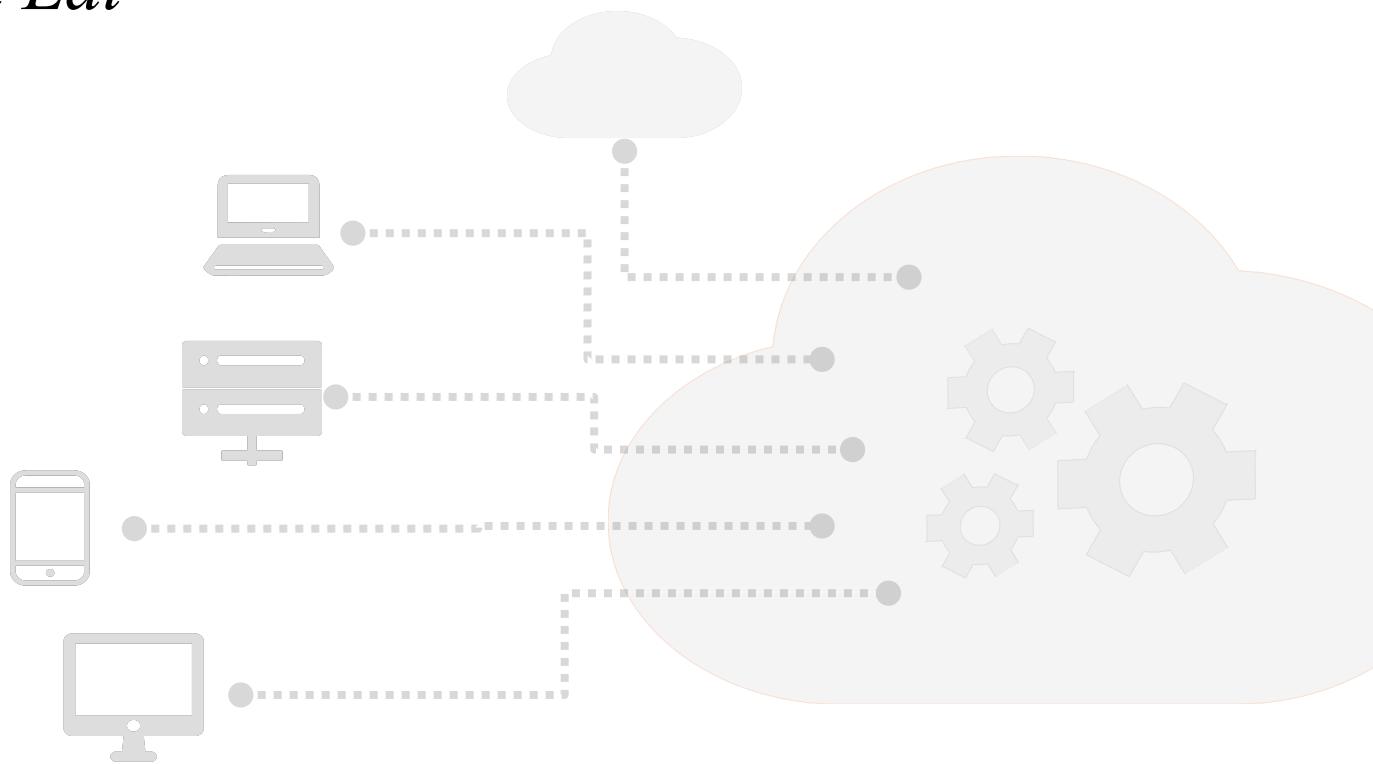


I ILLINOIS

CS 598: Systems for GenAI

Instructor: Fan Lai

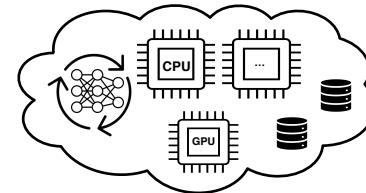


About Fan

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

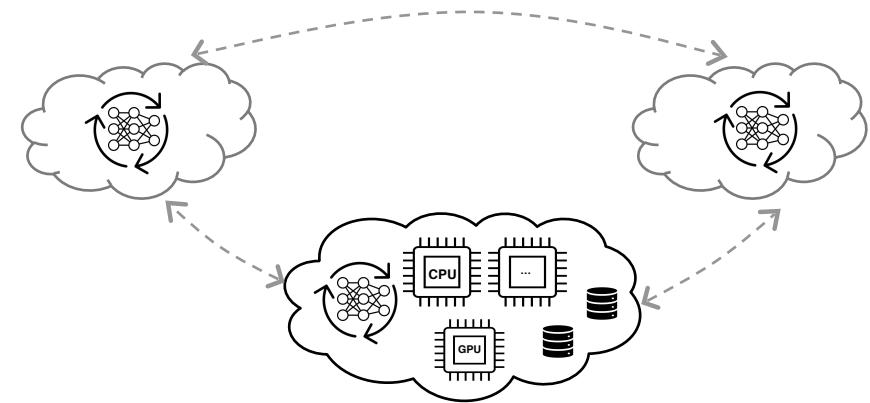
About Fan

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



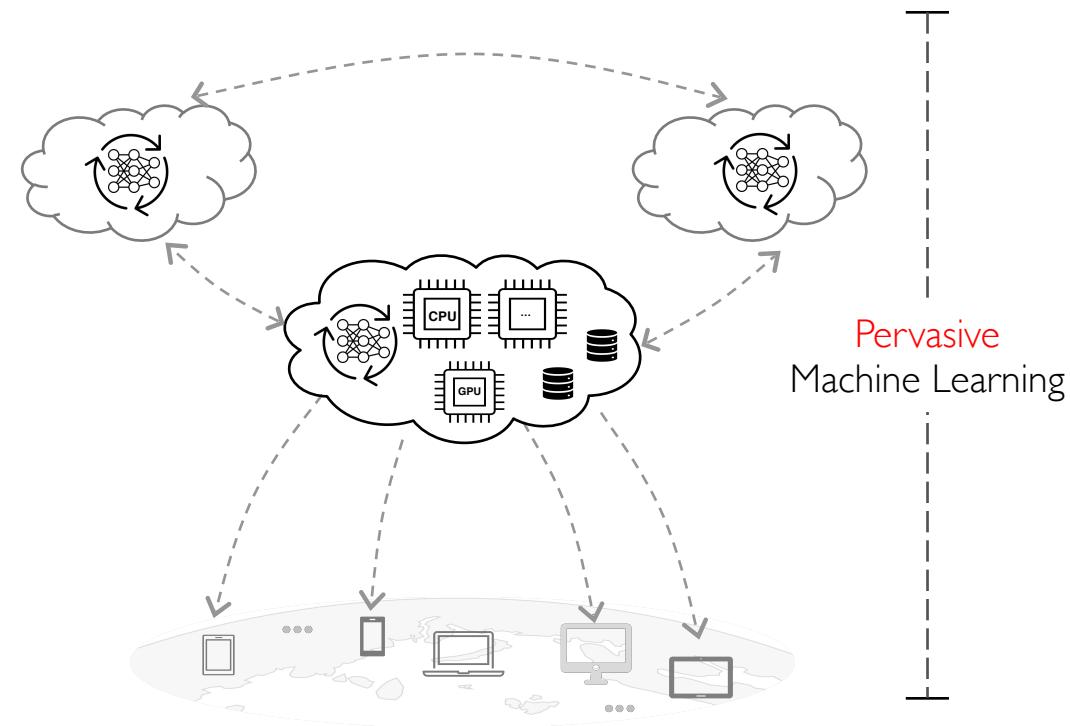
About Fan

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



About Fan

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



About Fan

- **Assistant Professor in CS**

- PhD'23: "Minimalist Systems for Pervasive Machine Learning"

- **Industry:**

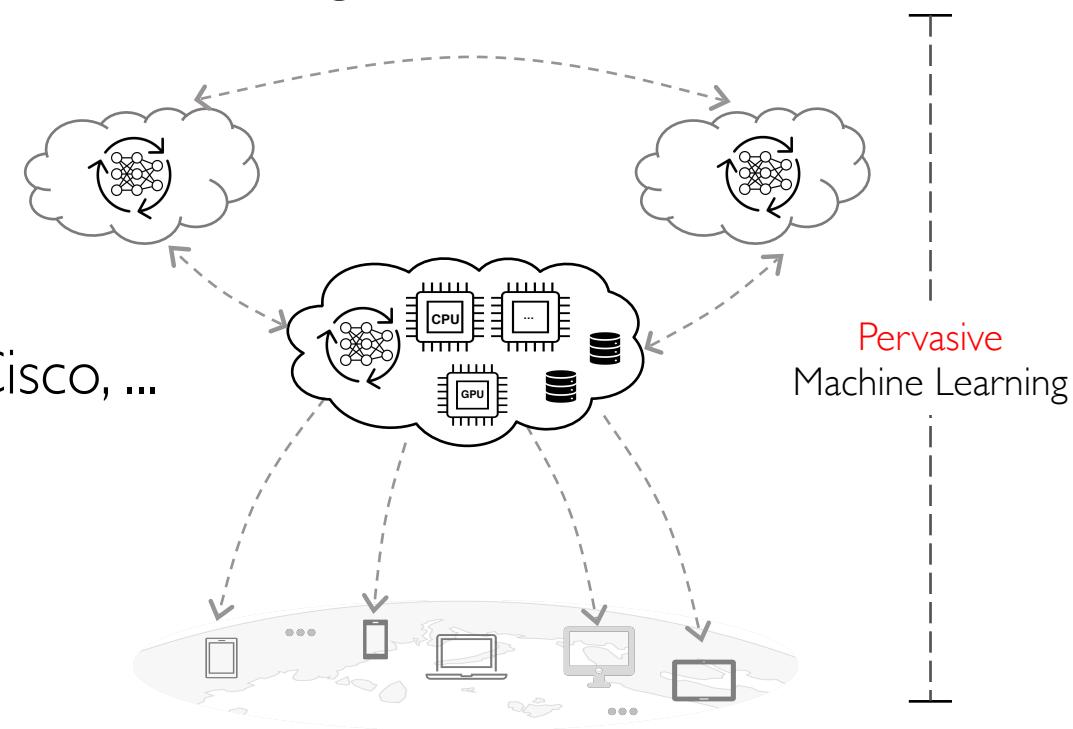
- Visiting Faculty@Google (2023-present)

- Meta AI (2022-2023)

- MLSys adoptions at Google, Meta, LinkedIn, Cisco, ...

- **Office hours**

- SC 3128, M 11 AM – 12 PM



About Chengsong Zhang (TA)

- **2rd-year Master's student**
 - 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](#), W 11:00 AM - 12:00 PM



Status and Prerequisites

- As of today: ~50 registered
 - If you are not planning to take the class, please drop ASAP

Status and Prerequisites

- **As of today: ~50 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 :)

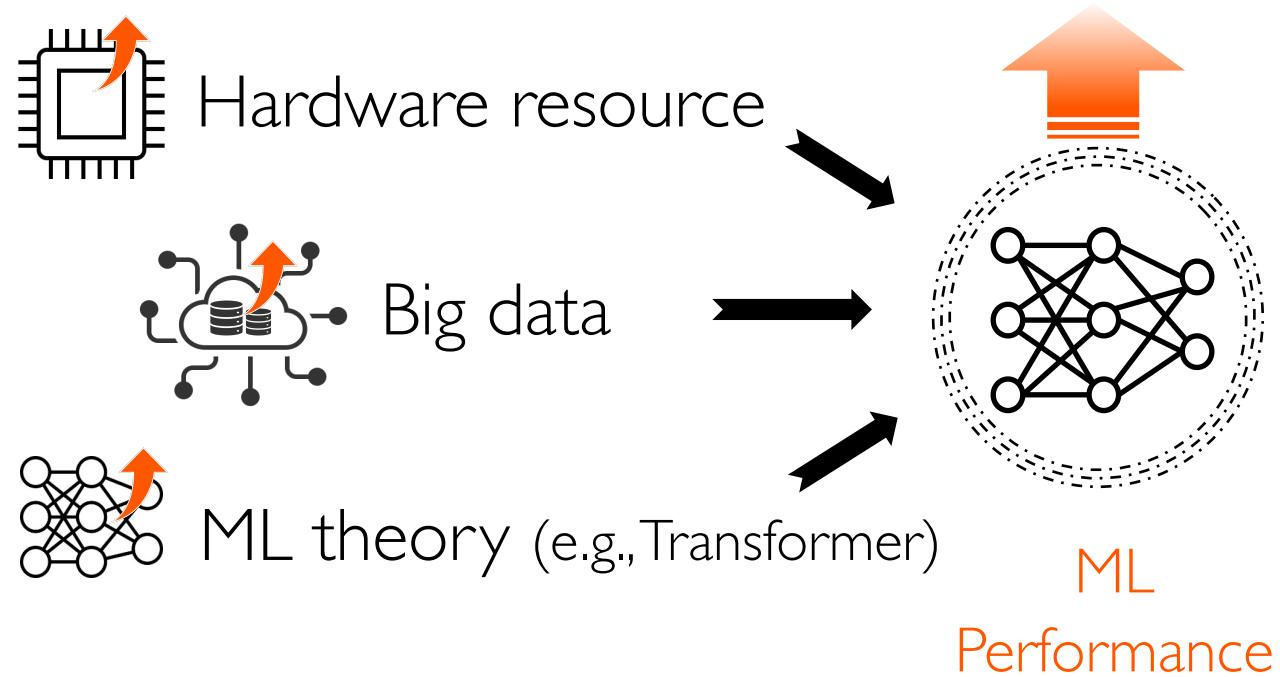
Agenda

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

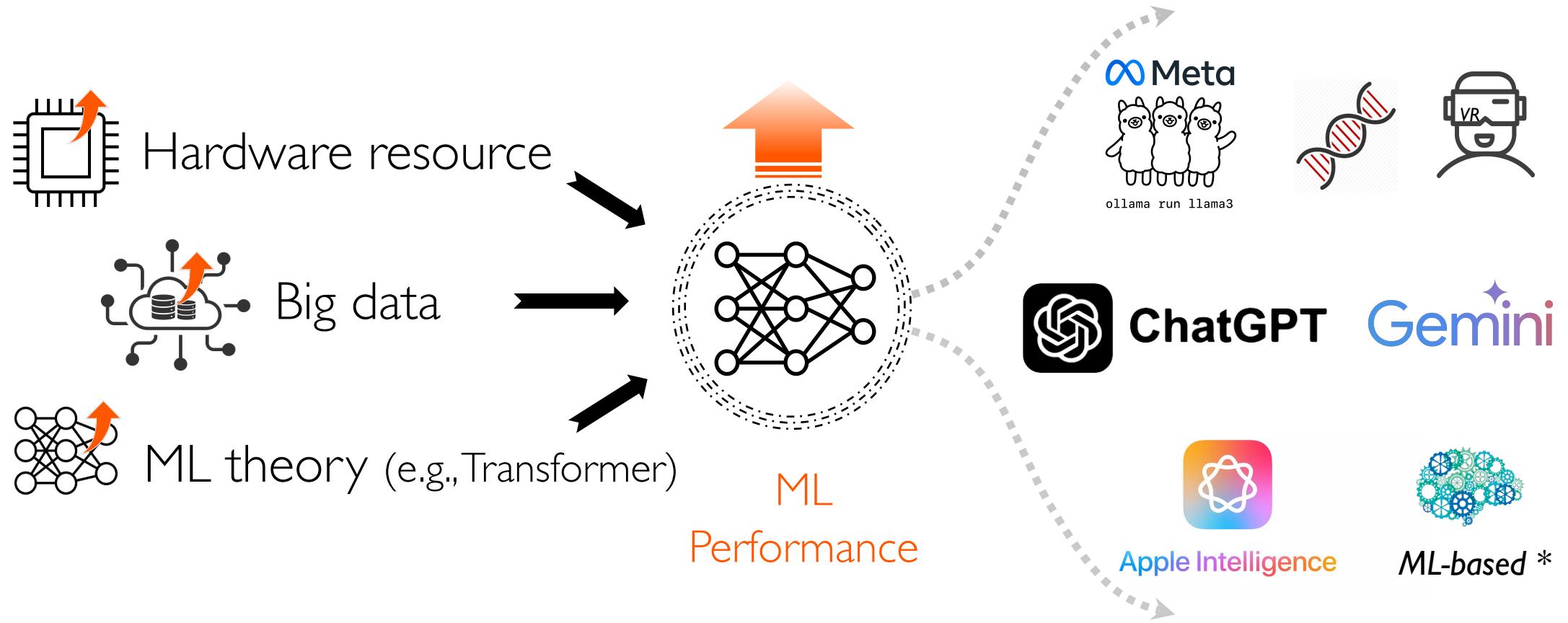


What Do We Talk About When We Talk About “**Systems** for GenAI”

Cloud Enabled Last Leap in Machine Learning (ML)

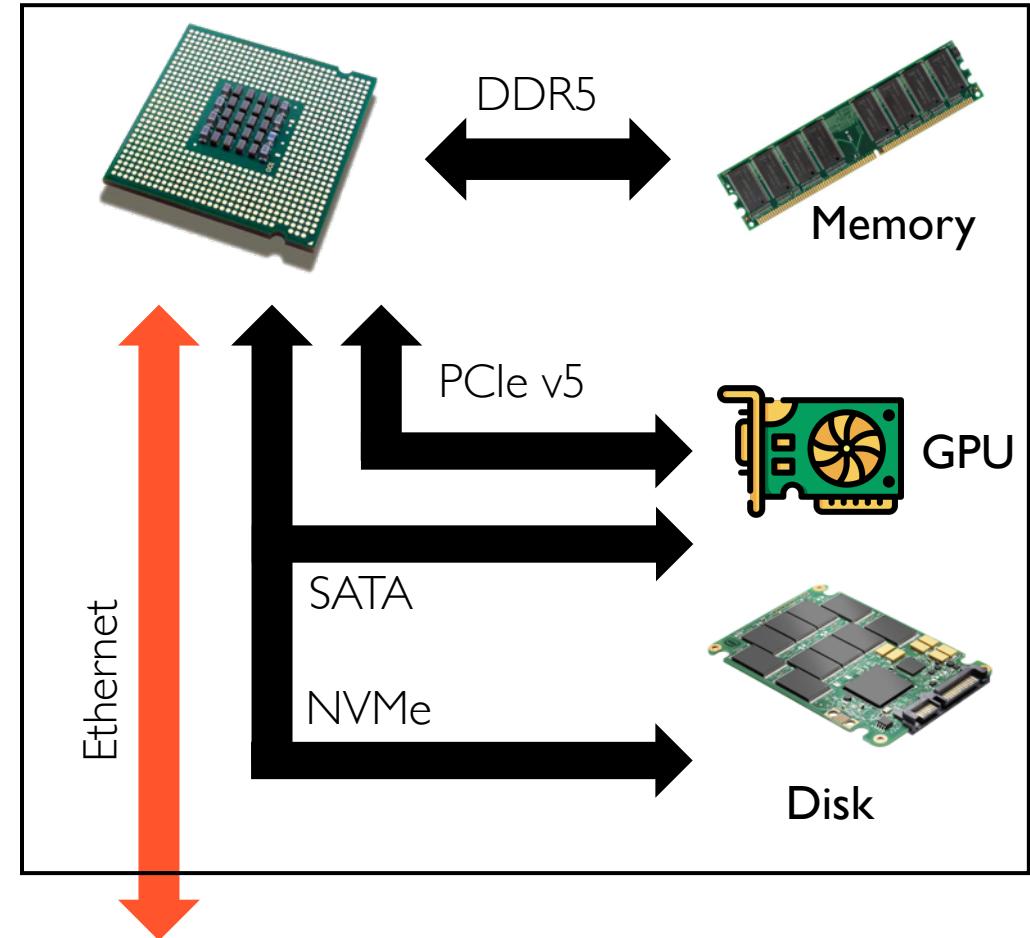


Cloud Enabled Last Leap in Machine Learning (ML)



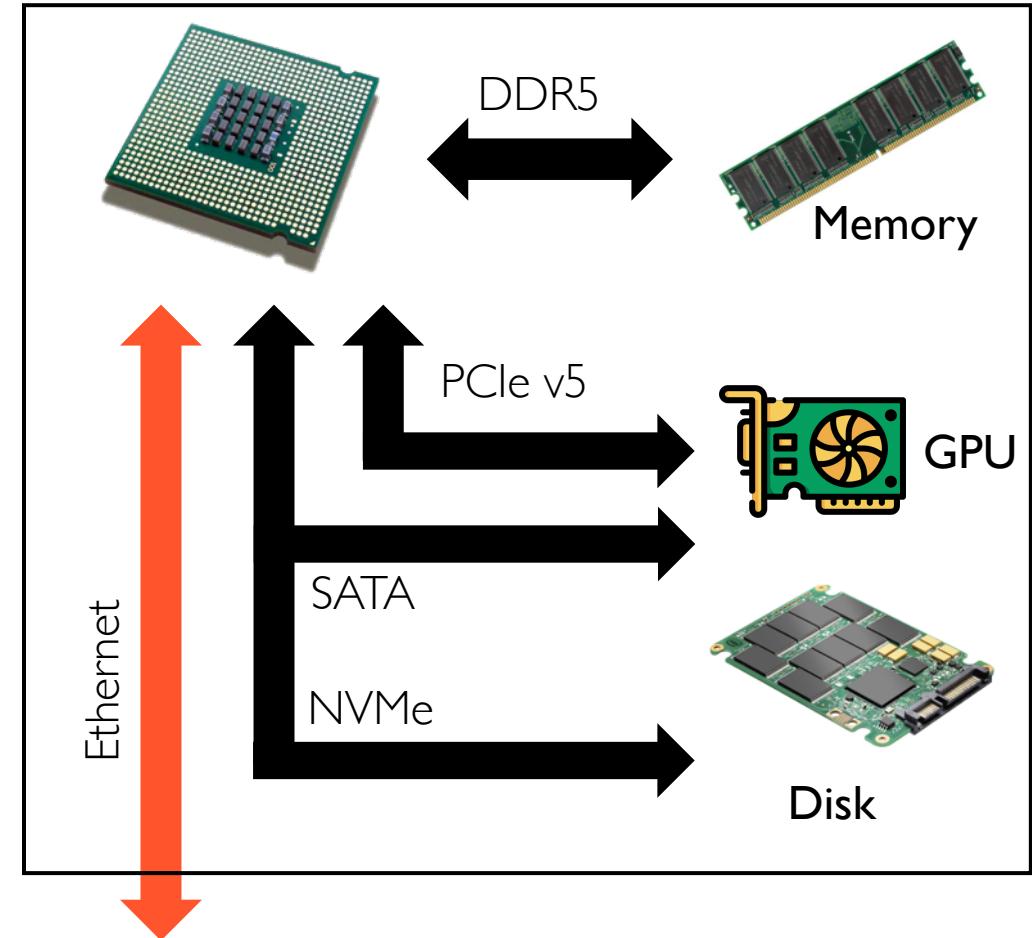
What's in a (Simplified) AI Server?

- Interconnected compute and storage resources



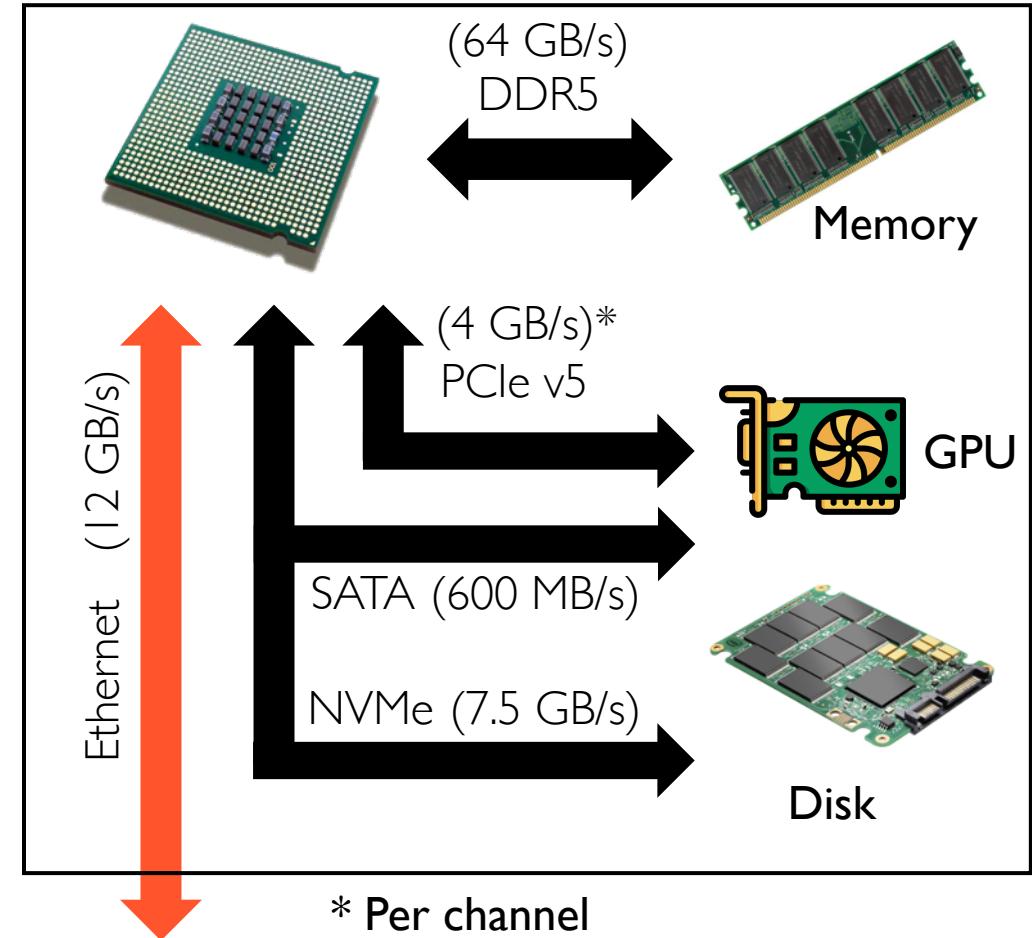
What's in a (Simplified) AI Server?

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



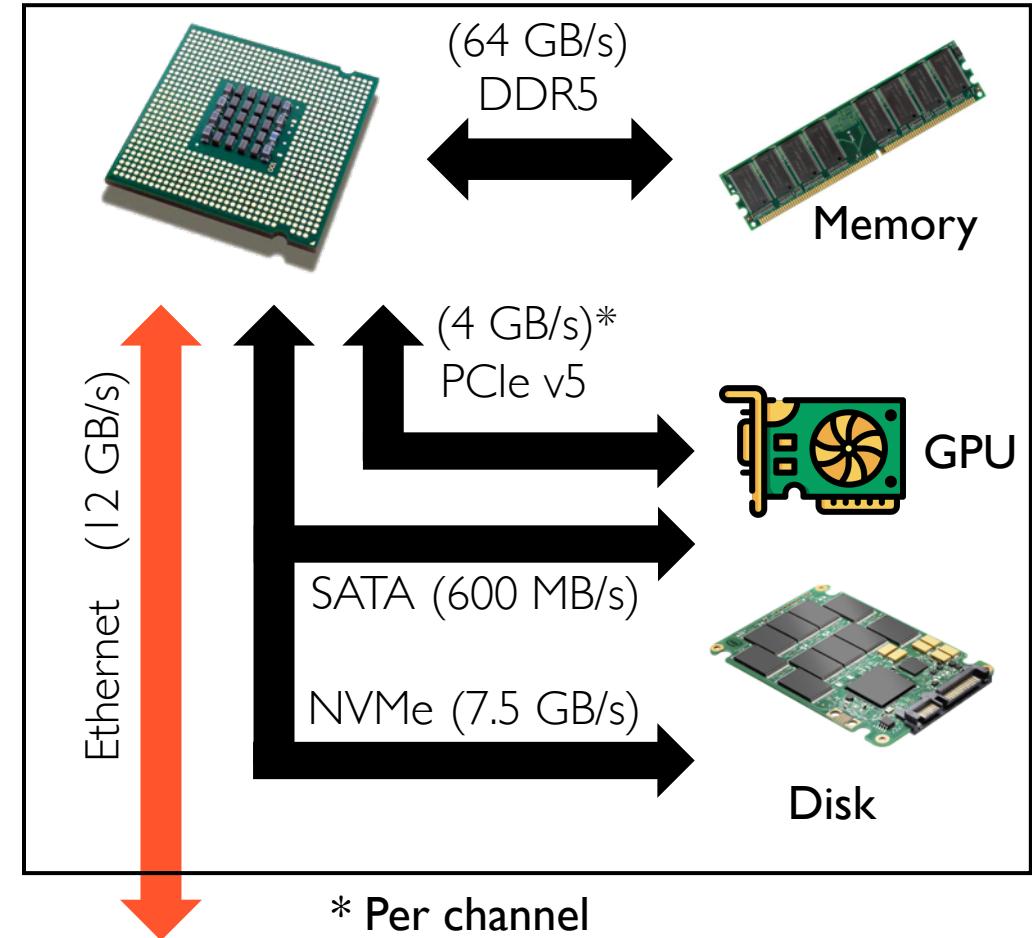
What's in a (Simplified) AI Server?

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

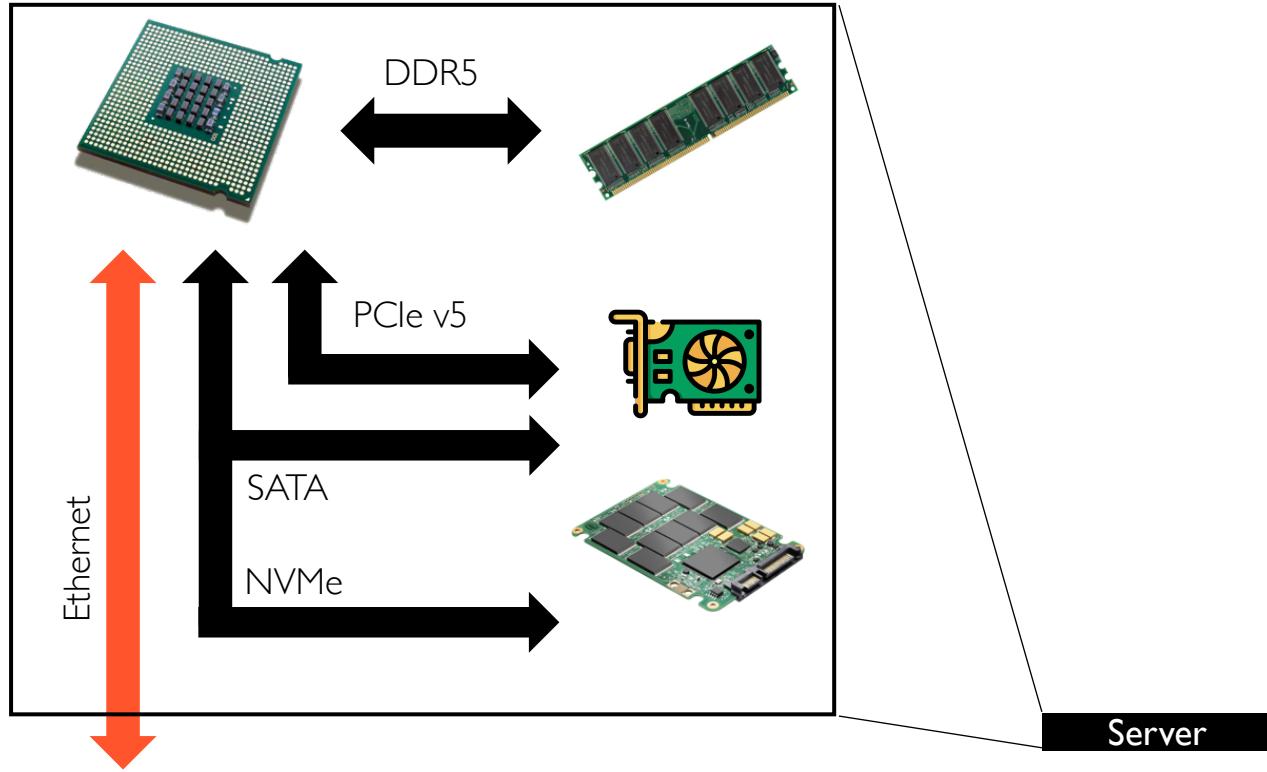


What's in a (Simplified) AI Server?

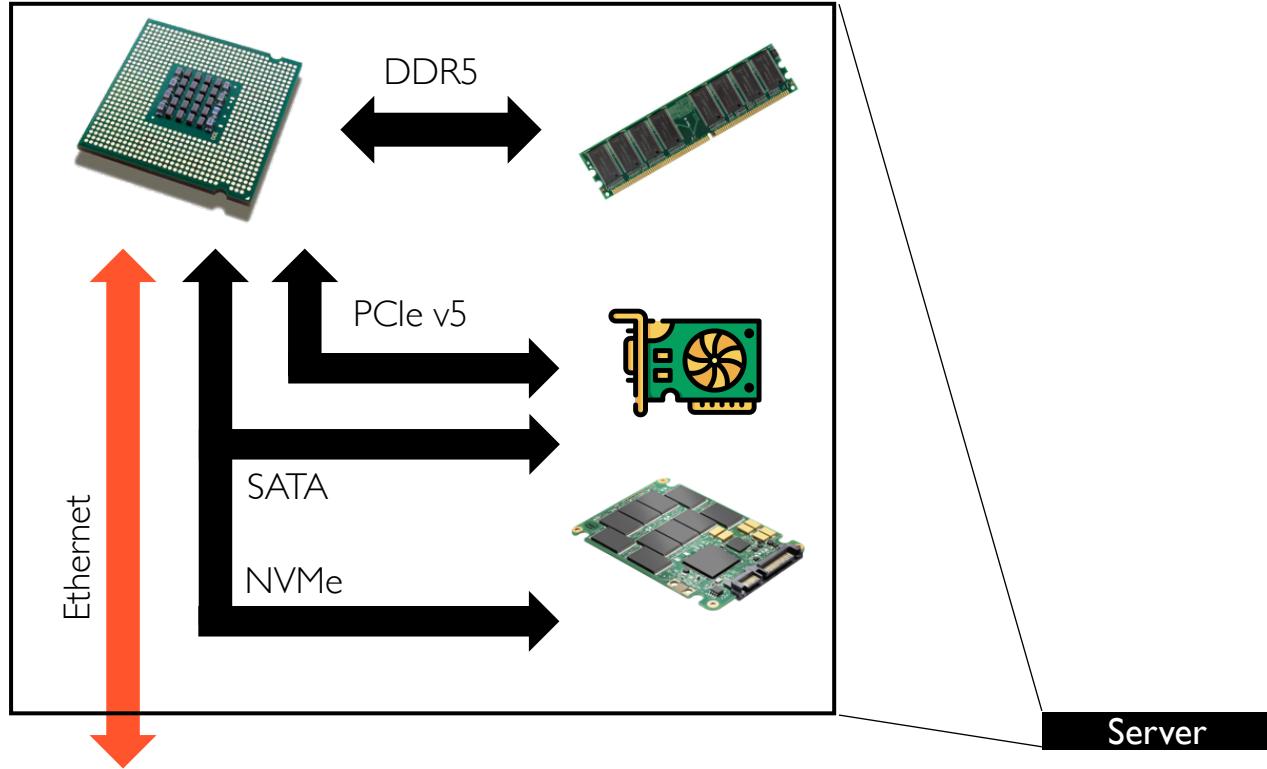
- **Interconnected compute and storage resources**
 - Different bandwidth and latency constraints
- **Simplified diagram**
 - Doesn't include faster networks such as RDMA, dedicated GPU interconnects such as NVLink



Cloud Datacenter: Scale-Out Servers

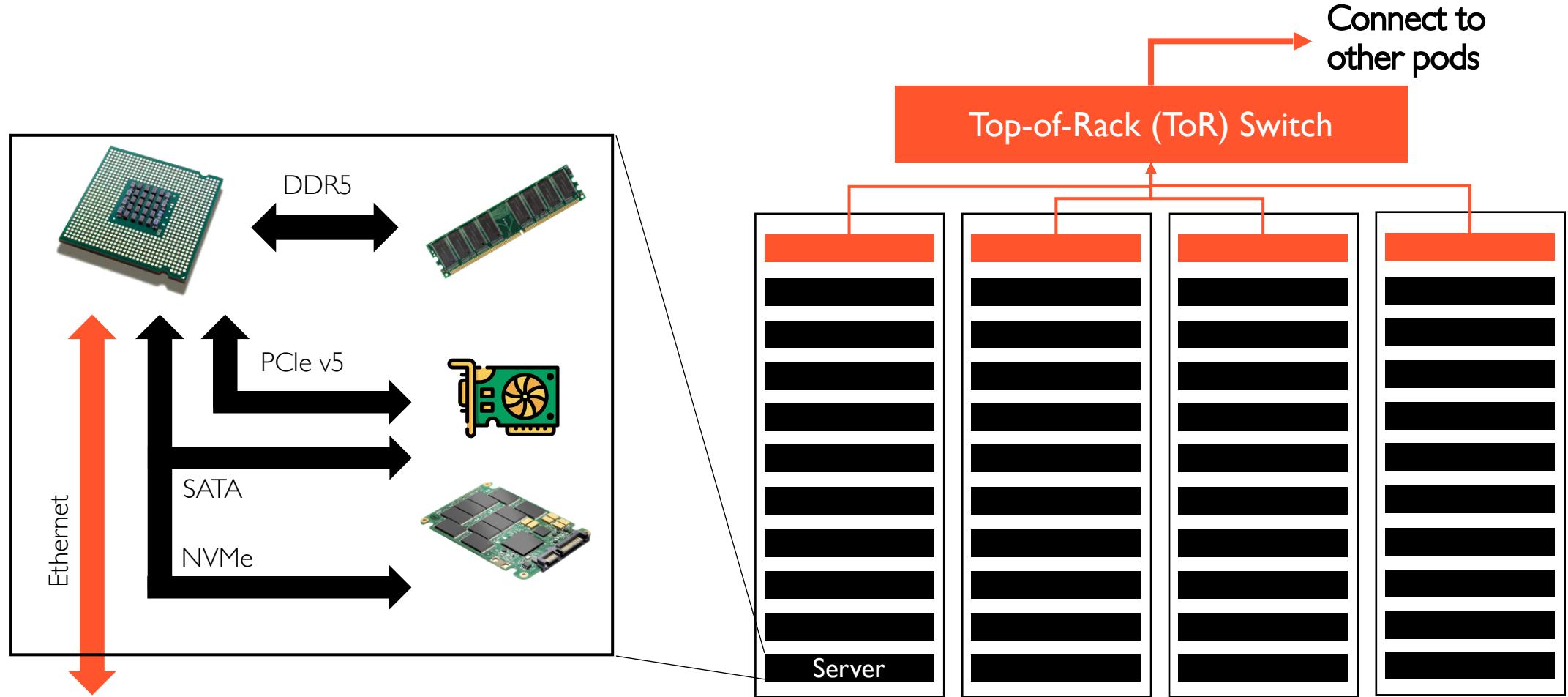


Cloud Datacenter: Scale-Out Servers



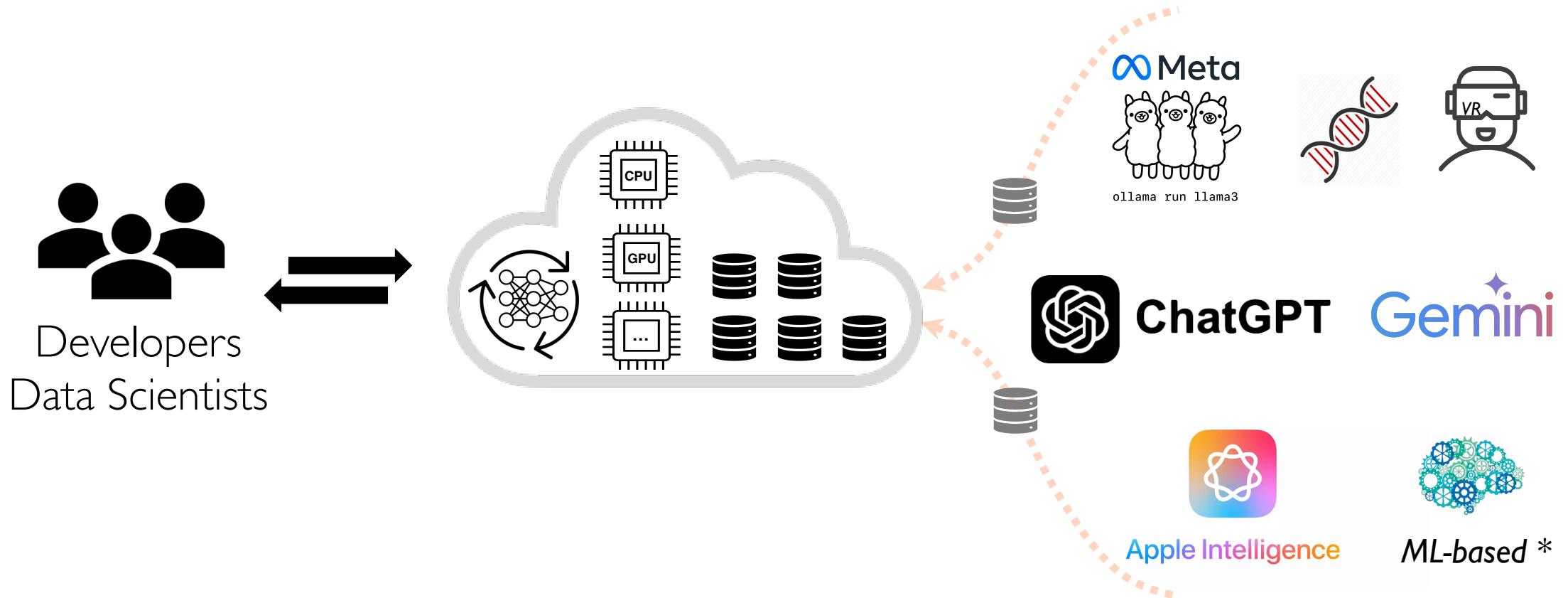
Meta's AI datacenter houses 16K of its 24K GPUs for training the latest Llama 3 model

Cloud Datacenter: Scale-Out Servers

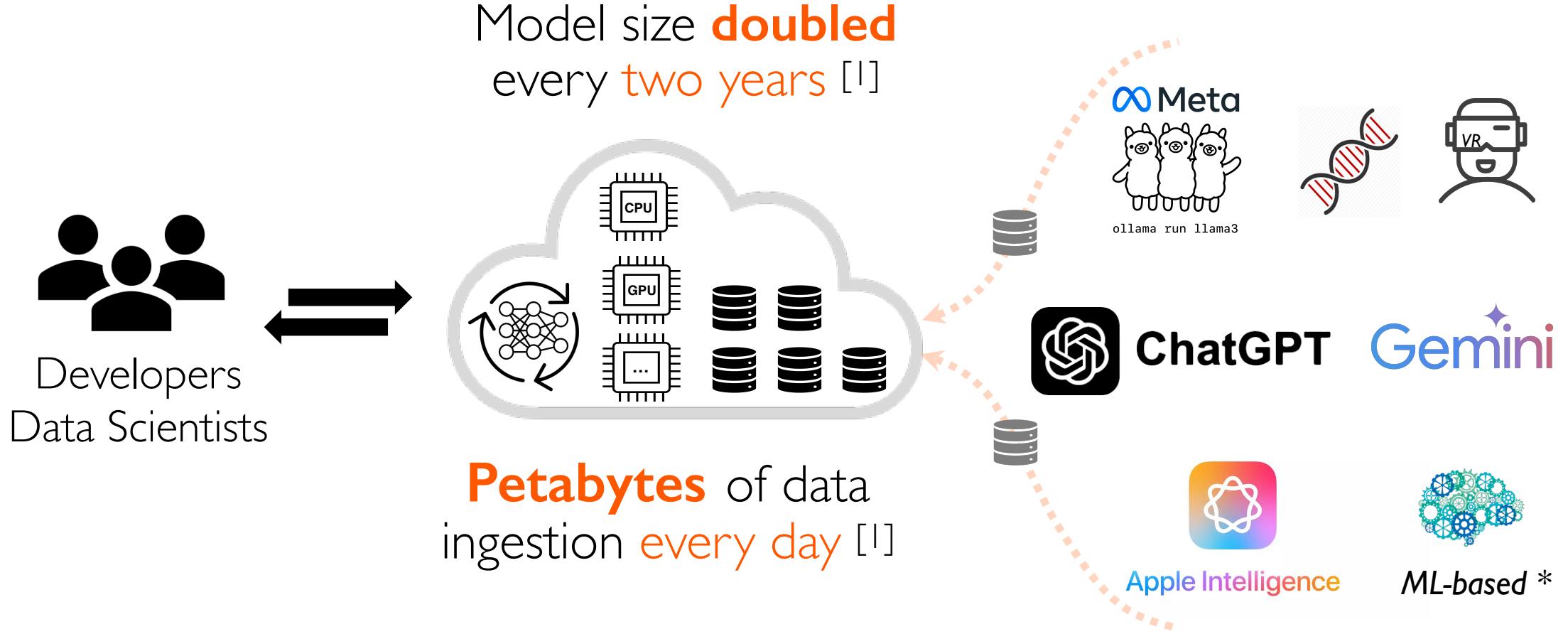


Meta's AI datacenter houses 16K of its 24K GPUs for training the latest Llama 3 model

Cloud Faces Skyrocketing GenAI Workloads



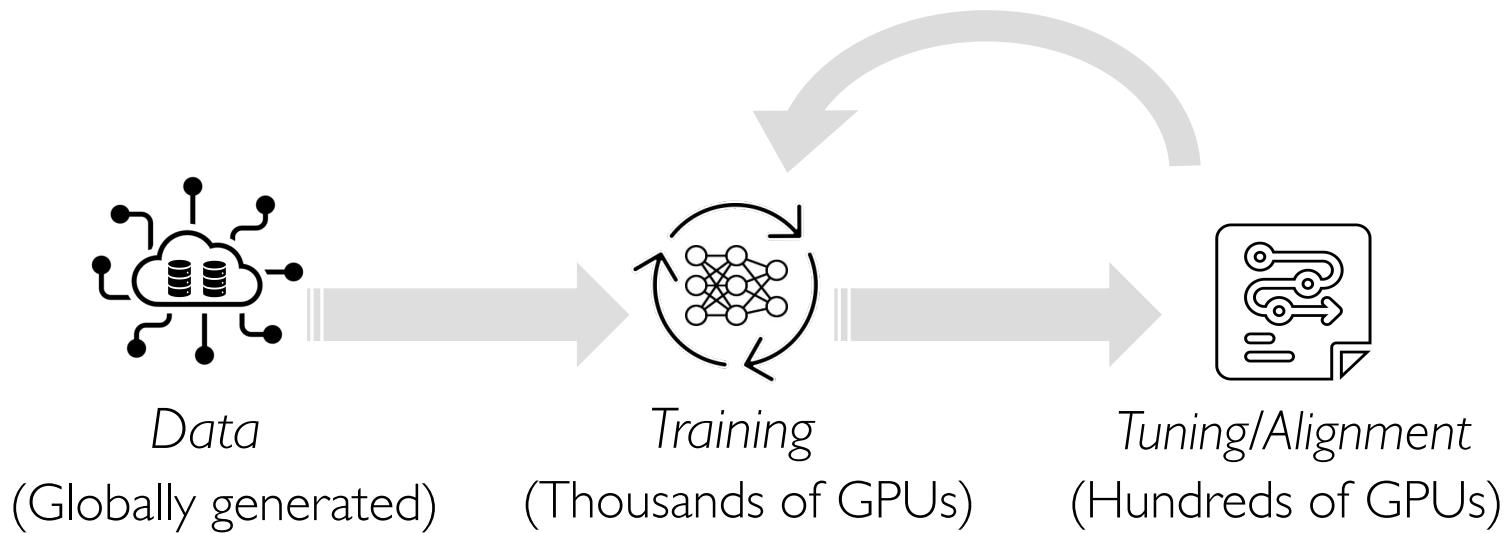
Cloud Faces Skyrocketing GenAI Workloads



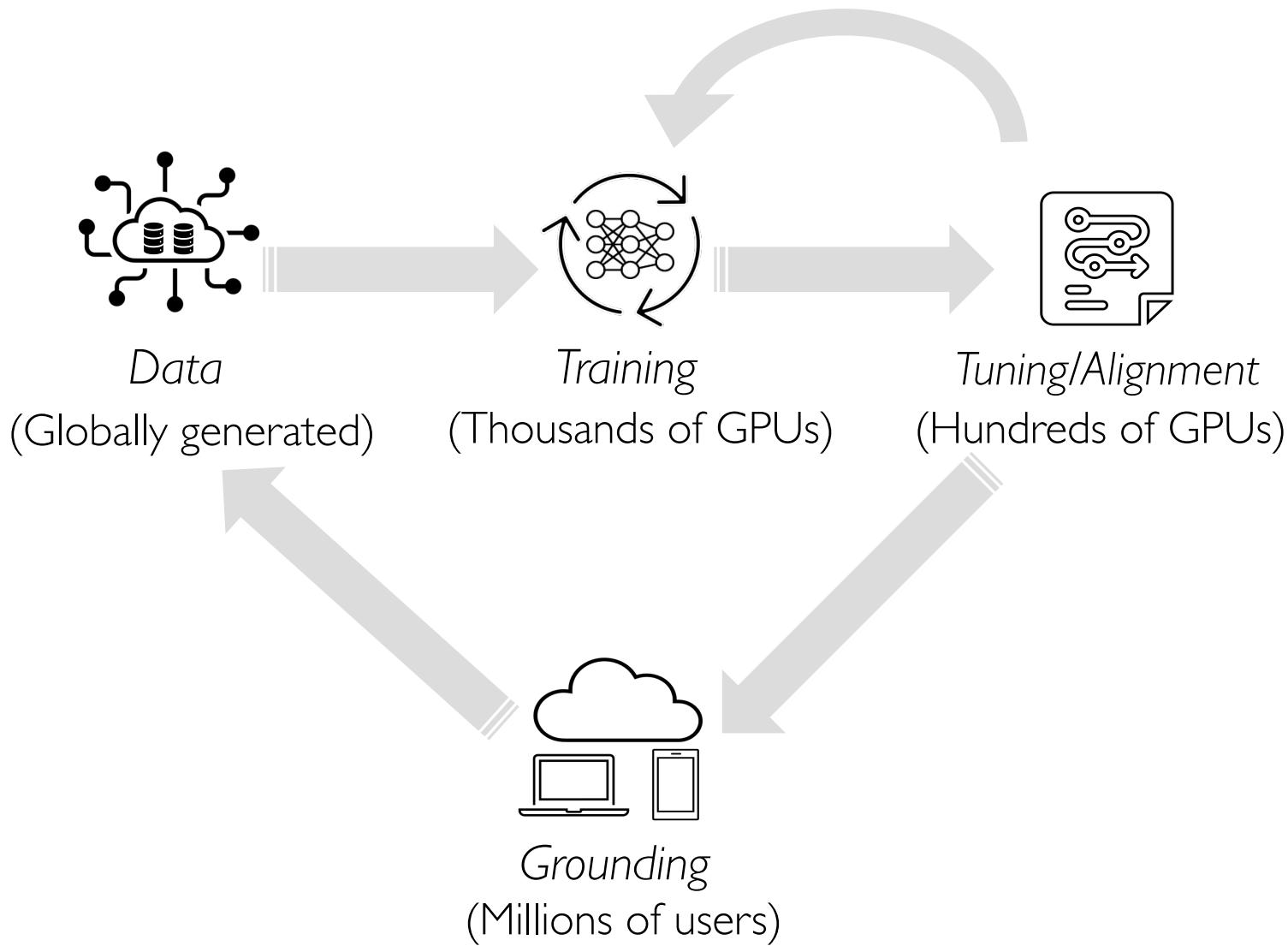
[!] "Sustainable AI: Environment Implications, Challenges and Opportunities", Meta, MLSys'22

GenAI Lifecycle Demands Great Systems Support

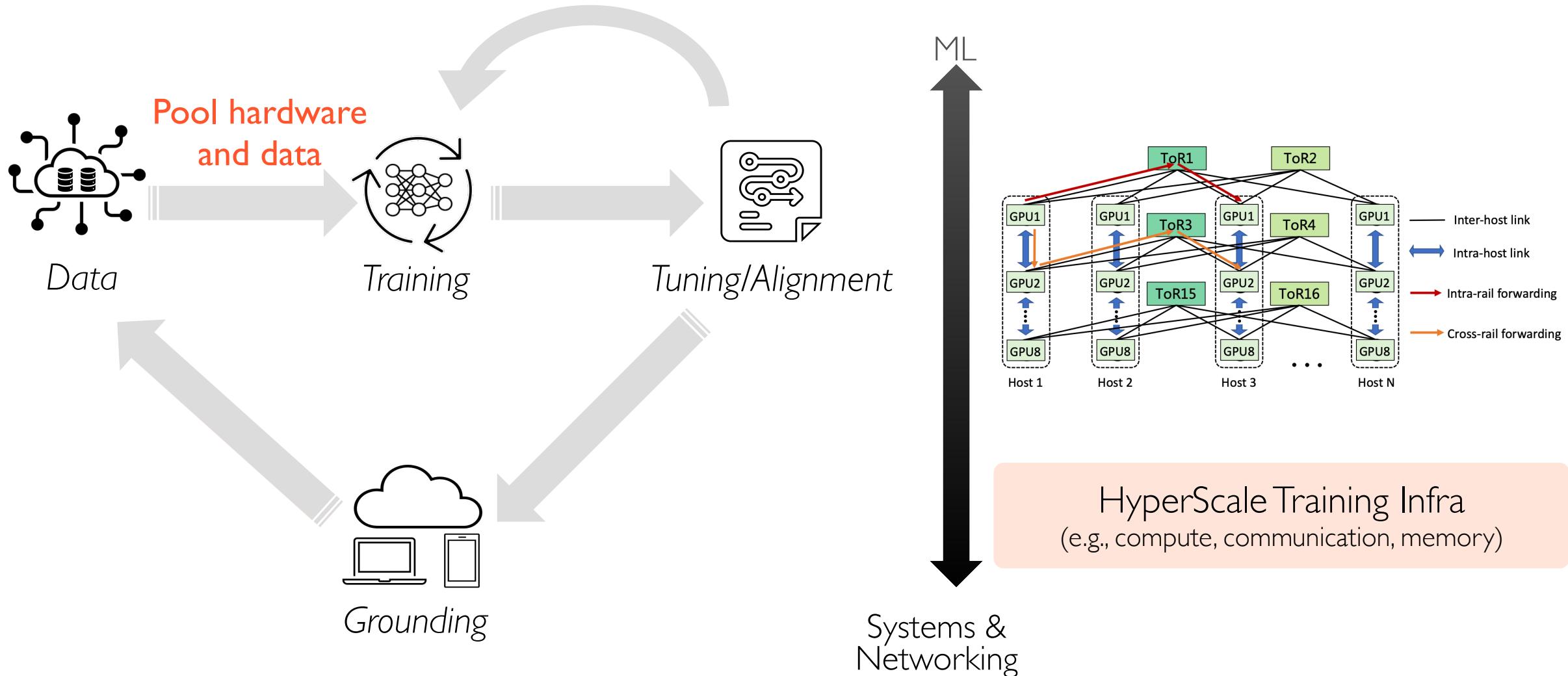
GenAI Lifecycle Demands Great Systems Support



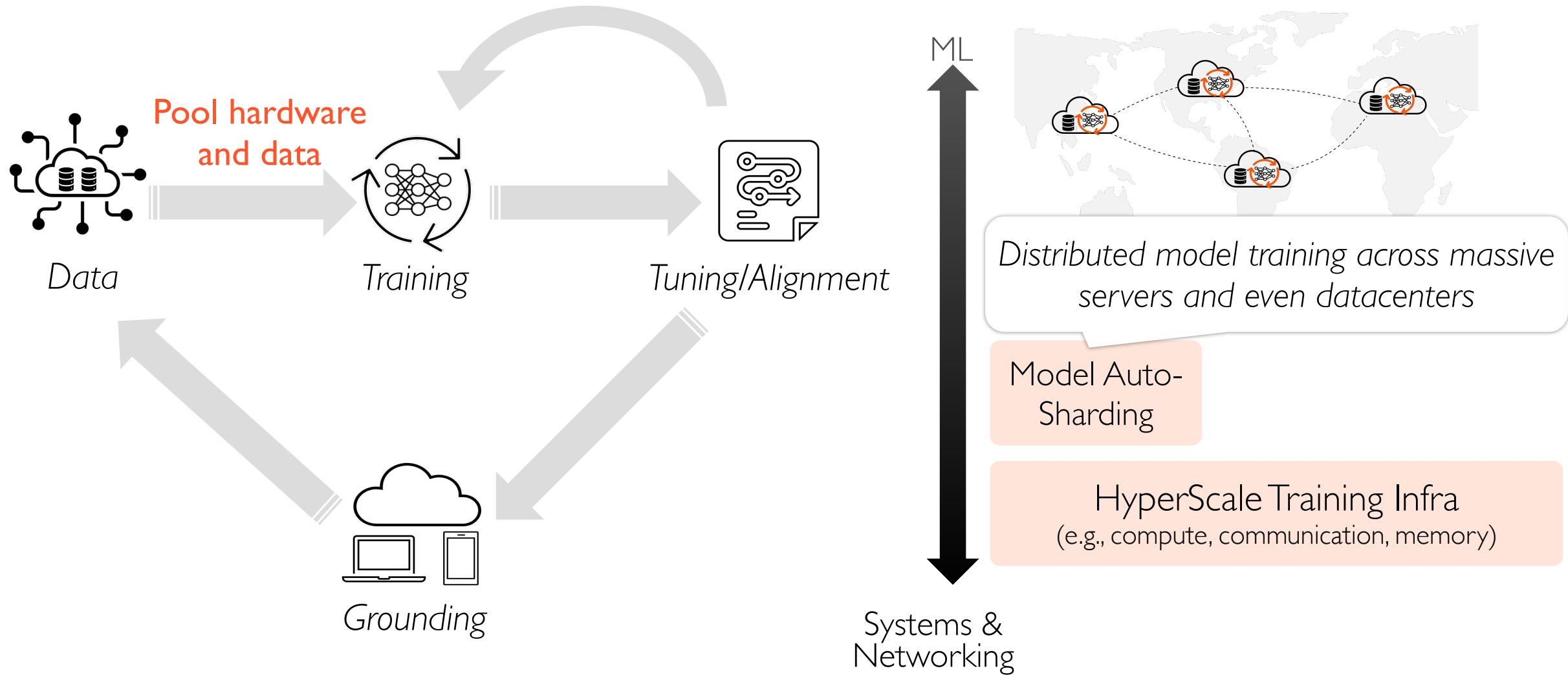
GenAI Lifecycle Demands Great Systems Support



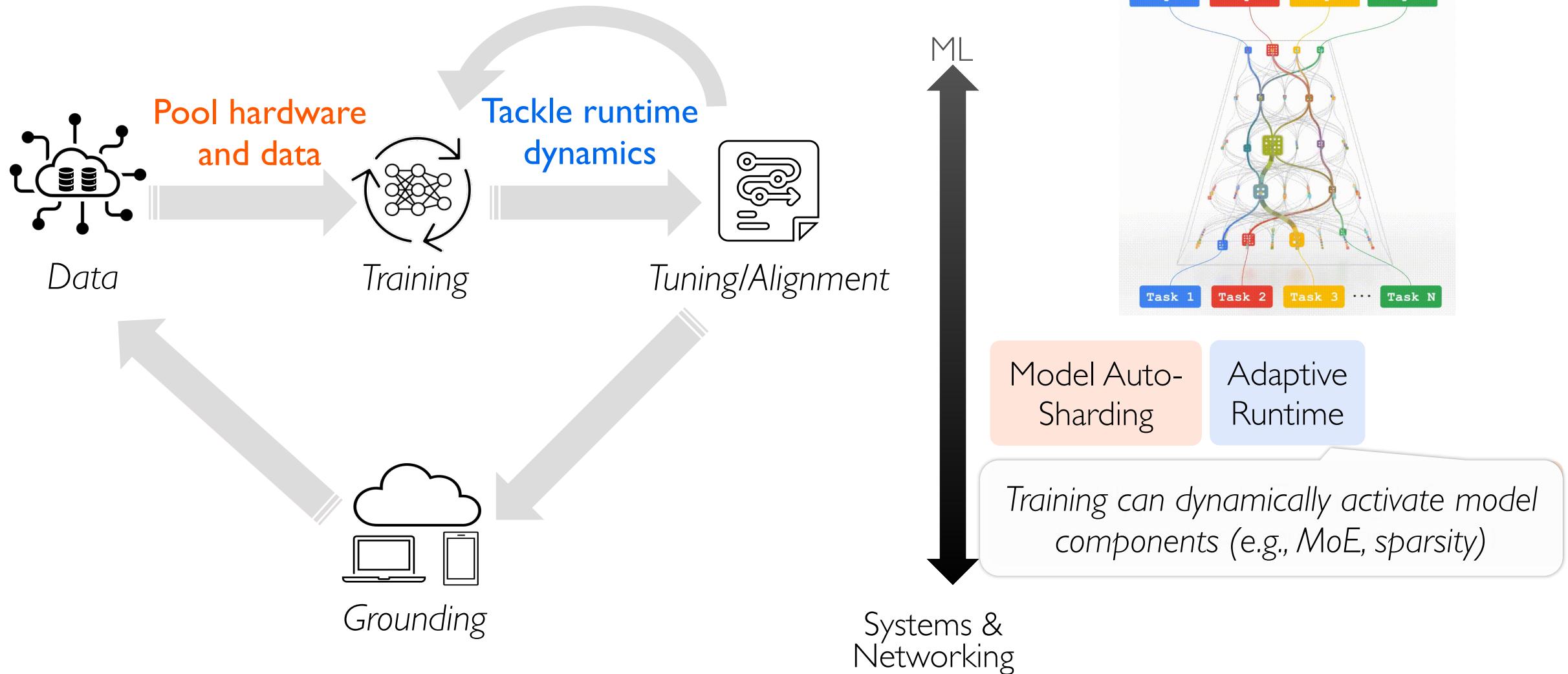
GenAI Lifecycle Demands Great Systems Support



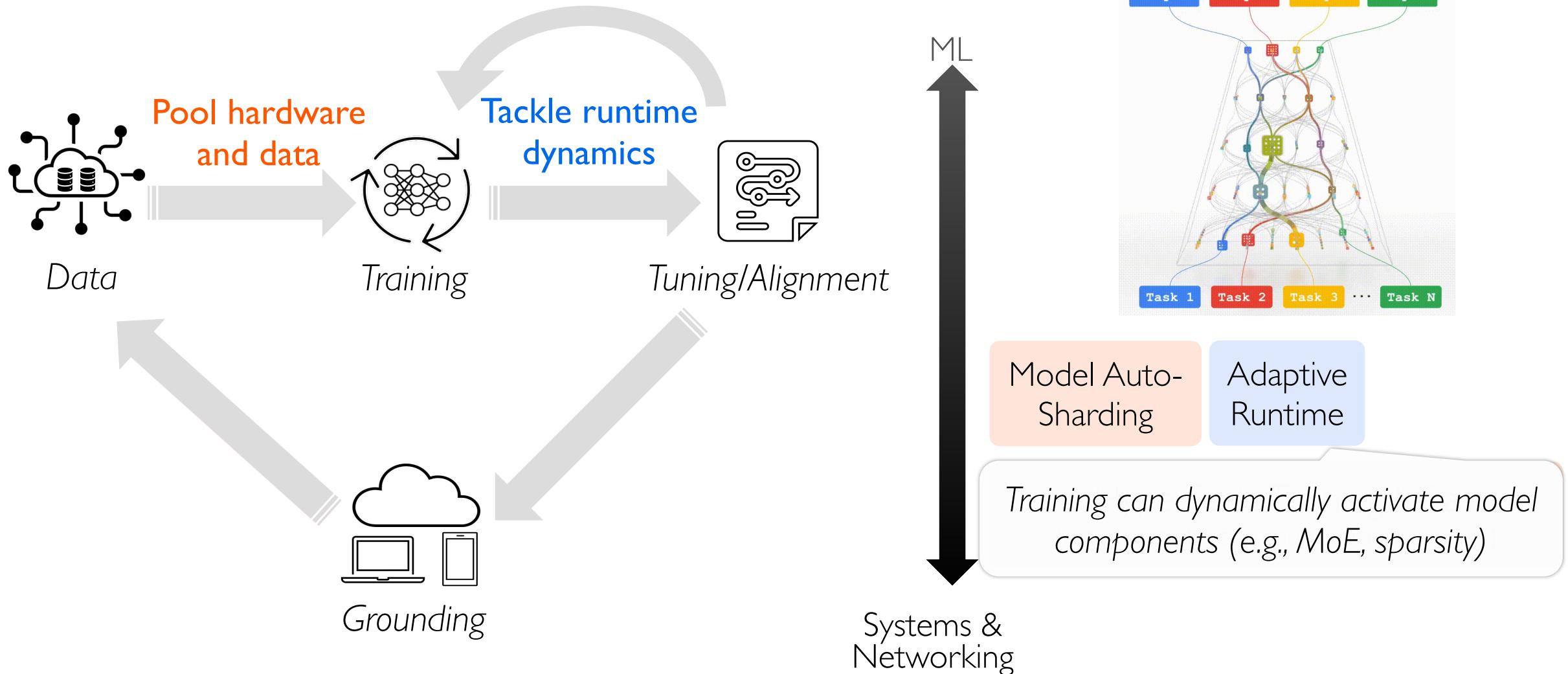
GenAI Lifecycle Demands Great Systems Support



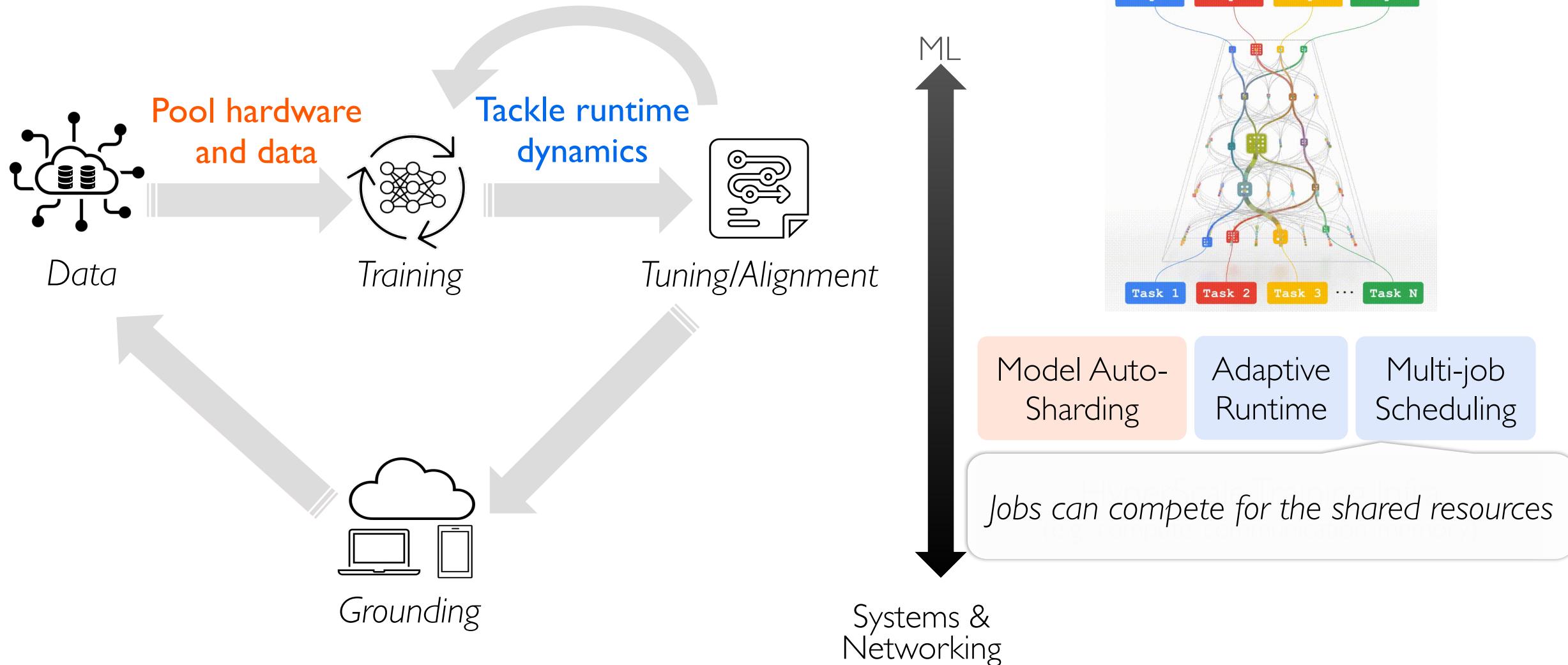
GenAI Lifecycle Demands Great Systems Support



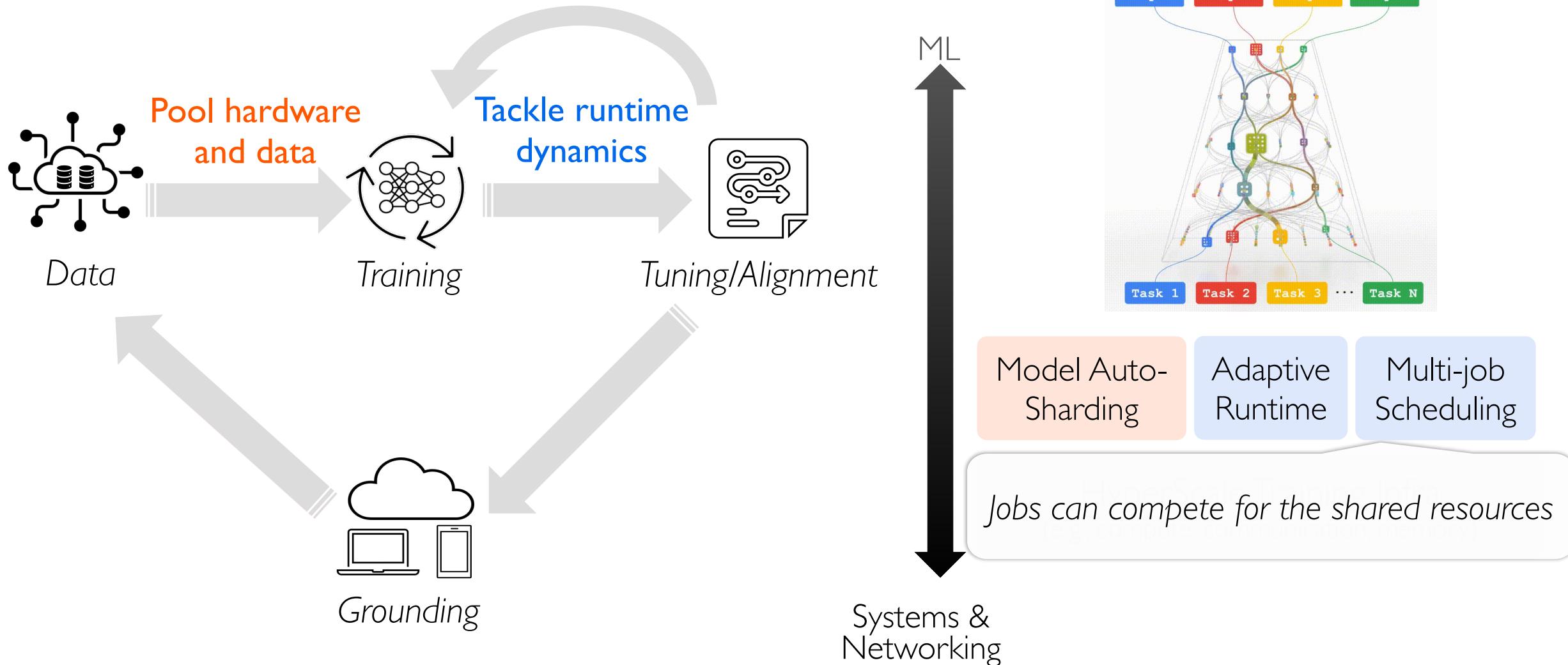
GenAI Lifecycle Demands Great Systems Support



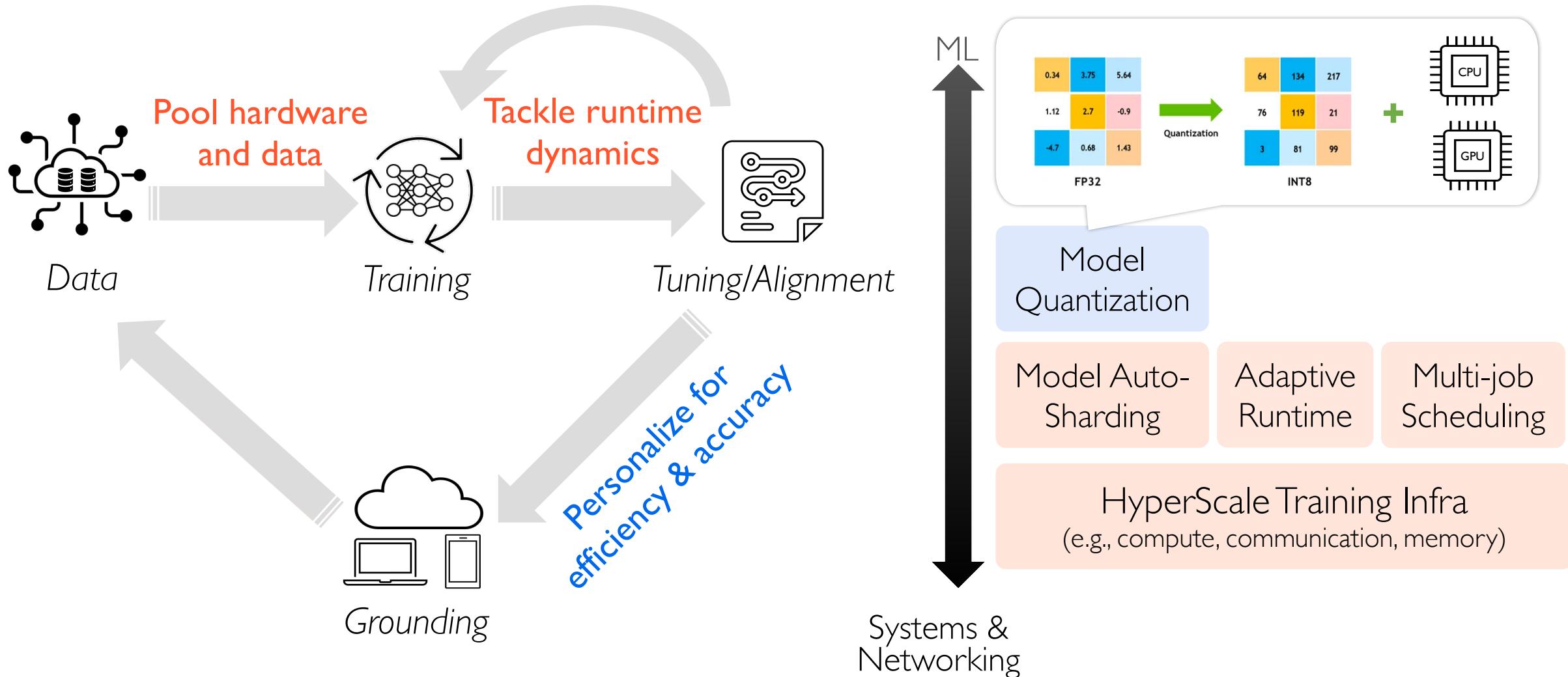
GenAI Lifecycle Demands Great Systems Support



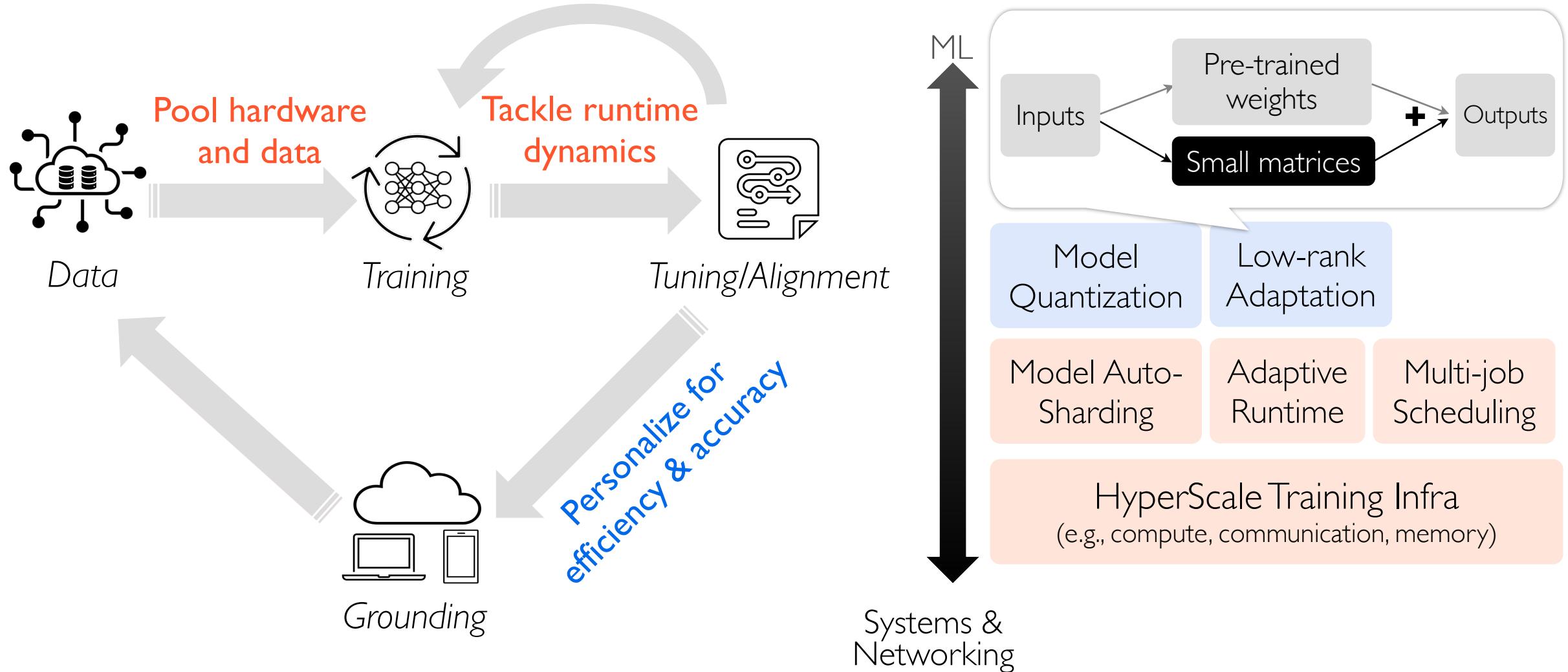
GenAI Lifecycle Demands Great Systems Support



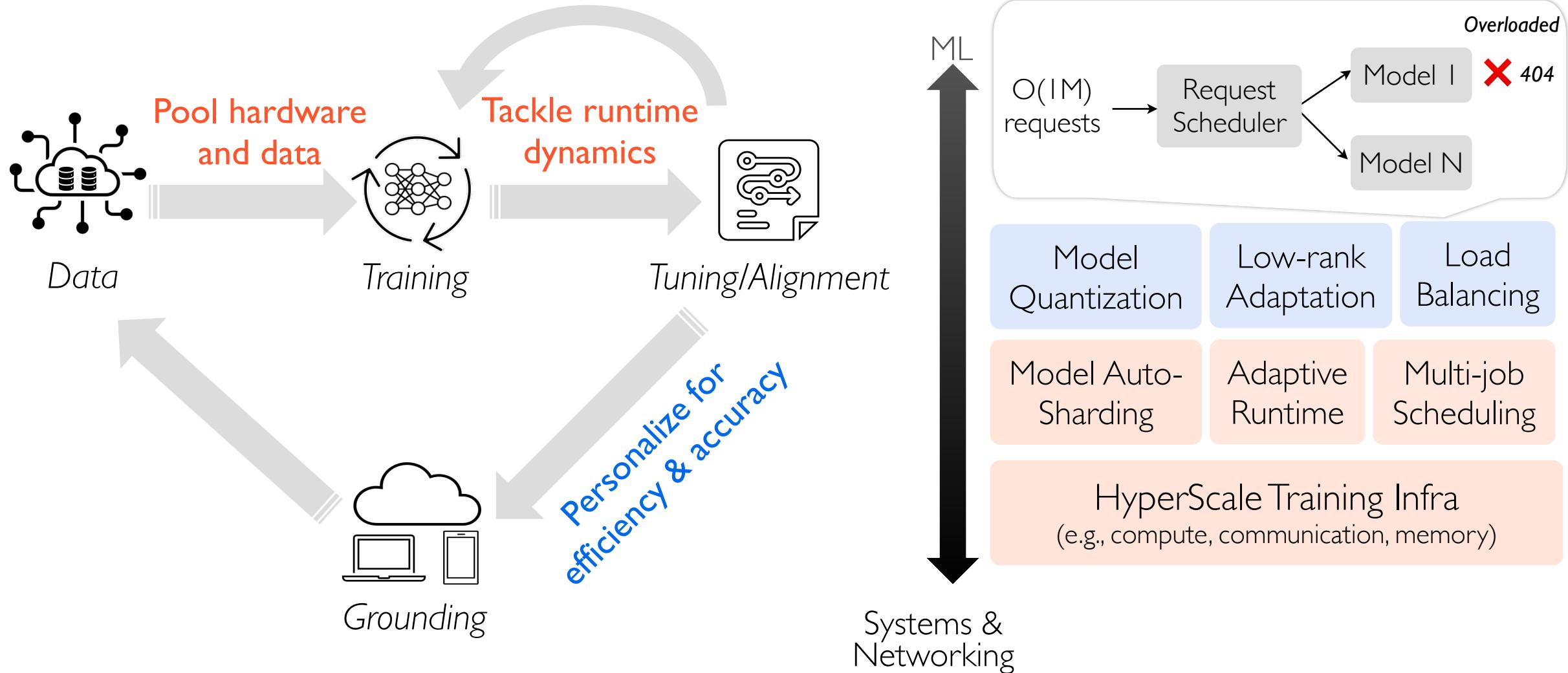
GenAI Lifecycle Demands Great Systems Support



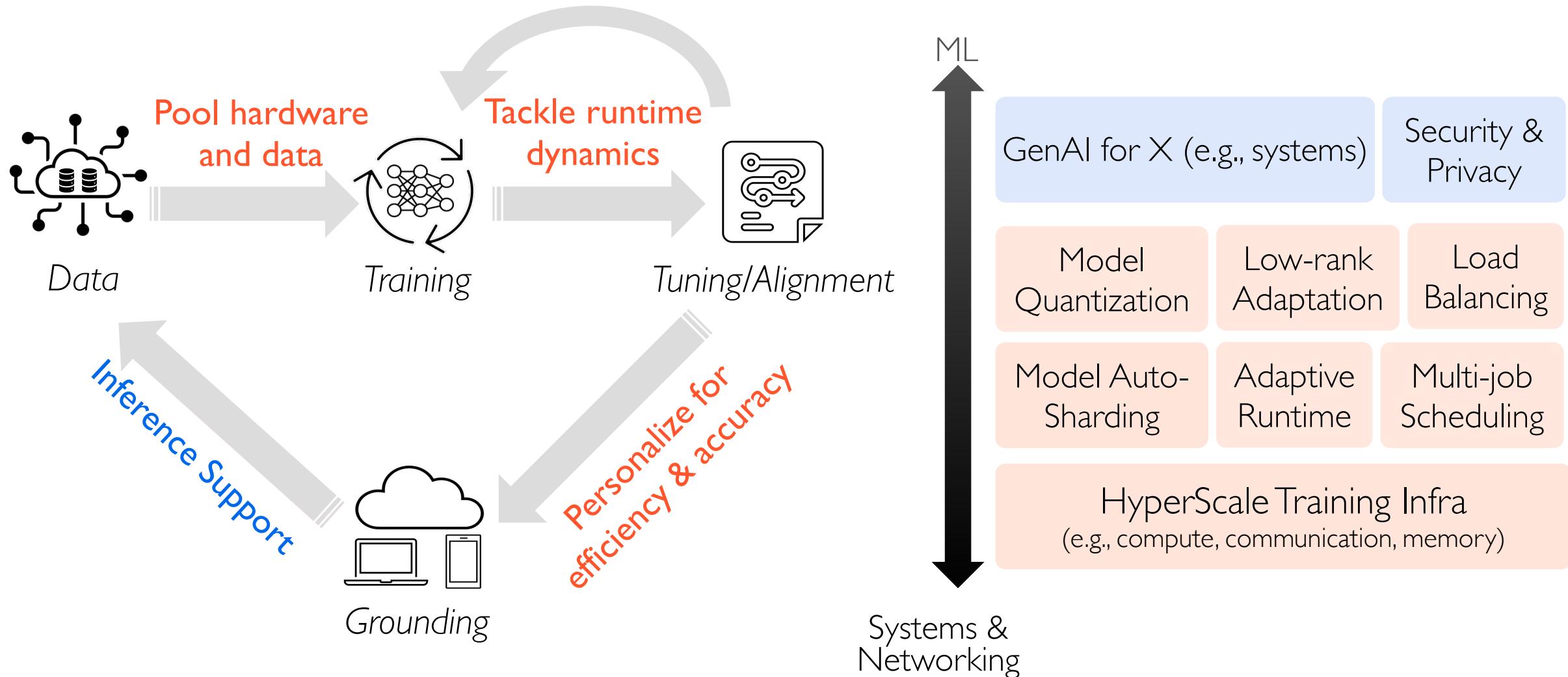
GenAI Lifecycle Demands Great Systems Support



GenAI Lifecycle Demands Great Systems Support



GenAI Lifecycle Demands Great Systems Support

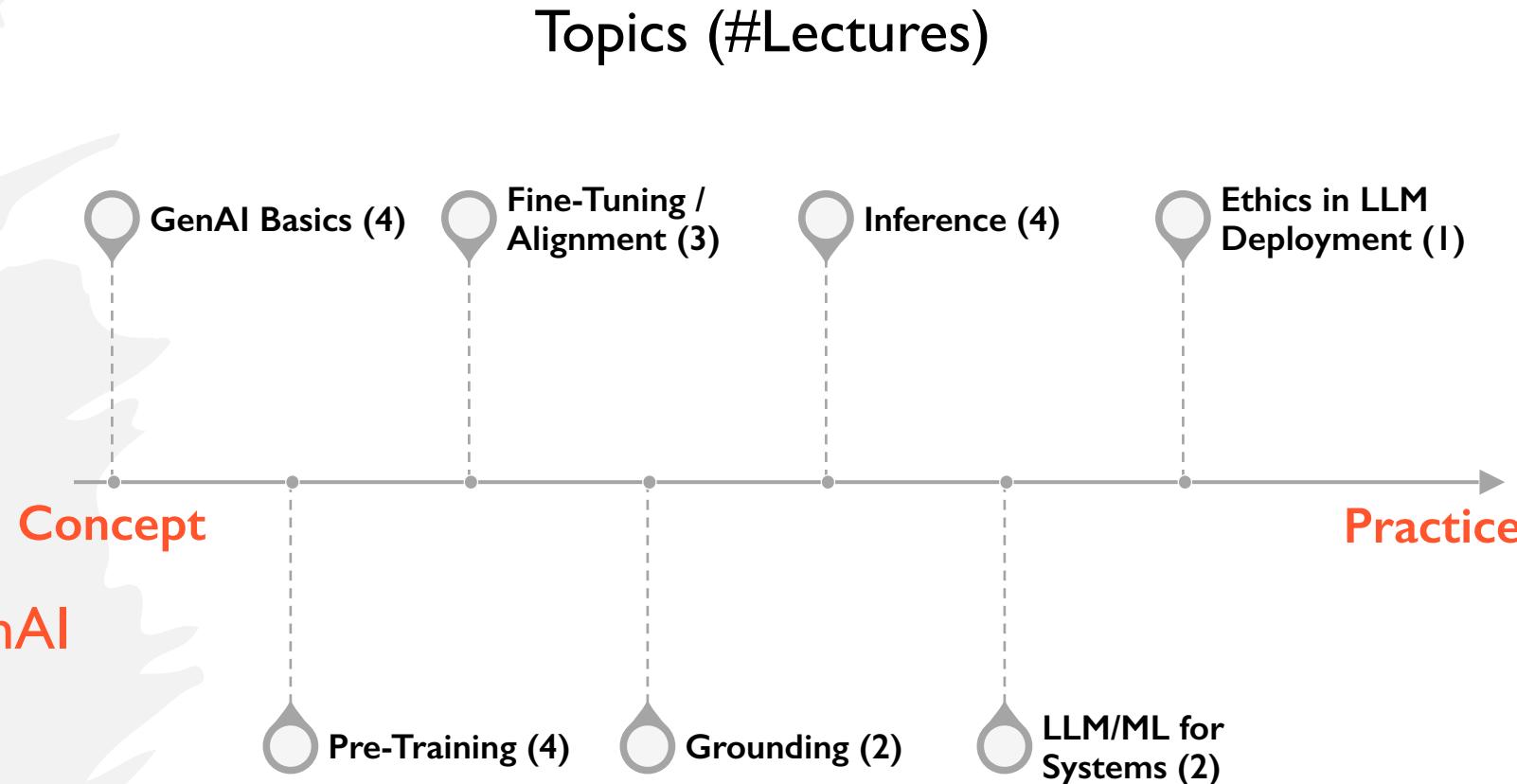


What will you learn in this course?

-- 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 Requirements (Tentative)

Participation	10%	Paper reading; Up to two absences
---------------	-----	-----------------------------------

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

Course Requirements (Tentative)

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

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

Course Requirements (Tentative)

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

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

Course Requirements (Tentative)

Participation	10%	Paper reading; Up to two absences
Paper Summary	15%	Two summaries
Paper Presentation & Discussion	15%	One Presentation
Project Presentations	20%	Mid-term + Final Poster
Project Report	40%	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 3
 - May allow a few smaller groups if/when students drop out
- Submit at <https://forms.gle/cmoy5Hi9JLRmtcbq9>
 - By Sept 5th the latest, but **right now** is better

Participation

- **Reading: 39 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

Participation

- **Reading: 39 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

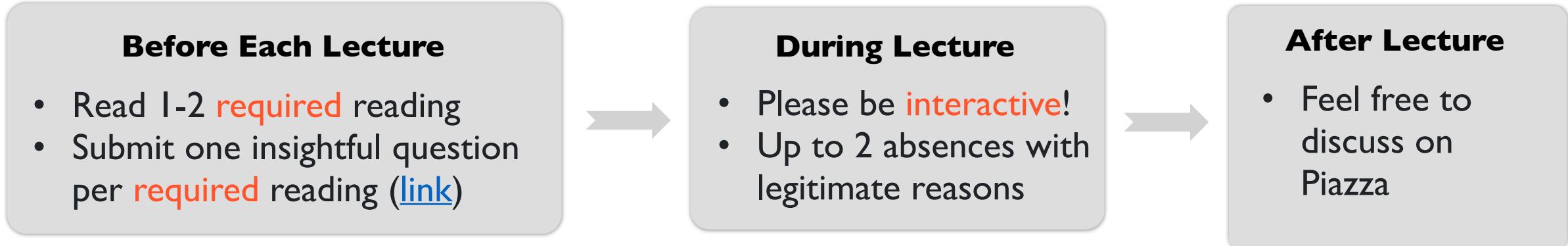
Before Each Lecture

- Read 1-2 **required** reading
- Submit one insightful question per **required** reading ([link](#))

Pre-Training	
Sept 18	The Llama 3 Herd of Models (Sec 1-4, Required) Gemini: A Family of Highly Capable Multimodal Models

Participation

- **Reading: 39 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



Paper Presentation

- **This is a seminar-style course**
 - Each group presents ~1 lectures covering required papers
 - 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
- **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 ~1 lectures covering required papers
 - 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
- **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@illinois.edu 24 hrs before the class

Paper Summaries

- ~2 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 should include the gist of class discussion

Share summary (in *.pdf) to cs598-aisys-staff@illinois.edu within 24 hrs of class 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 your final slides after class

Authors (Companion) Group

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

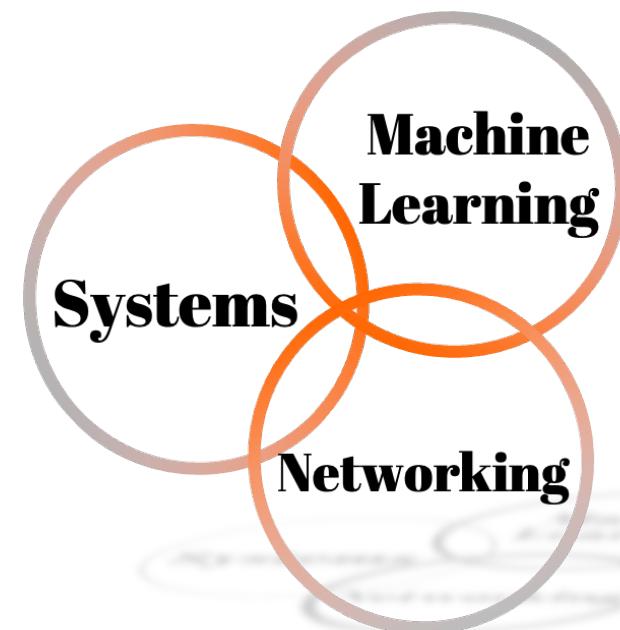
Reviewers Group

- 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 (Sept 26)

- Two pages, plus as many pages as needed for 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 (Oct 21, 23)

- **In-class short presentation over two days**
 - This is to make sure you are receiving feedback and making progress
- **Must include**
 - What is the problem?
 - Why is it important?
 - What are the most related work?
 - What's your hypothesis so far?
 - How are/will you evaluate it?

Poster & Paper (Dec 7, 19)

- **Poster session (Dec 7)**
- **Final report (7-8 pages excluding references):**
 - Extended from mid-term writeup
 - 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
09/05/24	Form Group	Find 3 like-minded teammates
09/26/24	Submit Proposal	Send your proposal by email to receive feedback either via email or in-person or both
10/21/24 10/23/24	Mid-Semester Presentations	Define and motivate a problem, overview, related work, and form initial hypothesis/idea
12/04/24 12/07/24	In-Class or Poster Presentations	Present your research findings
12/19/24	Research paper	Submit a report like the papers you read

NO extensions! Delayed submission will receive NO CREDIT

Thank you!

Form Groups of 3; submit by *Sept 5*

Reviewers Group

- Write one paper summary
- Participate in panel discuss.

Authors (Companion) Group

- Write one paper summary
- Participate in panel discuss.