# ML/AI for Cloud Computing

Tessema Mengistu(PhD)

mengistu@vt.edu

# Outline

- Resource Management in the Cloud
- Application of ML/AI in the Cloud:
  - Cloud Resource Management
    - Demand Forecast
    - Energy consumption
  - Data Center Networking
  - Security and Compliance

# Resource Management in the Cloud

- Cloud Computing service providers usually have millions of servers in their datacenters
  - AWS had 1.4 million servers in 2014 [1]
    - 36 regions, 114 AZ - as of 4/29/2025
  - Microsoft Azure had 4 million servers in 2021[2]
  - IBM operates over 60 Cloud Data Centers in various locations around the world [3]
- Warrants intelligent automation

# Resource Management in the Cloud

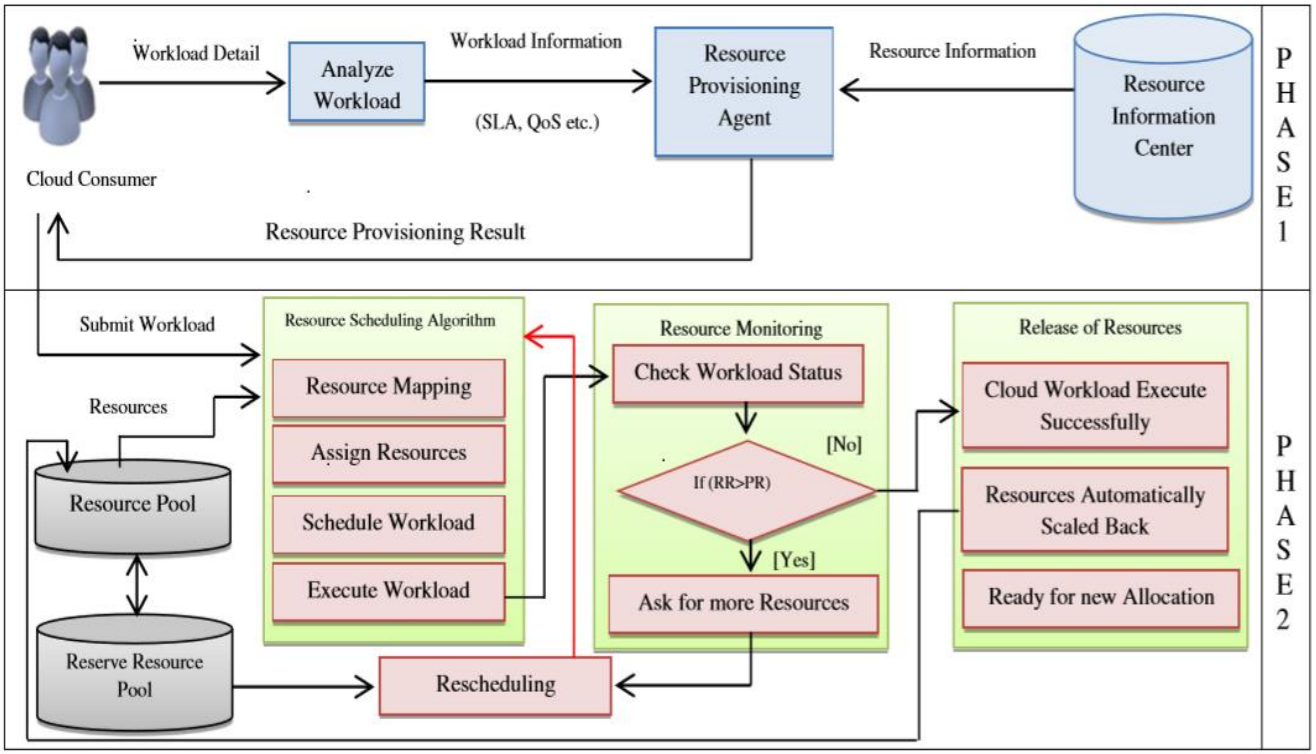
- Cloud workload
  - An abstraction of work of an instance or set of instances to be executed in the cloud infrastructure
  - For example, a web application or an interactive game
- Resource Management
  - The process of allocating resources to a set of applications (workloads), in a manner that seeks to jointly meet the performance objectives of the **applications**, the infrastructure (i.e., data center) **providers** and the **users** of the cloud resources
  - It involves :
    - Resource discovery
    - Resource provisioning
    - Resource scheduling
    - Resource monitoring

# Resource Management in the Cloud

- In general , resource management in a cloud can be stated as:
  - Given a set of resources - $\{r_1, r_2, r_3, r_4, \ldots, r_n\}$

    users‘ workloads - $\{w_1, w_2, w_3, \ldots\ldots w_m\}$,
  - Find an optimal assignment of workloads to resources taking into account the Service Level Agreement (SLA) and optimization objectives
    - The focus of optimization objectives can be:
      - Resource utilization
      - Monitory unit
      - Energy consumption
      - SLA violations
      - QoS
      - . . .

# Resource Management in the Cloud

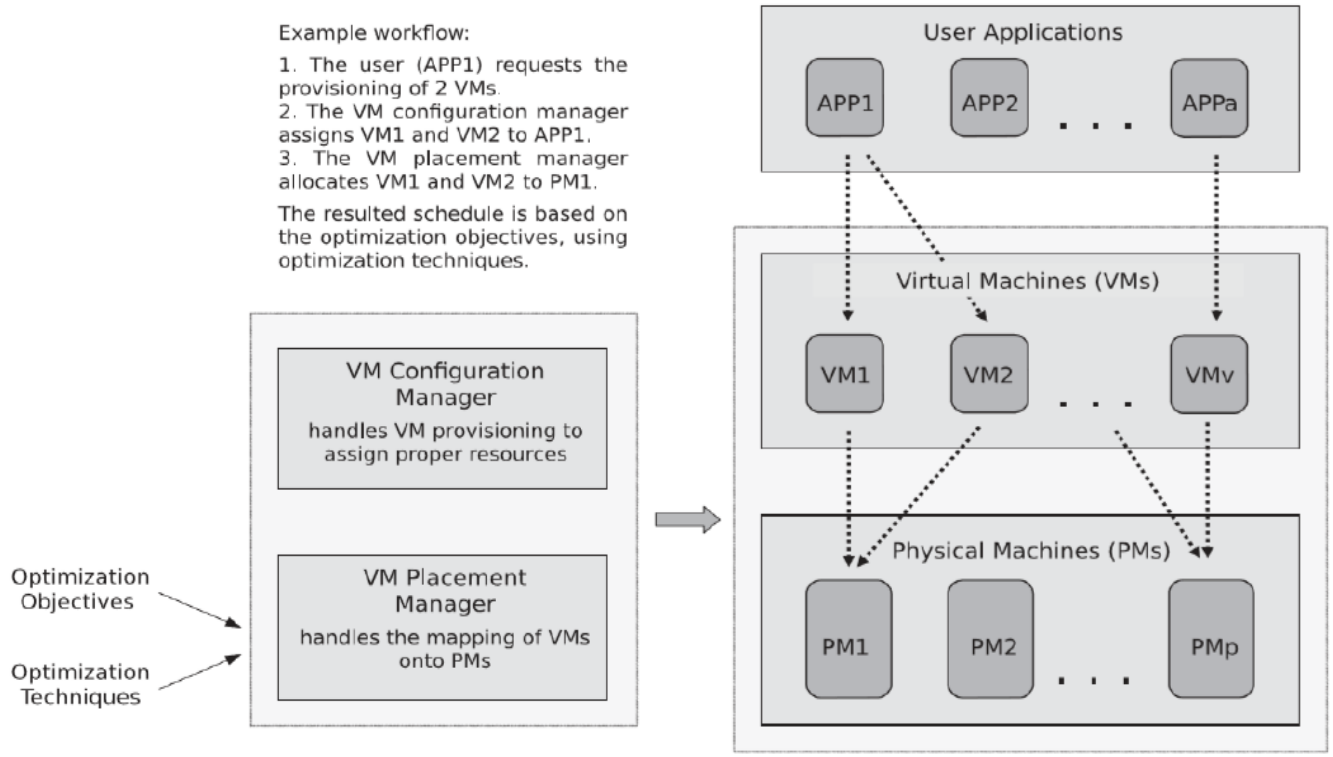



Source: https://link.springer.com/article/10.1007/s10723-015-9359-2

# Resource Management in the Cloud

- Executing workloads in the cloud involves two steps:
  - The allocation of workloads to VMs
    - Task Scheduling
    - NP-complete problem
    - Can be **static** or **dynamic** – based on resource reequipments
    - Avoid under-provisioning and/or over-provisioning.
  - The mapping of VMs to physical machines
    - aka VM placement
    - NP-complete problem
    - can be:
      - Event-driven, periodic, or hybrid
      - Proactive or reactive
      - . . .
  - Bin packing vs. knapsack

# Resource Management in the Cloud




Source: https://dl.acm.org/doi/10.1145/2983575

# Application of ML/AI: Demand Forecast

- Task scheduling
  - Managing dynamic workloads requires adaptive scheduling techniques
    - To effectively handle varying resource demands
  - Complex tasks are decomposed into multiple subtasks to form task flows
    - Allocated to processors for parallel processing
    - As certain tasks' computations depend on the results of previous tasks and priority constraints exist among tasks
      - Directed acyclic graphs (DAGs) can be utilized to abstract and model the workflow

# Application of ML/AI: Demand Forecast

- ML techniques:
  - Supervised learning
  - Deep Learning
    - Long Short-Term Memory (LSTM)  Recurrent Neural Network (RNN)
  - Reinforcement Learning
- Microsoft Azure
  - A framework, **resource central,** to provide online forecasts of different workloads using various ML Gradient Boosting Trees [Microsoft, 2020]

# Application of ML/AI: Demand Forecast

- Predicting future resource demands based on past trends and application requirements
  - CPU usage patterns
  - User behavior
  - IoT device activity
  - . . .
- A data-driven Machine Learning (ML) model - forecast future workload demand and control auto-scaling of resources accordingly
  - Intelligent VM consolidation
    - Higher Physical Machine(PM) utilization
  - Reliable host-overload detection

# Application of ML/AI: Demand Forecast

- ML techniques:
  - Supervised learning
  - Deep Learning
  - Reinforcement Learning
    - Optimal resource allocation strategies based on rewards like application performance

# Application of ML/AI: Energy Consumption

- Reducing the energy consumption of cloud computing centers with heterogeneous computing resources is a significant challenge
  - From 2011 to 2035, the energy demand on computing centers will increase by more than 66%, and the energy consumption of idle server accounts for 70% of the maximum energy consumption [ACM, 2018]
  - Data center electricity consumption in the U.S. alone is emitting nearly 100 million tons of carbon pollution per year [IEEE Tran. 2020]
    - Significant sources of energy consumption is cooling energy - about 38%

# Application of ML/AI: Energy Consumption

- Energy-saving algorithms/techniques used in data centers include:
  - Server level
    - Dynamic Voltage Frequency Scaling – DVFS
      - CPU operates at a frequency f, the energy consumption is proportional to $f^3$ , whereas the task executing time is roughly proportional to f for CPU-intensive tasks
  - Rack level
    - Servers + networking devices
  - Data center level

# Application of ML/AI: Energy Consumption

- Google optimizes fan speeds and other energy kobs using a neural network [Google, 2014]
- Estimating energy consumption while real-time running applications is a challenge
  - Ensemble learning
- ML techniques:
  - Supervised learning
    - Forecast energy consumption of workloads
  - RL
    - Rewarding energy-efficient scheduling and penalizing inefficient scheduling

# Data Center Networking

- Data center traffic has diverse communication patterns and requires efficient flow control mechanisms
  - Incast problem
    - When several hosts send data to a single receiver host
      - Congestion at the switch buffer of the receiver link
    - ML techniques:
      - Supervised learning - Support Vector Regressor, Decision tree, …
        - Dynamic buffer allocation, routing optimization, load balancing, . . .
      - RL

# Data Center Networking

- Data center traffic has diverse communication patterns and is a mix of varying flow sizes that have different objectives
  - Some flows (mice) are latency-sensitive
    - Web search and online gaming
  - Some flows (elephant) are more throughput-oriented
    - Virtual machine migration and data backup
  - ML techniques:
    - Supervised learning – classification algorithms
    - Deep learning

# Data Center Networking

- Data centers host a vast amount of user data
  - Network security is a key requirement
    - Firewalls
    - Deep Packet Inspection (DPI)
    - . . .
  - Intrusion detections systems (IDS) and traffic classification
    - ML techniques:
      - Supervised learning – classification
      - DL – Deep Packet (2019)

# Security and Compliance

- Anomaly detection
  - Recognizing unusual resource consumption patterns indicating system bottlenecks or malicious activities
- Security automation and intelligence
  - Streamline security operations and enhance threat detection and response capabilities through automated processes and intelligent decision-making
    - Automated incident response
    - Automated policy enforcement
- Identifying insider threats
  - Analyze user behavior, profile users, and process natural language to identify anomalous activities

# Security and Compliance

- Cloud compliance
  - Adhering to regulatory requirements, industry standards, and organizational policies within cloud environments to ensure data protection, privacy, and security
  - Example: GDPR, HIPAA, PCI DSS, . . .
  - Challenges
    - Diverse regulatory frameworks
    - Dynamic infrastructure
    - Security risks
      - Shared security model

# Security and Compliance

- Automated Compliance Checks
  - Automated Configuration Auditing
  - Behavioral Analytics
- Real-time Monitoring and Response
- Predictive Analytics for Compliance Risk Assessment

# Challenges

- Data availability
- Decision latency

# References


- Tahseen Khan, et al., Machine learning (ML)-centric resource management in cloud computing: A review and future directions, Journal of Network and Computer Applications, Volume 204, 2022,
  - https://www.sciencedirect.com/science/article/pii/S1084804522000649