Taming the Big Data Tidal Wave" 2012
(Bill Franks, Chief Analytics Officer Teradata) Applications

- **Web Data** ("the original big data")
  - Analyze customer web browsing of e-commerce site to see topics looked at etc.
- **Auto Insurance** (telematics monitoring driving)
  - Equip cars with sensors
- **Text data in multiple industries**
  - Sentiment analysis, identify common issues (as in eBay lamp example), Natural Language processing
- **Time and location (GPS) data**
  - Track trucks (delivery), vehicles(track), people(tell them nearby goodies)
- Retail and manufacturing: **RFID**
  - Asset and inventory management,
- Utility industry: **Smart Grid**
  - Sensors allow dynamic optimization of power
- Gaming industry: **Casino Chip tracking (RFID)**
  - Track individual players, detect fraud, identify patterns
- **Industrial engines and equipment**: sensor data
  - See GE engine
- **Video games: telemetry**
  - This is like monitoring web browsing but rather monitor actions in a game
- Telecommunication and other industries: **Social Network data**
  - Connections make this big data.
  - Use connections to find new customers with similar interests

“Taming the Big Data Tidal Wave” 2012 (Bill Franks, Chief Analytics Officer Teradata) Technology

- Parallel Computing, Clouds, Grids, MapReduce,
- Sandbox for separate standalone Analytics experimentation

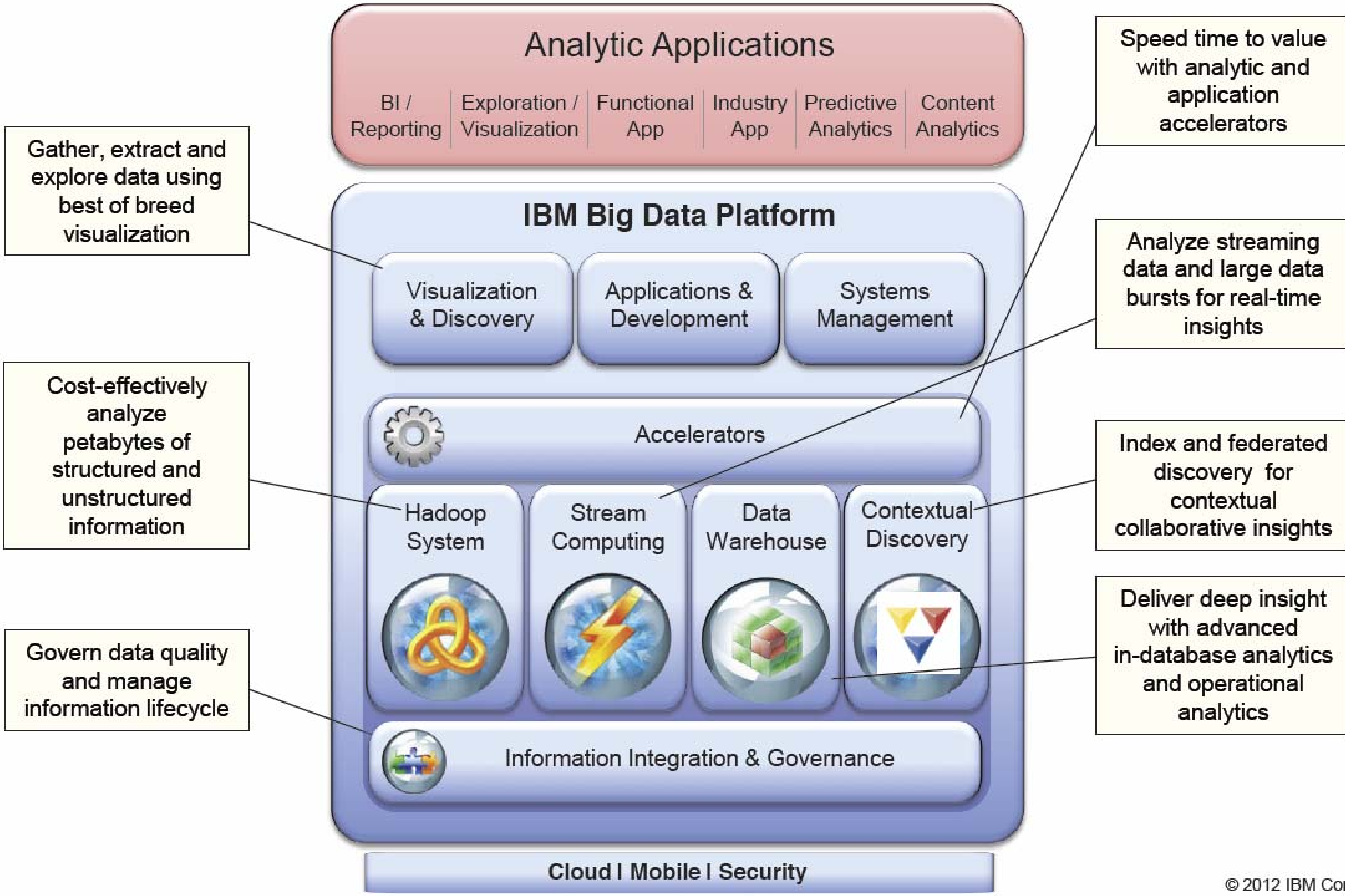
Anjul Bhambhri, VP of Big Data, IBM http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

More Mission-Critical Apps Ride on Big Data Platforms

- Integrate and manage the full ***variety***, ***velocity*** and ***volume*** of data
- Apply ***advanced analytics*** to information in its ***native*** form
- Visualize all available data for ***ad-hoc analysis and discovery***
- Development environment for ***building new analytic applications***
- Integration and deploy applications with enterprise grade ***availability***, ***manageability***, ***security***, and ***performance***

Anjul Bhambhri, VP of Big Data, IBM http://fisheritcenter.haas.berkeley.edu/Big_Data/index.html

Big Data Platform and Application Framework



Big Data Platform – Internet Scale Analytics

Open source components

Platform Capabilities

- Built-in analytics
  - Text analytics engine, annotators, Eclipse tooling
  - Interface to project R (statistical platform)
- Deep integration with enterprise software stack
- Analytical tool for analysts
- Ready-made business process accelerators
- Integrated installation of supported open source and other components
- Web Console for admin and application access
- Platform enrichment: additional security, performance features, . . .
- World-class support
- Full open source compatibility

Business benefits

- Quicker time-to-value due to IBM technology and support
- Reduced operational risk
- Enhanced business knowledge with flexible analytical platform
- Leverages and complements existing software

Deep Analytics Appliance – Revolutionizing Analytics

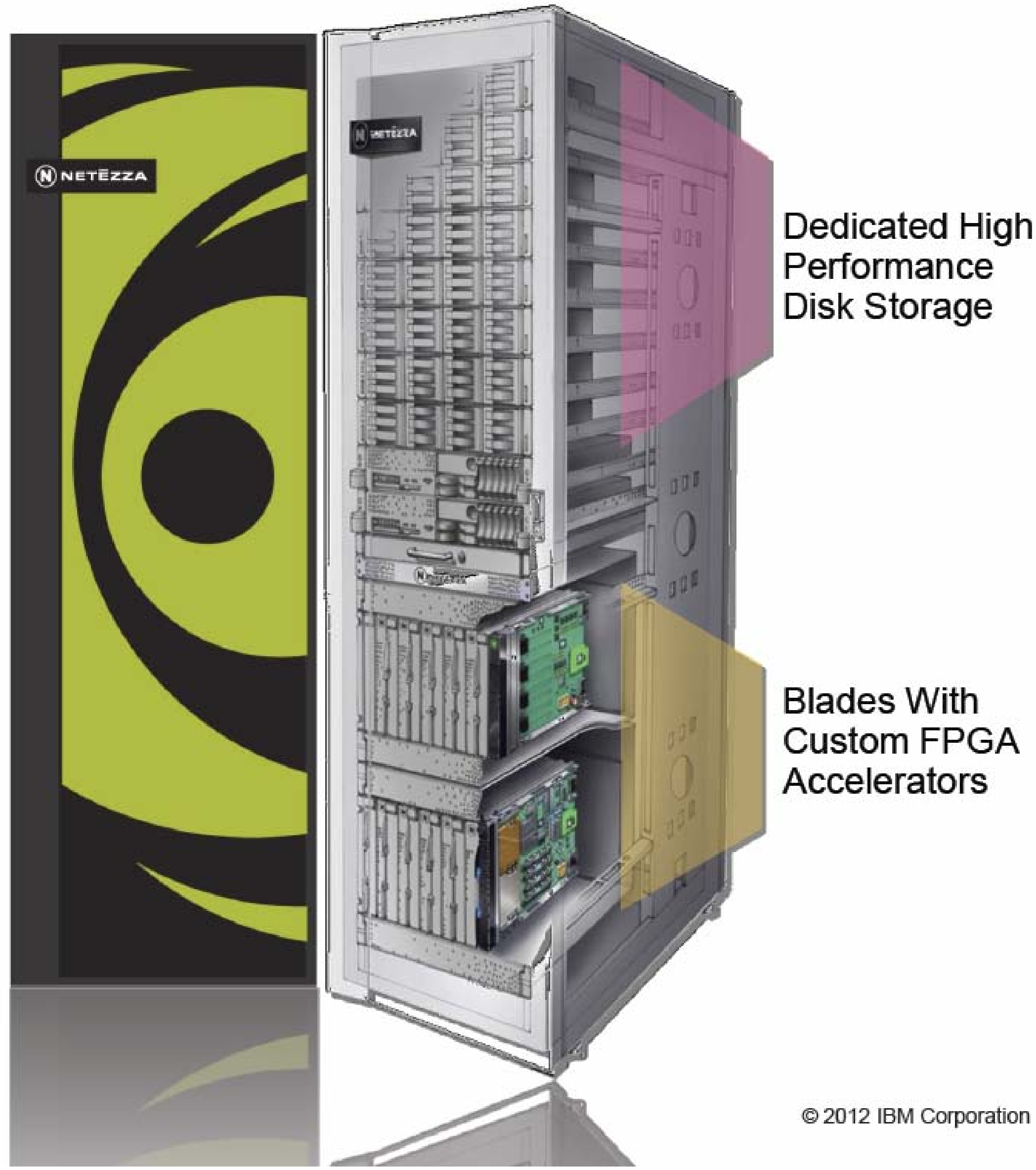
Purpose-built analytics appliance

Speed: 10-100x faster than traditional systems

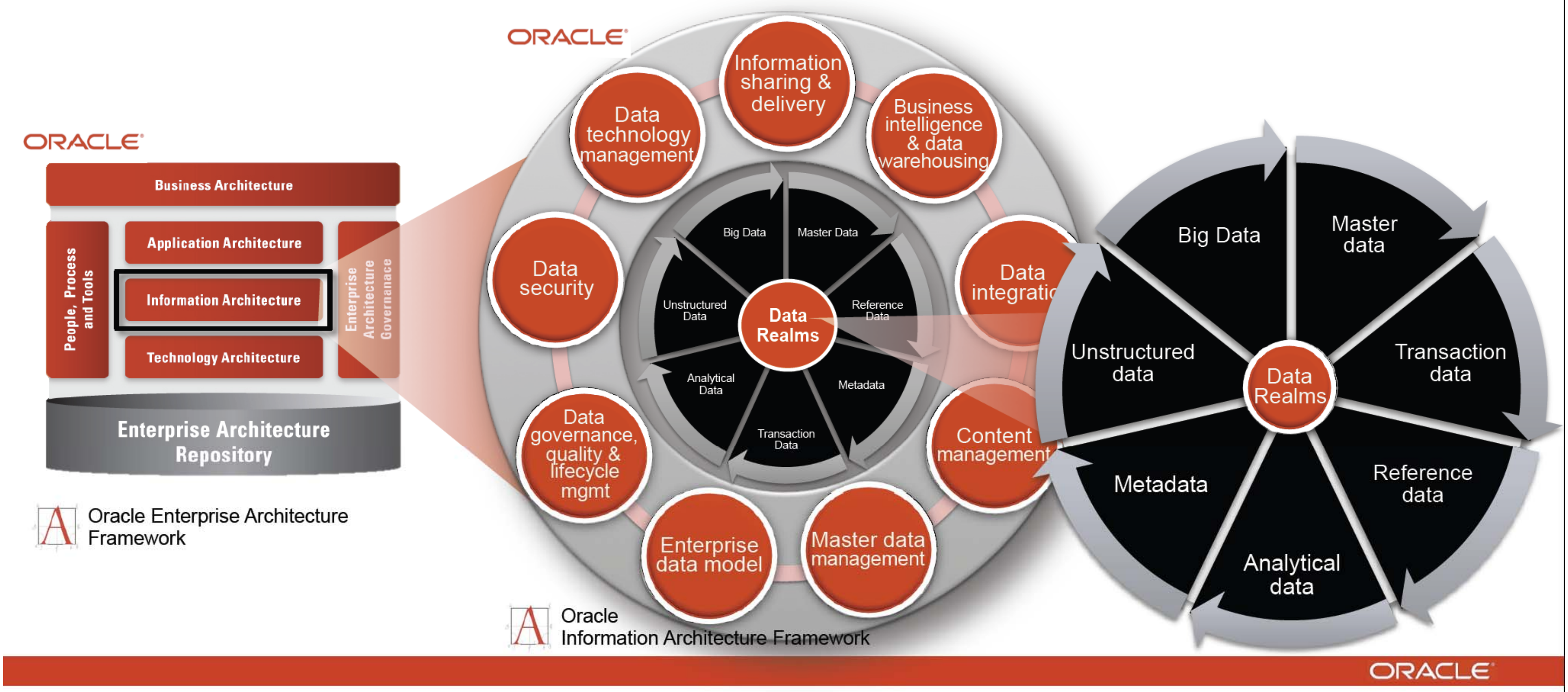
Simplicity: Minimal administration and tuning

Scalability: Peta-scale user data capacity

Smart: High-performance advanced analytics



Information Architecture Capability Model



http://cs.metrostate.edu/~sbd/slides/Sun.pdf Oracle

http://fisheritcenter.haas.berkeley.edu/Big_Data/Williams.pdf

ONE SIZE DOESN'T FIT ALL

- There isn't one solution for driving an organization with big data:
  - **Hadoop** is for:
    - Engineers, batch (asynchronous), map reduce (divide and conquer), unstructured, flexible problems
  - **HBase** is for:
    - Engineers, real-time, large data blob, unstructured, key lookup, flexible problems
  - **Teradata** (or another data warehousing solution) is for:
    - Analysts, real-time or batch, structured, flexible problems
  - **Cassandra** (or **MongoDB** or …) is for:
    - Engineers, real-time, smaller data blob, unstructured, key lookup, flexible problems
  - Some problems warrant specialized solutions

http://fisheritcenter.haas.berkeley.edu/Big_Data/Williams.pdf