ECS 289D Seminar in OS: Datacenter Systems for LLMs

Yang Zhou
Sep 25, 2025

With slides adapted from Prof. Amanda Raybuck last-year offering

About me

- Yang Zhou
    - BS in CS from Peking University, China in 2018
    - PhD in CS from Harvard University, USA in 2024
    - Postdoc from UC Berkeley Sky Computing lab in 2025
    - First-year assistant professor at UC Davis
- Research focus:
    - Equal interests in core systems and ML systems research,
        - e.g., efficient LLMs, GPU communication, heterogeneous computing.
    - Currently working on UCCL for GPU communication
        - https://github.com/uccl-project/uccl

Agenda for today

Introduction to datacenter systems for LLMs

Course logistics and more

Topic overview

How to give a good talk

Introduction to datacenter systems for LLMs

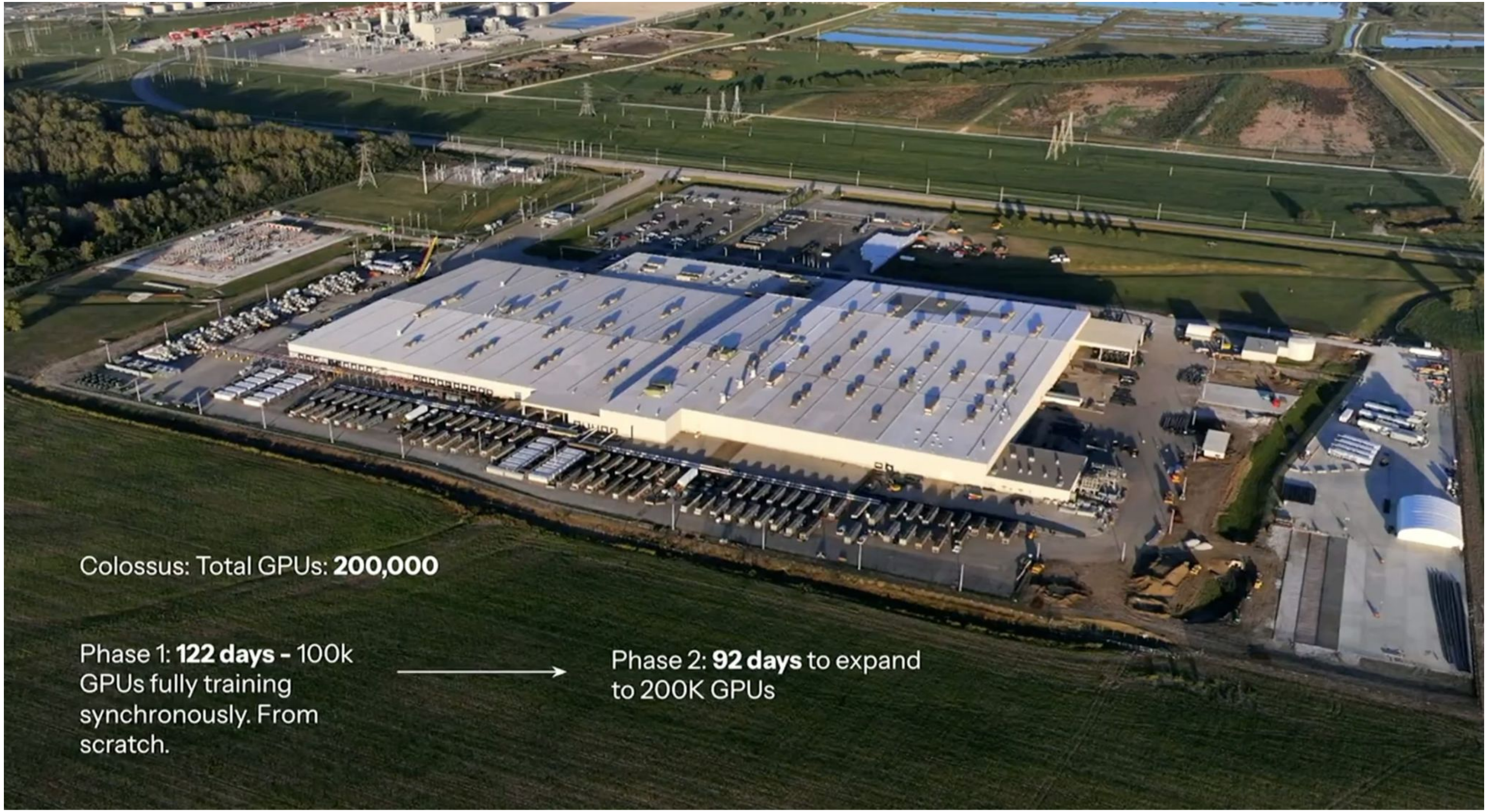
LLM Booming: Companies & Models

- LLM companies are rapidly emerging
- Many new LLM models are being developed
- The LLM landscape includes diverse offerings
- Interest in large language models is increasing

LLM Use Cases

- LLMs power intelligent chatbots and virtual assistants
- They enable advanced content generation and summarization
- LLMs are used for complex data analysis and extraction
- They facilitate code generation and debugging assistance

xAI Colossus for LLMs: 300MW, 200k GPUs



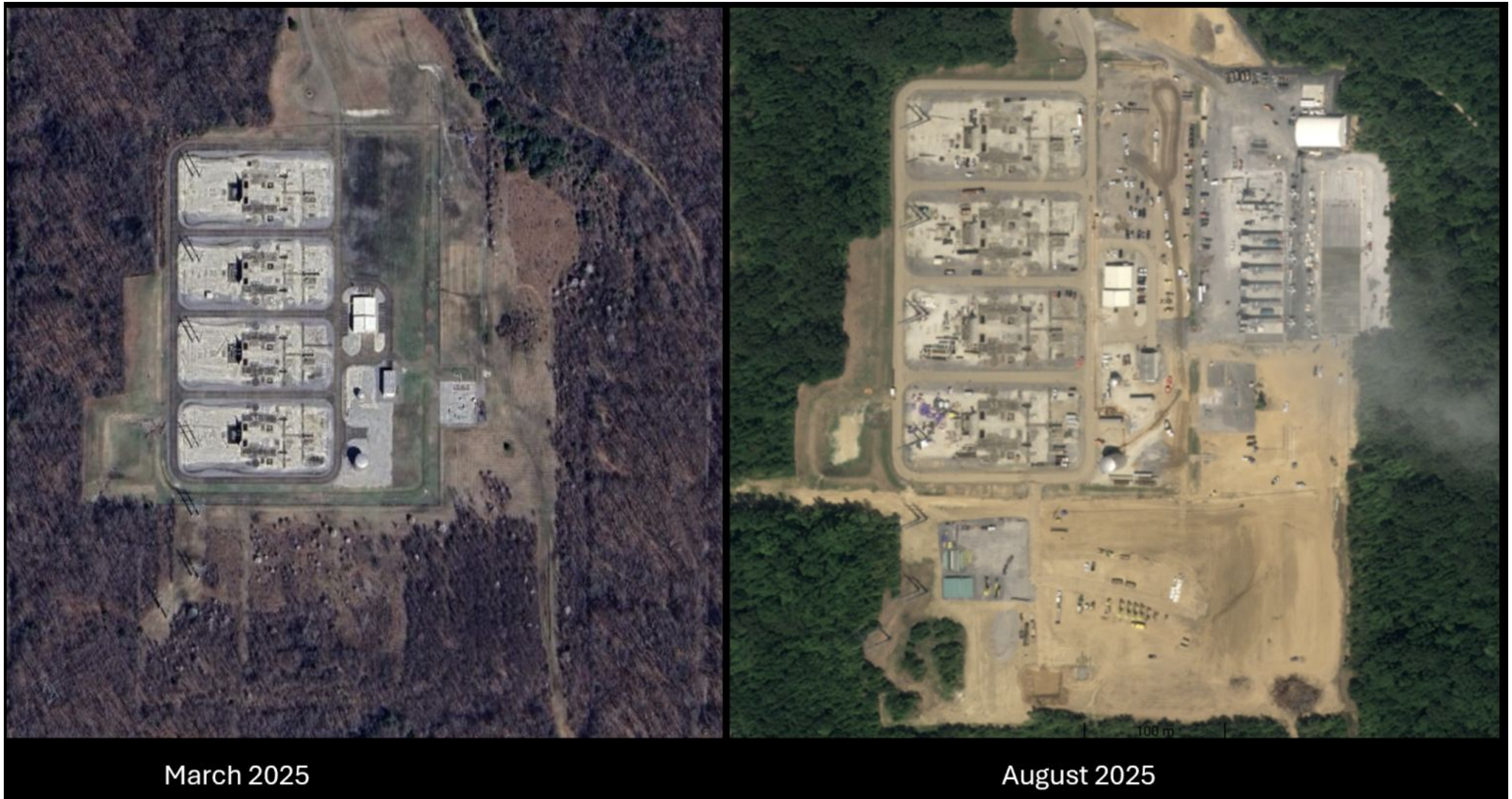
Colossus: Total GPUs: **200,000**

Phase 1: **122 days -** 100k GPUs fully training synchronously. From scratch.

→

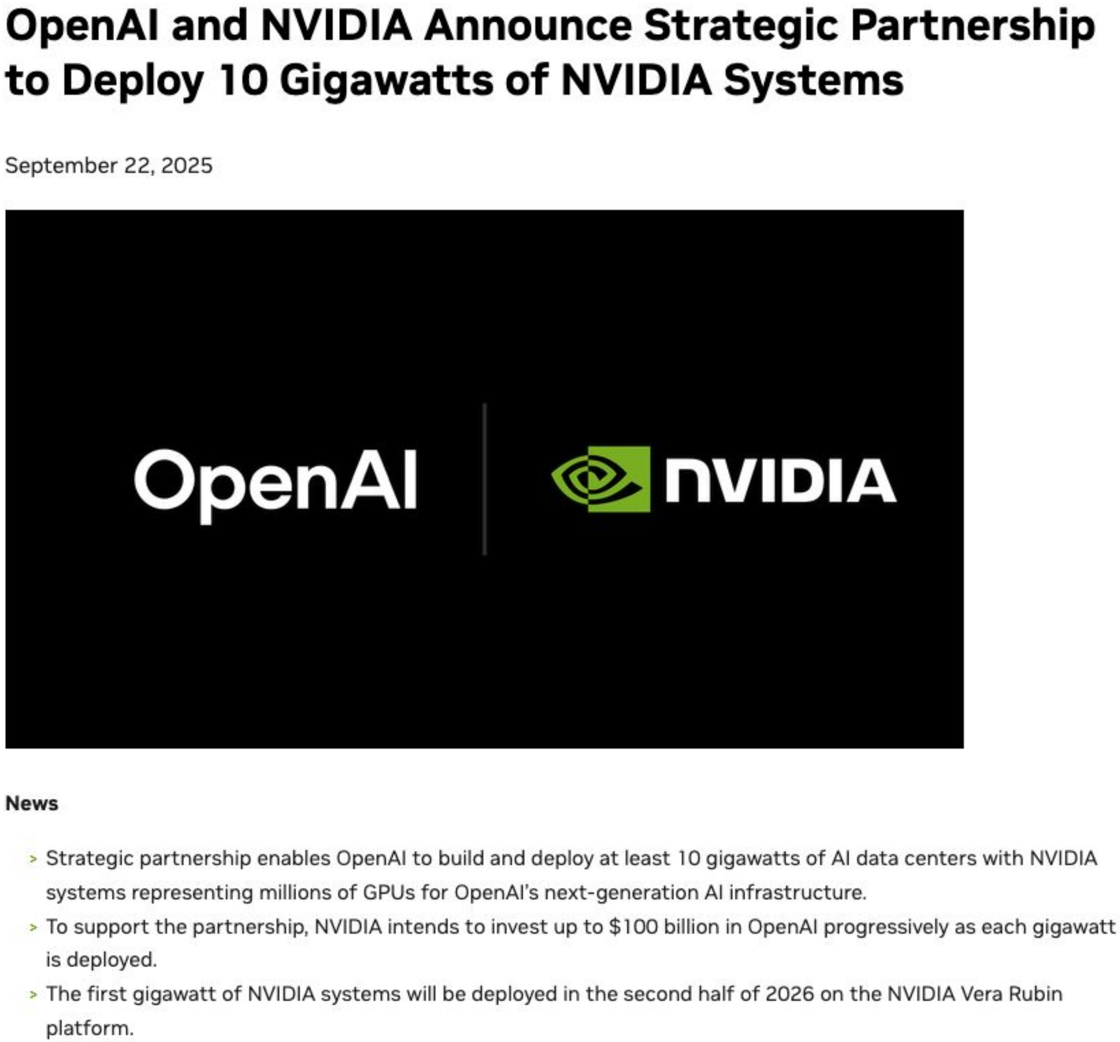
Phase 2: **92 days** to expand to 200K GPUs

xAI Colossus 2 for LLMs: 1.1GW, 550k GPUs



March 2025

August 2025



OpenAI and NVIDIA Announce Strategic Partnership to Deploy 10 Gigawatts of NVIDIA Systems

September 22, 2025

News

- Strategic partnership enables OpenAI to build and deploy at least 10 gigawatts of AI data centers with NVIDIA systems representing millions of GPUs for OpenAI's next-generation AI infrastructure.
- To support the partnership, NVIDIA intends to invest up to $100 billion in OpenAI progressively as each gigawatt is deployed.
- The first gigawatt of NVIDIA systems will be deployed in the second half of 2026 on the NVIDIA Vera Rubin platform.

What is a datacenter?

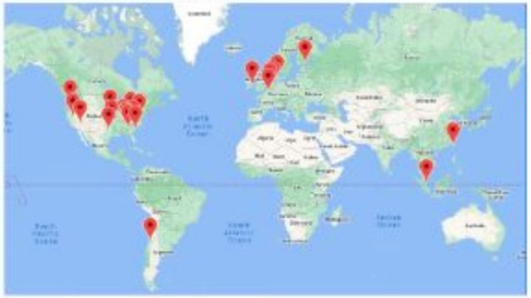
- Large facility housing primarily commodity computers
  - 10,000s - millions of machines
  - Commodity servers (e.g., Dell, Open Compute Project (OCP))
- Interconnected by a commodity network
  - Ethernet
- Commodity?
  - Cheap! Reduces operating costs (vs. custom parts)
  - “One-size-fits-all” components
  - All good as long as they improve (Moore’s Law)
  - Strong drive by DC operators to commoditize all infrastructure

Why Datacenters?

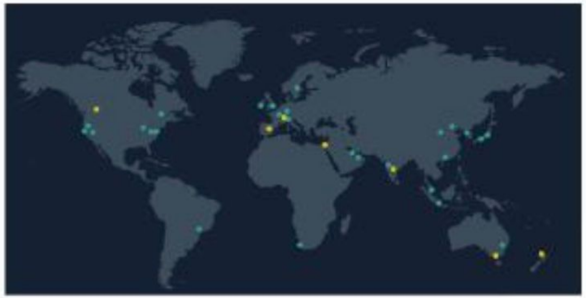
- Consolidation
  - Run many people’s workloads on the same infrastructure
  - Use infrastructure more efficiently (higher utilization)
  - Leverage workload synergies (e.g., caching)
- Virtualization
  - Build your own (virtual) private infrastructure quickly and cheaply
  - Move it around + scale it up/down anywhere, anytime
- Outsourcing
  - No need to maintain an on-premise set of servers
  - Expertise is provided by the datacenter vendor

Datacenters Today

- Over 8000 datacenters globally
- Over 2600 datacenters in the US
- Huge energy consumers – almost 2% of global energy use
  - Usually built near energy sources (hydroelectric, wind, solar)



Google datacenters



Amazon AWS datacenters

Why researching datacenters systems for LLMs?

- Datacenters have become the cornerstone of the LLM workloads
  - The world’s most valuable companies are datacenter operators
- They comprise the state-of-the-art in computing
  - Scale, technology, applications
- They combine networking, systems, and economic lessons
  - Increasingly important knowledge
- They are rarely discussed in class
  - Fairly recent topic
  - Crosses disciplines

Course logistics and more

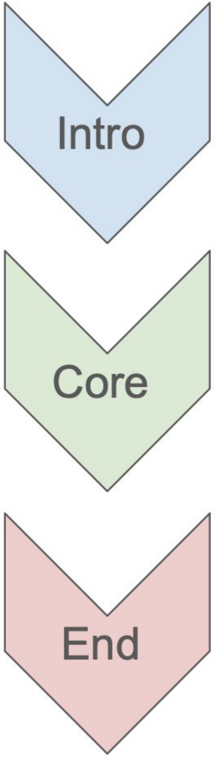
Course logistics

- Graduate-level, seminar-based, research-focused course
- The goals of the course are:
  - To learn about classic and cutting-edge datacenter systems for LLMs
  - To practice reading and discussing research/technical papers
  - To conduct a research project
- Three main components:
  - Reading list + paper reviews + discussion
  - Leading a lecture
  - Research project

Assumptions

- Programming in C/C++ and Python
- Basic OS and Networking concepts
  - Virtual address, isolation
- Basic knowledge about computer architecture
  - GPU, registers, DMA engines

Schedule



Intro

- Class Introduction
- Define projects and find groups
- Decide on paper(s) to present

Core

- Research paper presentations & discussions
- Work on project

End

- Present project
- Submit project and source code

Paper Readings

- Most of our material is not (or insufficiently) covered in textbooks
- Readings will be **research papers or famous tech blogs**
  - Seminal papers in the field, research papers that show what is possible
  - Details of key ideas in datacenter systems for LLMs
- We will read & discuss 2-4 papers per week
  - Each session will cover
    - 1 required reading, 1-2 optional readings
      - 5% bonus points if you also write (good) reviews for optional readings
    - 2 presentations + discussions, 40 min each
      - 1 presentation for required reading, 1 for optional readings
      - Presentations and discussions can blend together

Paper Reviews: What you need to do

- Read the paper before class
- Write a short review that includes:
  - A short (3-4 sentence) summary of the paper
  - Some advantages or disadvantages of the approach
  - Any other questions or comments you have
- Submit on canvas by **11:59pm the night before class**
  - See canvas assignments

Paper Presentations: What you need to do

- Select a paper to present
  - First-come, first-served, due by **11:59 pm Tue 9/30**
  - Select a paper via Canvas comment under “Paper Assignments” post
    - Available today noon
  - If everything is taken, up to 2 students may present 1 paper
- When it is your time, prepare and present the paper for discussion
  - See schedule for order of papers
  - Send me a copy of your slides after the session, so I can put on the course website
- Participate in discussion!
  - Worth 10% of your grade

Projects: What you need to do

- Teams of 3-5 students work on each project
- Define your project
  - Whatever related to systems and LLMs
  - Talk with me on the compute resources you need
- Prepare a proposal
  - Paper introduction defining the project
  - Evaluation plan and timeline/milestones
- **Start working!**
  - Check in with me regularly (mandatory mid-quarter check in by 11/06)

Project Discussion

- Post (and answer) questions on Canvas
  - Gets you better, faster, collaborative answers
- Check in with me regularly
  - Use office hours
  - 1 mandatory mid-quarter check in (11/06)

Grading

Straight grading scale (>=90% A, 87-89 B+, 83-86 B, 80-82 B-, etc.)

- Project: 50%
  - 1% project proposal
  - 4% mid-term report
  - 10% final project presentation
  - 35% final report and code
- Class presentation: 25%
- Paper reviews: 15%
- Class participation: 10%
  - Class attendance, in-class discussions, and online discussions

Academic Integrity and Generative AI Policy

- Talking with others in the class about papers, projects is OK, encouraged
- Give credit to any resources you borrow for presentations, write-ups
  - E.g,. if you use slides from a conference presentation in your paper, give credit
  - When in doubt, cite it!
- I am willing to allow the use of Generative AI, but within reason (& ACK use)
  - Using it for proposal/project writing:
    - You may use it for a “first draft” but I expect human-made revisions over what the model produces
    - OK to use for grammar and clarity improvements
  - Using it for code generation:
    - OK , but acknowledge use in write-ups
  - You (and your group) are responsible for understanding anything you turn in!

Topic overview

Four main topics

- Datacenter networking
- Host networking
- LLM Inference
- LLM Training

Datacenter networking

The Tail at Scale

optional Attack of the Killer Microseconds

A Scalable, Commodity Data Center Network Architecture

optional VL2: A Scalable and Flexible Data Center Network

Data Center TCP (DCTCP)

optional Swift: Delay is Simple and Effective for Congestion Control in the Datacenter

Design Guidelines for High Performance RDMA Systems

optional Deconstructing RDMA-enabled Distributed Transactions: Hybrid is Better!

Host networking

RDMA over Ethernet for Distributed AI Training at Meta Scale

optional An Extensible Software Transport Layer for GPU Networking

IX: A Protected Dataplane Operating System for High Throughput and Low Latency

optional Arrakis: The Operating System is the Control Plane

Shenango: Achieving High CPU Efficiency for Latency-sensitive Datacenter Workloads

optional Snap: a Microkernel Approach to Host Networking

Demystifying NCCL: An In-depth Analysis of GPU Communication Protocols and Algorithms

optional MSCCL++: Rethinking GPU Communication Abstractions for Cutting-Edge AI Applications

LLM Inference

Efficient Memory Management for Large Language Model Serving with PagedAttention

optional vAttention: Dynamic Memory Management for Serving LLMs without PagedAttention

DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving

optional Optimizing SLO-oriented LLM Serving with PD-Multiplexing

FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving

optional XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models

NanoFlow: Towards Optimal Large Language Model Serving Throughput

optional Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve

LLM Training

FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

optional FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning

optional FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision

ZeRO: Memory Optimizations Toward Training Trillion Parameter Models

optional PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel

optional Everything about Distributed Training and Efficient Finetuning

Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures

optional DeepSeek Open Infra

Gemini: Fast Failure Recovery in Distributed Training with In-Memory Checkpoints

optional Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning

How to give a good talk

Slides and advice adapted from Thorsten Hoefler (ETH Zurich) and Simon Peter (UWashington)

Talking about research

- A good researcher can express their knowledge well
  - This assumes we have that knowledge!
  - So only advance to this step once you understand the paper :)
- Why is talking about research useful?
  - Order your thoughts, think about how to explain them
  - Communication with other researchers (your classmates!)
  - Gather feedback
  - Establish relationships
  - Eventually build a career

Why do you care?

- Presenting will be important for your career!
  - This is a good way to convince people to:
    - Give you good grades
    - Give you a job/promotion
    - Give you money/resources
    - Think you’re smart
    - Like you, recommend you
- But presentation skills are hard to acquire
  - No one is good at this right away
  - Practice practice practice!

A good research talk

- Is centered around the audience (not you)
  - **Teaches**, engages, provokes, excites listeners
- Provides intuitions to the audience
  - Take away messages, surprises, “wow” effects
- It does not need to
  - Tell every detail (not possible anyway in the time limit)
  - Show off how smart you are

Tip: focus on clearly defined goals

- Pick your goals carefully
  - What do you want to communicate? What should people remember?

Anatomy of a talk

- Motivation, placement
  - 20%
- Key ideas
  - 70-80%
- Evaluation results
  - 0-10%
- Do not present results without an explanation
  - Don’t just say “the authors made Application A run 50% faster”
  - Instead say “the authors present the FOO method relying on intuition BAR and that achieves 50% improvement for Application A”

The beginning of your talk

- You have about 2 minutes before your audience dozes off or starts reading email
    - Use them! Make every second count
    - Good approach: Present an abstract of your talk (or an elevator pitch):
        - Problem / motivation
        - Approach / idea
        - Experiments / results
        - Broader meaning / impact
- Answer these questions within two minutes:
    - What is the problem?
    - Why is this talk interesting? Why should I listen?

Communicating the key idea

- Pick a goal for the talk
  - Plan and make key points in your head. Organize the whole talk around these key points. Pick no more than three (better: one)
  - Be explicit, very explicit: “If you remember nothing else from this talk, remember this”
  - Repeat, repeat (but don’t be annoying)
- Do NOT be shallow, be deep
  - Avoid overviews
  - Do NOT ramble
  - Get to the meat quickly (assume we’ve all read the paper already)

Examples

- Are your main weapon
- Make your own examples, try to avoid just using the paper’s
- Ideally have a motivating example at the beginning
  - Maybe pose a question to get the audience thinking:
    - e.g., “What is the maximum speedup if we can solve this problem?”
- Illustrate the idea in action
  - From different perspectives
  - Show corner cases, highlight shortcomings
- Images say more than 1000 words!

What to use

- Enthusiasm
  - Be excited, pull the audience with you
  - Don’t be afraid to move around, but don’t pace or fidget
- Your brain!
  - Review/polish your slides before the talk
  - Have the storyline down
  - Focus on the key ideas
- Animations and graphics
  - Can be helpful (aesthetically as well as informative)
  - Nice slides make people more receptive

What to omit

- Do not present excessive related work
  - You can mention it in your slides, or have backup slides
  - Give credit
- Do not present too many technicalities
  - Audience probably won't follow
  - Put details in the backup slides in case you do get a question on them
- Do not exaggerate with animations
  - Animations are good, but too many are difficult to follow
- Do not clutter your slides

How to present

- Be (or at least appear) confident
  - Don’t forget to breathe! :)
- Make eye contact
  - Look around, don’t stare at a single person
  - Tip: identify a nodder (these people always exist). They’ll give you confirmation
- Watch the audience
  - Sometimes they ask questions, don’t let them interrupt you but serve their questions
  - Questions are great, ask some and answer them!
- Finish on time!
  - Skip slides if necessary, never ask “should I continue?”
    - No polite person would ever say, “no, thanks”

Miscellaneous

- Standard stuff
  - Aviod erorrs no sldies
  - Face the audience
  - Check your laptop before
- Practice, practice, practice
  - Give your talk a couple of times before you present it publicly
  - Present to your group mates, me during office hours, your friends, your cat…
- You’ll attend more talks here than you’ll give
  - Engage, help your fellow students!
  - Ask questions, participate in the discussion, be awake

Miscellaneous

- You may borrow slides
    - But, be aware of the context of the slides
        - E.g., conference slides are intended for a very specialized audience
    - May have to adapt the slides for a more general audience (and the time frame)
    - And give credit
- Timing and cadence
    - Between 1-2 minutes per slide is common
    - Try not to rush too quickly through slides or spend too much time on a slide

Next Tuesday, 9/30

Paper Presentation Selection Due