CSE 4/587

Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Intro to Cloud Computing

Announcements

- HW #2 due tonight
- Phase 3 due in a week – Friday is a workshop day
- Course evaluations are up, please fill them out!
    - If 85% of the class responds, 2% extra credit (487 and 587 are independent)

What is Data Intensive Computing?

- The phrase was initially coined by National Science Foundation (NSF)
- The four V's of DIC/Big Data
  - Volume, velocity, variety, veracity (uncertainty)
- What do you expect to extract by processing this large data?
  - Intelligence for decision making
- What is different now?
  - Storage models, processing models
  - Big Data, analytics and cloud infrastructures

What is Data Intensive Computing?

- The phrase was initially coined by National Science Foundation (NSF)
- The four V's of DIC/Big Data
  - Volume, velocity, variety, veracity (uncertainty)
- What do you expect to extract by processing this large data?
  - Intelligence for decision making
- **What is different now?**
  - Storage models, processing models
  - **Big Data, analytics and cloud infrastructures**

What is Different Now?

With increasing prevalence of technology, data is everywhere...

…but now, the tools for analyzing that data are more available than ever

Cloud Computing

Cloud is a facilitator for Big Data computing and is indispensable in this context

It provides processors, software, operating systems, storage, monitoring, load balancing, clusters and other requirements as a service

*Cloud offers **accessibility** to Big Data computing*

Cloud computing models:

- Software (SaaS), Google Apps, OneDrive, Gaming platforms, etc
- Platform (PaaS), Microsoft Azure, Google App Engine (GAE)
- Infrastructure (IaaS), Amazon web services (AWS)
- Services-based application programming interface (API)

Layers of a Cloud Environment

1. **Hardware:** The servers, storage, network devices, etc
2. **Virtualization:** Abstraction layer that creates a virtual representation of physical computing and storage resources
   a. Allows multiple applications to use the same resources
3. **Application and service:** Coordinates and supports requests from the clients, and provides services depending on the particular model

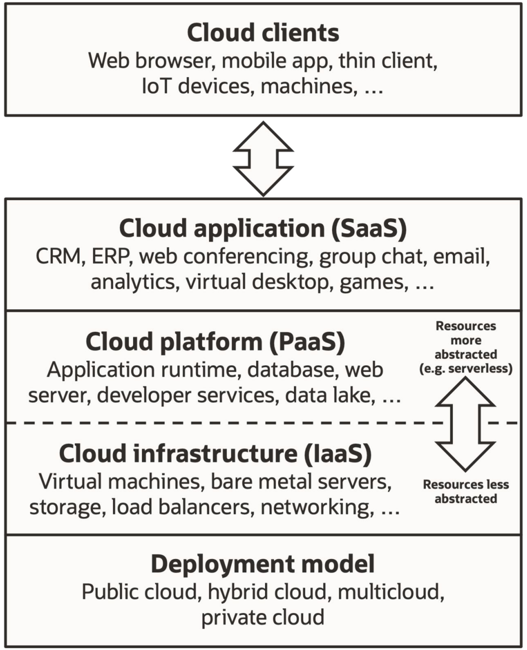
https://cloud.google.com/learn/what-is-cloud-architecture

Cloud Computing Models

NIST (National Institute of Standards and Technology) has defined three different common models of cloud computing

1. Software as a Service (SaaS)
2. Platform as a Service (PaaS)
3. Infrastructure as a Service (IaaS)

Resources get less abstracted



Cloud clients
Web browser, mobile app, thin client, IoT devices, machines, ...

Cloud application (SaaS)
CRM, ERP, web conferencing, group chat, email, analytics, virtual desktop, games, ...

Cloud platform (PaaS)
Application runtime, database, web server, developer services, data lake, ...

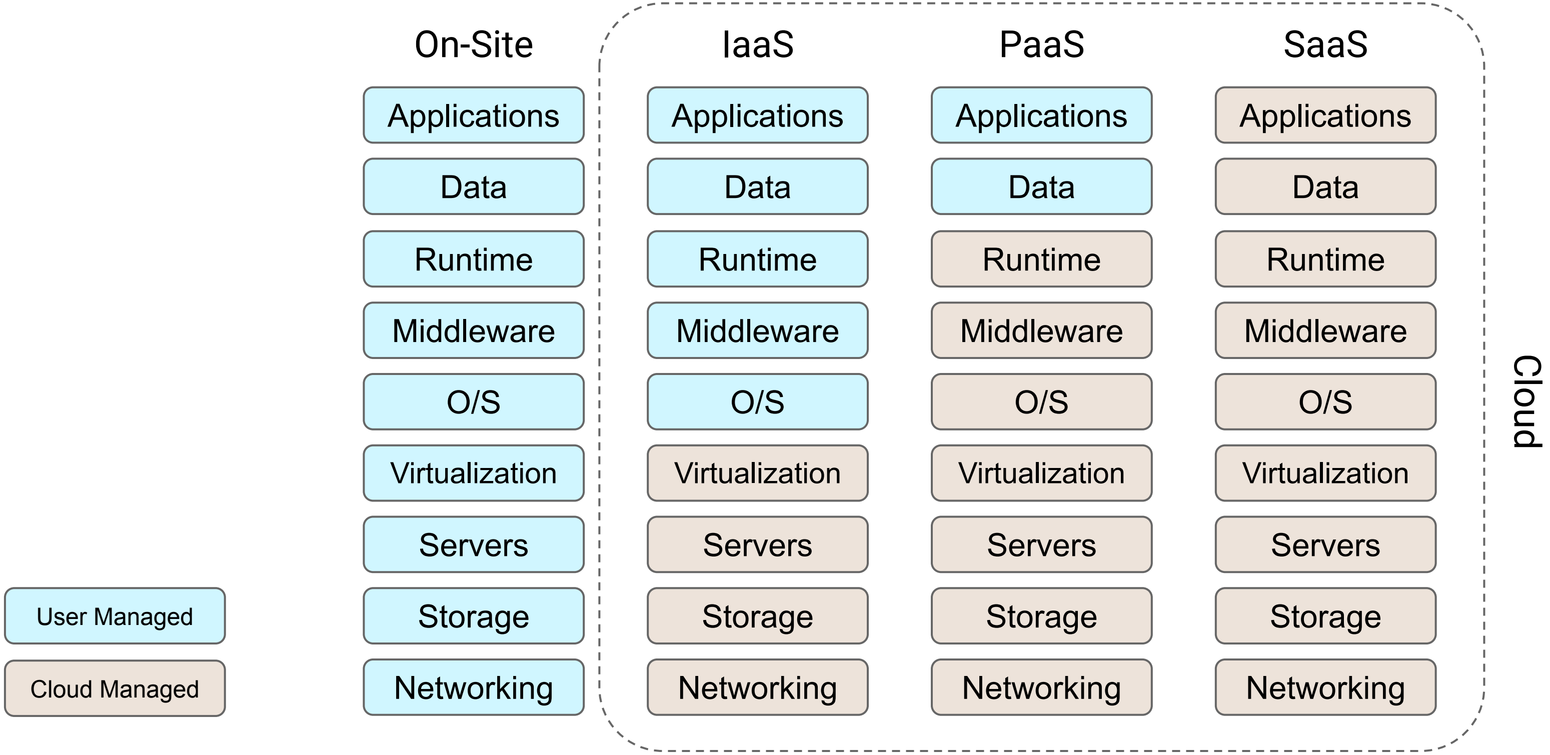
Cloud infrastructure (IaaS)
Virtual machines, bare metal servers, storage, load balancers, networking, ...

Deployment model
Public cloud, hybrid cloud, multicloud, private cloud

Resources more abstracted (e.g. serverless)
Resources less abstracted

https://en.wikipedia.org/wiki/Cloud_computing

Cloud Computing Models



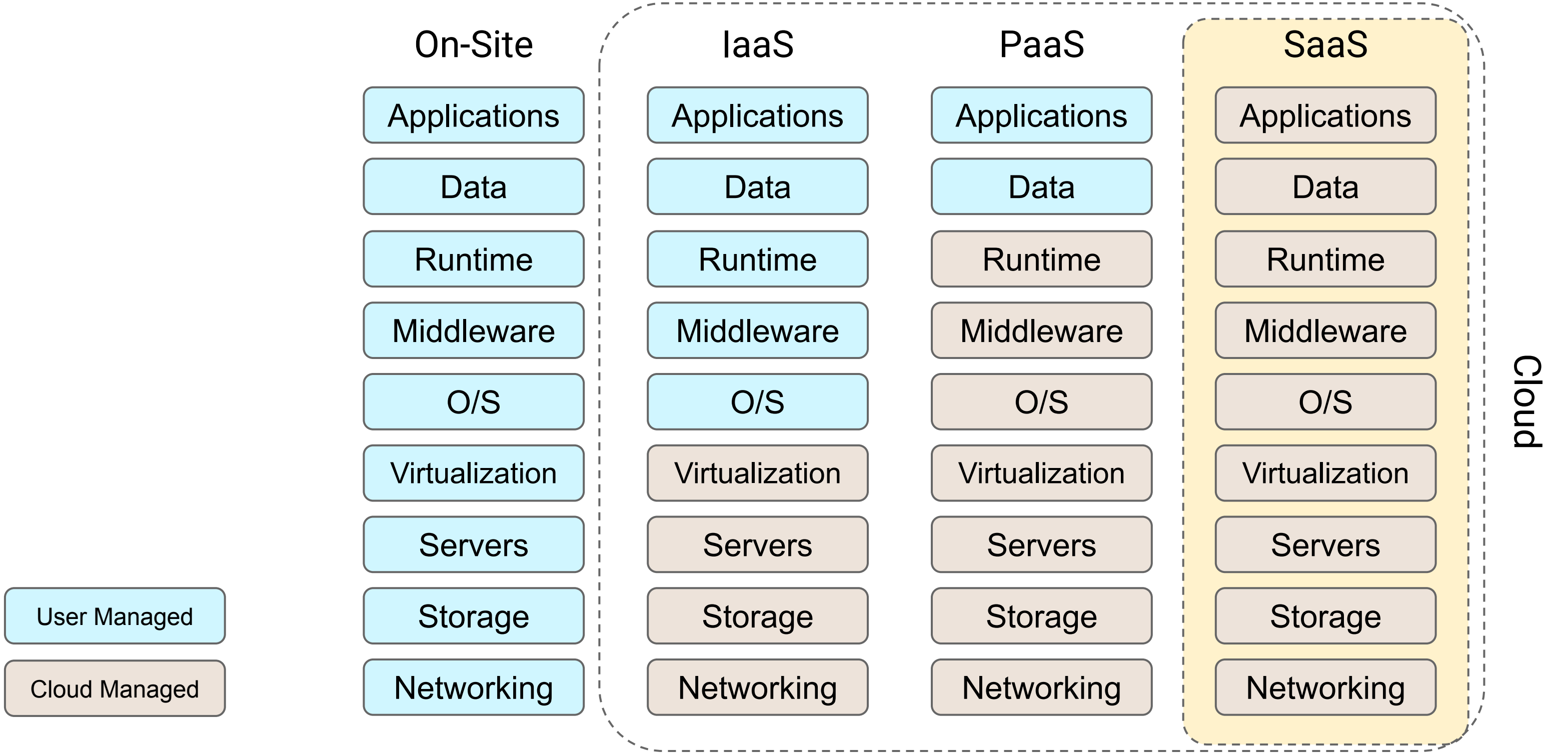
On-Site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟫 Applications
🟦 Data	🟦 Data	🟦 Data	🟫 Data
🟦 Runtime	🟦 Runtime	🟫 Runtime	🟫 Runtime
🟦 Middleware	🟦 Middleware	🟫 Middleware	🟫 Middleware
🟦 O/S	🟦 O/S	🟫 O/S	🟫 O/S
🟦 Virtualization	🟫 Virtualization	🟫 Virtualization	🟫 Virtualization
🟦 Servers	🟫 Servers	🟫 Servers	🟫 Servers
🟦 Storage	🟫 Storage	🟫 Storage	🟫 Storage
🟦 Networking	🟫 Networking	🟫 Networking	🟫 Networking

🟦 User Managed

🟫 Cloud Managed

https://www.redhat.com/en/topics/cloud-computing/what-is-paas

Software as a Service (SaaS)



On-Site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟫 Applications
🟦 Data	🟦 Data	🟦 Data	🟫 Data
🟦 Runtime	🟦 Runtime	🟫 Runtime	🟫 Runtime
🟦 Middleware	🟦 Middleware	🟫 Middleware	🟫 Middleware
🟦 O/S	🟦 O/S	🟫 O/S	🟫 O/S
🟦 Virtualization	🟫 Virtualization	🟫 Virtualization	🟫 Virtualization
🟦 Servers	🟫 Servers	🟫 Servers	🟫 Servers
🟦 Storage	🟫 Storage	🟫 Storage	🟫 Storage
🟦 Networking	🟫 Networking	🟫 Networking	🟫 Networking

🟦 User Managed

🟫 Cloud Managed

https://www.redhat.com/en/topics/cloud-computing/what-is-paas

Software as a Service (SaaS)

"The capability provided to the consumer is to ***use the provider's applications*** running on a cloud infrastructure.", NIST

- Highest level of resource abstraction
  - User does not need to (nor can they) manage underlying infrastructure
  - Network, Operating System, File System, etc
- The software is managed and installed by the cloud service provider
- Usually have some sort of fee associated with use
- **Examples:** Google Drive (Docs, Photos, etc), OneDrive, Amazon AWS

Cloud Application Requirements

What does the software running in the cloud do differently from non-cloud applications?

Challenges:

- Scalability – how easily can we add/remove tasks from the job?
- Elasticity – what if more (or less) resources become available?
- Load Balancing – how do we balance the tasks across cloud nodes?
- Multi-Tenancy – how do we handle multiple tasks on the same node?

Cloud Application Requirements

What does the software running in the cloud do differently from non-cloud applications?

Challenges:

- Scalability – how easily can we add/remove tasks from the job?
- Elasticity – what if more (or less) resources become available?
- Load Balancing – how do we balance the tasks across cloud nodes?
- Multi-Tenancy – how do we handle multiple tasks on the same node?

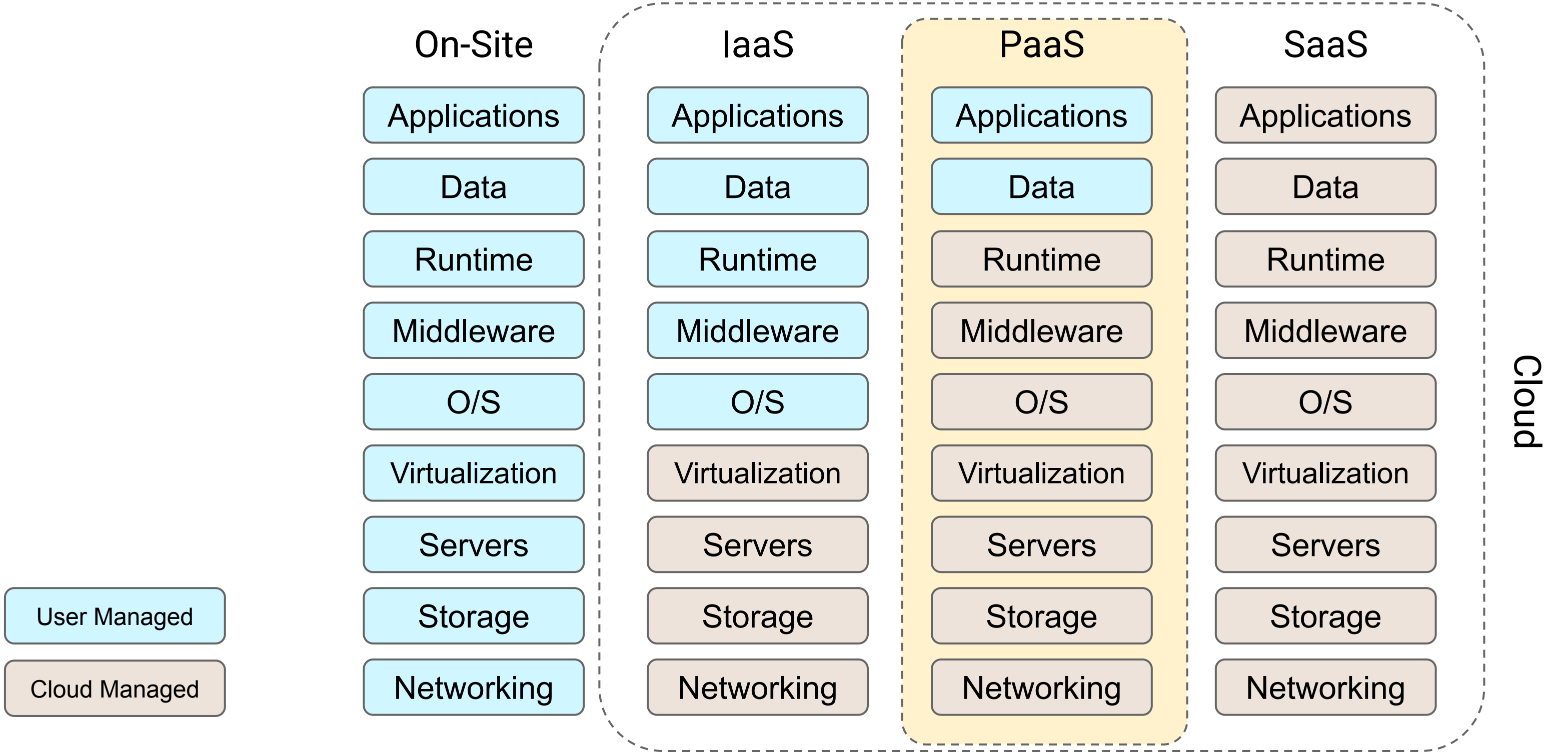
Specifically we want to address these challenges transparently

Cloud Application Requirements

- Many of these challenges have been addressed by techniques we've already explored this semester **(MapReduce and Spark)**
    - MapReduce and Spark both scale up linearly and transparently
    - If more resources become available, new tasks can be spawned
    - Tasks can be moved to balance computation
    - Multiple tasks (in VMs) can be run on a single node
- **Sidenote:** in other areas like High-Performance Computing (HPC) there's a lot of research being done to take advantage of Cloud Computing while addressing these challenges (ie see Charm++)

http://charm.cs.illinois.edu/research/cloud

Platform as a Service (SaaS)



On-Site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟫 Applications
🟦 Data	🟦 Data	🟦 Data	🟫 Data
🟦 Runtime	🟦 Runtime	🟫 Runtime	🟫 Runtime
🟦 Middleware	🟦 Middleware	🟫 Middleware	🟫 Middleware
🟦 O/S	🟦 O/S	🟫 O/S	🟫 O/S
🟦 Virtualization	🟫 Virtualization	🟫 Virtualization	🟫 Virtualization
🟦 Servers	🟫 Servers	🟫 Servers	🟫 Servers
🟦 Storage	🟫 Storage	🟫 Storage	🟫 Storage
🟦 Networking	🟫 Networking	🟫 Networking	🟫 Networking

🟦 User Managed

🟫 Cloud Managed

https://www.redhat.com/en/topics/cloud-computing/what-is-paas

Platform as a Service (PaaS)

"The capability provided to the consumer is to ***deploy onto the cloud infrastructure*** consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider." - NIST

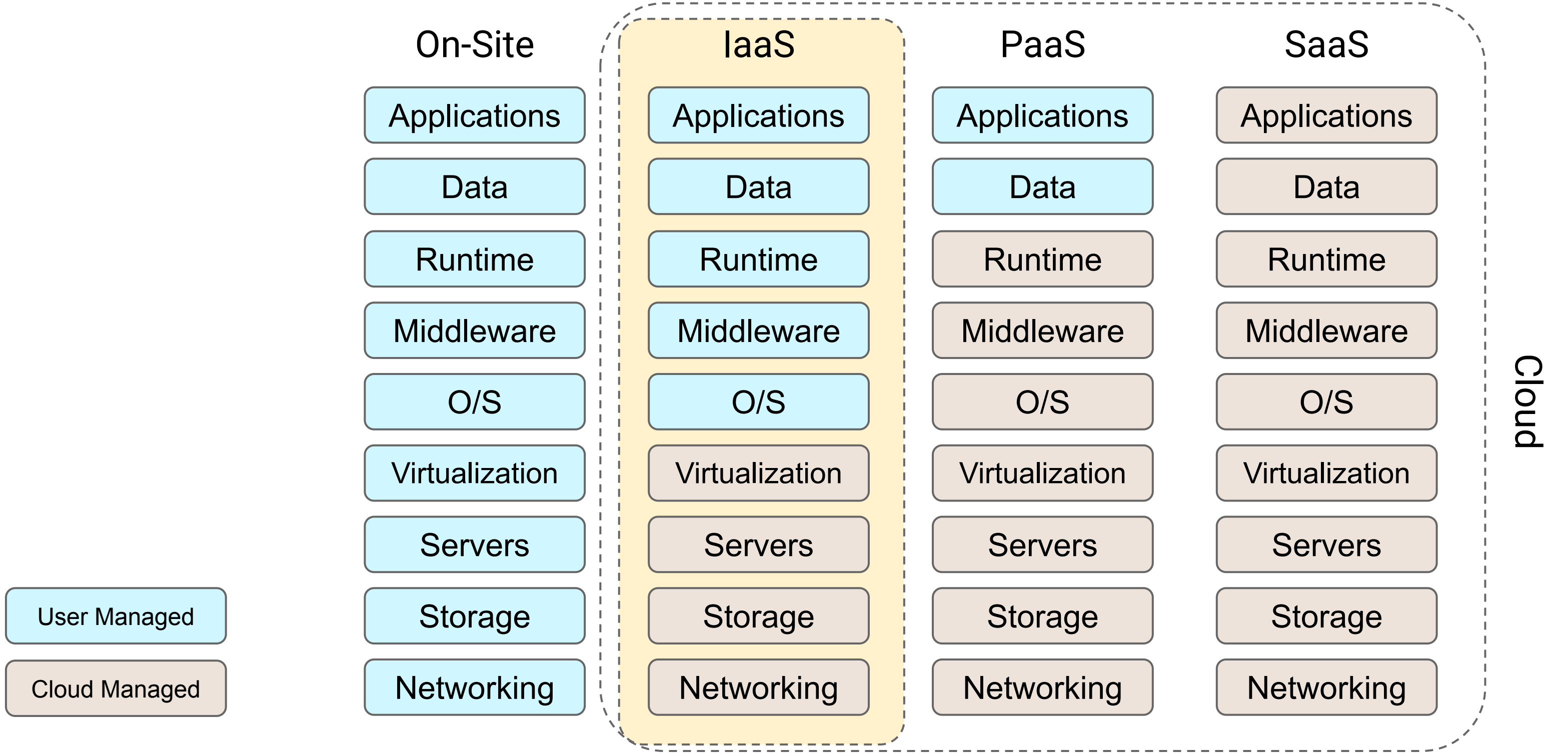
- Middle level of resource abstraction
  - Users can develop and deploy their own applications to the cloud
  - They still do not manage the underlying infrastructure
- Cloud providers provide the OS, hardware, execution environment, etc
- Users provide the software…allows them to deploy server-side software without buying, maintaining, and managing the hardware
- **Examples:** Google App Engine, Microsoft Azure, Heroku, etc

Example: Dataproc

Dataproc is Google's cloud service for deploying Apache Spark and Apache Hadoop applications to a cloud environment

- Integration with both Spark and Hadoop – take your applications as written for small clusters or single node, and scale to the cloud
- Automatic scaling/resizing – elastic resource management can scale your application automatically as resources become available
- Utilize existing Spark/Hadoop libraries for ML, SQL, Streaming, etc

Infrastructure as a Service (SaaS)



On-Site	IaaS	PaaS	SaaS
🟦 Applications	🟦 Applications	🟦 Applications	🟫 Applications
🟦 Data	🟦 Data	🟦 Data	🟫 Data
🟦 Runtime	🟦 Runtime	🟫 Runtime	🟫 Runtime
🟦 Middleware	🟦 Middleware	🟫 Middleware	🟫 Middleware
🟦 O/S	🟦 O/S	🟫 O/S	🟫 O/S
🟦 Virtualization	🟫 Virtualization	🟫 Virtualization	🟫 Virtualization
🟦 Servers	🟫 Servers	🟫 Servers	🟫 Servers
🟦 Storage	🟫 Storage	🟫 Storage	🟫 Storage
🟦 Networking	🟫 Networking	🟫 Networking	🟫 Networking

🟦 User Managed

🟫 Cloud Managed

https://www.redhat.com/en/topics/cloud-computing/what-is-paas

Infrastructure as a Service (IaaS)

"the consumer is able to deploy and run arbitrary software, which can include operating systems and applications" -NIST

- Lowest level of resource abstraction
  - Users can deploy and run arbitrary software, including OS
  - May also have some limited control over network components
  - (Still doesn't control the actual hardware/network/etc)
- Cloud providers provide the hardware
- **Examples:** Amazon AWS, Google Compute Engine, Azure, IBM Cloud