Edge AI and Ultra-Low Power Computing

Break-Out Session, 2025 NSF CSR PI Meeting

Thanks to all participants!

Leaders: Maria Gorlatova, Duke University

Mahadev Satyanarayanan, Carnegie Mellon University

Mobility Comes at a High Price

- **Weight / size** (can't be too heavy, too big, or too awkward)
- **Thermal dissipation** (can't burn the user!)
- **Energy efficiency** (long battery life)

} **Severe constraints on SWaP devices**

Leads to *"Mobility Penalty"*

- ***resource-poor relative to non-mobile hardware***
- **persists over time, in spite of Moore's Law**
- **both mobile & non-mobile hardware improve**
- **gap remains**

How to Achieve Real-Time AI on Mobile Devices?

Catalyst for ASIC Discovery: Edge Offload

Tier-3

SWaP Devices

Drones

Video Cameras

Smartphones

AR/VR Devices

Tier-2

Cloudlet

WiFi or cellular

Low Latency
< 20 ms RTT
short tail

Internet

Tier-1

Cloud

- Implement the application purely in software first
- Use cloudlet offload to overcome mobility penalty
- No special ASIC needed up front
- Create the ASIC only after the win is clear

Some Questions to Frame Discussion

Cloud LMMs are all-capable. Is mobile computing dead?

No! Reasons include: privacy, operation without cloud connectivity.

LMM advances will drive an explosion in mobile applications.

Previously developed mobile-edge-cloud architectures can be repurposed for LMM personalization

AI is a rapidly changing field – are custom AI accelerators obsolete upon delivery?

Proofs of concept can be valuable.

Core advances in AI are cloud-driven. For mobile systems, “sweet spot” closer to programmability than hard-coding.

Some Questions to Frame Discussion

In 5 years, will we still need edge offloading?

Yes!

Flexibility offered by edge is important for mobile AI applications.

Keeping data local remains important, including in regulatory terms.

Robotics and AI communities leading the realization of pervasive robotic platforms – where can computer systems researchers "move the needle"?

CSR researchers can offer system-level vision.

Understand complex multi-dimensional resource constraints and tradeoffs --- across computing, communications, energy, …

Some Questions to Frame Discussion

User experience with AR and VR systems is not captured by traditional metrics – what approaches could drive system and algorithm design?

Opportunity to infuse awareness of computer system metrics and resource restrictions into XR platforms and applications.

Is energy harvesting compatible with edge AI?

AI is ever-more energy hungry.

Opportunity for CSR researchers to define which part of sensor-to-cloud pipeline can be done on an energy harvesting budget.