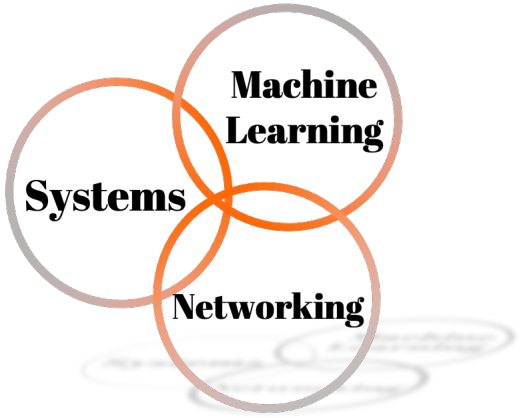
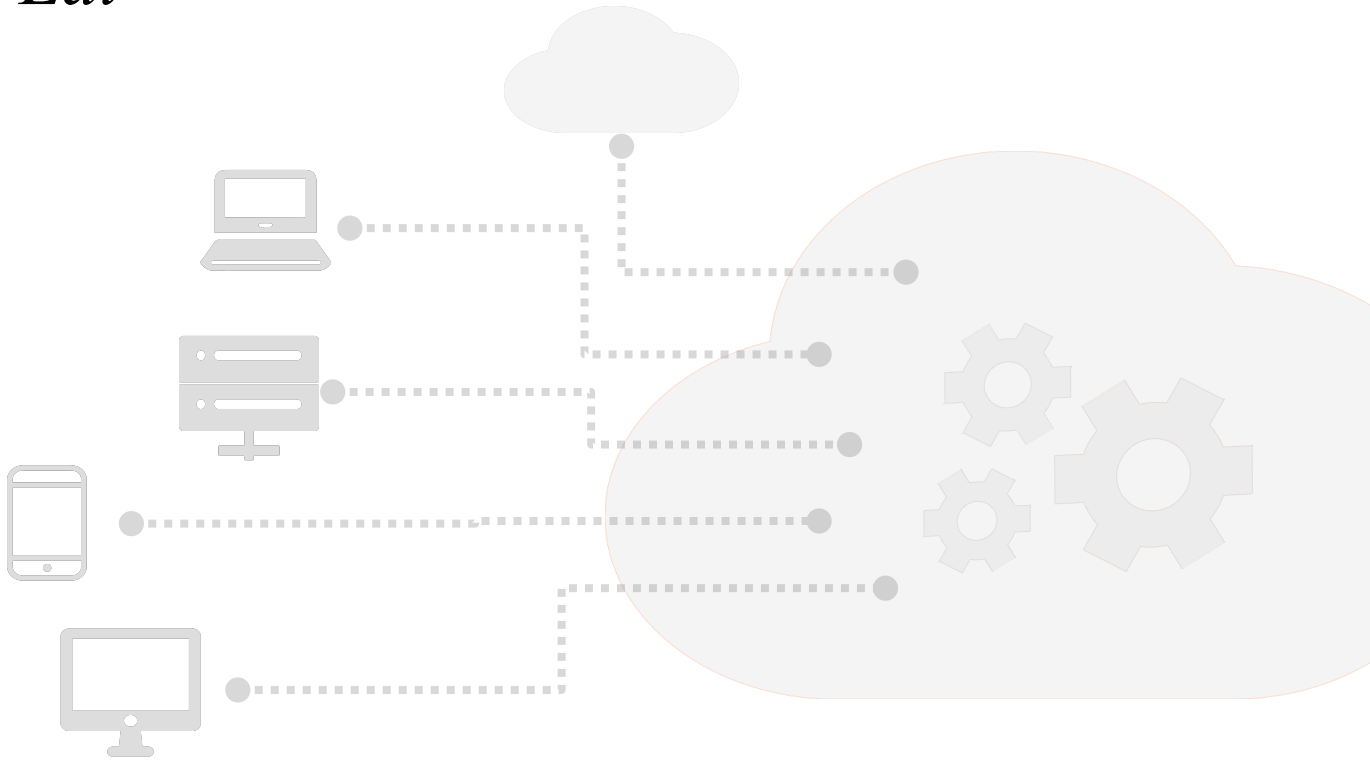


CS 598: Systems for GenAI

Instructor: Fan Lai



ILLINOIS

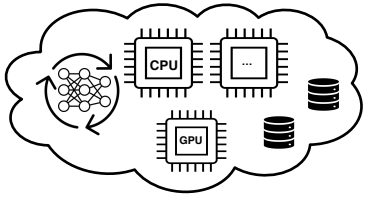
About Fan

- **Assistant Professor in CS**
  - PhD'23: "Minimalist Systems for Pervasive Machine Learning''
- **Research interest**
  - Cloud ML (LLMs, Image/Video Gen)
  - Collaborative ML (on-device AI), ML4Sys

About Fan

- **Assistant Professor in CS**
  - PhD'23: "Minimalist Systems for Pervasive Machine Learning"

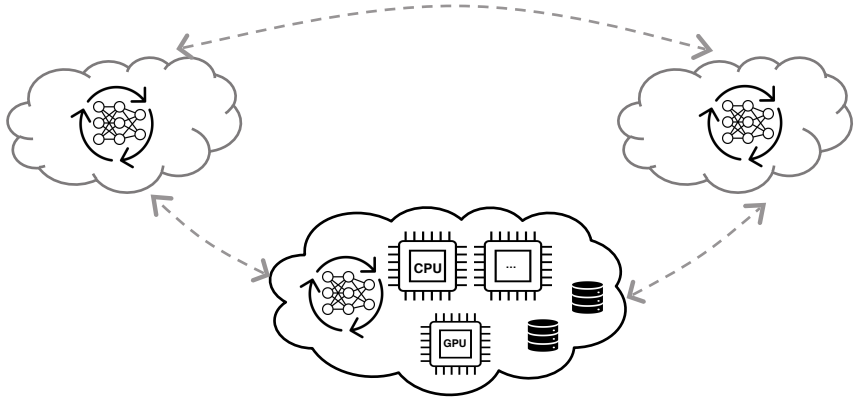
- **Research interest**
  - Cloud ML (LLMs, Image/Video Gen)
  - Collaborative ML (on-device AI), ML4Sys



About Fan

- **Assistant Professor in CS**
  - PhD'23: "Minimalist Systems for Pervasive Machine Learning"

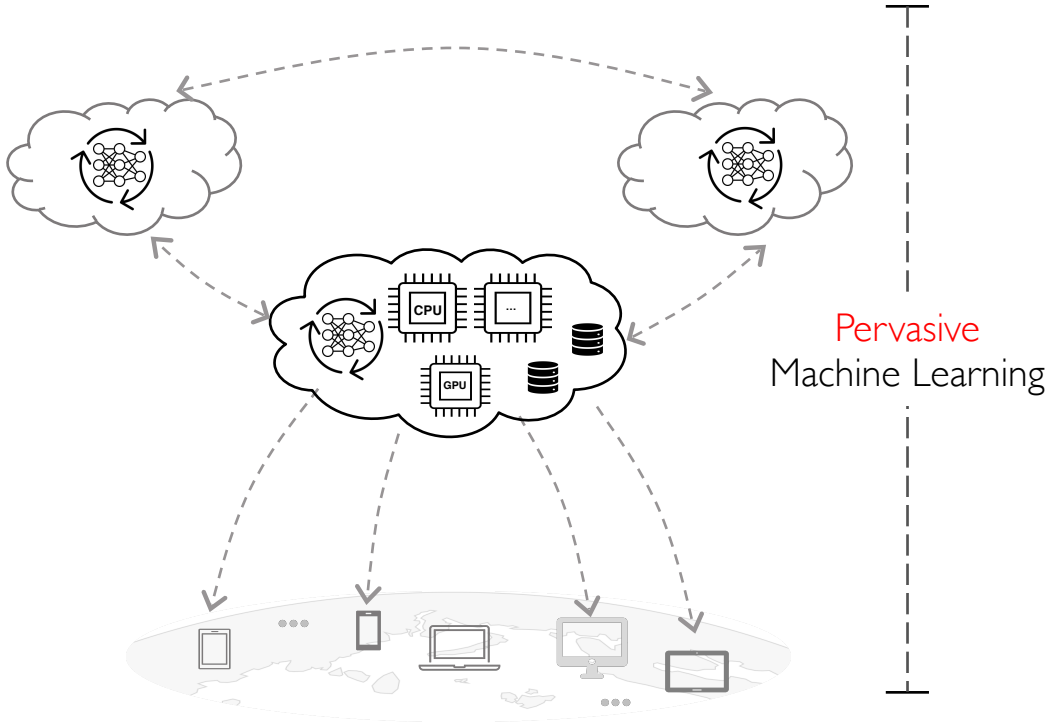
- **Research interest**
  - Cloud ML (LLMs, Image/Video Gen)
  - Collaborative ML (on-device AI), ML4Sys



About Fan

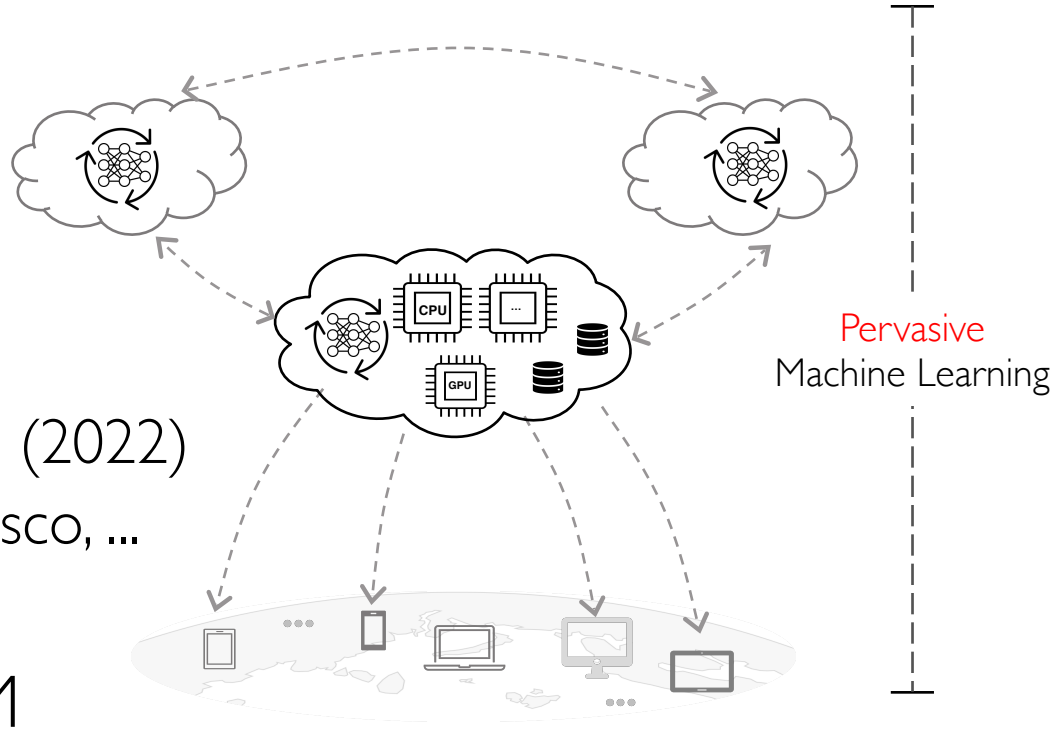
- **Assistant Professor in CS**
  - PhD'23: "Minimalist Systems for Pervasive Machine Learning"

- **Research interest**
  - Cloud ML (LLMs, Image/Video Gen)
  - Collaborative ML (on-device AI), ML4Sys



About Fan

- **Assistant Professor in CS**
  - PhD'23: "Minimalist Systems for Pervasive Machine Learning"
- **Research interest**
  - Cloud ML (LLMs, Image/Video Gen)
  - Collaborative ML (on-device AI), ML4Sys
- **Industry**
  - Visiting Faculty@Google (2023-2024), Meta AI (2022)
  - MLSys adoptions at Google, Meta, LinkedIn, Cisco, ...
- **Office hours:** SC 3128, Friday 2 PM – 3 PM



About Chengsong Zhang (TA #1)

- **Research on image generation**
- **Created several Stable Diffusion WebUI Extensions**
  - Segment Anything (Github 3.3k stars)
  - AnimateDIFF extension (Github 3k stars)
- **Office hours from next week**
  - Zoom, Wednesday 7 PM - 8 PM

About Jimmy Shong (TA #2)

- **Research on hyper-scale model training**
- **Worked on On-Device LLM Chatbot at MIT**
  - TinyVoiceChat (Github ~800 stars)
- **Office hours from next week**
  - Zoom, Wednesday 7 PM - 8 PM

Status and Prerequisites

- **As of today: ~92 registered**
  - If you are not planning to take the class, please drop ASAP

Status and Prerequisites

- **As of today: ~92 registered**
  - If you are not planning to take the class, please drop ASAP
- **Systems Courses: OS/databases/distributed systems/networking**
  - Equivalent courses are acceptable as well
- **ML/AI course is helpful but not required**
- **Good programming skills** (e.g., PyTorch, TensorFlow, ...)
  - Build systems for course project

Course Schedule

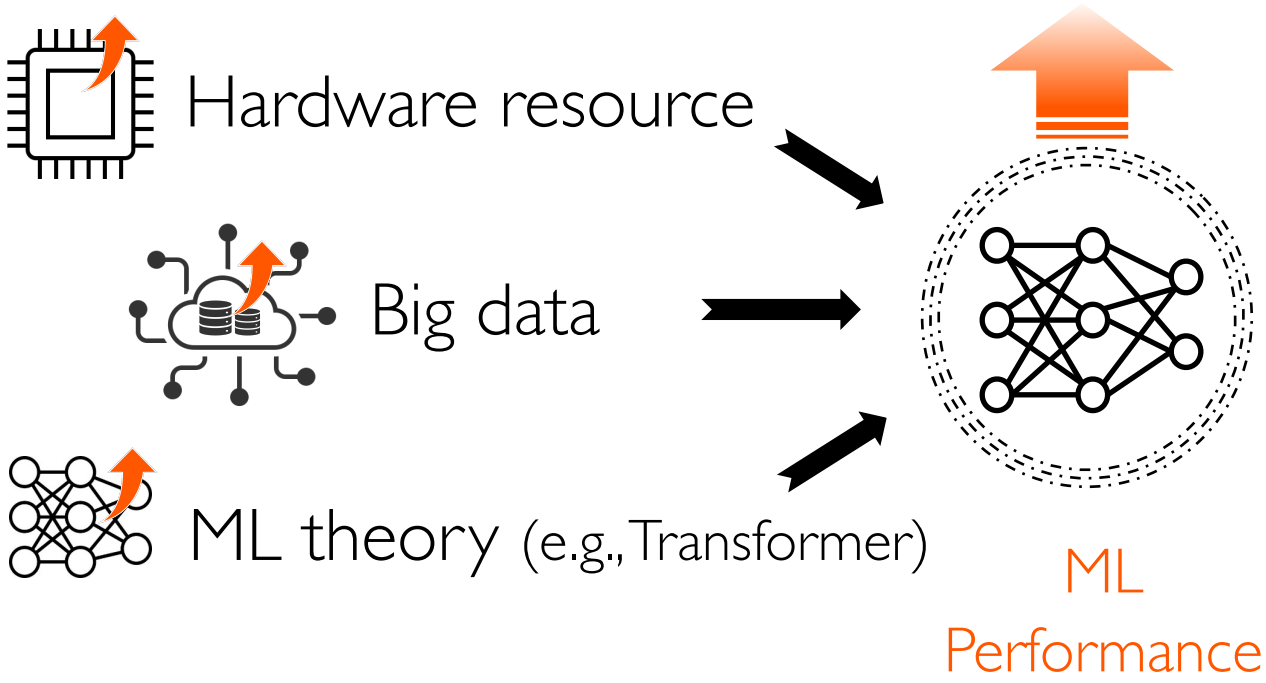
- Webpage: https://github.com/fanlai0990/cs598
- Discussion, Questions: Please Use Piazza
- In-person lectures, discussion
- Feel free to drop us email too :)

Agenda

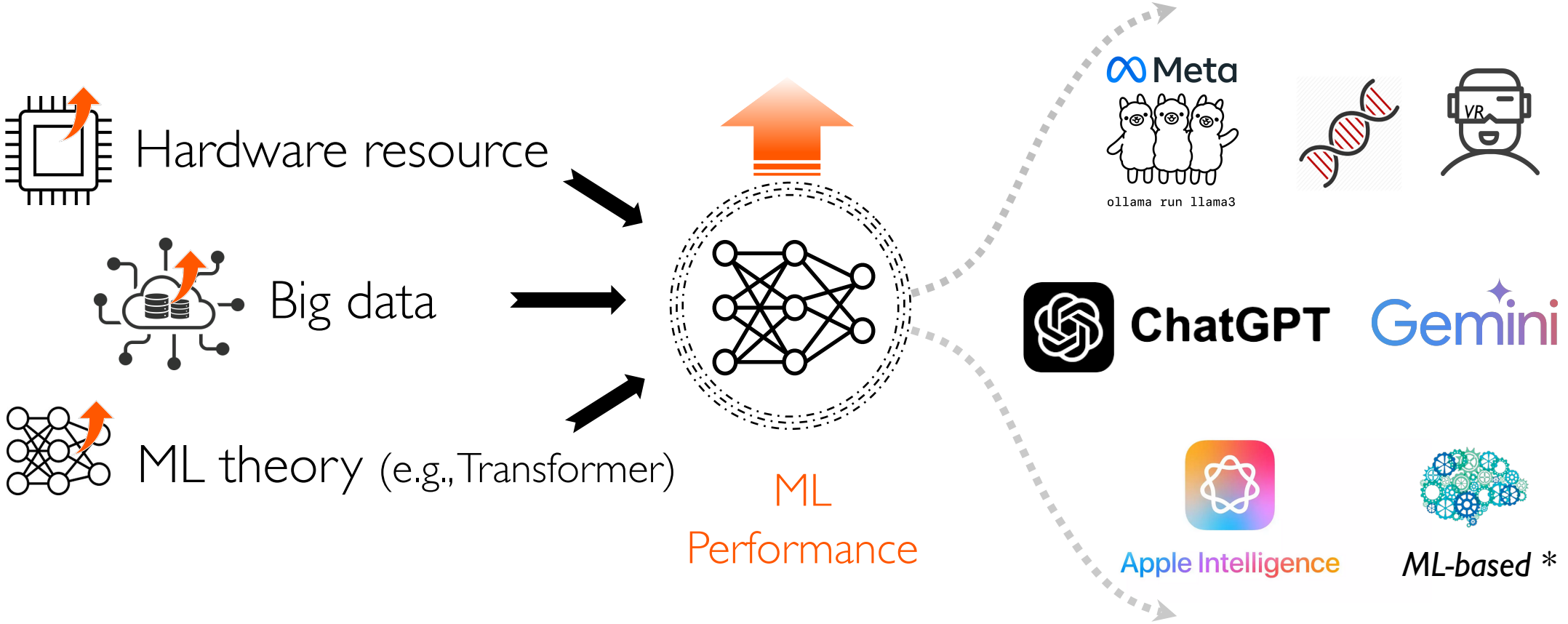
- What will you learn in this course?
  - GenAI, systems, networks, ...
- What will you do in this course?
  - Participation, presentation, research project, ...

What Do We Talk About When We Talk About "Systems for GenAI"

Cloud Enabled Last Leap in Machine Learning (ML)

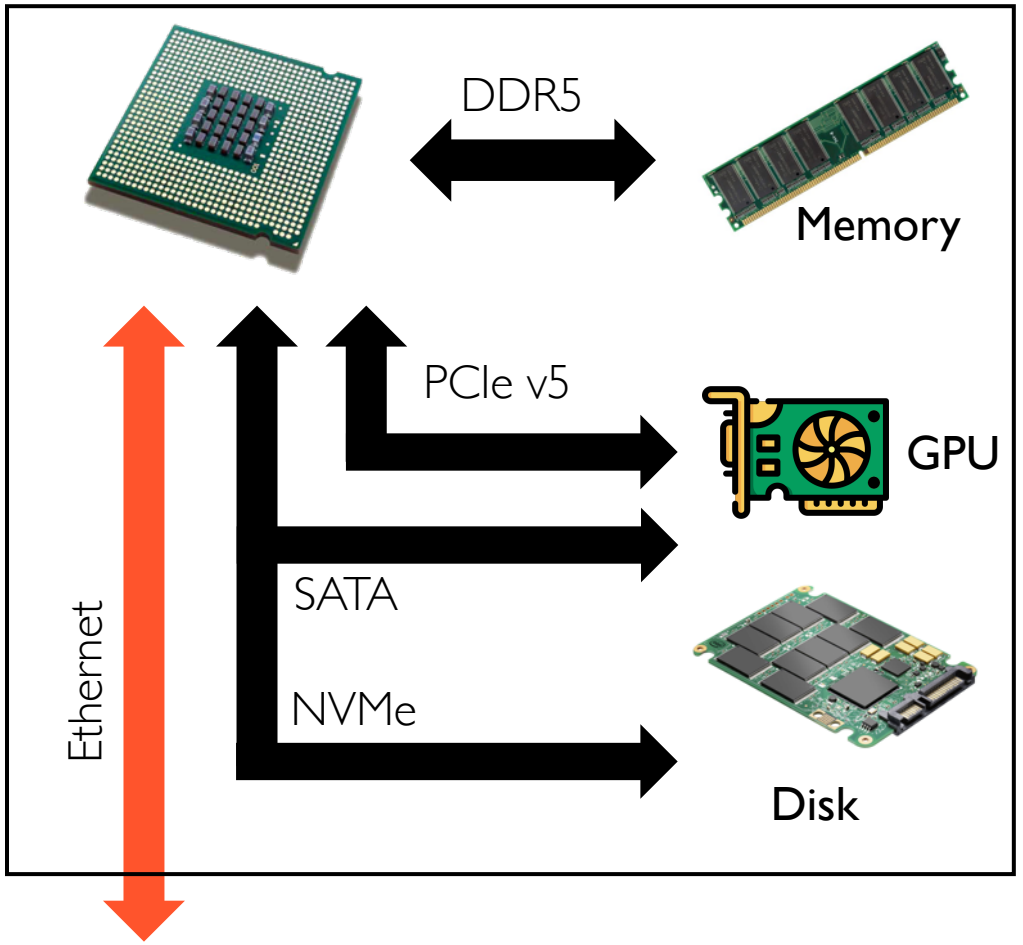


Cloud Enabled Last Leap in Machine Learning (ML)



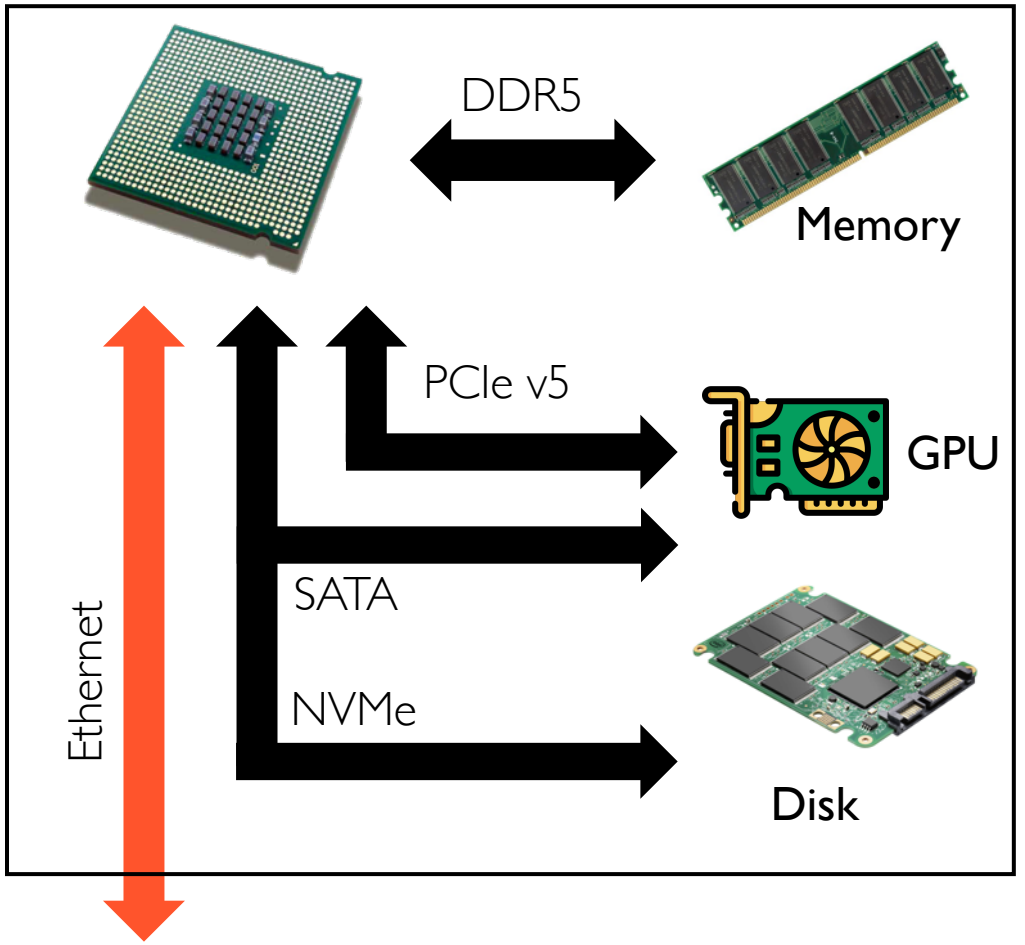
What's in a (Simplified) AI Server?

- **Interconnected compute and storage resources**



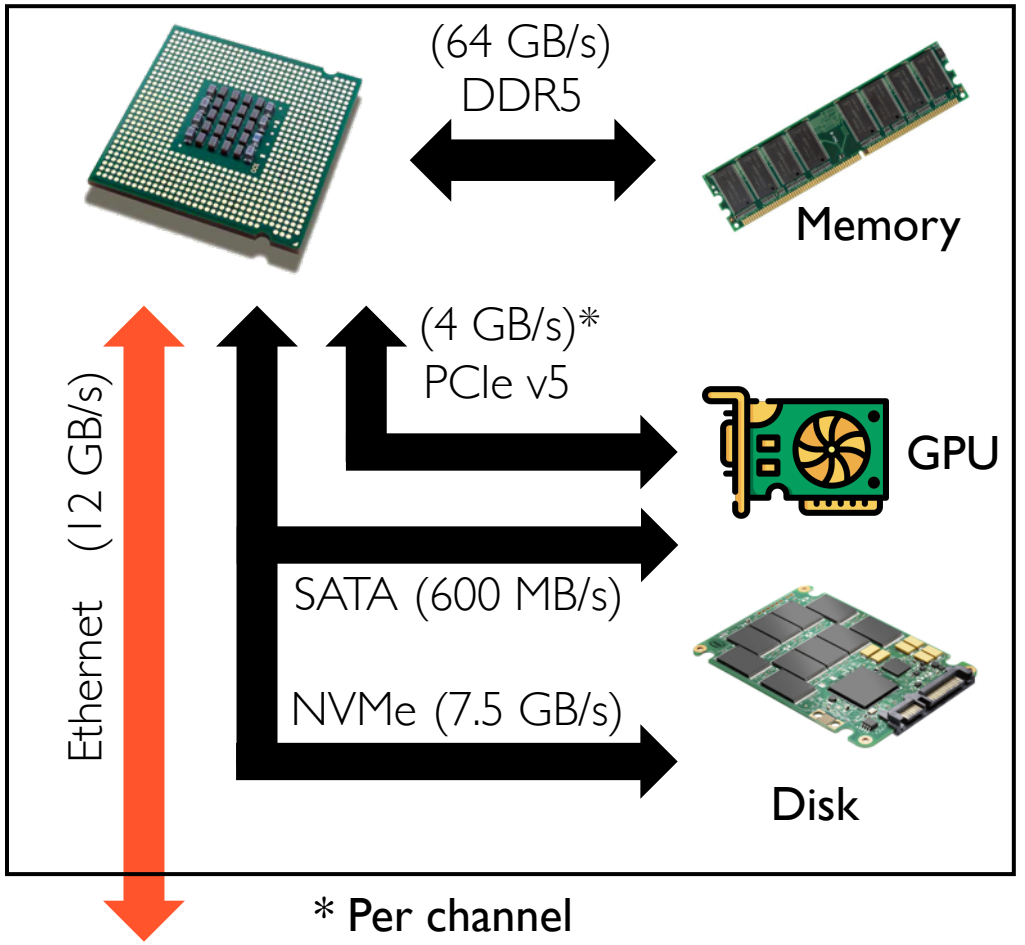
What's in a (Simplified) AI Server?

- **Interconnected compute and storage resources**
  - Different bandwidth and latency constraints



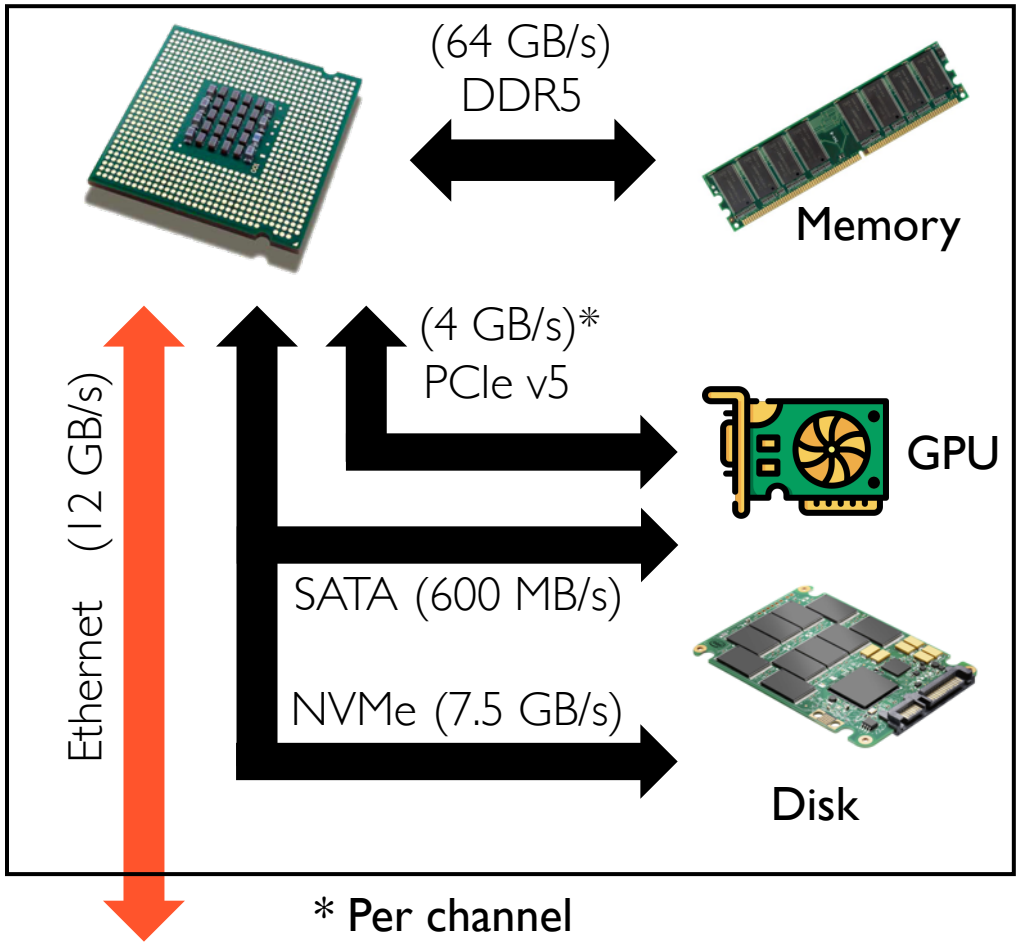
What's in a (Simplified) AI Server?

- **Interconnected compute and storage resources**
  - Different bandwidth and latency constraints

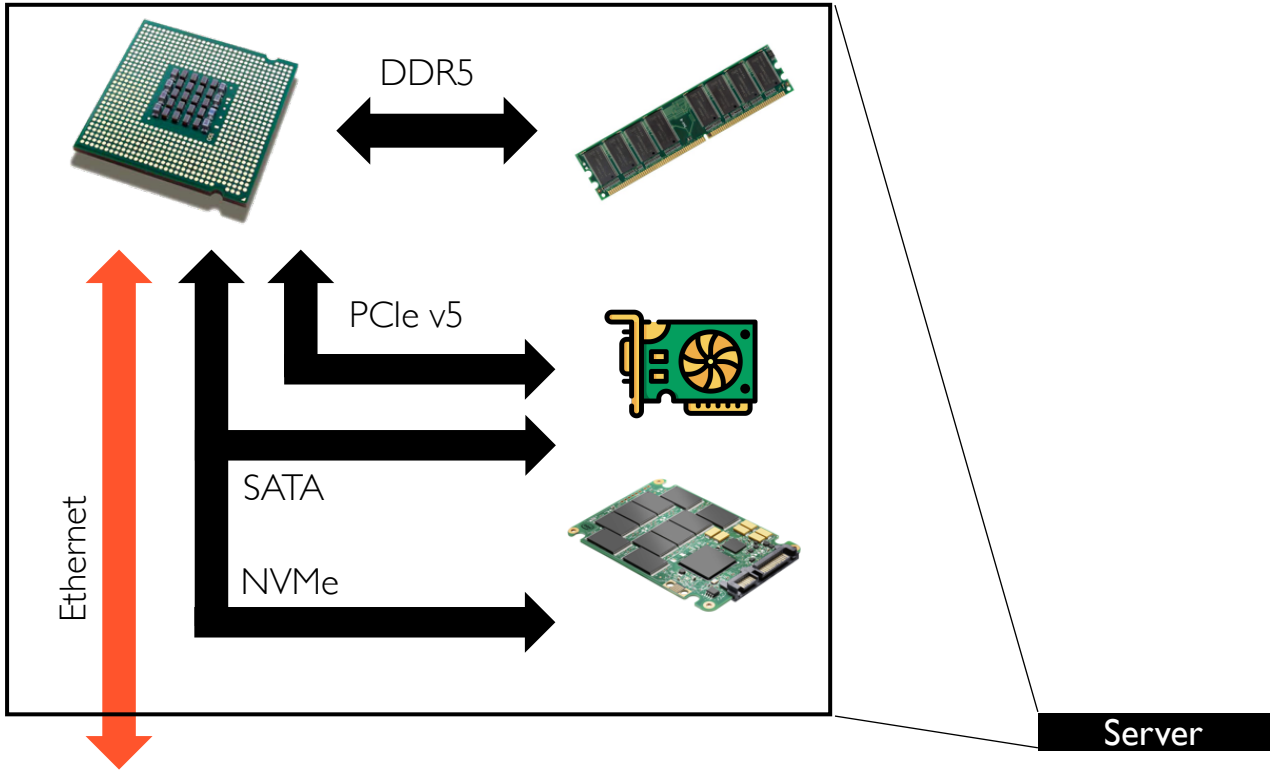


What's in a (Simplified) AI Server?

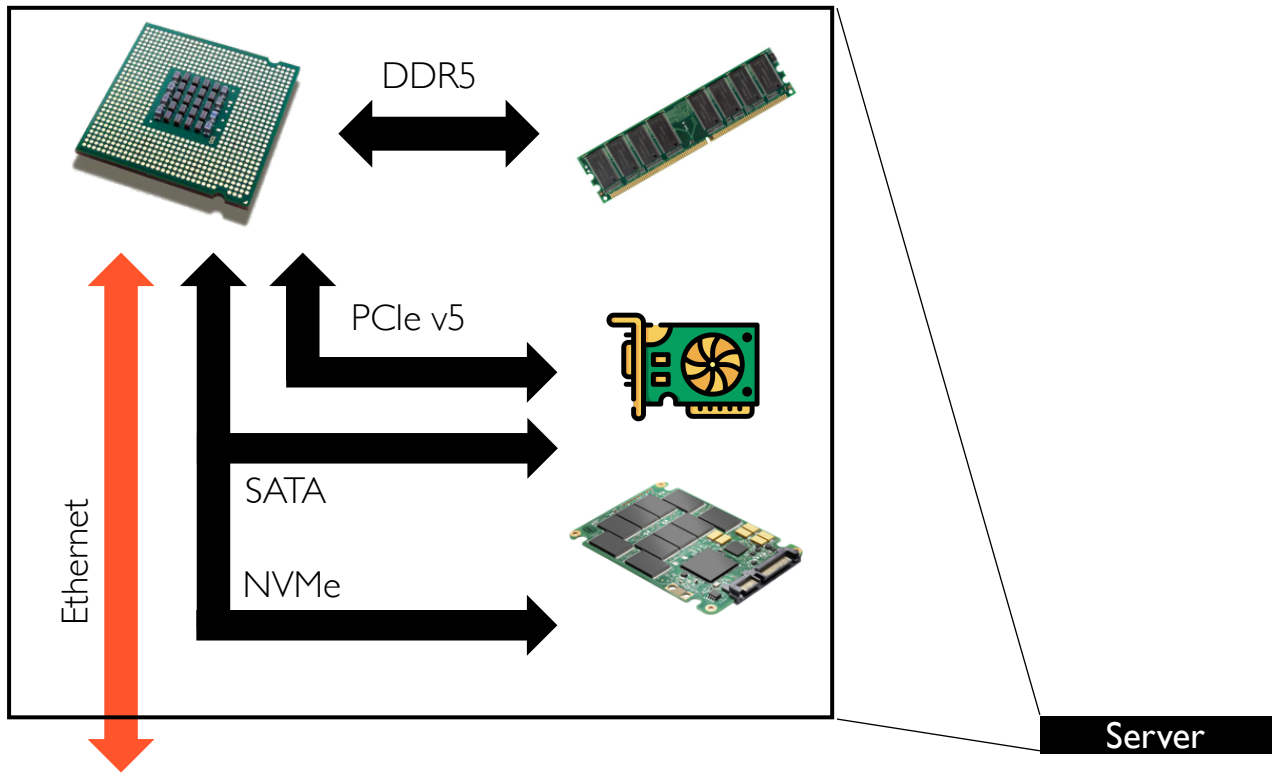
- **Interconnected compute and storage resources**
  - Different bandwidth and latency constraints
- **Simplified diagram**
  - Doesn't include faster networks such as RDMA, dedicated GPU interconnects such as NVLink



Cloud Datacenter: Scale-Out Servers

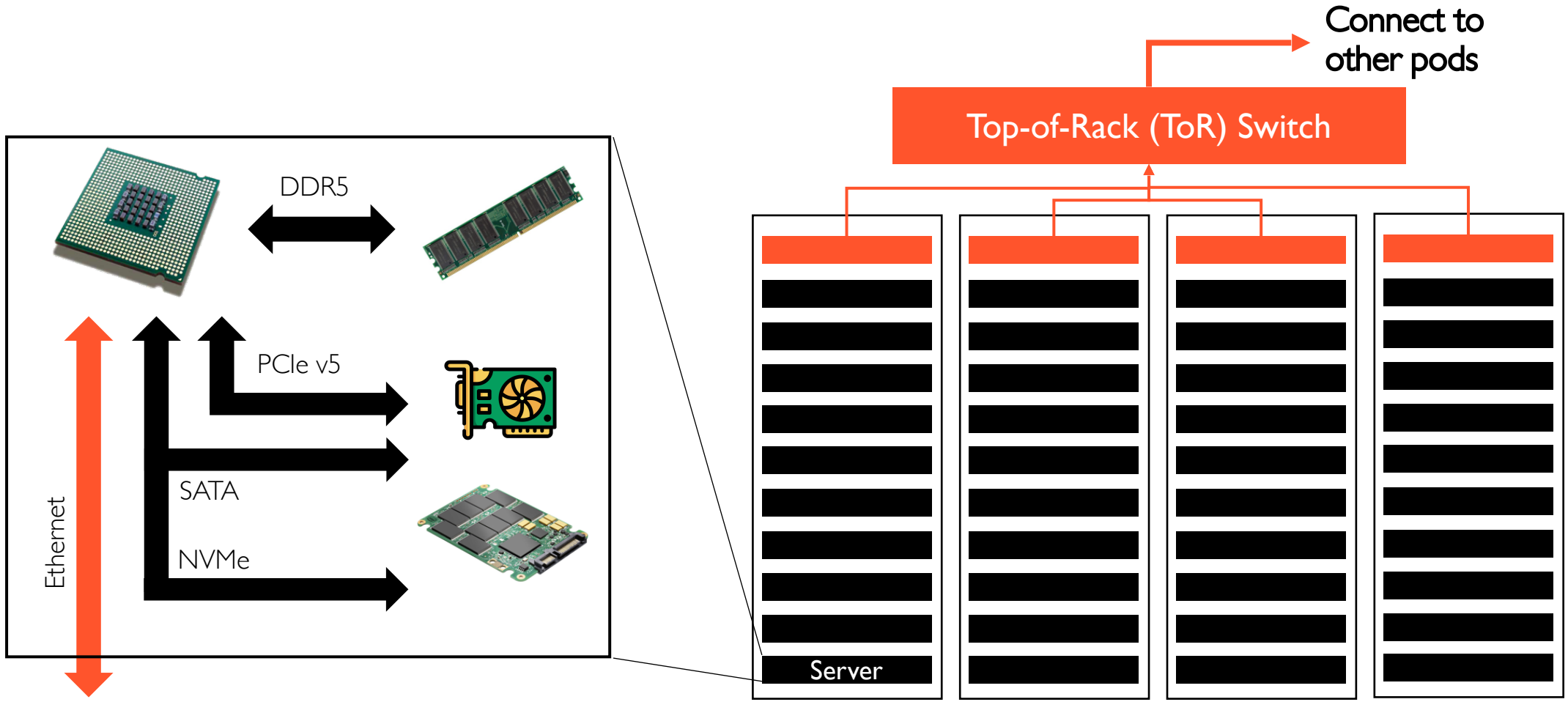


Cloud Datacenter: Scale-Out Servers



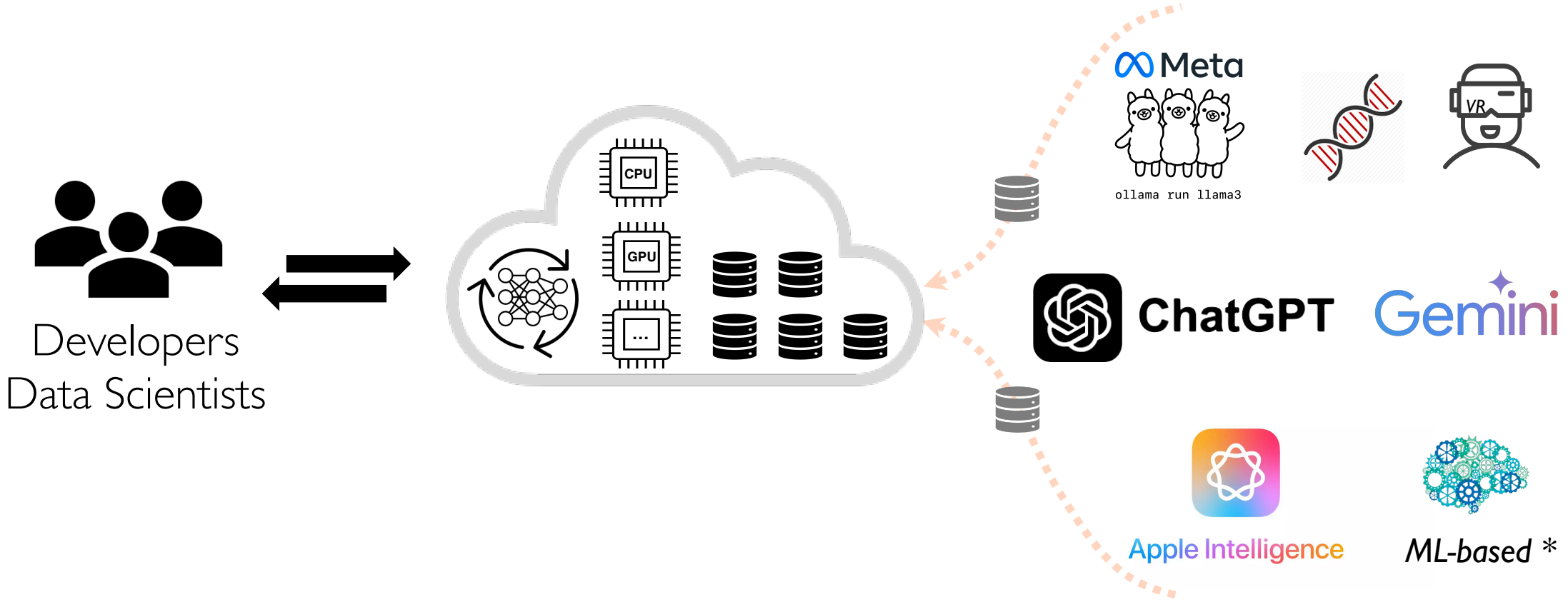
Meta's AI datacenter houses 16K of its 24K GPUs for training the latest Llama 3 model

Cloud Datacenter: Scale-Out Servers



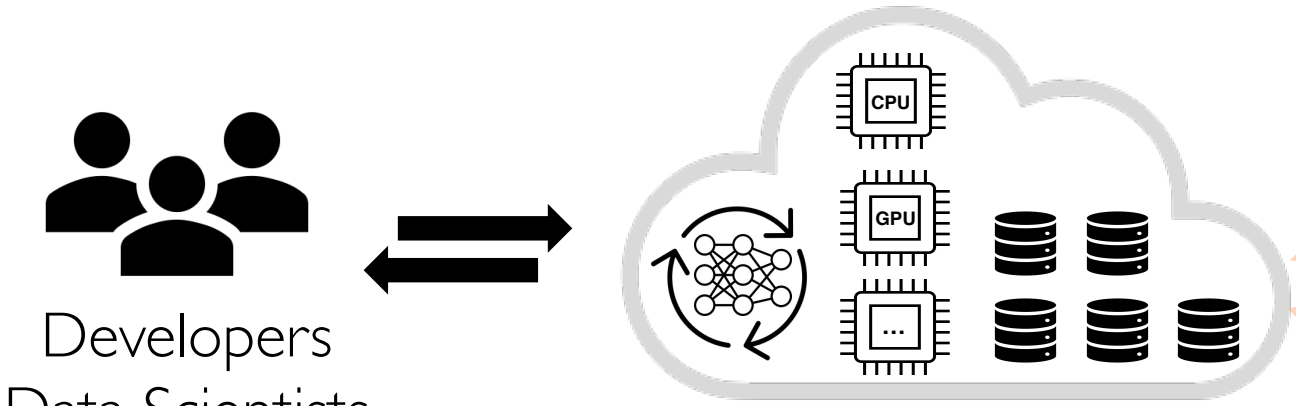
Meta's AI datacenter houses 16K of its 24K GPUs for training the latest Llama 3 model

Cloud Faces Skyrocketing GenAI Workloads



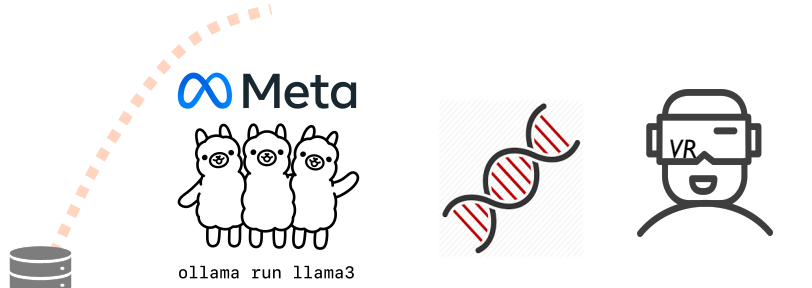
Cloud Faces Skyrocketing GenAI Workloads

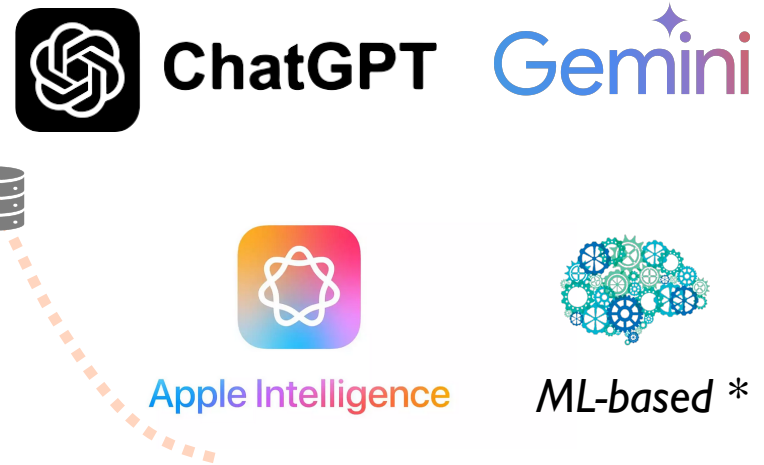
Model size **doubled** every two years [1]



Developers
Data Scientists

Petabytes of data ingestion every day [1]

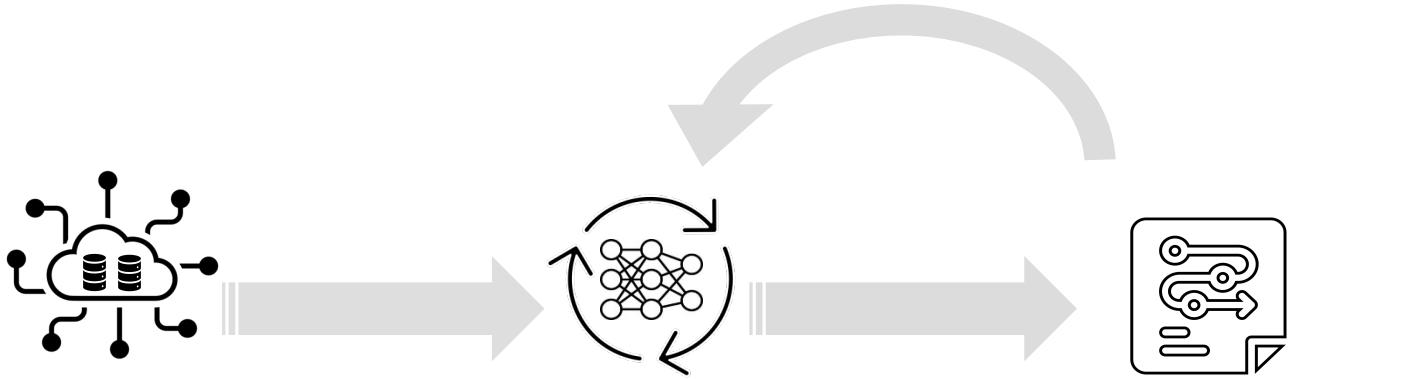




[1] "Sustainable AI: Environment Implications, Challenges and Opportunities", Meta, MLSys'22

GenAI Lifecycle Demands Great Systems Support

GenAI Lifecycle Demands Great Systems Support

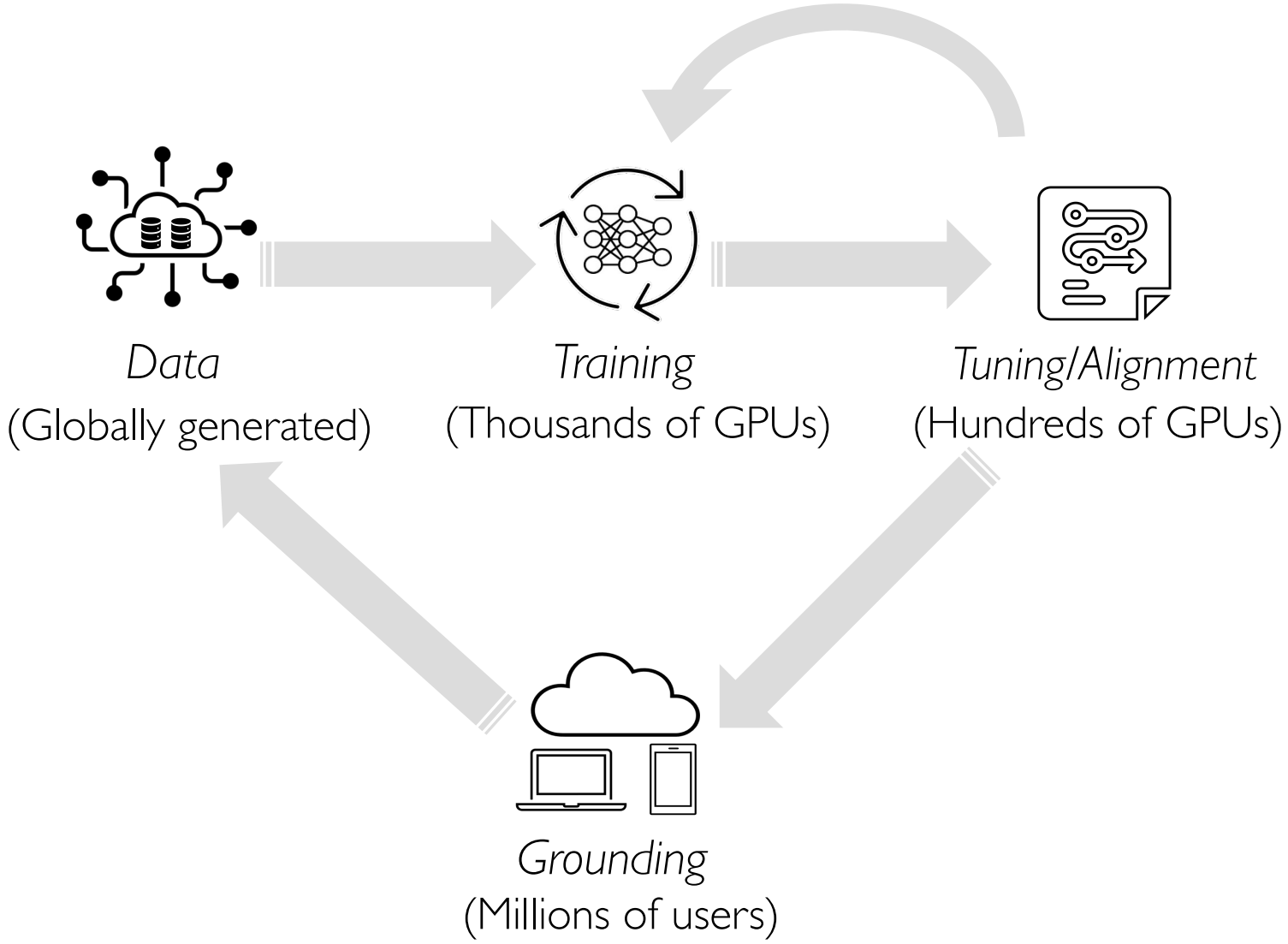


Data
(Globally generated)

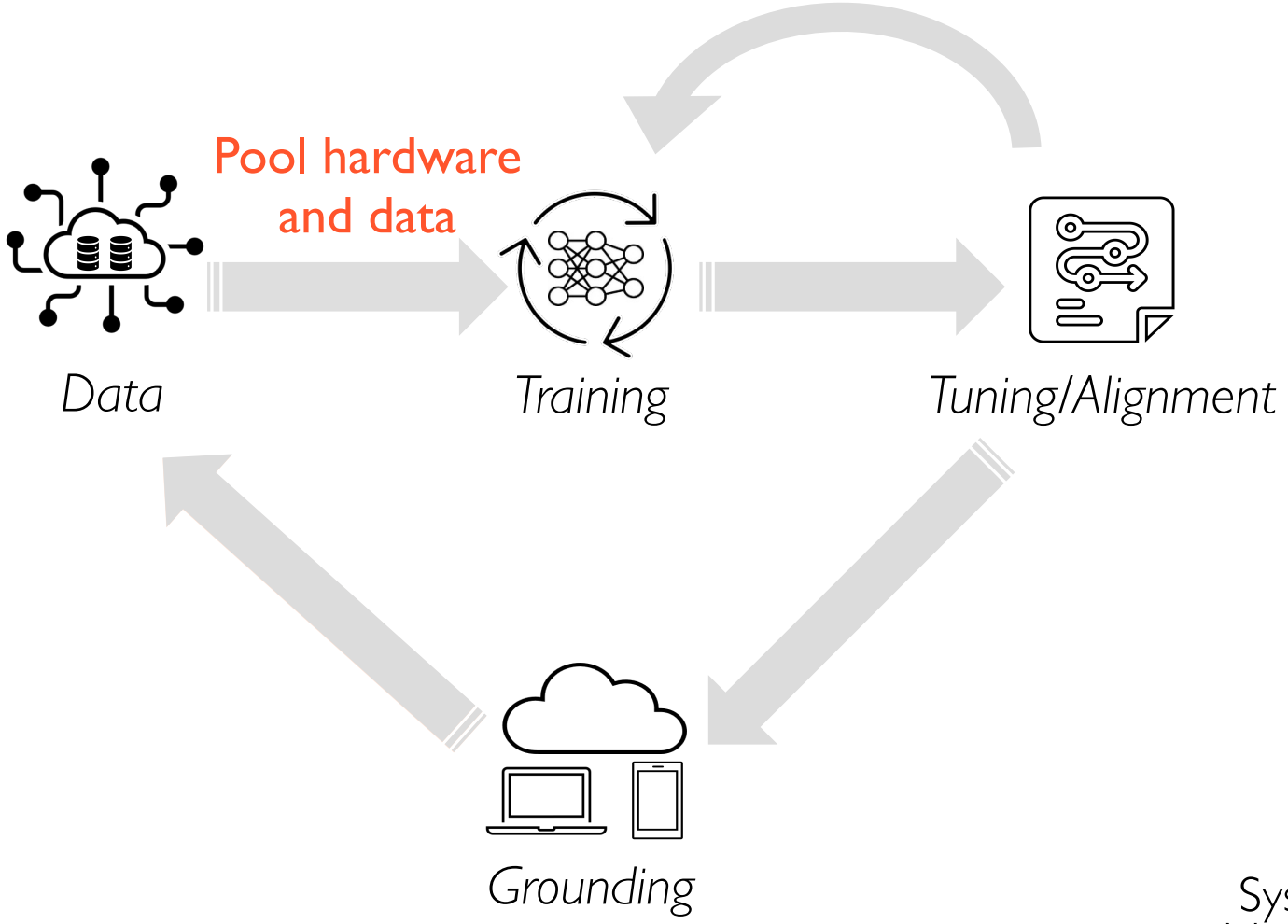
Training
(Thousands of GPUs)

Tuning/Alignment
(Hundreds of GPUs)

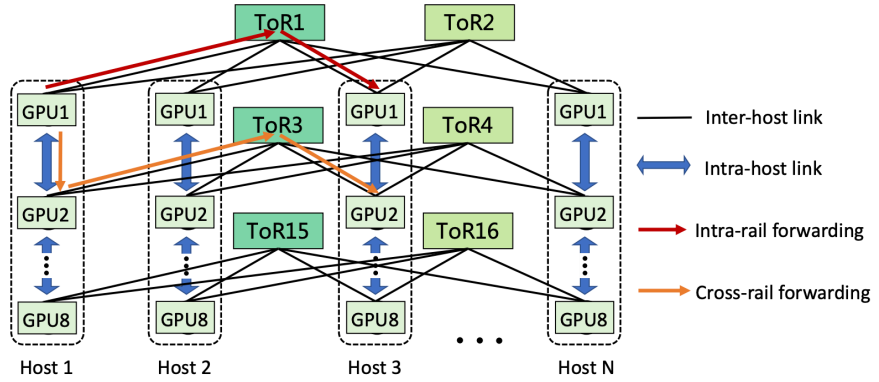
GenAI Lifecycle Demands Great Systems Support



GenAI Lifecycle Demands Great Systems Support



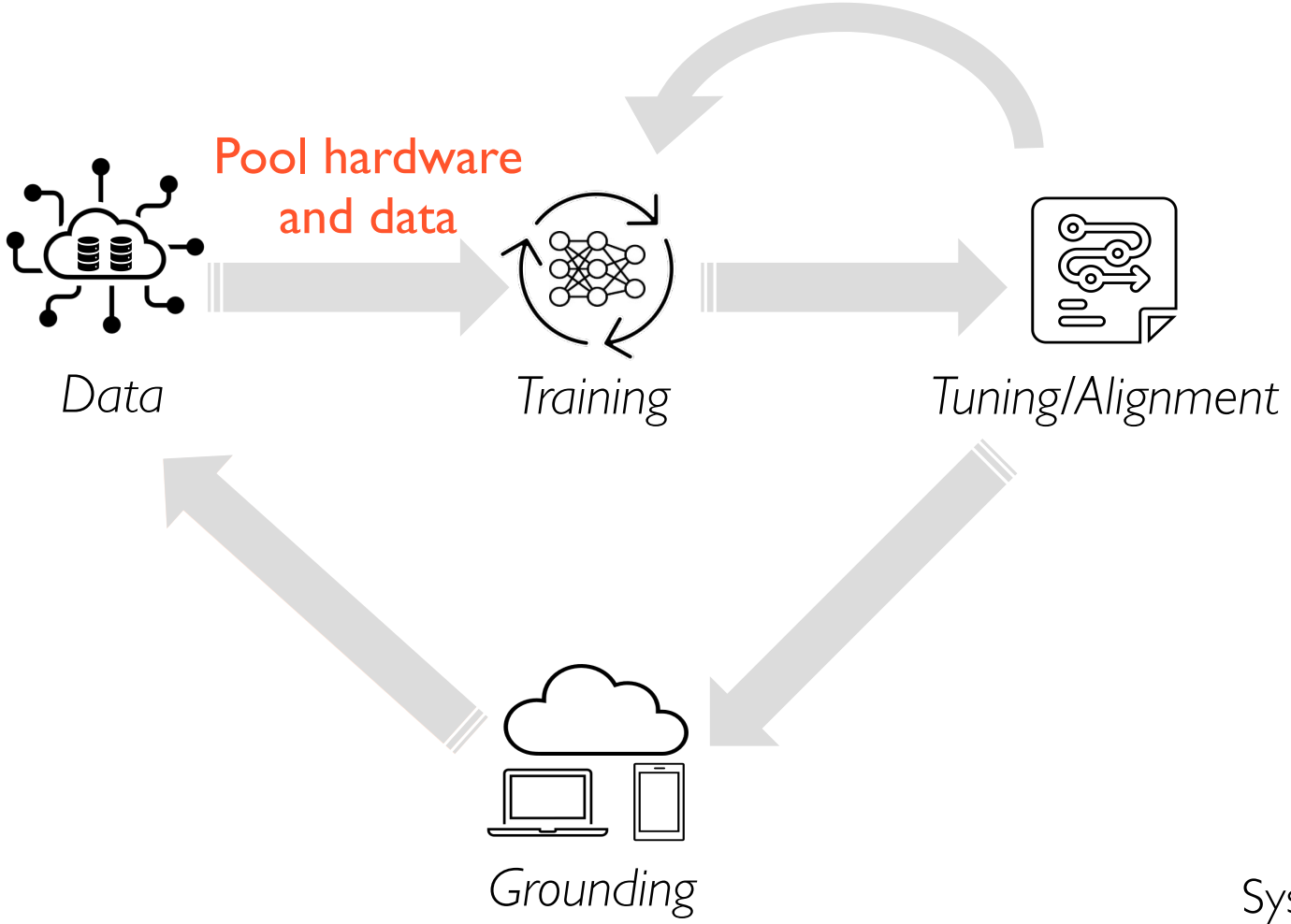
ML



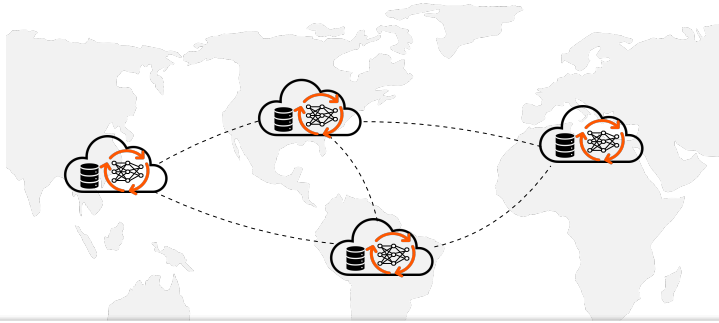
HyperScale Training Infra
(e.g., compute, communication, memory)

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



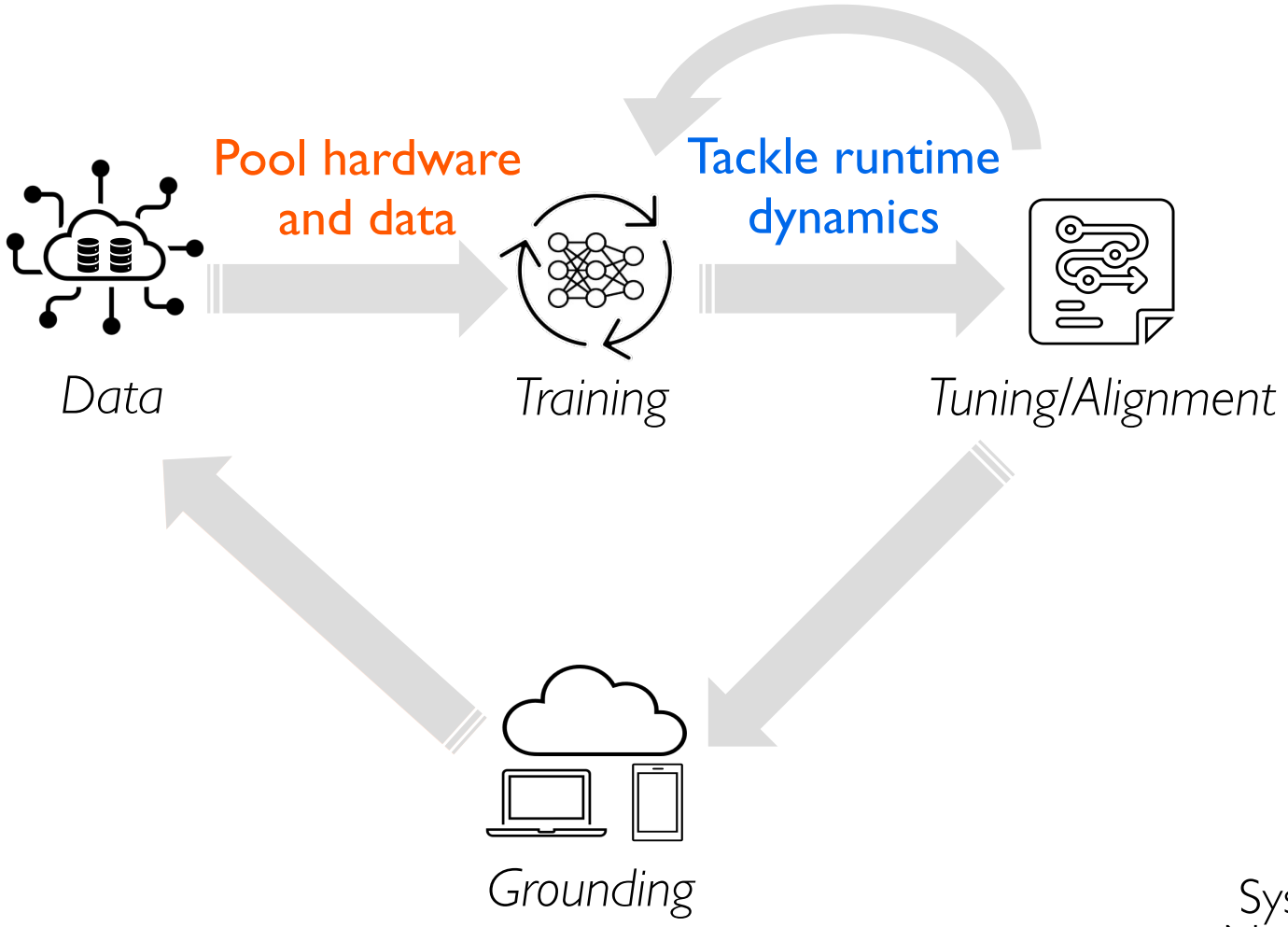
Distributed model training across massive servers and even datacenters

Model Auto-Sharding

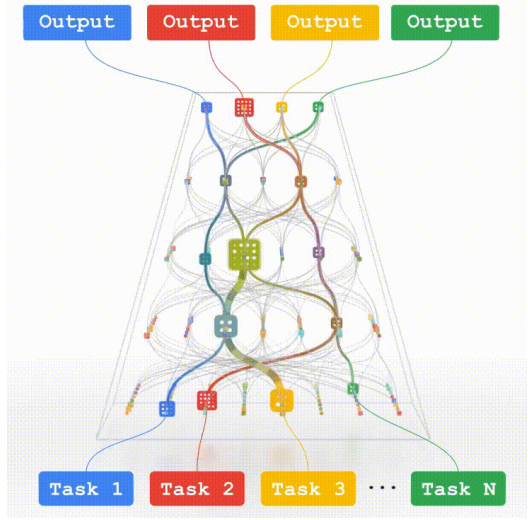
HyperScale Training Infra
(e.g., compute, communication, memory)

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



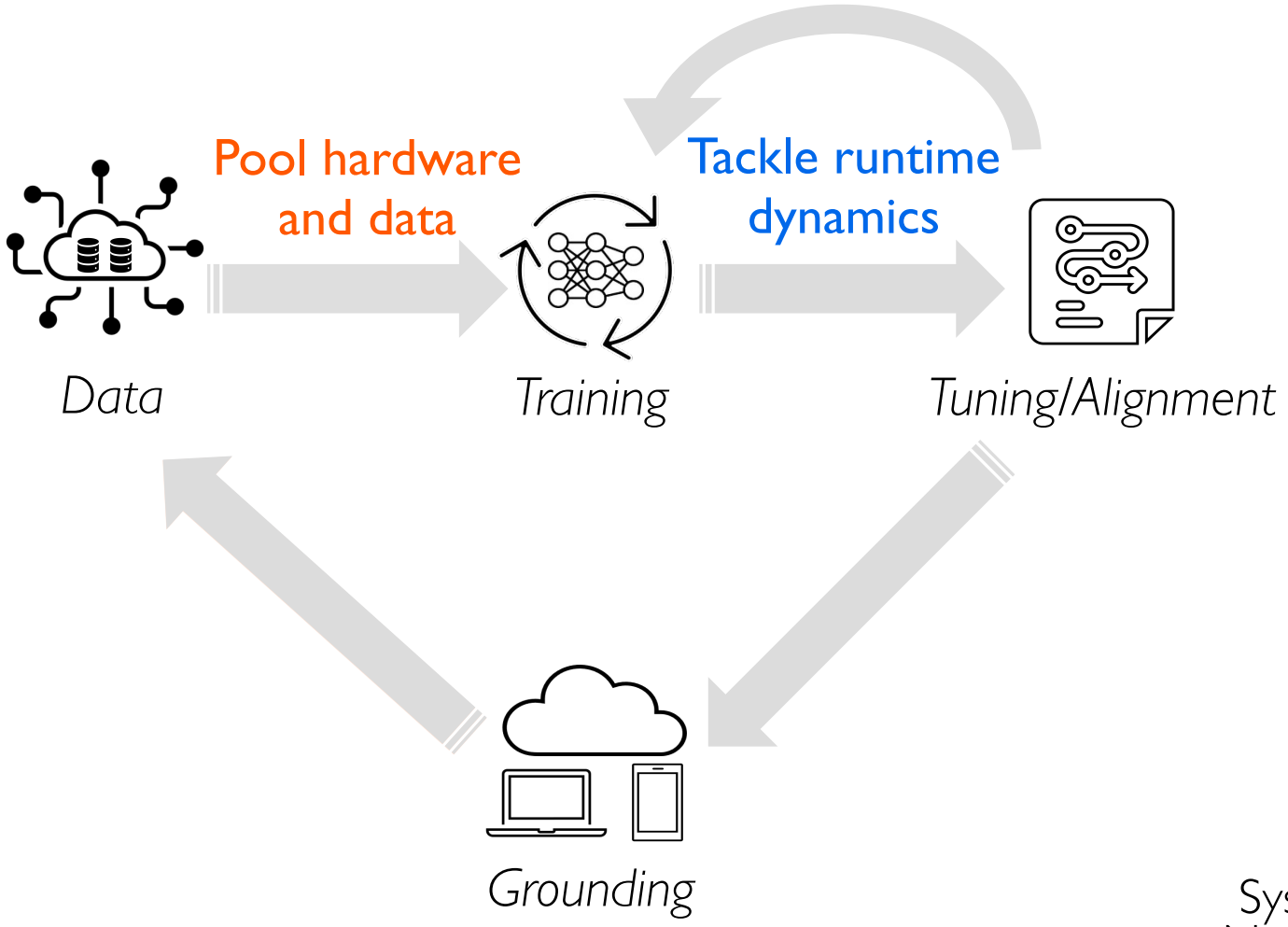
Model Auto-Sharding

Adaptive Runtime

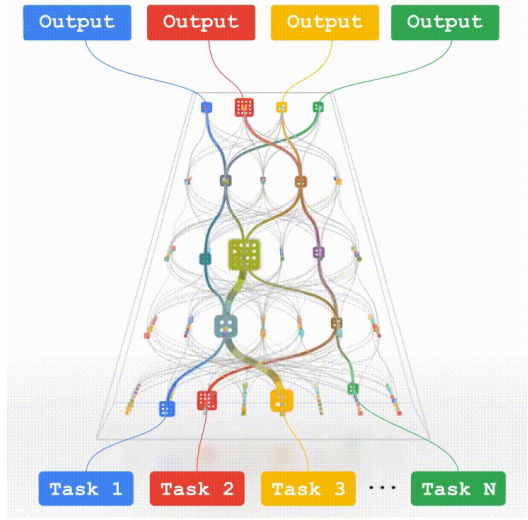
Training can dynamically activate model components (e.g., MoE, sparsity)

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



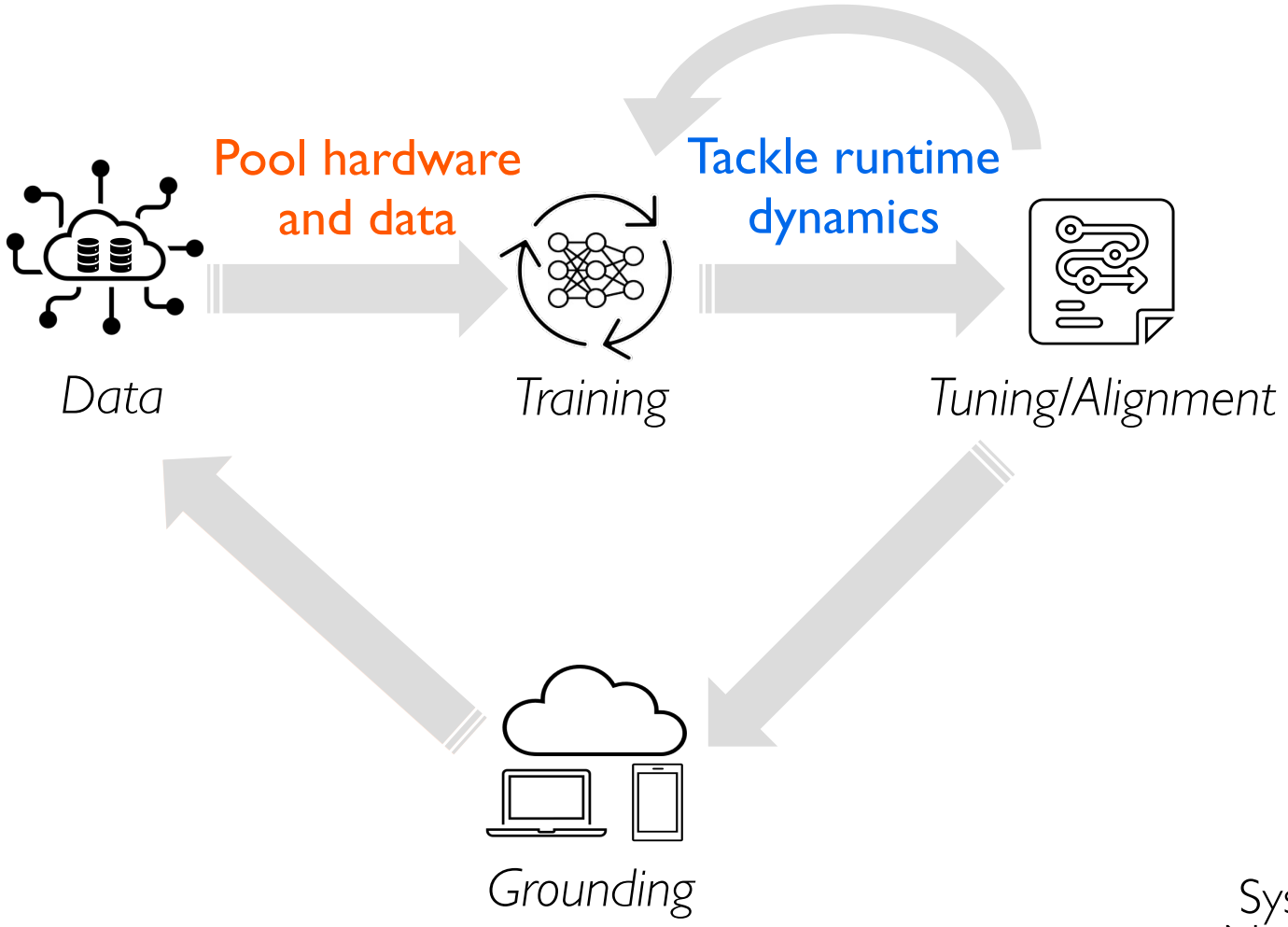
Model Auto-Sharding

Adaptive Runtime

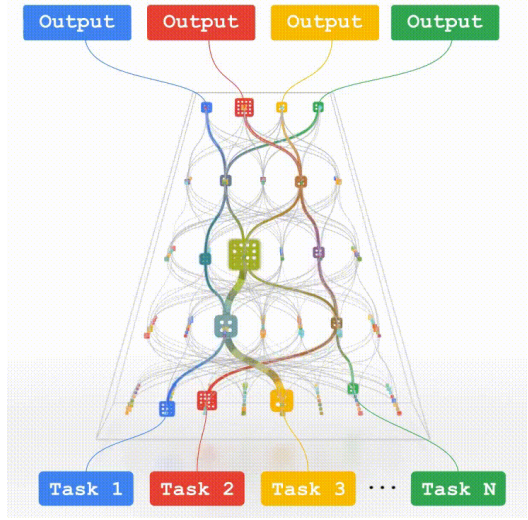
Training can dynamically activate model components (e.g., MoE, sparsity)

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



Model Auto-Sharding

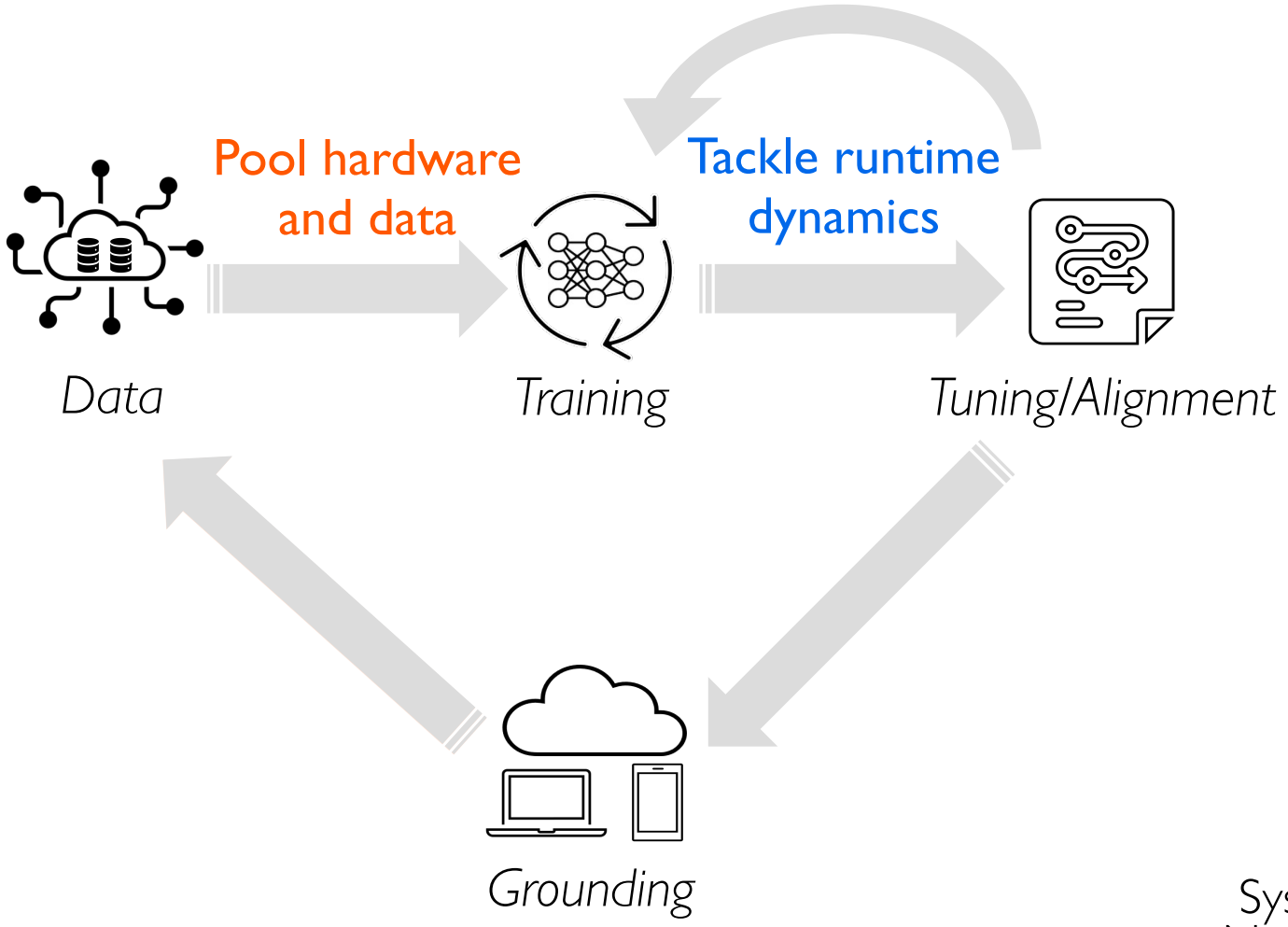
Adaptive Runtime

Multi-job Scheduling

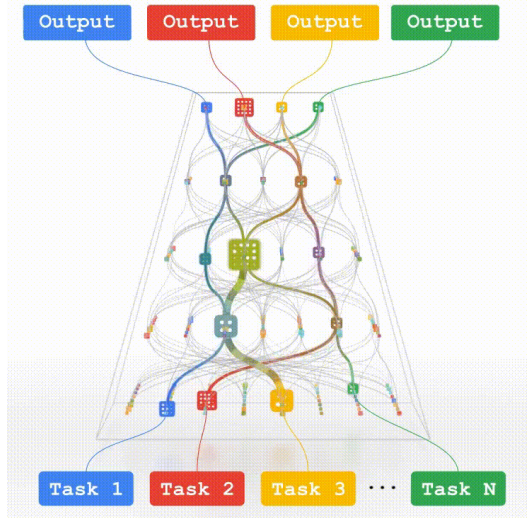
Jobs can compete for the shared resources

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



Model Auto-Sharding

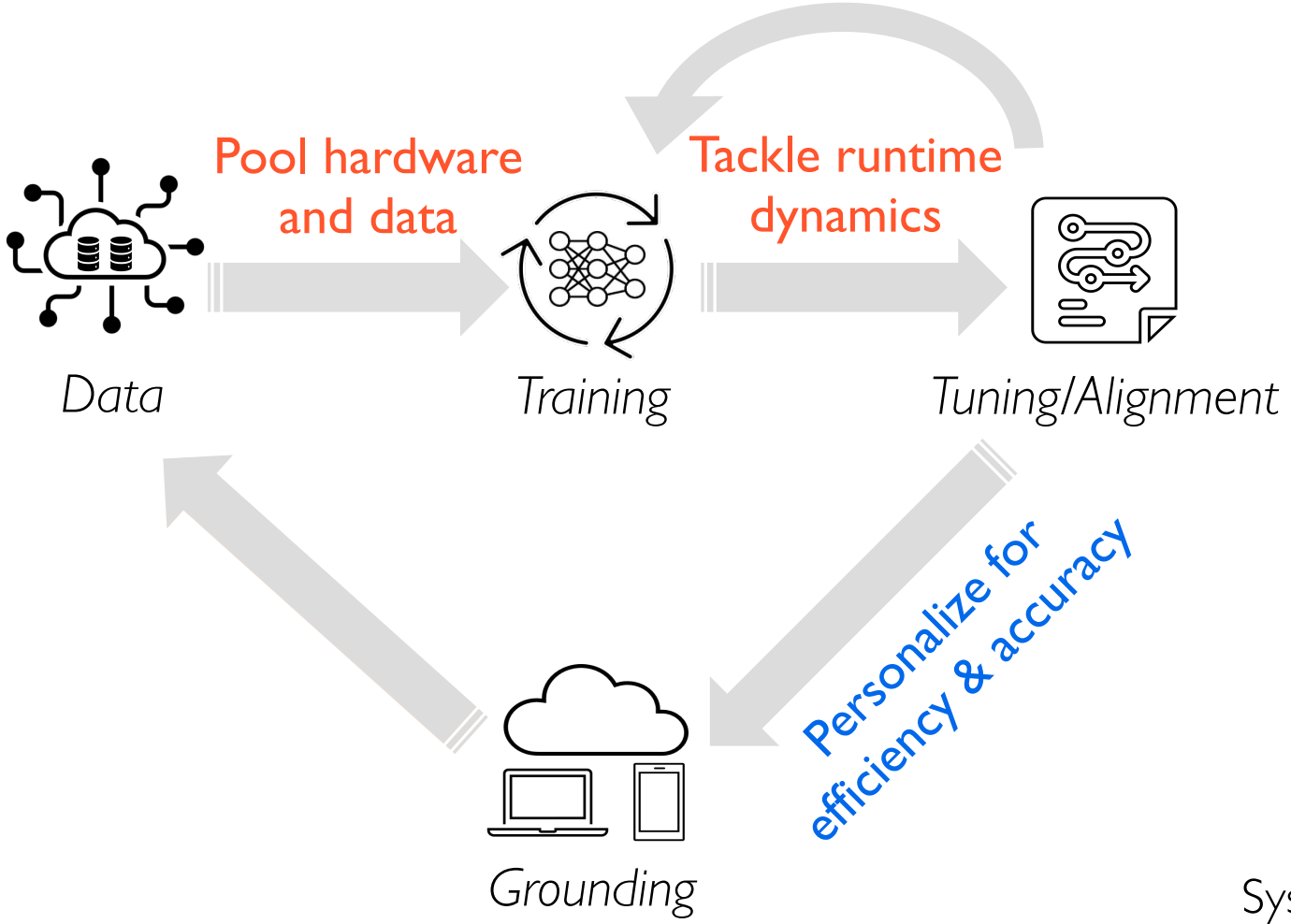
Adaptive Runtime

Multi-job Scheduling

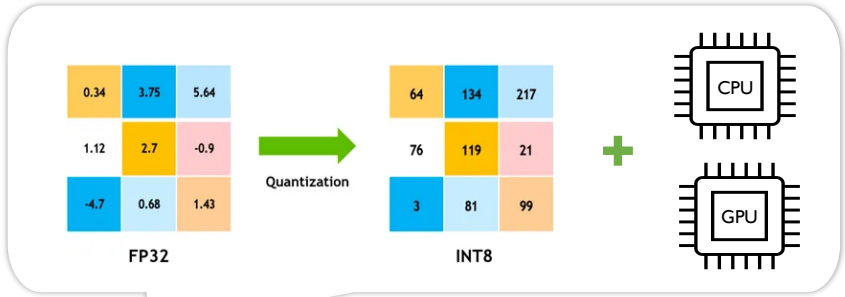
Jobs can compete for the shared resources

Systems & Networking

GenAI Lifecycle Demands Great Systems Support



ML



Model Quantization

Model Auto-Sharding

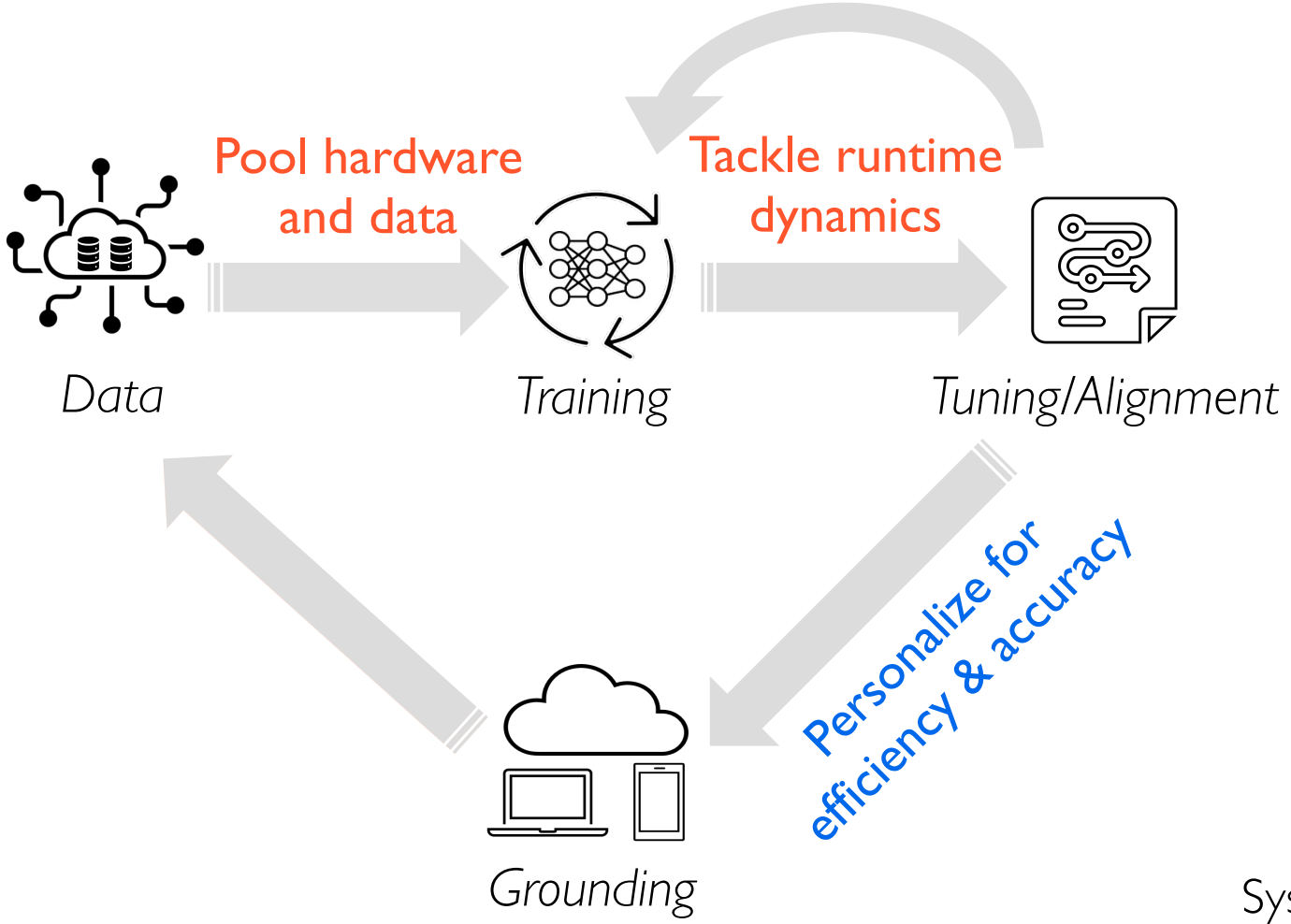
Adaptive Runtime

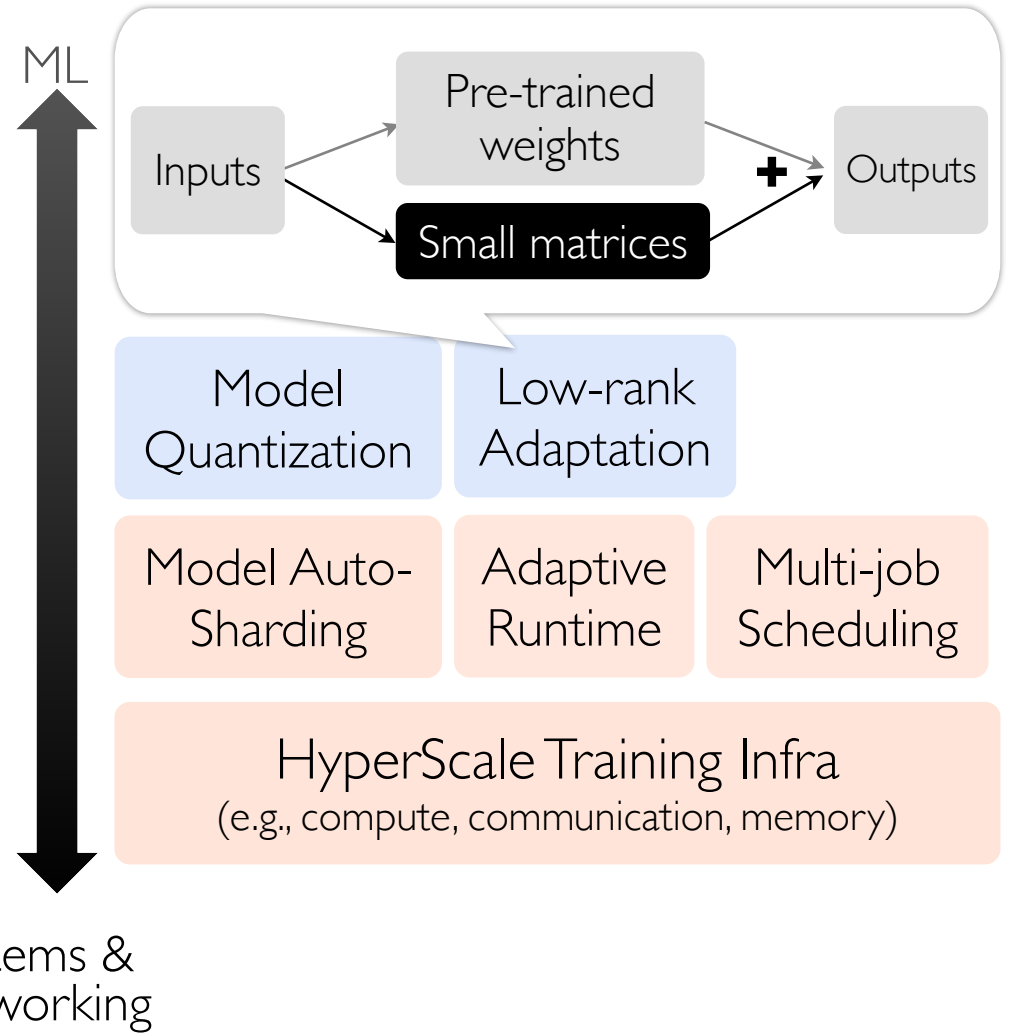
Multi-job Scheduling

HyperScale Training Infra
(e.g., compute, communication, memory)

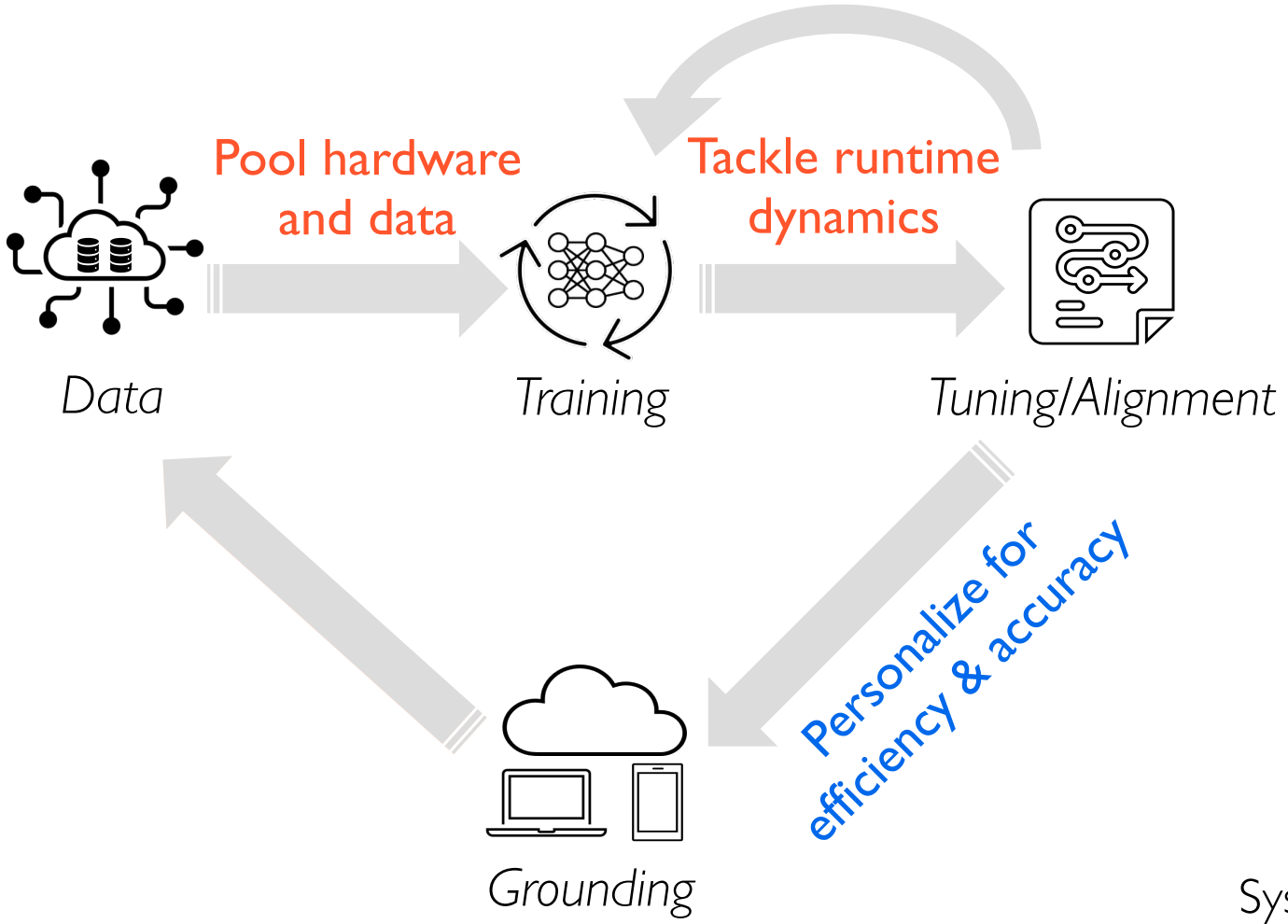
Systems & Networking

GenAI Lifecycle Demands Great Systems Support





GenAI Lifecycle Demands Great Systems Support



ML

Systems & Networking

Overloaded

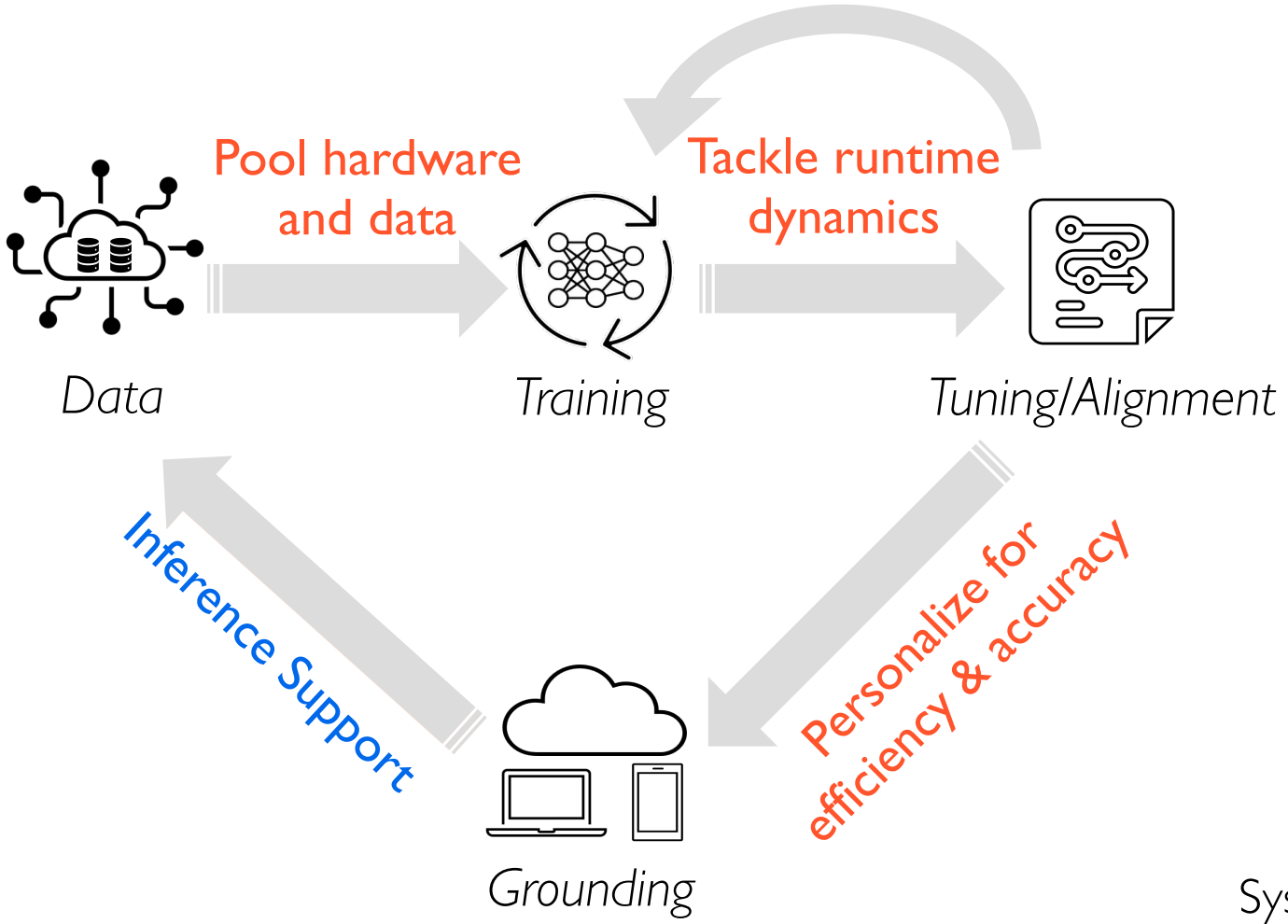
O(1M) requests → Request Scheduler → Model 1 ✗ *404*

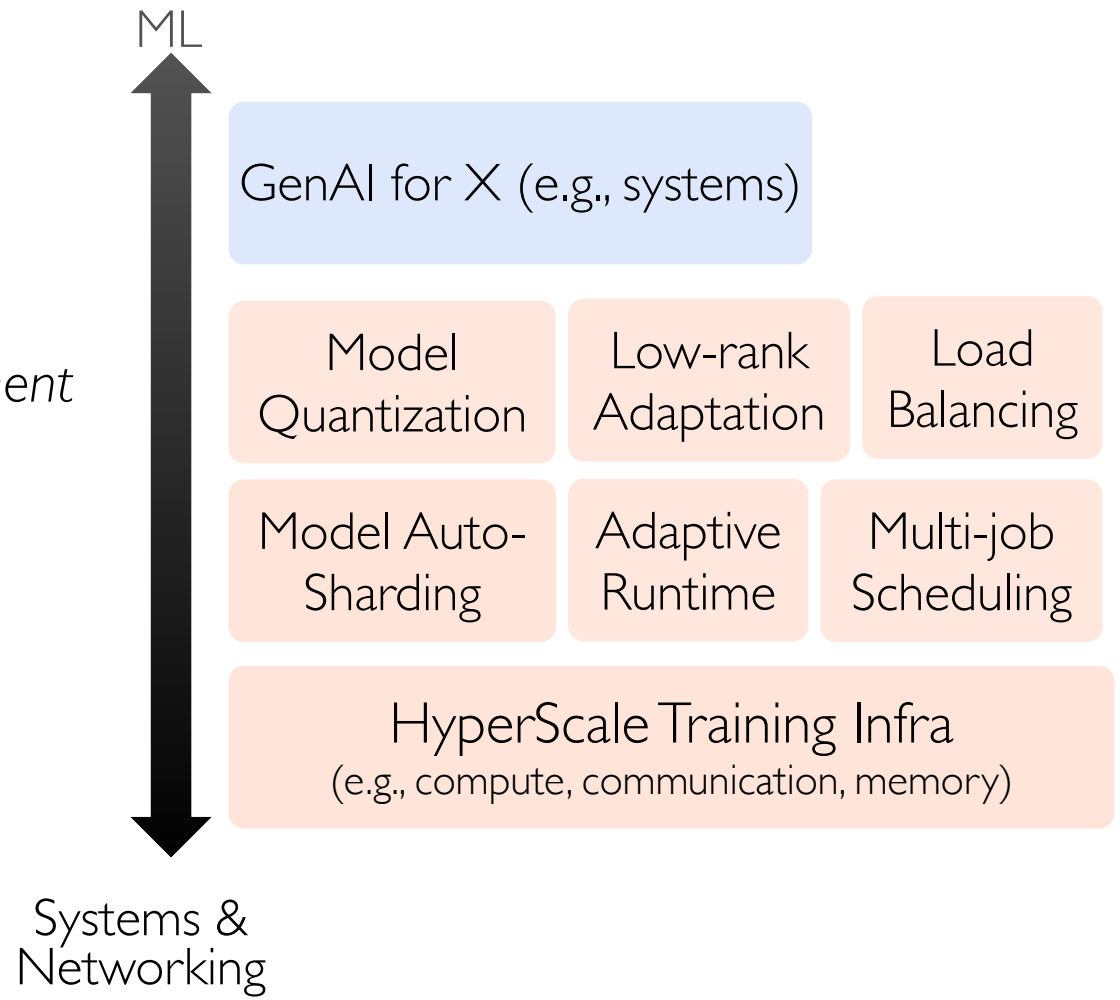
Request Scheduler → Model N

Model Quantization	Low-rank Adaptation	Load Balancing
Model Auto-Sharding	Adaptive Runtime	Multi-job Scheduling

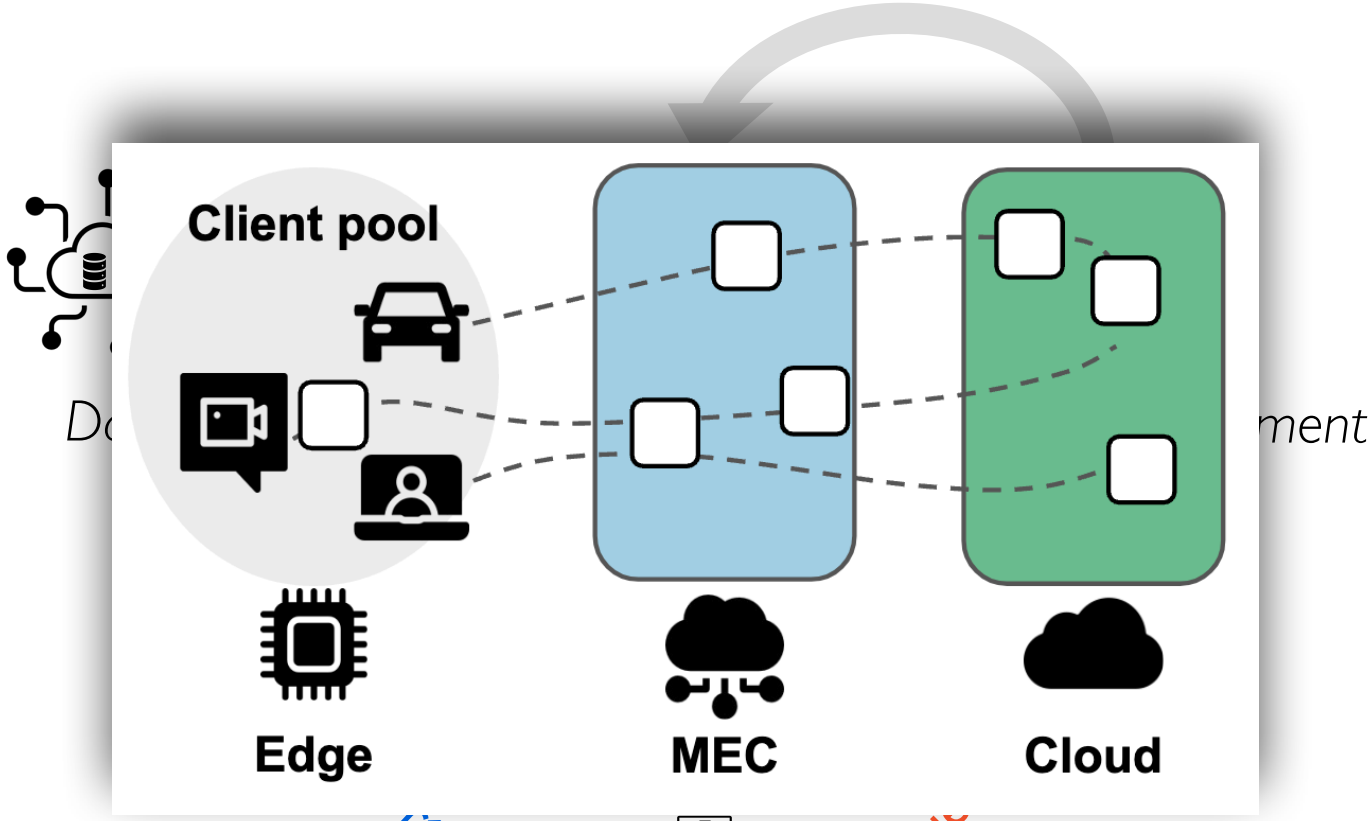
HyperScale Training Infra
(e.g., compute, communication, memory)

GenAI Lifecycle Demands Great Systems Support





GenAI Lifecycle Demands Great Systems Support



ML

Systems & Networking

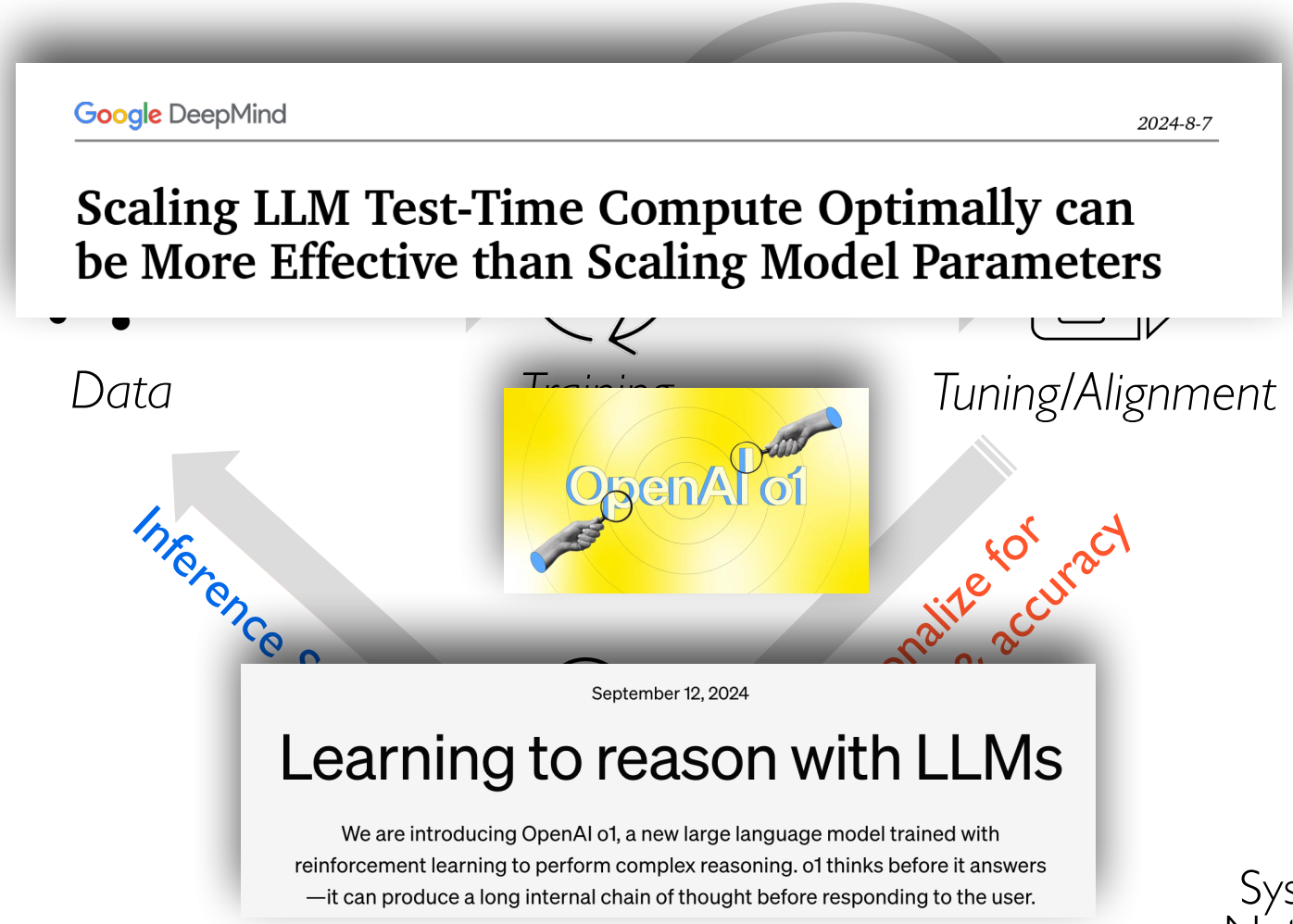
GenAI for X (e.g., systems)

Model Quantization | Low-rank Adaptation | Load Balancing

Model Auto-Sharding | Adaptive Runtime | Multi-job Scheduling

HyperScale Training Infra
(e.g., compute, communication, memory)

GenAI Lifecycle Demands Great Systems Support



ML

GenAI for X (e.g., systems)

Scaling Reasoning

Model Quantization

Low-rank Adaptation

Load Balancing

Model Auto-Sharding

Adaptive Runtime

Multi-job Scheduling

HyperScale Training Infra
(e.g., compute, communication, memory)

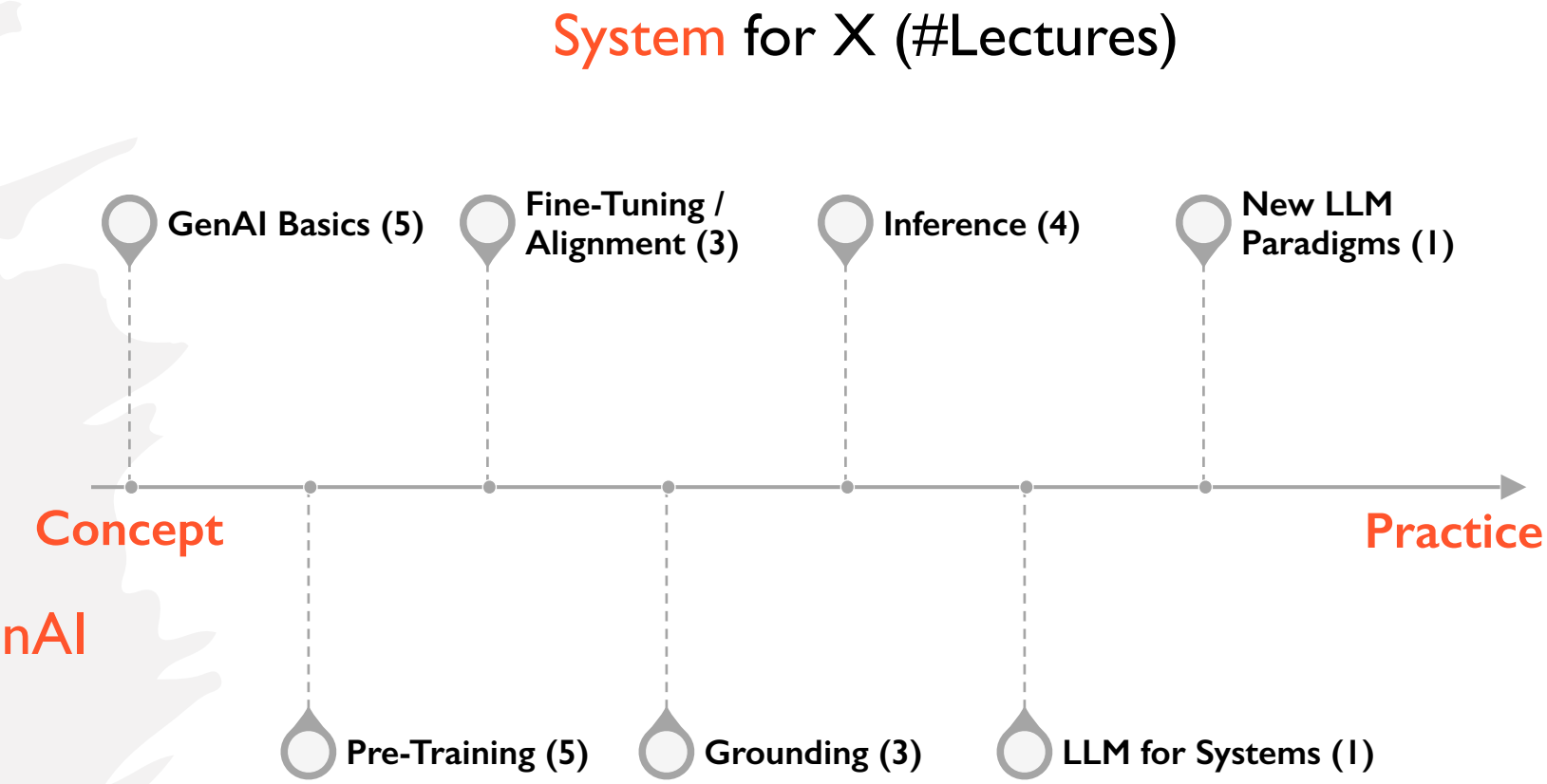
Systems & Networking

What will you learn in this course?

-- How to Support Practical GenAI

What will you learn in this course?

-- How to Support Practical GenAI



What will you learn in this course?

-- How to Support Practical GenAI

Learning Outcomes

After completing this course you should be able to

- Articulate and critique latest GenAI Sys
- Be prepared to perform new research
- Communicate your research findings

What will you learn in this course?

-- How to Support Practical GenAI

Learning Outcomes

After completing this course you should be able to

- Articulate and critique latest GenAI Sys
- Be prepared to perform new research
- Communicate your research findings

This is a Research Seminar Course.
The more active the better.

Agenda

- What will you learn in this course?
  - GenAI, systems, networks, ...
- **What will you do in this course?**
  - Participation, presentation, research project, ...

Course Expectations (Grading)

Participation	15%	Paper reading; Up to two absences

ALL activities will be done in groups except for your own participation

Course Expectations (Grading)

Participation	15%	Paper reading; Up to two absences
Paper Summary	10%	Two summaries

ALL activities will be done in groups except for your own participation

Course Expectations (Grading)

Participation	15%	Paper reading; Up to two absences
Paper Summary	10%	Two summaries
Paper Presentation & Discussion	15%	One Presentation

ALL activities will be done in groups except for your own participation

Course Expectations (Grading)

Participation	15%	Paper reading; Up to two absences
Paper Summary	10%	Two summaries
Paper Presentation & Discussion	15%	One Presentation
Project Proposal	5%	2-page Proposal
Project Presentations	20%	Mid-term + Final Presentation
Project Report	35%	One Final Report

ALL activities will be done in groups except for your own participation

Form Groups ASAP

- Use Piazza to find potential group members
- Group size should be 4-5
  - May allow a few smaller groups if/when students drop out
- Submit at https://forms.gle/h8W3VrkN5ViZkg5G8
  - By Jan 31st the latest, but **right now** is better

Participation

- **Reading: 43 papers/articles across**
  - Primarily from systems venues like SOSP, OSDI, NSDI, EuroSys, and MLSys
  - Some from traditional AI/ML venues but still with systems flavor

Participation

- **Reading: 43 papers/articles across**
  - Primarily from systems venues like SOSP, OSDI, NSDI, EuroSys, and MLSys
  - Some from traditional AI/ML venues but still with systems flavor

Before Each Lecture

- Read 1-2 required reading
- Submit one insightful question per required reading (link)

Feb 14 (Training Infra)	The Llama 3 Herd of Models (Sec 1-4.2, Required) DeepSeek-V3 Technical Report (Sec 3.1-3.2, 3.4, Required) Gemini: A Family of Highly Capable Multimodal Models

Participation

- **Reading: 43 papers/articles across**
  - Primarily from systems venues like SOSP, OSDI, NSDI, EuroSys, and MLSys
  - Some from traditional AI/ML venues but still with systems-y flavor

Before Each Lecture

- Read 1-2 required reading
- Submit one insightful question per required reading (link)

During Lecture

- Please be interactive!
- Up to 2 absences with legitimate reasons

After Lecture

- Feel free to discuss on Piazza

Paper Presentation

- **This is a seminar-style course**
  - Each group presents one lecture covering required papers
  - (Optional) Include a few slides to skim through companion papers

- **The entire class will be dedicated to the assigned paper(s)**
  - Aim for 40-min presentation w/o interruption, but there will be intermittent discussions
  - No more than 55 minutes in total (w/ interruption)

- **Lead the discussion** (details)
  - Go through the required paper in details, along with its strengths and weaknesses

Paper Presentation

- **This is a seminar-style course**
  - Each group presents one lecture covering required papers
  - (Optional) Include a few slides to skim through companion papers

- **The entire class will be dedicated to the assigned paper(s)**
  - Aim for 40-min presentation w/o interruption, but there will be intermittent discussions
  - No more than 55 minutes in total (w/ interruption)

- **Lead the discussion** (details)
  - Go through the required paper in details, along with its strengths and weaknesses

Share slides (in *.ppt format) to cs598-aisys-staff@lists.cs.illinois.edu 24 hrs before the class

Paper Summaries

- **Two summaries per-group, covering required and companion papers**
- **Each summary must follow the template and address the following**
  - What is the problem and why is it important?
  - What is the hypothesis of the work?
  - What is the proposed solution, and what key insight guides their solution?
  - What is one (or more) drawback/limitation, and how will you improve it?
- **Summary (template) should include the gist of class discussion**

Share summary (in *.pdf) to cs598-aisys-staff@lists.cs.illinois.edu within 24 hrs of presentation

Panel Discussion (right after presentation)

- **The Authors (assigned)**
  - 'Companion' group that writes summary and answer questions from 'Reviewers'
- **The Reviewers (assigned)**
  - 'Reviewer' group that writes summary and poses challenging **research** questions
- **Rest of the Class & Presenters**
  - Ask questions directly too

Each group will be assigned to these slots at least once; 1 presentation & 2 summaries (only! 😊)

Summary of Your Role

Presenter Group

- Share slides 24 hrs before class
- Maximum 55-minute presentation
- Respond to questions on Piazza
- Share you final slides after class

Authors (Companion) Group

- Share summary within 24 hrs after class presentation
- Respond to Reviewers' questions

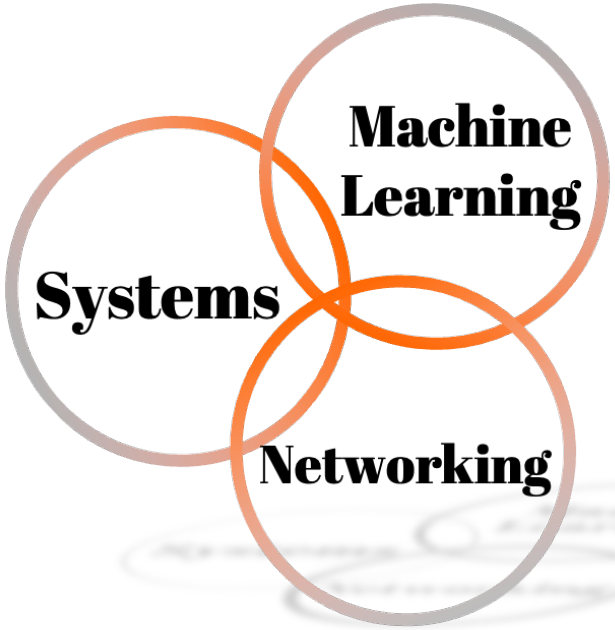
Reviewers Group

- Share summary within 24 hrs after class presentation
- Pose challenging research questions

Audience Group

- Read required papers
- Submit one insightful question per required reading before each lecture

Systems for GenAI Projects



Research-Oriented Course!

- **The final project accounts for 60% of total grades**

- **What can and cannot be a project?**
  - Just surveys are not allowed
  - Measurements of new environments or of existing solutions in new environments are acceptable
  - Proposing new research problems or addressing existing ones is highly preferred!

Research-Oriented Course!

- **The final project accounts for 60% of total grades**
- **What can and cannot be a project?**
  - Just surveys are not allowed
  - Measurements of new environments or of existing solutions in new environments are acceptable
  - Proposing new research problems or addressing existing ones is highly preferred!
- **An ideal project should answer the questions you asked during paper reviews and points you cared about for presentations**

How to Approach it?

1. **Find a problem and motivate why this is worth solving**
2. **Quickly survey background and related work**
   - Might require you to go back to the first step
3. **Form/update your hypothesis**
4. **Test your hypothesis**
   - Go back to 3 until you are happy
5. **Present your findings on poster and in writing**
   - Discuss known limitations

Draft Proposal (Feb 26)

- **Two pages (template), plus references, must include**
  - ***Problem:*** What is the problem?
  - ***Motivation****:* Why is it important?
  - ***Solution overview:*** Any initial thoughts?
  - ***Evaluation plan:*** How would you evaluate it?

- **Submit as a group, including your group information** *(form a group ASAP)*

- **Approved by the instructor and agreed upon by you**
  - Forms the basis of expectation

Mid-Semester Checkpoint (Apr 2, 4)

- **In-class short presentation over two lectures**
  - This is to make sure you are receiving feedback and making progress

- **Must include**
  - What is the problem?
  - Why is it important? (Motivating experiments)
  - What are the most related work?
  - What's your hypothesis so far?
  - How are/will you evaluate it?

Poster & Paper (Apr 30, May 2, 7)

- **Final Presentation/Poster (Apr 30, May 2, 7)**

- **Final report (7-8 pages excluding references by May 15):**
  - Should be written like the papers you've read
  - As if you'd submit it to a workshop with ~3 more months of work or to a conference after ~6 more months of work
  - How to Write a Great Research Paper by Simon Peyton Jones

Project Ideas

- **Some potential project ideas (will share more on Piazza soon)**
  - Multi-tier memory, including CXL and RDMA, for long-context LLMs
  - Energy-efficient GenAI serving on end-user devices
  - Systems support for reinforcement learning from human feedback
  - Making chatbots real by allowing them to cut in
  - Repurposing spatiotemporal redundancy in video generation
  - …
- **Any of your crazy ideas!**

Any Questions?

Date	Milestone	Details
01/31/25	Form Group	Find 4-5 like-minded teammates
02/26/25	Submit Proposal	Send your proposal by email to receive feedback either via email or in-person or both
04/02/25 04/04/25	Mid-Semester Presentations	Define and motivate a problem, overview, related work, and form initial hypothesis/idea
04/30/25 05/02/25 05/07/25	In-Class or Poster Presentations	Present your research findings
05/15/25	Research paper	Submit a report like the papers you read

NO extensions! Delayed submission will receive NO CREDIT

Thank you!

Audience Group

- Read required papers
- Submit short questions

Presenter Group

- Present one lecture
- 55-minute presentation

Form Groups of 4-5; submit by *Jan 31st*

Reviewers Group

- Write one paper summary
- Participate in panel discuss.

Authors (Companion) Group

- Write one paper summary
- Participate in panel discuss.