

- * On-demand delivery of computing power: *utility computing*
 - Similar to traditional public utility services such as water, electricity, gas, and telephony
- * Technologies: cluster, grid, and cloud computing

- Definitions of cloud computing

- * Buyya *et al.*:

Cloud is a parallel and distributed computing system consisting of a collection of interconnected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers

- * Vaquero *et al.*:

Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized Service Level Agreements.

- * Key characteristics summarized by U.C. Berkeley:

- (1) *The illusion of infinite computing resources available on demand*
- (2) *The elimination of an up-front commitment by cloud users*
- (3) *The ability to pay for use of computing resources on a short-term basis as needed*

- * National Institute of Standards and Technology (NIST):

- * Software resources are packaged as “services” that
 - # are well-defined, self-contained modules that provide standard business functionality
 - # are independent of the state or context of other services → loosely coupled
 - # follow industry standards, such as HTTP, XML, SOAP, ...
- * Development, maintenance, and usage of the SOA
 - # Reuse, granularity, modularity, composability, componentization, and interoperability
 - # Standards-compliance (both common and industry-specific)
 - # Services identification and categorization, provisioning and delivery, and monitoring and tracking
- Web 2.0
 - * Information sharing, user-centric design, cooperation, collaborative Writing
 - * Social networking
 - * Read-write webs (Prosumer web): users provide information, instead of the web sites
- Service mashup: providers’ information and services may be aggregated at the consumers’ web
 - * E.g.: Amazon, del.icio.us, Facebook, and Google make their service APIs (Application programming interface) publicly accessible using standard protocols such as SOAP and REST
 - * Integration of various info. and services, e.g., Facebook like, Google map, Google+, subscription, ...
- Grid computing

- SaaS

- * Started around 2000
- * Traditional desktop applications are provided as a service in the Web
- * Alleviates the burden of software maintenance for customers and simplifies development and testing for providers
- * Examples: Google Apps, Salesforce CRM, Zoho CRM, Clarizen, ...

- PaaS

- * Started around 2011
- * Offers an environment in which developers create and deploy applications
- * Users does not need to know how many processors or how much memory that applications are using
- * Multiple programming models and specialized services (e.g., data access, authentication, and payments) are offered
- * Example: Force.com, Google App Engine, Heroku, Cloud Foundry, Microsoft Azure, ...

- IaaS

- * Provides on-demand provisioning of servers running several choices of operating systems and a customized software stack
- * Offers virtual resources on demand: computation, storage, and communication
- * Bottom layer of cloud computing systems
- * Example: Amazon EC2, Google Compute Engine, Rackspace, GoGrid, Microsoft, HP, AT&T, OpSource, ...

- * Automatic scaling and load balancing
 - # Allow users' applications to scale up and down, based on application-specific metrics such as transactions per second, number of simultaneous users, request latency, ...
 - # When the number of virtual servers is increased by automatic scaling, incoming traffic must be automatically distributed among the available servers
- * Service-level agreements (SLA)
 - # IaaS providers' commitment to delivery of a certain QoS as a warranty
 - # Usually include availability and performance guarantees
 - # Metrics must be agreed upon by all parties as well as penalties for violating these expectations
 - # E.g., Amazon EC2: <http://aws.amazon.com/ec2-sla/>
- * Hypervisor and operating system choice
 - # Traditionally: IaaS was based on heavily customized open-source Xen deployments
 - IaaS providers needed expertise in Linux, networking, virtualization, metering, resource management, ...
 - # Recently: other IaaS platforms (VMWare vCloud, Citrix Cloud Center (C3), ...) lowered the barrier of entry for IaaS competitors, leading to a rapid expansion of the market

- IaaS examples