Cloud and Machine Learning

CSCI-GA.3033-085 Spring 2024

Prof. Hao Yu

Prof. I-Hsin Chung

Welcome

- Introductions

- Goals of the course
- Logistics and administrative items
- Syllabus
- Introduction to cloud computing and ML on Cloud

Who are we?

- I-Hsin Chung, Ph.D. (ihchung@nyu.edu)
- Research Staff Member, Manager @ IBM T. J. Watson Research Center
- Research interests: Cloud computing, High Performance Computing and Disaggregated Systems
- 2o+ years of Experience
- Enjoy working with students

- Hao Yu, Ph.D. (hy2467@nyu.edu)
- Research Staff Member @ IBM T. J. Watson Research Center
- Research interests: Operating system, Compiler, and Performance analysis
- 20+ years of Experience
- Loves to coach soccer and Science Olympiad

Introduce yourself

- Your name and preferred name
- What do you do?
- What stage in your program, which school?
- What is the best class you have ever taken and why is it the best class?

Welcome

- Introductions
- Goals of the course
- Logistics and administrative items
- Syllabus
- Introduction to cloud computing and ML on Cloud

What this course is about…

- It is about how to build cloud systems for machine learning workloads
  - The internals of cloud computing systems
  - The unique characteristics of machine learning workloads
  - It is not about how to use cloud for general application development (but you will use it for machine learning training projects)
  - It is not about the details of machine learning (e.g., matrix calculations of the deep neural network)
- We'll learn about machine learning on existing cloud systems
- We'll discuss architecture design considerations for cloud systems to enable machine learning
- We'll build and use cloud service to run machines learning workloads

Goal of study

- Understand the structure of the cloud computing
- Understand the machine learning workload characteristics and the considerations for computation environment/performance needed
- Through active participation, have a broad view of current offerings of cloud computing with focus on machine learning
- With the projects, get hands-on experience to apply/verify the knowledge learned in class
- Enough background knowledge needed to dive deeper into specific areas/topics of cloud computing and machine learning

Welcome

- Introductions
- Goals of the course
- Logistics and administrative items
- Syllabus
- Introduction to cloud computing and ML on Cloud

Logistics

- Pre-reqs:
  - FUNDAMENTAL ALGORITHMS
  - PROGRAMMING LANGUAGES
  - OPERATING SYSTEMS
  - PYTHON PROGRAMMING
  - Linux skills and software tools
- Workload: Approximately 10 hours per week in addition to class time
- Material
  - Most material is made available electronically, no single text book
  - Bring laptop to lecture for hands-on or potential quiz.

Logistics

- Communication: Slack
  - https://nyuspring24cloudml.slack.com
- GA information:
  - Hao Li (hl5262@nyu.edu) and Chanukya Vardhan Gujjula (crg9968@nyu.edu)
- Office hours:
  - 4:05pm Thursday at WWH 308 in person, or
  - by appointment (ping on slack and can discuss via zoom https://nyu.zoom.us/j/95081079803)
- Course website: https://cs.nyu.edu/courses/spring24/CSCI-GA.3033-085/
- Project groups and rules:
  - **1-2** students per group (class presentation and term/2nd project)
  - Team can reshape for different projects/presentation.
  - Instructors will assign the groups, if students cannot find a partner.

Survey info

- Experience with
  - Programming
  - ML/DL
  - Cloud
  - Systems knowledge
  - Linux
  - Shell scripting
- https://docs.google.com/forms/d/e/1FAIpQLScNqRq5EcwatmaLVjq-iDpgZKVm5Hz5dcu_O5-QVMW_RDEtOw/viewform

Grading

- Class participation: **10%**
- Home works (5): **30% (every HM requires a report:** good writing etiquette**)**
- Class presentation(s): **10%** 1-2 students per group
- Projects: **50%**
  - Project 1:
    Performance analysis and modeling of DNN workload (**20%**), report required.
  - Project 2 (term project):
    Identify a problem that DNN training can be applied. Implement the training in the cloud computing (**30%**)
    - Apply the techniques learned from the semester: Cloud, DNN, Performance analysis.
    - 1 page proposal (optional) for feedback.
    - Graded material: group presentation + report.

Grading

- For the report – quality rather than quantity is the key
- Grading is mostly subjective, will provide rubrics
  - Re-grading request needs mis-grading evidence.
- Deadline extension w/o penalty: only for "**unexpected**" incidents

Submission time	Penalty
On time	Full credits - No penalty
Within 3 days	5%
Up to 1 week	10%
Every week after 1 week	Additional 10% per week
More than 3 weeks	30% (Welcome "eventual" submission)

Welcome

- Introductions
- Goals of the course
- Logistics and administrative items
- Syllabus
- Introduction to cloud computing and ML on Cloud

Syllabus

- Introduction to cloud computing
- Introduction to machine learning on the cloud: Domains, Frameworks, Use cases
- Getting started with ML/DL on the cloud
- Infrastructure: Data center organization
- Orchestration: Container/Docker, Kubernetes
- Deep learning, machine learning (services)
- Performance characterization and Analysis of DL workloads
- Scalable/Distributed deep learning
- Desegregated systems and computing
- Invited talks about one of these topics (Machine learning, data oriented systems,...)
- class and project presentations

Homework/Project Environments

- We will use Google Cloud for projects and homework.
- We will use NYU HPC IT resources for some projects and home works.
- For project #2: we may collaborate with External professionals for projects
- We may add IBM Cloud for later homework/projects.

Welcome

- Introductions
- Goals of the course
- Logistics and administrative items
- Syllabus
- Introduction to cloud computing and ML on Cloud

What is cloud computing?

- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

What is cloud computing?

- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

What is cloud computing?

- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

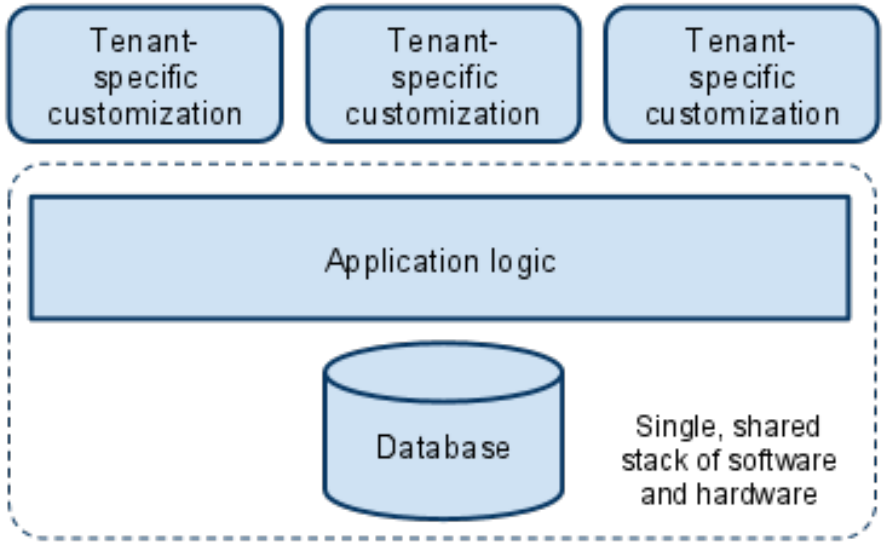
What is cloud computing?

- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

Multi-tenancy

- In a multi-tenant environment, a **single application** can be used and **customized** by **different organization** as if they each have a separate instance.



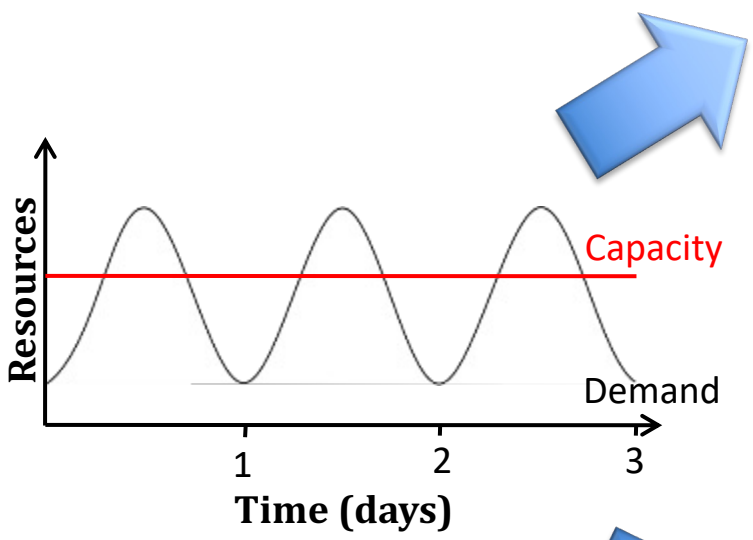
Slide credit: Che-Rung Lee

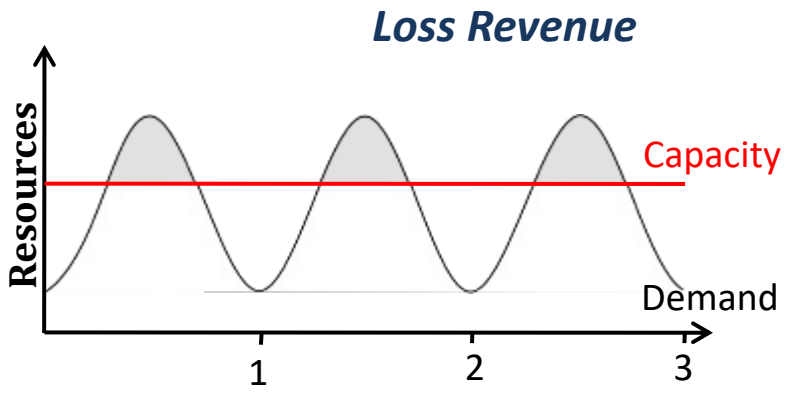
What is cloud computing?

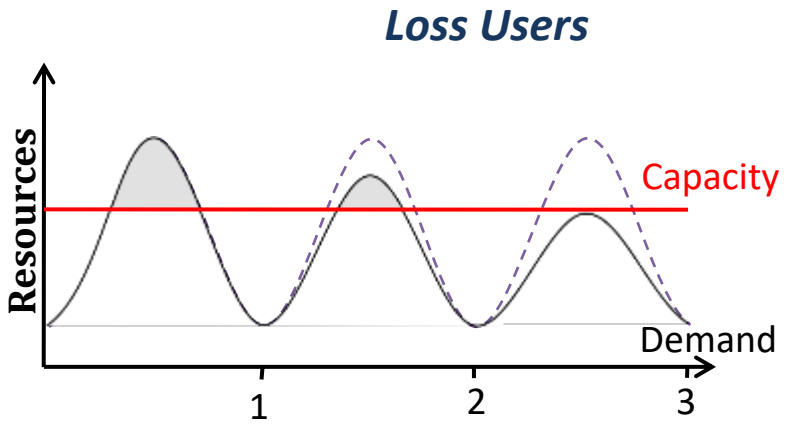
- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

Under provision

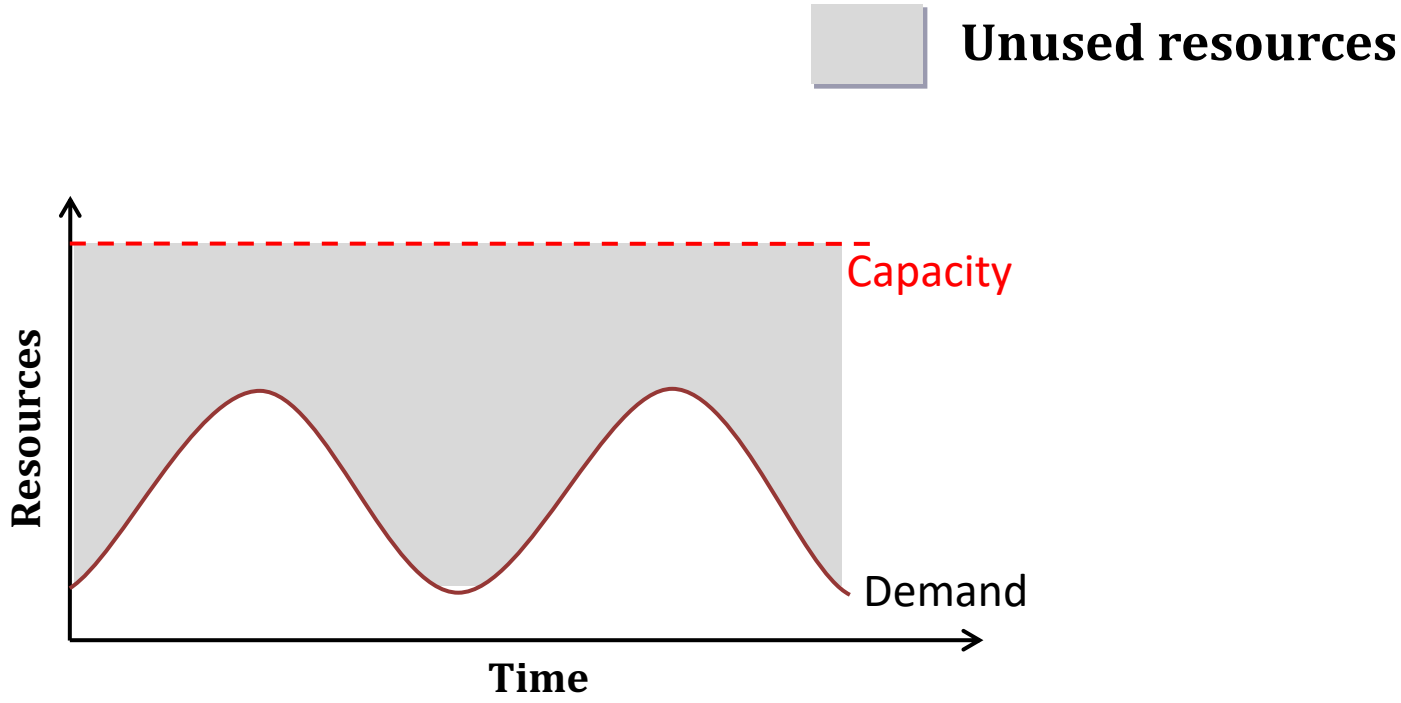






Slide credit: Che-Rung Lee

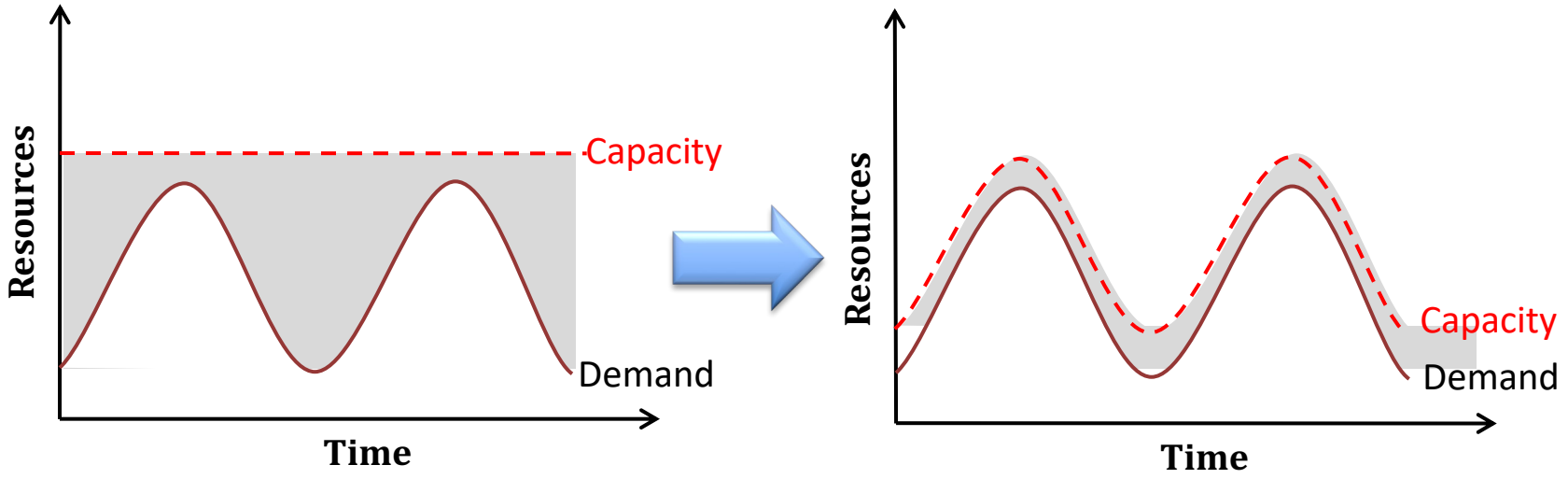
Low utilization



Slide credit: Che-Rung Lee

Dynamic provisioning

- Target for efficient hosting: the computational resources can be adjusted dynamically



Slide credit: Che-Rung Lee

What is cloud computing?

- NIST definition: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.”*

* https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

Cloud computing essential elements

- *On-demand self-service*
- *Broad network access*
- *Resource pooling*
- *Rapid elasticity*
- *Measured service*

Three service models

Cloud Infrastructure as a Service (IaaS).

- To provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software.

Cloud Platform as a Service (PaaS).

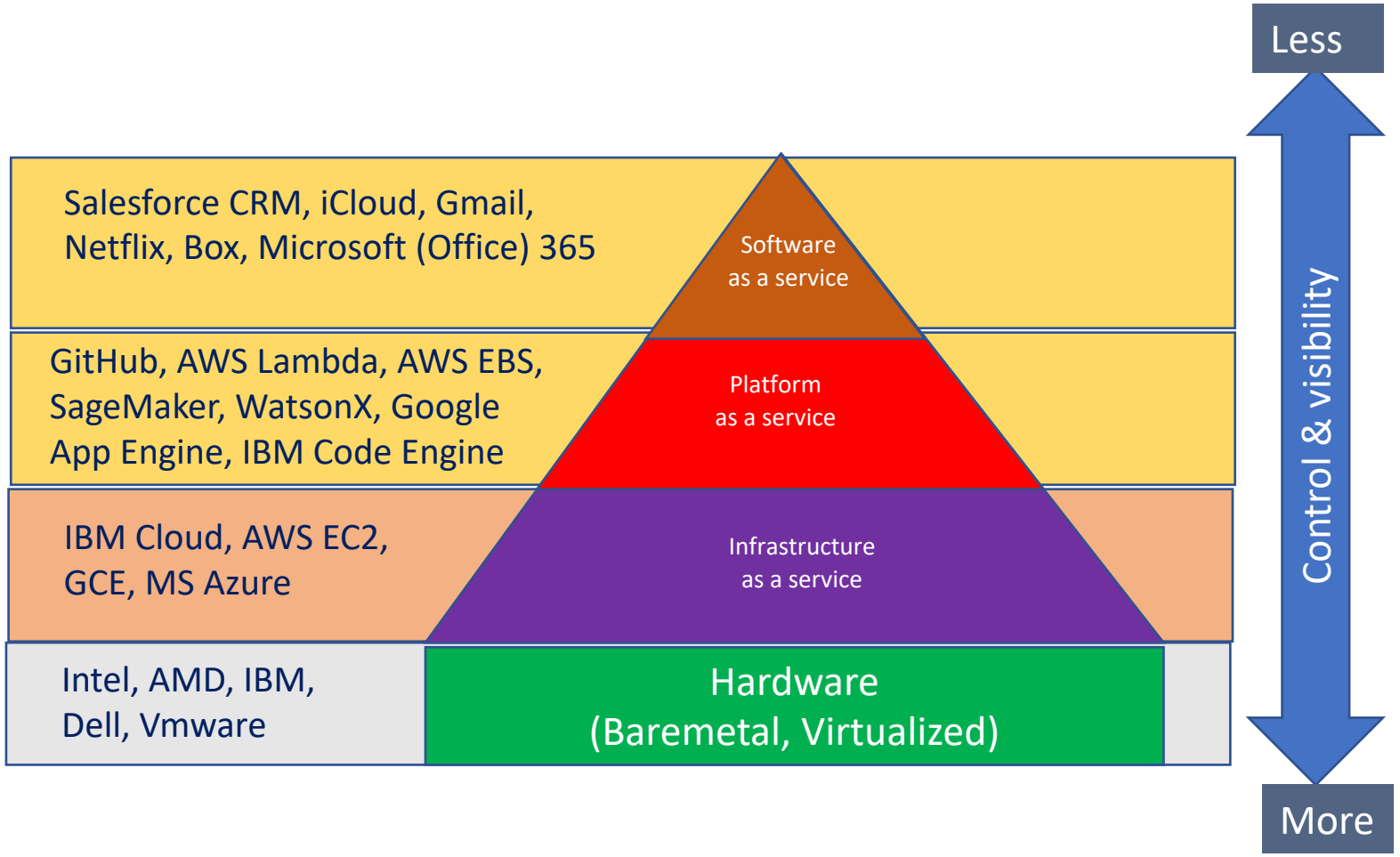
- To deploy onto the cloud infrastructure consumer-created or acquired applications.

Cloud Software as a Service (SaaS).

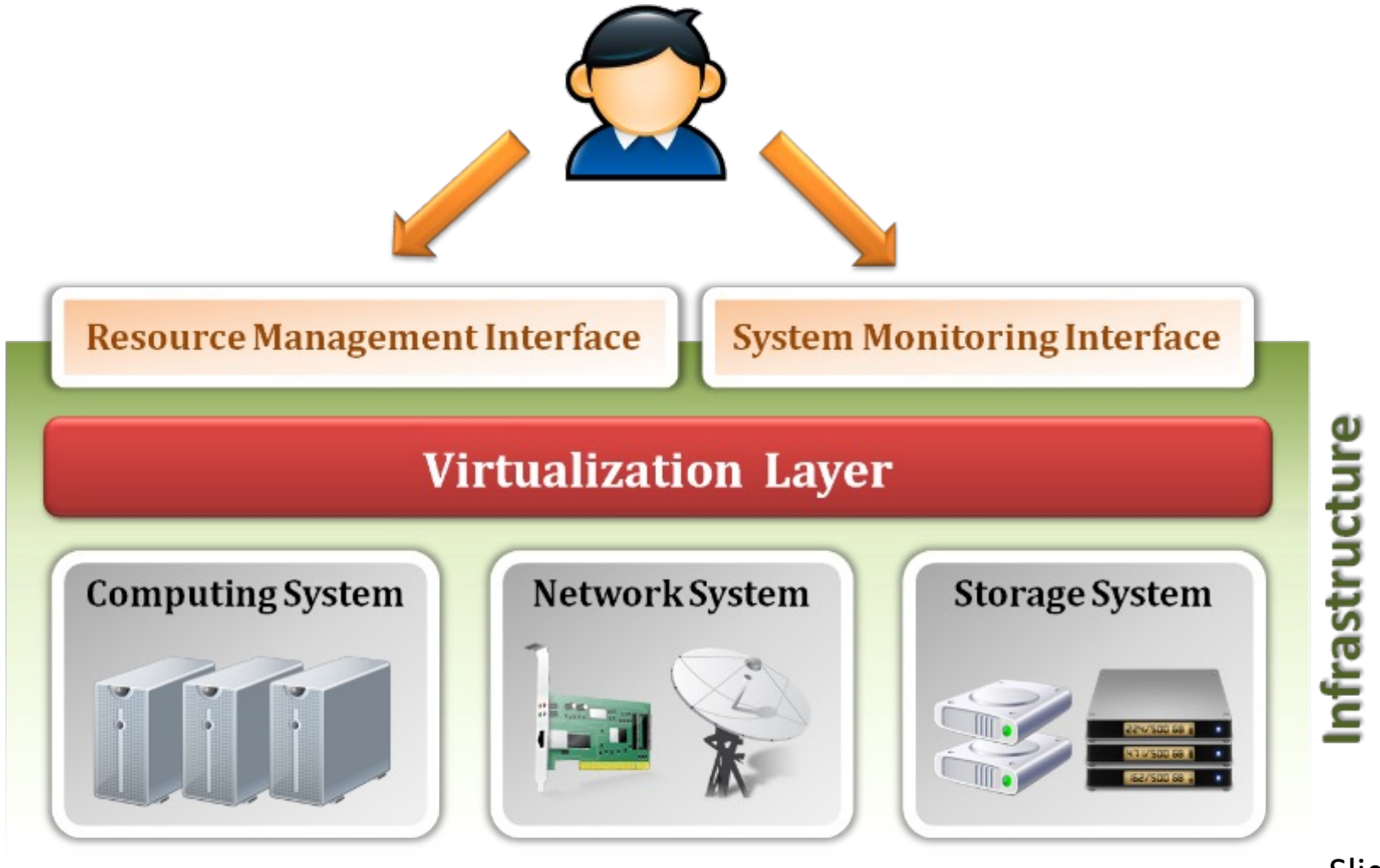
- To use the provider's applications running on a cloud infrastructure.

Slide credit: Che-Rung Lee

Cloud Computing Models: IaaS, PaaS, SaaS

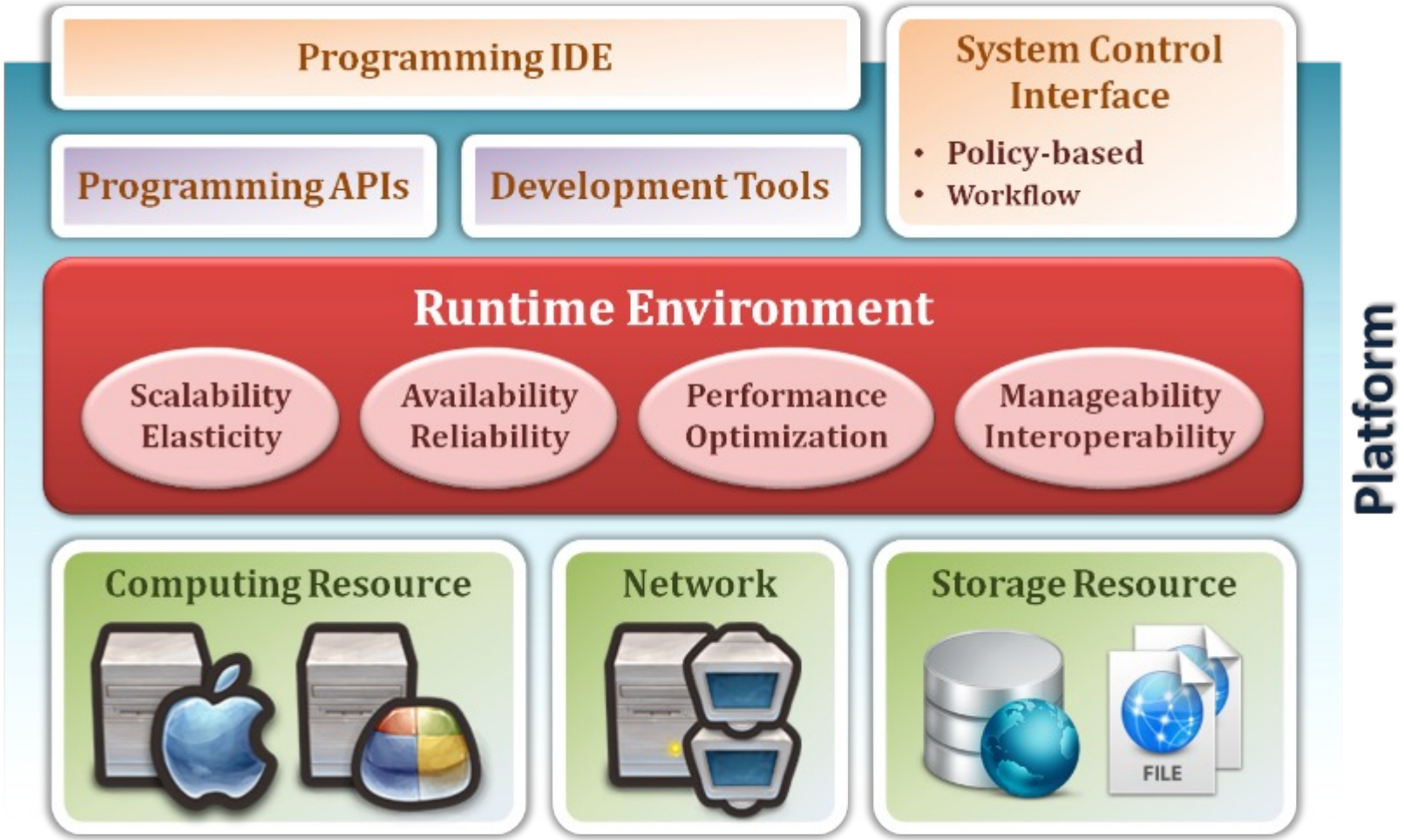


Infrastructure as a Service



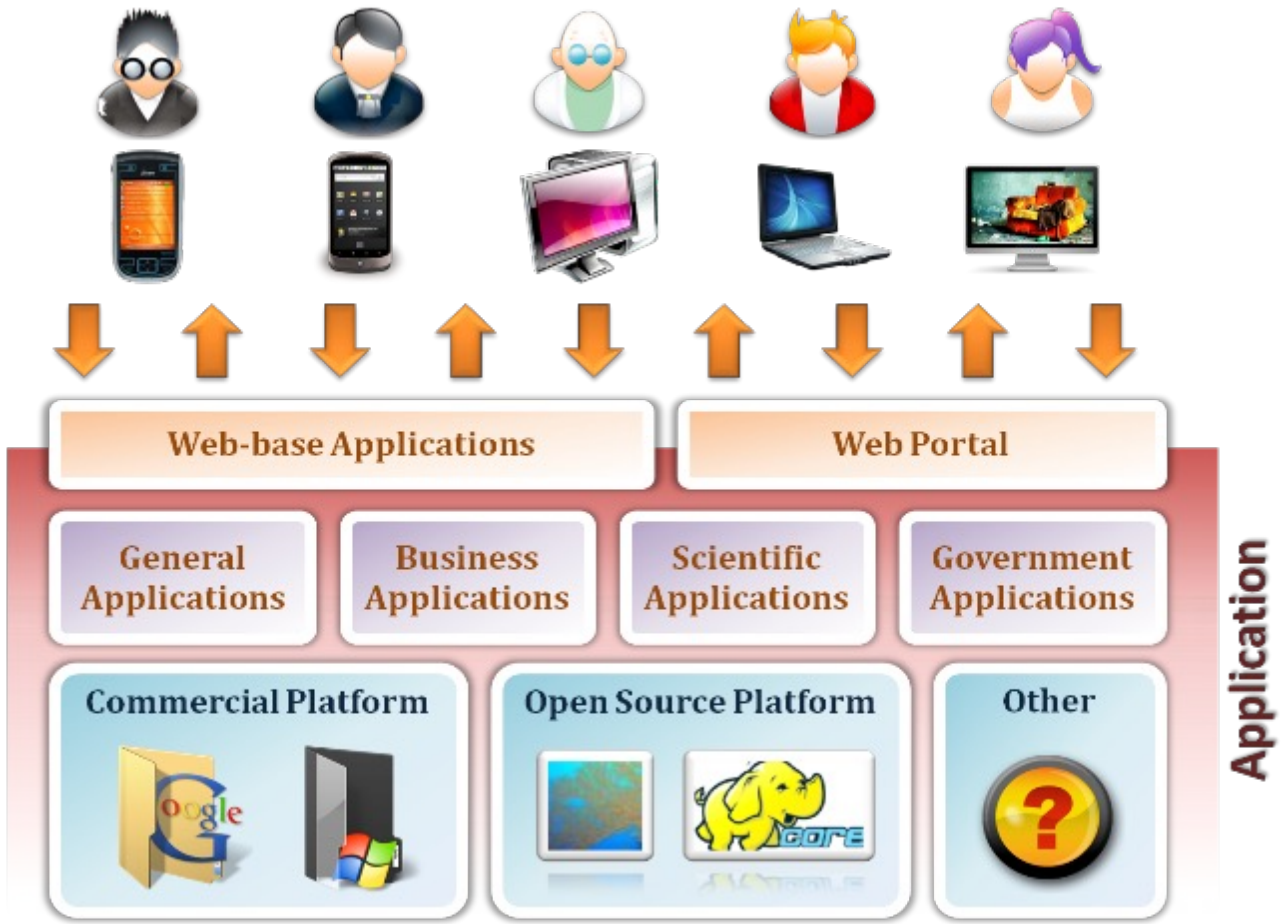
Slide credit: Che-Rung Lee

Platform as a Service



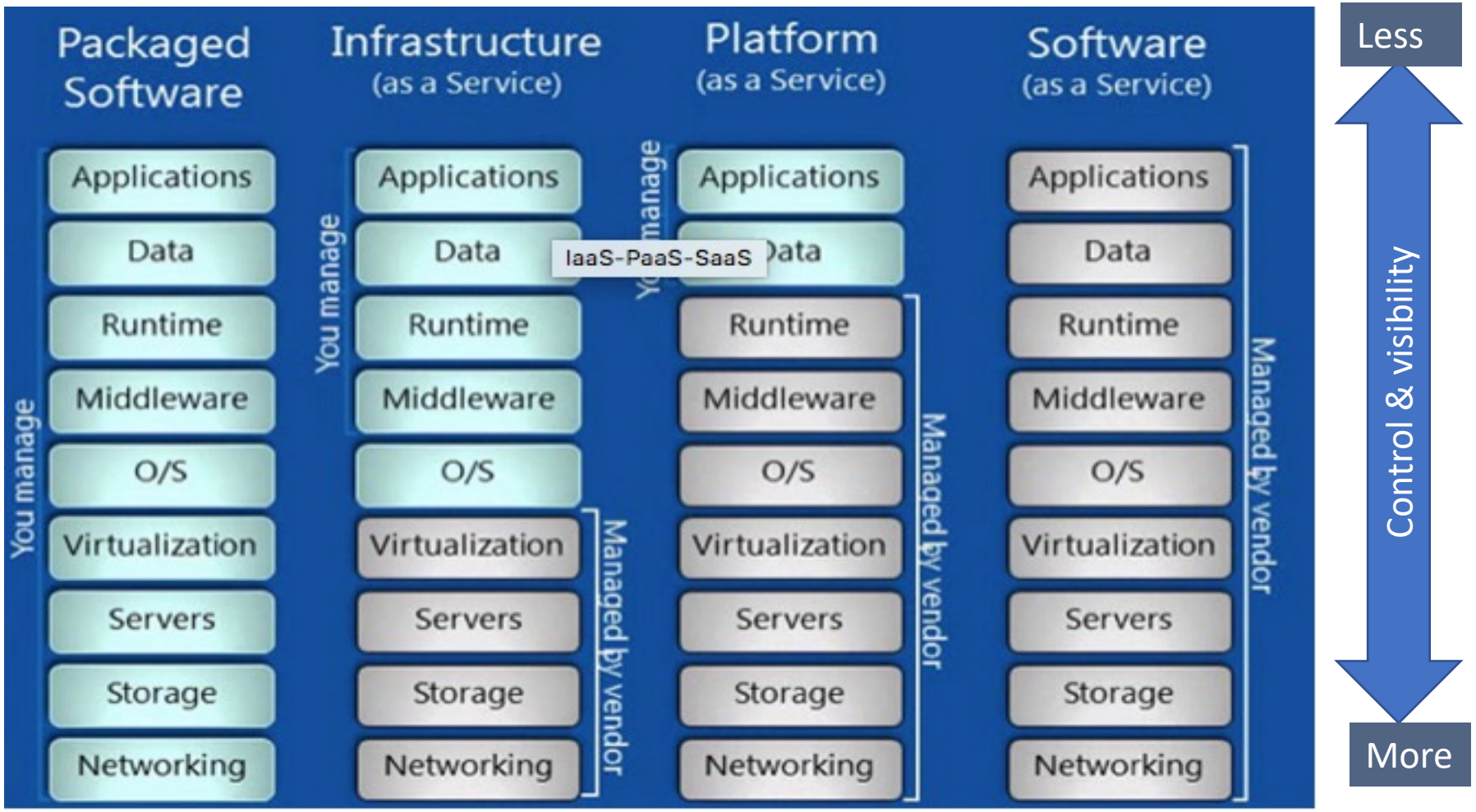
Slide credit: Che-Rung Lee

Software as a Service



Slide credit: Che-Rung Lee

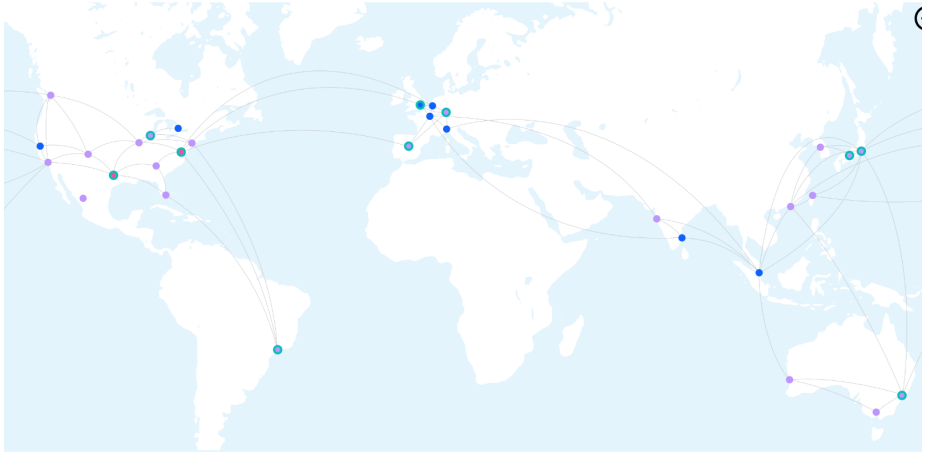
Layers of Cloud Computing



Public cloud locations (Jan. 2024)

A

B



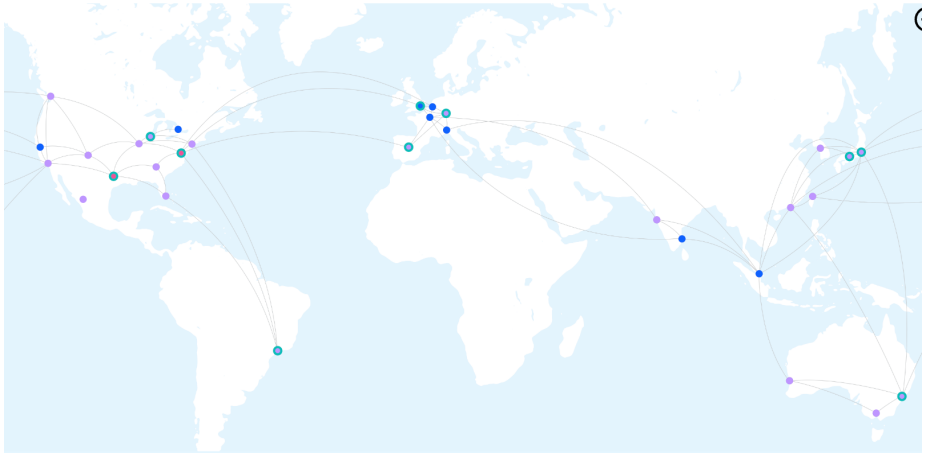
C

D

Public cloud locations (Jan. 2024)

AWS (33 regions, 105 zones)

GCE (39 regions, 118 zones)



IBM

https://www.ibm.com/cloud/data-centers

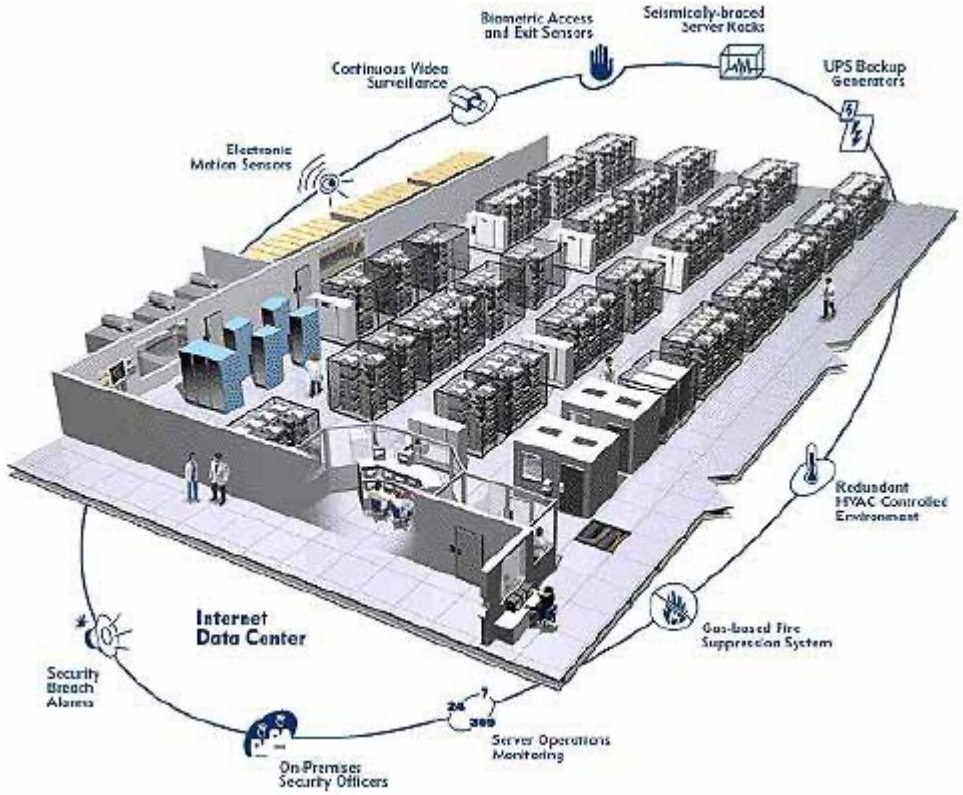
AZURE

https://azure.microsoft.com/en-us/global-infrastructure/)

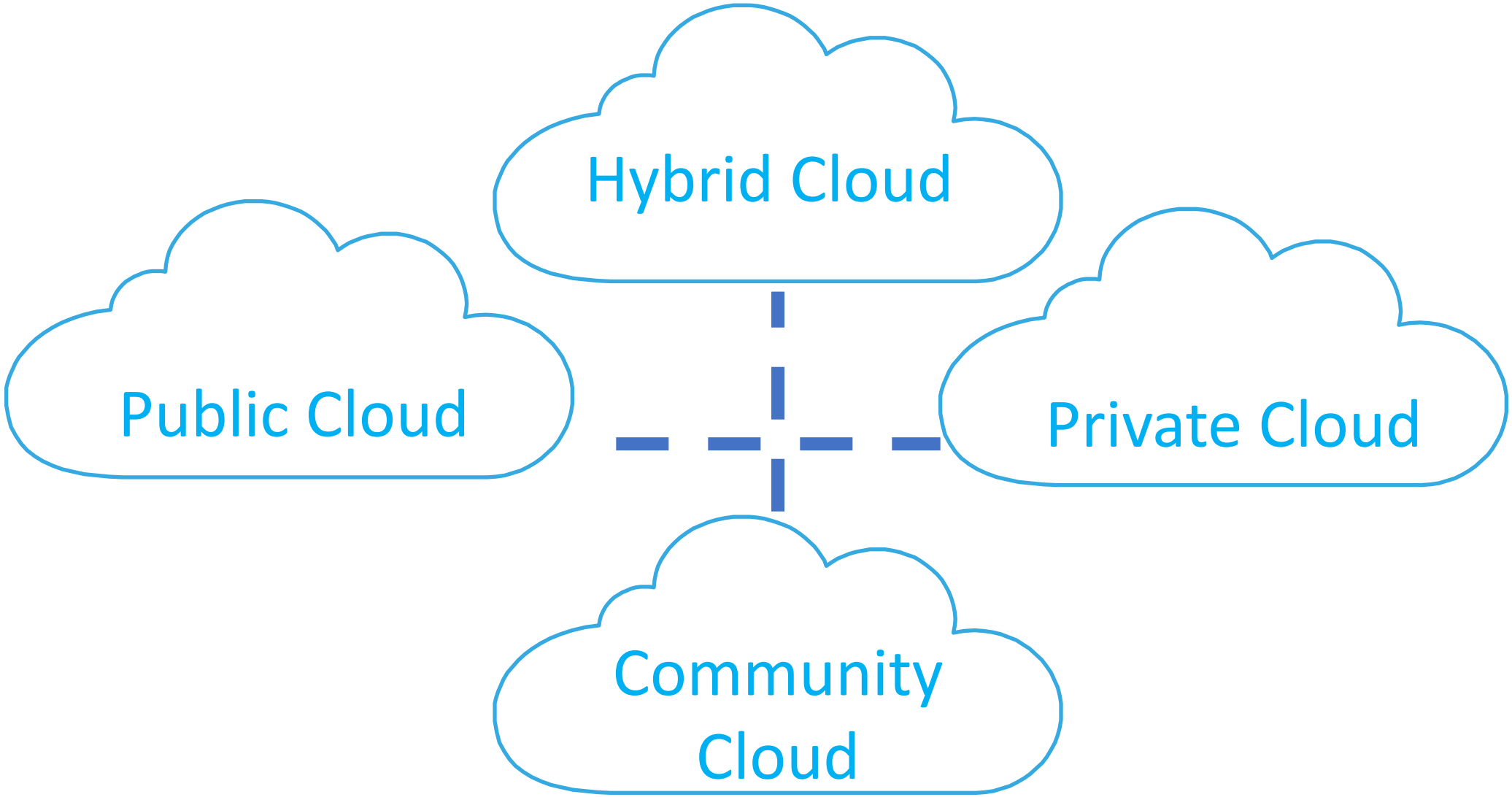
https://datacenterlocations.com/microsoft-azure/

What are data centers?

https://www.youtube.com/watch?v=XZmGGAbHqa0



Cloud Computing Deployment models



Four deployment models

Private cloud.

- The cloud infrastructure is operated solely for an organization.

Community cloud.

- The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns. A variation of private cloud

Public cloud.

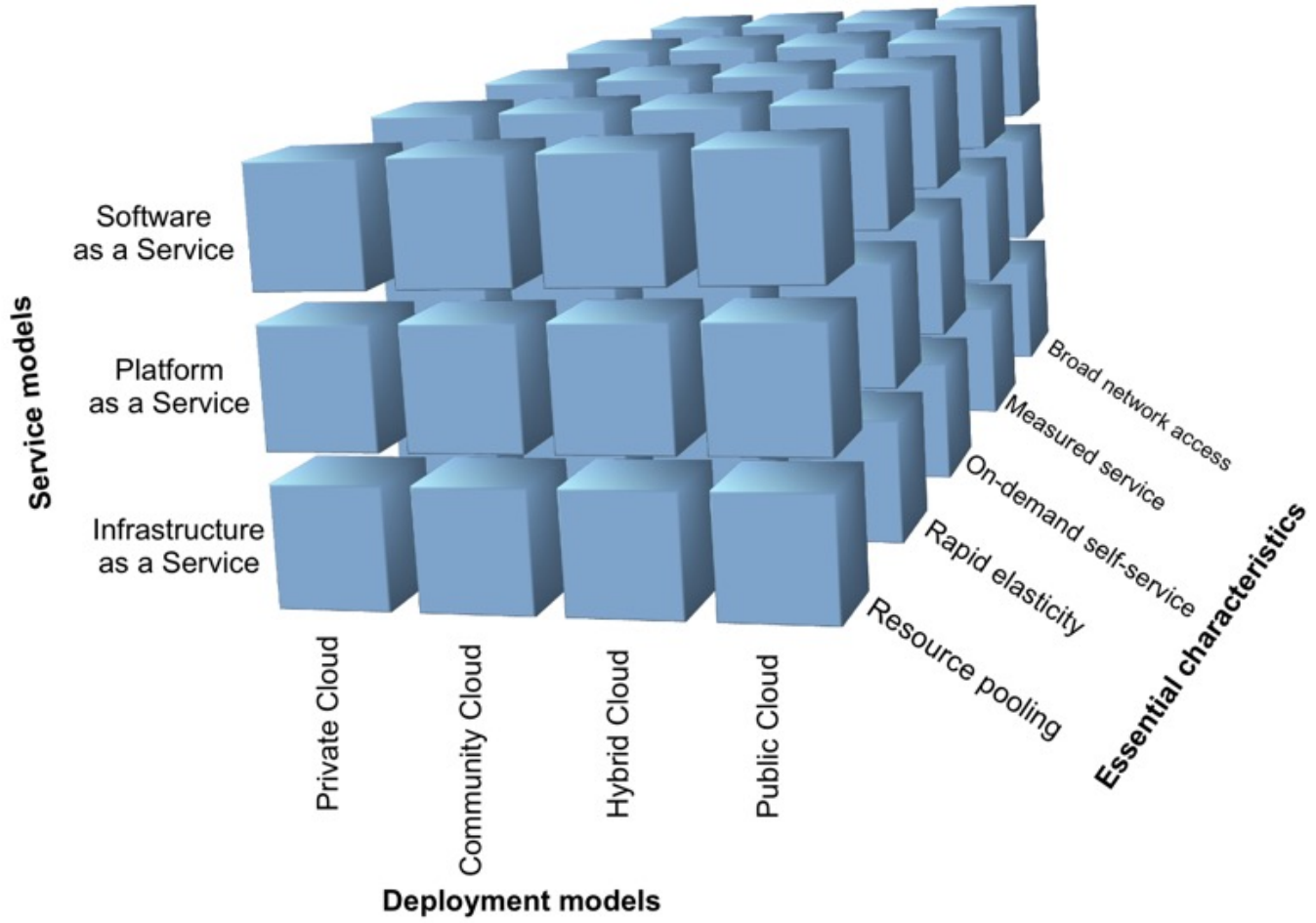
- The cloud infrastructure is made available to the general public

Hybrid cloud.

- The cloud infrastructure is a composition of two or more clouds.

Slide credit: Che-Rung Lee

Three Aspect of Cloud Computing



Life On Cloud (Cloud Computing), Amar K. 2015, retrieved 9/7/2023, https://www.linkedin.com/pulse/life-cloud-amar-kashari/

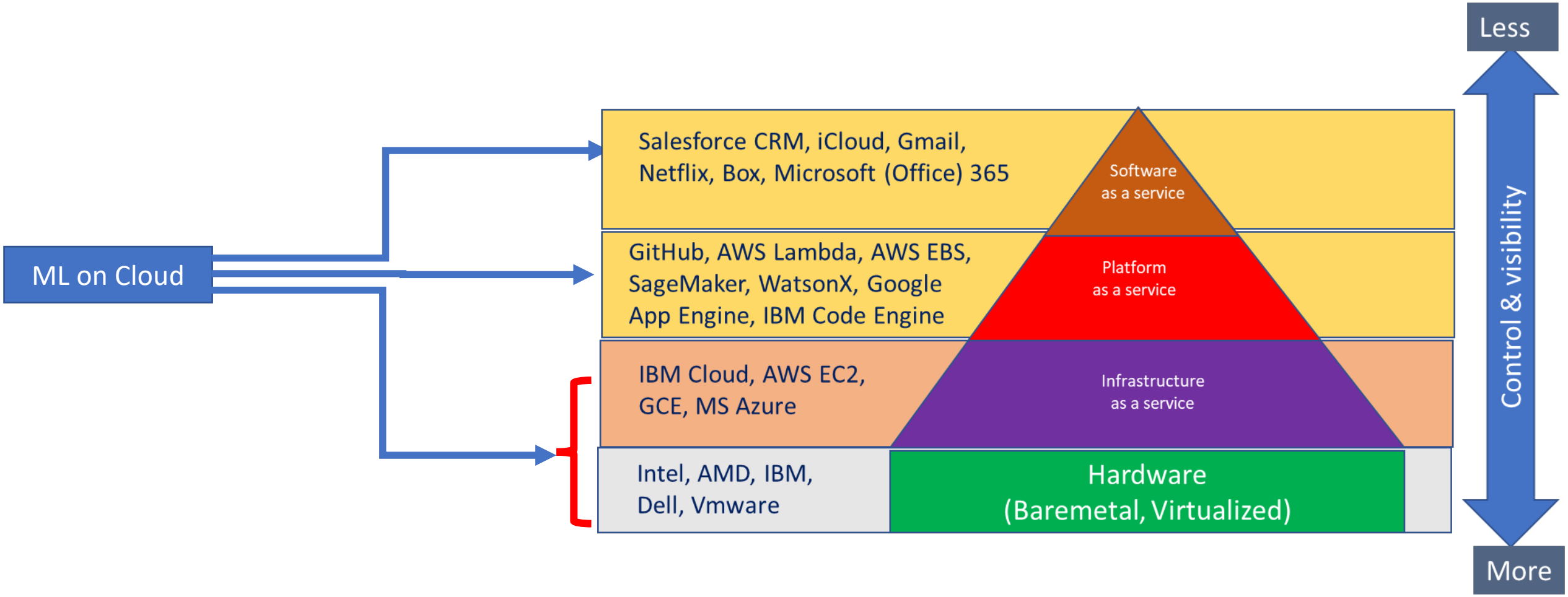
Cloud computing benefits

- Cloud enables new business models
- Time to deploy services
- Cost control and ability to scale on-demand
- Cap(ital)Ex vs Op(eration)Ex: quarterly vs hourly (ref)
- Business agility
- Simplified usage model
- Resource efficiency
- ……..

Cloud concerns

- Persistent Data
- Security
- Regulation
- Cost
- Service Level Agreements (SLA’s)
- Loss of control
- Business continuity
- ……..

Machine learning on cloud



Home setup and preview of next class

- Use your NYU NetID and password to connect to GCP
  - Optional: install the Google Cloud SDK, which provides the gcloud cli for interacting with GCP. Instructions for installing the SDK can be found: https://cloud.google.com/sdk/docs/install-sdk
  - When exercise, or doing homework, project, please release/delete resource right after your work/exercise is done. There is limited fund shared by the whole class.
    - If need discuss an operational problem, use screen capture (e.g. snipping tool in windows) to show the problem.

- Alternatively, get access to Google Cloud free tier services at here.
- AWS has a collection of free tier services, try sign up AWS Educate follow this page.

Summary

- Introductions
- Grading, logistics and syllabus
- Introduction to Cloud and ML on Cloud
- 3 homework assignments
  - Access to slack
  - Sign up for Google Cloud
  - Report on a brief survey of AI/ML/DL services from 3 public cloud providers (Graded, due Feb 16.)