

# Topics on the Edge

Pushing the fundamental limits of federated learning on the mobile edge

- Federated Learning and Mobile Edge Computing
- Hierarchical FL
- Compute-aware Client Selection

# Federated Learning and the Mobile Edge

# The Mobile Edge

## “Enduring Challenges” of Pervasive Computing

- Resource poverty
- Communication Uncertainty
- Finite Energy
- *Multi-modal interaction*
- *Scarce user attention*
- Lower privacy, security, robustness

# Federated Learning: Natural Advantages

## “Enduring Challenges” of Pervasive Computing

- Resource poverty
- Communication Uncertainty Client Selection
- Finite Energy
- *Multi-modal interaction*
- *Scarce user attention*
- Lower privacy, security, robustness Federated Training

# Federated Learning: Natural Disadvantages

## “Enduring Challenges” of Pervasive Computing

- Resource poverty
  - Communication Uncertainty
  - Finite Energy
  - *Multi-modal interaction*
  - *Scarce user attention*
  - Lower privacy, security, robustness
- There will always be a strong case for centralization!

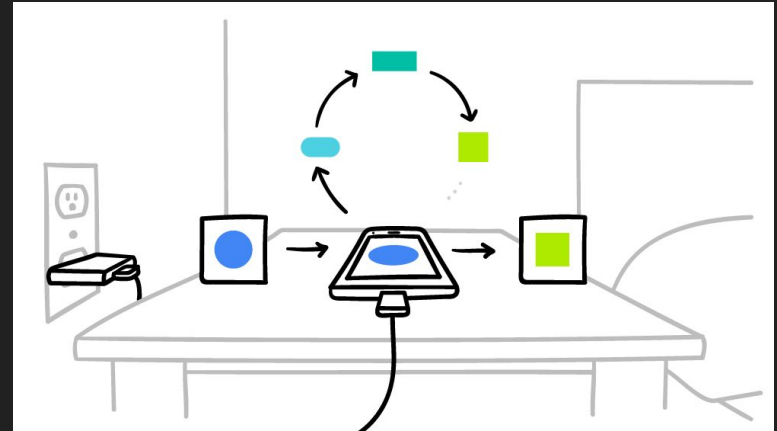
# Current Federated Learning

## Current Applications:

- Next word prediction
- Recommender systems

## Not very performance sensitive

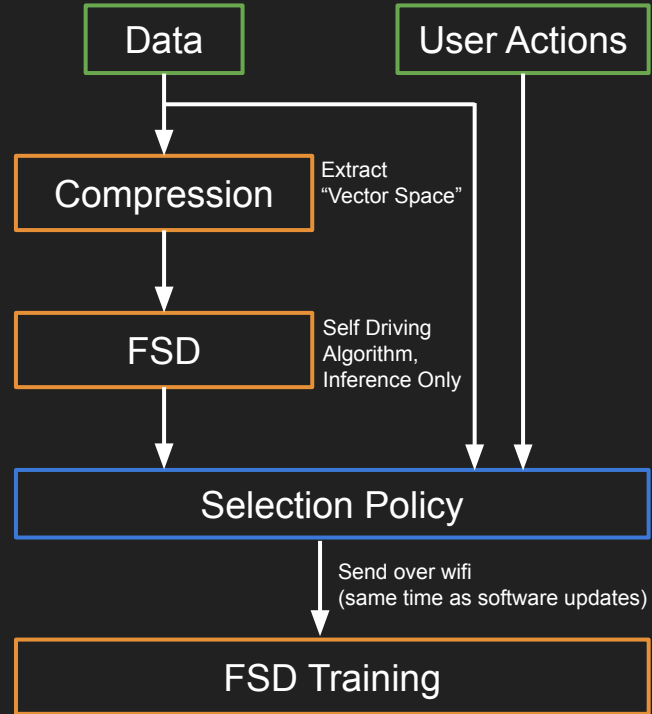
- Recruit only plugged-in, wifi-connected clients
- Research focus is mostly on accuracy, data heterogeneity, etc
- Does not hit the fundamental limits of mobile computing



# Why Mobile Edge?

# Why (Not) Mobile Edge?

## Tesla Full Self Driving Training

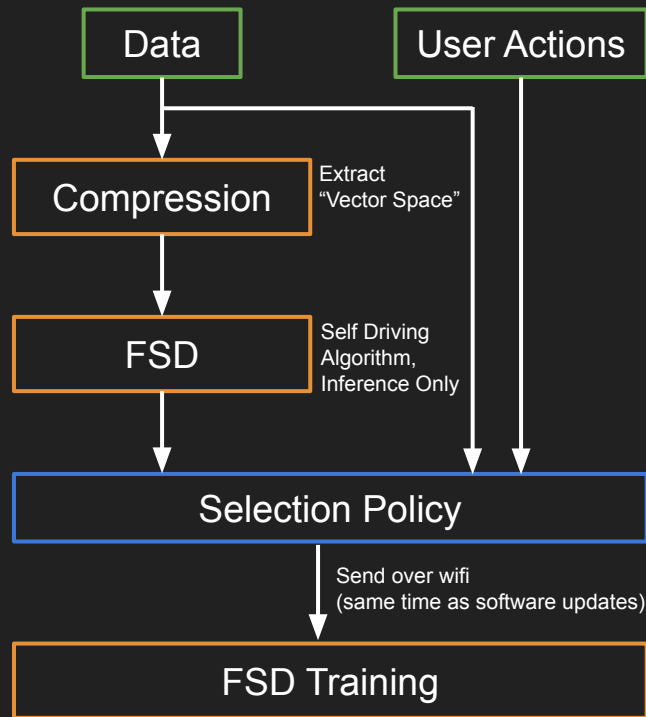




# Why (Not) Mobile Edge?

## Tesla Full Self Driving Training

- Edge training is hard/expensive
- Users don't know about privacy
- Users don't care about privacy



# No articles about privacy on the first page!

People are only mildly concerned: [Self-Driving Cars and Data Collection: Privacy Perceptions of Networked Autonomous Vehicles](#) (2017, Lujó Bauer's group)

Google tesla autopilot training data

Q All News Videos Images Shopping More Tools

About 12,800 results (0.27 seconds)

- FutureCar**  
Toyota's Woven Planet is Now 'Training' its Self-Driving ...  
Toyota's Woven Planet is Now 'Training' its Self-Driving Development Vehicles Using Data From Low-Cost Cameras, an Approach Used by Tesla.  
2 weeks ago
- Automotive IQ**  
IQ News: Toyota and Tesla's Self-Driving Approach Leads to more Affordable AVs | Automotive IQ  
Tesla has proven that collecting data with self-driving cars through ... of training data to support fully self-driving functionalities...  
2 weeks ago
- carandbible**  
Elon Musk Says In TED Video FSD Beta Has 100,000 Users  
... to the amount of training data it was gathering via real word usage. ... FSD is Tesla's successor to AutoPilot which has been largely...  
12 hours ago
- Not a Tesla App**  
2022.4.5.20 Official Tesla Release Notes - Software Updates  
For maximum safety and accountability, use of Full Self-Driving (Beta) will ... the data size of the next-gen autolabeler, training network...  
3 weeks ago
- Not a Tesla App**  
2022.4.5.21 Official Tesla Release Notes - Software Updates  
A disengagement is when the Autopilot system disengages for the remainder of ... the data size of the next-gen autolabeler, training network...  
2 weeks ago
- Analytics Insight**  
Want to Land Your Dream ML Job at Tesla? Here's How?  
The Simulation team realizes these goals through generating synthetic datasets for neural network training, building tools that enable Autopilot...  
1 week ago
- VentureBeat**  
What is autonomous AI? A guide for enterprises  
Join AI and data leaders for insightful talks and exciting networking ... One of the developers of Tesla's autopilot software, for instance...  
3 weeks ago

Google tesla autopilot training data

Q All News Videos Images Shopping More Tools

About 404,000 results (0.44 seconds)

- <https://towardsdatascience.com/tesla-ai-day-2021-revi...>  
**Tesla AI Day 2021 Review — Part 2: Training Data. How Does ...**  
Sep 23, 2021 — Tesla is combining manual labeling, auto labeling, and simulation to create real-world datasets for fully self-driving cars.
- <https://towardsdatascience.com/teslas-deep-learning-at...>  
**Tesla's Deep Learning at Scale: Using Billions of Miles to ...**  
May 7, 2019 — Tesla's advantage in training data implies an advantage in object detection, prediction, and path planning/driving policy.

**People also ask**

- Does Tesla Autopilot use deep learning?
- Is Tesla self-driving machine learning?
- Does Tesla Autopilot collect data?
- What programming language does Tesla Autopilot use?

<https://www.tesla.com> > ...  
**Artificial Intelligence & Autopilot | Tesla**  
Build AI training chips to power our Dojo system. Implement bleeding-edge technology from the smallest training nodes to the multi-die training tiles.  
[Autopilot](#) [FSD Chip](#) [Tesla Australia](#)

**Videos**

- Tesla FULL self driving explained by an engineer (with Elon ...**  
YouTube · CNET Highlights  
Aug 19, 2021
- 10 key moments in this video
- From 02:59 Biological Visual Cortex Wiring
- From 04:57 Neural Network Backbone
- From 06:12 Detection Head
- From 09:48 Problem: Per-Camera Detection...
- From 11:1 Multi-Cam Vector Spac Predictions
- Tesla AI Day**  
YouTube · Tesla  
Aug 19, 2021
- 10 key moments in this video
- Andrej Karpathy - AI for Full-Self Driving at Tesla**  
YouTube · Matroid

# Why Mobile Edge?

- Privacy and Liability
  - Increased awareness of data privacy
  - GDPR and “[Personal Data Sovereignty](#)”
  - Liability to data breaches and the difficulty in obtaining [cyber insurance](#)

# Why Mobile Edge?

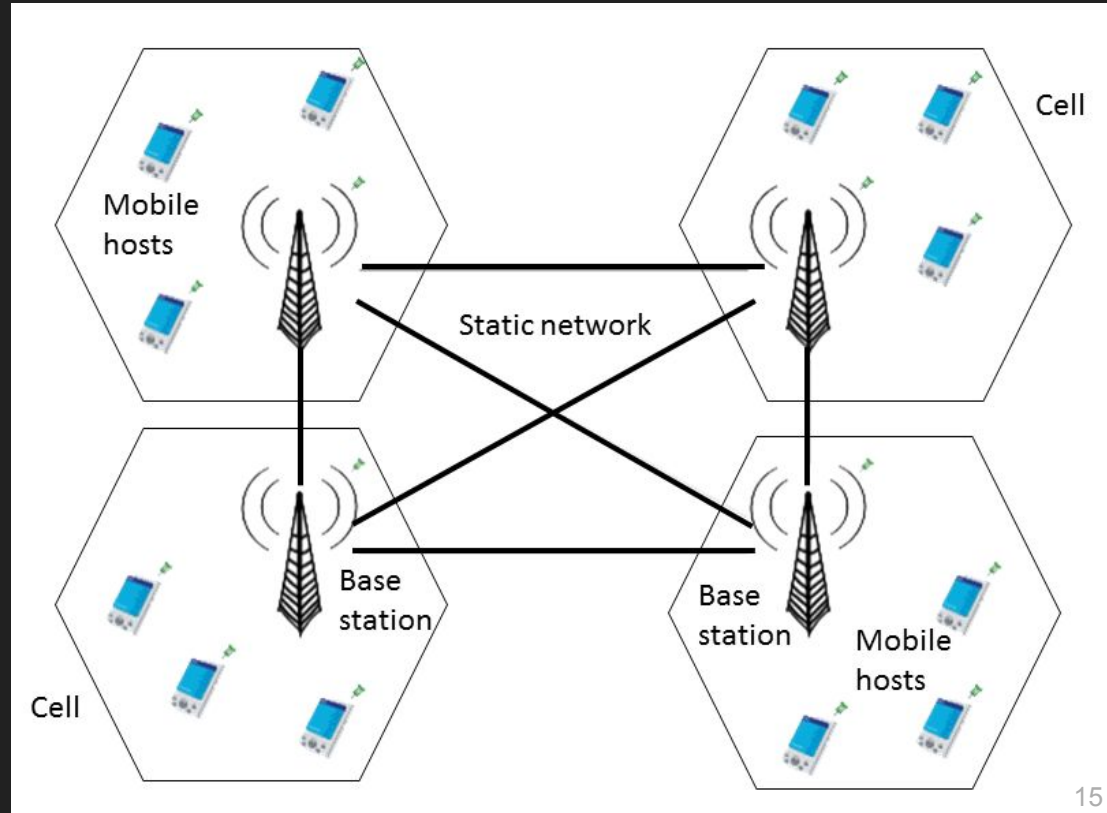
- Privacy and Liability
  - Increased awareness of data privacy
  - GDPR and “[Personal Data Sovereignty](#)”
  - Liability to data breaches and the difficulty in obtaining [cyber insurance](#)
- Communication Constraints
  - High data rate, i.e. video, LIDAR (easily 100s of gbps)
  - More domain-specific tasks where you need more data from each client

# Why Mobile Edge?

- Privacy and Liability
  - Increased awareness of data privacy
  - GDPR and “[Personal Data Sovereignty](#)”
  - Liability to data breaches and the difficulty in obtaining [cyber insurance](#)
- Communication Constraints
  - High data rate, i.e. video, LIDAR (easily 100s of gbps)
  - More domain-specific tasks where you need more data from each client
- **Rapid Iteration**
  - Relative to training time
  - Without rapid iteration, none of these constraints matter!

# Hierarchical FL

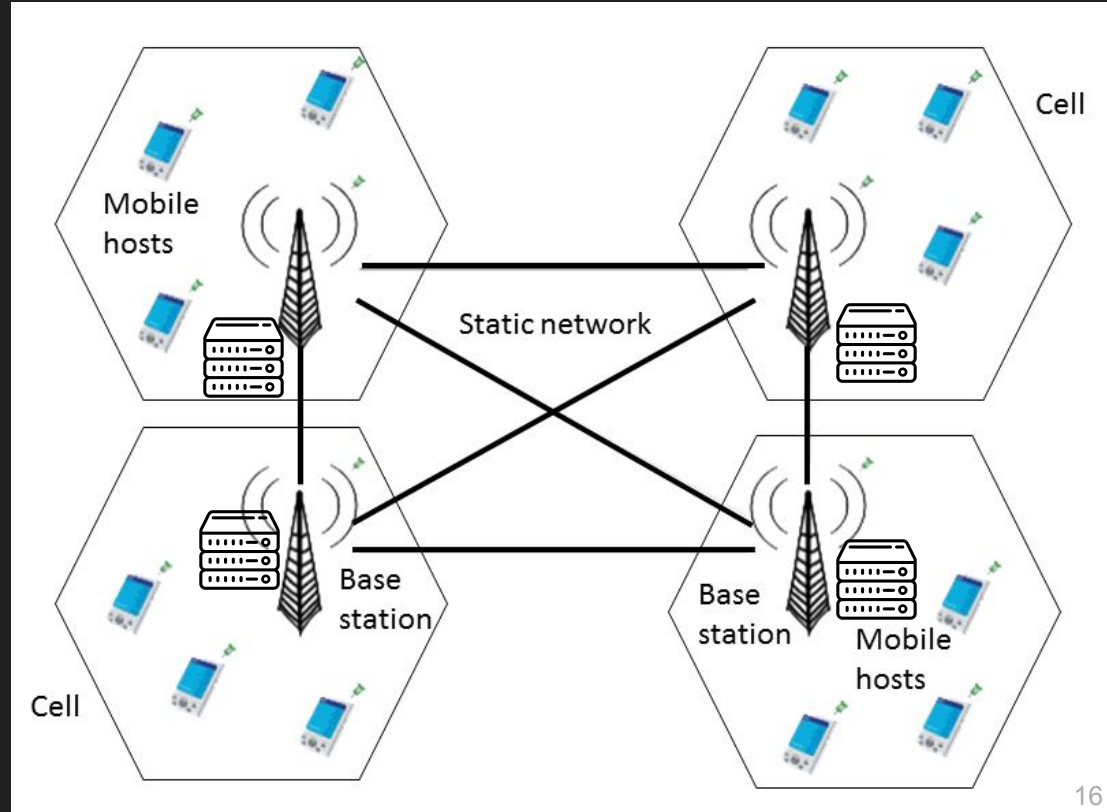
# Physical Architecture



# Just add Edge Servers!

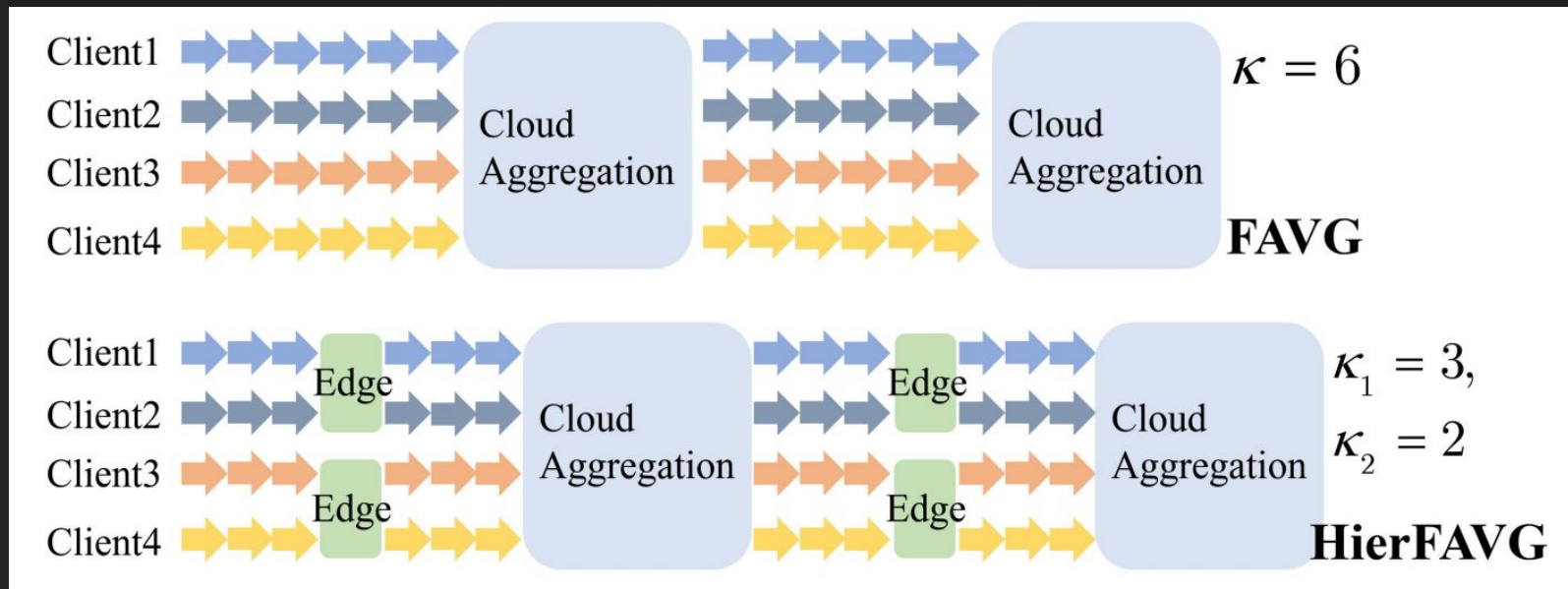
## Verizon 5G Edge

(It's an AWS virtual machine connected to the cellular network somewhere)





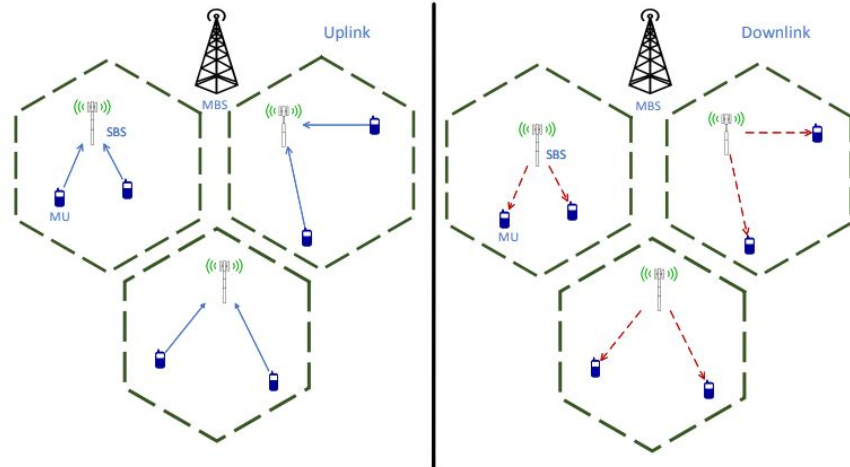
# Hierarchical Federated Learning



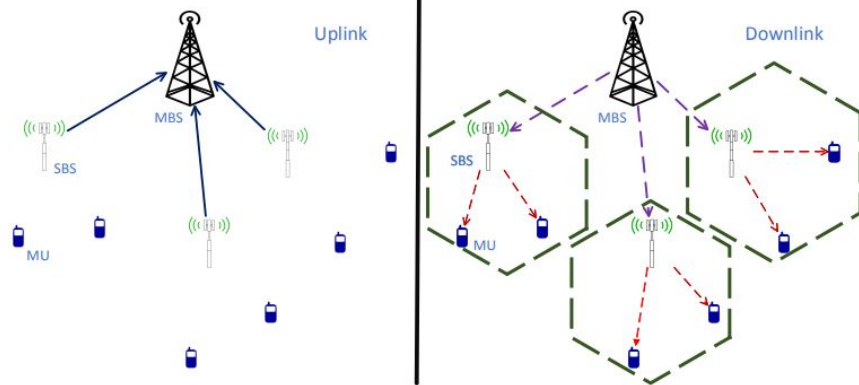
[Client-Edge-Cloud Hierarchical Federated Learning](#) (2019)

# Hierarchical Federated Learning

Hierarchical Federated Learning  
Across Heterogeneous Cellular  
Networks (2019)

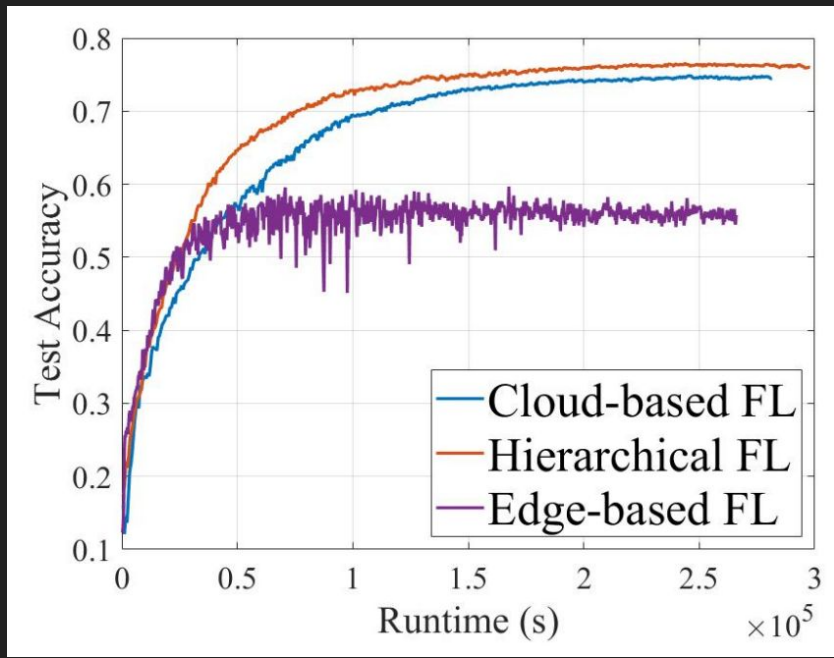


(a) Local gradient update

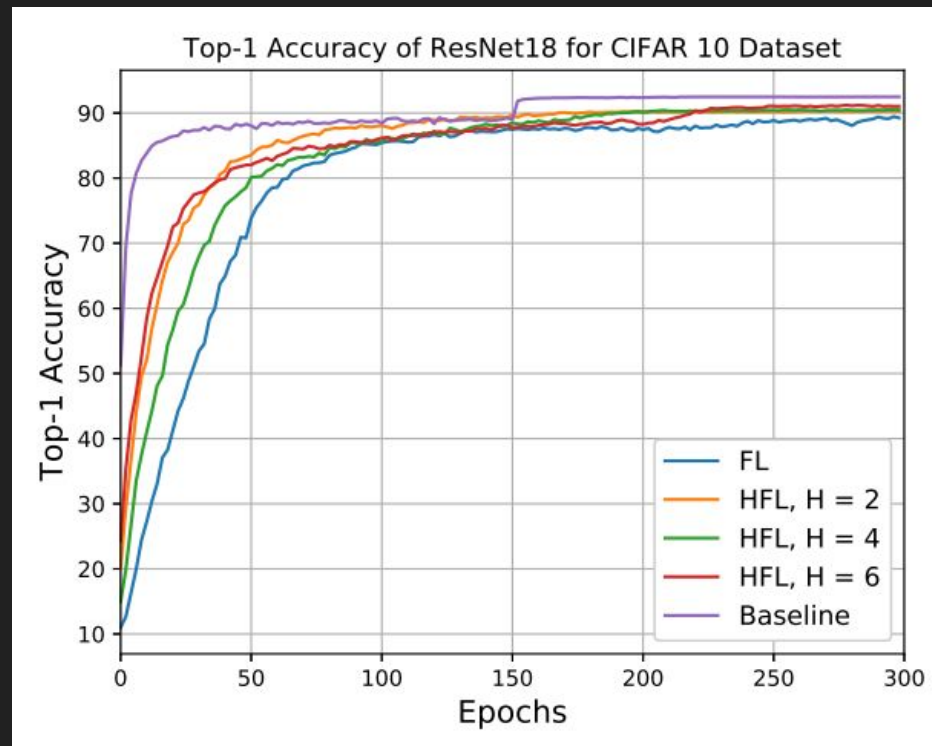


(b) Global model averaging

# More Synchronization = Faster Convergence



\*experiments use 50 and 28 clients, respectively



# Hierarchical Federated Learning

Extensions:

- [Arbitrarily many levels](#) (2022)
- [Edge aggregation server selection and scheduling](#) (2020)

Possible Ideas:

- Per-cluster adaptation (number of steps, gradient compression, etc.)
- Cluster selection and client selection
- Mixed hierarchical, non-hierarchical FL

# Why Not Now?

Hierarchical FL is “canonical,” but...

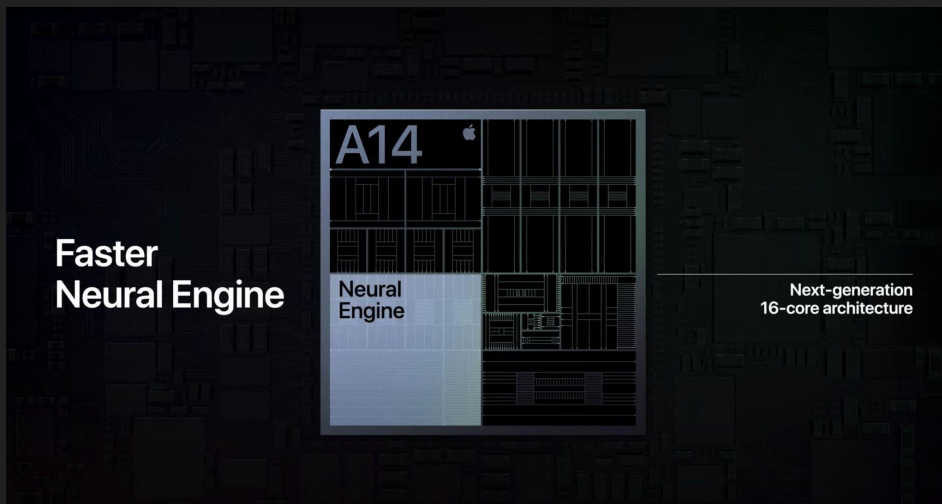
# Why Not Now?

Hierarchical FL is “canonical,” but...

- Infrastructure is not fully there yet
- Edge server deployment is difficult
- Data is billed to the mobile user, not the developer. No discount for edge communication.

# Compute-Aware Client Selection

# Today's Server, Tomorrow's Edge





# Compute-Aware Client Selection

Most obvious approach: pick all clients that will complete in time.

[Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge](#), 2019

---

**Algorithm 3** Client Selection in Protocol 2

---

**Require:** Index set of randomly selected clients  $\mathbb{K}'$

- 1: **Initialization**  $\mathbb{S} \leftarrow \{\}$ ,  $T_{\mathbb{S}=\emptyset}^d \leftarrow 0$ ,  $\Theta \leftarrow 0$
  - 2: **while**  $|\mathbb{K}'| > 0$  **do**
  - 3:    $x \leftarrow \arg \max_{k \in \mathbb{K}'} \frac{1}{T_{\mathbb{S} \cup k}^d - T_{\mathbb{S}}^d + t_k^{\text{UL}} + \max\{0, t_k^{\text{UD}} - \Theta\}}$
  - 4:   remove  $x$  from  $\mathbb{K}'$
  - 5:    $\Theta' \leftarrow \Theta + t_x^{\text{UL}} + \max\{0, t_x^{\text{UD}} - \Theta\}$
  - 6:    $t \leftarrow T_{\text{cs}} + T_{\mathbb{S} \cup x}^d + \Theta' + T_{\text{agg}}$
  - 7:   **if**  $t < T_{\text{round}}$  **then**
  - 8:      $\Theta \leftarrow \Theta'$
  - 9:     add  $x$  to  $\mathbb{S}$
  - 10:   **end if**
  - 11: **end while**
  - 12: **return**  $\mathbb{S}$
- 

In summary, Client Selection is formulated by the following maximization problem with respect to  $\mathbb{S}$ :

$$\begin{aligned} \max_{\mathbb{S}} \quad & |\mathbb{S}| \\ \text{s.t.} \quad & T_{\text{round}} \geq T_{\text{cs}} + T_{\mathbb{S}}^d + \Theta_{|\mathbb{S}|} + T_{\text{agg}}. \end{aligned} \tag{4}$$

# Compute-Aware Client Selection

[FedMCCS: Multicriteria Client Selection Model for Optimal IoT Federated Learning](#), 2021

## B. Problem Formulation

We formulate our problem as a bilevel maximization with knapsack and other constraints as follows:

$$\max_{X_S} |X_S|$$

subject to

$$\begin{cases} \forall X_{f_z^i} \sum_{r \in \{\text{CPU, Memory, Energy}\}} \text{Util}_{r}^{X_{f_z}} < \text{Budget}_r^{X_{f_z}} [co_1] \\ \forall X_{f_z^i} \sum \left( T_d^{X_{f_z}} + \text{Util}_{r=T_{ud}}^{X_{f_z}} + T_{ul}^{X_{f_z}} \right) < T [co_2] \end{cases}$$

subject to

$$\max ER_{X_{f_z^i}} = \left[ \frac{|X_{f_z} \cdot l_A|}{|X_{f_z} \cdot l_A| + |X_{f_z} \cdot l_N|} \times 100 \right] [co_3]. \quad (1)$$

# Compute-Aware Client Selection

[FedMCCS: Multicriteria Client Selection Model for Optimal IoT Federated Learning](#), 2021

Select as many clients as possible, such that:

(1) we do not exceed the resource budget

(2) we do not exceed the round time

(3) selection also maximizes clients with minority classes

## B. Problem Formulation

We formulate our problem as a bilevel maximization with knapsack and other constraints as follows:

$$\max_{X_S} |X_S|$$

subject to

$$\begin{cases} \forall X_{f_z^i} \sum_{r \in \{\text{CPU, Memory, Energy}\}} \text{Util}_{r}^{X_{f_z}} < \text{Budget}_r^{X_{f_z}} [co_1] \\ \forall X_{f_z^i} \sum (T_d^{X_{f_z}} + \text{Util}_{r=T_{ud}}^{X_{f_z}} + T_{ul}^{X_{f_z}}) < T [co_2] \end{cases}$$

subject to

Percent of “abnormal” samples

$$\max ER_{X_{f_z^i}} = \left[ \frac{|X_{f_z} \cdot l_A|}{|X_{f_z} \cdot l_A| + |X_{f_z} \cdot l_N|} \times 100 \right] [co_3]. \quad (1)$$

# Compute-Aware Client Selection

Open questions:

- How do you select the budget and round time?
- How do you reconcile compute-aware selection with fairness if data is correlated to compute in hard-to-quantify ways?
- How do you know (i.e. predict) the resource usage ahead of time, especially if a device hasn't participated recently or has never participated?

Related work: [Runtime Performance Prediction for DL](#), 2021

# Why Not Now?

- Under-studied (probably because it's really hard to study!)
- Performance-sensitive Federated Learning on the mobile edge not really used or needed yet
- Google: devices are limited in diversity, well-profiled in advance, and developers have root access

# Conclusion

## Hierarchical FL

- Obvious, simple, useful
- Very easy to implement in the lab
- Very hard to implement in the wild

## Compute-aware Client Selection

- Logical, but not so simple
- Not needed until FL becomes more popular (especially by non-privileged parties)