2. Cloud Computing (2/4)

Hyunchan, Park

http://oslab.jbnu.ac.kr

Division of Computer Science and Engineering

Jeonbuk National University

Agenda

- What is Cloud Computing ?
  - Different perspectives
  - Properties and characteristics
  - Benefits from cloud computing
- Service and deployment models
  - Three service models
  - Four deployment models

Properties and Characteristics

Scalability
Elasticity
- Dynamic provision
- Multi-tenant design

Scalability & Elasticity

Give me the world without limitation!!

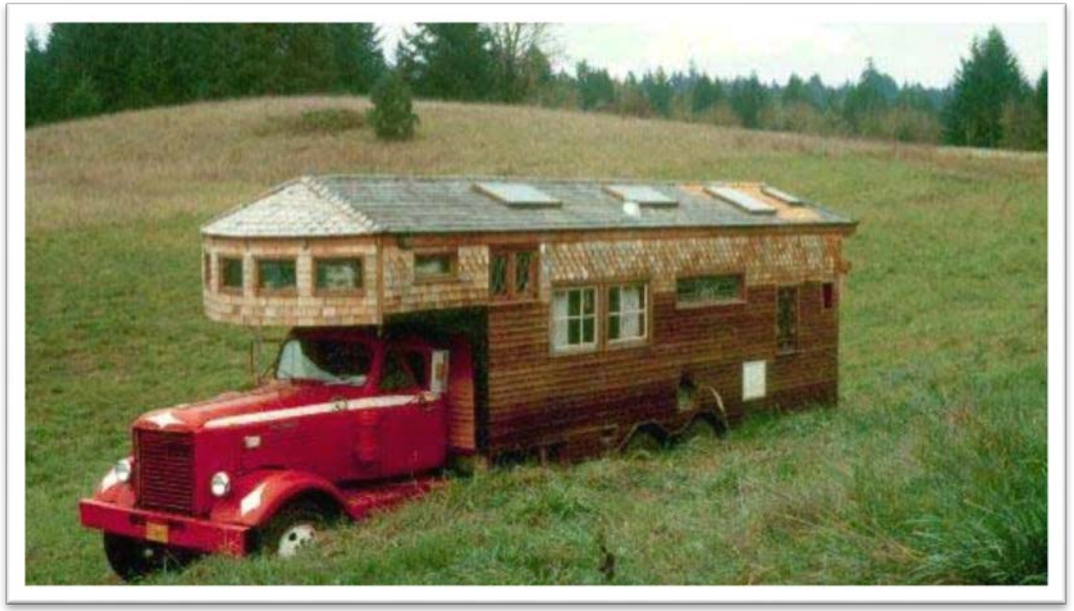
Scalability & Elasticity

- What is scalability ?
  - A desirable property of a system, a network, or a process, which indicates its ability to either handle growing amounts of work in a graceful manner or to be readily enlarged.
- What is elasticity ?
  - The ability to apply a quantifiable methodology that allows for the basis of an adaptive introspection with in a real time infrastructure.
- But how to achieve these properties ?
  - Dynamic provisioning
  - Multi-tenant design

Scalability
Elasticity
• Dynamic provision
• Multi-tenant design

Dynamic Provisioning

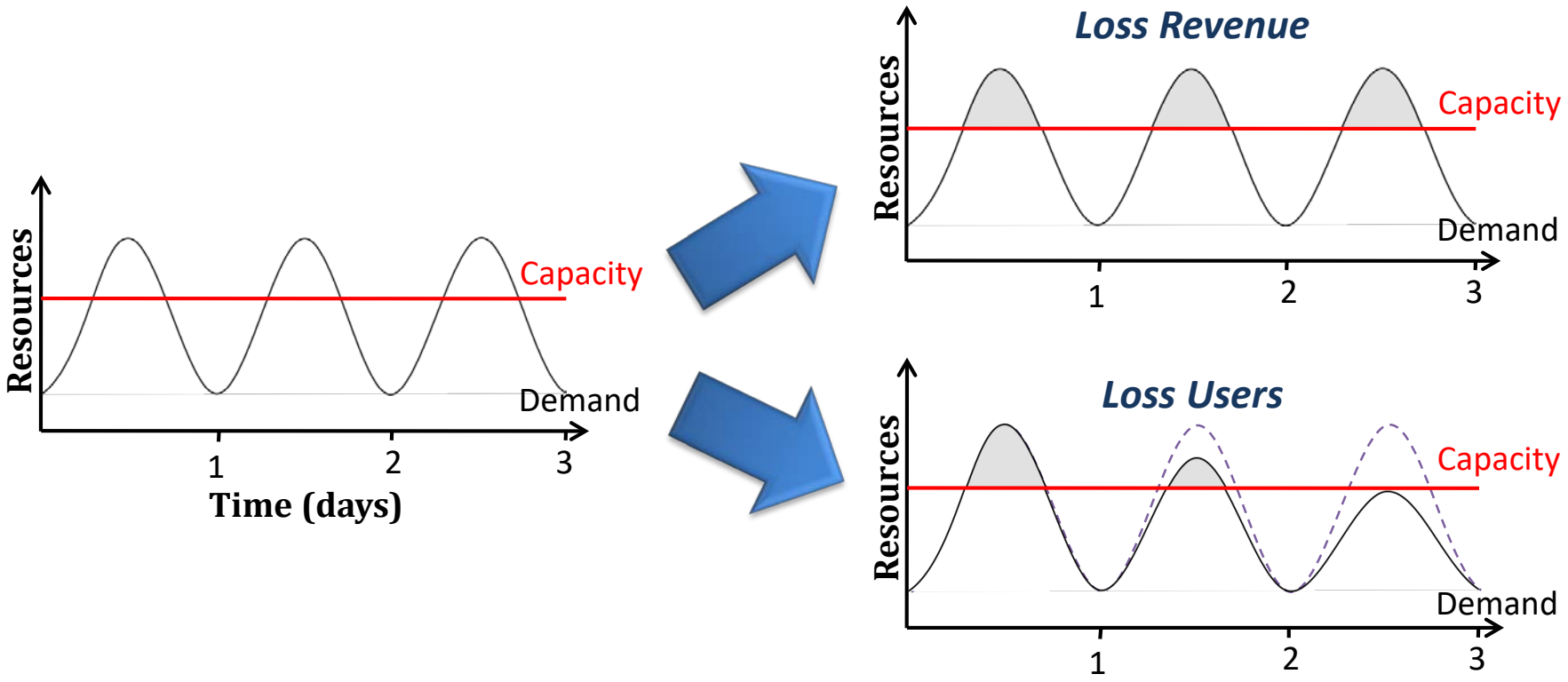
- What is dynamic provisioning ?
  - Dynamic Provisioning is a simplified way to explain a complex networked server computing environment where server computing instances are provisioned or deployed from a administrative console or client application by the server administrator, network administrator, or any other enabled user.



Scalability
Elasticity
• Dynamic provision
• Multi-tenant design

Dynamic Provisioning

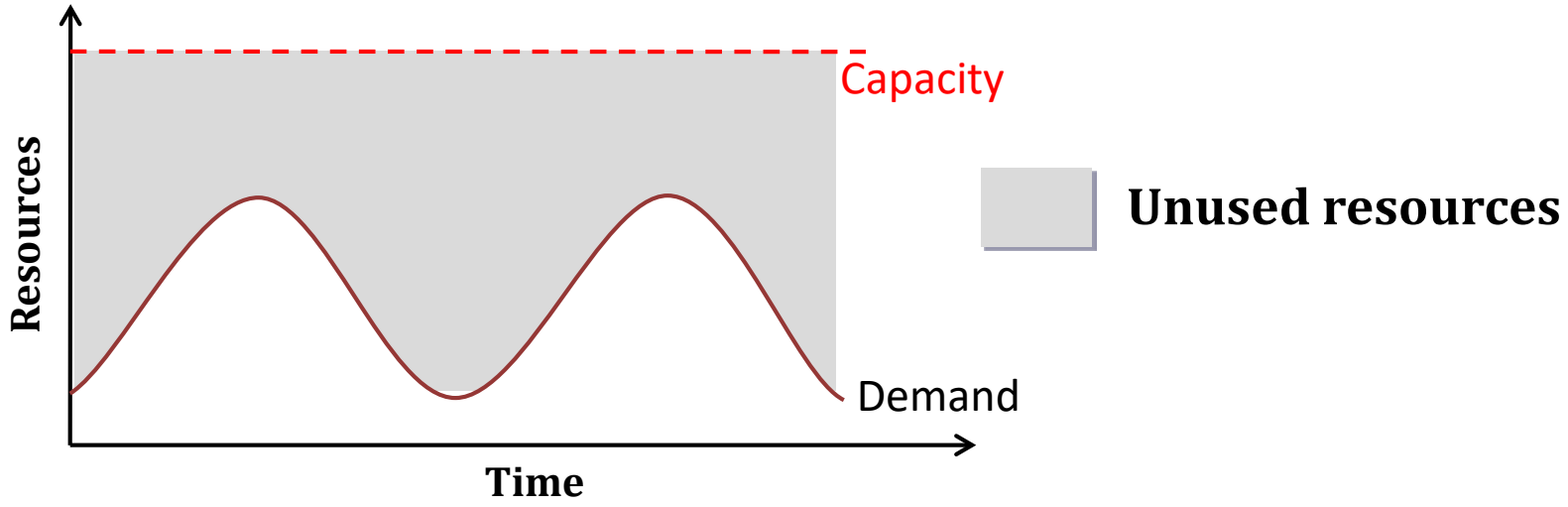
- In traditional computing model, two common problems :
  - Underestimate system utilization which result in under provision



Scalability
Elasticity
• Dynamic provision
• Multi-tenant design

Dynamic Provisioning

- Overestimate system utilization which result in low utilization

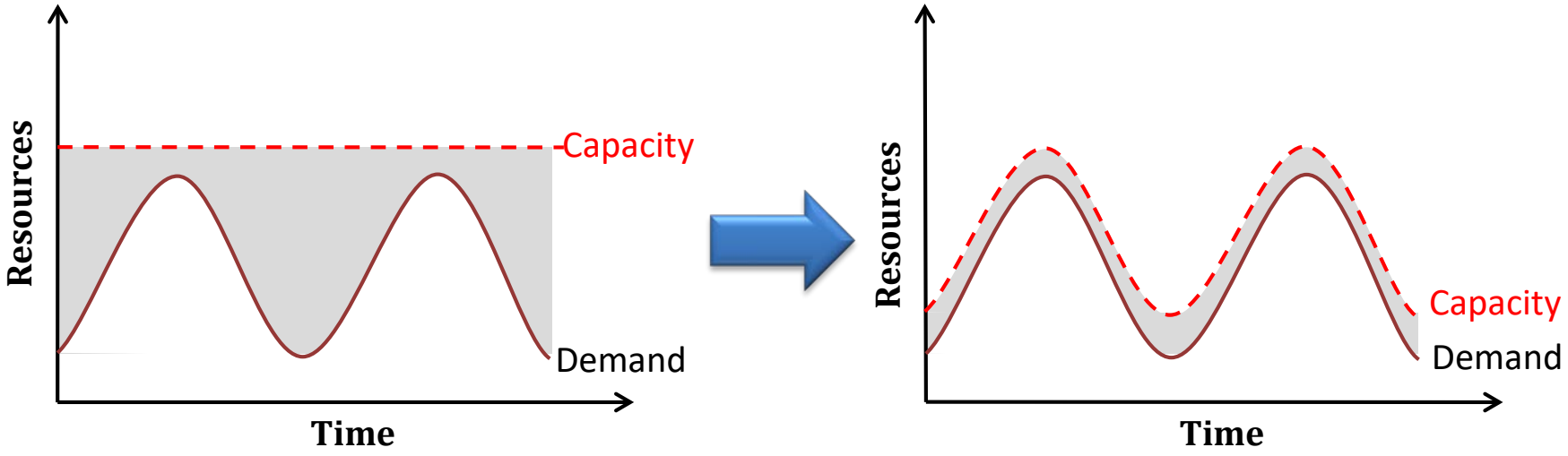


- How to solve this problem ??
  - Dynamically provision resources

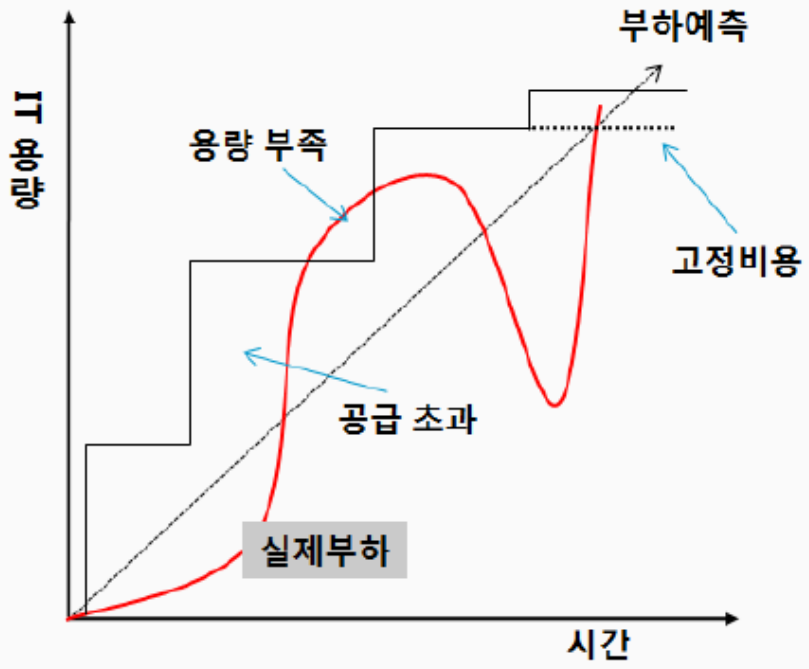
Scalability
Elasticity
• Dynamic provision
• Multi-tenant design

Dynamic Provisioning

- Cloud resources should be provisioned dynamically
  - Meet seasonal demand variations
  - Meet demand variations between different industries
  - Meet burst demand for some extraordinary events

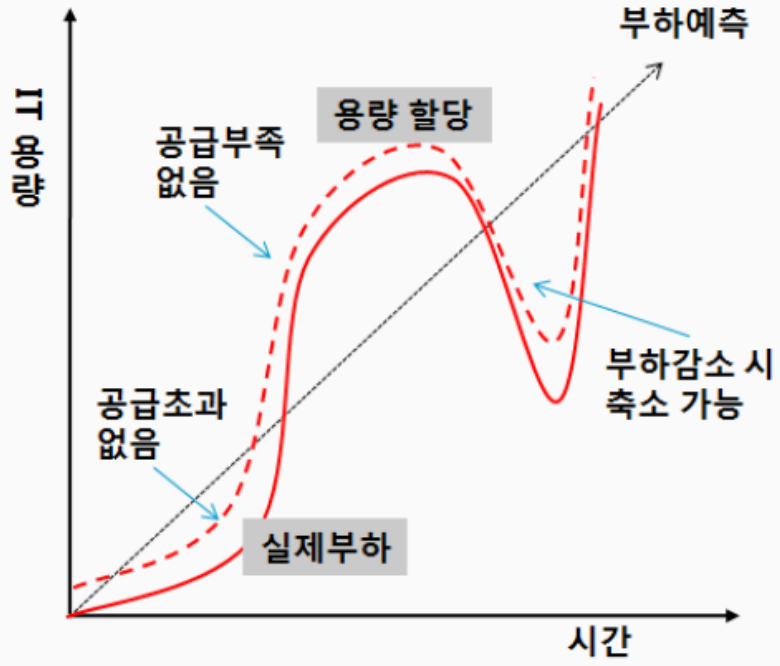


과거 비효율적 패턴



- 정확한 부하 예측의 어려움
- 공급 초과로 인한 비효율적 IT 비용 지출
- 예상치 못한 업무 증가에 따른 품질 저하
- 인프라 유지를 위한 고정비용 지속 상승

효율적 이용 패턴



- 부하 변동에 따른 탄력적 대응
- 공급 초과로 인한 초과 비용 지출 방지
- 예상치 못한 업무 증가에 유연한 증설
- 인프라 보유에 따른 고정비용 감소

Multi-tenant Design

- What is multi-tenant design ?
  - Multi-tenant refers to a principle in software architecture where a single instance of the software runs on a server, serving multiple client organizations.
  - With a multi-tenant architecture, a software application is designed to virtually partition its data and configuration thus each client organization works with a customized virtual application instance.

- Client oriented requirements :
  - Customization
    - Multi-tenant applications are typically required to provide a high degree of customization to support each target organization's needs.
  - Quality of service
    - Multi-tenant applications are expected to provide adequate levels of security and robustness.

Availability
Reliability

- Fault tolerance
- System resilience
- System security

Availability & Reliability

Data Never Loss

Machine Never Fail

Availability & Reliability

- What is availability ?
  - The degree to which a system, subsystem, or equipment is in a specified operable and committable state at the start of a mission, when the mission is called for at an unknown time.
  - Cloud system usually require high availability
    - Ex. “Five Nines” system would statistically provide 99.999% availability
- What is reliability ?
  - The ability of a system or component to perform its required functions under stated conditions for a specified period of time.
- But how to achieve these properties ?
  - Fault tolerance system
  - Require system resilience
  - Reliable system security

Fault Tolerance

- What is fault tolerant system ?
  - Fault-tolerance is the property that enables a system to continue operating properly in the event of the failure of some of its components.
  - If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a naively-designed system in which even a small failure can cause total breakdown.

- Four basic characteristics :
  - No single point of failure
  - Fault detection and isolation to the failing component
  - Fault containment to prevent propagation of the failure
  - Availability of reversion modes

Availability
Reliability
- Fault tolerance
- System resilience
- System security

Fault Tolerance

- Single Point Of Failure (SPOF)
  - A part of a system which, if it fails, will stop the entire system from working.
  - The assessment of a potentially single location of failure identifies the critical components of a complex system that would provoke a total systems failure in case of malfunction.
- Preventing single point of failure
  - If a system experiences a failure, it must continue to operate without interruption during the repair process.

Availability
Reliability
- Fault tolerance
- System resilience
- System security

Fault Tolerance

- Fault Detection and Isolation (FDI)
  - A subfield of control engineering which concerns itself with monitoring a system, identifying when a fault has occurred and pinpoint the type of fault and its location.

- Isolate failing component
  - When a failure occurs, the system must be able to isolate the failure to the offending component.

Fault Tolerance

- Fault Containment
  - Some failure mechanisms can cause a system to fail by propagating the failure to the rest of the system.
  - Mechanisms that isolate a rogue transmitter or failing component to protect the system are required.

- Available of reversion modes
  - System should be able to maintain some check points which can be used in managing the state changes.

System Resilience

- What is resilience ?
  - Resilience is the ability to provide and maintain an acceptable level of service in the face of faults and challenges to normal operation.
  - Resiliency pertains to the system's ability to return to its original state after encountering trouble. In other words, if a risk event knocks a system offline, a highly resilient system will return back to work and function as planned as soon as possible.

- Some risk events
  - If power is lost at a plant for two days, can our system recover ?
  - If a key service is lost because a database corruption, can the business recover ?

Availability
Reliability
- Fault tolerance
- System resilience
- System security

System Resilience

- Disaster Recovery
  - Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of technology infrastructure critical to an organization after a natural or human-induced disaster.

- Some common strategies :
  - Backup
    - Make data off-site at regular interval
    - Replicate data to an off-site location
    - Replicate whole system
  - Preparing
    - Local mirror systems
    - Surge protector
    - Uninterruptible Power Supply (UPS)

Availability
Reliability
- Fault tolerance
- System resilience
- System security

System Security

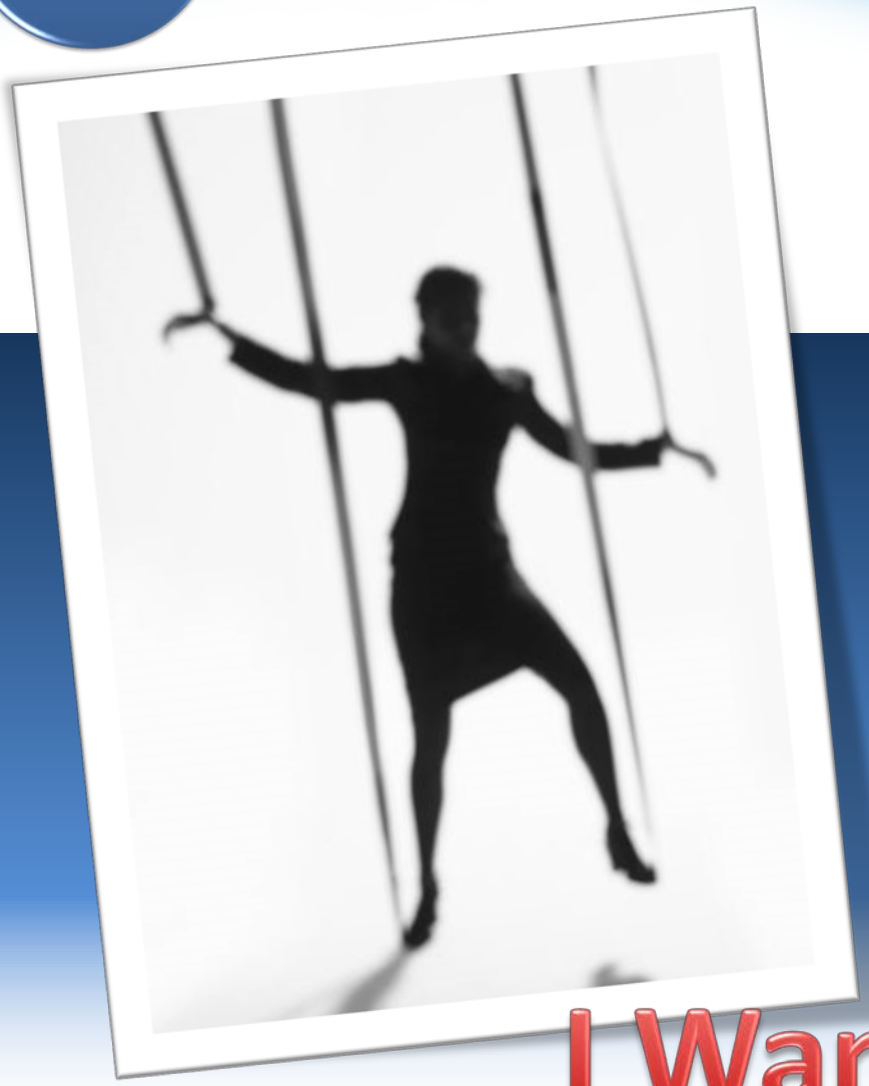
- Security issue in Cloud Computing :
  - Cloud security is an evolving sub-domain of computer security, network security, and, more broadly, information security.
  - It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing.

System Security

- Important security and privacy issues :
  - Data Protection
    - To be considered protected, data from one customer must be properly segregated from that of another.
  - Identity Management
    - Every enterprise will have its own identity management system to control access to information and computing resources.
  - Application Security
    - Cloud providers should ensure that applications available as a service via the cloud are secure.
  - Privacy
    - Providers ensure that all critical data are masked and that only authorized users have access to data in its entirety.

Manageability
Interoperability
- Control automation
- System monitoring
- Billing system

Manageability & Interoperability



I Want Full Control !!

Manageability & Interoperability

- What is manageability ?
  - Enterprise-wide administration of cloud computing systems. Systems manageability is strongly influenced by network management initiatives in telecommunications.
- What is interoperability ?
  - Interoperability is a property of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, without any restricted access or implementation.
- But how to achieve these properties ?
  - System control automation
  - System state monitoring

Control Automation

- What is Autonomic Computing ?
  - Its ultimate aim is to develop computer systems capable of self-management, to overcome the rapidly growing complexity of computing systems management, and to reduce the barrier that complexity poses to further growth.

- Architectural framework :
  - Composed by Autonomic Components (AC) which will interact with each other.
  - An AC can be modeled in terms of two main control loops (local and global) with sensors (for self-monitoring), effectors (for self-adjustment), knowledge and planer/adapter for exploiting policies based on self- and environment awareness.

Control Automation

- Four functional areas :
  - Self-Configuration
    - Automatic configuration of components.
  - Self-Healing
    - Automatic discovery, and correction of faults.
  - Self-Optimization
    - Automatic monitoring and control of resources to ensure the optimal functioning with respect to the defined requirements.
  - Self-Protection
    - Proactive identification and protection from arbitrary attacks.

Manageability
Interoperability
- Control automation
- System monitoring
- Billing system

System Monitoring

- What is system monitor ?
  - A System Monitor in systems engineering is a process within a distributed system for collecting and storing state data.
- What should be monitored in the Cloud ?
  - Physical and virtual hardware state
  - Resource performance metrics
  - Network access patterns
  - System logs
  - … etc
- Anything more ?
  - Billing system

Manageability
Interoperability
- Control automation
- System monitoring
- Billing system

Billing System

- Billing System in Cloud
  - Users pay as many as they used.
  - Cloud provider must first determine the list of service usage price.
  - Cloud provider have to record the resource or service usage of each user, and then charge users by these records.
- How can cloud provider know users' usage ?
  - Get those information by means of monitoring system.
  - Automatically calculate the total amount of money which user should pay. And automatically request money from use's banking account.

Performance Optimization

- Parallel processing
- Load balancing
- Job scheduling

Performance & Optimization

High Performance Improvement

Performance & Optimization

- Performance guarantees ??
  - As the great computing power in cloud, application performance should be guaranteed.
  - Cloud providers make use of powerful infrastructure or other underlining resources to build up a highly performed and highly optimized environment, and then deliver the complete services to cloud users.
- But how to achieve this property ?
  - Parallel computing
  - Load balancing
  - Job scheduling

Parallel Processing

- Parallel Processing
  - Parallel processing is a form of computation in which many calculations are carried out simultaneously, operating on the principle that large problems can often be divided into smaller ones, which are then solved concurrently.

- Parallelism in different levels :
  - Bit level parallelism
  - Instruction level parallelism
  - Data level parallelism
  - Task level parallelism

Performance Optimization
- Parallel processing
- Load balancing
- Job scheduling

Parallel Processing

- Hardware approaches
  - Multi-core computer
  - Symmetric multi-processor
  - General purpose graphic processing unit
  - Vector processor
  - Distributed computing
    - Cluster computing
    - Grid computing
- Software approaches
  - Parallel programming language
  - Automatic parallelization

Performance Optimization
- Parallel processing
- Load balancing
- Job scheduling

Load Balancing

- What is load balancing ?
  - Load balancing is a technique to distribute workload evenly across two or more computers, network links, CPUs, hard drives, or other resources, in order to get optimal resource utilization, maximize throughput, minimize response time, and avoid overload.

- Why should be load balanced ?
  - Improve resource utilization
  - Improve system performance
  - Improve energy efficiency

Unbalanced →

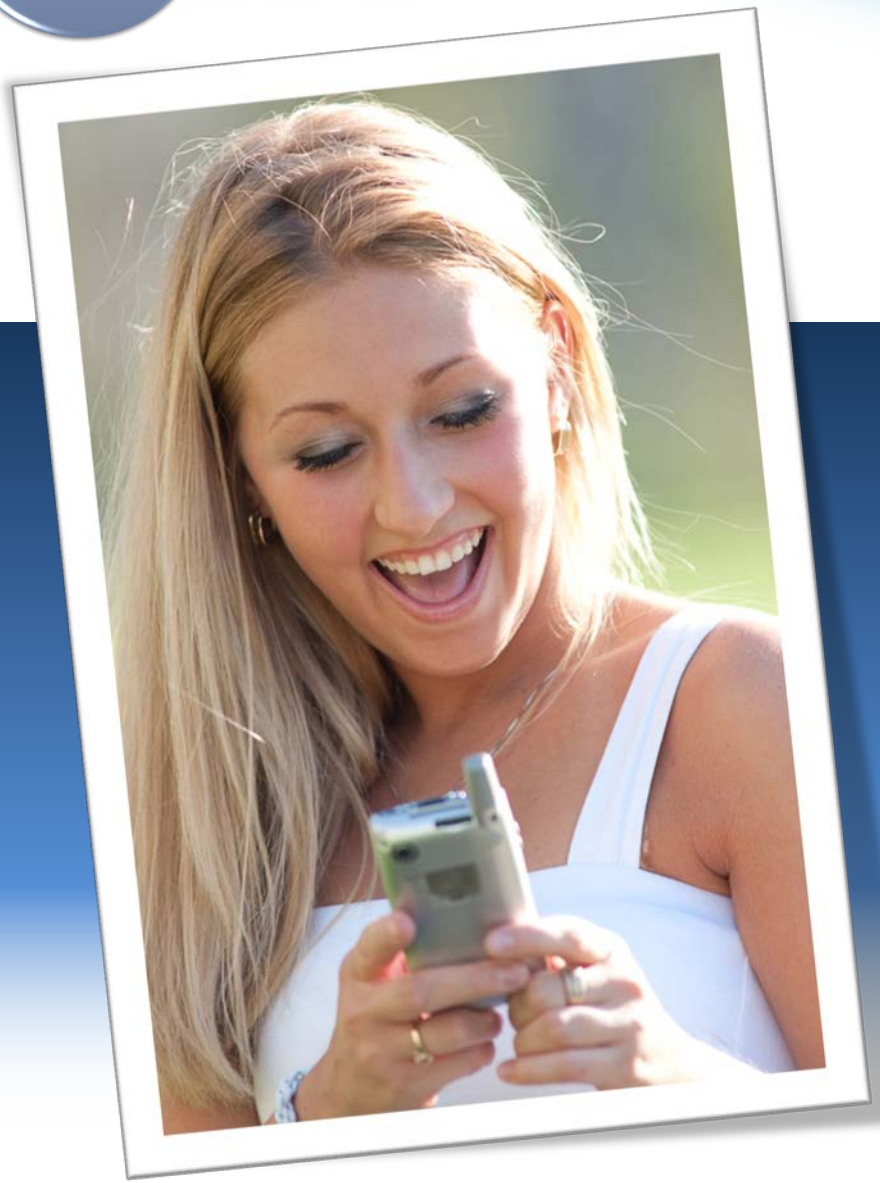
Job Scheduling

- What is job scheduler ?
  - A job scheduler is a software application that is in charge of unattended background executions, commonly known for historical reasons as batch processing.
- What should be scheduled in Cloud ?
  - Computation intensive tasks
  - Dynamic growing and shrinking tasks
  - Tasks with complex processing dependency
- How to approach ?
  - Use pre-defined workflow
  - System automatic configuration

Accessibility
Portability

- Uniform access
- Thin client

Accessibility & Portability



Anyone !

Anytime !

Anywhere !

Accessibility & Portability

- What is accessibility ?
  - Accessibility is a general term used to describe the degree to which a product, device, service, or environment is accessible by as many people as possible.
- What is service portability ?
  - Service portability is the ability to access services using any devices, anywhere, continuously with mobility support and dynamic adaptation to resource variations.
- But how to achieve these properties ?
  - Uniform access
  - Thin client

Accessibility
Portability
• Uniform access
• Thin client

Uniform Access

- How do users access cloud services ?
  - Cloud provider should provide their cloud service by means of widespread accessing media. In other word, users from different operating systems or other accessing platforms should be able to directly be served.
  - Nowadays, web browser technique is one of the most widespread platform in almost any intelligent electronic devices. Cloud service take this into concern, and delivery their services with web-based interface through the Internet.

Thin Client

- What is thin client ?
  - Thin client is a computer or a computer program which depends heavily on some other computer to fulfill its traditional computational roles. This stands in contrast to the traditional fat client, a computer designed to take on these roles by itself.
- Characteristics :
  - Cheap client hardware
    - While the cloud providers handle several client sessions at once, the clients can be made out of much cheaper hardware.
  - Diversity of end devices
    - End user can access cloud service via plenty of various electronic devices, which include mobile phones and smart TV.
  - Client simplicity
    - Client local system do not need complete operational functionalities.