AI Model Deployment

Hyperscaler Vendor Offerings

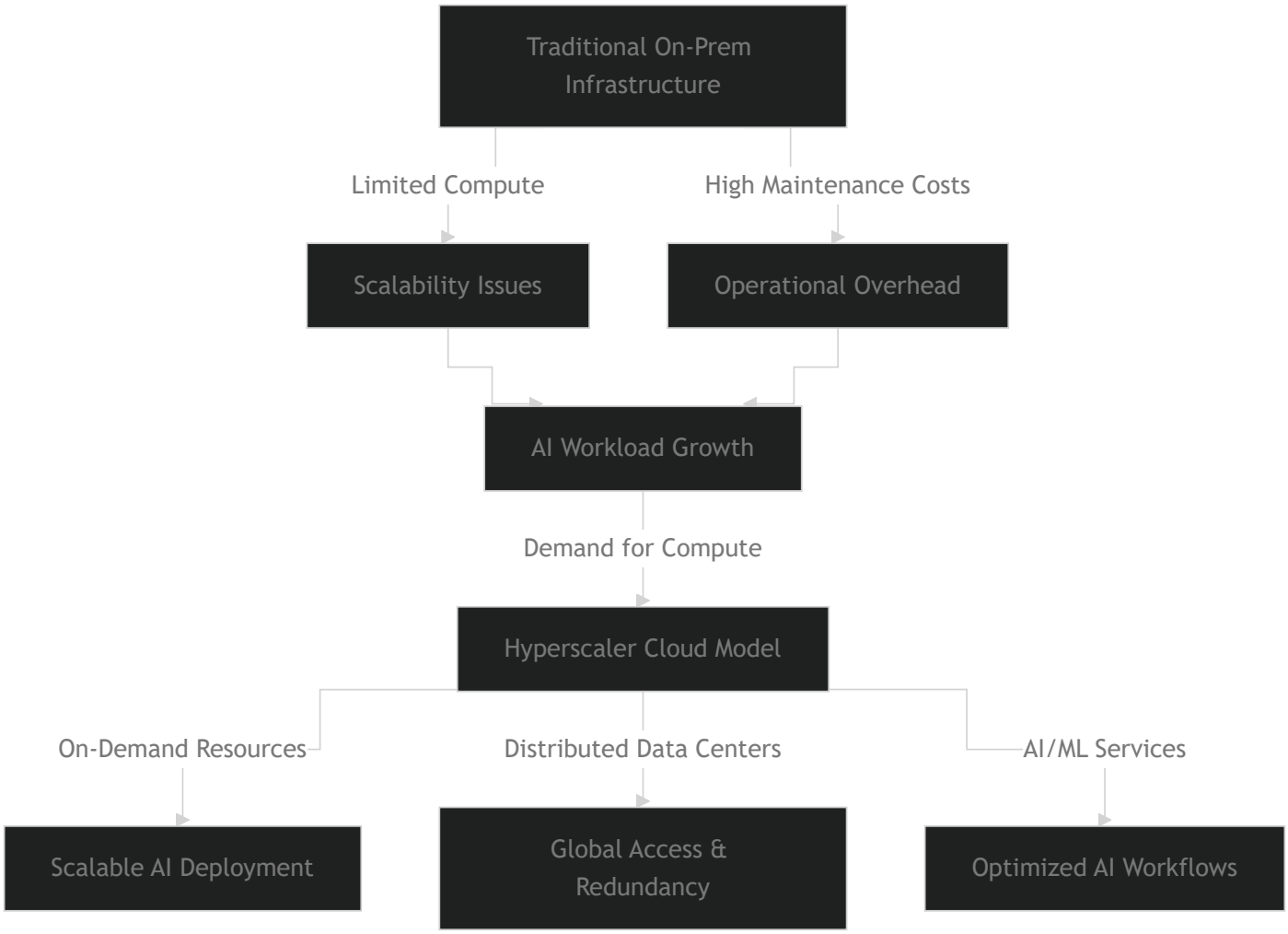
What is a Hyperscaler?

Hyperscalers are large cloud providers that offer scalable, on-demand computing resources across global data centers. These companies—**Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP)**—enable enterprises to deploy AI and machine learning (ML) models at scale, leveraging their vast infrastructure and AI-specific services.

Key Characteristics of Hyperscalers:

- **Global Scalability:** Data centers worldwide ensure low-latency access and redundancy.
- **On-Demand Resources:** Flexible computing power and storage, billed based on usage.
- **AI/ML Services:** Pre-built and customizable tools for deploying machine learning models.
- **Security & Compliance:** Industry-standard security protocols and regulatory compliance (e.g., GDPR, HIPAA).

Where does the need for hyperscalers come from?



Comparing AI/ML Offerings

The three major hyperscalers provide distinct AI/ML deployment options. Below is an overview of their capabilities:

Cloud Provider	AI/ML Deployment Services	Key Strengths	Key Limitations
AWS	SageMaker, Lambda for inference, EKS for containerized models	Broadest AI/ML service ecosystem, strong security, global reach	Pricing complexity, potentially expensive for high-compute workloads
Azure	Azure Machine Learning, AKS for model hosting, Functions for serverless inference	Strong enterprise integration (Active Directory, DevOps), hybrid cloud capabilities	Less mature AI tooling compared to AWS

Cloud Provider	AI/ML Deployment Services	Key Strengths	Key Limitations
GCP	Vertex AI, Cloud Run, TensorFlow Serving, TPU accelerators	Best for AI-first workloads, strong model training and scaling capabilities	Fewer enterprise integrations compared to AWS and Azure

Choosing the Right Cloud Provider

Each hyperscaler has strengths and trade-offs. The best choice depends on the **business requirements, budget, and performance needs** of the AI solution.

- **Cost vs. Performance:** Does the project prioritize cost savings, or is high performance essential?
- **Integration Needs:** Does the company already use a particular cloud provider for other services?
- **Scalability:** Will the model need to handle fluctuating or high-demand workloads?
- **Security & Compliance:** Does the organization operate in a highly regulated industry (e.g., finance, healthcare)?

Hyperscaler Decision-Making Activity

In this scenario based discussion activity you will:

- Apply your knowledge to a real-world consulting scenario
- Use the Hyperscaler Decision-Making Activity Worksheet to evaluate a case study and select the best cloud provider
- Discuss trade-offs and justify your recommendation with peers.

Be prepared to discuss your descisions with the class when we return from breakout rooms!