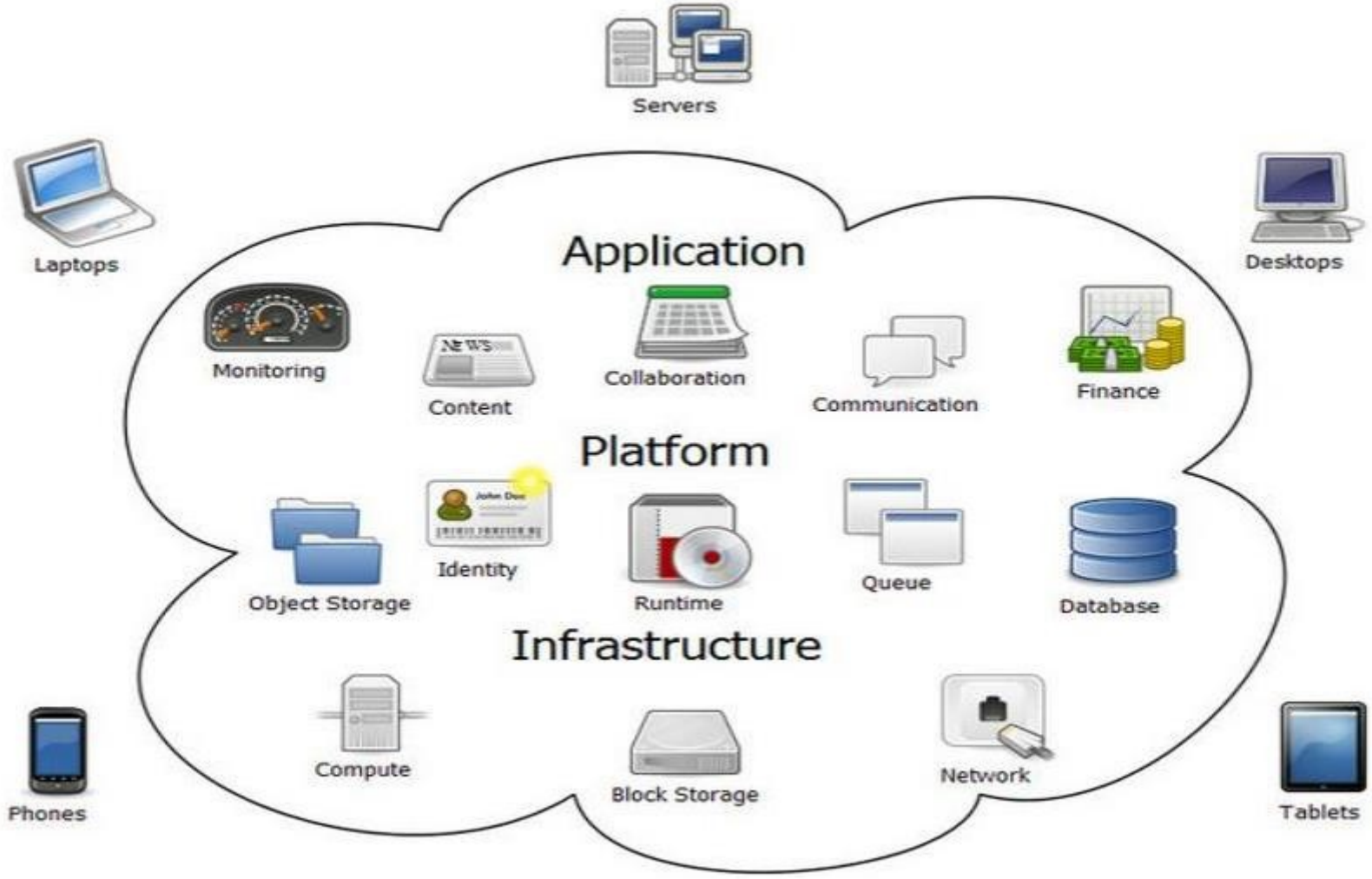
Cloud Resource Management

Different Resources in Computing



Resources types

- Physical resource
  - Computer, disk, database, network, scientific instruments
- Logical resource
  - Execution, monitoring, communicate application

Resources Management

- The term **resource management** refers to the operations used to control how capabilities provided by Cloud resources and services cane be made available to other entities, whether users, applications, services in an efficient manner.

Data Center Power Consumption

- Currently it is estimated that servers consume 0.5% of the world’s total electricity usage.
- Server energy demand doubles every 5-6 years.
- This results in large amounts of CO 2 produced by burning fossil fuels.
- Need to reduce the energy used with minimal performance impact.

Motivation for Green Data Centers

- **Economic**
  - New data centers run on the Megawatt scale, requiring millions of dollars to operate.
  - Recently institutions are looking for new ways to reduce costs
  - Many facilities are at their peak operating stage, and cannot expand without a new power source.
- **Environmental**
  - Majority of energy sources are fossil fuels.
  - Huge volume of CO_2 emitted each year from power plants.
  - Sustainable energy sources are not ready.
  - Need to reduce energy dependence

Green Computing ?

- Advanced scheduling schemas to reduce energy consumption.
  - Power aware
  - Thermal aware
- Performance/Watt is not following Moore’s law.
- Data center designs to reduce Power Usage Effectiveness.
  - Cooling systems
  - Rack design

Research Directions

- How to conserve energy within a Cloud environment.
  - Schedule VMs to conserve energy.
  - Management of both VMs and underlying infrastructure.
  - Minimize operating inefficiencies for non-essential tasks.
  - Optimize data center design.

VM scheduling on Multi-core Systems

- There is a nonlinear relationship between the number of processes used and power consumption
- We can schedule VMs to take advantage of this relationship in order to conserve power

Power-aware Scheduling

- Schedule as many VMs at once on a multi-core node.
  - Greedy scheduling algorithm
  - Keep track of cores on a given node
  - Match VM requirements with node capacity

Algorithm 1 Power based scheduling of VMs

```
FOR i = 1 TO i ≤ |pool| DO
  pe_i = num cores in pool_i
END FOR

WHILE (true)
  FOR i = 1 TO i ≤ |queue| DO
    vm = queue_i
      FOR j = 1 TO j ≤ |pool| DO
        IF pe_j ≥ 1 THEN
        IF check capacity vm on pe_j THEN
          schedule vm on pe_j
          pe_j − 1
        END IF
      END IF
    END FOR
  END FOR
  wait for interval t
END WHILE
```

VM Management

- Monitor Cloud usage and load.
- When load decreases:
  - Live migrate VMs to more utilized nodes.
  - Shutdown unused nodes.
- When load increases:
  - Use WOL to start up waiting nodes.
  - Schedule new VMs to new nodes.

Minimizing VM Instances

- Virtual machines are loaded!
  - Lots of unwanted packages.
  - Unneeded services.
- Are multi-application oriented, not service oriented.
  - Clouds are based off of a Service Oriented Architecture.
- Need a custom lightweight Linux VM for service oriented science.
- Need to keep VM image as small as possible to reduce network latency.

Resource Management for IaaS

- Infrastructure-as-a-Service (IaaS) is most popular cloud service
- In IaaS, cloud providers offer resources that include computers as virtual machines, raw (block) storage, firewalls, load balancers, and network devices.
- One of the major challenges in IaaS is resource management.

Resource Management - Objectives

- Scalability
- Quality of service
- Optimal utility
- Reduced overheads
- Improved throughput
- Reduced latency
- Specialized environment
- Cost effectiveness
- Simplified interface

Resource Management – Challenges (Hardware)

- CPU (central processing unit)
- Memory
- Storage
- Workstations
- Network elements
- Sensors/actuators

Resource Management – Challenges (Logical resources)

- Operating system
- Energy
- Network throughput/bandwidth
- Load balancing mechanisms
- Information security
- Delays
- APIs/(Applications Programming Interfaces)
- Protocols

Resource Management Aspects

- Resource provisioning
- Resource allocation
- Resource requirement mapping
- Resource adaptation
- Resource discovery
- Resource brokering
- Resource estimation
- Resource modeling

Resource Management

- Resource provisioning
  - Allocation of a service provider's resources to a customer
- Resource allocation
  - Distribution of resources economically among competing groups of people or programs
- Resource adaptation
  - Ability or capacity of that system to adjust the resources dynamically to fulfill the requirements of the user

Resource Management

- Resource mapping
  - Correspondence between resources required by the users and resources available with the provider
- Resource modeling
  - Resource modeling is based on detailed information of transmission network elements, resources and entities participating in the network.
  - Attributes of resource management: states, transitions, inputs and outputs within a given environment.
  - Resource modeling helps to predict the resource requirements in subsequent time intervals

Resource Management

- Resource estimation
  - A close guess of the actual resources required for an application, usually with some thought or calculation involved
- Resource discovery and selection
  - Identification of list of authenticated resources that are available for job submission and to choose the best among them.
- Resource brokering
  - It is the negotiation of the resources through an agent to ensure that the necessary resources are available at the right time to complete the objectives
- Resource Scheduling

Resource Provisioning Approaches

- Nash equilibrium approach using Game theory
- Network queuing model
- Prototype provisioning
- Resource (VM) provisioning
- Adaptive resource provisioning
- SLA oriented methods
- Dynamic and automated framework
- Optimal cloud resource provisioning (OCRP)

Resource Allocation Approaches

- Market-oriented resource allocation
- Intelligent multi-agent model
- Energy-Aware Resource allocation
- Measurement based analysis on performance
- Dynamic resource allocation method
- Real time resource allocation mechanism
- Dynamic scheduling and consolidation mechanism

Resource Mapping Approaches

- Symmetric mapping pattern
- Load-aware mapping
- Minimum congestion mapping
- Iterated local search based request partitioning
- SOA API
- Impatient task mapping
- Distributed ensembles of virtual appliances (DEVAs)
- Mapping a virtual network onto a substrate network

Resource Adaptation Approaches

- Reinforcement learning guided control policy
- Web-service based prototype
- OnTimeMeasure service
- Virtual networks
- DNS-based Load Balancing
- Hybrid approach

Performance Metrics for Resource Management

- Reliability
- Ease of deployment
- QoS
- Delay
- Control overhead