CSE 585: Advanced Scalable Systems for GenAI

Mosharaf Chowdhury

Today's Agenda

- Administrivia
- Topics
- Projects

About Mosharaf

- Associate Professor of CSE
  - http://www.mosharaf.com/
  - https://symbioticlab.org/
- Office hours:
  - Appointment-only

ViNEYard

virtual network embedding (2008-2012)

in-memory computing (2009-2014)

Coflow

data-parallel communication (2010-2016)

Infiniswap

software memory disaggregation (2016-2022)

Salus

GPU resource management (2017-2022)

FedScale

systems for federated learning (2019-)

AI energy optimization (2021-)

●●●

About Insu Jang (GSI)

- 4th-year PhD student at SymbioticLab

- Office hours from next week
  - 4828 BBB, 1230PM-130PM Fridays
  - No office hours this week
  - insujang@umich.edu

Status

- As of today: ~60 registered or w/ override
- If you are not planning to take the class, drop ASAP
  - Existing overrides that have not converted will be revoked

Course Schedule

- **Webpage:** https://github.com/mosharaf/cse585
- **Meetings**
  - 10:30 AM – 12PM (**T/Th** for lectures and seminars)
  - 1:30 PM – 2:30 PM (**Fri** for makeups and projects)
- **Pay attention to the online announcements and schedule**
  - On average, two meetings per week
  - Friday makeups will be added on a need-to-add basis

Prerequisites

- **EECS 482 / 484 / 489 / 491**
  - Equivalent courses are acceptable as well
- **Good programming skills**
  - Build substantial systems for course project

Course Requirements

Paper Summary	15%
Paper Presentation	15%
Participation	10%
Project Report	40%
Project Presentations	20%

Topics (#Lectures)

- GenAI Basics (3)
- Pre-Training (4)
- Post-Training (1)
- Inference (6)
- Grounding (1)
- GenAI (for) Systems (1)
- Power and Energy (1)
- Ethical Considerations (1)

Group-Based Work

- ALL activities will be done in groups except for participation
  - Paper presentation
  - Paper summary
  - Research projects

Form Groups ASAP

- **Submit at** https://forms.gle/iHkfmPvBtz5gXjTb7
  - By September 5 the latest, but **right now** is better
    - We need a group to pickup duties for Sep 5!!!
  - Use piazza to find group members
  - Group size should be 3 to 4

Readings

- **40 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 must present at least one lecture (required papers and the rest)
  - Paper presentation account for **15%** of the total grade
- **The entire class will be dedicated to the assigned paper(s)**
  - Aim for 40-minute presentation without interruption
  - But there will be intermittent discussions
- **Lead the discussion**
  - Go through the paper in details, along with its strengths and weaknesses
  - Include companion papers and other related papers

Paper Presentation

- Share your slides to cse585-staff@umich.edu 24 hours before the class
  - Use Google Sheets so we can provide in-place comments/feedback
- Prepare early
- Practice a lot
- Also, read
  - How to Give a Bad Talk, by David A. Patterson

Paper Summaries

- **This is a paper-reading course**
  - Paper summaries account for **15%** of the total grade
- **Roughly 1-2 summary per-group (assigned)**
- **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 or limitation of the proposal, and how will you improve it?
- **Summary must include the gist of class discussion**

Paper Summaries

- **Reviews must be shared to cse585-staff@umich.edu within 24 hours of class presentation**
  - Use Google docs so we can provide in-place comments/feedback
- **Delayed submission will receive NO CREDIT**
  - There will be NO extensions

Panel Discussion

- **The Authors**
  - Groups that present and write summary
- **The Reviewers**
  - Each group will be assigned to at least one of these slots
  - Will have their own questions to ask to the authors
  - Will receive questions raised by the class (described below) from the GSI before the lecture
- **Rest of the Class**
  - Submit one insightful question for each presented papers by 3PM the day before
  - Ask questions directly too

In general,

- No extensions
- Everyone must come to class after reading the required papers of the day

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

Resource-Centric View

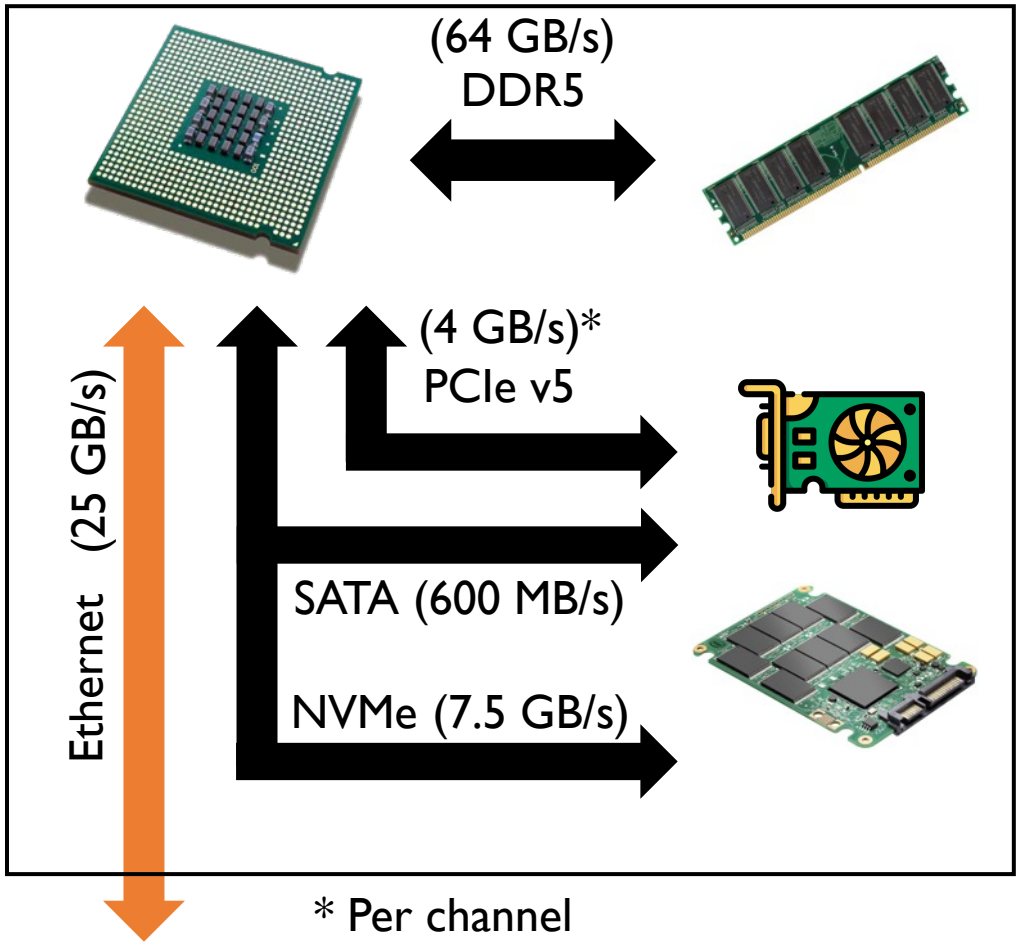
What's in a (Simplified) Server?

Interconnected compute and storage resources

- Different bandwidth and latency constraints

Simplified diagram

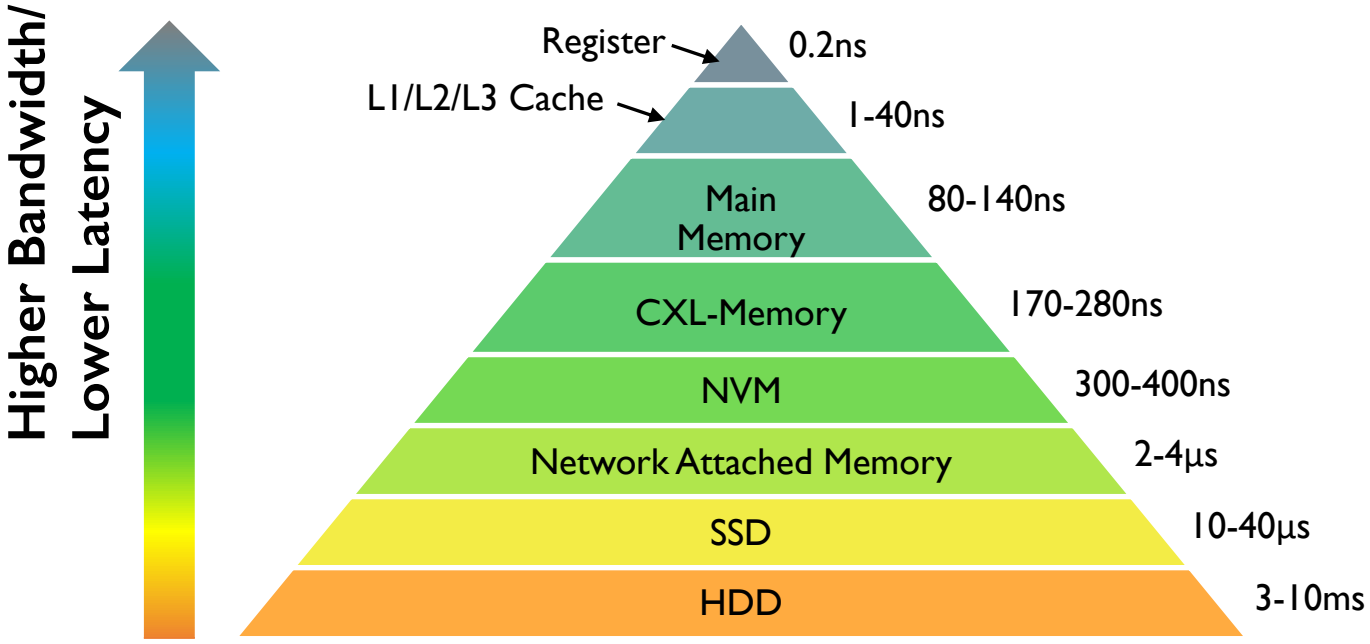
- Doesn't include faster networks such as RDMA, dedicated GPU interconnects such as NVlink, etc…



Typical Memory/Storage Hierarchy

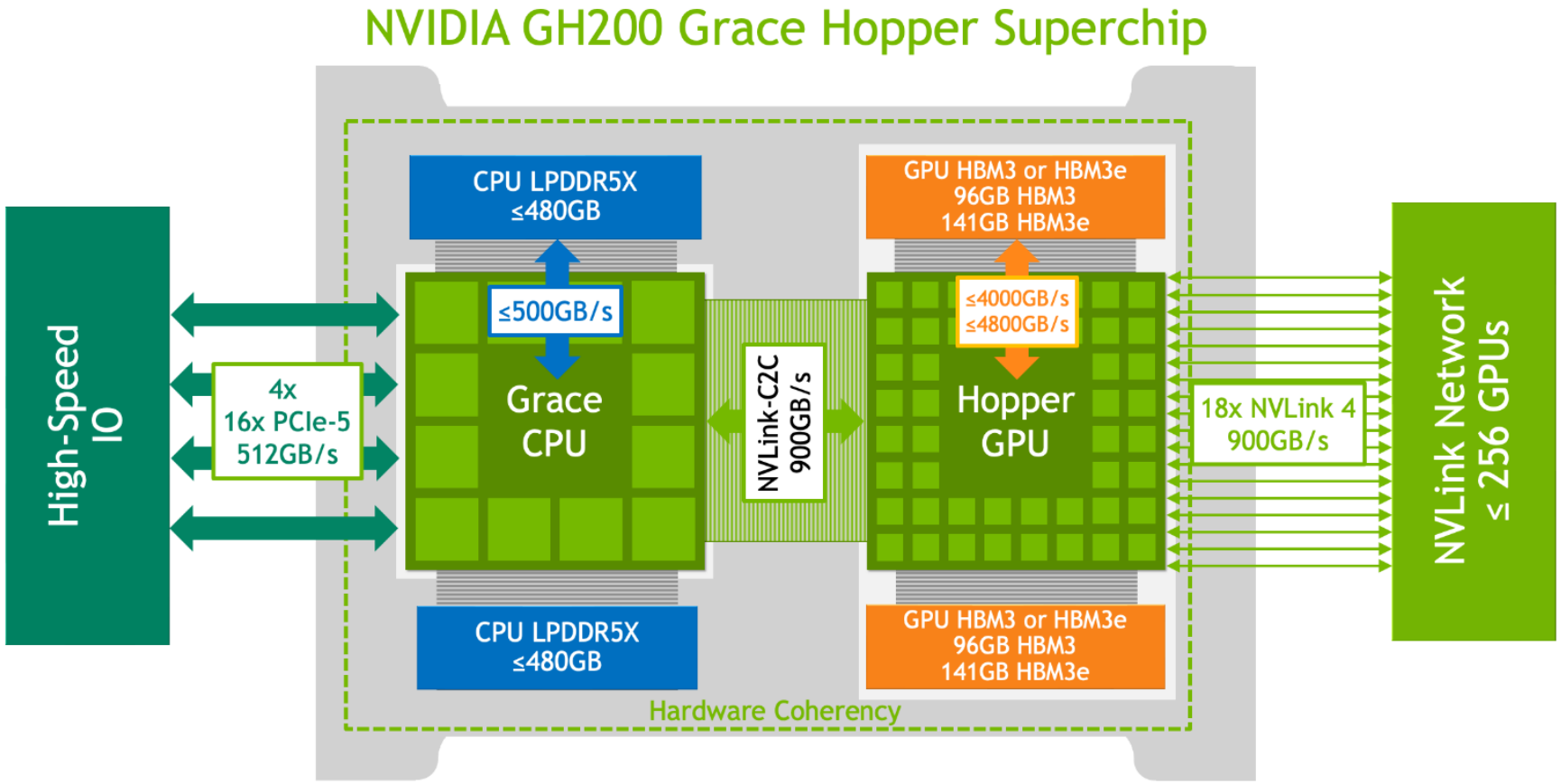
Fundamental Goals of (SW/HW) System Design

- Minimize time to access data
- Maximize compute utilization
- **Balanced System**



Maruf et al, SIGMETRICS 2023

What's in a Modern AI Server?



https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

Scale Out: Warehouse-Scale Computer (WSC)

Single organization

Homogeneity (to some extent)

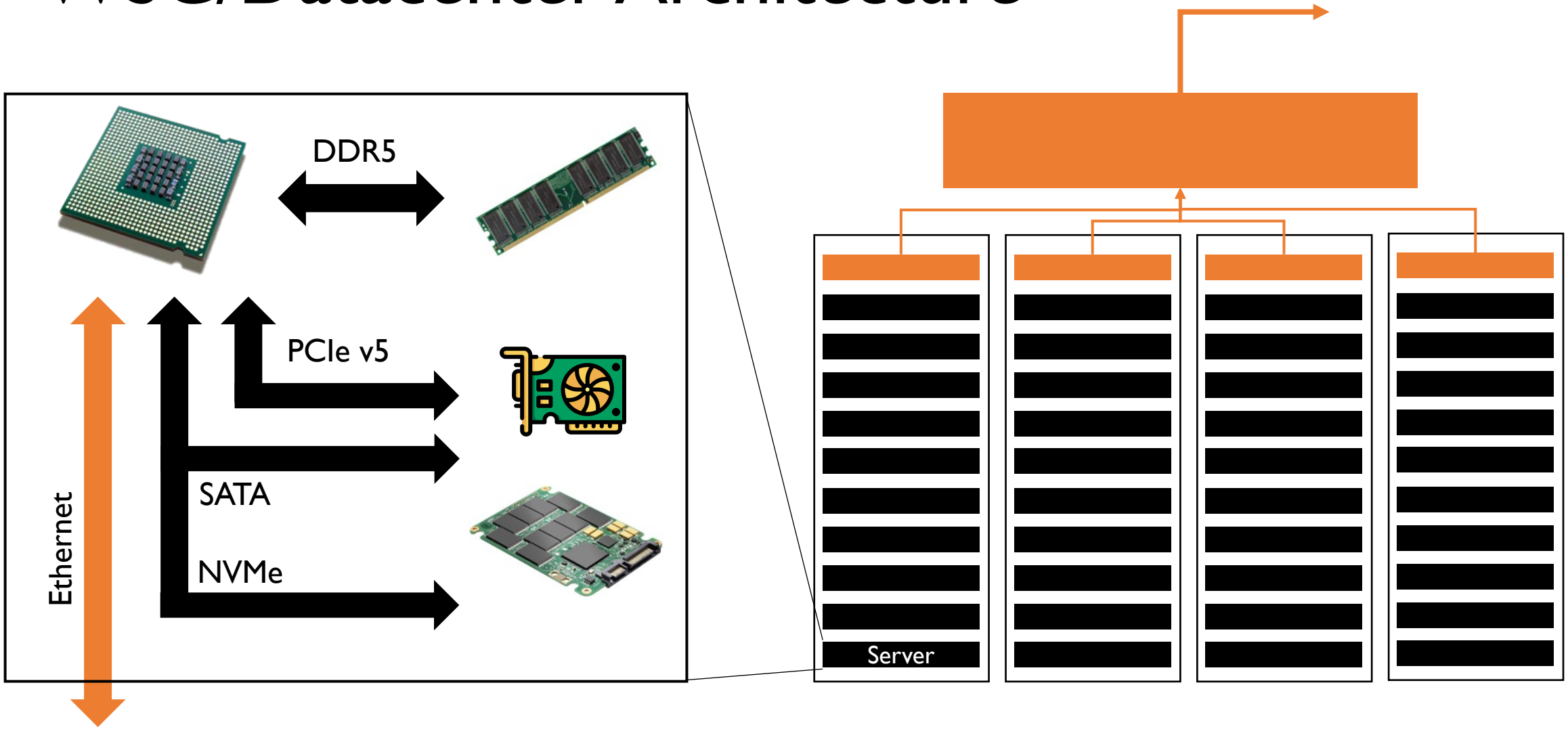
Cost efficiency at scale

- Multiplexing across applications and services
- Rent it out!

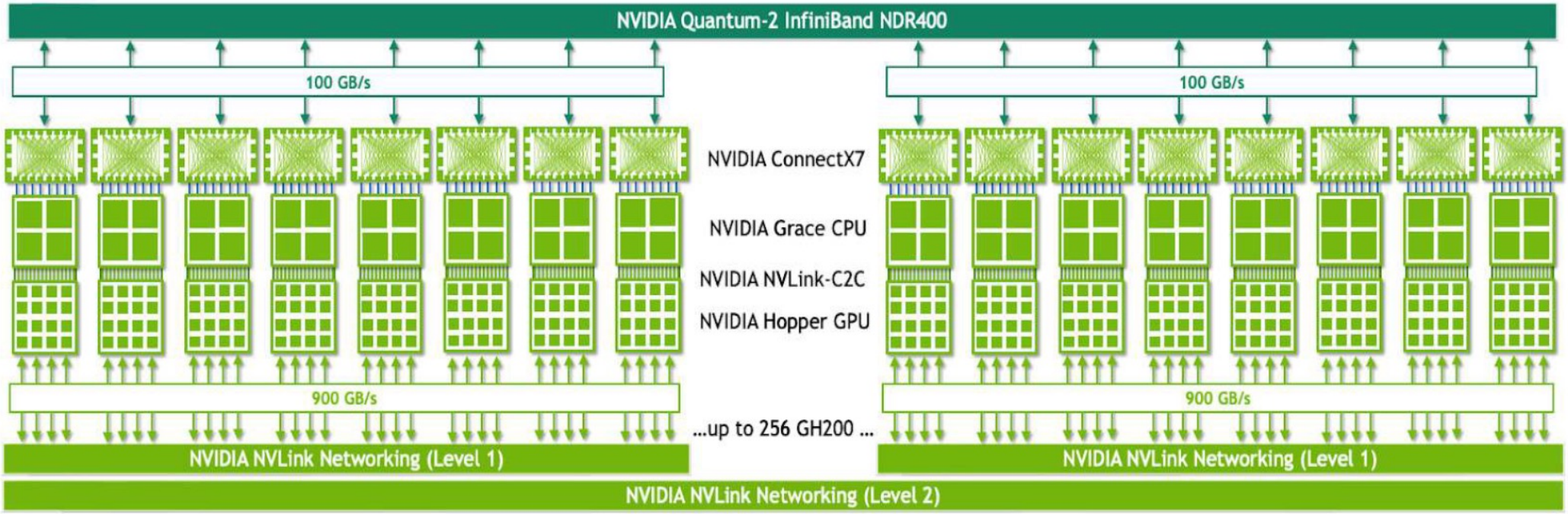
Many concerns

- Infrastructure
- Networking
- Storage
- Software
- Power/Energy
- Failure/Recovery
- …

WSC/Datacenter Architecture



Example: Scaling Out Using NVIDIA GH200



https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

Datacenter Needs an Operating System

Datacenter is a collection of

- Compute
- Memory
- All connected by an interconnect

Not unlike a computer

Some differences

1. VERY high level of parallelism
2. VERY large scale
3. Diversity of workloads
4. Resource heterogeneity
5. Failure is the norm

Three Categories of Software

1. **Platform-level**
   - Software firmware that are present in every machine
2. **Cluster-level**
   - Distributed systems to enable everything
3. **Application-level**
   - User-facing applications built on top

Common "Systems" Techniques

Technique	Performance/Efficiency	Availability/Resilience
Replication & Erasure coding	X	X
Sharding/partitioning	X	X
Scheduling & Load balancing	X	
Health & Integrity checks		X
Compression & Quantization	X	
Centralized controller	X	
Canaries		X
Speculation & Redundant execution	X	

Break!

Workload-Centric View

The Llama 3 Herd of Models

Three key levers in the development of high-quality foundation models

1. Data
   - Both quality and quantity matters
   - Both pre-processing and post-processing matters
   - Llama 3 was pre-trained on a corpus of about 15T multilingual tokens.
2. Scale:
   - Compute-optimal training for the biggest and overtraining for smaller ones
   - Pre-trained using $3.8*10^{25}$ FLOPs
3. Managing complexity:
   - Simplicity for scale

Pre-Training

1. Curation and filtering of a large-scale training corpus
2. Development of a model architecture and corresponding scaling laws for determining model size
3. Development of techniques for efficient pre-training at large scale
4. Development of a pre-training recipe

Scaling Laws

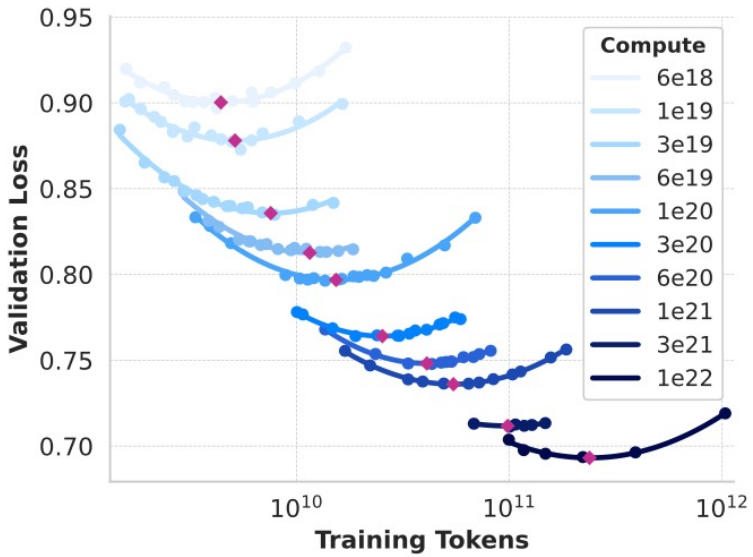


Figure 2 Scaling law IsoFLOPs curves between 6×10^{18} and 10^{22} FLOPs. The loss is the negative log-likelihood on a held-out validation set. We approximate measurements at each compute scale using a second degree polynomial.

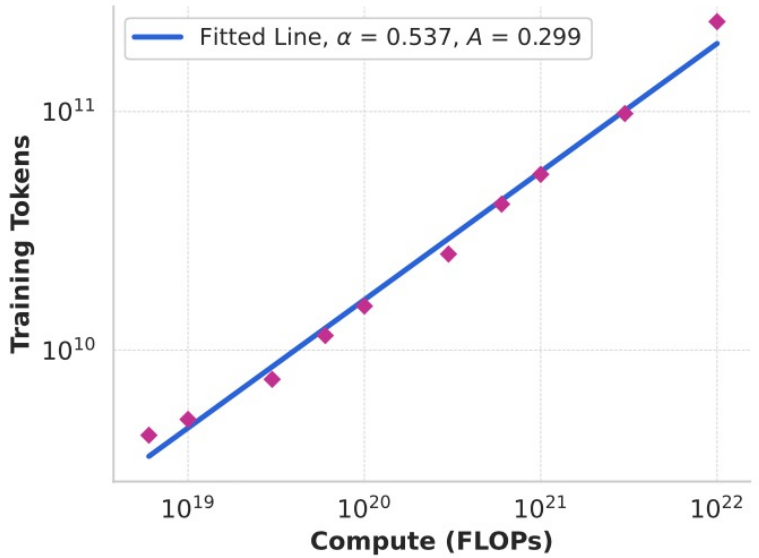


Figure 3 Number of training tokens in identified compute-optimal models as a function of pre-training compute budget. We include the fitted scaling-law prediction as well. The compute-optimal models correspond to the parabola minimums in Figure 2.

https://llama.meta.com/

Training Infrastructure

- **Compute**
  - Up to 16K H100 GPUs, each running at 700W TDP with 80GB HBM3
  - Each server is equipped with eight GPUs and two CPUs. Within a server, the eight GPUs are connected via NVLink
- **Storage**
  - Tectonic: 240 PB of storage out of 7,500 servers equipped with SSDs, and supports a sustainable throughput of 2 TB/s and a peak throughput of 7 TB/s
  - Major challenge: highly bursty checkpoint writes that saturate the storage fabric for short durations
    - Ranging from 1 MB to 4 GB per GPU, for recovery and debugging
    - Minimize GPU pause time during checkpointing and increase checkpoint frequency to reduce the amount of lost work after a recovery

Network Infrastructure

- **400 Gbps interconnects between GPUs**
  - Llama 3 405B used RDMA over Converged Ethernet (RoCE) fabric
  - Smaller models were trained using Nvidia Quantum2 InfiniBand fabric
- **Network Topology**
  - RoCE-based cluster has 24K GPUs connected by a three-layer Clos network
  - 16 GPUs/rack x 192 racks/pod x 8 pods
  - Full BB within each pod and 1:7 oversubscription across pods
  - Cluster software are topology-aware
- **Load balancing and CC**
  - See paper

4D Parallelism

- **Tensor, Pipeline, Context, and Data parallelism**
  - See Figure 5
- **GPU utilization**
  - Careful configuration of the parallelism configuration, hardware, and software
- **Network-aware parallelism configuration**

Reliability and Operational Challenges

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

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 on new environments are acceptable
  - Reproducing results from existing solutions is also acceptable
- 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

Milestones

Date	Milestone	Details
09/05/24	Form Group	Find like-minded students
09/19/24	Submit Proposal	Send your proposal by email to receive feedback either via email or in-person or both
10/22/24 10/24/24	Mid-Semester Presentations	Define and motivate a problem, overview related work, and form initial hypothesis and idea
12/03/24 12/05/24	In-Class or Poster Presentations	Present your findings
12/16/24	Research paper	Submit a report like the papers you read

Draft Proposal (Sep 19)

- **Two pages including references that must include**
  - What is the problem?
  - Why is it important to solve?
  - Any initial thoughts on what you want to do?
  - How would you evaluate your solution?
- **Include team members**
  - Meaning, form a group ASAP
- **Approved by the instructor and agreed upon by you**
  - Forms the basis of expectation

Mid-Semester Checkpoint (Oct 22,24)

- **In-class short presentation over two days**
  - This is to make sure you are 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?

Presentation & Paper (Dec 3, 5, 16)

- **Research paper**
  - The key part
  - 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
- **Extended from the mid-semester checkpoint**

Project Ideas

Some project suggestions

- https://docs.google.com/document/d/1sxhnw_443IYerYXBo0LmkgcQusAjnpQfsBRhMBW53a0/edit#heading=h.wlyqsjqv97gf

You can propose your own projects too!

Next Class…

Read the required readings

Form groups of 3-4 and fill out https://forms.gle/iHkfmPvBtz5gXjTb7 by *Sep 5*

- Decide if you'll drop, **before** you fill it
- If you are to drop, drop immediately

Sign up for Sep 5 presentation slot for extra benefits!