ATHENA

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.

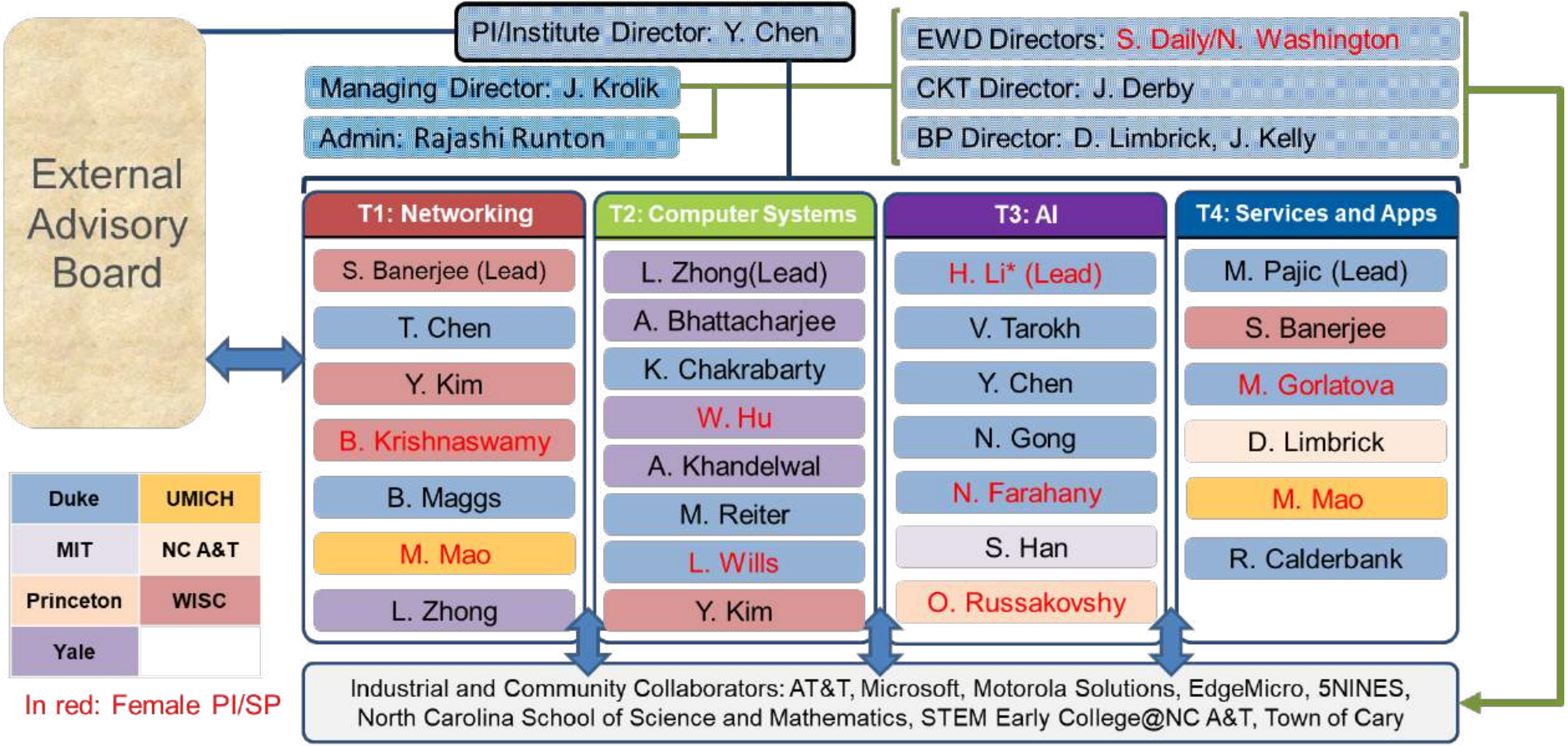
A Team Composed of 30 World-class Scholars

Suman Banerjee	Abhishek Bhattacharjee	Vahid Tarokh	Krishnendu Chakrabarty	Yiran Chen	Tingjun Chen	Shaundra Daily	Jeffrey Derby	Nita Farahany	Neil Gong
Maria Gorlatova	Song Han	Wenjun Hu	Anurag Khandelwal	Younghyun Kim	Bhuvana Krishnaswamy	Jeffery Krolik	Helen Li	Daniel Limbrick	Bruce Maggs
Morley Mao	Miroslav Pajic	Mike Reiter	Olga Russakovsky	Leandros Tassiulas	Nicki Washington	Lisa Wu Wills	Lin Zhong	Robert Calderbank	John Kelly

Duke	NC A&T	MIT	Princeton	UMICH	WISC	Yale

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. Krolik

Admin: Rajashi Runton

EWD Directors: S. Daily/N. Washington

CKT Director: J. Derby

BP Director: D. Limbrick, J. Kelly

T1: Networking	T2: Computer Systems	T3: AI	T4: Services and Apps
S. Banerjee (Lead)	L. Zhong(Lead)	H. Li* (Lead)	M. Pajic (Lead)
T. Chen	A. Bhattacharjee	V. Tarokh	S. Banerjee
Y. Kim	K. Chakrabarty	Y. Chen	M. Gorlatova
B. Krishnaswamy	W. Hu	N. Gong	D. Limbrick
B. Maggs	A. Khandelwal	N. Farahany	M. Mao
M. Mao	M. Reiter	S. Han	R. Calderbank
L. Zhong	L. Wills	O. Russakovshy	
	Y. Kim		

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

Duke	UMICH
MIT	NC A&T
Princeton	WISC
Yale	

In red: Female PI/SP

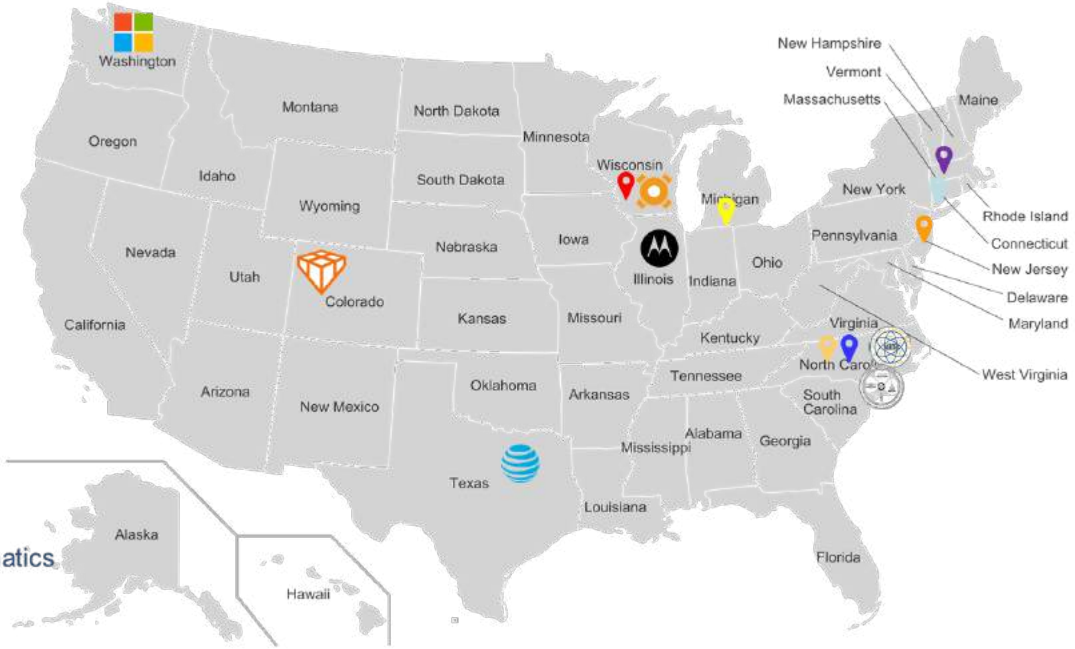
The Geographical Distribution of the Participated Institutions

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



External Advisory Board

Victor Bahl
Technical Fellow, CTO
Azure for Operators at Microsoft

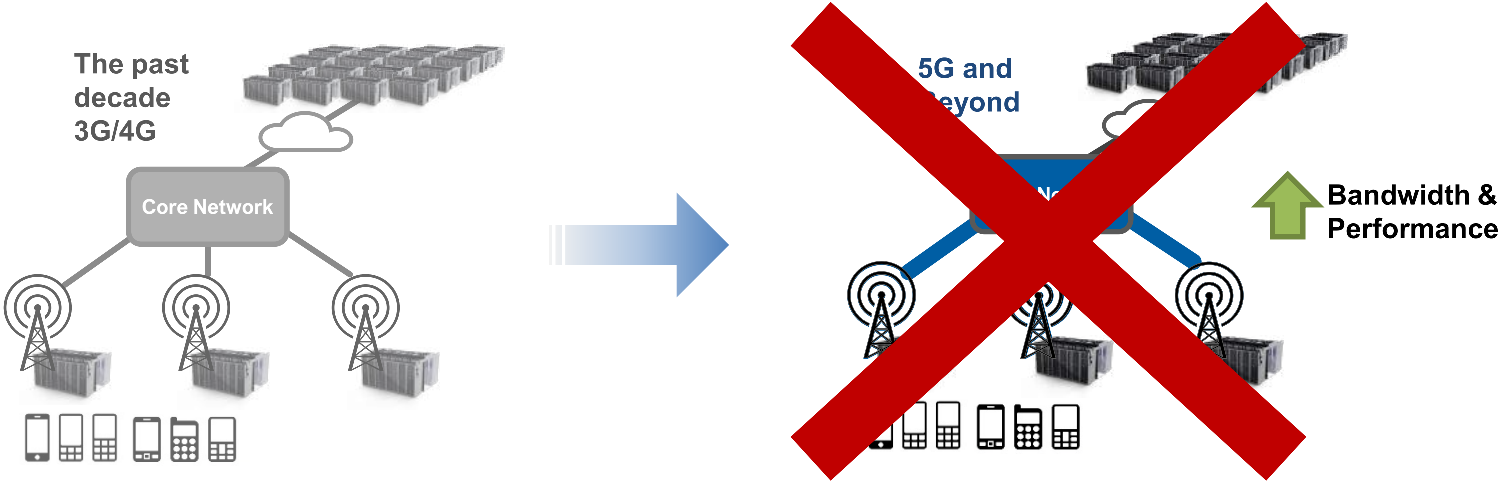
Jilei Hou
VP & Head of AI Research
Qualcomm

Jehan Wickramasuriya
VP, AI & Data Engineering
Motorola Solutions

Charlie Zhang
SVP
Samsung Research America

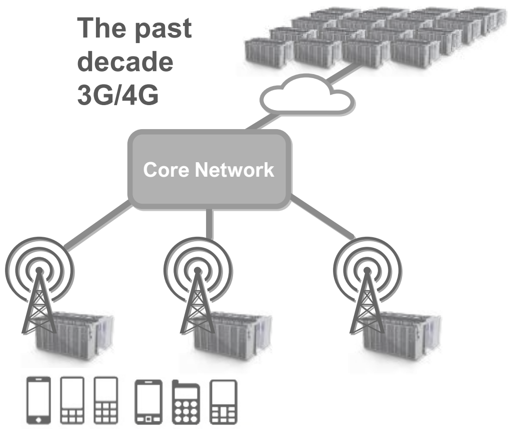
Victor Zhirnov
Chief Scientist
Semiconductor Research Corporation

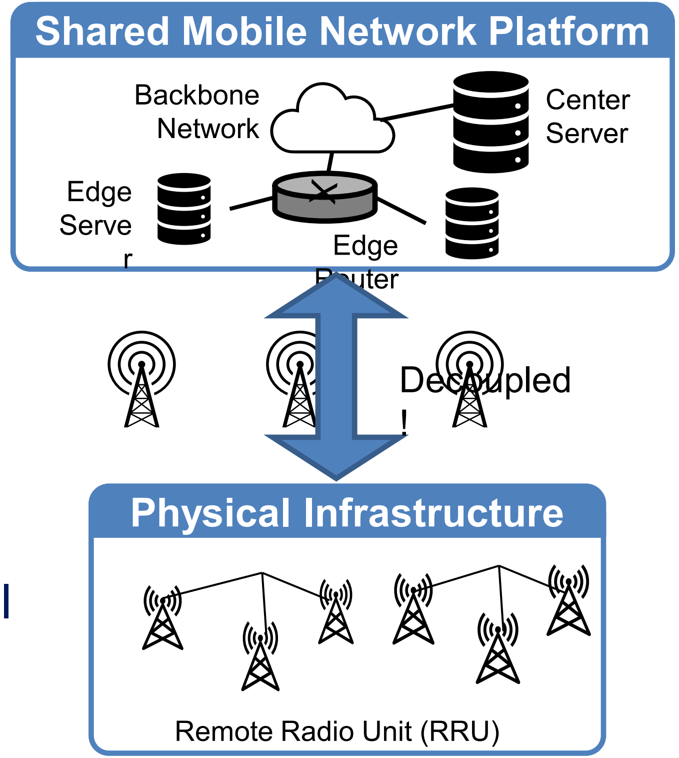
Current Generational Upgrade of Mobile Networks



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

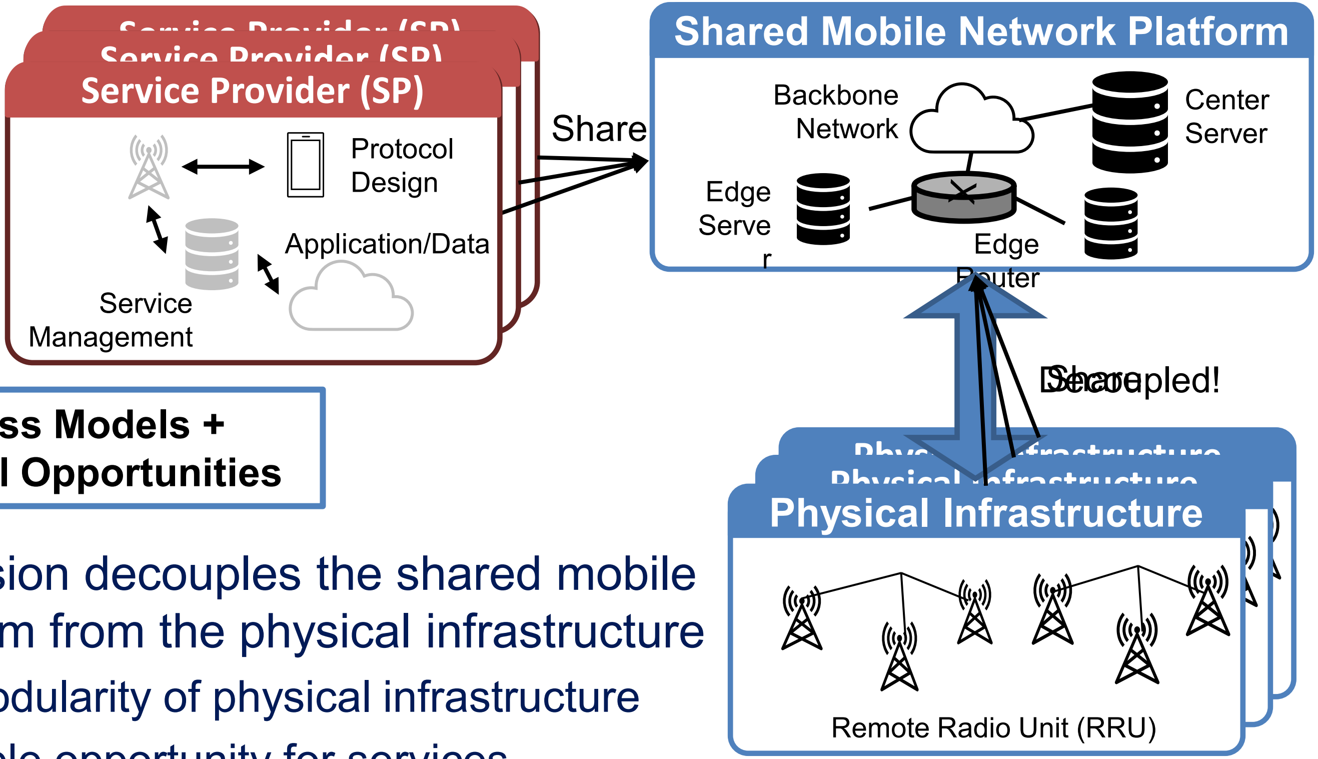
Decoupling of Network Platform and Infrastructure





- Instead, our vision decouples the shared mobile network platform from the physical infrastructure

Decoupling of Network Platform and Infrastructure



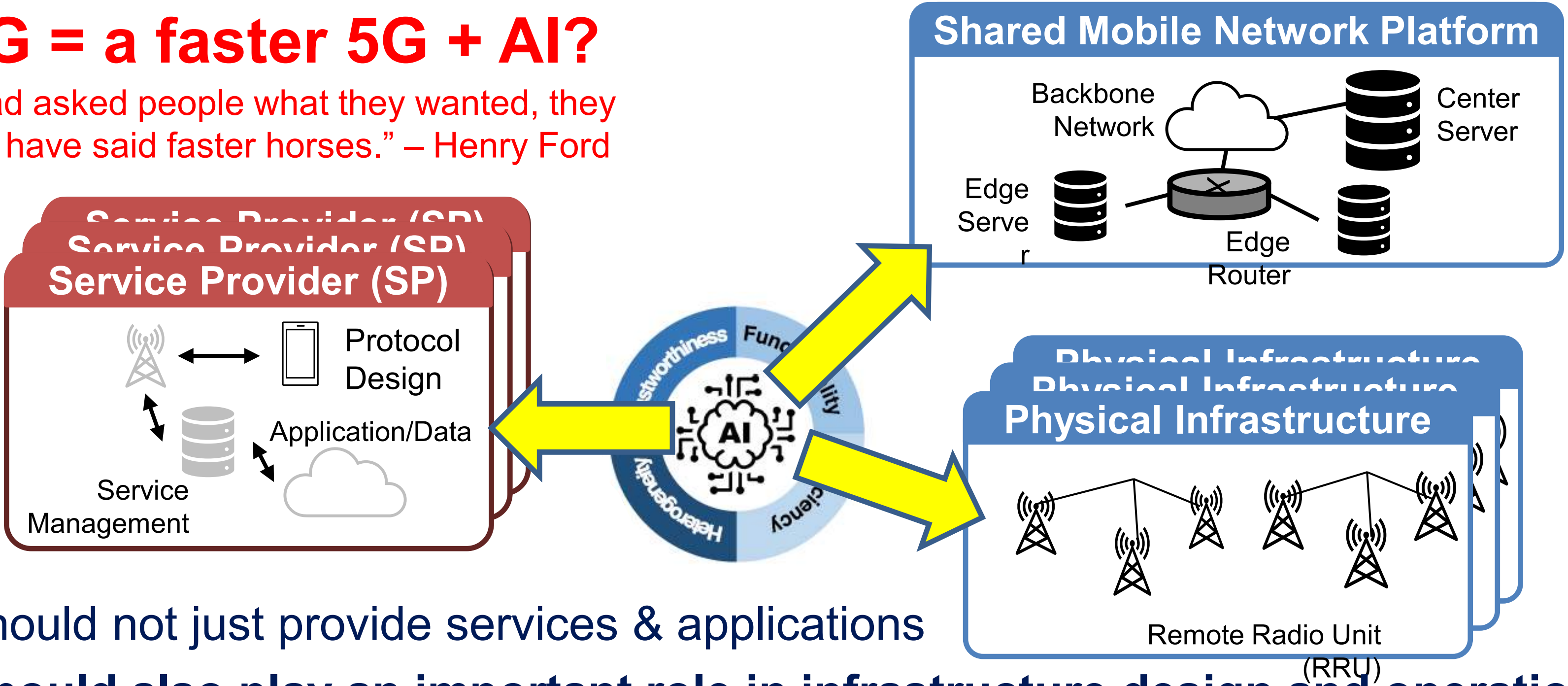
New Business Models +
Entrepreneurial Opportunities

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

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

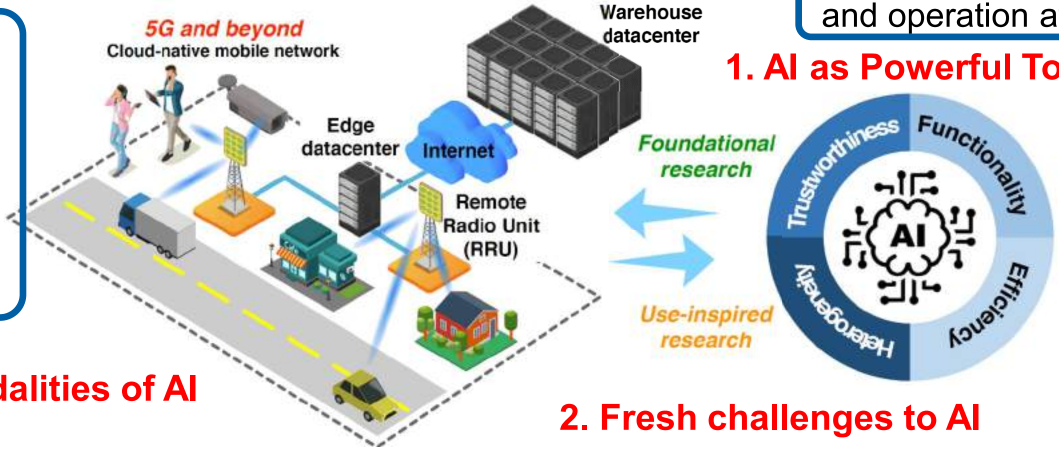


- AI should not just provide services & applications
- **AI should also play an important role in infrastructure design and operation**

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



1. AI as Powerful Tools

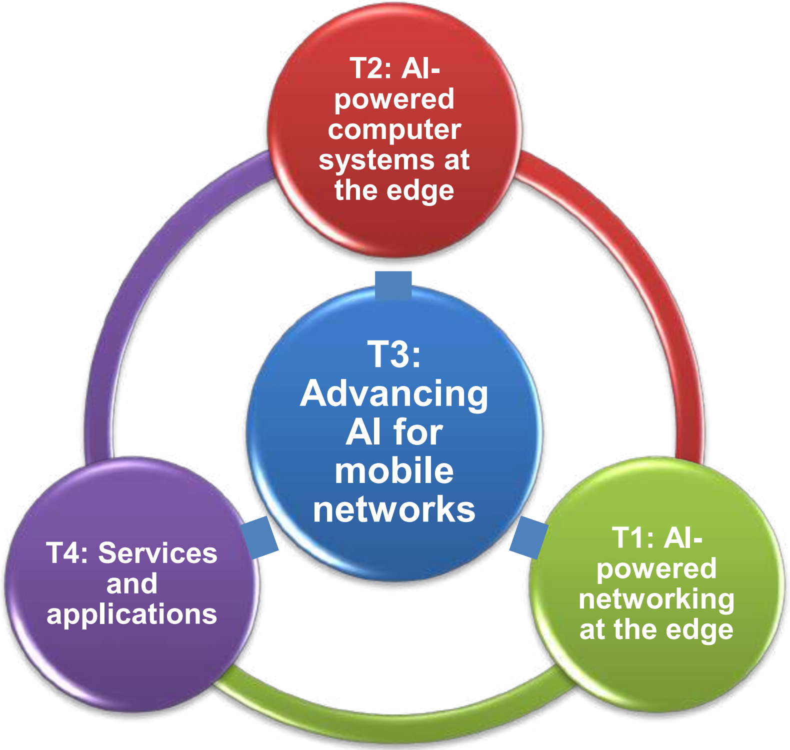
2. Fresh challenges to AI

3. New modalities of AI

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

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


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

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

Suman Banerjee (Lead)

Tingjun Chen

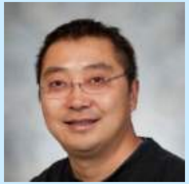
Younghyun Kim

Bhuvana Krishnaswamy

Bruce Maggs

Morley Mao

Leandros Tassiulas

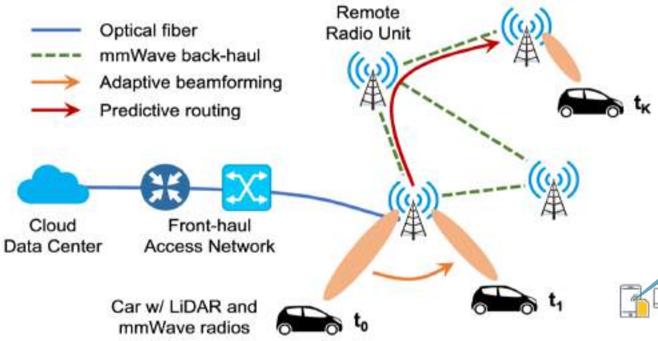


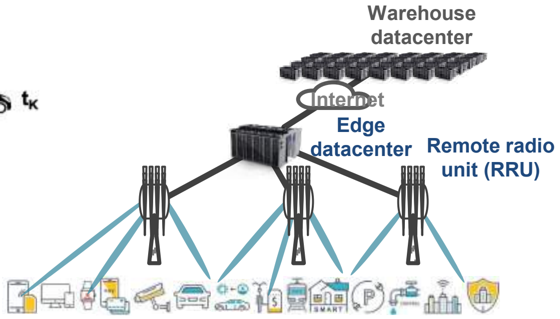
Lin Zhong

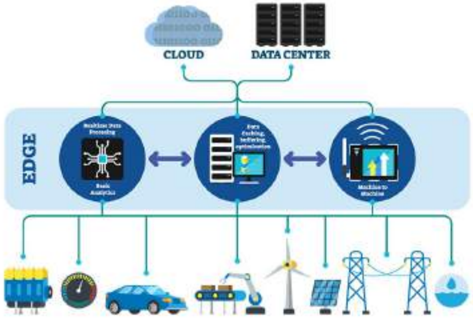
Duke UMICH Yale WISC

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







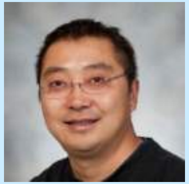
AI-Powered Computer Systems at Edge

Goals

- Efficiency
- Programmability
- Availability

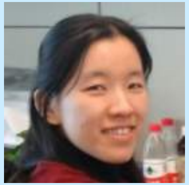
Outcomes

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


Lin Zhong (Lead)

Abhishek Bhattacharje

Krishnendu Chakrabarty


Wenjun Hu

Anurag Khandelwal

Younghyun Kim

Mike Reiter

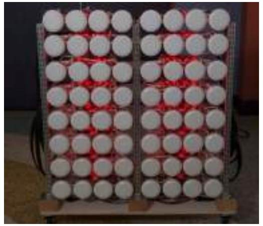
Lisa Wu Wills

Duke WISC Yale

Edge Datacenters

Shouldering the computational need of mobile networks & services

- Challenges
  - Efficiency & Availability
  - Diversity of requirements



Argos

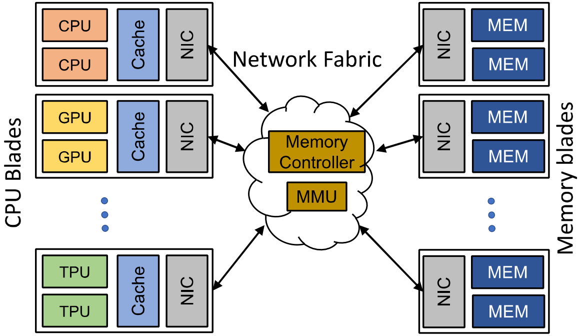
L. Zhong, MobiCom '12

Agora

L. Zhong, CoNEXT '20

World's first massive MIMO: From FPGA to software

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



AI Foundations

Goals

- Functionality
- Efficiency
- Heterogeneity
- Trustworthiness

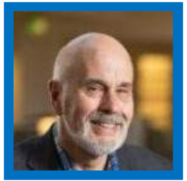
Outcomes

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

Yiran Chen

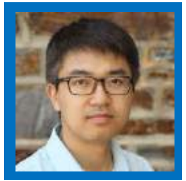
Helen Li

Olga Russakovsky



Robert Calderbank

Vahid Tarokh



Neil Gong

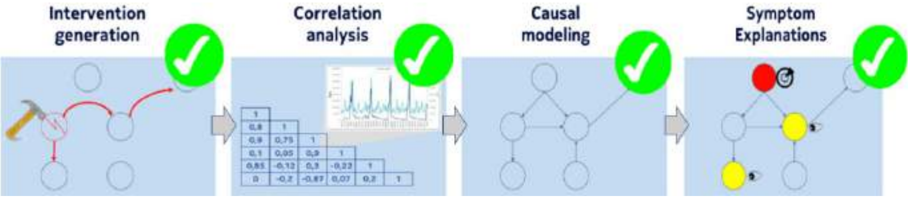
Nita Farahany

Song Han

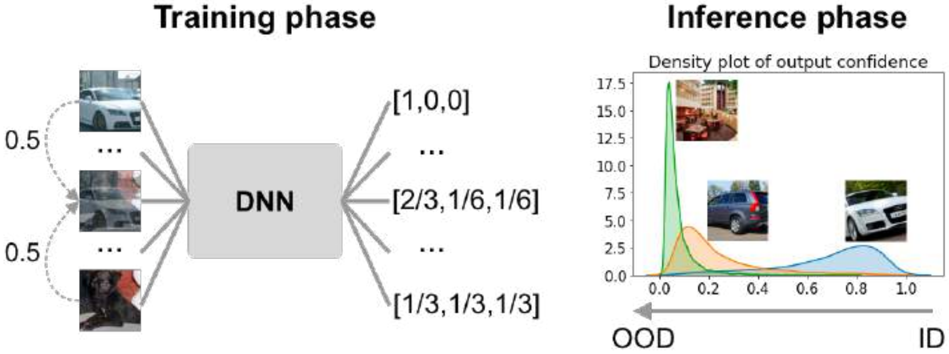
Duke	MIT	Princeton

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



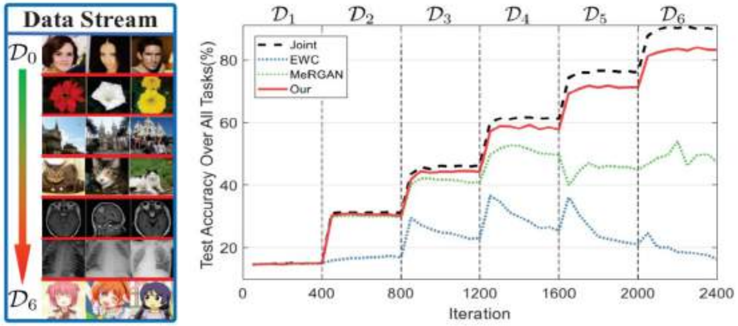
Causal analysis and interpretable AI (L. Carin, ICLR'21, NeuIPS'20)



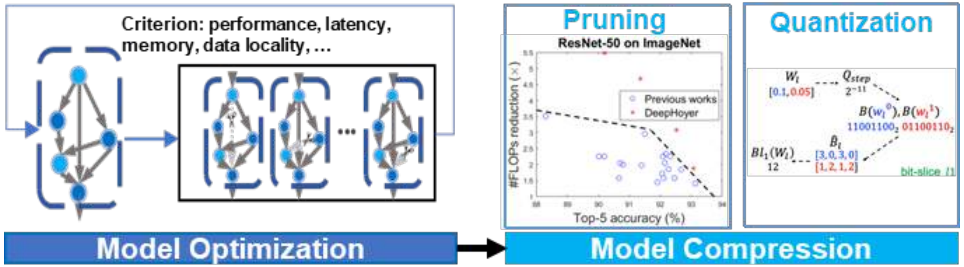
OOD data prediction (Y. Chen & H. Li, '21)

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)



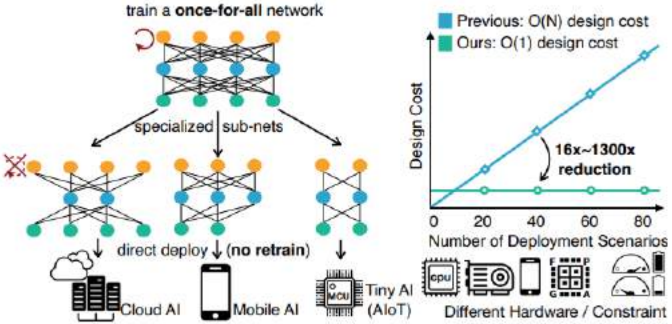
Compact model design and NAS (H. Li, AAAI'20, ICLR'21)

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

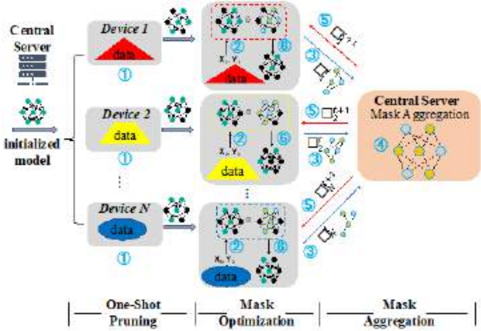
- Horizontal heterogeneity
  - Personalization for each device

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



train a once-for-all network
specialized sub-nets
direct deploy (no retrain)
Cloud AI
Mobile AI
Tiny AI (AIoT)
Previous: O(N) design cost
Ours: O(1) design cost
Design Cost
16x~1300x reduction
0
20
40
60
80
Number of Deployment Scenarios
cpu
Different Hardware / Constraint

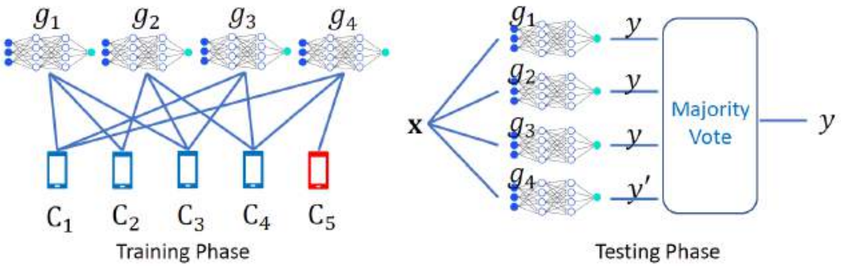
Personalization (S. Han, ICLR'20; Y. Chen, TCPS'21, ICLMA'19 Best Paper)



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

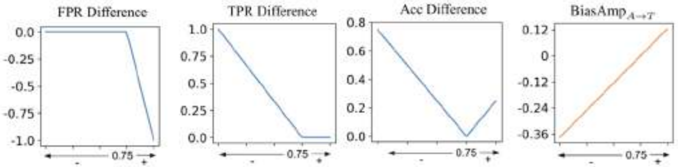


Robustness and security (N. Gong, AAAI'21; NDSS'19 Distinguished Paper Award honorable mention; H. Li, NeurIPS'20 Oral)

$$\text{BiasAmp}_{\rightarrow} = \frac{1}{|\mathcal{A}||\mathcal{T}|} \sum_{a\in\mathcal{A}, t\in\mathcal{T}} y_{at}\Delta_{at} + (1 - y_{at})(-\Delta_{at})$$

$$y_{at} = \mathbb{1}\left[P(A_a = 1, T_t = 1) > P(A_a = 1)P(T_t = 1)\right]$$

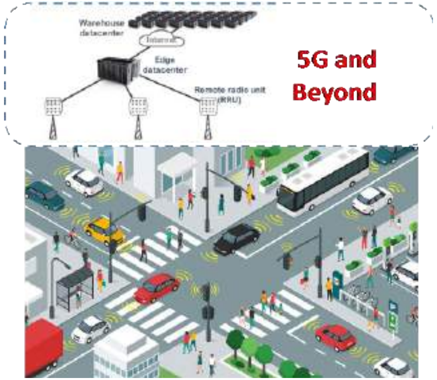
$$\Delta_{at} = \begin{cases} P(\hat{T}_t = 1|A_a = 1) - P(T_t = 1|A_a = 1) \\ \text{if measuring } A \rightarrow T \\ P(\hat{A}_a = 1|T_t = 1) - P(A_a = 1|T_t = 1) \\ \text{if measuring } T \rightarrow A \end{cases}$$



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



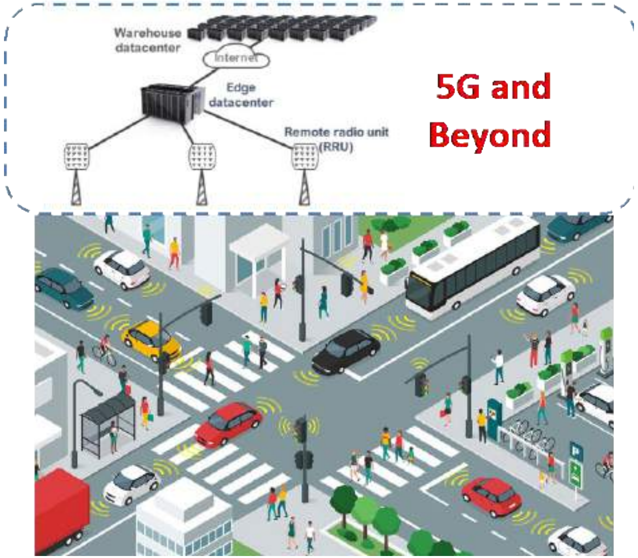
Miroslav Pajic (Lead)	Maria Gorlatova	Daniel Limbrick	Morley Mao	Suman Banerjee

Duke	NC A&T	UMICH	WISC

- 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

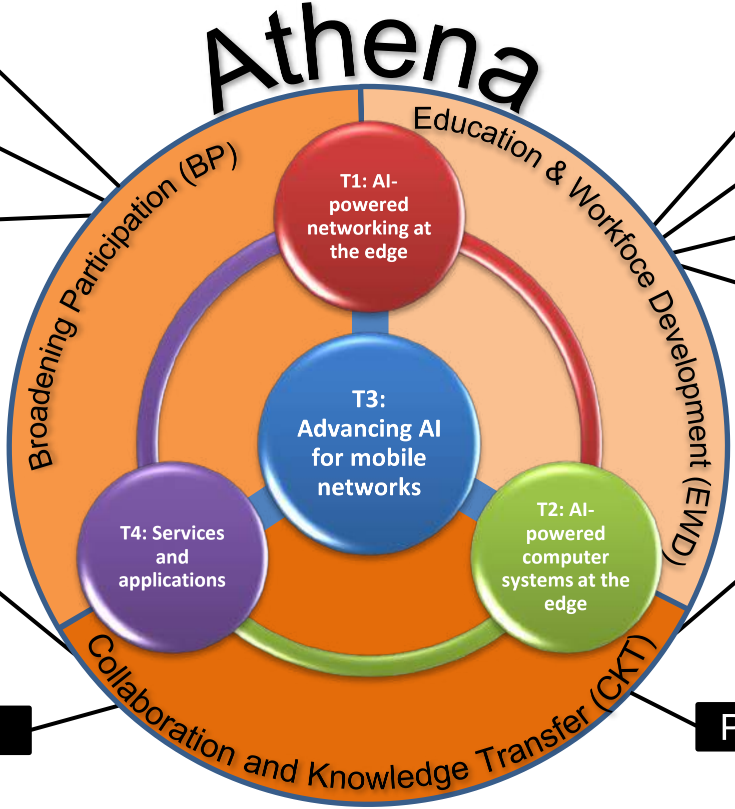
- Our Focus
  - Robust situational-awareness at the edge
  - High-assurance autonomy at the edge

The Roles of Other Components

Women

Minoritized

Other Underrepresented Groups


Athena
Broadening Participation (BP)
Education & Workfoce Development (EWD)
Collaboration and Knowledge Transfer (CKT)
T1: AI-powered networking at the edge
T3: Advancing AI for mobile networks
T4: Services and applications
T2: AI-powered computer systems at the edge

K-12

Undergrads

Graduate Students

Post-Doc

The STEM Early College at N.C. A&T

North Carolina School of Science and Mathematics

Yale University

PRINCETON UNIVERSITY

Duke

Academic

WISCONSIN

UNIVERSITY OF MICHIGAN

Massachusetts Institute of Technology

Town of Cary

Community

MOTOROLA SOLUTIONS

AT&T

Microsoft

Industrial

5NINES

EDGEMICRO Get Closer

Professional

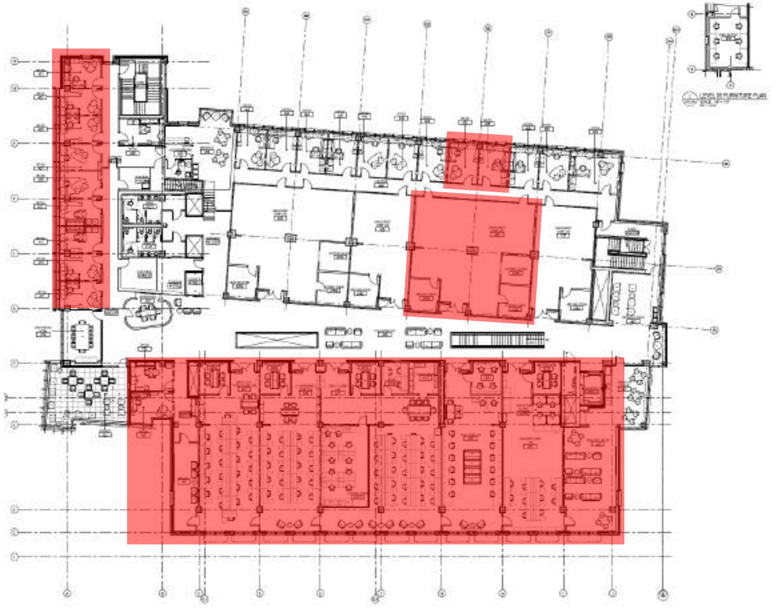
IEEE

acm

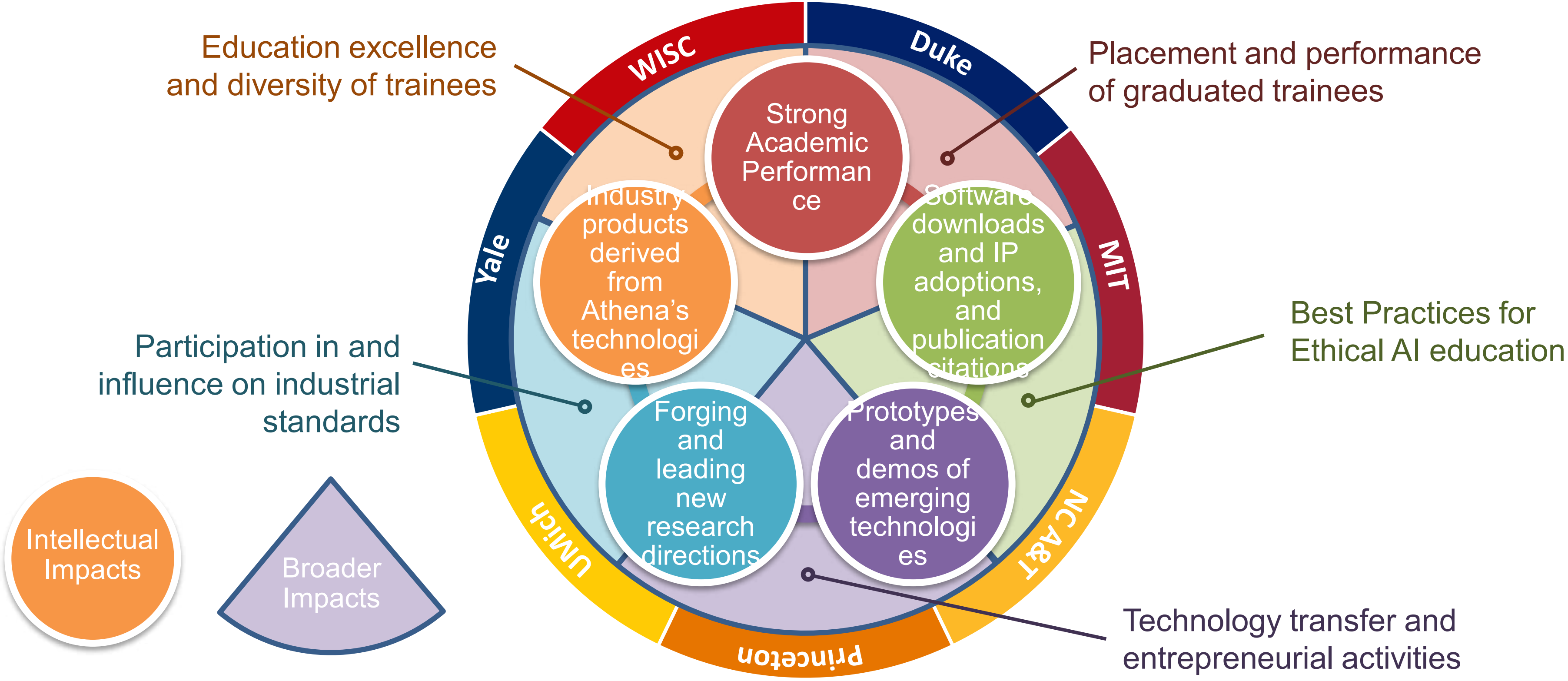
Office and Lab Space, Computing Resources

- 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.

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



Metrics of Athena's Success



Education excellence and diversity of trainees

Placement and performance of graduated trainees

Participation in and influence on industrial standards

Best Practices for Ethical AI education

Technology transfer and entrepreneurial activities

Strong Academic Performance

Industry products derived from Athena's technologies

Software downloads and IP adoptions, and publication citations

Forging and leading new research directions

Prototypes and demos of emerging technologies

Intellectual Impacts

Broader Impacts

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!