
University of
St Andrews
www.st-andrews.ac.uk

University of St Andrews

Edge Machine Learning:

The Opportunities for Systems Research

Blesson Varghese

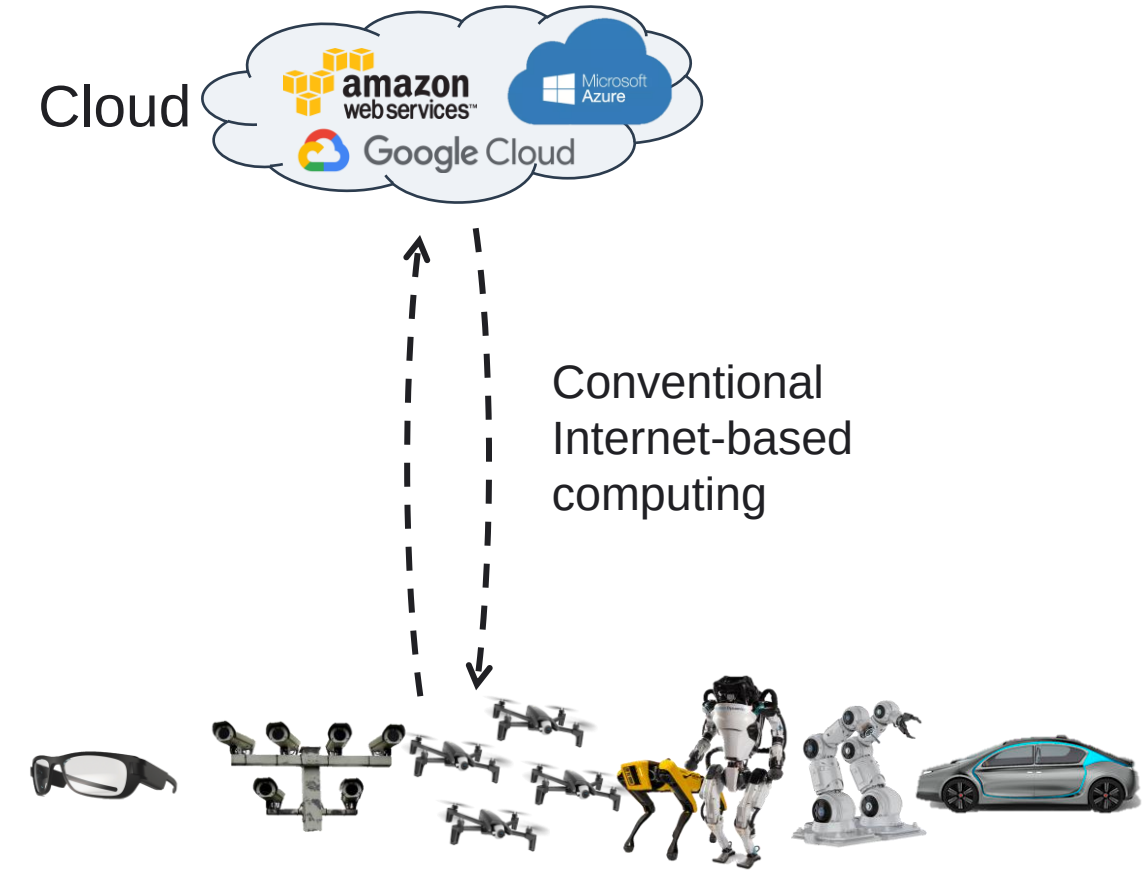
blesson@st-andrews.ac.uk

www.blessonv.com

Our Systems

- Key issues
  - Privacy
  - Responsiveness
  - Bandwidth

Internet systems yesterday



Our Systems - Trends

>6 billion

<4 billion

- 4-8GB RAM
- 32-128GB storage and extendable

Apollo Guidance Computer

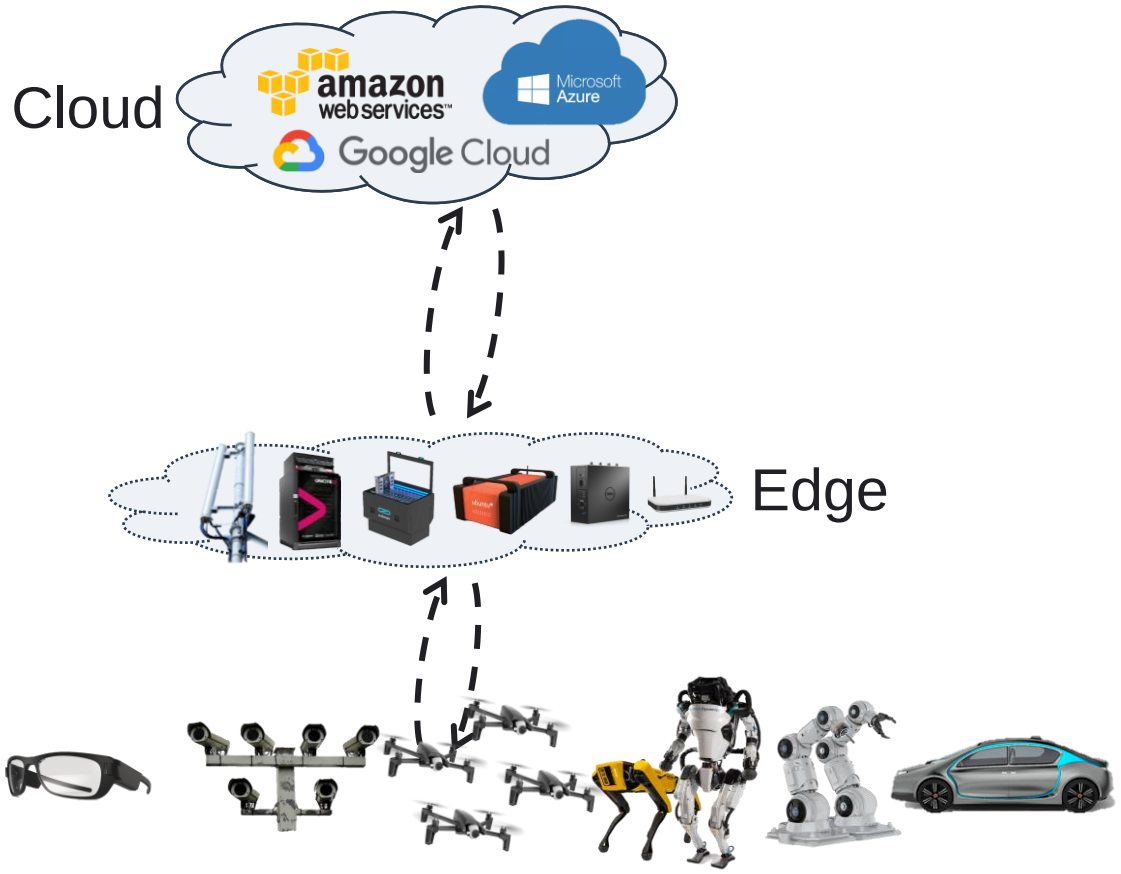
- 4KB RAM
- Some are GPU powered
- 32KB storage

Our Systems

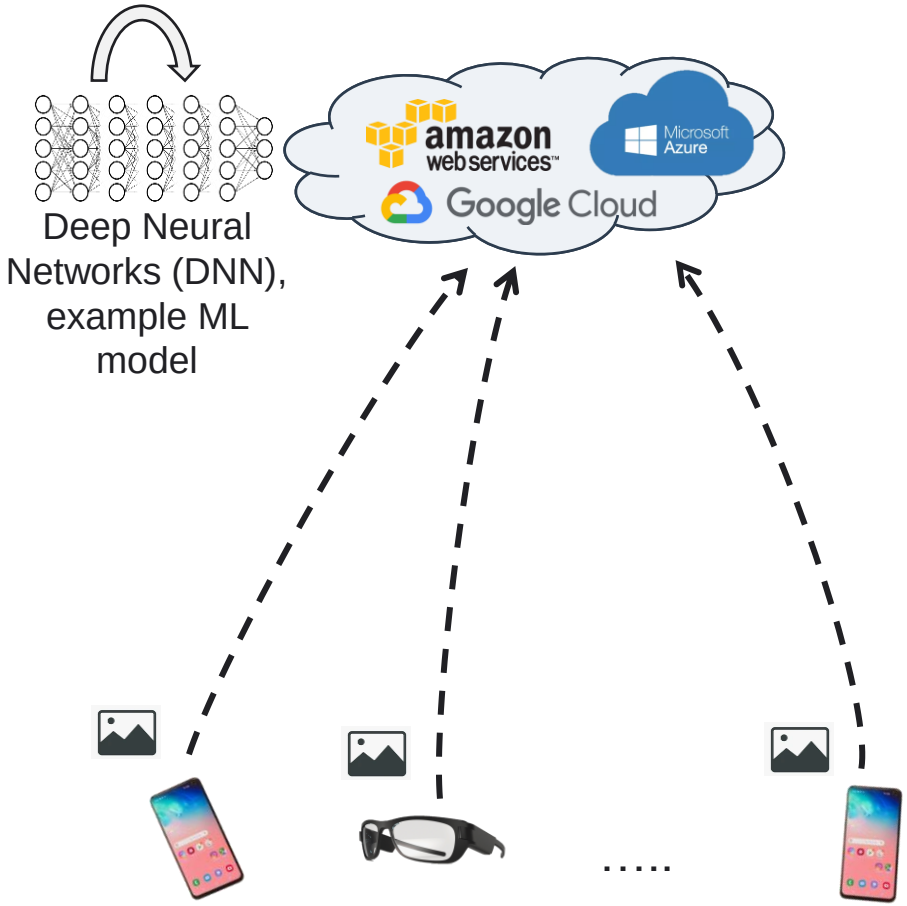
- Key benefits
  - Privacy preserving
  - Improves responsiveness
  - Reduces ingress bandwidth demand

Internet systems tomorrow

Extreme edge - devices, including sensors/ actuators



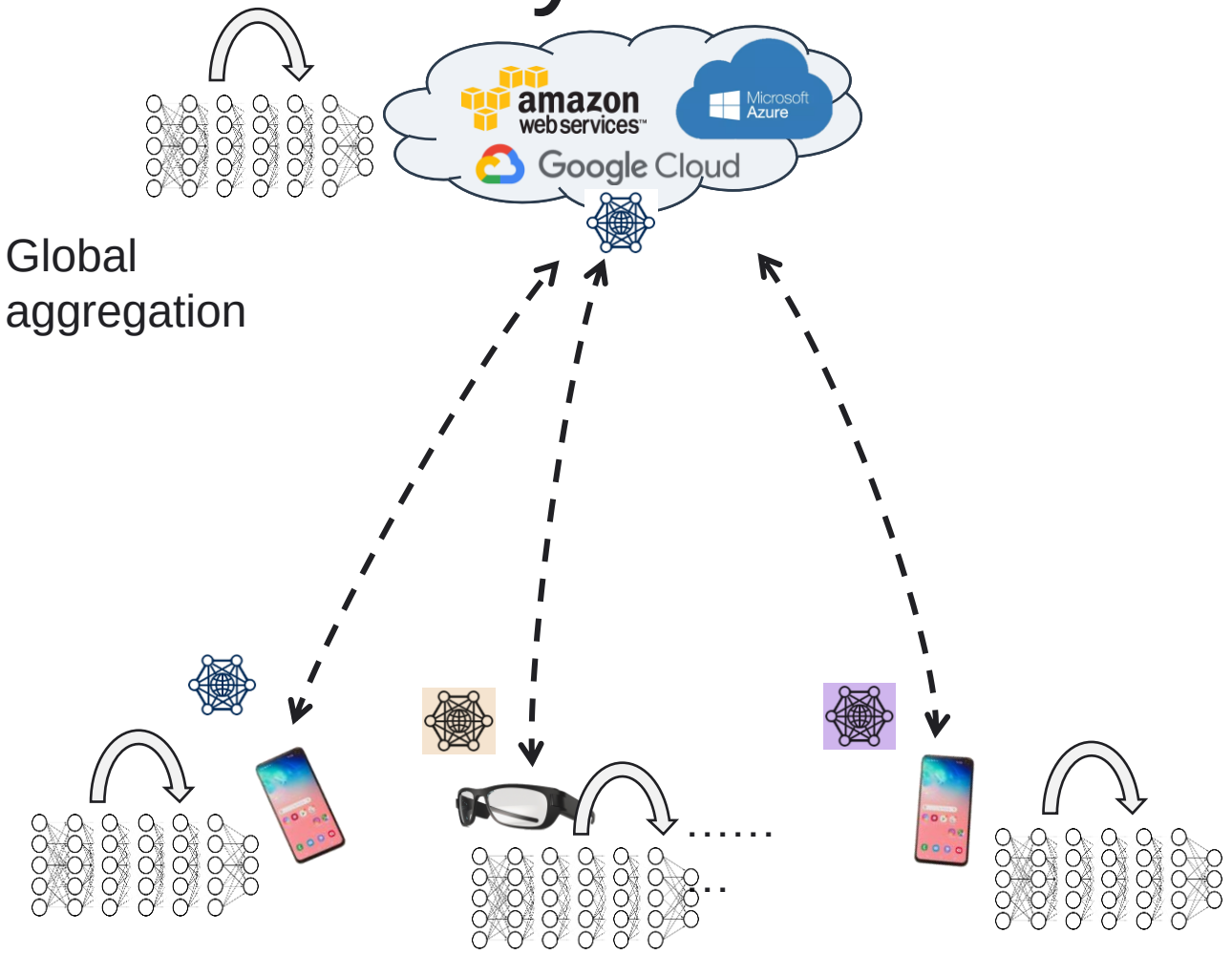
ML and Systems



- Essential techniques required for making sense of the data that is generated
  - *Interpreting signals from sensors*
  - *Making predictions about the environment*

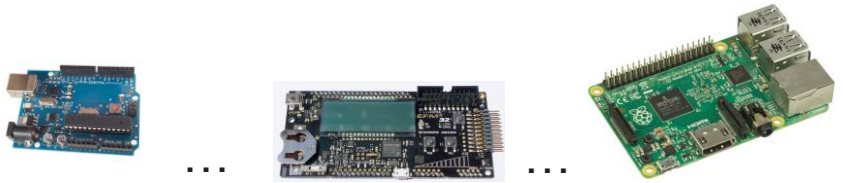
- Key Points
  - Not privacy preserving
  - Bandwidth intensive
  - Not responsive for real-time applications

ML and Systems



- Classic Federated Learning
- Key Points
  - Privacy preserving
  - Less bandwidth intensive
  - Responsive for real-time

A Few Opportunities…



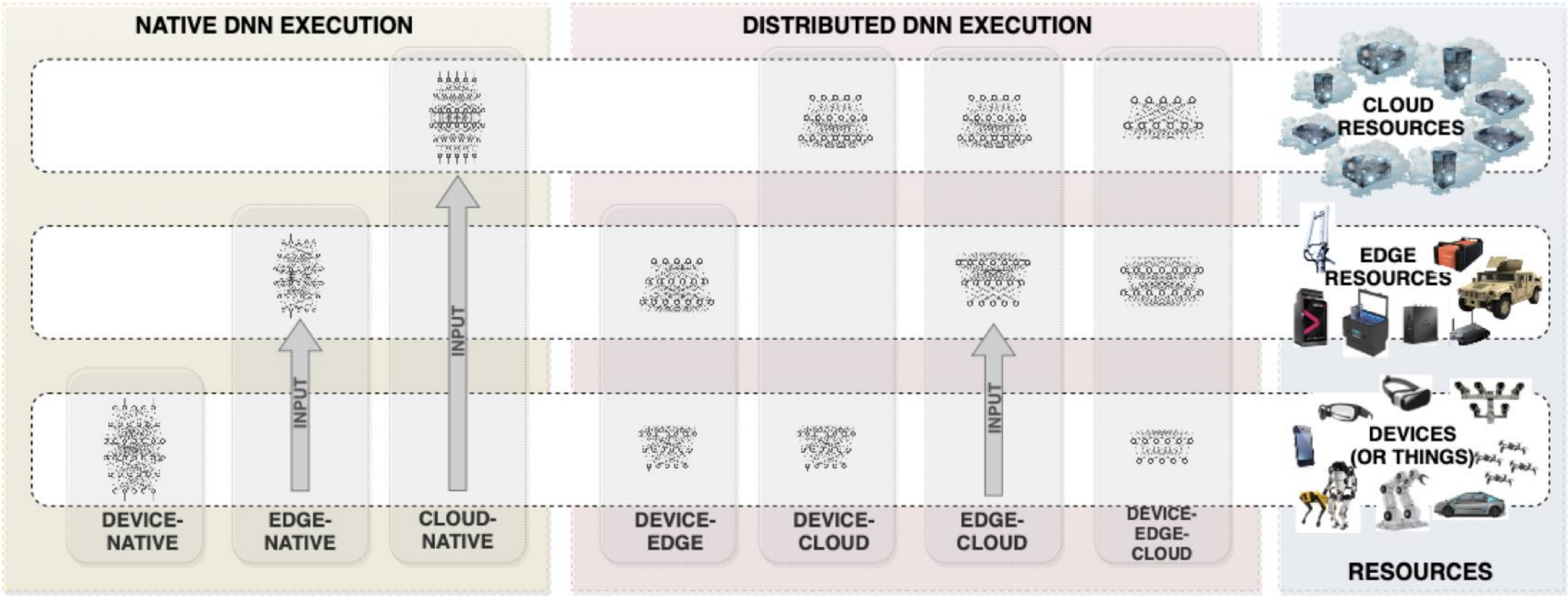
- Key areas (we have been working on):
  - *ML systems that work in resource constrained environments*
  - *ML systems that respond to changing operational requirements*
  - *ML systems that are performance efficient*

Edge ML - Techniques

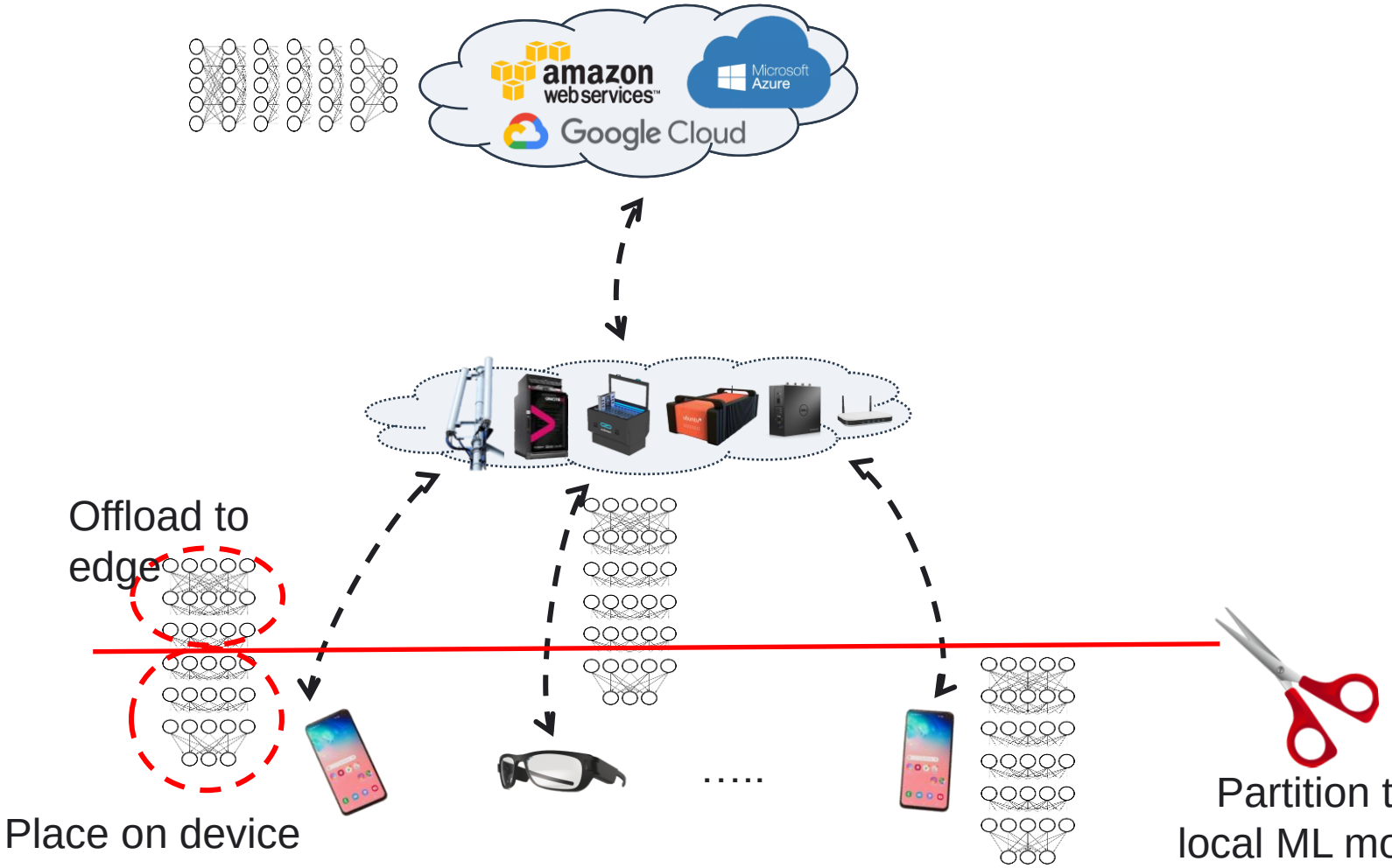
- New build – ‘Create from Scratch’
- Miniaturisation – ‘Squeeze’
  - Compression
  - Quantization
- Offloading – ‘Shift’
  - Partitioning

Preserves privacy and accuracy

Offloading in ML



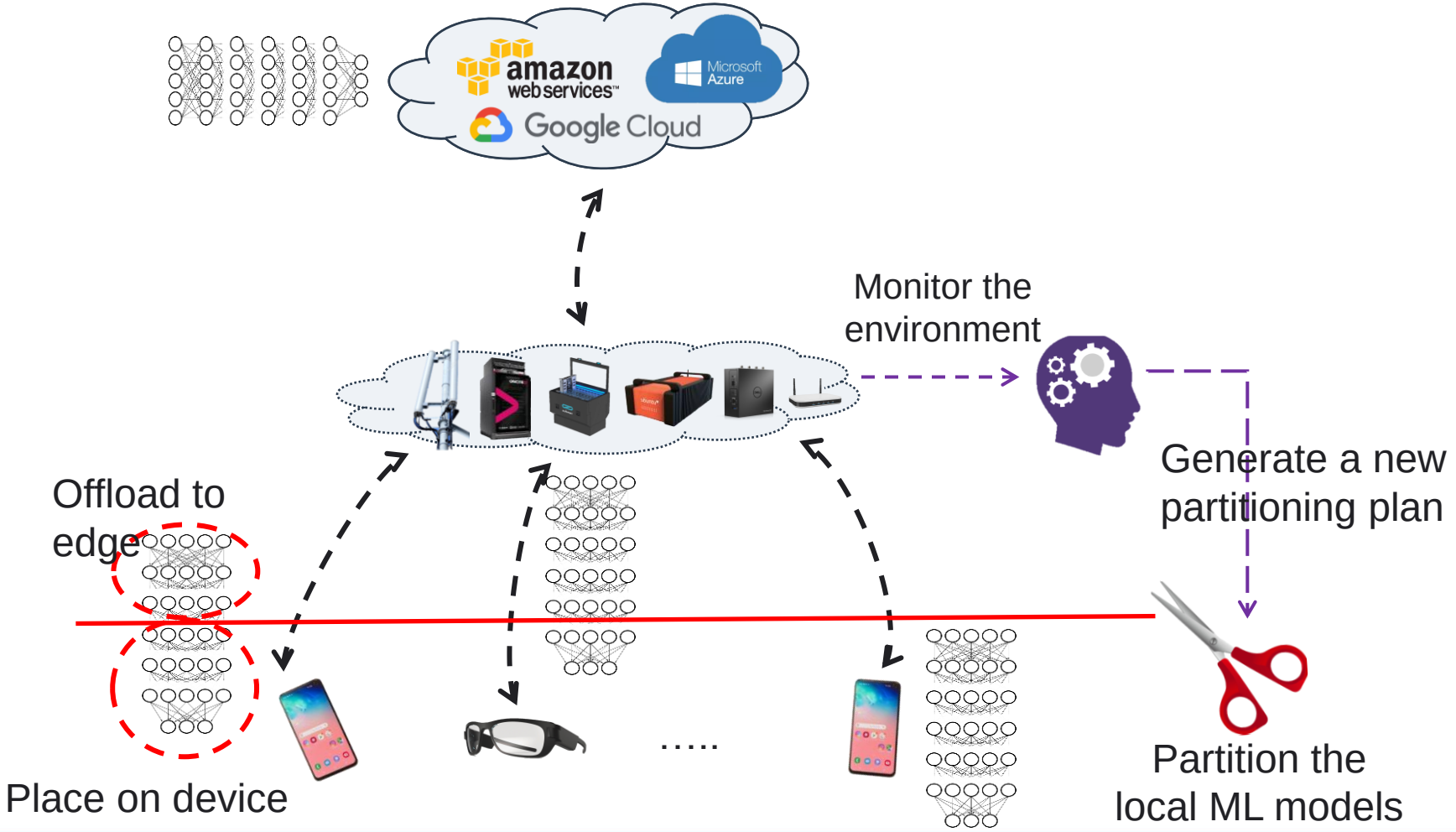
Offloading in Federated Learning



- **Technique 1**: Partition ML models and offload to edge
  - Alleviates computation burden on devices
  - Mitigates the problem of stragglers

FedAdapt

Offloading in Federated Learning



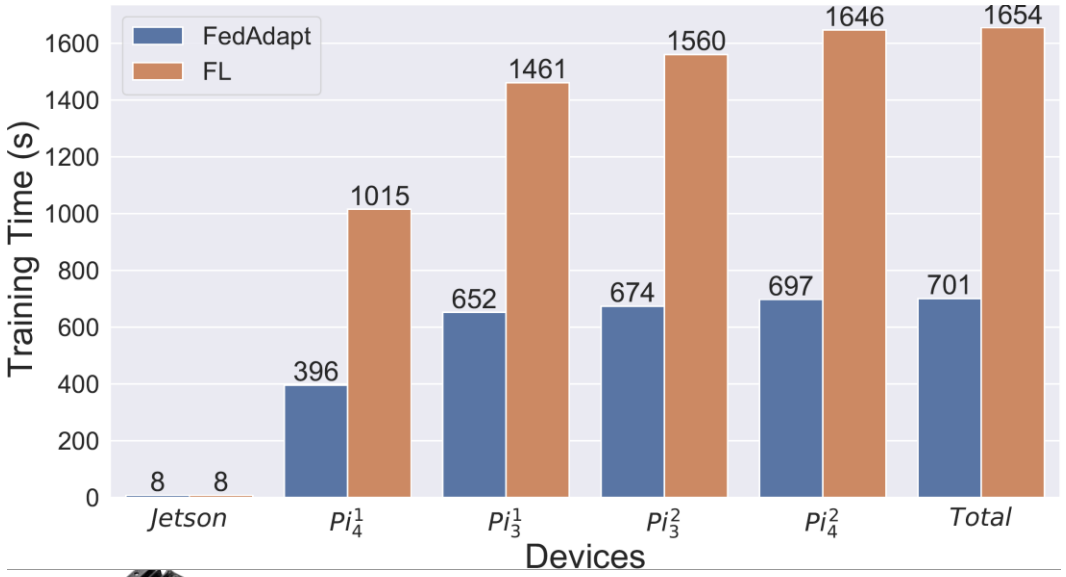
- **Technique 2**: Adaptive partitioning using reinforcement learning
  - Generates a partitioning plan that adapts to changing network bandwidth

FedAdapt

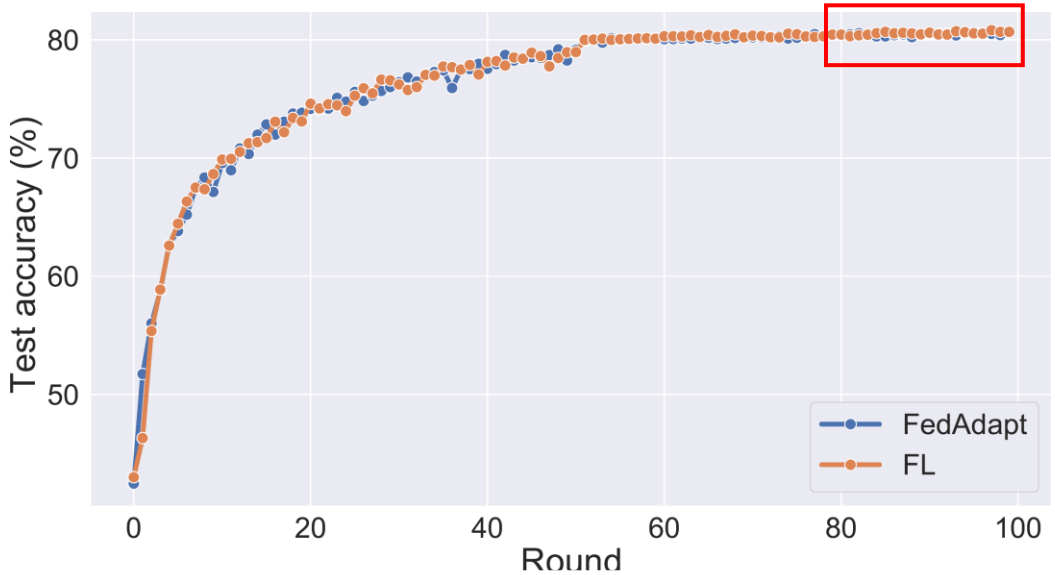
A Few Experimental Results

FedAdapt

- Used the VGG-5 and VGG-8 DNN model for Federated Learning



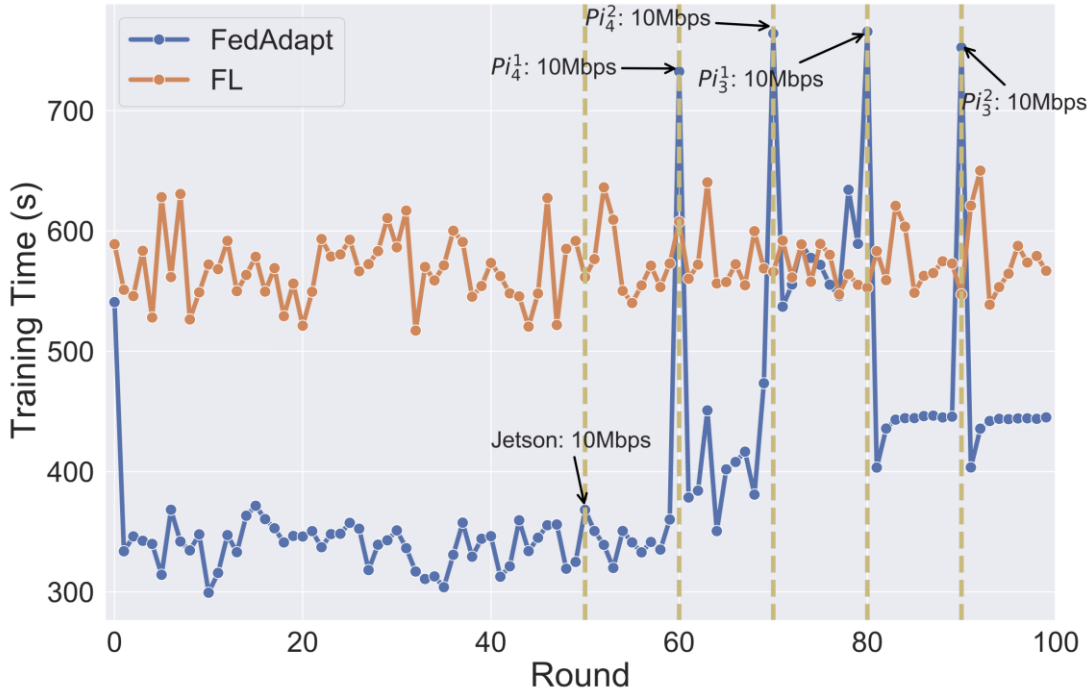
2.35x speedup over classic Federated Learning



Same overall accuracy as Federated Learning

A Few Experimental Results

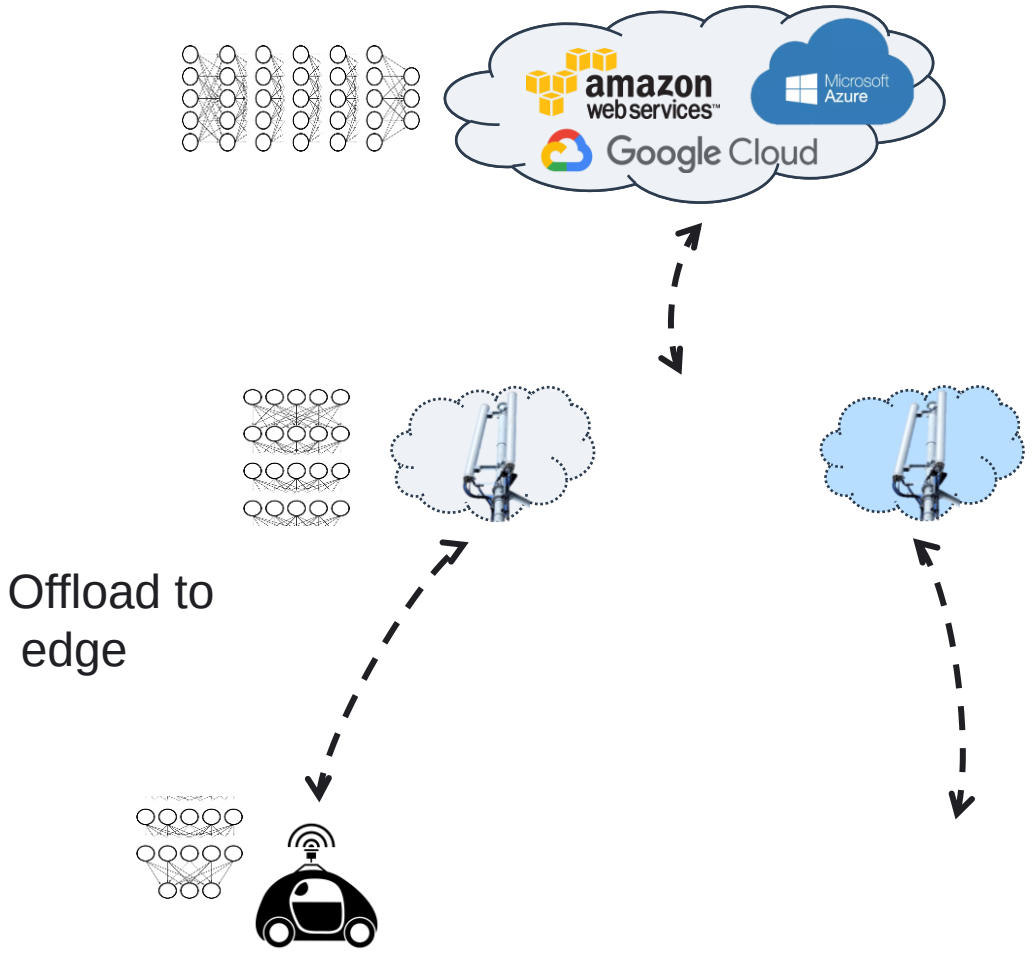
- The use of reinforcement learning for varying network connections; each vertical line is when the network connection drops to 10Mbps.



Reduces the overall training time by up to 40%

D. Wu et al. “FedAdapt: Adaptive Offloading for IoT Devices in Federated Learning,” IEEE Internet of Things Journal, 2022.

Migration in Federated Learning



- For mobile devices how do we resume training without loosing data from prior training?
  - *Building resilience into the training process*

R. Ullah et al. “FedFly: Towards Migration in Edge-based Distributed Federated Learning,” IEEE Communications Magazine, 2022.

Conclusions

- Many opportunities for systems research in edge machine learning
- Computational and communication related bottlenecks need to be addressed

- What must we do to bring training times down to sub-second without compromising accuracy?

- *Acknowledgement*
  - FedAdapt and FedFly were sponsored by Rakuten


University of
St Andrews
www.st-andrews.ac.uk