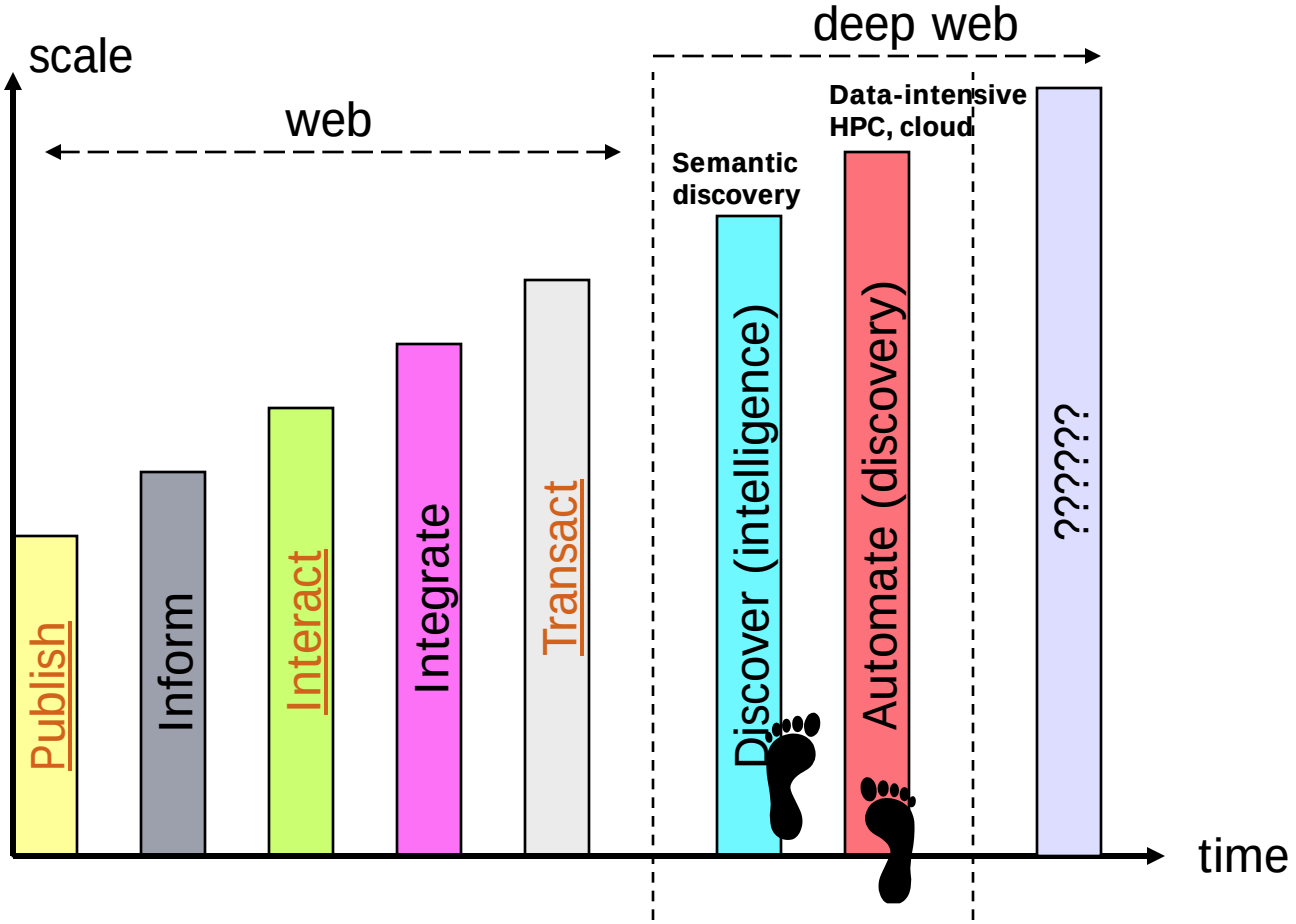
Cloud : The Next Generation Computer

Bina Ramamurthy
Computer Science and Engineering Department
University at Buffalo (UB)
This work is partially funded by grant NSF-DUE-CCLI-Phase II: 0920335
bina@buffalo.edu

Generations

- Vacuum tube computer
- Transistors based computer
- Integrated circuits
- Main frame
- Minicomputer
- Microcomputer
- Desk tops, PCs, Laptops, Palmtops
- Internet and the web
- Mobile devices (cell phone and PDA)
- Cloud computing

Evolution of Internet Computing



Evolution

- Industrial
  - Automation
  - Just-in-time
  - Advances in supply chain
  - …
- Business
  - Remote operations
  - Heterogeneity
  - Scale
  - Integration (application, data)
  - E-commerce
  - …

Evolution (contd.)

- Information technology
  - Internet
  - World-wide web
  - Grid
  - Mobile and wireless
  - Devices
  - Software, platforms
  - Search engines
  - Tremendous advances in application: it is an app-world

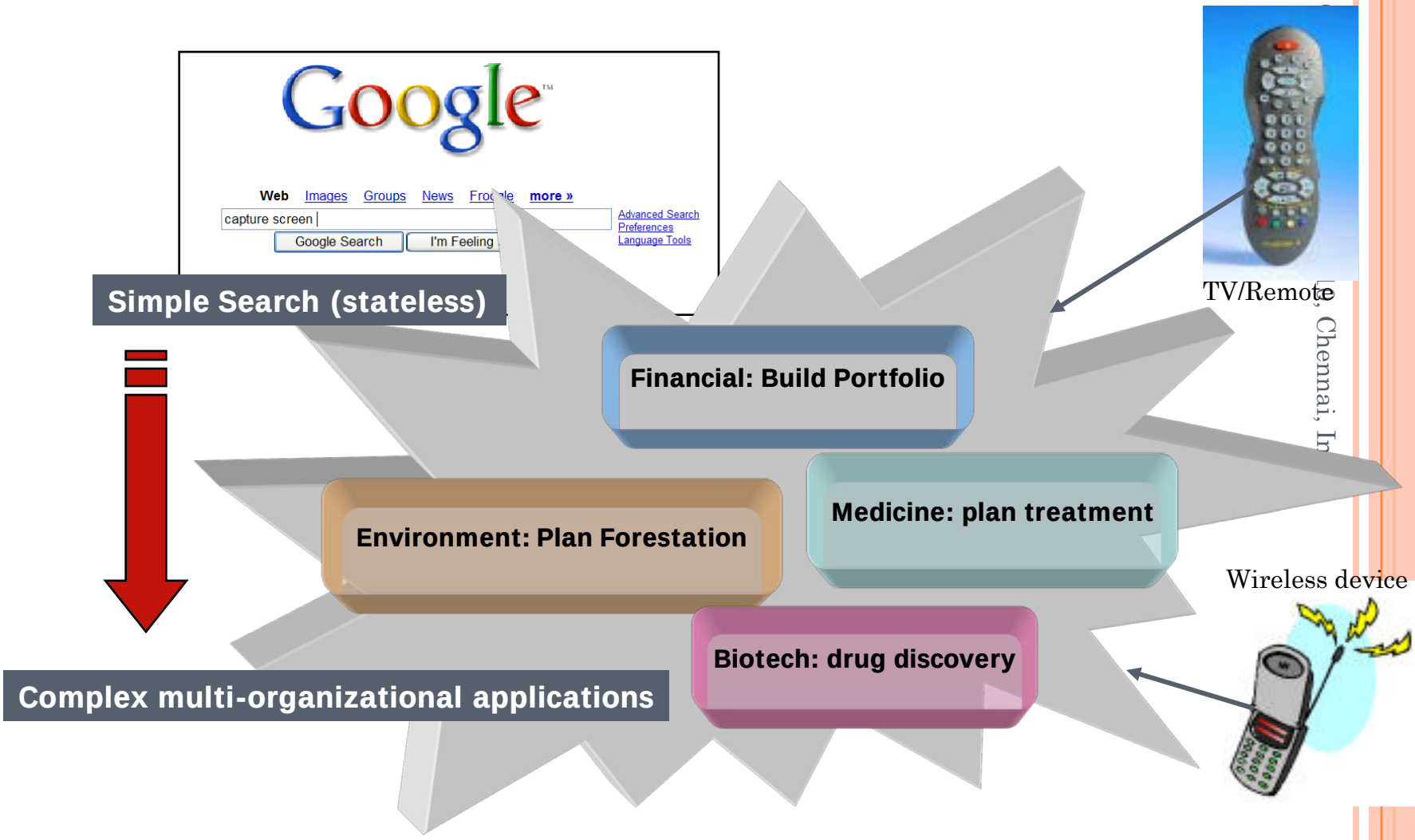
Evolution (contd.)

- Computing research
  - Programming languages
  - RISC vs. CISC architectures
  - Memory capacity
  - Computing power
  - Simple program→Object →Component→…
- Environment
  - Accessibility
  - Globalization (outsourcing, markets)
  - …

Evolution (contd.)

- Society
  - IT users not exclusive to Computer Science
  - Digital media
  - ipod, iphone, idog,..
  - Youtube, myspace, social networking
  - Blogs,wikies, podcasts
  - Facebook, orkut, twitter

Beyond Search Engines: Enabling Information Technology and Scientific Applications

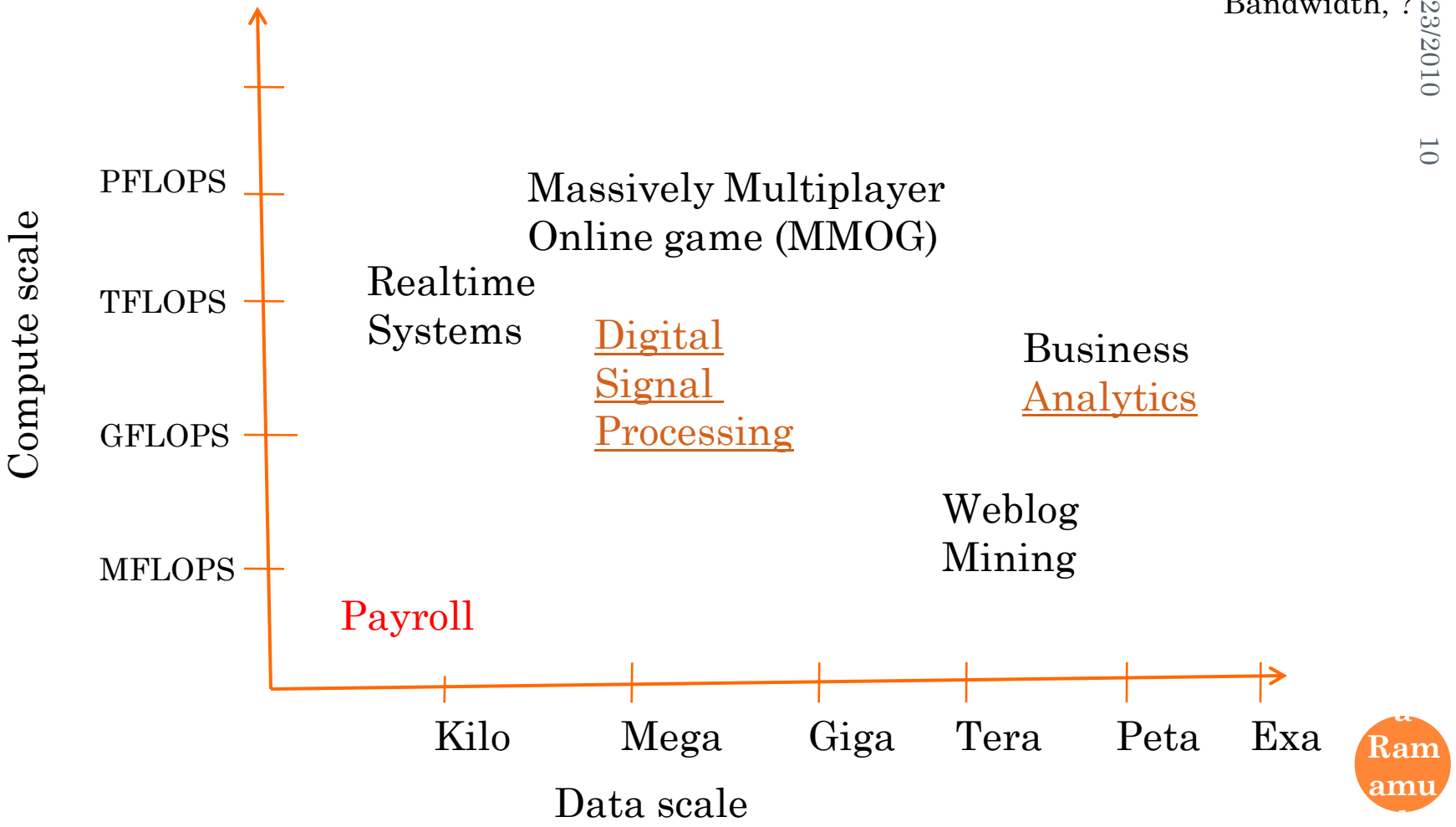


Background

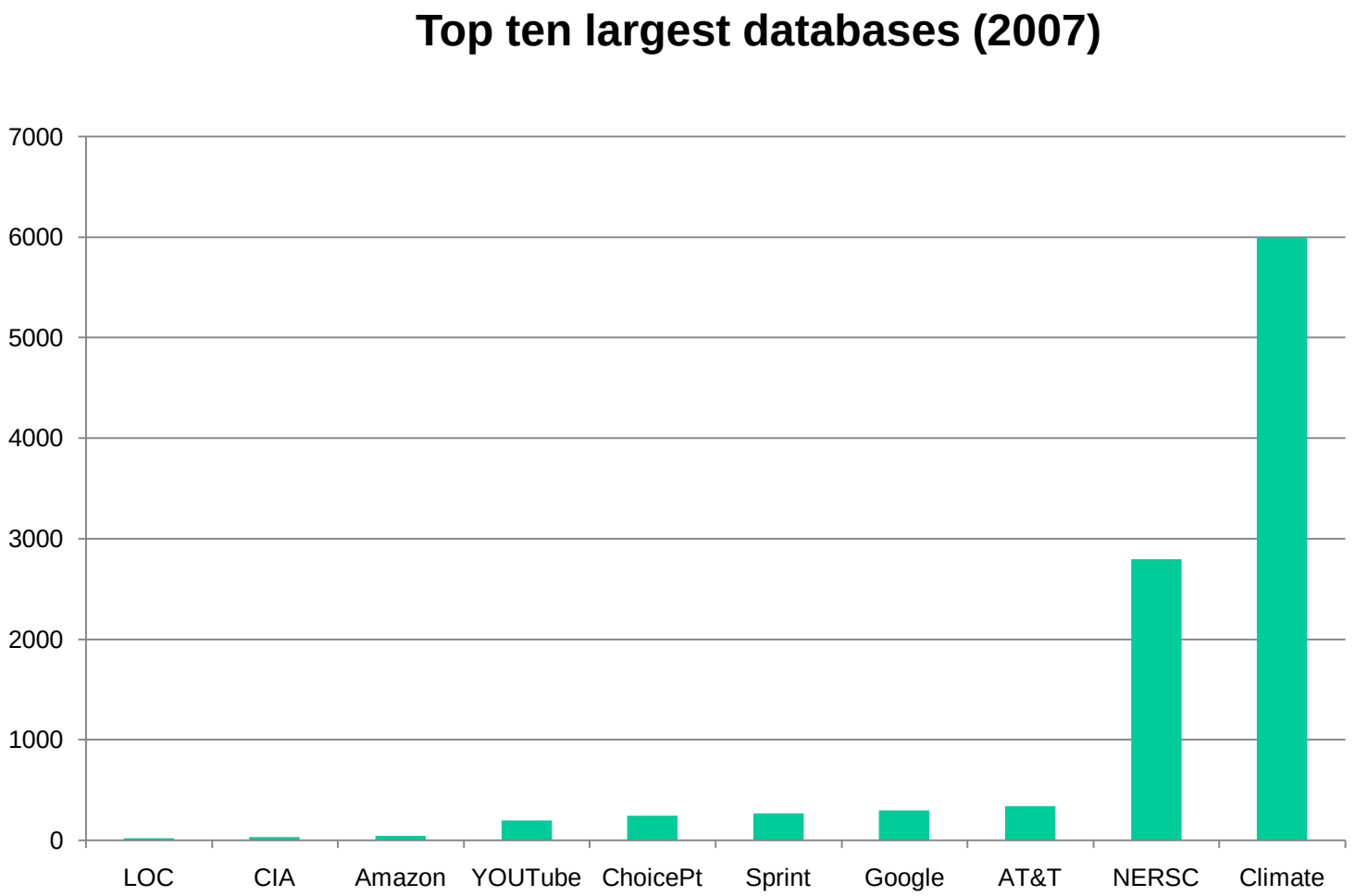
- **Problem Space: explosion of data**
- **Solution space: emergence of multi-core, virtualization, cloud computing**
- **Inability of traditional file system to handle data deluge**
- **The Big-data Computing Model**
  - **MapReduce Programming Model (Algorithm)**
  - **Google File System; Hadoop Distributed File System (Data Structure)**
  - **Microsoft Dryad**
- **This talk is about Cloud Computing and its Relevance to Big-data and Data-intensive computing**

Problem Space

Other variables:
Communication
Bandwidth, ?



Top Ten Largest Databases



Ref:
http://www.businessintelligencelowdown.com/2007/02/top_10_largest_.html

Challenges

- Need transformative solutions such as the Internet and the Search
- Alignment with the needs of the business / user / non-computer specialists / community and society
- Need to address the scalability issue: large scale data, high performance computing, automation, response time, rapid prototyping, and rapid time to production
- Need to effectively address (i) ever shortening cycle of obsolescence, (ii) heterogeneity and (iii) rapid changes in requirements
- Transform data from diverse sources into intelligence and deliver intelligence to right people/user/systems

Enter the cloud

- **Cloud computing** is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like the electricity grid.
- The cloud computing is a culmination of numerous attempts at large scale computing with seamless access to virtually limitless resources.
  - on-demand computing, utility computing, ubiquitous computing, autonomic computing, platform computing, edge computing, elastic computing, **grid computing**, …

"Grid Technology: A slide from my presentation to Industry (2005)

- Emerging enabling technology.
- Natural evolution of distributed systems and the Internet.
- Middleware supporting network of systems to facilitate sharing, standardization and openness.
- Infrastructure and application model dealing with sharing of compute cycles, data, storage and other resources.
- Publicized by prominent industries as on-demand computing, utility computing, etc.
- Move towards delivering "computing" to masses similar to other utilities (electricity and voice communication)."
- Now,

Hmmm…sounds like the definition for cloud computing!!!!!

IT IS A CHANGED WORLD NOW…

- Explosive growth in applications: biomedical informatics, space exploration, business analytics, web 2.0 social networking: YouTube, Facebook
- Extreme scale content generation: e-science and e-business data deluge
- Extraordinary rate of digital content consumption: digital gluttony: Apple iPhone, iPad, Amazon Kindle
- Exponential growth in compute capabilities: multi-core, storage, bandwidth, virtual machines (virtualization)
- Very short cycle of obsolescence in technologies: Windows Vista→ Windows 7; Java versions; C→C#; Phython
- Newer architectures: web services, persistence models, distributed file systems/repositories (Google, Hadoop), multi-core, wireless and mobile
- Diverse knowledge and skill levels of the workforce
- You simply cannot manage this complex situation with your traditional IT infrastructure:

Answer: The Cloud Computing?

- Typical requirements and models:
  - platform (PaaS),
  - software (SaaS),
  - infrastructure (IaaS),
  - Services-based application programming interface (API)
- A cloud computing environment can provide one or more of these requirements for a cost
- Pay as you go model of business
- When using a public cloud the model is similar to renting a property than owning one.
- An organization could also maintain a private cloud and/or use both.

Topics for Discussion

- Popular Cloud models
- Cloud Characteristics
  1. Extreme-scale
  2. Economies of scale
  3. Elastic resource
  4. Reliability and Availability
  5. Newer programming and application models
- Potential benefits
  1. Accessibility to masses
  2. Organic disaster mitigation and recovery
  3. Green Computing
  4. Rapid prototyping and deployment
- Demo
- References
- Conclusion

Public Cloud Computing Models

- Amazon Elastic Compute Cloud (**EC2**) suite
- Microsoft's Windows **Azure** Cloud Platform
- Google Application Engine (**GAE**)

- NASA's **Nebula** cloud computing platform: primarily for use by NASA

- Elastic Utility Computing Architecture Linking Your Programs To Useful Systems (**EUCALYPTUS**): open source cloud enabler that simulates Amazon EC2+S3; used by Nebula

Cloud Models

- Azure, EC2 and GAE are complementary models:
  - In EC2 you can work at the low levels and have control over the infrastructure
  - Azure is abstract in that it provides a fabric of compute cycles and storage and operates at the enterprise level
  - GAE is a Phython-based environment for design, develop and deploy.
- All have well-defined pricing model that is quite reasonable

Windows Azure

- Enterprise-level on-demand capacity builder
- Fabric of cycles and storage available on-request for a cost
- You have to use Azure API to work with the infrastructure offered by Microsoft

Amazon EC2

- Amazon EC2 is one large complex web service.
- EC2 provided an API for instantiating computing instances with any of the operating systems supported.
- It can facilitate computations through Amazon Machine Images (AMIs) for various other models. I will illustrate this with MapReduce
- Signature features: S3, Cloud Management Console, MapReduce Cloud, Amazon Machine Image (AMI)

Google App Engine

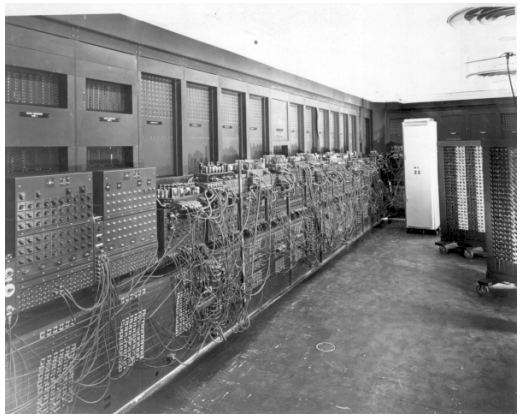
- This more a web interface for a development environment that offers a one stop facility for design, development and deployment Java and Phython-based applications in Java and Phython.
- Google offers the same reliability, availability and scalability at par with Google's own applications
- Interface is software programming based
- Comprehensive programming platform irrespective of the size (small or large)
- Signature features: templates and appspot

EXTREME-SCALES

Devices ⟵ Size ⟶ Computing services

Smaller

Larger



Eniac Computer 1964

Cloud Data center 2009

Economies of Scale

- Economies of scale: Increases use/production leads to increased efficiency and lowers unit cost
- Large organization have perfected the management and administration of large scale infrastructure for their own demanding operations;
- Example: amazon.com claims 99.95% availability;
- Why not build a business model around it by renting the resources?
- Three major players: amazon.com 40%, Google App Engine 40%, and Microsoft Azure 20% of the market

Elastic Resources

- Move, shift, balance. load, unload, juggle
- You can request only as many resources as you want and return them when you no longer need them
- Automatic load balancing possible
- On-demand

Non-functional Characteristics

- Availability: Amazon.com claims it is 99.95% available
- Reliability:
- Complex data relationships: Consider the customer data of a Bank application and the customer data for a Facebook application
  - Events, friends, groups, links among this: collocation of data is simply impossible; frequent updates
- In many cases ACID property of dbms is dropped! That's alright since most of the data is write once read many (WORM): web logs, historical data

Newer Programming and Application Models

- MapReduce programming model
- Google File System (GFS), Hadoop File System (HDFS), Collossus, BigTable (stores data in <key,value> pairs)
- Dryad from Microsoft
- Social networking management frameworks
- Many more ideas in the recent Symposium on Cloud Computing (SOCC 2010, USA) that I attended last week.

Accessibility to the Masses

- No software to write (at least for the common user); Take a look at the logo of SaleForce.com
- Buy cycles using a credit faster than you can go to a supermarket and get bread.
- The cloud environment will be usable similar to how gadgets were used
- NoSQL mostly; often substituted by simpler abstractions of storage

Disaster Mitigation

- Iceland's volcanic ash cloud illustrates the case for cloud computing: http://cloudcomputing.sys-con.com/node/1377030
- Cloud storage and on-demand services can help in setting up the backup as well as operation capacity to help in mitigating disasters.
- After the disaster has moved on, the cloud can be used to recover and restore; once the operation is restored, the cloud capacity is no longer needed and services can be terminated.

Facilitates Green Computing

- The cloud hardware is well located and optimized for efficient operation.
- Easy to control ramp up and ramp down the capacity, turn off and on according to green algorithms;
- Organization can establish policies and enforce them using green processes such as automatic on/off.
- Though on/off is available for individual PC/laptops, power saving is up to the user.

Rapid Prototyping and Deployment

- It is possible to quickly set up and ramp up to production.
- Example: Pop!World, popuation biology evolutionary tool.
  - Though it was developed as a stand alone we have deployed it on Google Apps Engine to be used by many first level undergraduates
- Example 2: HealthAdvisor application that was developed as JSP-Java-Tomcat-MySQL enterprise application was ported to Google App Engine.
- Example 3: MapReduce application on Amazon.com

Demo

- VMware simulated Hadoop and MapReduce demo
- Remote access to NEXOS system at my Buffalo, NY lab (Inaccessible?)
  - 5-node HDFS running HDFS on Ubuntu 8.04
  - 1 –name node and 5 data-nodes
  - Each is an old commodity PC with 512 MB RAM, 120GB – 160GB external memory
  - Zeus (namenode), datanodes:
    hermes, dionysus, aphrodite, athena, theos
- Amazon Elastic Cloud Computing (EC2) supported MapReduce

Demo Sampler

- Multi-core programming demo
- Machine instantiations and connecting to the instances
- Amazon machine images
- MapReduce on Amazon EC2.
- Our applications on Google App Engine

References

1. Cloud Computing: Characteristics, economies and architectures: http://en.wikipedia.org/wiki/Cloud_Computing
2. Eucalyptus Cloud Platform (ECP). http://open.eucalyptus.com/documents/ccgrid2009.pdf
3. http://docs.amazonwebservices.com/AWSEC2/latest/DeveloperGuide/
4. First ACM Symposium on Cloud Computing (SOCC) 2010, Indianapolis Indiana, http://research.microsoft.com/en-us/um/redmond/events/socc2010/