Computing in the Cloud

Ioan Raicu
Computer Science Department
Illinois Institute of Technology

CS 553: Cloud Computing
September 15th, 2025

Logistics

- Read chapter 4, 5, 6, 7 from textbook (skipped chapter 2 and 3 for now)

Cloud Ecosystem and Enabling Technologies

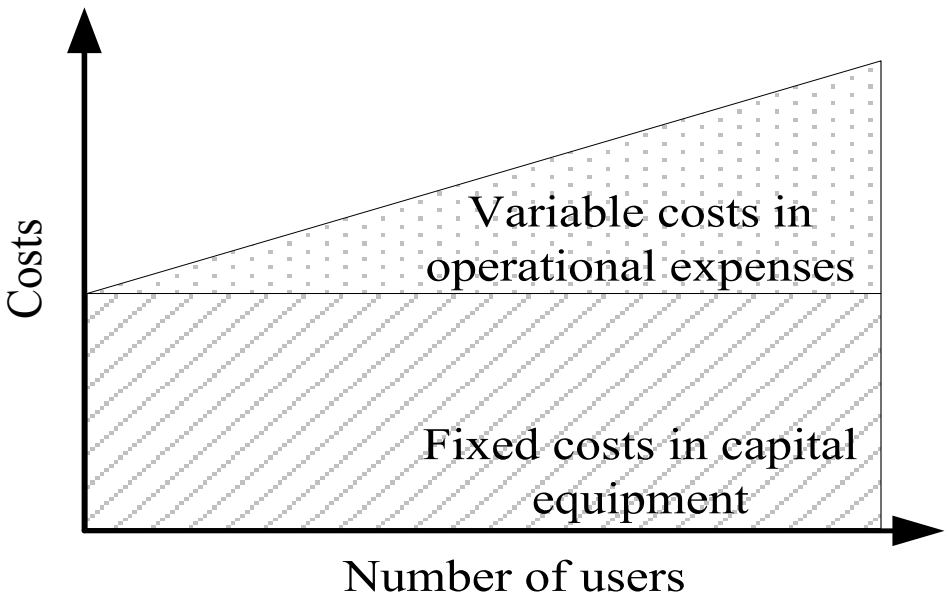
Classical Computing
(Repeat the following cycle every 18 months)

Buy and own
Hardware, system software, applications to meet peak needs
Install, configure, test, verify, evaluate, manage
- - - -
Use (Finally)
- - - -
Pay $$$$$ (High cost)

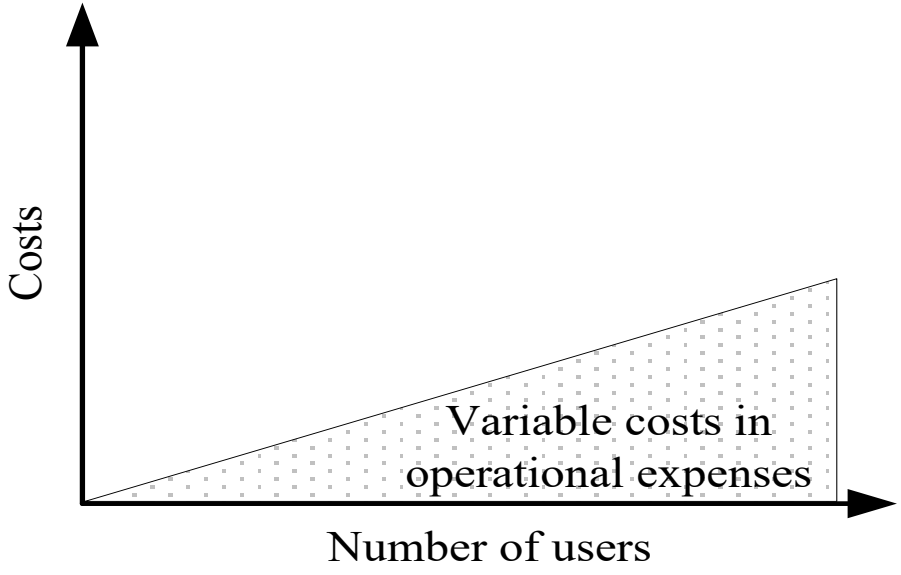
Cloud Computing
(Pay as you go per each service provided)

Subscribe
- - - -
Use (Save about 80-15% of the total cost)
- - - -
$ - Pay for what you use
based on the QoS

Cloud Ecosystem and Enabling Technologies



Traditional IT Cost Model



Cloud Computing Cost Model

Public, Private, and Hybrid Clouds

- Public Cloud:
  - Built over the Internet and can be accessed by any user who has paid for the service
  - Owned by service providers and are accessible through a subscription, typically offered on a flexible price-per-use basis
  - Examples: Google App Engine (GAE), Amazon Web Services (AWS), Microsoft Azure, IBM Blue Cloud, and Salesforce.com's Force.com
  - Promote standardization, preserve capital investment, and offer application flexibility
- Private Cloud:
- Hybrid Cloud:

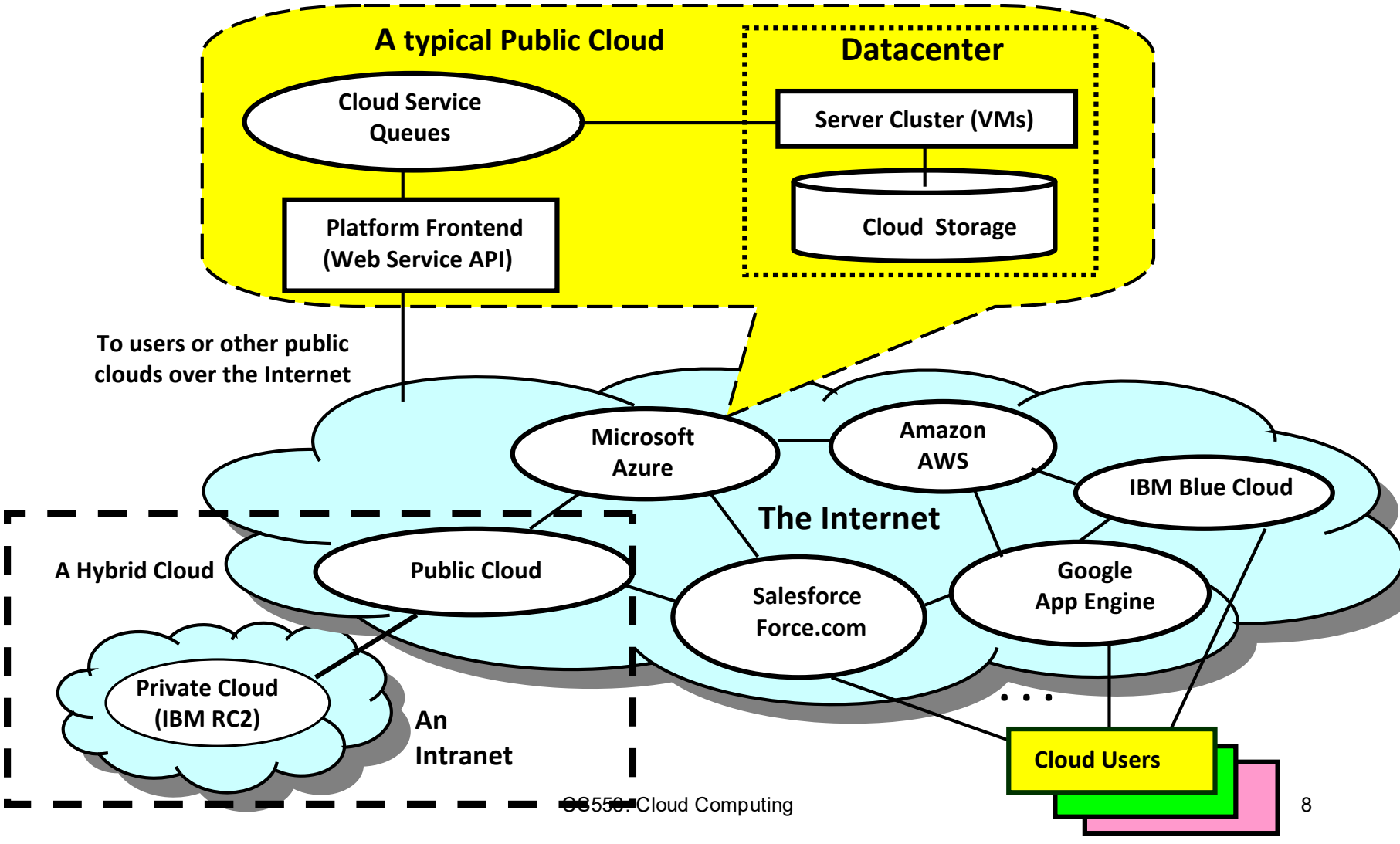
Public, Private, and Hybrid Clouds

- Public Cloud:
- Private Cloud:
  - Built within the domain of an intranet owned by a single organization
  - It is client owned and managed, and its access is limited to the owning clients and their partners
  - Its deployment was not meant to sell capacity over the Internet through publicly accessible interfaces
  - Private clouds give local users a flexible and agile private infrastructure to run service workloads within their administrative domains
  - Aimed to deliver more efficient and convenient cloud services
  - Example: *Research Compute Cloud* (RC2)
    - Built by IBM interconnecting the computing and IT resources at eight IBM Research Centers scattered throughout the United States, Europe, and Asia
  - Attempt to achieve customization and offer higher efficiency, resiliency, security, and privacy
- Hybrid Cloud:

Public, Private, and Hybrid Clouds

- Public Cloud:
- Private Cloud:
- Hybrid Cloud:
  - Built with both public and private clouds
  - Private clouds can also support a hybrid cloud model by supplementing local infrastructure with computing capacity from an external public cloud
  - Provides access to clients, the partner network, and third parties
  - Operate in the middle between private and public clouds, with many compromises in terms of resource sharing

Public, Private, and Hybrid Clouds



Amazon EC2

- Amazon was the first company to introduce VMs in application hosting
- Customers can rent VMs instead of physical machines to run their own applications
- By using VMs, customers can load any software of their choice
- The elastic feature of such a service is that a customer can create, launch, and terminate server instances as needed, paying by the hour for active servers
- Amazon provides several types of preinstalled VMs
- Instances are often called *Amazon Machine Images (AMIs)* which are preconfigured with operating systems based on Linux or Windows, and additional software

Amazon EC2 Execution Environment

- Three types of AMI

Image Type	AMI Definition
Private AMI	Images created by you, which are private by default. You can grant access to other users to launch your private images.
Public AMI	Images created by users and released to the AWS community, so anyone can launch instances based on them and use them any way they like. AWS lists all public images at http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=171.
Paid QAMI	You can create images providing specific functions that can be launched by anyone willing to pay you per each hour of usage on top of Amazon's charges.

Amazon EC2 Execution Environment

- Types of instances
  - **Standard instances** are well suited for most applications.
  - **Micro instances** provide a small number of consistent CPU resources and allow you to burst CPU capacity when additional cycles are available. They are well suited for lower throughput applications and Web sites that consume significant compute cycles periodically.
  - **High-memory instances** offer large memory sizes for high-throughput applications, including database and memory caching applications.
  - **High-CPU instances** have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications.
  - **Cluster compute instances** provide proportionally high CPU resources with increased network performance and are well suited for high-performance computing (HPC) applications and other demanding network-bound applications. They use 10 Gigabit Ethernet interconnections.

Logistics

- HW3 (Understanding the Cost of Cloud Computing) deadline extension to Friday 9/26
- HW4 will be posted Friday 9/26

Seminars

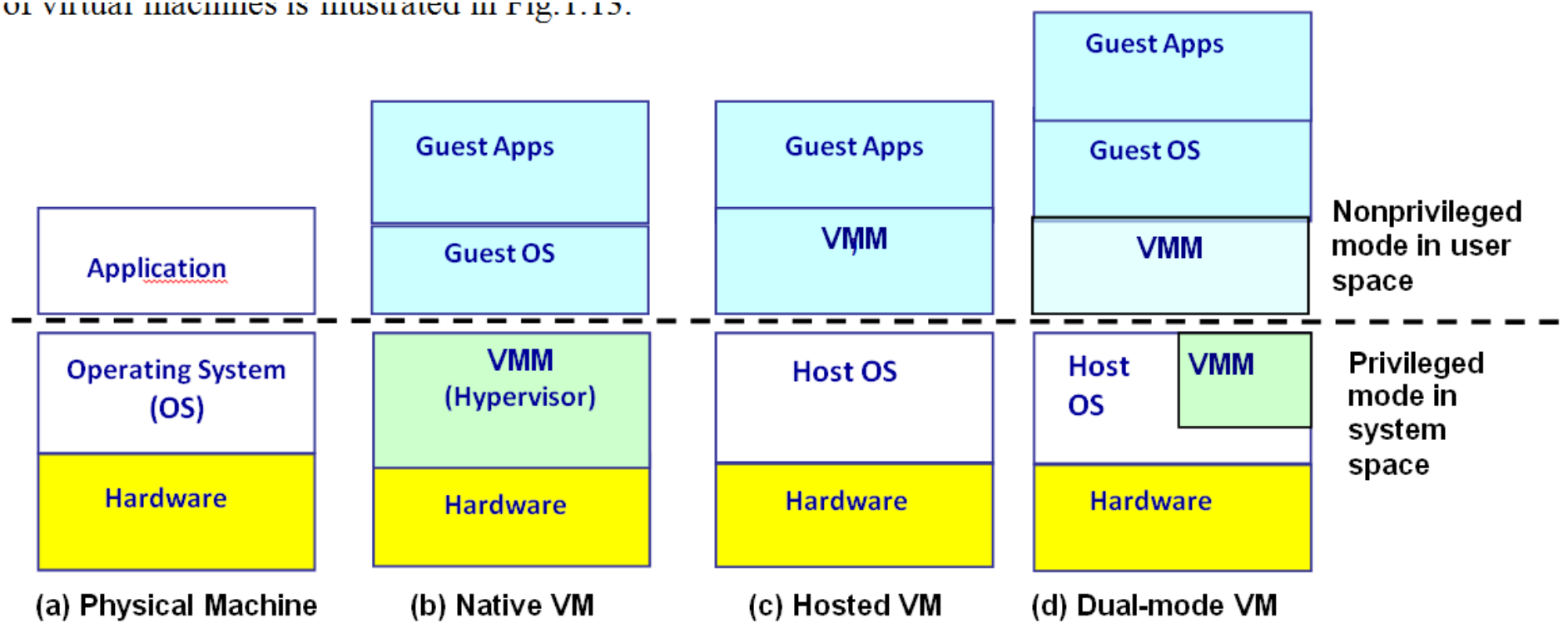
- Tuesday, September 30, 2025
  - Who: Tyler J. Skluzacek (https://tylerskluzacek.com)
  - From: Oak Ridge National Lab
  - Research area: Workflow and Ecosystem Services
- Tuesday, October 7, 2025
  - Who: Simone Silvestri (https://silvestri.engr.uky.edu/)
  - From: University of Kentucky
  - Research area: Internet of Things + Network Management
- Monday, October 20, 2025
  - Who: Alexandru Orhean (https://www.cdm.depaul.edu/Faculty-and-Staff/Pages/faculty-info.aspx?fid=1648)
  - From: DePaul
  - Research area: Distributed Systems

Types of Instances

- On-demand
- Spot instances
- Reserved instances
- Dedicated hosts
- Billing granularity

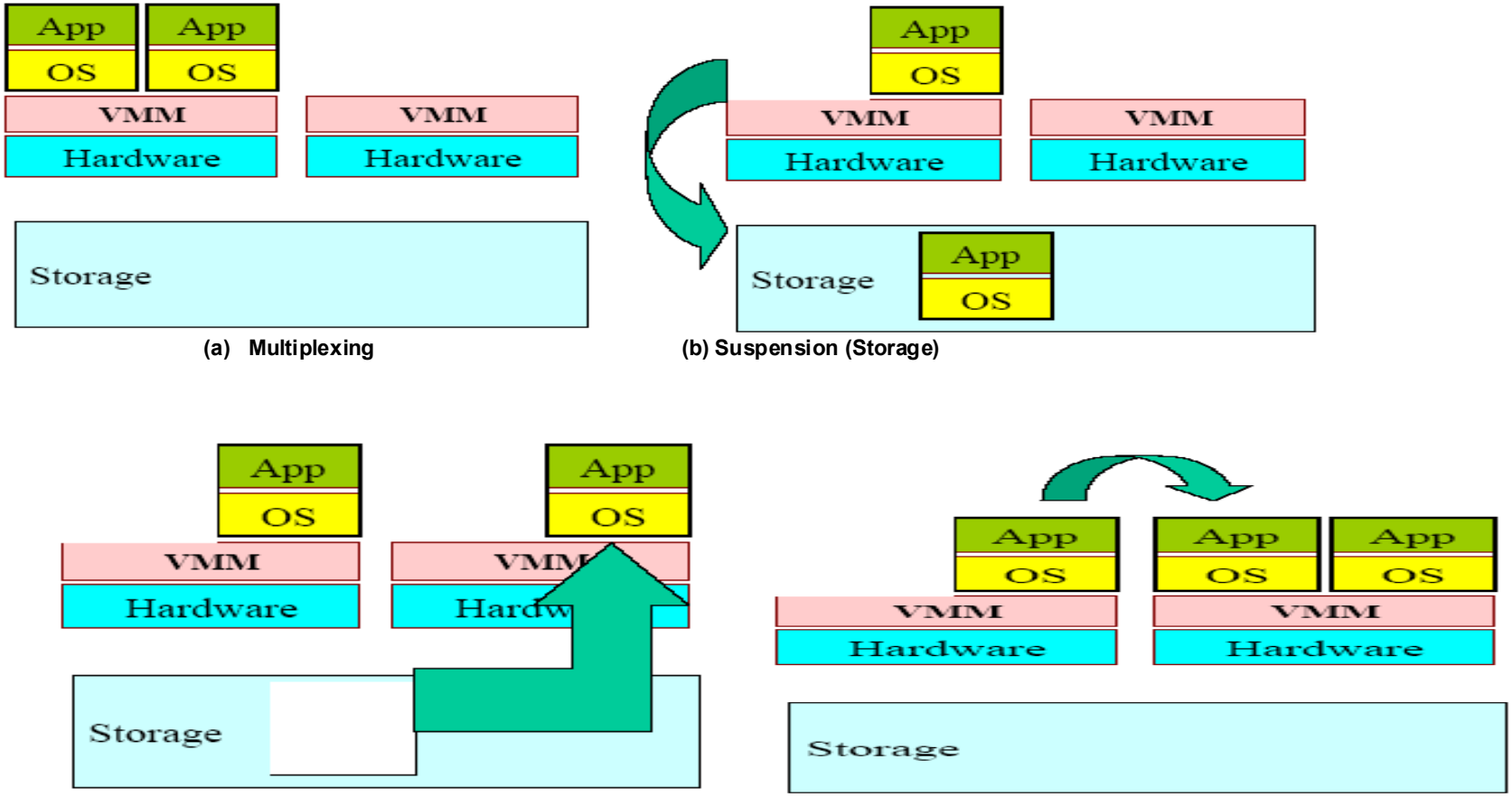
Virtual Machines and Virtualization Middleware

of virtual machines is illustrated in Fig.1.13.



- Three virtual machine (VM) architectures in Parts (b-d), compared with the traditional physical machine shown in Part (a)

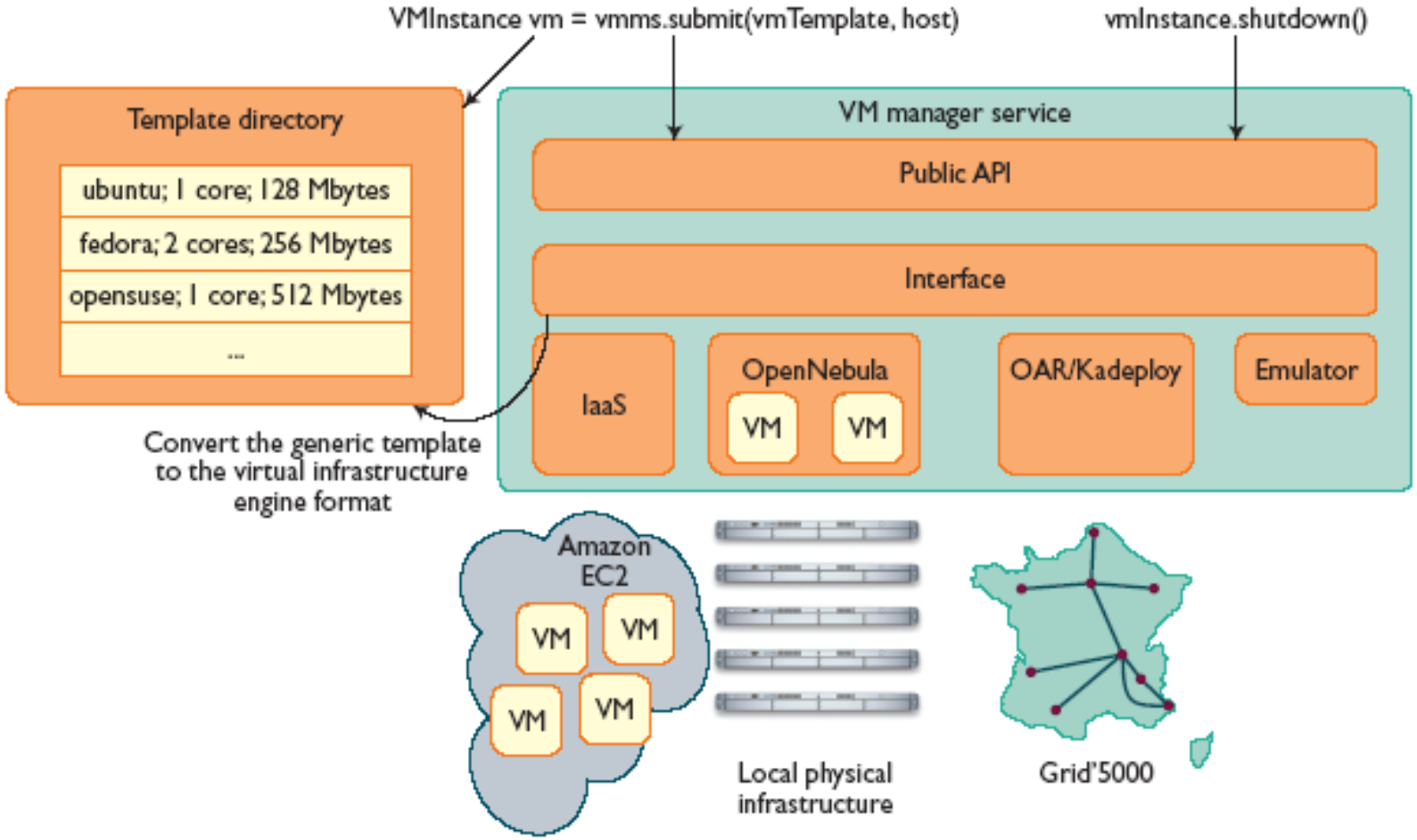
Virtual Machines and Virtualization Middleware



- Virtual machine multiplexing, suspension, provision, and migration in a distributed computing environment

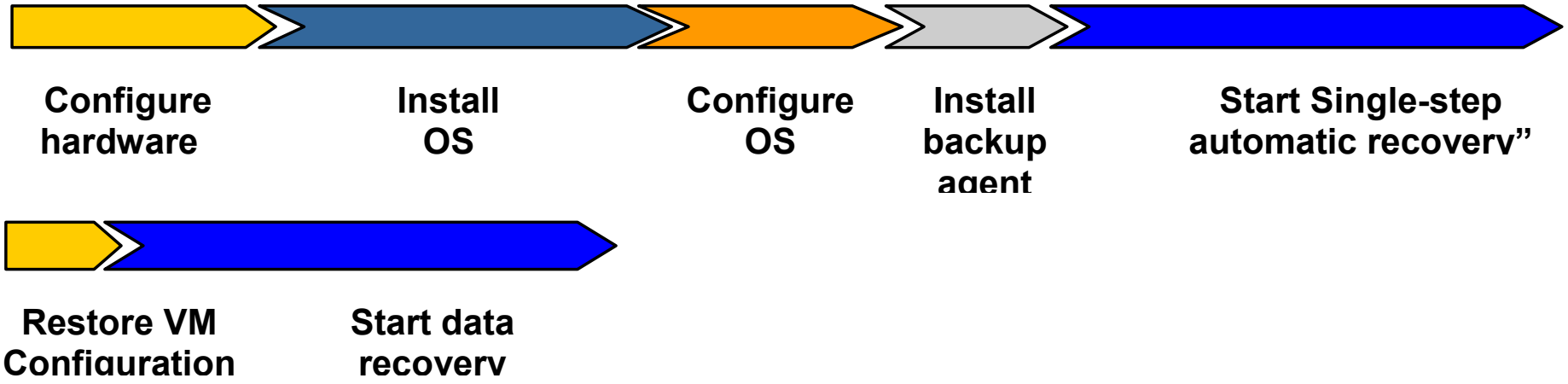
Virtual Machine Creation and Management

- Interactions among VM managers for cloud creation and management
- The manager provides a public API for users to submit and control the VMs



Virtualization Support and Disaster Recovery

- Recovery overhead of a conventional disaster recovery scheme, compared with that required to recover from live migration of VMs

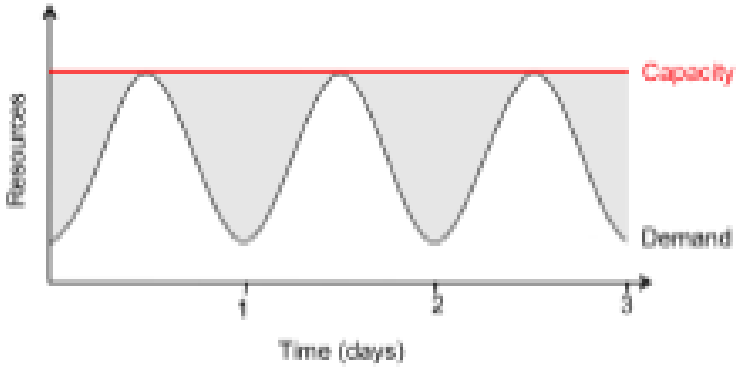


Resource Provisioning and Platform Deployment

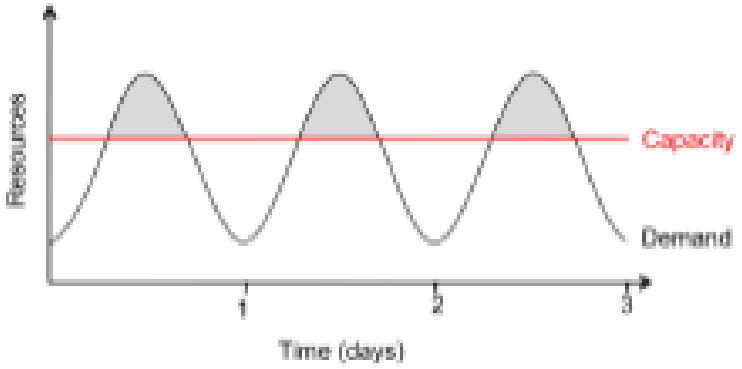
- Providers supply cloud services by signing SLAs with end users
- The SLAs must commit sufficient resources such as CPU, memory, and bandwidth that the user can use for a preset period
- Under-provisioning of resources will lead to broken SLAs and penalties
- Overprovisioning of resources will lead to resource underutilization, and consequently, a decrease in revenue for the provider
- Deploying an autonomous system to efficiently provision resources to users is a hard problem
  - The difficulty comes from the unpredictability of consumer demand, software and hardware failures, heterogeneity of services, power management, and conflicts in signed SLAs between consumers and service providers

Resource Provisioning and Platform Deployment

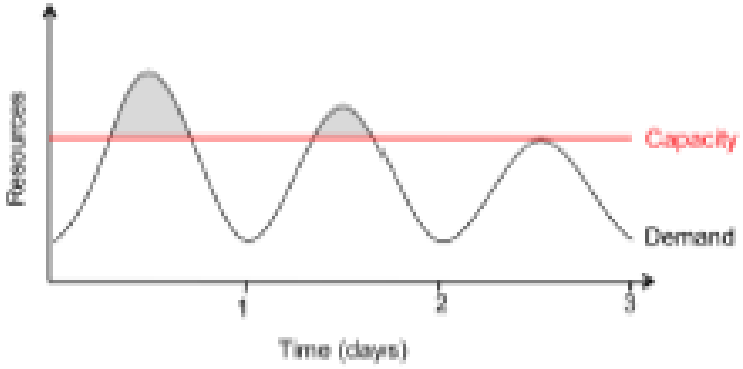
- Three cases of cloud resource provisioning without elasticity: (a) heavy waste due to overprovisioning, (b) underprovisioning and (c) under- and then overprovisioning



(a) Provisioning for peak load



(b) Underprovisioning 1



(c) Underprovisioning 2

Resource Provisioning and Platform Deployment

- **Demand-Driven Resource Provisioning:**
  - This method adds or removes computing instances based on the current utilization level of the allocated resources
  - For example: the demand-driven method automatically allocates two Xeon processors for the user application, when the user was using one Xeon processor more than 60 percent of the time for an extended period
  - In general, when a resource has surpassed a threshold for a certain amount of time, the scheme increases that resource based on demand
  - When a resource is below a threshold for a certain amount of time, that resource could be decreased accordingly
  - Amazon implements such an auto-scale feature in its EC2 platform
  - Pros: This method is easy to implement
  - Cons: The scheme does not work out right if the workload changes abruptly

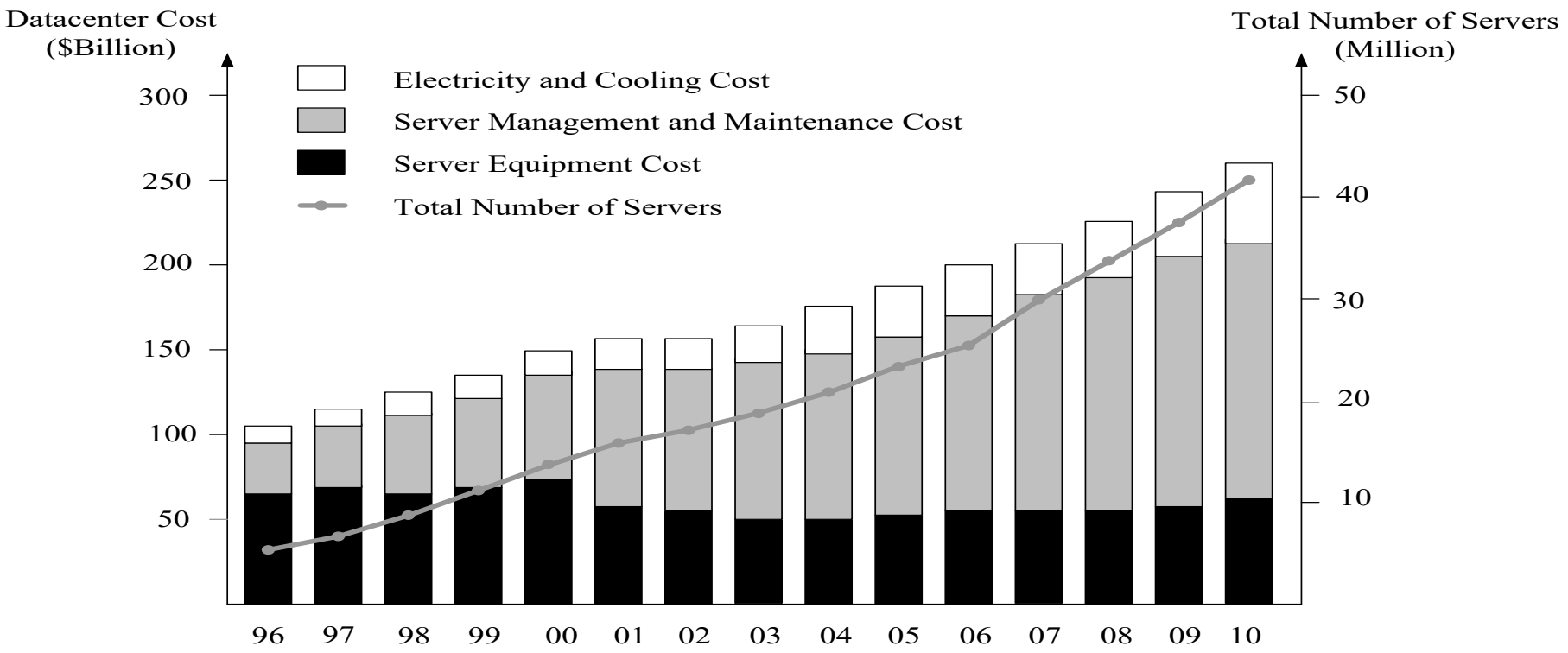
Resource Provisioning and Platform Deployment

- **Event-Driven Resource Provisioning:**
  - This scheme adds or removes machine instances based on a specific time event
  - The scheme works better for seasonal or predicted events such as Christmastime in the West and the Lunar New Year in the East
  - During these events, the number of users grows before the event period and then decreases during the event period
  - This scheme anticipates peak traffic before it happens
  - The method results in a minimal loss of QoS, if the event is predicted correctly
  - Otherwise, wasted resources are even greater due to events that do not follow a fixed pattern

Resource Provisioning and Platform Deployment

- **Popularity-Driven Resource Provisioning:**
  - Popularity is determined from the Internet, and instances are created by popularity demand
  - The scheme anticipates increased traffic with popularity
  - The scheme has a minimal loss of QoS, if the predicted popularity is correct
  - Resources may be wasted if traffic does not occur as expected

Datacenter Virtualization for Cloud Computing



- Growth and cost breakdown of datacenters over the years

Data-Center Design and Interconnection Networks

- A data center is often built with a large number of servers through a huge interconnection network
- Study the design of large-scale data centers and small modular data centers that can be housed in a 40-ft truck container
- Study the interconnection of modular data centers and their management issues and solutions

Warehouse-Scale Data-Center Design

- A huge data center that Is 11 times the size of a football field, housing 400,000 to 1 million servers

Warehouse-Scale Data-Center Design

- Data centers are built economics of scale - meaning lower unit cost for larger data centers
- Small data centers could have 1,000 servers
- The larger the data center, the lower the operational cost
- The approximate monthly cost to operate a 400-server data center is estimated by:
  - Network cost $13/Mbps
  - Storage cost $0.4/GB
  - Administration costs
- The network cost to operate a small data center is about seven times greater and the storage cost is 5.7 times greater

Warehouse-Scale Data-Center Design

- **Data-Center Construction Requirements:**
  - Most data centers are built with commercially available components
  - An off-the-shelf server consists of a number of processor sockets, each with a multicore CPU and its internal cache hierarchy, local shared and coherent DRAM, and a number of directly attached disk drives
  - The DRAM and disk resources within the rack are accessible through first-level rack switches and all resources in all racks are accessible via a cluster-level switch
  - Consider a data center built with 2,000 servers, each with 8 GB of DRAM and four 1 TB disk drives. Each group of 40 servers is connected through a 1 Gbps link to a rack-level switch that has an additional eight 1 Gbps ports used for connecting the rack to the cluster-level switch.
- Disk bandwidth changes drastically between local and off-rack access
  - Local disks is 200 MB/s, whereas the bandwidth from off-rack disks is 25 MB/s via shared rack uplinks
  - The total disk storage in the cluster is almost 10 million times larger than local DRAM
  - A large application must deal with large discrepancies in latency, bandwidth, and capacity
- In a very large-scale data center, components are relatively cheaper, and are very different from those in building supercomputer systems

Upcoming Seminars

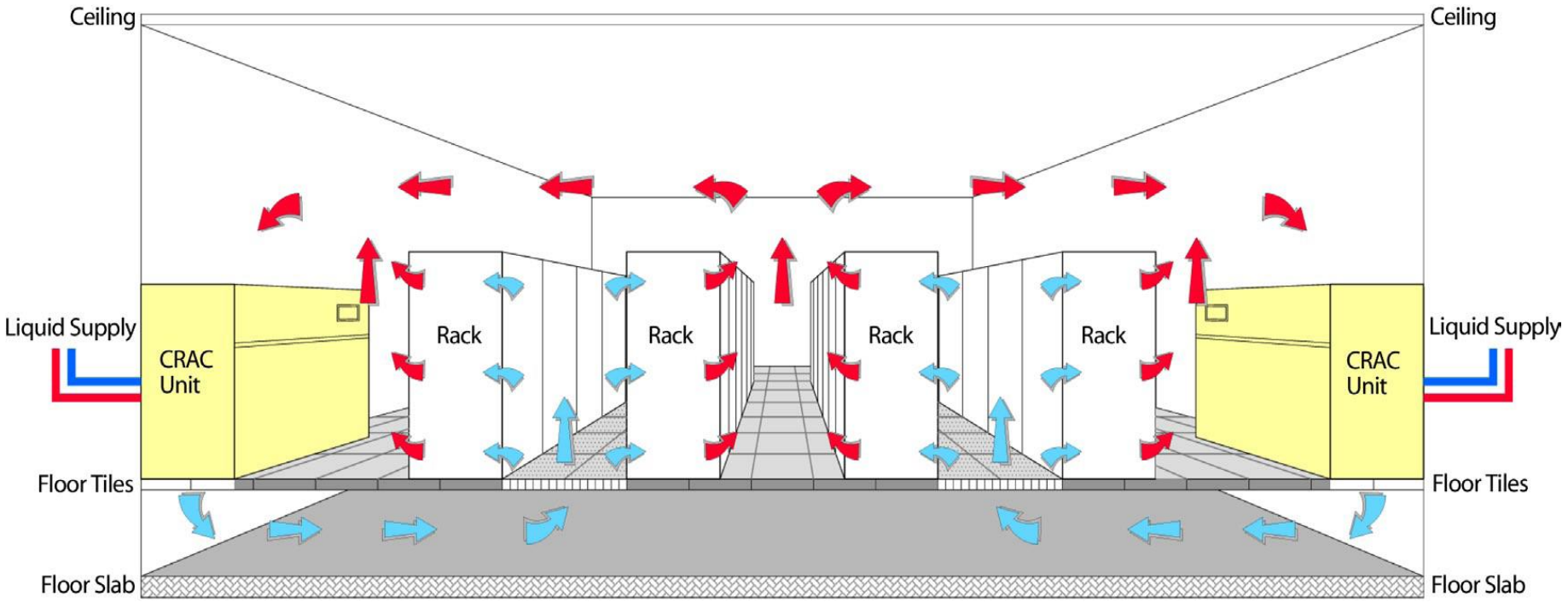
- Today
  - Towards Secure and Safe AI-enabled Systems Through Optimizations
  - Guanhong Tao, Computer Science @ Purdue University

Upcoming Homeworks

- HW4 on benchmarking, due 3/18
- HW5 (external sort) & HW6 (Hadoop/Spark) ~ 1.5 weeks each
- HW7 (scheduling & load balancing) ~ 2 weeks
- No HW8, will adjust weight of final exam to 40% and homeworks down to 60%

Warehouse-Scale Data-Center Design

- **The cooling system in a raised-floor data center with hot-cold air circulation supporting water heat exchange facilities**

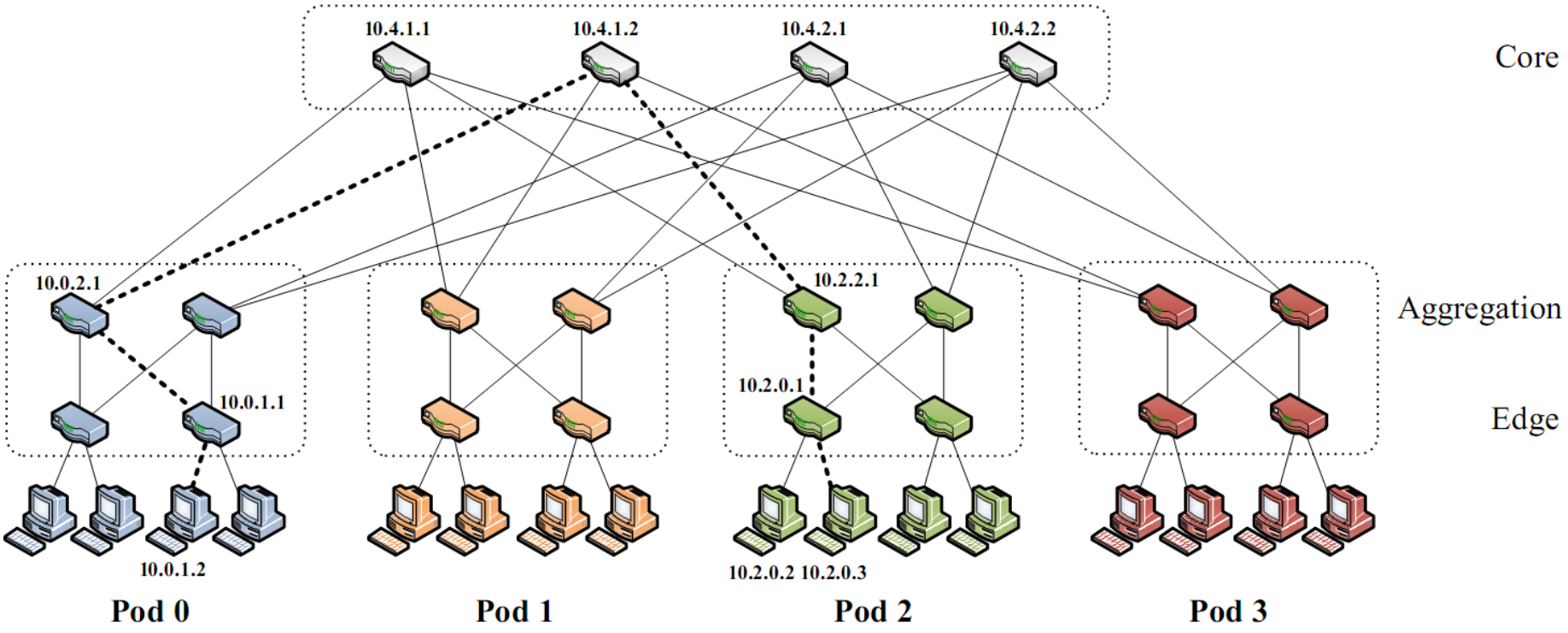


Warehouse-Scale Data-Center Design

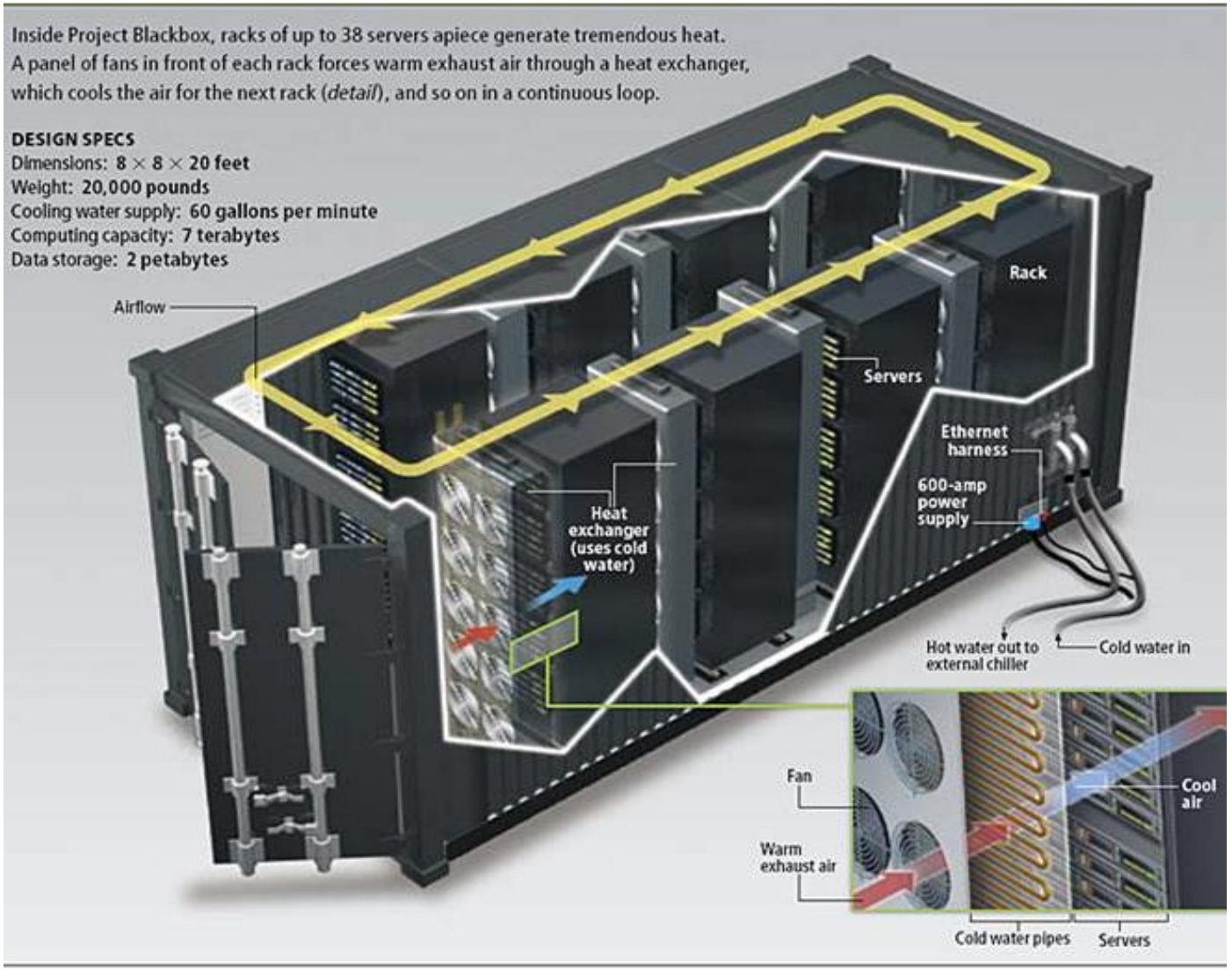
- With a scale of thousands of servers, concurrent failure, either hardware failure or software failure, of 1 percent of nodes is common
- Many failures can happen in hardware
  - CPU failure, disk I/O failure, and network failure
- It is possible that the whole data center does not work in the case of a power crash
- Some failures are software related
- The service and data should not be lost in a failure situation
- Reliability can be achieved by redundant hardware
- The software must keep multiple copies of data in different locations and keep the data accessible while facing hardware or software errors

Data-Center Interconnection Networks

- **A fat-tree interconnection topology for scalable data-center construction**



Modular Data Center in Shipping Containers



Questions