THE OHIO STATE UNIVERSITY

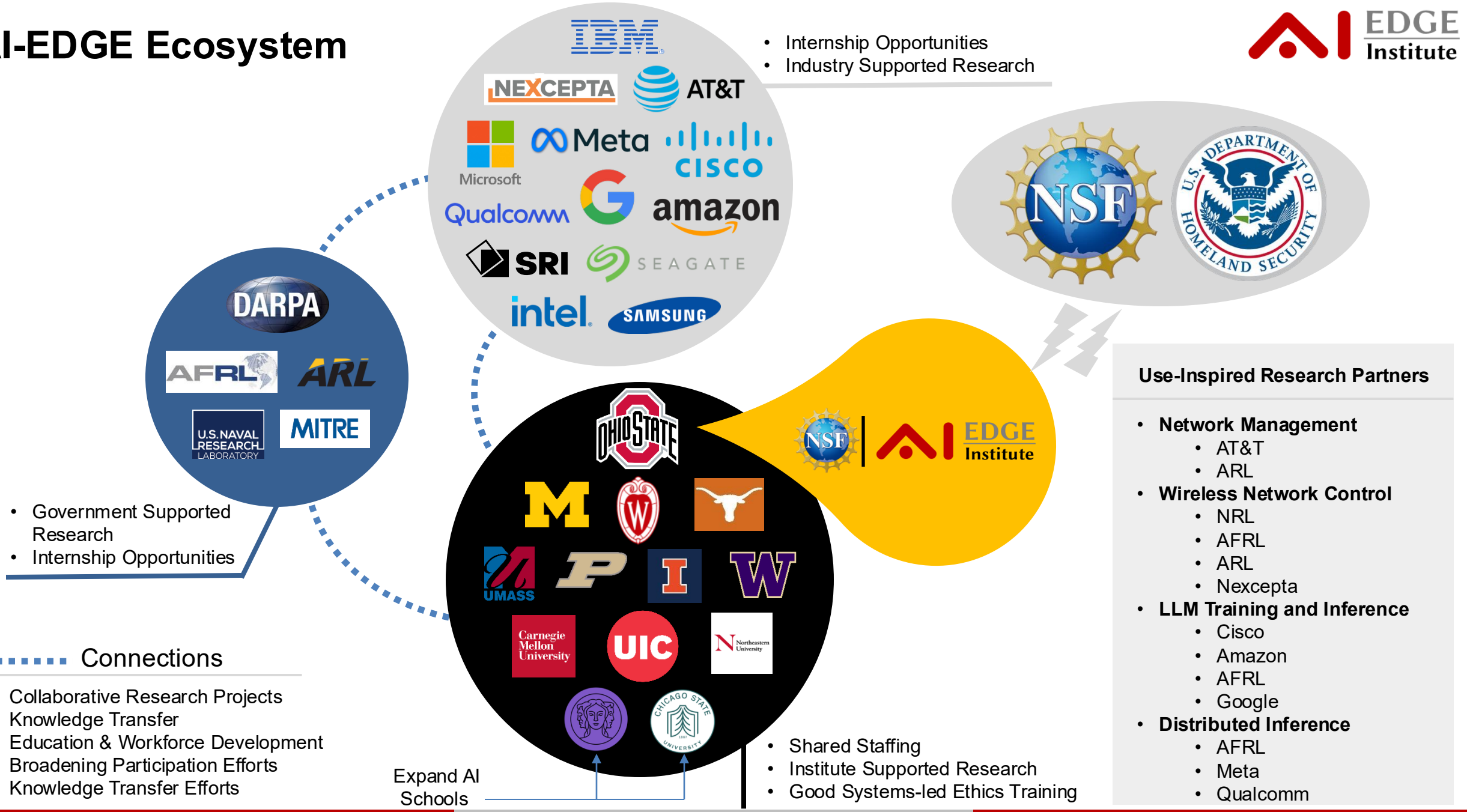
NSF AI INSTITUTE FOR FUTURE EDGE NETWORKS AND DISTRIBUTED INTELLIGENCE

Federated Multi-Objective Learning: Democratizing LLMs with the Edge

Jia (Kevin) Liu
The Ohio State University

AI EDGE Institute

AI-EDGE Ecosystem



- Internship Opportunities
- Industry Supported Research

- Government Supported Research
- Internship Opportunities

Connections

- Collaborative Research Projects
- Knowledge Transfer
- Education & Workforce Development
- Broadening Participation Efforts
- Knowledge Transfer Efforts

Expand AI Schools

- Shared Staffing
- Institute Supported Research
- Good Systems-led Ethics Training

Use-Inspired Research Partners

- **Network Management**
  - AT&T
  - ARL
- **Wireless Network Control**
  - NRL
  - AFRL
  - ARL
  - Nexcepta
- **LLM Training and Inference**
  - Cisco
  - Amazon
  - AFRL
  - Google
- **Distributed Inference**
  - AFRL
  - Meta
  - Qualcomm

Organization and Key Personnel: Academia

EDGE Institute

PI: Ness Shroff

Expertise: Net. Theory, Bandits, RL, Optimization, Algorithms, MDP, Games

Co-PI: Elisa Bertino

Expertise: Information Security, Database, Privacy and Trust

Co-PI: Gauri Joshi

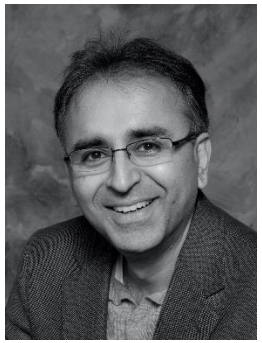
Expertise: Distributed Learning, Bandits, Bayesian Optimization

Co-PI: Jim Kurose

Expertise: Computer Network Arch. & Protocols, Network Measurements

Co-PI: Rob Nowak

Expertise: Machine Learning, Stat. Signal Processing, Statistics



SP: Anish Arora

Expertise: Network Systems Scalability & Dependability

SP: Kaushik Chowdhury

Expertise: Network Systems, 5G, Protocols, Experiments At-Scale

SP: Mingyan Liu

Expertise: Net. Resource Allocation, Sequential Decision Theory

SP: Sanjay Shakkottai

Expertise: Net. Optimization, Stat. Learning and Wireless Communication

SP: Arnob Ghosh

Expertise: Reinforcement Learning, Bandits, MDP

SP: Raef Bassily

Expertise: Privacy-Preserving Data Analysis, ML, Optimization, Info. Theory

SP: Constantine Caramanis

Expertise: Decision Making in Complex Systems, High Dim. Statistics, Optimization

SP: Eylem Ekici

Expertise: Cognitive Radio, Vehicular Communication, Net. Resource Management

SP: Atilla Eryilmaz

Expertise: Stochastic Network Optimization, Bandits, Control

SP: Stratis Ioannidis

Expertise: Distributed Systems, Networking, Optimization, ML, Privacy

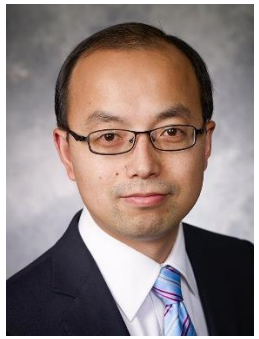
Organization and Key Personnel: Academia

SP: Nan Jiang

Expertise: Reinforcement Learning, Online Learning

SP: Yingbin Liang

Expertise: Info. Theory, Wireless Communications, Optimization, Statistical SP



SP: Zhiqiang Lin

Expertise: Security, Trusted Computing, Program Analysis,

SP: Jia (Kevin) Liu

Expertise: ML, Distri. Optimization, Stochastic Network Optimization,

SP: Tommaso Melodia

Expertise: Wireless Networks, Cognitive Radio Experiments at Scale

SP: Aryan Mokhtari

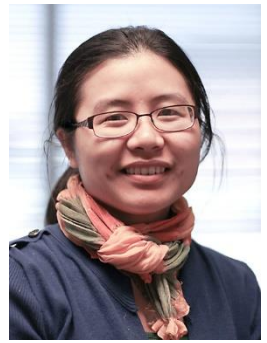
Expertise: Convex and Non-convex Optimization, Large-scale ML & Data Science

SP: Sewoong Oh

Expertise: Theoretical ML, Robust Statistics, Social Comp., Diff. Privacy

SP: Srini Parthasarathy

Expertise: Data Analytics, Graph Analytics, Network Science ML, Database Systems



SP: Chunyi Peng

Expertise: Mobile Networking Systems, Security, 5G, 6G Systems

SP: Hulya Seferoglu

Expertise: Coded Comp., IoT, Anomaly Detection in Video Streaming

SP: Kannan Srinivasan

Expertise: WirelessSys, Protocols, Measurements, Communication Security

SP: Aylin Yener

Expertise: Info. Theory, Cybersecurity, Wireless Comm., Optimization, Learning

SP: Lei Ying

Expertise: Complex Stochastic Systems, Big Data, Graph Data Mining

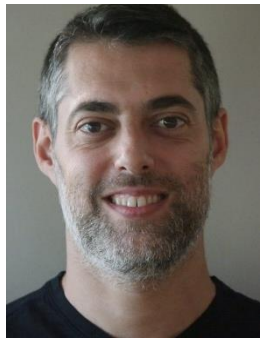
SP: Jie Wei

Expertise: ML, remote sensing Multimodal computing, Medical imaging

Organization and Key Personnel: Industry/DoD

Microsoft: Victor Bahl

Expertise: Edge Comp., 5G, Mobile Computing, Wireless Sys., Cloud Comp.



IBM: Lior Horesh

Expertise: Optimization, Applied Inverse Problems, Large-Scale Simulations, ML

NRL: Clement Kam

Expertise: Information Freshness, Scheduling, Cognitive Radio, ML, AoI

Qualcomm: Junyi Li

Expertise: Wireless Communication, Mobile Broadband, OFDMA

AT&T: Milap Majmundar

Expertise: Mobile Netw., Radio Access Network, Spectrum Strategy

AFRL: Chris Myers

Expertise: Computational Cognitive Models for Complex Tasks

AFRL: Lee Seversky

Expertise: Autonomy, Command & Control Systems

IBM: Mark Squillante

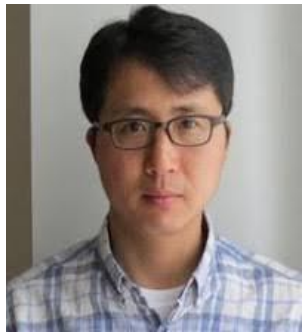
Expertise: Mathematical Foundation of Complex Sys. Modeling and Analysis

Mitre: Venki Ramaswamy

Expertise: Cellular Nets, 5G Mobile, Blockchains, AI/ML

Nexcepta: Sastry Kompella

Expertise: Network Optimization, Scheduling, Cognitive Radio, ML, AoI



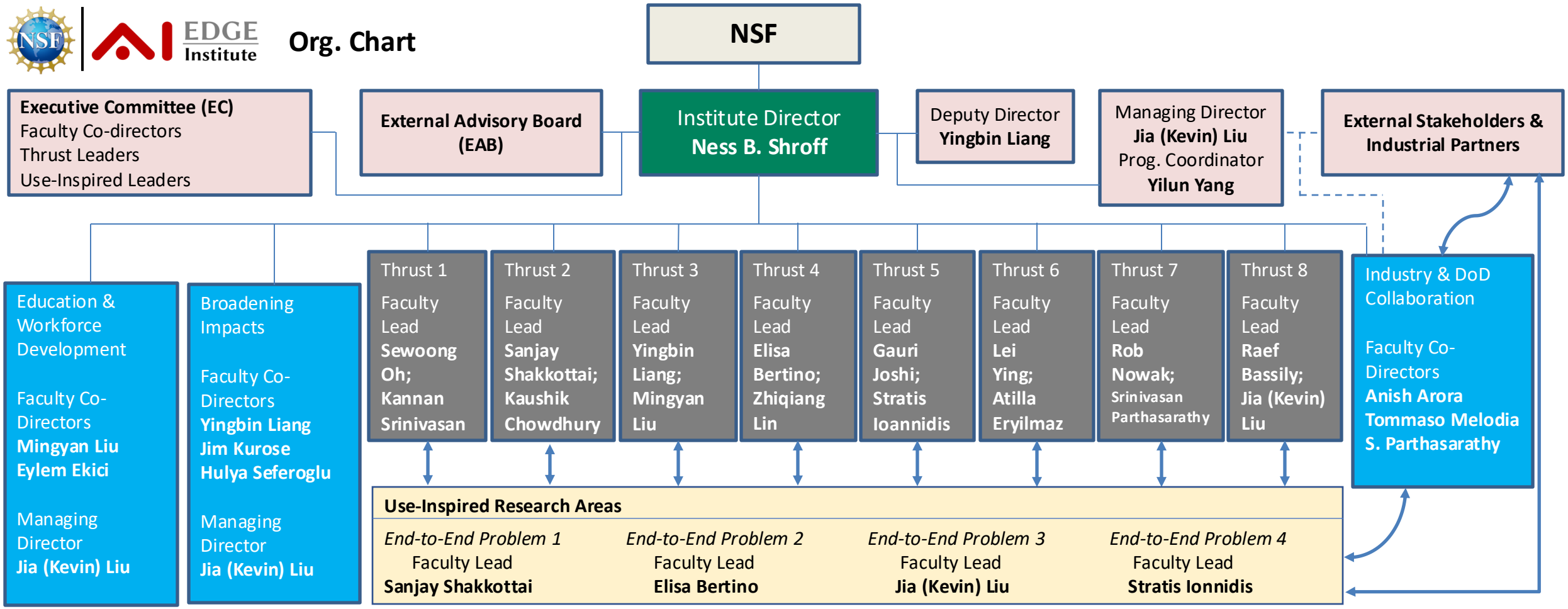
Cisco Myungjin Lee

Expertise: Network AI infrastructure, Federated Learning, Scheduling

ARL: Anathram Swami

Expertise: Network Science, Signal Processing, Wireless Communications

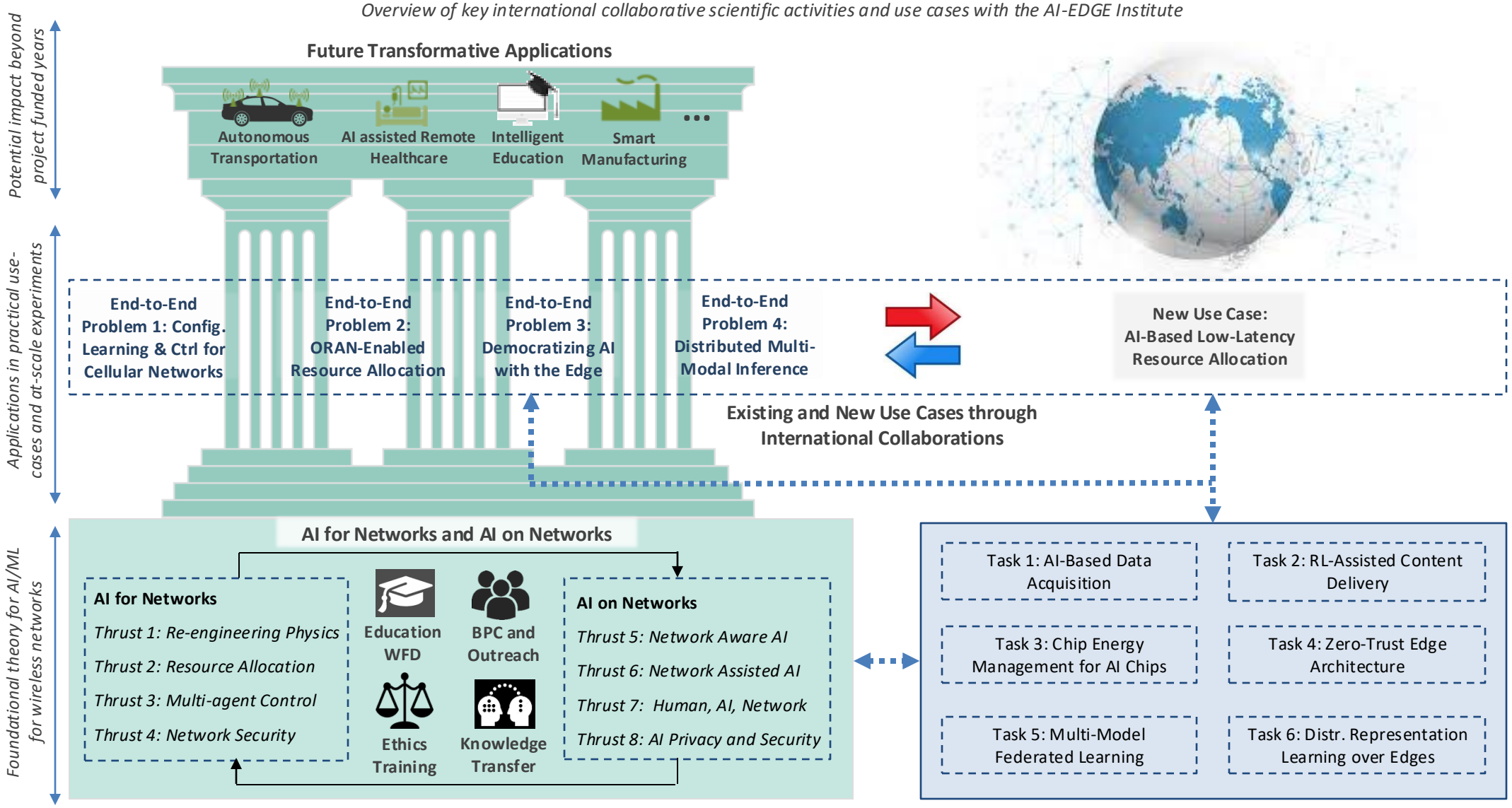
Management Structure



AI-EDGE goes Global

AI-EDGE Institute International Collaborations

Overview of key international collaborative scientific activities and use cases with the AI-EDGE Institute



AI-EDGE's Current Research Thrusts

New Proposed Tasks through International Collaboration

International Partners

AI EDGE Institute

Indian Institute of Technology Bombay

서울대학교 SEOUL NATIONAL UNIVERSITY

IIT Madras

KAIST AI Graduate School of AI

성균관대학교 SUNG KYUN KWAN UNIVERSITY

KOREA UNIVERSITY

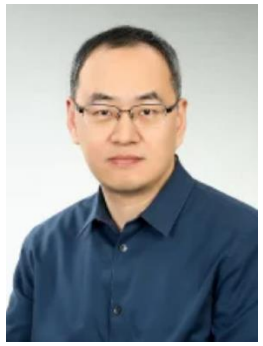
YONSEI UNIVERSITY 연세대학교

New Research collabs. in 2025

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

टामूअसं tifr
Tata Institute of Fundamental Research

International Collaborators



Korea Univ.: Changhee Joo

Expertise: Wireless Netw., ML, network economics, content delivery networks

KAIST: Song Chong

Expertise: Wireless Netw., mobile comput., ML, learning-based adaptive network control

SNU: Kyunghan Lee

Expertise: Mobile network and computing, learning for application layers

IIT Bombay: D. Manjunath

Expertise: Stoch. model for network systems, economics of CDN

IIT Bombay: Nikhil Karamchandani

Expertise: Networking, information theory, online learning, stat. inference

IIT Bombay: Jayakrishnan Nair

Expertise: Modeling, performance eval., online learning, queueing systems, comm. networks

IIT Bombay: Gaurav Kasbekar

Expertise: Comm. network, network security, game theory, privacy, ML

IIT Bombay: Sharayu Moharir

Expertise: Communication, networking, online learning, fed. learning

IIT Madras: Balaraman Ravindran

Expertise: RL for social networks, data mining, learning based network analysis

Yonsei University: Jang-Won Lee

Expertise: Networking, Optimization, Stochastic Control Cyber-physical Systems

Sungkyunkwan University: Hyunseung Choo

Expertise: Comm. network, Internet of Things, Cloud Computing, ML

IIT Hyderabad: Bheemarjuna Reddy Tamma

Expertise: Radio Access technologies, M2M, IoT, Network Security

AI-EDGE End-to-End Problem

Democratizing LLMs with the Edge

Key Team Members (29)

Ohio State (15):

Kevin Liu, Ness B. Shroff, Yingbin Liang, Srini Parthasarathy, Aylin Yener, Ziyue Luo, Atena Nourzad, Xue Zheng, Rohith Sudha Krishnan, Jiaxuan Cai, Peiwen Qiu, Srijith Nair, William Wu, Sungjae Lee (ICICLE), Yinglun Xia

Wisconsin (3):

Rob Nowak, Jifan Zhang, Rishabh Sharma

UT Austin (3):

Sanjay Shakkottai, Kaushik Chowdhury, Sundar Srinivasan

Carnegie Mellon (2):

Gauri Joshi, Siddharth Shah

RIT (2):

Haibo Yang, Zhe Li

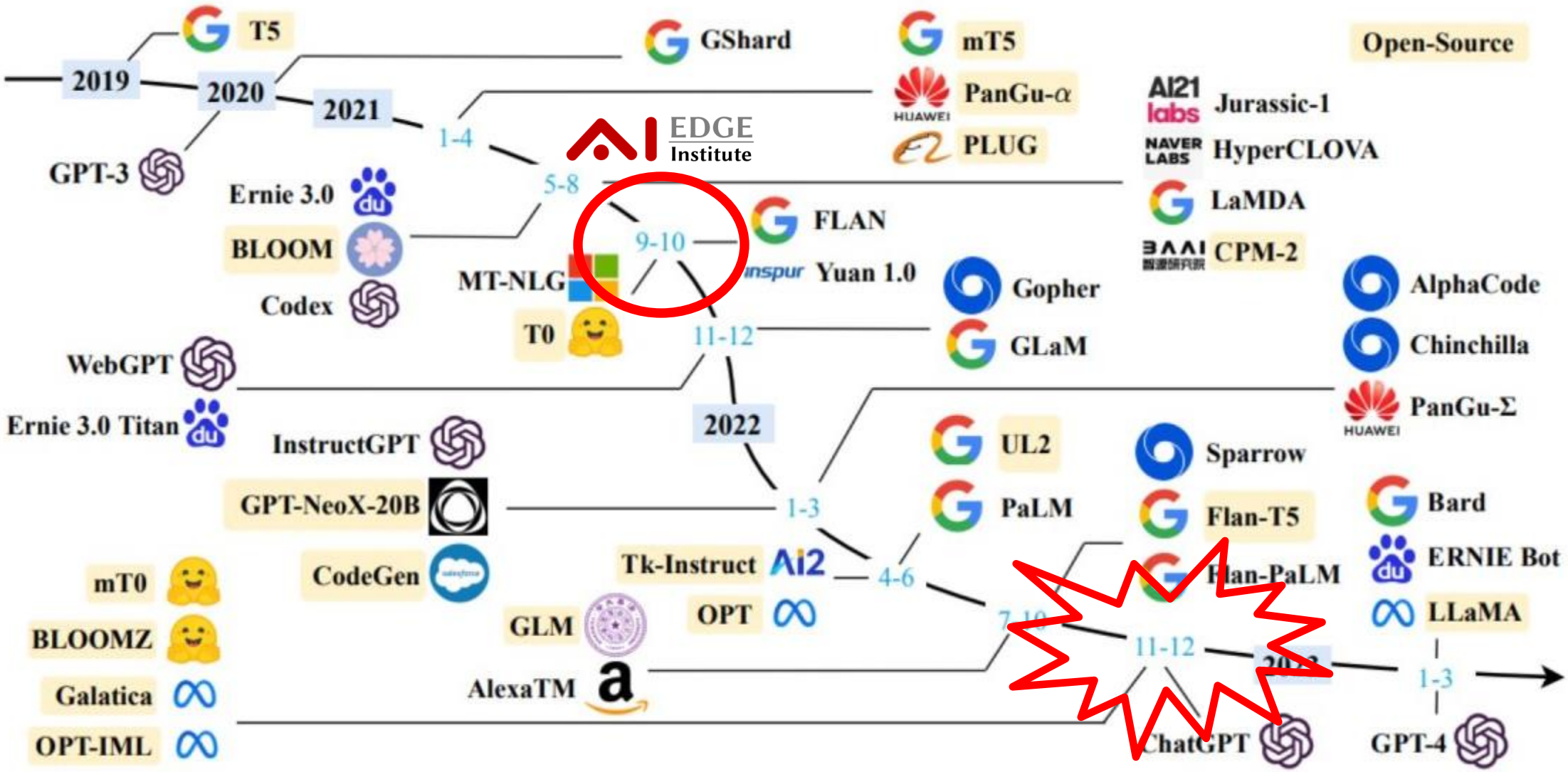
IIT Madras (2):

Saurav Prakash, B. Ravindran

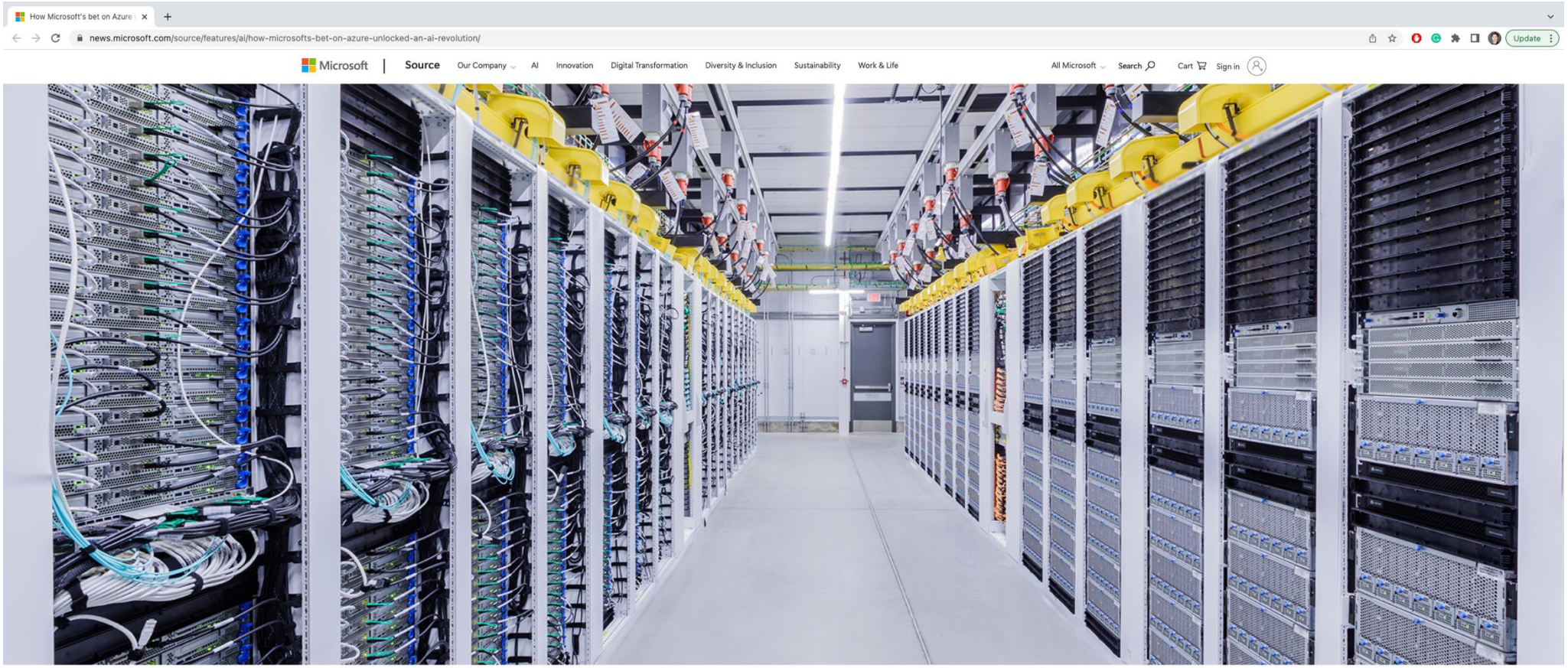
Industry (2):

Bicheng Ying (Google), Zidong Liu (ComboCurve)

LLMs Take the World by Storm



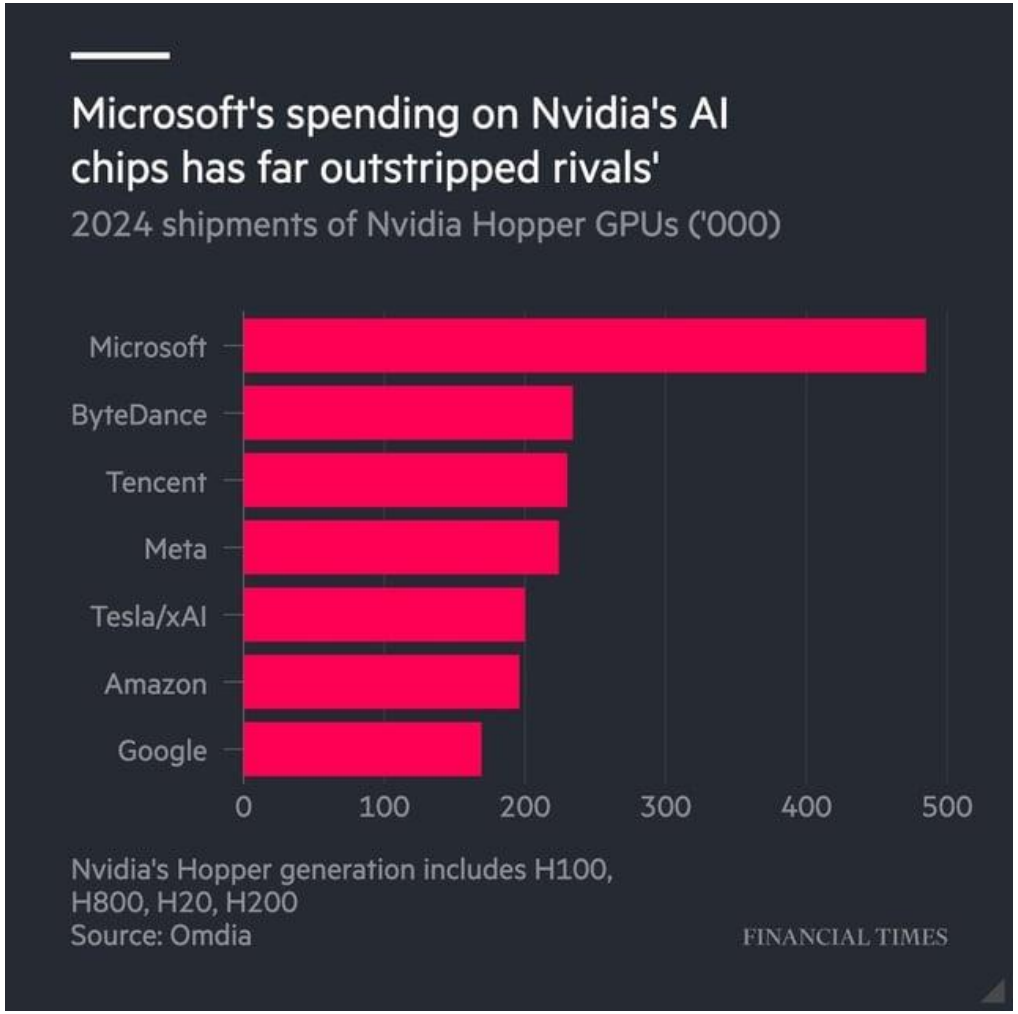
At Microsoft…

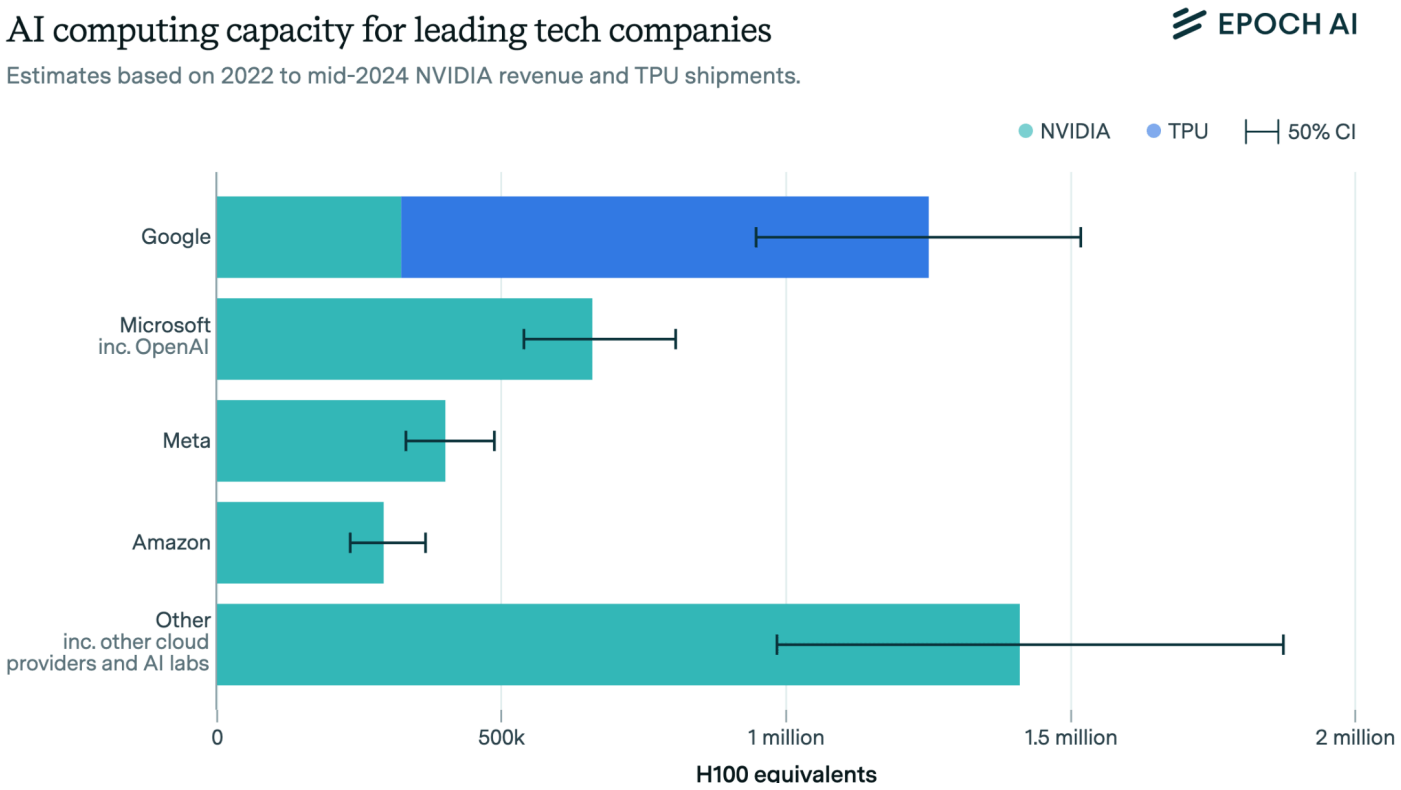


How Microsoft's bet on Azure unlocked an AI revolution

About five years ago, artificial intelligence research organization OpenAI pitched Microsoft on a bold idea that it could build AI systems that would forever change how people interact

Big Tech's Spending Frenzy on AI Infra



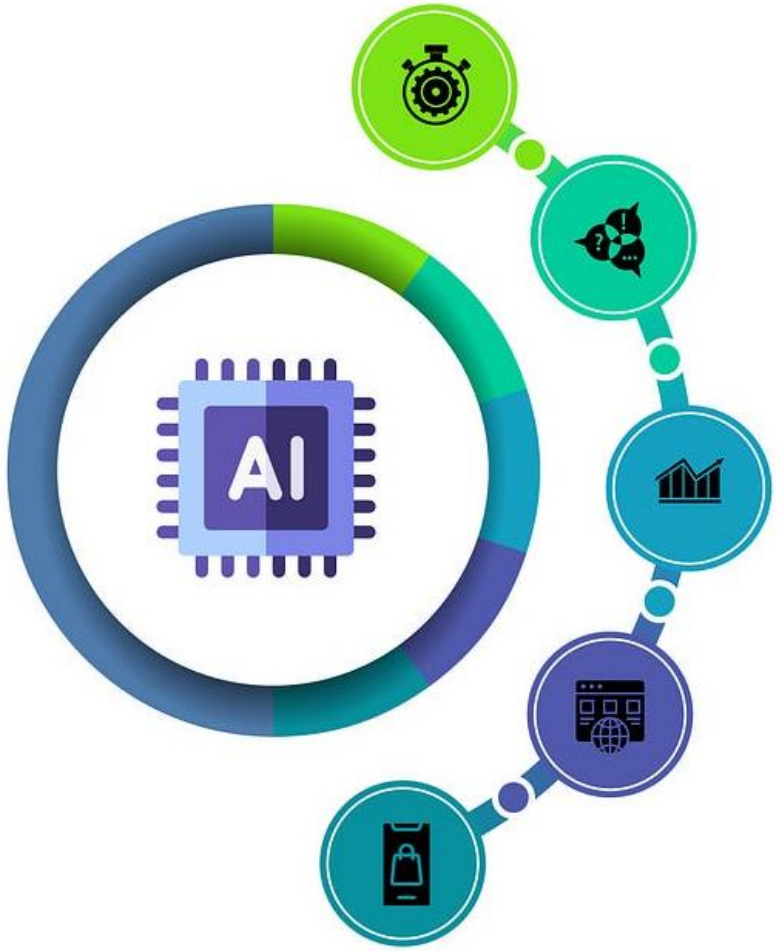


A Compelling Need for Democratizing LLMs

- Growing concerns about domination of big tech giants in **training** LLMs
  - Non-transparency of state-of-the-art LLM technologies
  - Security and trustworthiness due to closed training processes
  - Environmental sustainability risks of centralized huge computing centers

Question: Can we open LLM training in a distributed network environment? (i.e., enabling LLM training to "anywhere and everywhere?")

Democratize LLM Training with the Edge!



LLM Training = Pretraining + Finetuning

Vision: A unified learning framework over broadly defined edge networks for both LLM training phases

Key Idea:

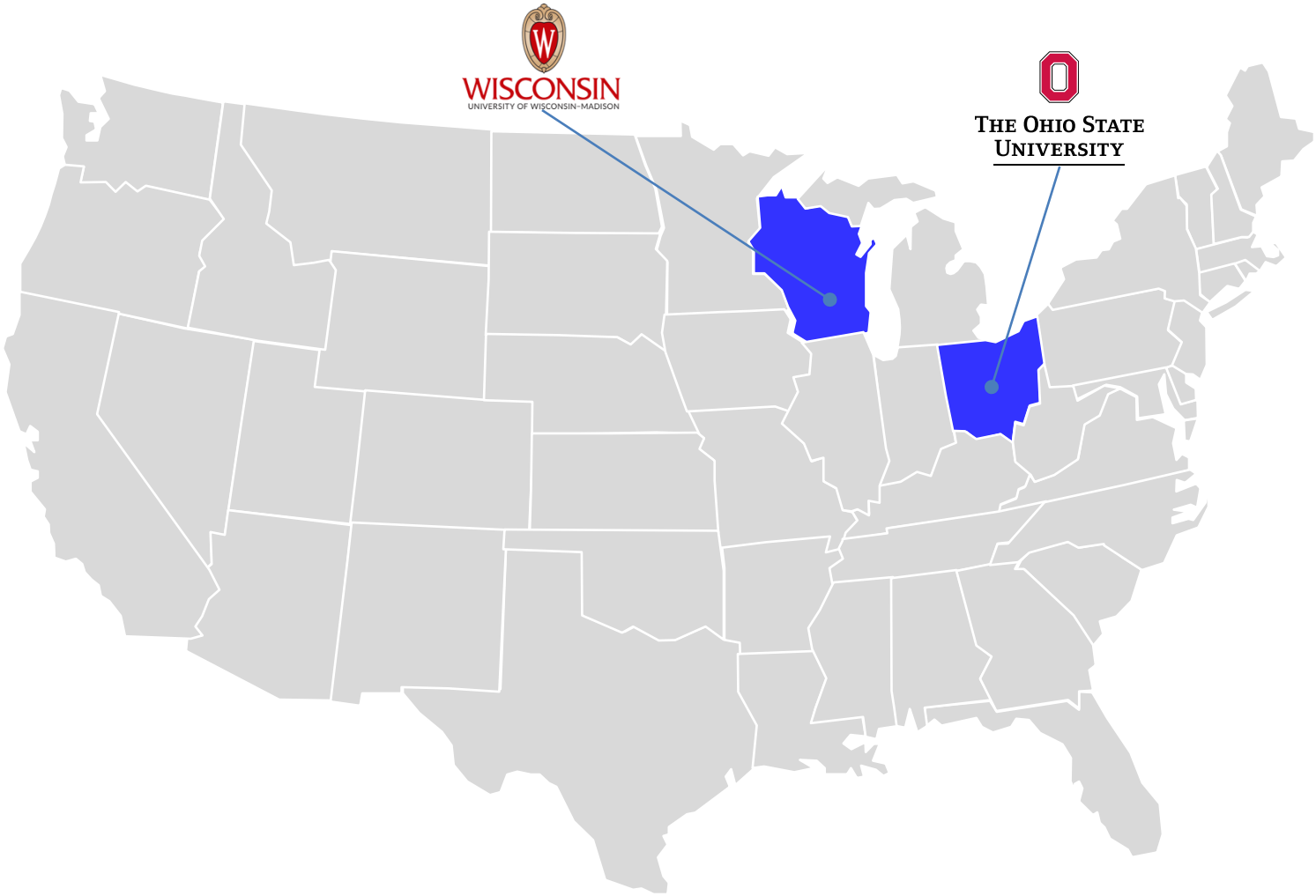
- Leverage lower-end but abundant, under-utilized, and spare GPUs across multiple institutions to perform both LLM pretraining and finetuning in a distributed fashion

Challenges of Democratize LLMs with the Edge!

- Lots of Networking Challenges to overcome:
  - Pretraining:
    - Handling large network delays
    - Transmitting a large amount of data over limited bandwidth
    - Handling servers and workers heterogeneity
    - Packet losses

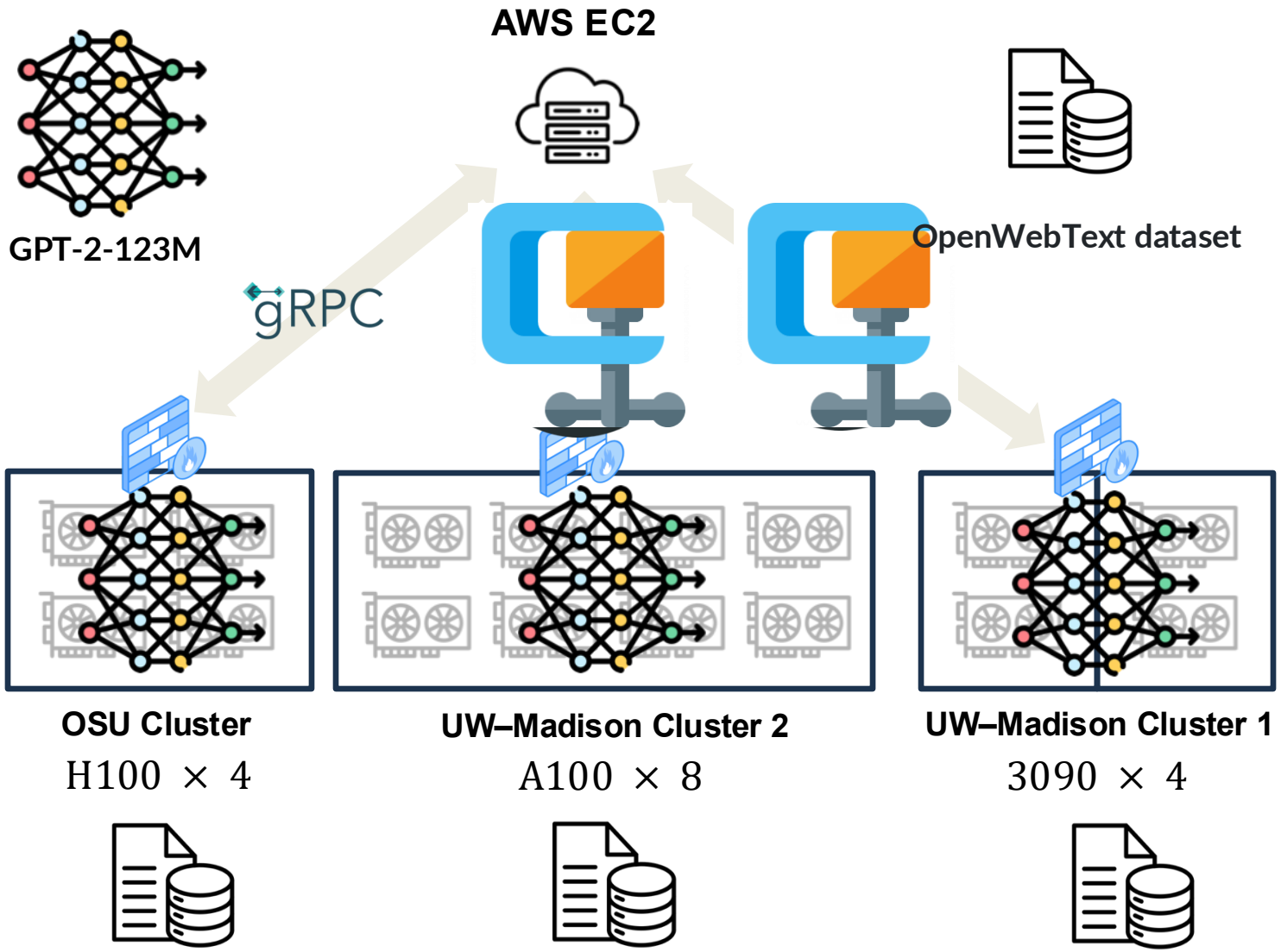
  - Finetuning:
    - In addition to similar challenges as in pretraining, we also need to manage interference, noise, and wireless resources if finetuning at the wireless edge

Proof of Concept 1: Distributed LLM Pretraining

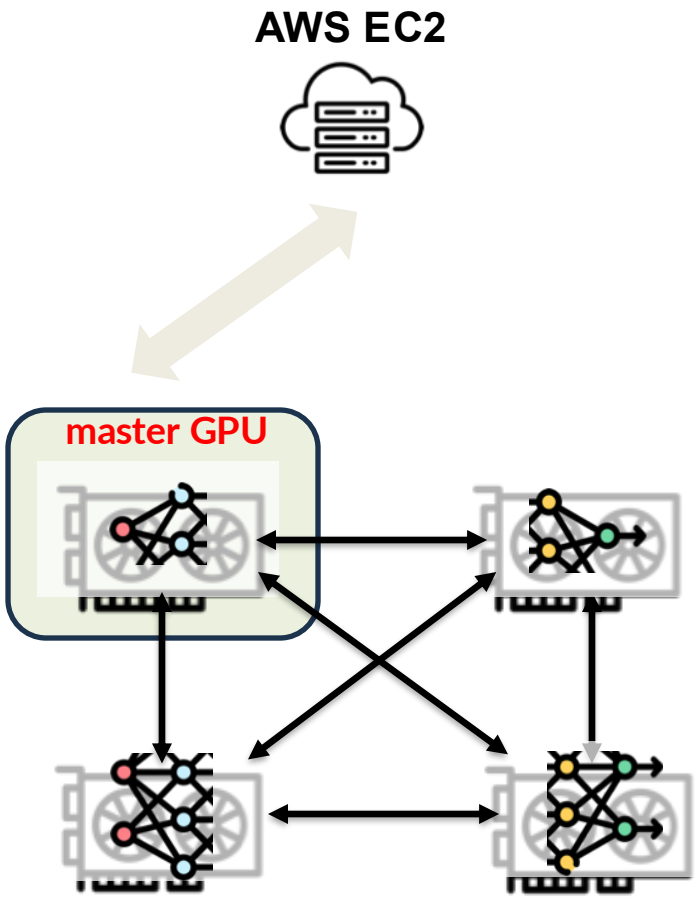


Started a Cross-Institution Pretraining System in Year 3

Cross-Institution Pretraining System



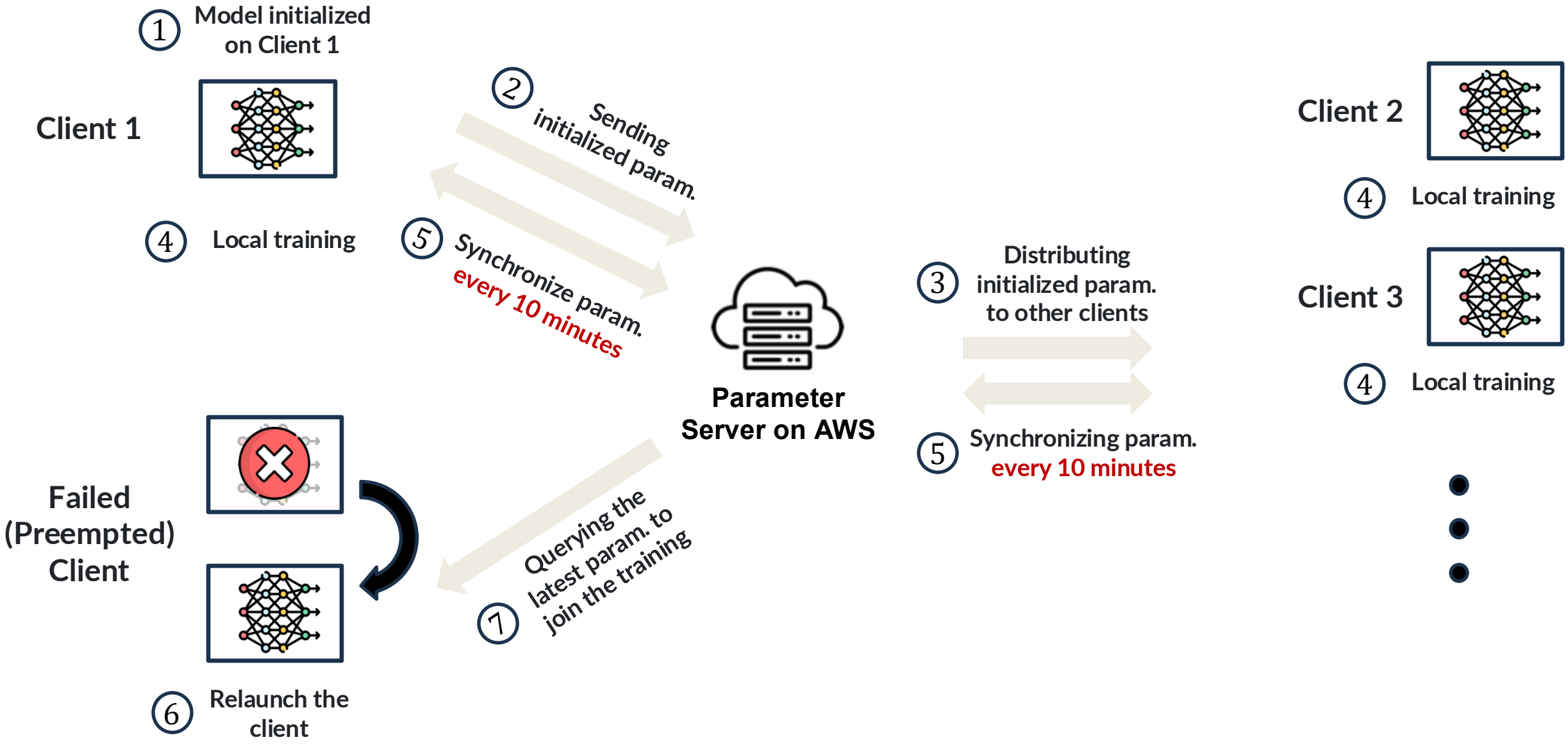
Cross-Institution Pre-Training System Architecture



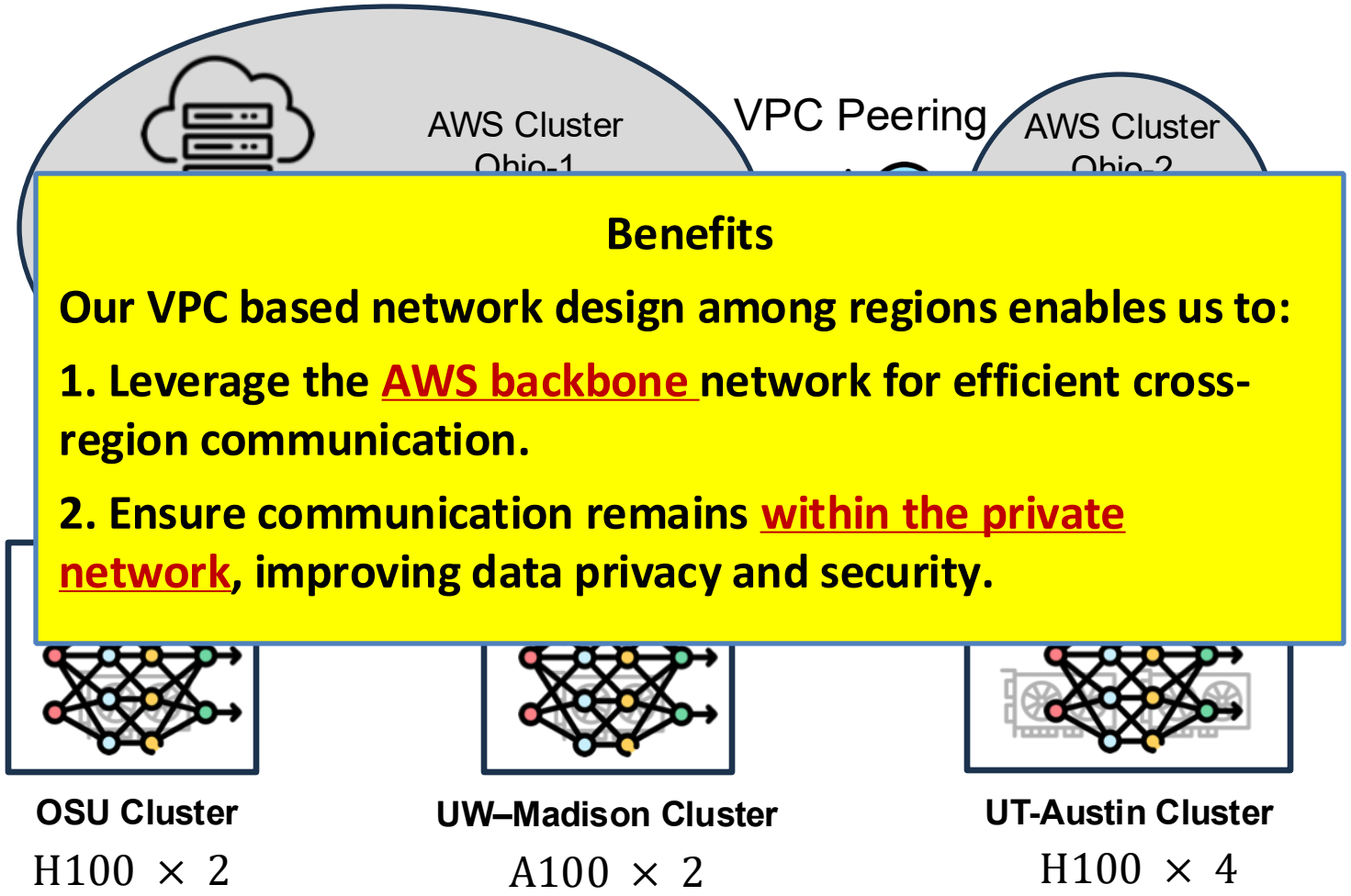
LLM is trained with
Fully Sharded Data Parallel (FSDP)
Within Each Cluster

Cross-Institution Pretraining System: The Algorithm

EDGE Institute



Idea 1: Hierarchical Synchronization

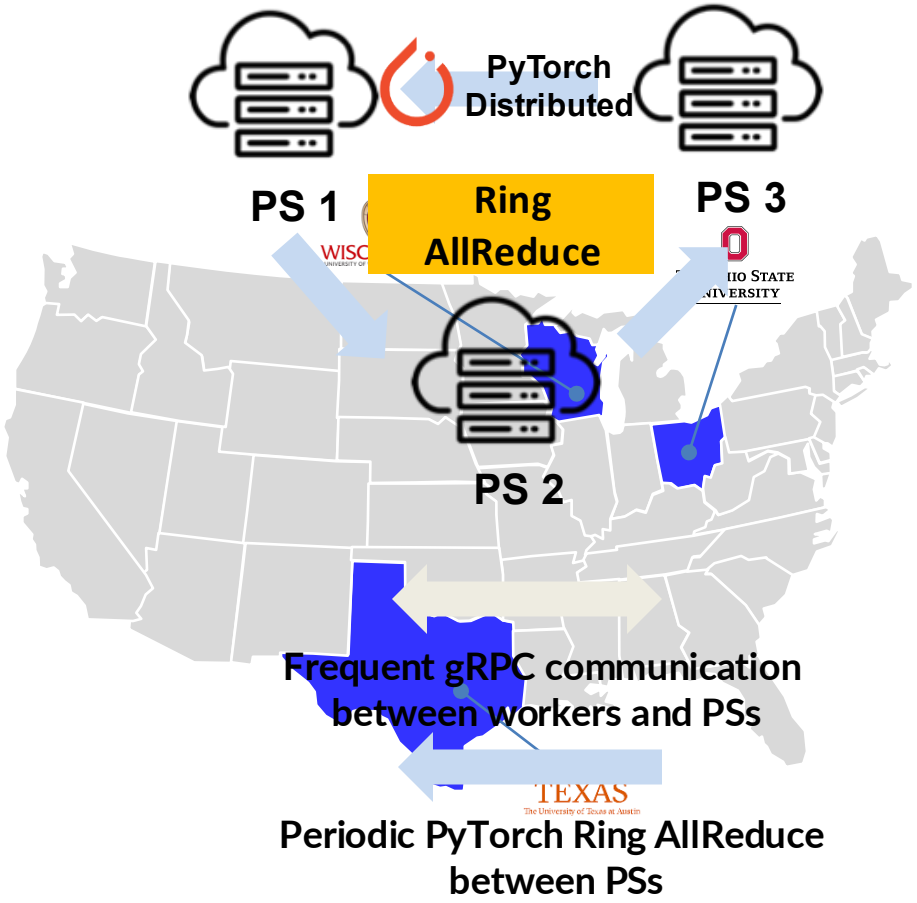


Benefits

Our VPC based network design among regions enables us to:

1. Leverage the AWS backbone network for efficient cross-region communication.

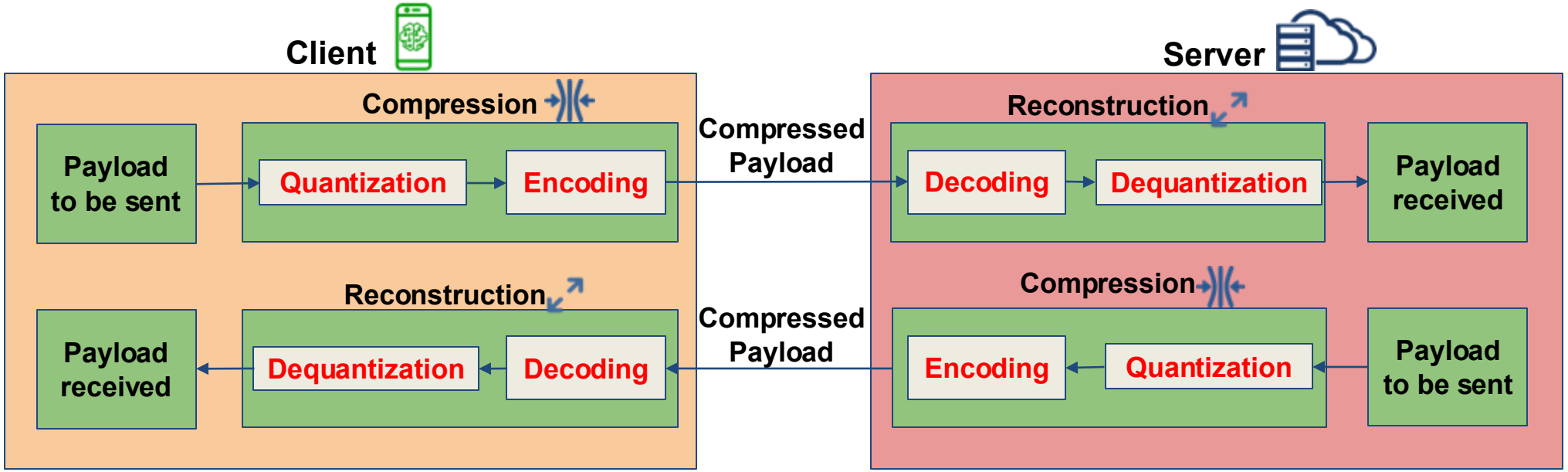
2. Ensure communication remains within the private network, improving data privacy and security.



Organize workers into groups through virtual private cloud (VPC) with near-by parameter servers (PSs) for fast intra-group sync, and let PSs to aggregate across groups on a slower cadence.

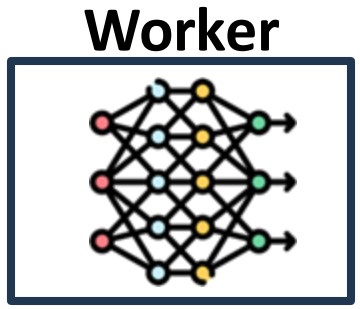
Idea 2: Two-Level Data Compression

- Motivation
  - Reduce payload sent across the network.



- A Two-Level Data Compression System
  1) Data quantization with lower precision floating point representations
  2) Coding compression based on quantized levels (transmit only codebook index)

Huffman-Based Gradient Compression

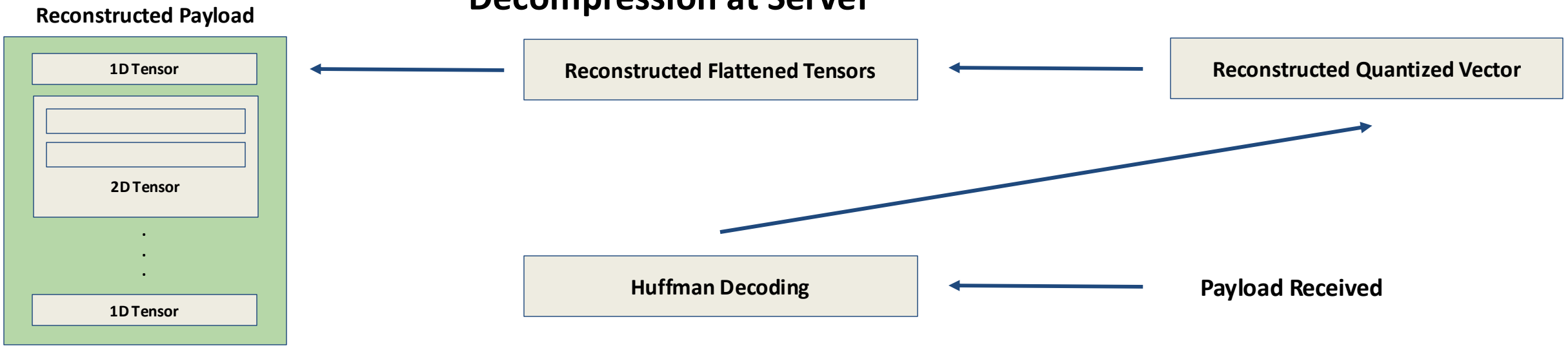


David A. Huffman (1925-1999)

BS, MS, The Ohio State University

DSc, Massachusetts Institute of Technology

Decompression at Server



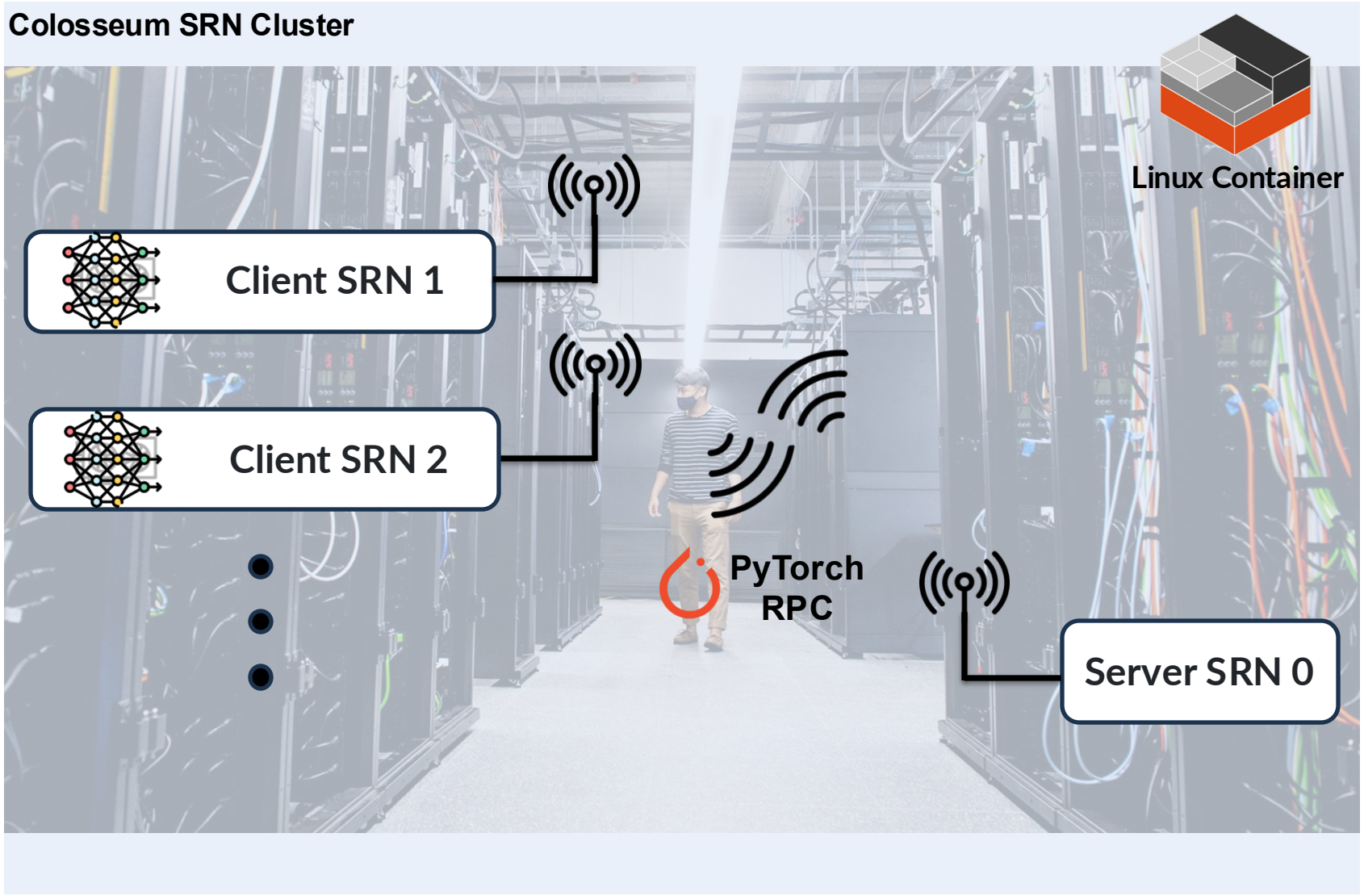
Video Demo of Distributed LLM Pretraining

Pretraining Video Demo

Proof of Concept 2: Distributed LLM Finetuning

A Colosseum-Based Finetuning System

Colosseum-Based Finetuning Design: Iteration 1



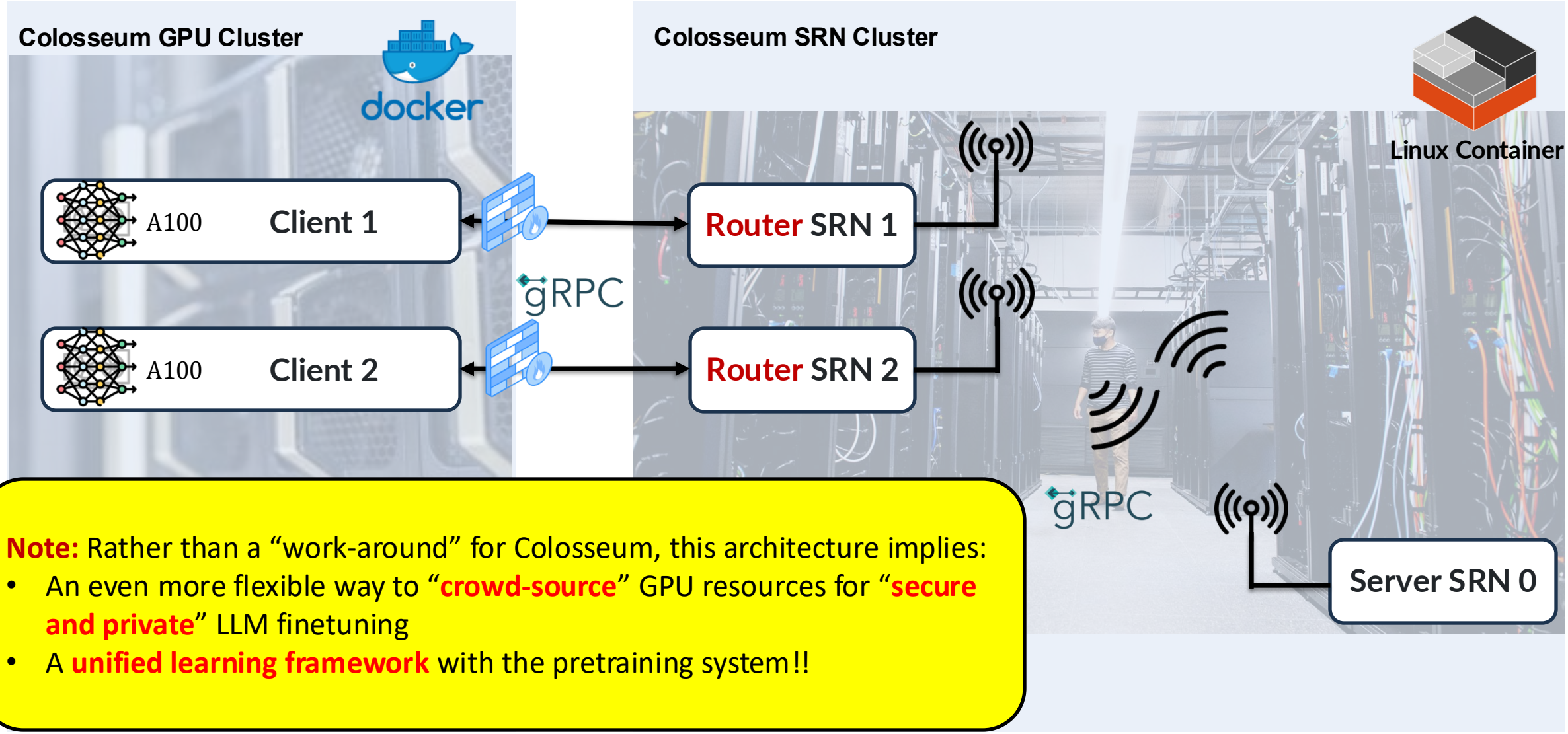
Benefits:

- Wireless communication enabled via SRNs.
- Utilizes PyTorch RPC for efficient communication.

Obstacles:

- SRN nodes have low-end and GPU software support is unsuitable for LLM training.
- Dedicated DGX GPU cluster is only connected to SRNs via wired connections.
- Firewalls between the GPU and SRN clusters block the use of PyTorch RPC.

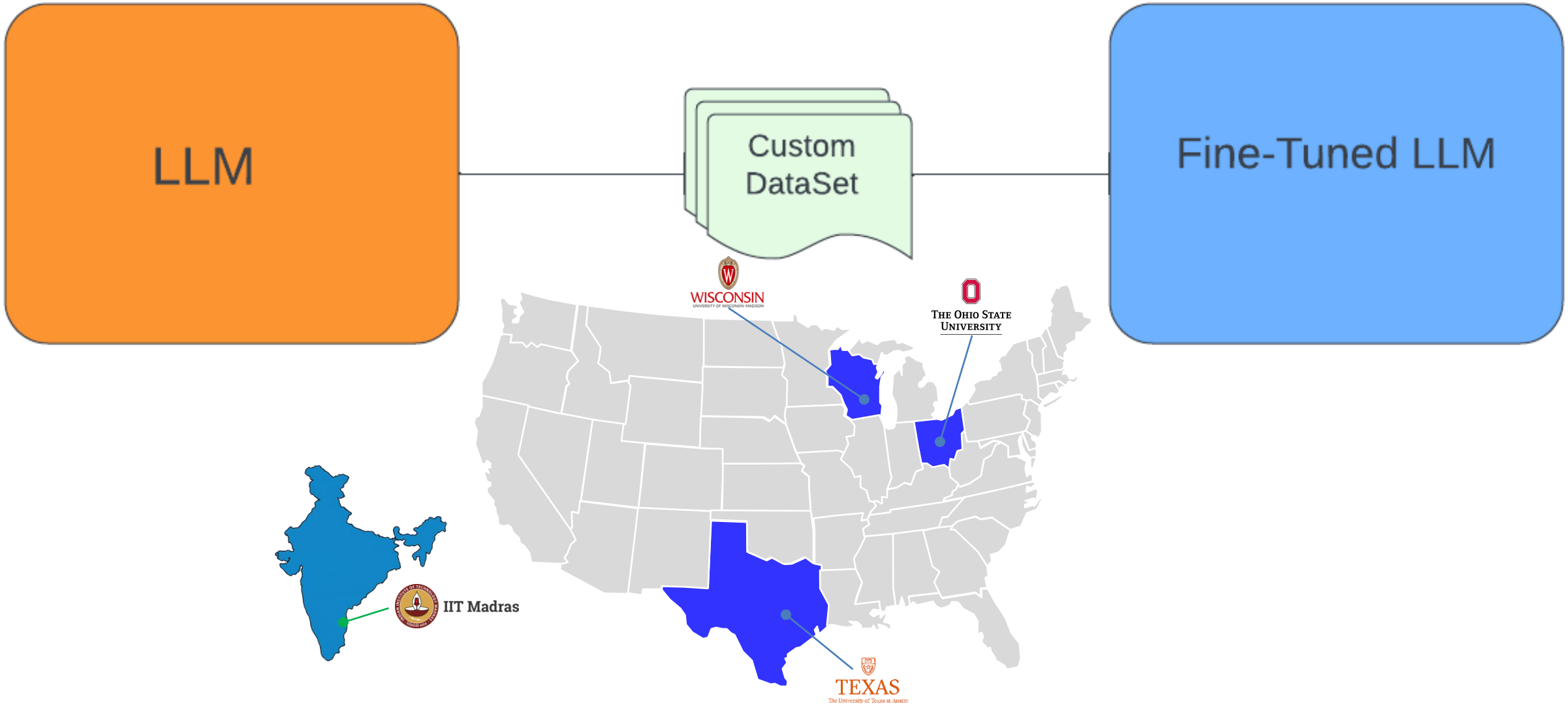
Colosseum-Based Finetuning Design: Iteration 2



Note: Rather than a "work-around" for Colosseum, this architecture implies:

- An even more flexible way to "**crowd-source**" GPU resources for "**secure and private**" LLM finetuning
- A **unified learning framework** with the pretraining system!!

An International Cross-Institution Finetuning System



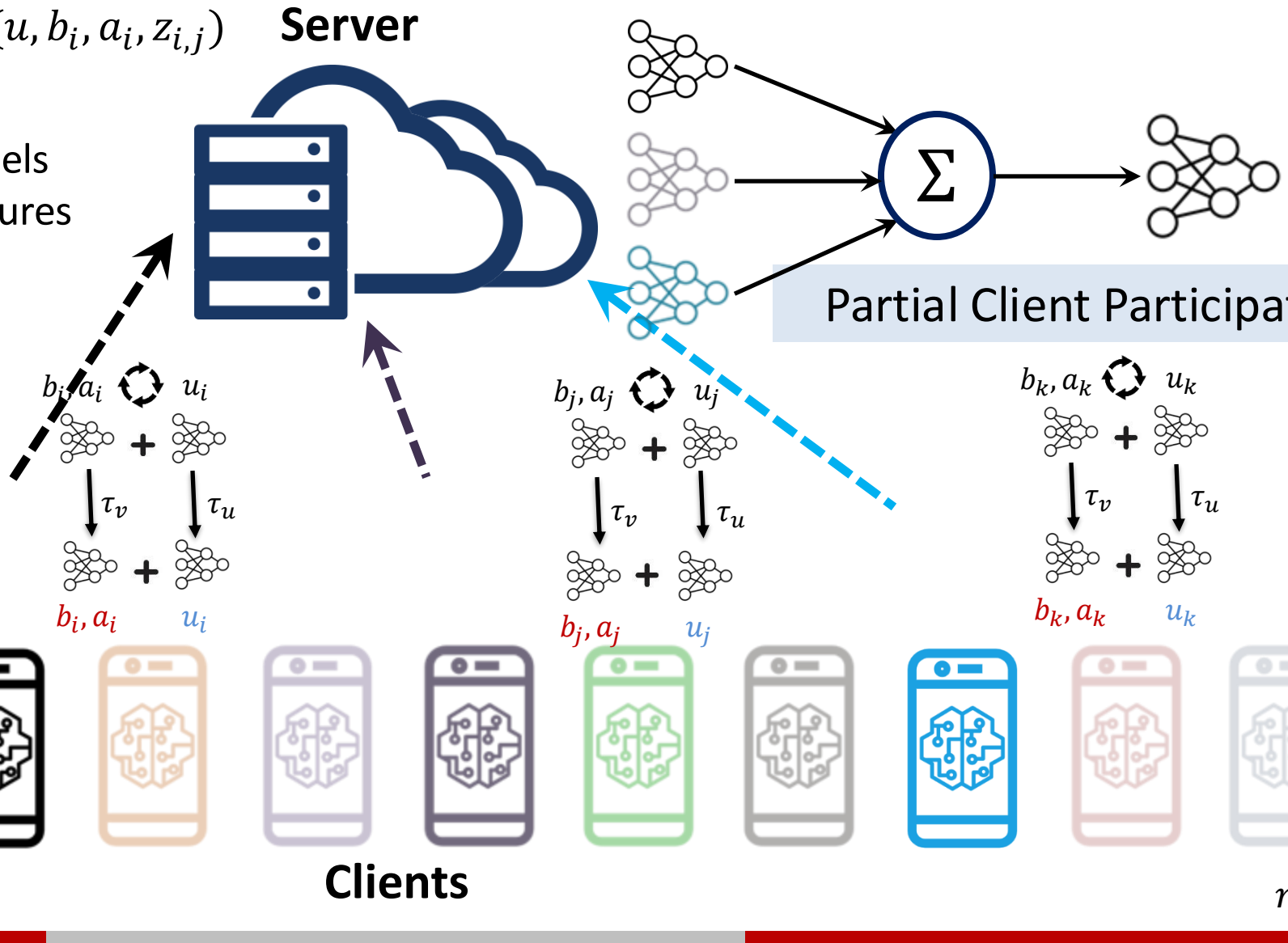
Hybrid LoRA in the Wild

$$\min_{u,\{b_i\}_{i=1}^n,\{a_i\}_{i=1}^n} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\frac{1}{N_i} f_i(u, b_i, a_i, z_{i,j})$$

- Server aggregates local LLM models
- Update private parameters (a,b) at a faster timescale
- Update shared parameters (u) at a slower timescale

Server: sends global model (u) to selected clients

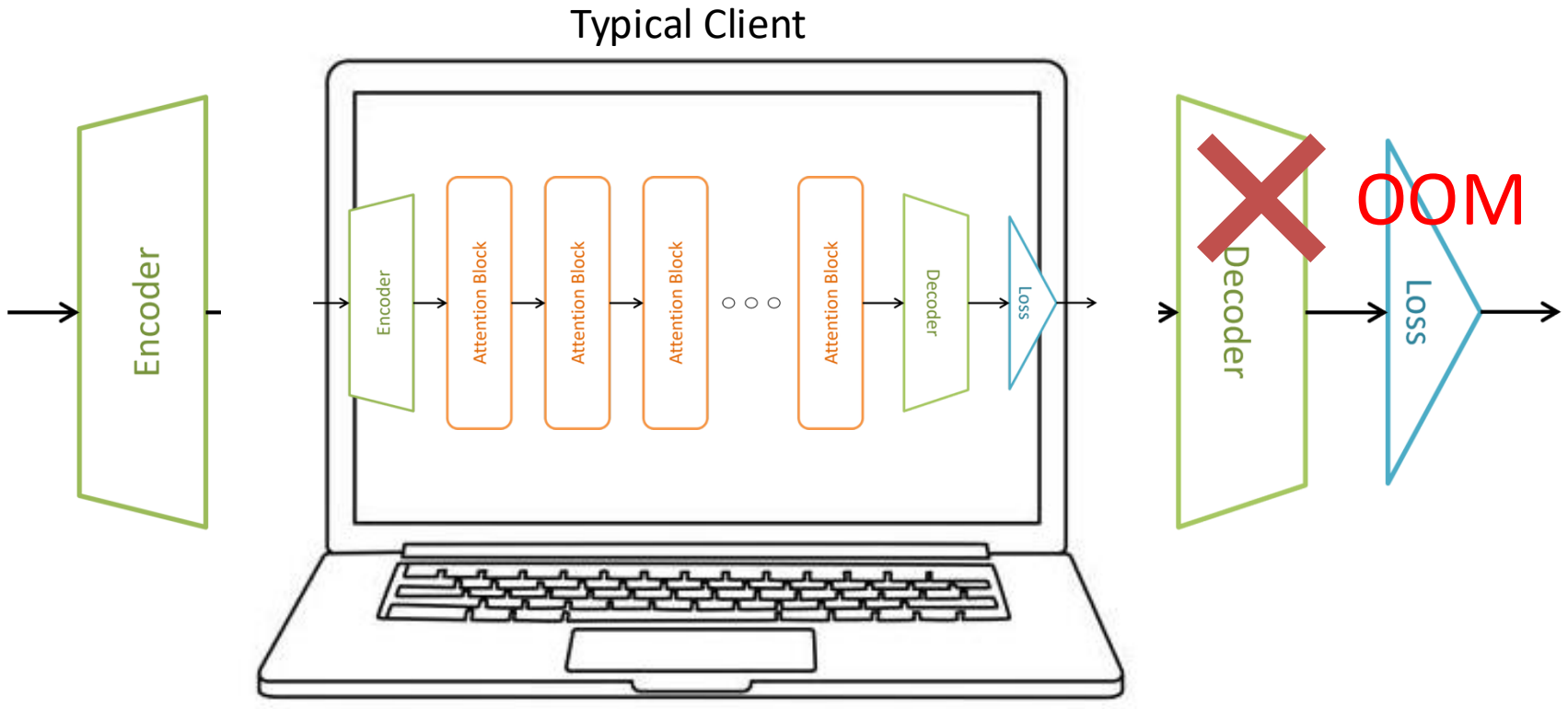
Clients: Keeping LoRA (b, a) local ensures privacy *preservation*. Model gets updated and finishes the current round.



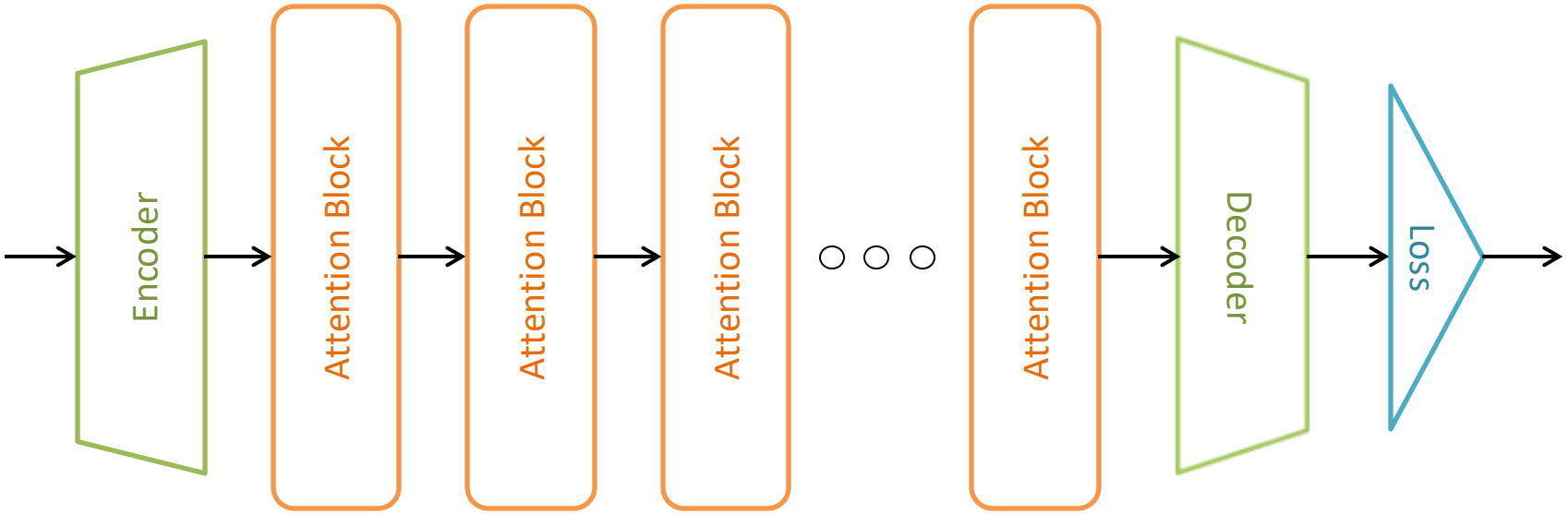
Limitation of FedAvg-Type Methods

LLM model size could easily exceed the storage capacity of edge devices:

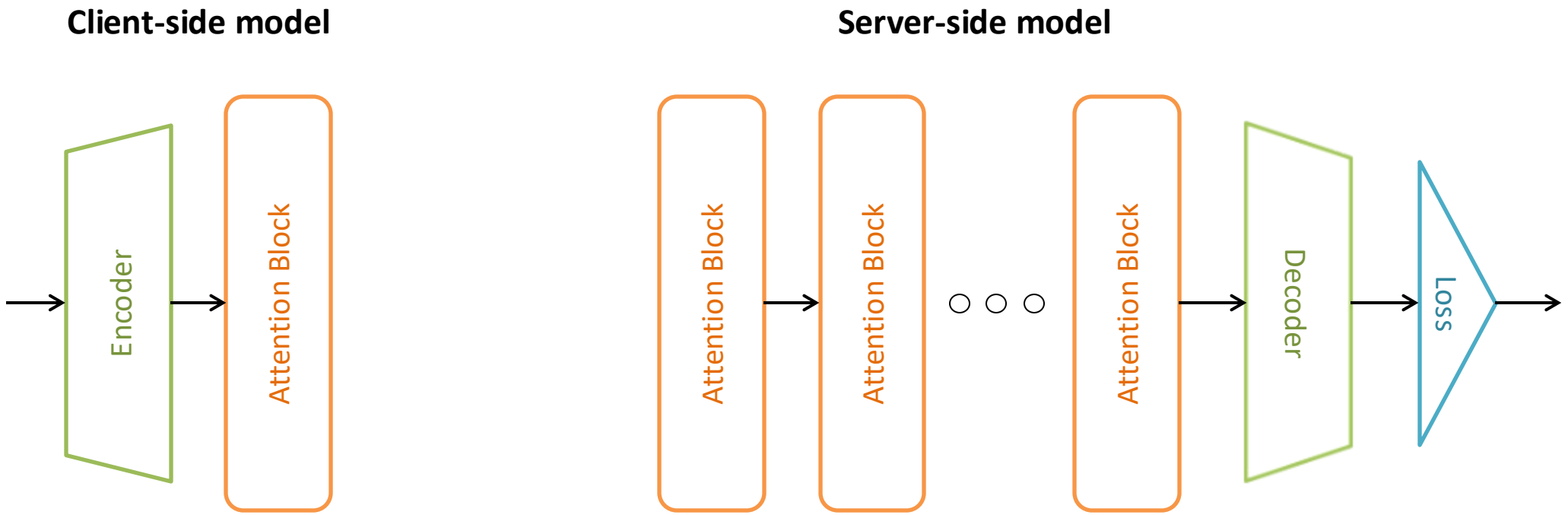
- Contains encoder, decoder, and a cascade of a series of self-attention blocks
- Model size is typically several GBs



Solution: Federated Split Learning



Solution: Federated Split Learning



Learns a model jointly by splitting it into two pieces:

- Smaller client-side model trained by clients
- Larger server-side model trained by server

Video Demo of Distributed LLM Finetuning

Finetuning Video Demo

Summary

- **Goal:** *Democratize LLMs with the Edge*

- **What We Have Achieved This Year:**
  - A cross-institution LLM pretraining system between OSU and Wisconsin
  - A Colosseum-based LLM finetuning system
  - **Sharing a unified learning framework**

- **Proofs of Concepts and Impacts:**
  - Cross-institution pretraining over the edge **that overcomes networking challenges**
  - Wireless edge-based finetuning with advanced **comm.-efficient** finetuning methods
  - Both pretraining and finetuning over the edge achieve **competitive performances**

Takeaway: **AI-EDGE's innovations →The 1st "Anywhere & Everywhere" LLM Training @ Edge**

Roadmap

	Year 1-2	Year 3 (this year)	Year 4	Year 5
Cross-Institution LLM Pretraining	Theoretical and algorithmic foundation of mixed parallelisms for LLM pretraining	(i) Joint learning and scheduling with FSDP (ii) Async FL-based cross-institutional pretraining	Pretraining for larger LLM; Increase the # of institutions; Compression; Mixed parallelism.	Test and verify all communication-efficient methods developed by AI-EDGE researchers
Cross-Institution LLMs Finetuning	Communication-efficient federated learning algorithmic foundation	FL-based finetuning on Colosseum (i) Basic LoRA; (ii) Hybrid LoRA; (iii) Zeroth-order method	(i) More complex LLM finetuning tasks (e.g., RLHF alignment); (ii) System heterogeneity	(i) Security & privacy; (ii) anarchic FL for finetuning; (iii) Over-the-air FL for finetuning;
Milestone Proof-of-Concept		Proofs-of-Concepts: (1) Federated cross-institution LLM pre-training (2) FL-based LLM finetuning on Colosseum	Train-test-improve both LLM pretraining and finetuning Demo of larger-scale LLM pretraining Demo of larger-scale LLM finetuning	Train-test-improve both LLM pretraining and finetuning Demo of mixed parallelisms in LLM pretraining Defense against security & privacy attacks in LLM finetuning


thank you
danke
謝謝
ngiyabonga
рахмат
teşekkür ederim
спасибо
gracias
tapadh leat
bedankt
dziękuję
obrigado
merci
go raibh maith agat
sukriya
kop khun krap
terima kasih
dakujem
arigatô
grazie
mauruuru
hvala
dank je
mochchakkeram
Баярлалаа
faafetai lava
kiitos
dankie
dhanyavad
köszönöm
enkosi
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
takk
sagolun
dèkuji
sobodi
mèsi
najis tuke
didi madloba
kam sah hamnida
rahmat
rahmet
tanemirt
ありがとう
감사합니다
xièxie
ευχαριστώ
diolch
dhanyavadagalu
shukriya
мерси
trugarez
mamnun
дякую
djiere dieuf
tau
paldies
grazzi
matondo
misaotra
blagodaram
спасибі
vinaka
barka
mersi
welalin
tack
spas
mahalo
хвала
asante
manana
obrigada
murakoze
chokrane
tenki
akun
dankon
aciu
nanni
nandri
bayarlalaa
gràçie

धन्यवाद