Module 11

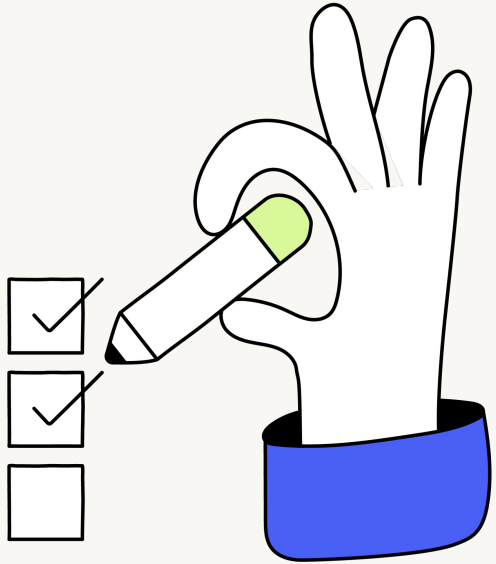
Workshop 2: Optimising ML Inference Scaling for Performance and Cost

Icebreaker

One giant or many small?

If you had to serve one million predictions per second, would you rather use one super-powerful computer, or many smaller computers?

Share your answers in the chat.



Module 11

Workshop 2: Optimising ML Inference Scaling for Performance and Cost

Agenda

- **Review:** Recap key concepts from async unit 3.
- **Practice:** Simulating inference scaling trade-offs.
- **Closing:** Key takeaways and next steps.

Learning objectives

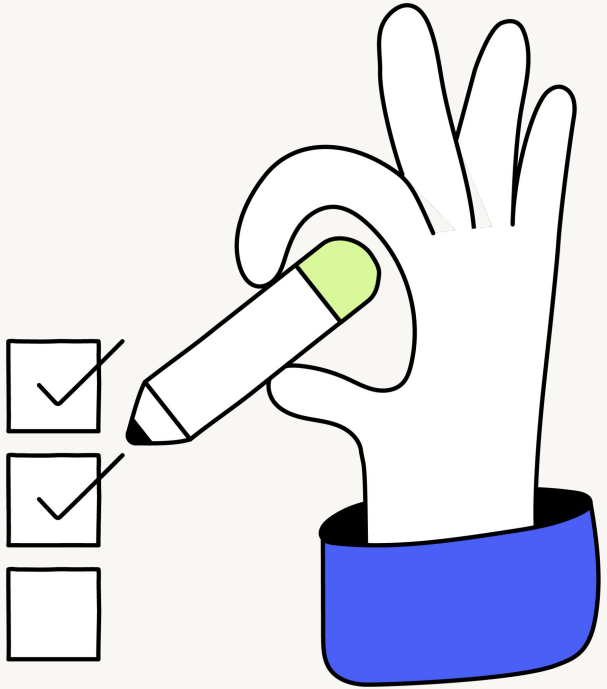
- **Understand the fundamental trade-offs** between latency, throughput, and resource allocation for ML inference workloads.
- **Simulate the impact** of different scaling strategies (vertical vs. horizontal) on the performance and theoretical cost of an ML inference service.
- **Critically assess** how various computational resources relate to operational requirements in ML/AI systems.
- **Reinforce concepts** of performance optimisation and cost-efficiency in designing scalable ML platform architectures.

Async review: Units 3

Go to My Multiverse, Module 11
Workshop 2

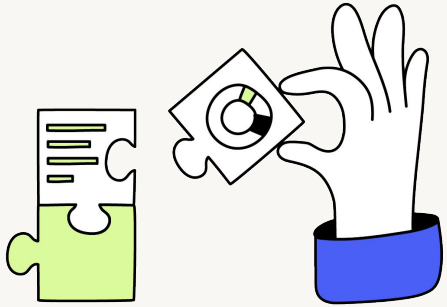
Navigate to the **“Async review”** page

Review the summary of Unit 3



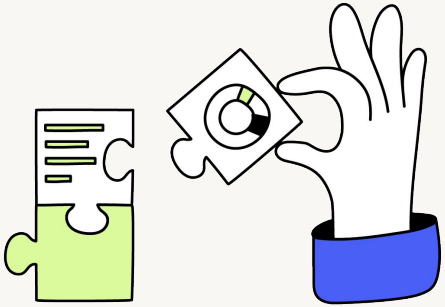
Unit 3: ML/AI Platform architecture and resource allocation

- **ML/AI platform architecture basics:** The essential components that support model deployment at scale—compute, storage, model serving layers, and data pipelines.
- **Operational performance metrics:** How latency (time per prediction) and throughput (predictions per second) guide deployment decisions, especially for real-time applications.
- **Scaling strategies:** The difference between vertical scaling (using more powerful machines) and horizontal scaling (distributing load across many machines), and how each impacts performance and cost.



Unit 3: ML/AI Platform architecture and resource allocation

- **Resource selection:** The trade-offs between CPUs, GPUs, and accelerators, and how different workloads (e.g., batch vs. real-time inference) require different hardware strategies.
- **Cost-performance balancing:** How to align architectural choices with business goals by optimising for speed, scalability, and resource efficiency—ensuring models are not just accurate, but also affordable to run at scale.



The performance–cost balancing act

- ML platforms must deliver fast predictions (low latency) and many predictions (high throughput).
- But higher performance often means higher cost—through more powerful machines or more infrastructure.
- Smart deployment means balancing:
  - Latency – How fast is each prediction?
  - Throughput – How many predictions can run at once?
  - Cost – What resources are required to support both?

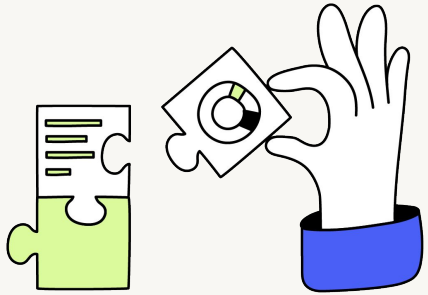
Scaling strategies: Vertical vs. horizontal

To balance these performance and cost trade-offs, teams typically scale their systems in one of two ways:

1. Vertical scaling (scale up)

Upgrade to more powerful hardware (e.g., faster CPUs, more RAM, better GPUs).

- Can improve latency for single predictions.
- Has limits and becomes expensive quickly.
- **Best for:** Low request volumes, latency-critical tasks.

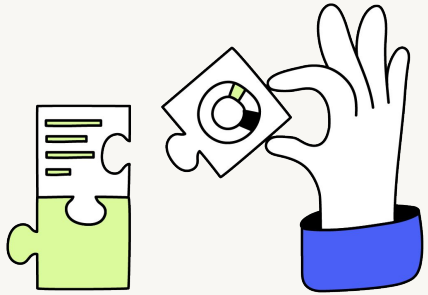


Scaling strategies: Vertical vs. horizontal

2. Horizontal scaling (scale out)

Add more machines or instances to handle traffic in parallel.

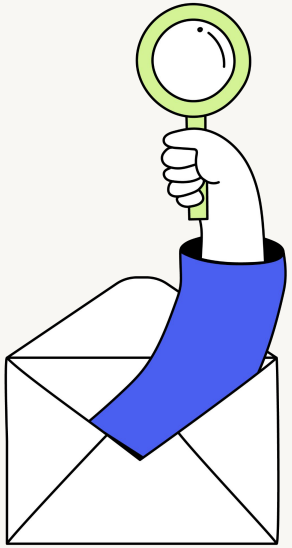
- Increases throughput by distributing load.
- Often more cost-efficient at scale.
- Requires load balancing and orchestration.
- **Best for:** High request volumes, scalable cloud systems.



Skills application: Simulating inference scaling trade-offs

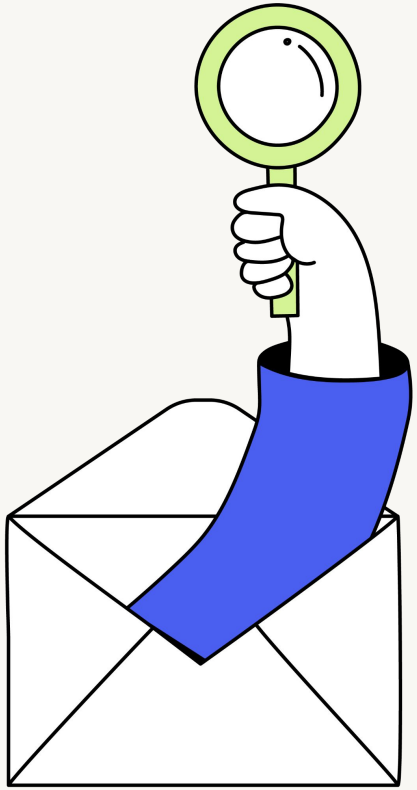
Skills application introduction

- **Objective:** In this skills application activity, you'll simulate how different infrastructure choices impact the performance and cost of an ML inference system. You'll test how latency and throughput respond to vertical and horizontal scaling strategies under varying request loads.
- **You'll learn how to:**
  - Make better decision-making when selecting compute resources helping manage cloud costs.
  - Ensure that ML services remain responsive at scale.



Skills application

- Work in breakout rooms in small groups.
- Use the Jupyter notebook and provided datasets.
- The challenge is designed to be solved individually, but collaboration is encouraged.
- Find all instructions and resources on Ariel.

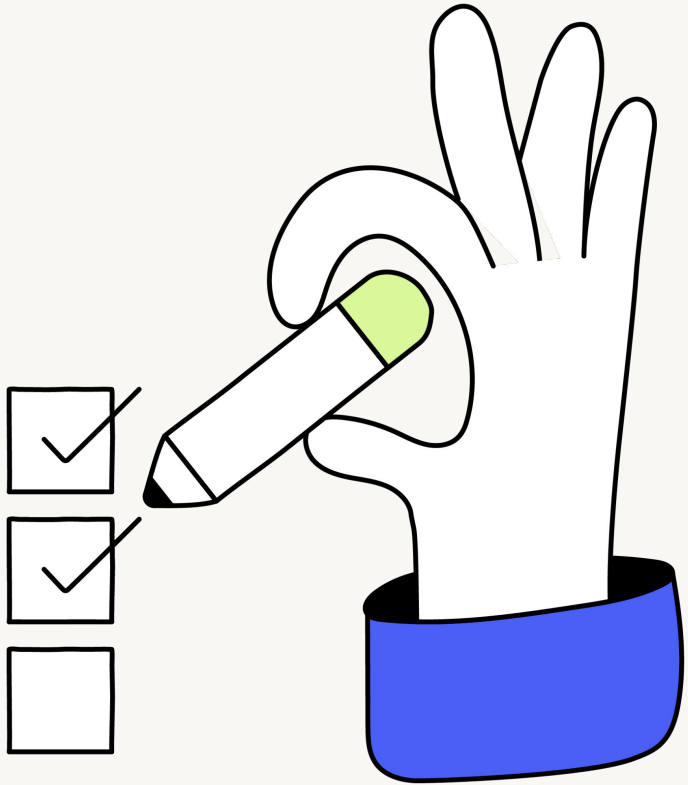


Access the skills application

Go to My Multiverse, Module 11
Workshop 2

Go to the Skills application page and follow the instructions provided

Be prepared to discuss your learnings when we come back together



Skills application debrief

Skills application debrief

Based on our simulations, what are the primary advantages and disadvantages of vertical scaling versus horizontal scaling for an ML inference service?

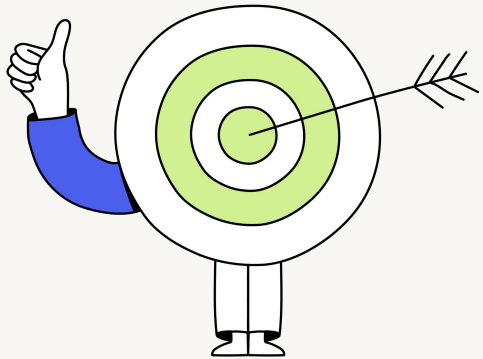
How does the "cost" aspect influence your architectural decisions for ML platforms, especially when considering performance targets?

Can you think of real-world cloud services or technologies that embody these vertical and horizontal scaling concepts?

Key takeaways

1. **Scaling isn't just technical—it's strategic:**

   Choosing between vertical and horizontal scaling isn’t just a question of speed. It’s about balancing latency, throughput, and cost to meet business goals efficiently. Smart scaling = smart architecture.



2. **The right infrastructure unlocks ML value:**

   By simulating real-world inference loads, you’ll learn how platform design impacts performance and budget—helping you make better decisions when deploying at scale.

Applying key takeaways to your organisation

Applying key takeaways

1. How could evaluating scaling strategies in advance help your team balance performance targets and infrastructure costs more effectively?

2. What parts of your ML infrastructure would benefit from a clearer understanding of latency, throughput, and resource allocation trade-offs?

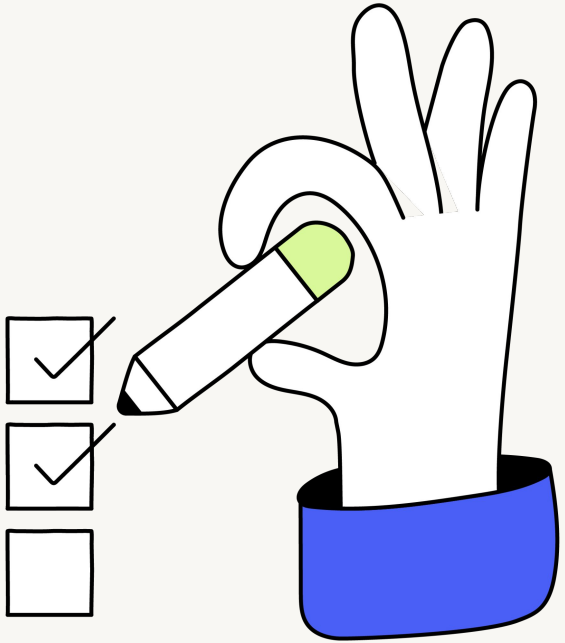
Share your plan to apply new skills

Go to My Multiverse, Module 11
Workshop 2

On the Apply new skills to your role page, find the discussion activity

Take 2 minutes to share how you will apply this skill in your role.

Take time this week to read what others have written and get new ideas!



Closing out

- **Next workshop:** Module 11 Group Coaching
- **Have questions?**
  - Chat with Atlas
  - Reach out to your Cohort Coach
  - Sign up for Tutoring
- **Log your OTJ hours**
- **Take the Workshop Survey** on My Multiverse

multiverse