Introduction

- Category of computing solutions to allow access to a service/technology on demand
  - Elastic computer
  - On-demand computing and storage
  - Enabling ubiquitous network access to a shared pool of configurable computing resources
  - Resources can be
    - Physical or virtual
    - Dedicated or shared
    - Accessed via modem/LAN/WAN/Internet
  - Provide various capabilities to process and store data in third party data centers
    - Sharing of resources to achieve coherence and economies of scale
    - Converged infrastructure and shared services
    - Focus on maximizing the effectiveness of shared resources
      - Shared by multiple users and dynamically reallocated on demand
      - Maximize usage of resources by possibly allocating in different time zones
  - Move away from CAPEX (capital expenditure) model to OPEX (operating expenditure) model
    - Do not buy the expensive hardware infrastructure
    - Buy only as much [virtual] infrastructure as needed and expand as business grows
- Characterized by self service interfaces
- Takes advantage of virtualization
- Save investment costs in infrastructure
- Back to the Future
  - Mainframes
  - Distributed client server model, based on PC
    - Distribute processing across multiple nodes without the need for mainframe gatekeepers
    - Parallel and distributed computing; data-intensive and network-centric model
    - Rapid deployment of applications without the assurance of proper security and controls
    - Nonstandard and insecure applications, leading to security breaches, identity thefts, and cyber threats
    - Complex challenge to manage enterprise
      - Problems with integration, interoperability, and widespread patching
      - Move from business enablement to IT maintenance
  - Network connecting the organization to the rest of the world through Internet
    - Integration of computers across organizations
    - Interoperability of the systems from suppliers and consumers to customers
    - Further increase in system complexity, with decrease in level of control and governance
    - High performance computing (HPC) vs high throughput computing (HTC)
      - HPC characterized by raw speed performance; current goal of exascale computing
      - HTC emphasizes high-flux computing; high-speed search and web services to millions of users simultaneously
      - HTC issues include cost, energy savings, security, and reliability at many data centers
      - Upgrade data centers with fast servers, storage systems, and high-bandwidth networks
      - Clusters, grids, and clouds

    - Peer-to-peer (P2P) networks for distributed file sharing and content delivery applications
        - P2P system built over multiple client machines
        - Peer machines may be globally distributed
  - HTC design objectives
    - Efficiency – Job throughput, data access, storage, power efficiency
    - Dependability – Quality of service (QoS) assurance, even under failure conditions
    - Adaptation in the programming model – Ability to support billions of job requests over massive data sets and virtualized cloud resources under various workload and service models
    - Flexibility in application deployment – Ability of distributed systems to run well
  - Current emphasis on mobile computing and ubiquitous computing
    - Ubiquitous computing uses pervasive devices at any place and time using wired or wireless communications
    - Internet of Things (IoT) is a networked connection of everyday objects, supported by cloud to achieve ubiquitous computing with any object
  - Cloud computing
    - Innovative collaboration of cloud technology and big iron
        - Best of mainframe technologies combined with the best of PC-enabled client-server plus the Internet
        - At scale, using a pay-as-you-go billing model
        - No need to buy expensive hardware or build data centers
    - Allows to pick resources as needed at every level – from hardware to applications
    - Commodity clouds
    - Enterprise-class clouds
  - Cloud service classification
    - Public cloud
        - Provided by some big players such as Amazon, Google, and Microsoft
        - Provide computing, storage, and other services to anyone willing to pay
        - Not regulated like public utilities
    - Private cloud
        - Operated by a private entity for a limited customer base
    - Public clouds operate at a very large scale compared to private cloud
        - They offer a broad range of powerful features: elasticity, fine-grained billing, high reliability due to geographic distribution, wide variety of resource types, and rich sets of platform services
    - Hybrid clouds
        - Combination of public and private cloud
        - Cloud burst
    - Community cloud
        - A private cloud to support a certain community
        - Academic cloud

- Cloud computing
  - Five essential characteristics of cloud computing identified by NIST: on-demand self-service, broad network access, resource pooling, rapid elasticity/expansion, and measured service
  - IT as a Service
    - Computers in the cloud configured to work together
    - Applications using collective computing power as if running on a single system
    - Flexibility from the allocation of resources on demand
    - Resources used as an aggregated virtual computer

  - Software as a Service (SaaS)
    * Meeting customer needs to be met over the web as an on-demand software solution
  - Platform as a Service (PaaS)
    * Platform to quickly develop scalable solutions without infrastructure costs
  - Infrastructure as a Service (IaaS)
    * Virtual data center to build scalable solutions at a lower cost

- Advantages of cloud computing
  - Reduced cost
    * Reduced capital expenses and operating expenses
    * Resources are only acquired when needed and paid for when used
  - Refined usage of personnel
    * Personnel focus on delivering value rather than maintaining hardware/software
  - Robust scalability
    * Allows for immediate scalability, both up and down, without long-term commitment

- Disadvantages
  - Unregulated marketplace
  - Not fully understood by professionals
    * No standards or best practices
    * Multiple definitions and interpretations of cloud-based models and frameworks in the IT literature
  - Wrong adoption decisions may affect the business adversely
  - Must make educated decisions about the scope of technology and its role in projects
    * The business goals should be well documented and fulfilled in a concrete and measurable manner at each phase of adoption

- Utility computing
  - Receive computing services from a paid service provider
  - Two major design objectives in any computing model: reliability and scalability
  - Models supported by QoS and SLAs
  - Users expect new network-efficient CPUs, scalable memory and storage schemes, distributed OSes, middleware for machine virtualization, new programming models, effective resource management, and application program development

- Internet of Things (IoT)
  - Networked connection of everyday objects, tools, devices, or computers
  - Sensors that interconnect all things in our daily life
  - Tag every object using RFID or sensor or other technology like GPS
  - Uses IPv6 to distinguish all objects and pervasive devices; universal addressability
  - Devices are interconnected and interact with each other in a meaningful way
    * Communication patterns from human-to-human (H2H), human-to-thing (H2T), and thing-to-thing (T2T)

- Cyber-physical systems (CPS)
  - Interactions between computational processes and physical world
  - CPS intergrates *cyber* (heterogeneous, asynchronous) with *physical* (concurrent, information-dense) objects

  - Merges computation, communications, and control into an intelligent closed feedback system
  - Exploration of VR applications in physical world

Technologies for network-based systems

- Multicore CPUs and multithreading
  - Processor speed measured in MIPS
  - Network bandwith measured in Mbps or Gbps; GE – 1 Gbps Ethernet bandwidth
  - Advances in CPU
    * Multicore architectures
    * Exploiting parallelism at ILP and TLP levels
    * Moore's Law – Number of transistors in a dense IC doubles approximately every two years
    * Clock rate increased as well but hit a limit on CMOS chips dues to power limitations; excessive heat generation with high frequency or high voltages
    * ILP makes up for frequency using multiple-issue superscalar architecture, dynamic branch prediction, and speculative execution
    * Rise of GPGPU
  - Multithreading
    * Simultaneous multithreaded processor (SMT)
    * Simultaneous scheduling of instructions from different threads in the same cycle
  - Power efficiency
    * About 2nJ/instruction on CPU; 200 pJ/instruction on GPU
    * CPU optimized for latency in caches and memory
    * GPU optimized for throughput with explicit management of on-chip memory
- Memory, storage, and wide-area networking
  - Disk and storage technology
    * Rapid growth in flash memory and SSD
    * SSD can handle large loads of read/write over a long time
  - System-area interconnects
    * Nodes in a small cluster connected by an Ethernet switch or a LAN
    * LAN connects client hosts to big servers
    * SAN connects servers to network storage
    * Network attached storage (NAS) connects clients hosts directly to network storage
  - Wide-area networking
    * Increases the capability to build massively distributed systems
    * Based on Gigabit Ethernet as interconnect in server clusters
- Virtual machines and virtualization middleware
  - Novel solution to underutilized resources, application inflexibility, software manageability, and security concerns in existing physical machines
  - Virtual machines
    * Host machine equipped with physical hardware
    * VM provisioned for any hardware system
    * VM built with virtual resources managed by a guest OS to run a specific application

    - Virtual machine monitor (VMM)
        - Middleware layer between host machine and VM
        - Hypervisor or bare metal VM – handles the bare hardware directly
        - Host VM – VMM runs in non-privileged mode; host OS need not be modified
        - Dual mode – part of VMM runs in user mode, another part runs in privileged mode
  - Hypervisor
    - Software to enable users to monitor and control servers built on hosted environments
    - Used to remotely allocate shared resources that can have a large impact on the efficiency of data transfer
  - VM primitive operations
    - VMM provides VM abstraction to the guest OS
    - With full virtualization, VMM exports a VM abstraction identical to the physical machine so that a standard OS can run just as it would on physical hardware
      1. Multiplex VMs between hardware machines
      2. VM suspended and stored in stable storage
      3. Suspended VM resumed or provisioned to new hardware platform
      4. Migrate VM from one hardware platform to another
  - Virtual infrastructure
    - Connects resources to distributed applications
    - Dynamic mapping of system resources to specific applications
    - Decrease in costs and increase in efficiency and responsiveness

- Data center virtualization for cloud computing
  - Cloud architectures built with commodity hardware and network devices
  - Data center design emphasizes price/preformance ratio over speed
  - Data center growth and cost breakdown
    - IT equipment – 30%
    - Chiller – 33%
    - UPS – 18%
    - Computer room air conditioning – 9%
    - Power distribution – 7%
  - Low-cost design philosophy
    - No need for high end switches and equipment
    - Software layer to handle network traffic balancing, fault tolerance, and expandability

System models for distributed and cloud computing

- Clusters of cooperative computers
  - Interconnected stand-alone computers working cooperatively as a single computing resource
  - Can handle heavy workloads with large data sest
  - Cluster architecture
    - Built around a low-latency high-bandwidth interconnection network
        - Loosely coupled node computers
        - Scalable with an increasing number of nodes
        - All resources on a node managed by its own OS
    - Cluster connected to Internet via a VPN gateway

        - Gateway IP address locates the cluster
    - Single-system image
        * Presents a collection of resources as a single, integrated, powerful resource
        * Makes cluster appear as a single machine to the user

Accessing the cloud: Web, APIs, SDKs

- Web interfaces, APIs, SDKs, and CLIs
    - Most clouds support access via web, with no local installation
    - Web interface can be tedious for repeated work
    - Cloud services support REST API – Representational State Transfer to permit request transmission via secure hypertext protocol (https), using `GET` and `PUT` commands
    - Cloud service providers give access to SDKs that allow the users to access REST APIs via programs in high level language
- Local and cloud-hosted applications
    - Should the application be run locally or in the cloud?