NSF AI Institute for Edge Computing Leveraging Next Generation Networks (Athena)

Yiran Chen, Duke University
PI and Director of Athena
Summary and Vision

- The Athena Institute advances Artificial Intelligence (AI) technologies to transform the design, operation, and service of future mobile networks.
- The research activities of Athena are organized under four synergistic thrusts: Networking, Computer Systems, AI, and Services.
- Athena is committed to educational and workforce development, cultivating a diverse next generation of mobile network leaders with the core values of ethics and fairness for AI.
- As a nexus point for community, Athena spearheads collaboration and knowledge transfer to translate its emerging technical capabilities to new business models and entrepreneurial opportunities, transforming the future competition model in both research and industry.
A Team Composed of 30 World-class Scholars

A multi-disciplinary team of scientists, engineers, statisticians, legal scholars, and psychologists.
Organization and Key Personnel

External Advisory Board

PI/Institute Director: Y. Chen
Managing Director: J. Krolak
Admin: Rajashi Runton

EWD Directors: S. Daily/N. Washington
CKT Director: J. Derby
BP Director: D. Limbrick, J. Kelly

T1: Networking
S. Banerjee (Lead)
T. Chen
Y. Kim
B. Krishnaswamy
B. Maggs
M. Mao
L. Zhong

T2: Computer Systems
L. Zhong (Lead)
A. Bhattacharjee
K. Chakrabarty
W. Hu
A. Khandelwal
M. Reiter
L. Wills
Y. Kim

T3: AI
H. Li* (Lead)
V. Tarokh
Y. Chen
N. Gong
N. Farahany
S. Han
O. Russakovsky

T4: Services and Apps
M. Pajic (Lead)
S. Banerjee
M. Gorlatova
D. Limbrick
M. Mao
R. Calderbank

Industrial and Community Collaborators: AT&T, Microsoft, Motorola Solutions, EdgeMicro, 5NINES, North Carolina School of Science and Mathematics, STEM Early College@NC A&T, Town of Cary

In red: Female PI/SP
The Geographical Distribution of the Participated Institutions

7 Participating Institutions:
- Duke University (Duke, Lead)
- Massachusetts Institute of Technology (MIT)
- North Carolina A&T University (NCAT)
- Princeton University
- University of Michigan (UMich)
- University of Wisconsin-Madison (WISC)
- Yale University (Yale)

8 National Collaborators:
- AT&T
- EdgeMicro
- Microsoft (MS)
- Motorola (Moto)
- North Carolina School of Science and Mathematics
- The STEM Early College at NCAT
- Town of Cary
- 5NINES

ATHENA
External Advisory Board

Victor Bahl
Technical Fellow, CTO
Azure for Operators at Microsoft

Jilei Hou
VP & Head of AI Research
Qualcomm

Jehan Wickramasuriya
VP, AI & Data Engineering
Motorola Solutions

Charlie Zhang
SVP
Samsung Research America

Victor Zhirnov
Chief Scientist
Semiconductor Research Corporation
Current Generational Upgrade of Mobile Networks

- Current generational upgrade: redesign and rebuild infrastructure to host higher performance and bandwidth for newer generation networks.
- This model is **inflexible, wasteful, and monopolistic**.
Decoupling of Network Platform and Infrastructure

- Instead, our vision decouples the shared mobile network platform from the physical infrastructure
Decoupling of Network Platform and Infrastructure

- Instead, our vision decouples the shared mobile network platform from the physical infrastructure
  - Enabling modularity of physical infrastructure
  - And of flexible opportunity for services

New Business Models + Entrepreneurial Opportunities
A New Comprehensive Role of AI

6G = a faster 5G + AI?

“If I had asked people what they wanted, they would have said faster horses.” – Henry Ford

- AI should not just provide services & applications
- AI should also play an important role in infrastructure design and operation
Our Three Key Insights and Main Goal

**Insight 1:** Advances in AI provides powerful tools to solve the problems in compute and network system designs and operation at scale.

**Insight 2:** Mobile networks and services pose fresh and intriguing challenges to AI’s theoretical advances and practical applications.

**Insight 3:** Mobile network infrastructures and their clients allow for new modalities of AI in clients, edge datacenters, and the Cloud (e.g., FL)

Athena not only addresses the challenges facing mobile networks but also advances the state-of-the-art of AI regarding functionality, efficiency, heterogeneity, and trustworthiness, enabling a shared intelligent network infrastructure and disrupting the ecosystem and business model.
Four Highly-Integrated Research Thrusts

**T1:** Create an adaptable, scalable, performance-aware mobile network infrastructure using a data-driven approach

**T2:** Design the next-generation edge data center with high efficiency, availability, and security, investigate system support for AI

**T3:** Develop AI techniques to fulfill the needs of the next-generation network in functionality, efficiency, heterogeneity, and trustworthiness

**T4:** Develop innovative services and applications for the next-generation network enabled and inspired by the AI technologies
AI-Powered Networking

**Goals**
- Adaptability
- Scalability
- Efficiency

**Outcomes**
- Wireless networks that learn
- Explainable configurations
- Trials with partners (EdgeMicro, 5NINES) and on PAWR platforms

**Fellows**
- Suman Banerjee (Lead)
- Tingjun Chen
- Younhyun Kim
- Bhuvana Krishnaswamy
- Bruce Maggs
- Morley Mao
- Leandros Tassiulas
- Lin Zhong

**Institutions**
- Duke
- UMICH
- Yale
- WISC
Fundamental Challenges

- **Adaptability**: Wireless environment continues to be highly dynamic (e.g., mobility, higher frequencies)

- **Scalability**: End devices growth is unabated (e.g., IoT, vehicles, multiple devices/person)

- **Efficiency**: High speeds require edge cloud services for faster processing, but need to deal with consequent growth in network traffic
AI-Powered Computer Systems at Edge

**Goals**
- Efficiency
- Programmability
- Availability

**Outcomes**
- Design, implementation, and evaluation of algorithm, software systems, & hardware

Lin Zhong (Lead)  Abhishek Bhattacharje  Krishnendu Chakrabarty  Wenjun Hu
Anurag Khandelwal  Younghyun Kim  Mike Reiter  Lisa Wu Wills

Duke  WISC  Yale
Edge Datacenters
Shouldering the computational need of mobile networks & services

• Challenges
  – Efficiency & Availability
  – Diversity of requirements

• Opportunities
  – Disaggregating resources for flexibility
  – AI-powered performance optimization
  – Hardware acceleration for AI

World’s first massive MIMO: From FPGA to software

Argos
L. Zhong, MobiCom ’12

Agora
L. Zhong, CoNEXT ’20

ATHENA
AI Foundations

Goals
- Functionality
- Efficiency
- Heterogeneity
- Trustworthiness

Outcomes
- AI methods for CNS systems
- Significant improvement in system metrics
- Theoretical breakthrough in AI foundations

Yiran Chen, Helen Li, Olga Russakovsky, Robert Calderbank, Vahid Tarokh, Neil Gong, Nita Farahany, Song Han

Duke, MIT, Princeton
AI Foundations in Functionality

- **Goals:** Gain *insights* into network and system operations while appropriately responding to both *foreseen* and *unforeseen* circumstances
  
  - Causal analysis
    - Causal inference
    - Invariant representations
  
  - Out-of-distribution (OOD) prediction
    - Calibration over input space via outlier exposure
    - Feature space analysis
AI Foundations in Efficiency

- **Goals:** Improve efficiency and scalability of AI models in networking and computing systems
- Efficient learning
  - Domain adaptation / knowledge transfer
  - Distributed optimization / federated learning
- AI model design / deployment / execution
  - Compact model design
  - Hardware-aware AutoML / neural architecture search (NAS)

Continual learning (L. Carin, NeurIPS’20, CVPR’21)
Distributed learning (H. Li, NeurIPS’17 Oral)

Compact model design and NAS (H. Li, AAAI’20, ICLR’21)
AI Foundations in Heterogeneity

Goals: Learn in the “physical world” via heterogeneous data and systems in distributed networks, and improve AI’s applications and services

- Horizontal heterogeneity
  - Personalization for each device

- Vertical heterogeneity
  - Heterogeneity-aware federated learning
  - Adaptive optimization

Personalization (S. Han, ICLR’20; Y. Chen, TCPS’21, ICLMA’19 Best Paper)

Federated leaning (Y. Chen, MobiCom’21; L. Carin & H. Li, NSF C-Accel)
AI Foundations in Trustworthiness

**Goals:** Ensure predictable and reliable quality-of-service (QoS) and measure fairness within an ethics framework

- Robustness at deployment
  - Robustness to adversarial attacks
  - Provable security and privacy protection

- Fairness, ethics and social implications
  - Algorithmic bias mitigation with practical solutions and theoretical guarantees
  - Ethics consultation framework

Robustness and security (N. Gong, AAAI’21; NDSS’19 Distinguished Paper Award honorable mention; H. Li, NeurIPS’20 Oral)

Algorithmic fairness (O. Russakovsky, ’21; CVPR’20; FAT**’20; ECCV’20 spotlight; CVPR’21)
AI-Powered Services & Apps at Edge

- Promise of Edge-Supported Autonomy

- Goal: **Assured, robust & resilient** services for **autonomous systems at the edge**

- Challenges:
  - Infrastructural (i.e., data/processing) requirements:
    - Low-latency, high-bandwidth network
    - Integrated computational support for collaborative AI
  - Algorithmic requirements:
    - (Scalable) AI methods for distributed/collaborative decision making
    - Strong performance (safety & functionality, robustness, trustworthiness) guarantees
AI-Powered Services and Apps at Edge

• **Goal:** *Assured, robust* & *resilient* services for autonomous systems at the edge
  - Exploiting the available (heterogenous) communication and computation resources
  - Provide strong safety & performance guarantees at design- and run-time, as the system & environment evolve
  - Rigorous design and analysis approaches for safety-critical systems

• **Our Focus**
  - Robust situational-awareness at the edge
  - High-assurance autonomy at the edge
The Roles of Other Components

Athena

T3: Advancing AI for mobile networks
T1: AI-powered networking at the edge
T2: AI-powered computer systems at the edge
T4: Services and applications

Broadening Participation (BP)

Education & Workforce Development (EWD)

Collaboration and Knowledge Transfer (CKT)

Academic
Industrial
Community
Professional

Women
Minoritized
Other Underrepresented Groups
K-12
Undergrads
Graduate Students
Post-Doc

Women
Minoritized
Other Underrepresented Groups
K-12
Undergrads
Graduate Students
Post-Doc

Community
Academic
Industrial
Professional
Office and Lab Space, Computing Resources

- The hub of Athena will be housed on the 4th floor of the new Wilkinson Building on Duke campus (10,000+sf).

- The project leads directly own 130+ servers with 600+ CPUs and have the accesses to institutional high-performance computing facilities of 60,000+ CPU cores and 10,000+ GPUs.

- Physical proximity of Aerpaw and MCity testbeds. Existing collaborations facilitate physical access to other PAWR testbeds.
Metrics of Athena’s Success

- **Strong Academic Performance**
- **Industry products derived from Athena’s technologies**
- **Forging and leading new research directions**
- **Prototypes and demos of emerging technologies**
- **Software downloads and IP adoptions, and publication citations**
- **Placement and performance of graduated trainees**
- **Best Practices for Ethical AI education**
- **Technology transfer and entrepreneurial activities**

**Education excellence and diversity of trainees**

**Participation in and influence on industrial standards**

**Intellectual Impacts**

**Broader Impacts**

**Princeton**

**WISC**

**Duke**

**Yale**

**MIT**

**NC A&T**

**Support from**

**Athena**
Summary

**Vision:** Athena envisions a virtualized mobile network powered by AI with unprecedented efficiency, reliability, and performance, and aims at realizing this vision with foundational and use-inspired research as well as advancing the SOTA of AI in both application and theory.

**Role of AI:** Instead of a mere important application, our developed AI technologies will also offer theoretical and technical foundations for future mobile networks in functionality, heterogeneity, scalability, and trustworthiness.

**Nexus Point:** Serving as the nexus point of community, Athena will facilitate the ecosystem of the emerging technologies and cultivate the diverse next-generation technical leaders having the values of ethics and fairness.

**Societal Impact:** The success of Athena will disrupt the future mobile network industries, create new business model and entrepreneurial opportunities, and transform the competition model of future mobile network industry and research.
Call for Collaborations!