BITS Pilani
Pilani|Dubai|Goa|Hyderabad

Module 7
Lecture 01
Architectures for the Cloud-1

Harvinder S Jabbal
SE ZG651/ SS ZG653 Software Architectures

innovate achieve lead

BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Architectures for the Cloud

Chapter Outline

- Basic Cloud Definitions
- Service Models and Deployment Options
- Economic Justification

Basic Cloud Definitions (from NIST)

innovate achieve lead

- *On-demand self-service.* A resource consumer can unilaterally provision computing services, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.
- *Ubiquitous network access.* Cloud services and resources are available over the network and accessed through standard networking mechanisms that promote use by a heterogeneous collection of clients.
- *Resource pooling.* The cloud provider's computing resources are pooled.
- *Location independence.* The location of the resources need not be of concern to the consumer of the resources.
- *Rapid elasticity. C*apabilities can be rapidly and elastically provisioned.
- *Measured service.* Resource usage can be monitored, controlled, and reported so that consumers of the services are billed only for what they use.
- *Multi-tenancy.* Applications and resources can be shared among multiple consumers who are unaware of each other.

Basic Service Models

- **Software as a Service (SaaS).** The consumer in this case is an end user. The consumer uses applications that happen to be running on a cloud. E.g. mail services or data storage services.
- **Platform as a Service (PaaS).** The consumer in this case is a developer or system administrator. The consumer deploys applications onto the cloud infrastructure using programming languages and tools supported by the provider.
- **Infrastructure as a Service (IaaS).** The consumer in this case is a developer or system administrator. The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications.

Deployment Models

- *Private cloud.* The cloud infrastructure is owned solely by a single organization and operated solely for applications owned by that organization.
- *Public cloud.* The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.
- *Community cloud.* The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns.
- *Hybrid cloud.* The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities.

Economic Justification

- Economies of scale
- Utilization of equipment
- Multi-tenancy

Economies of Scale

- Large data centers are cheaper to operate (per unit measure) than small data centers.
- *Large* in this context means 100,000+ servers
- *Small* in this context means <10,000 servers.

Reasons for Economies of Scale

innovate achieve lead

- *Cost of power.* The cost of electricity to operate a data center currently is 15 to 20 percent of the total cost of operation.
- Per-server power costs are lower in large data centers
  - Sharing of items such as racks and switches.
  - Negotiated prices. Large power users can negotiate significant discounts.
  - Geographic choice. Large data centers can be located where power costs are lowest.
  - Acquisition of cheaper power sources such as wind farms and rooftop solar energy.
- *Infrastructure labor costs. More efficient utilization of system administrators*
  - Small data center administrators service ~150 servers.
  - Large data center administrators service >1000 servers.

More Reasons for Economies of Scale

innovate achieve lead

- *Security and reliability.* Maintaining a given level of security, redundancy, and disaster recovery essentially requires a fixed level of investment. Larger data centers can amortize that investment over their larger number of servers.
- *Hardware costs.* Operators of large data centers can get discounts on hardware purchases of up to 30 percent over smaller buyers.

Utilization of Equipment

- Use of virtualization technology allows for easy co-location of distinct applications and their associated operating systems on the same server hardware. The effect of this co-location is to increase the utilization of servers.
- Variations in workload can be managed to increase utilization.
  - *Random access.* End users may access applications randomly. The more likely that the randomness of their accesses will end up imposing a uniform load on the server.
- *Time of day.*
  - Co-locate those services that are workplace related with those that are consumer related.
  - Consider time differences among geographically distinct locations.
- *Time of year.* Consider yearly fluctuations in demand.
  - Holidays, tax preparation season
- *Resource usage patterns.* Co-locate heavier CPU services with heavier I/O services.
- *Uncertainty.* Consider spikes in usage

Multi-tenancy

- Some applications such as salesforce.com use a single application for multiple different consumers.
- This reduces costs by reducing costs of
  - Help desk support
  - Upgrade once, simultaneously, for all consumers
  - Single version of the software from a development and maintenance perspective.

Summary

- The cloud provides a new platform for applications with some different characteristics.