RECENT INVESTIGATIONS IN THE INTERSECTION OF ML AND EDGE COMPUTING

Rajeev Shorey
(Ph.D., FINAE, Dist. Scientist ACM)
Distinguished Lecturer, IEEE Future Networks TC

CSE Department, IIT Delhi, India

(Formerly IBM, GM and TCS Research)

www.rajeevshorey.com

Illinois Institute of Technology, USA
21 November 2023

IEEE Future NETWORKS™

Enabling 5G and Beyond

IEEE

IEEE Future Networks – FutureNetworks.ieee.org

Collaboration

Content

IEEE Future Networks Tech Focus
Issue 16, June 2023

+ technical newsletter, podcasts, videos, articles

Events

+ more!

Research & Education

IEEE INGR International Network Generations Roadmap

+ eLearning, webinar series, white papers, tutorials

Join today! bit.ly/fntc-join

IEEE Future Networks

Be connected to IEEE Future Networks to shape future network requirements

Get monthly updates on technical workshops, summits, webinars, podcasts, and call for proposals, papers, and volunteer opportunities

Thousands are already members

Join today: bit.ly/fntc-join

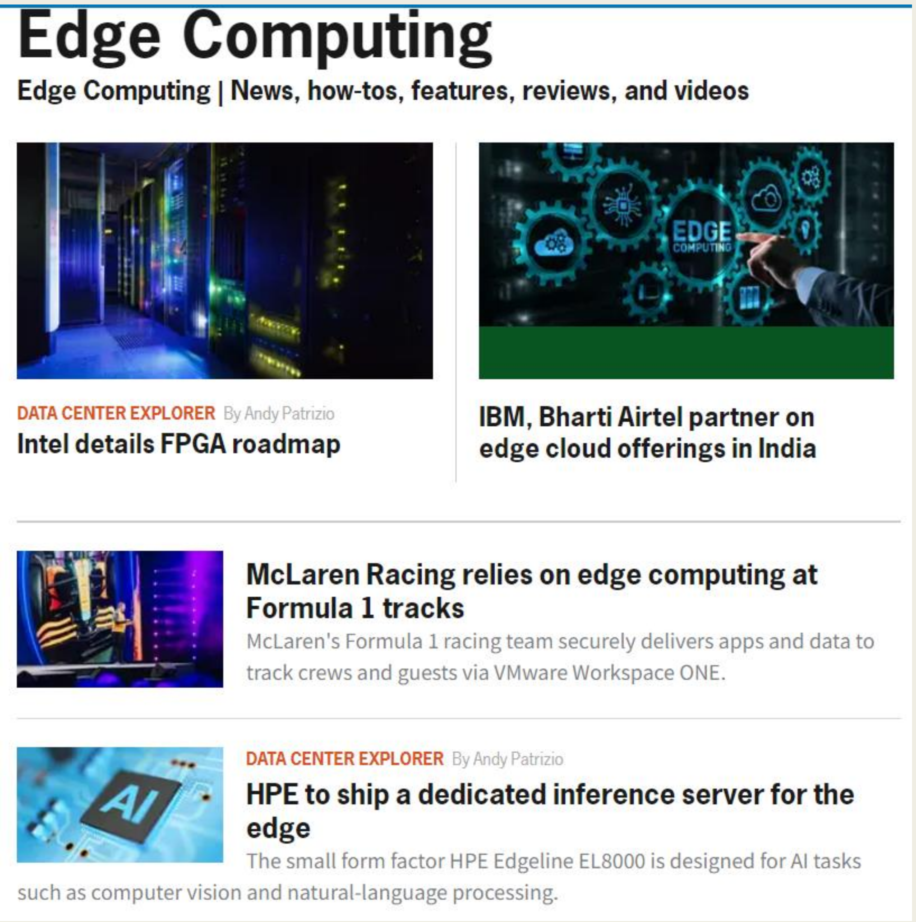
Agenda of the Talk

- Introduction & Motivation
- A Systems perspective
  - *Edge Intelligent Systems*
    - Federated Learning
- Recent Investigations in the Intersection of ML and Edge Computing
  - *Part 1 (Primary)*
    - **Federated Learning at the Edge Nodes**
  - *Part 2 (Snapshot)*
    - Splitting of CNNs on Resource Constrained Edge Devices
- Challenges and Future Research Directions
- Conclusion

Acknowledgements

- Computer Science & Engineering Department, IIT Delhi
  - *Manupriya Gupta, CMU*
  - *Pavas Goel, Graviton*
  - *Sourabh Bansal, FinTech*
  - *Manav Bansal, FinTech*
- Computer Science & Engineering Department, DTU, Delhi
  - *Tanmay Jain*
  - *Avaneesh*
  - *Ishan Prakash*
  - *Aniruddh Bansal*
- Dr. Rohit Verma, Intel Research, Bangalore
- Prof. Huzur Saran, CSE Department, IIT Delhi

The Buzz on Edge Computing



Edge Computing

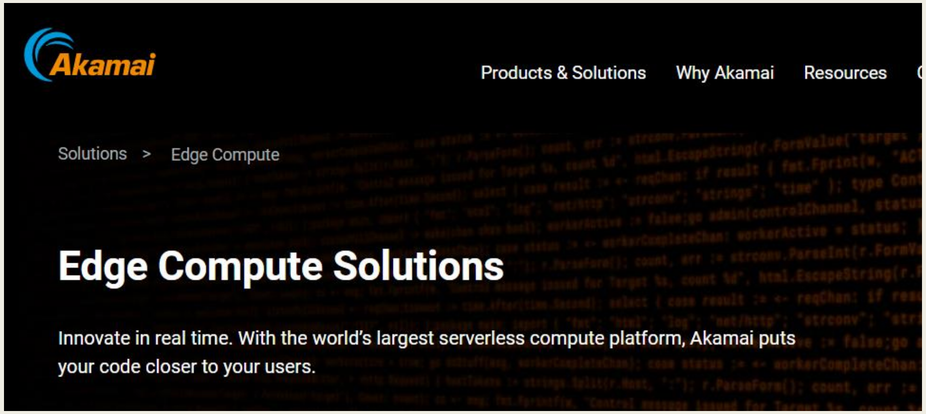
Edge Computing | News, how-tos, features, reviews, and videos

DATA CENTER EXPLORER By Andy Patrizio
Intel details FPGA roadmap

IBM, Bharti Airtel partner on edge cloud offerings in India

McLaren Racing relies on edge computing at Formula 1 tracks
McLaren's Formula 1 racing team securely delivers apps and data to track crews and guests via VMware Workspace ONE.

DATA CENTER EXPLORER By Andy Patrizio
HPE to ship a dedicated inference server for the edge
The small form factor HPE Edgeline EL8000 is designed for AI tasks such as computer vision and natural-language processing.



Akamai

Products & Solutions Why Akamai Resources

Solutions > Edge Compute

Edge Compute Solutions

Innovate in real time. With the world's largest serverless compute platform, Akamai puts your code closer to your users.

Edge Computing

Edge Computing | News, how-tos, features, reviews, and videos

CLOUD COMPUTING By David Linthicum
Cloud computing is reinventing cars and trucks

CLOUD COMPUTING By David Linthicum
The dirty little secret about edge computing

NEW TECH FORUM
Why edge computing matters for modern software development
The next stage of cloud computing brings computing power closer to users, paving the way to better user experiences and more intelligent applications.

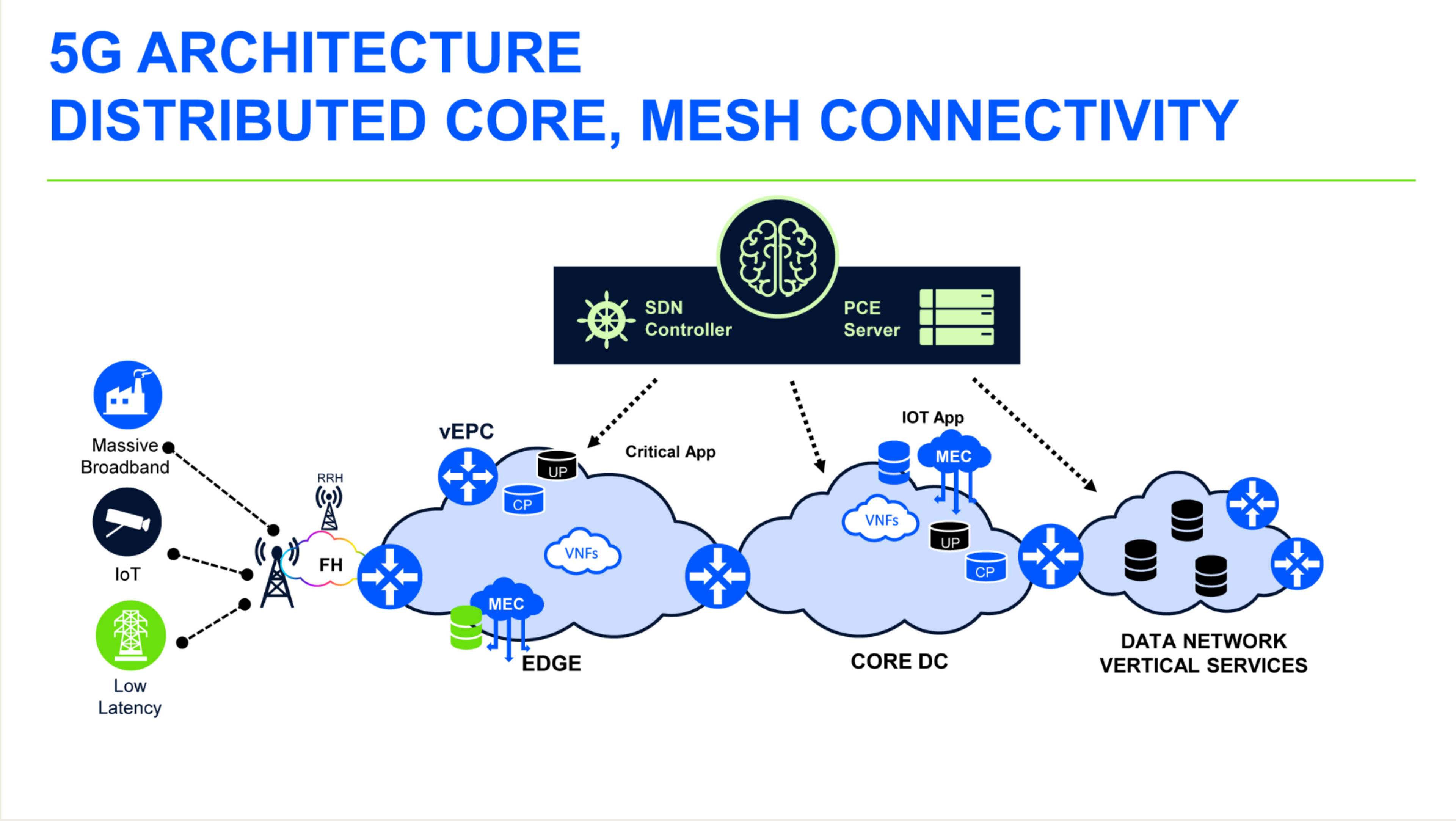
The 5G Vision: Three Broad Use Cases

The three broad use cases include enhanced mobile broadband, mission-critical services and massive IoT

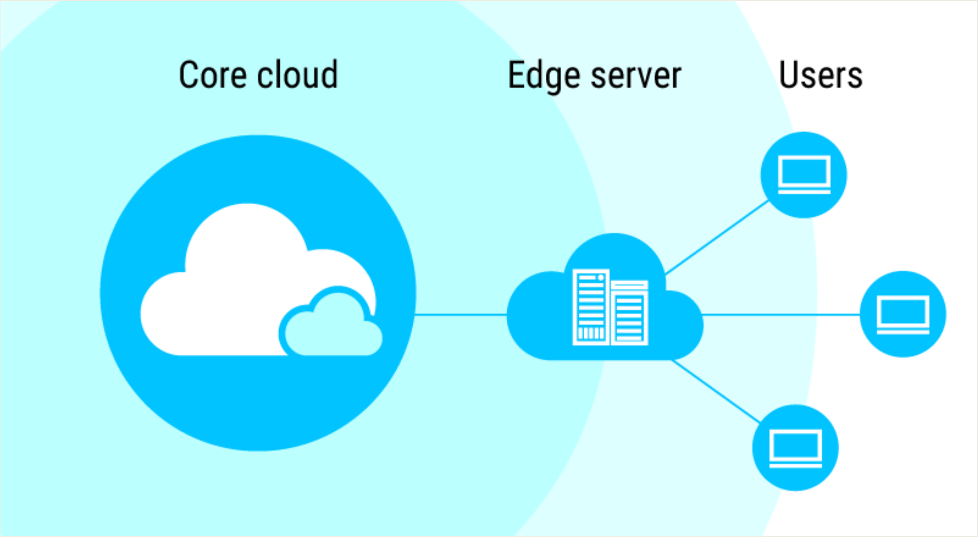
Ref: Leading the World to 5G, Qualcomm Technologies, Inc, 2016

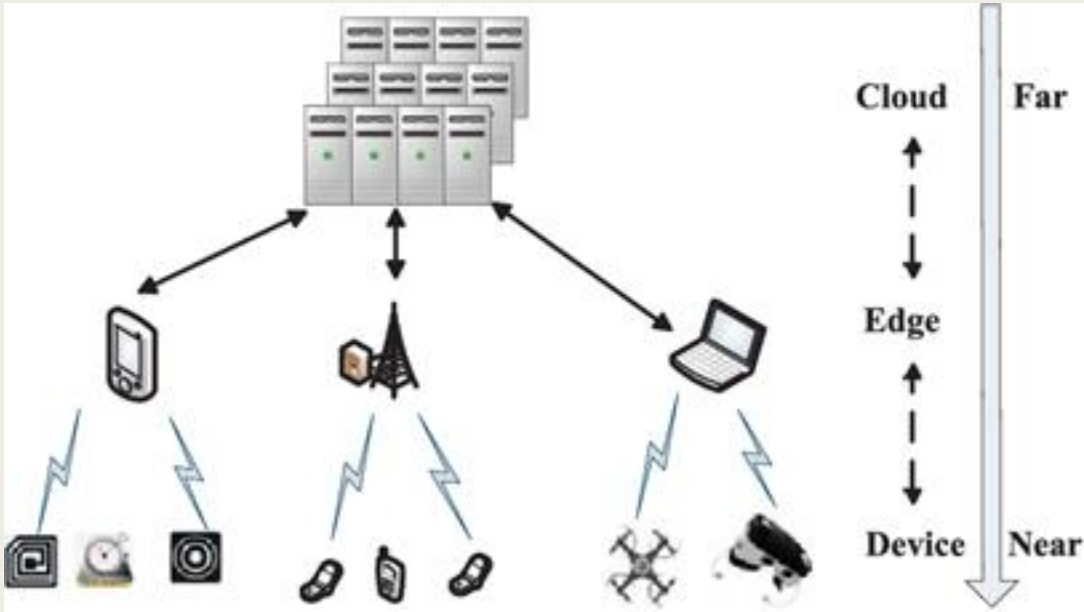
The three broad use cases are characterized by different metrics and parameters

The 5G Architecture



The Edge Nodes Play a Key Role in Enabling 5G



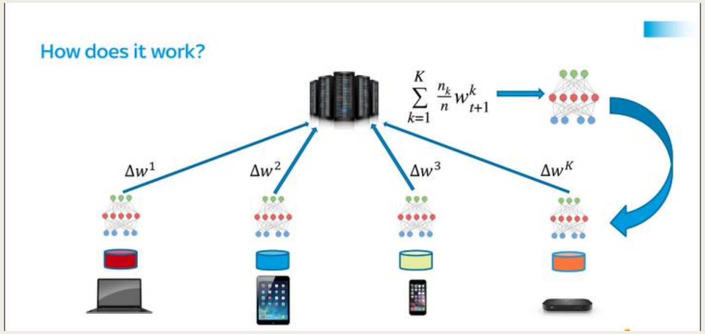


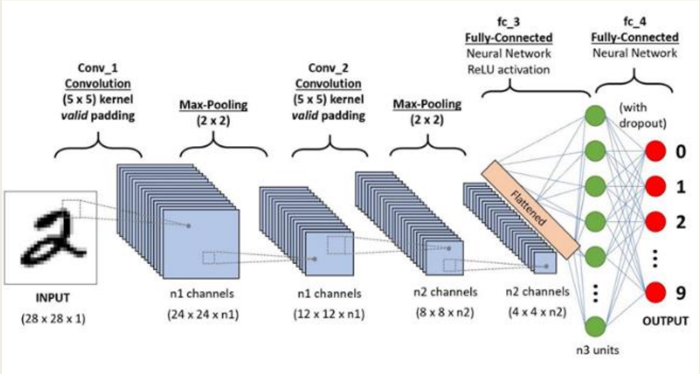
Edge Computing: Key Advantages

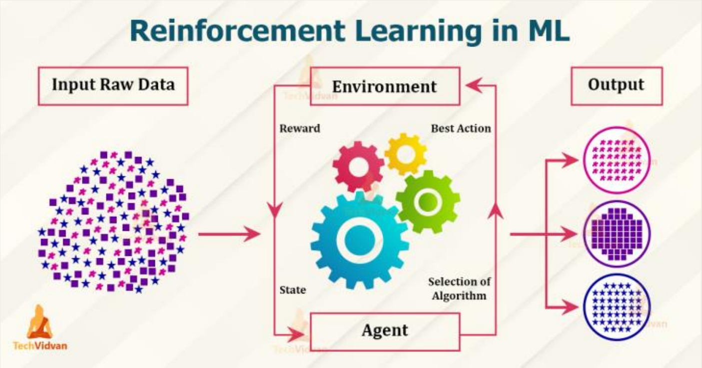
- Low Latency
- Backend Traffic Reduction
- Efficient Data Management
- Rapid Access to Data Analytics
- Network Load Reduction
- Cloud Cost Reduction

AI / ML / Deep Learning at the Edge Nodes

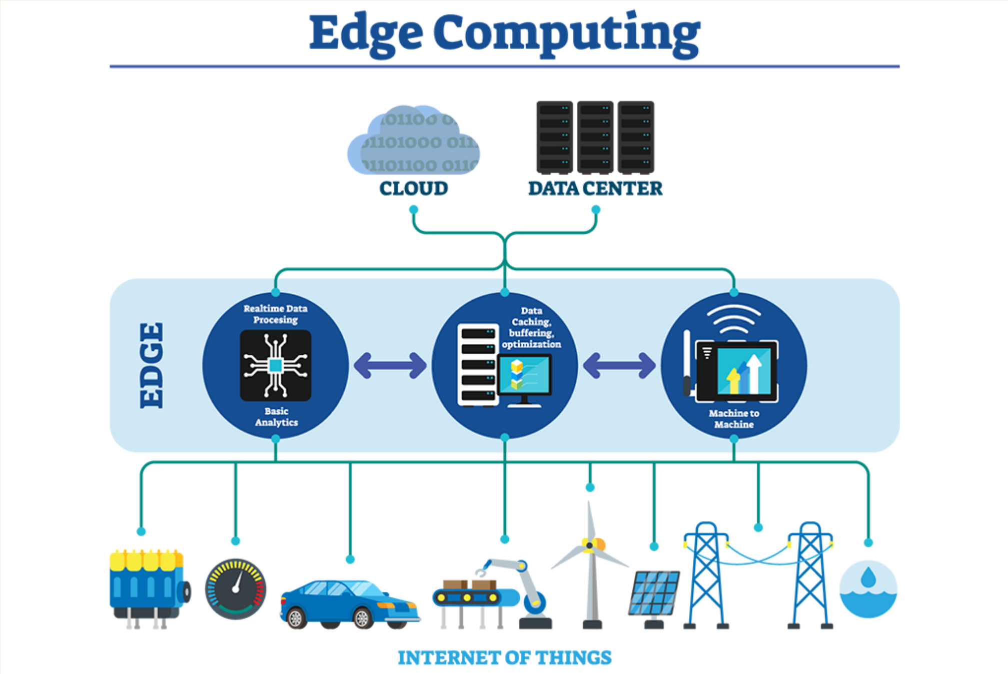
Learning at the Resource Constrained Edge Nodes





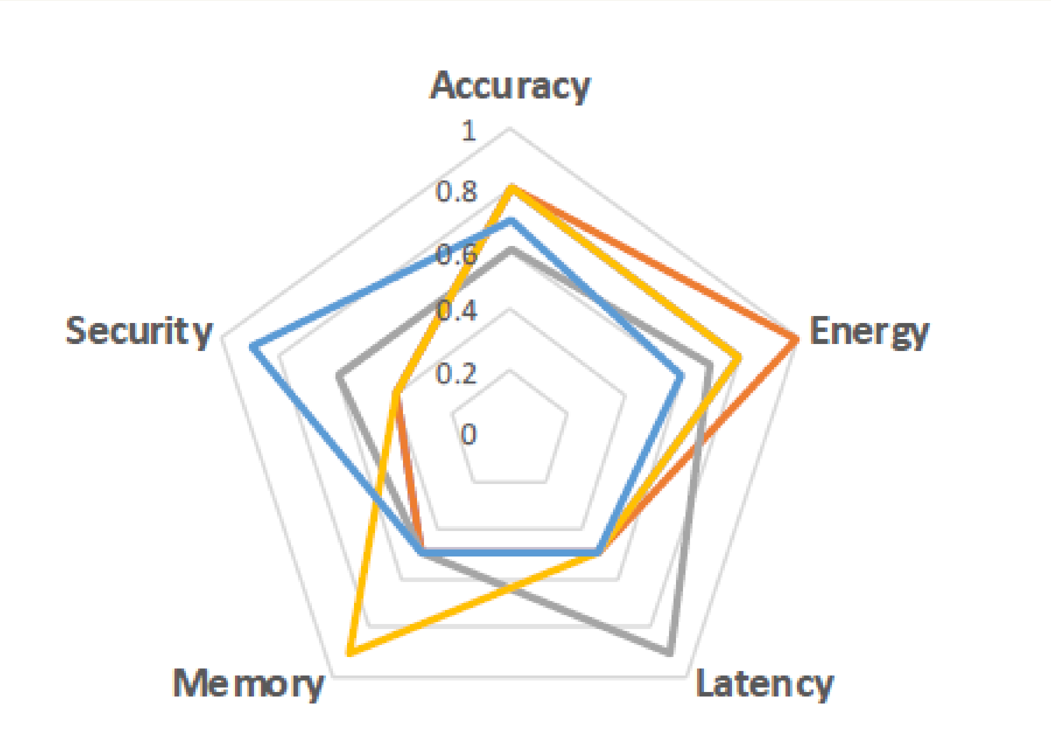


Resource Constrained Environment



Critical to understand the performance of the DL / FL / RL at the Edge Nodes

Design Space for Edge Intelligent Systems



FEDERATED LEARNING:

A PRIVACY PRESERVING PARADIGM

The Buzz on Federated Learning

Google is using federated learning to improve Assistant's "Hey Google" accuracy

ReportLinker

The Global Federated Learning Market size is expected to reach $198.7 Million by 2028, rising at a market growth of 11.1% CAGR during the forecast period

MIT News
ON CAMPUS AND AROUND THE WORLD

Collaborative machine learning that preserves privacy

Researchers increase the accuracy and efficiency of a machine-learning method that safeguards user data.

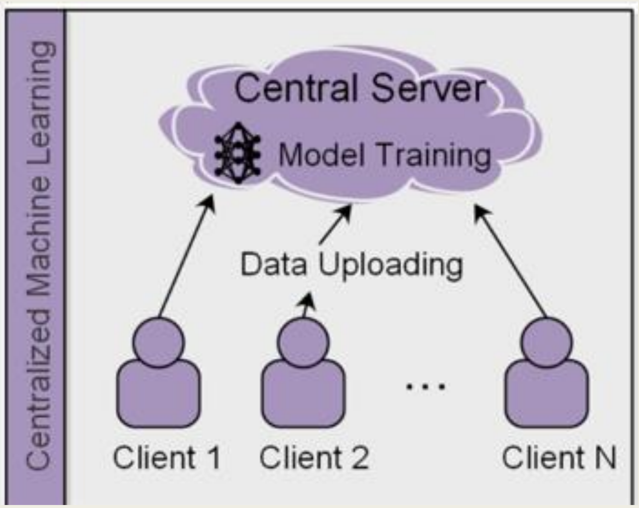
Adam Zewe | MIT News Office
September 7, 2022

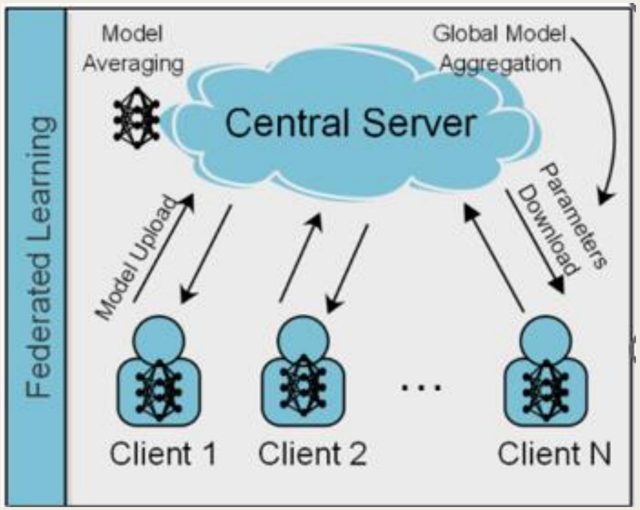
Applications of Federated Learning

- Application in the Healthcare Industry
- Applications for FinTech
- Applications in Insurance Sector
- Applications in IoT
- Application in other Industries and Technologies

CLASSICAL MACHINE LEARNING VERSUS FEDERATED LEARNING

- Central machine learning
  - move the data to the computation
- Federated (machine) learning
  - move the computation to the data



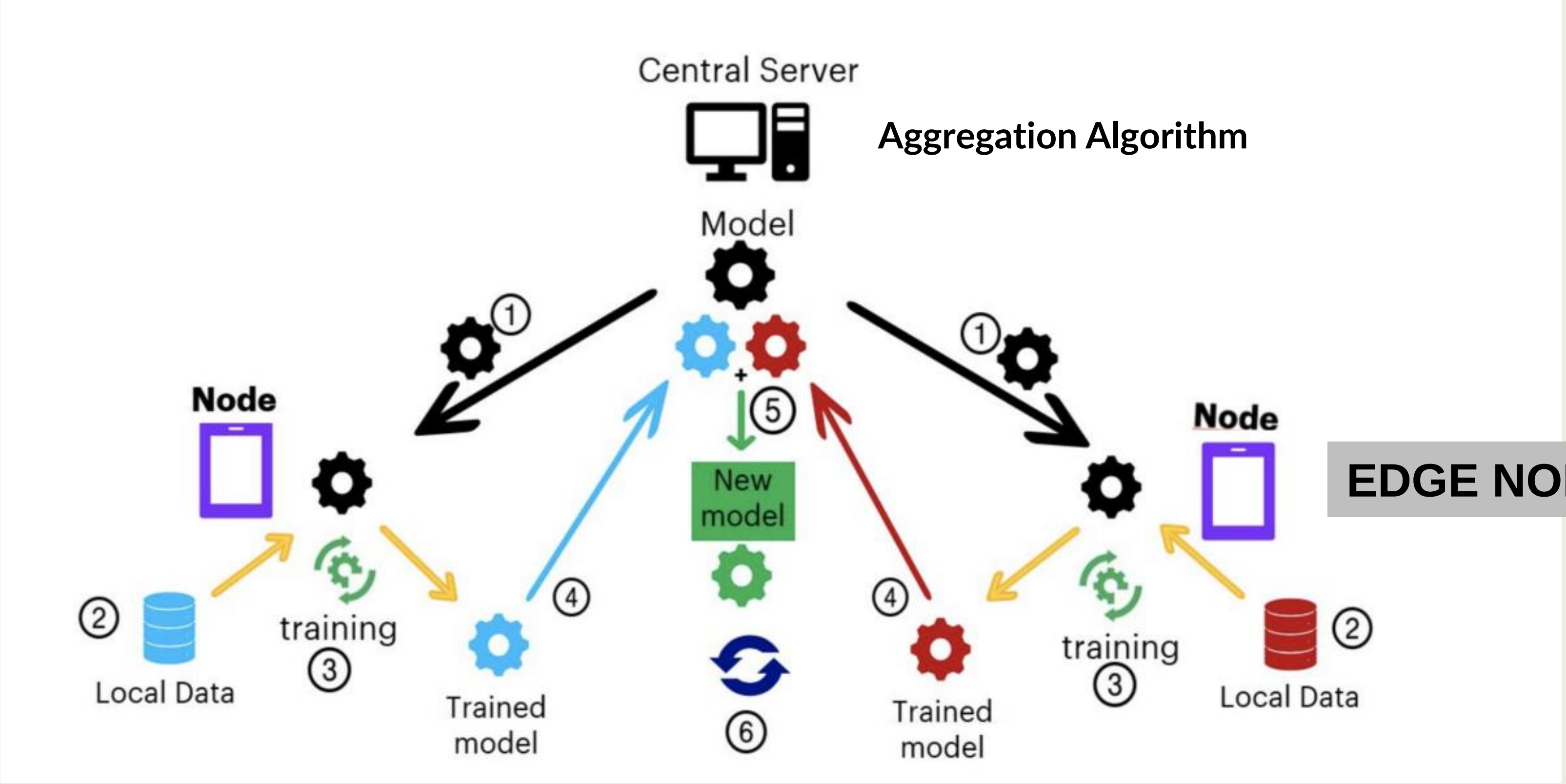


FEDERATED LEARNING IN A FAULTY EDGE ECOSYSTEM: ANALYSIS, MITIGATION AND APPLICATIONS

Work in Progress

Federated Learning

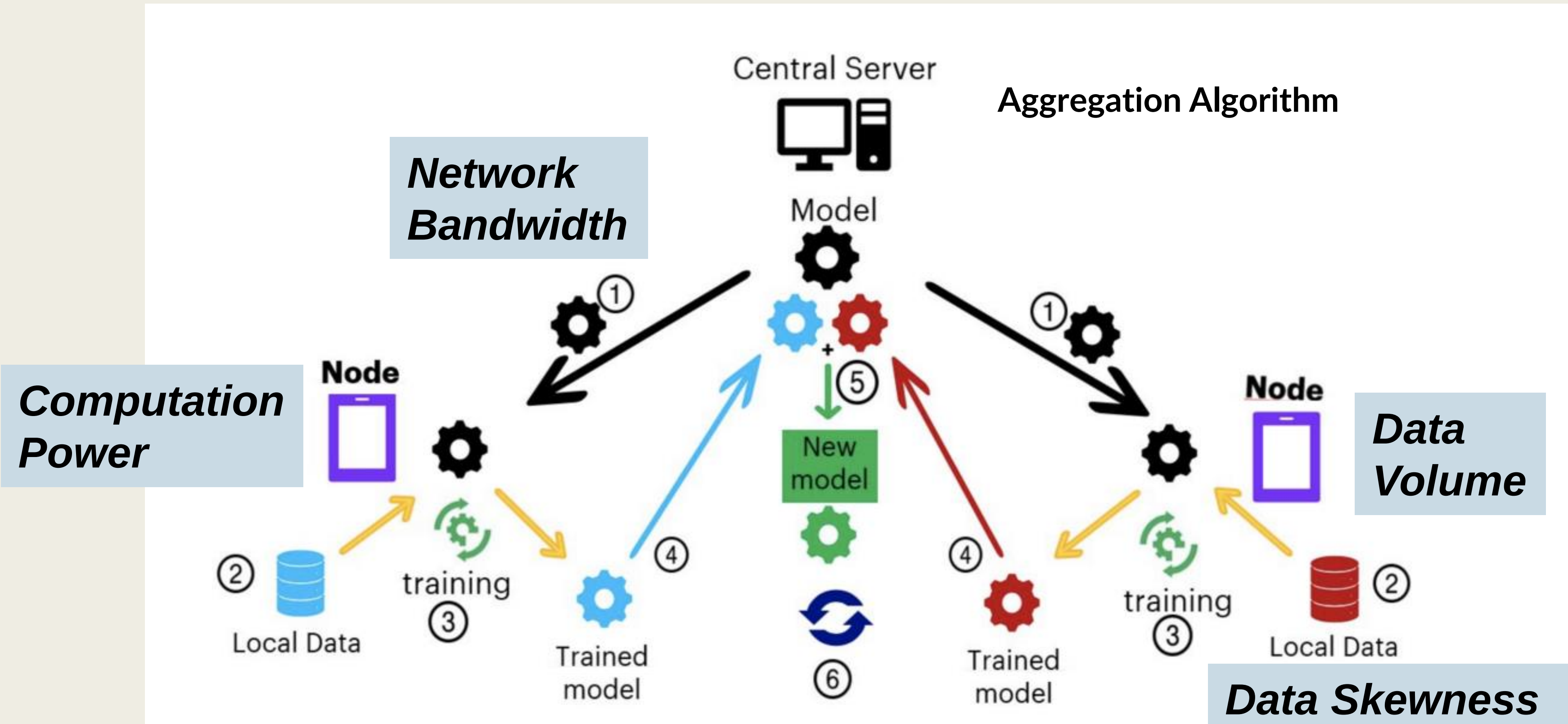
Distributed System with ML Model Exchange



EDGE NODES

FL Key Objective: Privacy Preserving Paradigm !

Federated Learning & Network Parameters



FL Performance is also a function of the System Parameters

WHAT IS THE PERFORMACE OF FEDERATED LEARNING?

ASSUMPTIONS

- “**Synchronous**” Federated Learning
- The FL system is “**Secure**”
- The architecture is “**Static**”

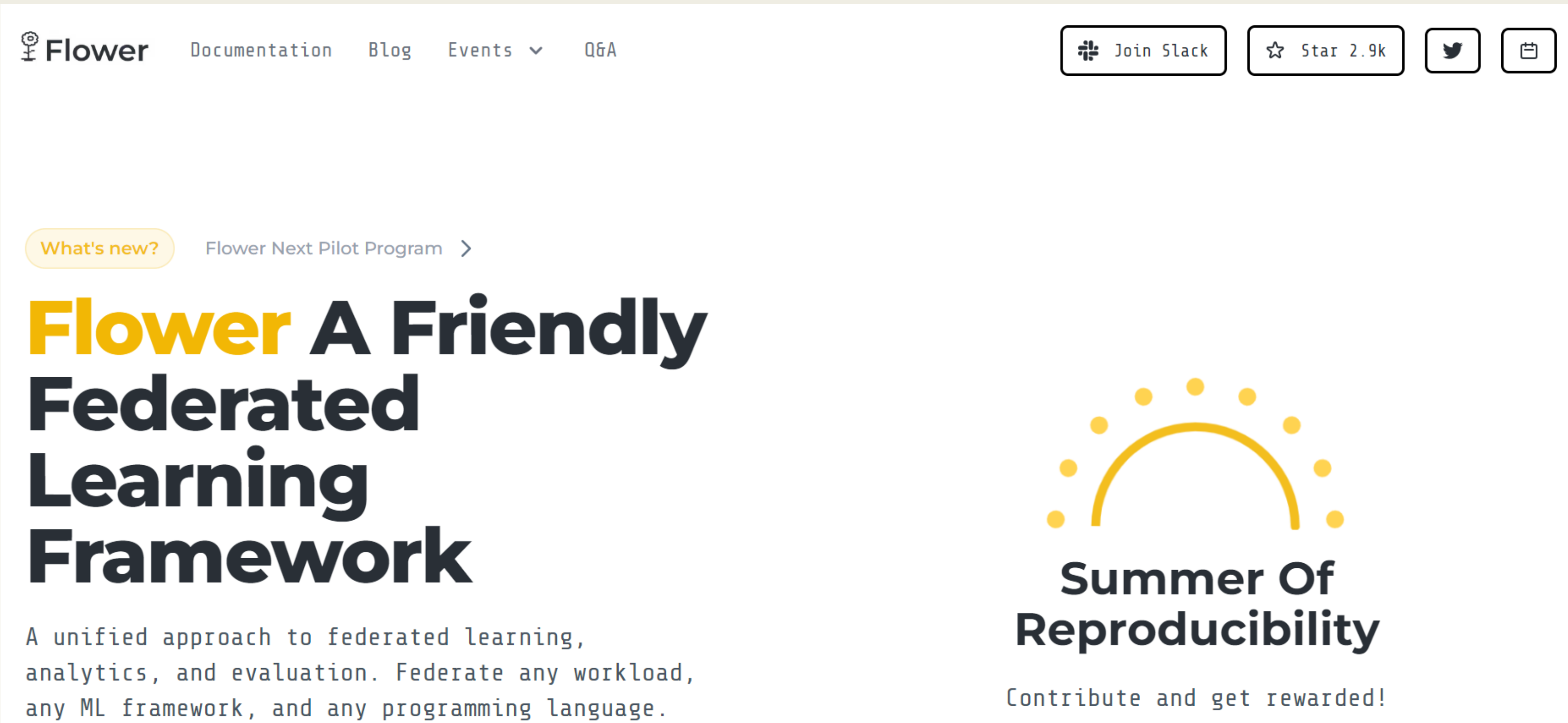
Metrics, Models and Data Sets

- **Metrics**
  - *Accuracy*
  - *Convergence Time*
- **Diverse Data Sets**
  - *MNIST*
    - Database of handwritten digits and contains 60,000 training images and 10,000 testing images
  - *CIFAR-10*
    - Consists of 60000 32x32 colour images in 10 classes, with 6000 images per class
  - *IoT Security Dataset*
    - From Kaggle
- **Diverse Models**
  - *AlexNet, ResNet, LeNet, …*

Simulation & Prototype Setup

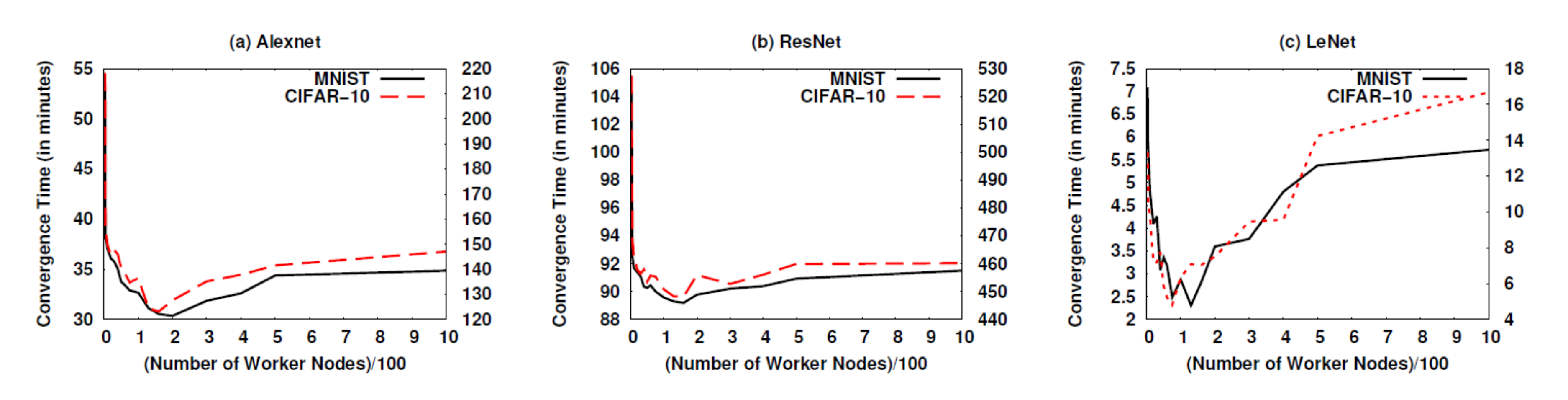
- **Simulation Setup**
  - *Pysyft*
  - *Simulations are run on an Ubuntu 20.04 system*
  - *12 GB RAM, Octa-core*
  - *1.5 GHz processor 16 GB Nvidia T4 GPU*
- **Prototype Setup**
  - *8 Raspberry Pi4 devices having 4 GB RAM quad-core 1.5 GHz processor*
  - *2 RPis have a storage of 8 GB*
  - *2 RPis have a storage of 4 GB*
  - *4 RPis have a storage of 2 GB*
  - *The aggregator is run on a Ubuntu 20.04 system with an 8 GB RAM and Octa-core 1.5 GHz processor*
  - *4 RPis (8 GB, 4 GB and two 2 GB) are connected to the aggregator over a WiFi network having a bandwidth of 10 Mbps*
  - *Other four are connected through an Ethernet line of 100 Mbps*

Flower: Federated Learning Framework



Impact of Worker Count on the Convergence Time for Different Learning Models

Left Y-axis: MNIST, Right Y-axis: CIFAR-10

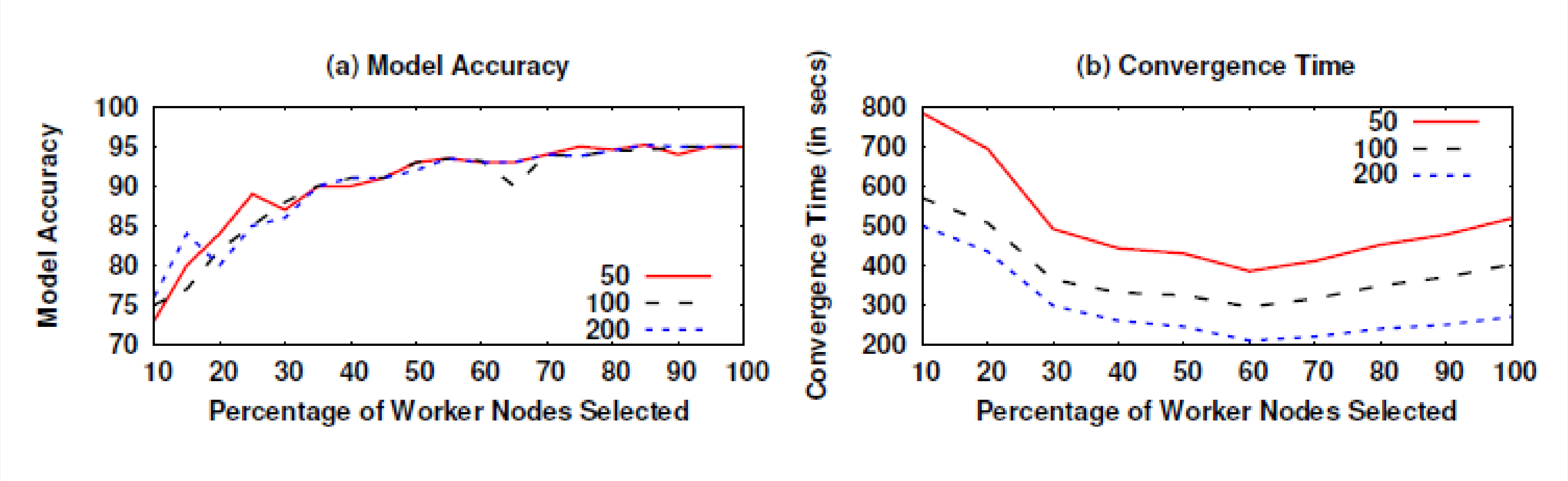


Homogeneous Data Distribution

Key Takeaways

- The number of worker nodes is crucial for FL model
- Optimal number of Worker Nodes for better working of the model

Model Accuracy and Convergence Time with % worker nodes selected



Homogeneous Data Distribution

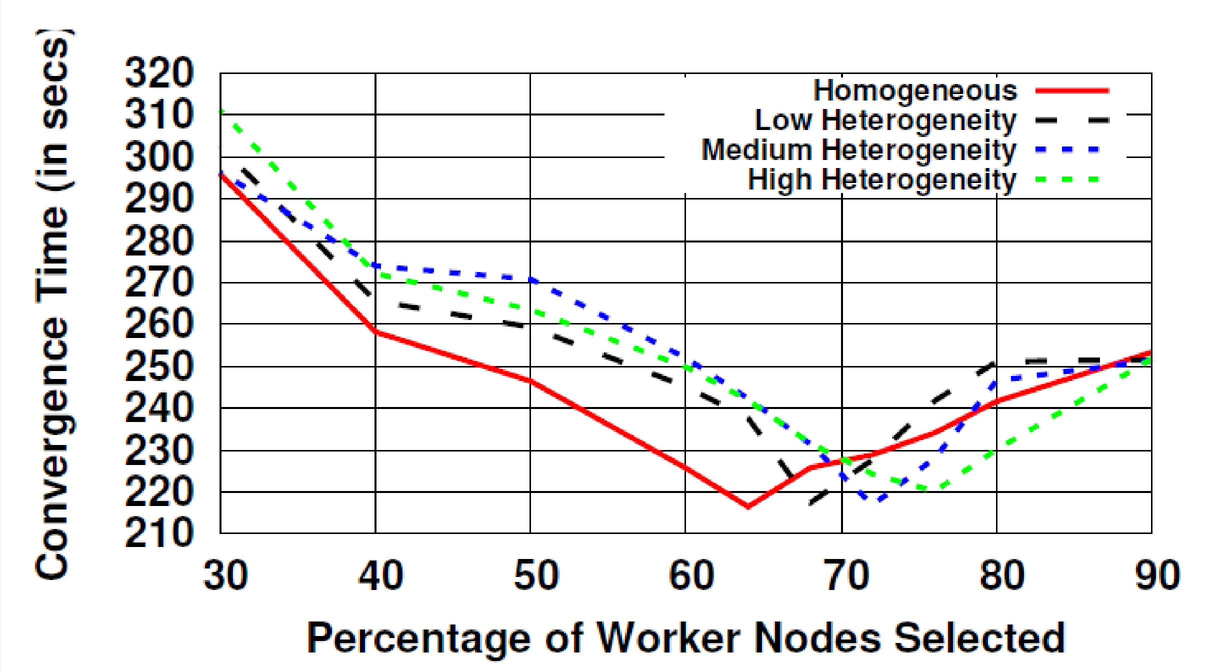
Key Takeaways

- At around 60% of worker nodes, *A* is almost similar to what it is at 100%
- On the contrary, the same 60% of nodes require *C* almost 25% less than what it takes when using all worker nodes

Hereafter, for all experiments we use 60% of the total worker nodes to contribute to the training process

WHAT HAPPENS WHEN WE HAVE HETEROGENEITY?

Variation of Convergence Time with % Worker Nodes Selected for Different Level of Heterogeneity



Heterogeneous Data Distribution:
Varying the Volume of Data at each Worker Node

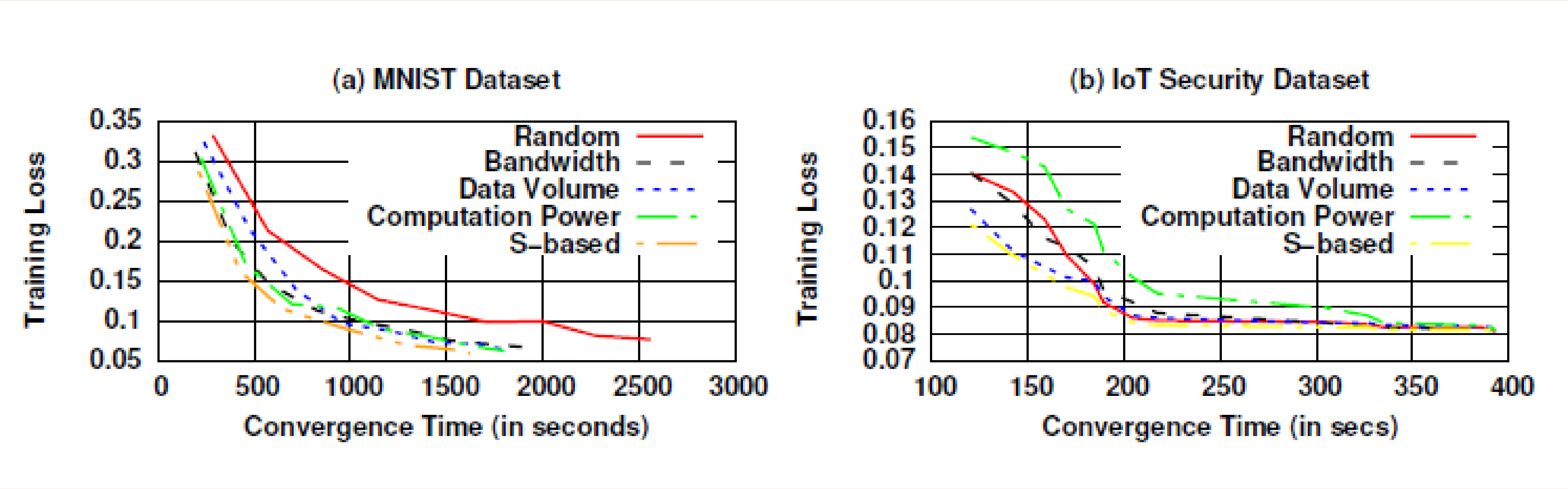
Key Takeaways

- The minimal convergence time shifts towards a higher % Worker Nodes as the heterogeneity increases

- *The degree of heterogeneity impacts the optimal number of worker nodes*

WHAT ARE THE RIGHT
EDGE NODE SELECTION STRATEGIES?

Convergence Time of the FL Model when the Top 60% Nodes are Selected for Five Selection Strategies



MNIST and IoT Security Datasets

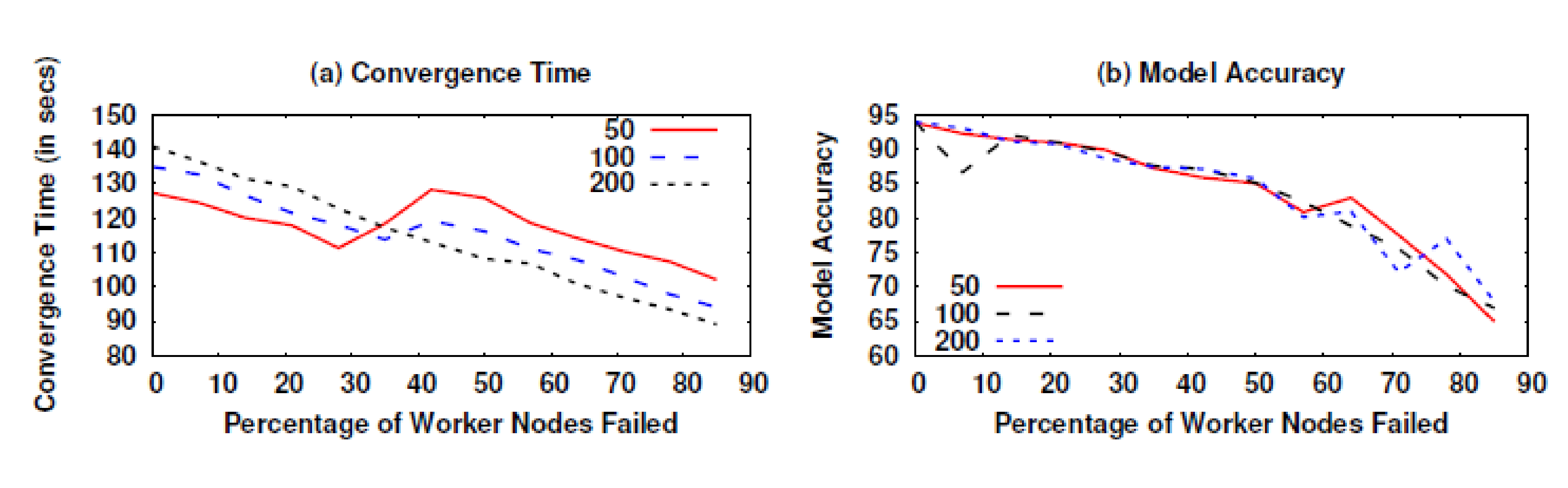
<u>Selection Score (S)</u>
Determines the top 60% Worker Nodes

$$\mathcal{S} = \left(\frac{\alpha}{\mathcal{B}} + \frac{\kappa * \mathcal{V}}{\mathcal{P}}\right) * \frac{1}{\mathcal{V}}$$

Key Takeaway

S – based selection strategy converges faster than the other naive strategies

Model Accuracy and Convergence Time for the FL Model when a % of Worker Nodes in set Fail

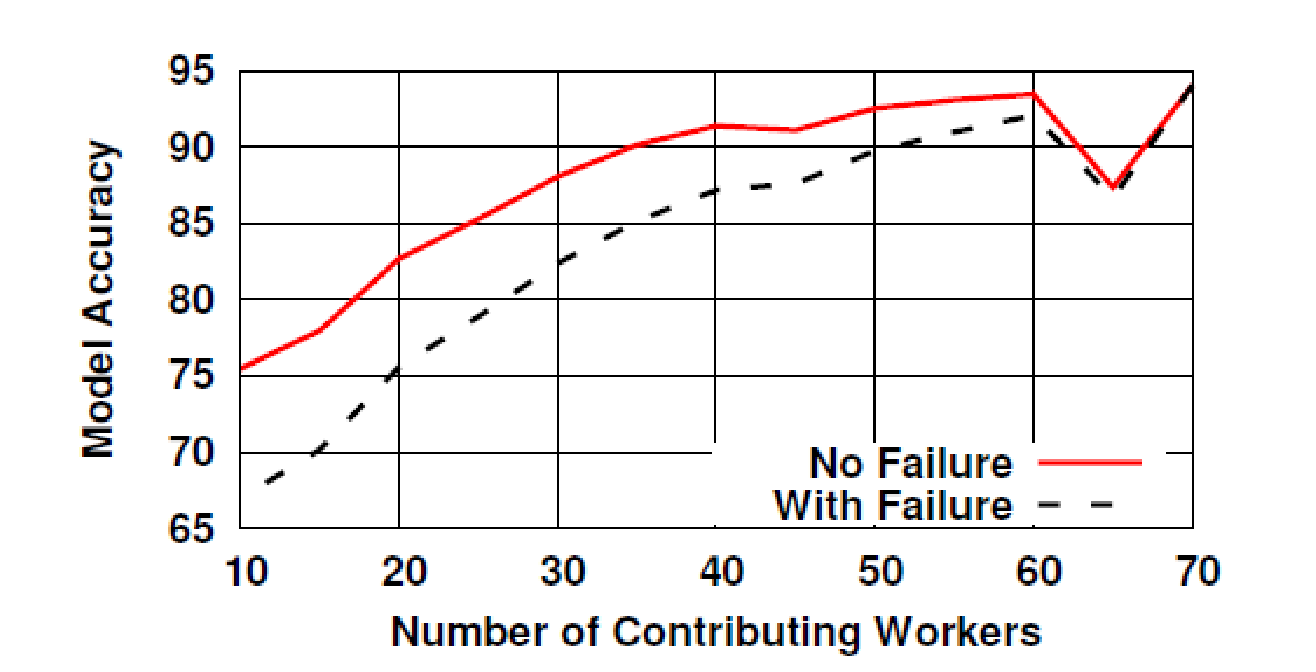


Worker Failure Analysis

Key Takeaways:

- *C* decreases with increasing W nodes that fail, however, *A* decreases too!
- The learning model does not converge to the state-of-the-art accuracy for the given model

Accuracy of the FL model for the same Number of Contributing Worker Nodes for Failure and No-failure Cases



Key Takeaways:

- We see lower accuracy in the scenario where nodes fail
- The failed nodes might have some crucial data samples which when removed due to worker node failure reduces *A*

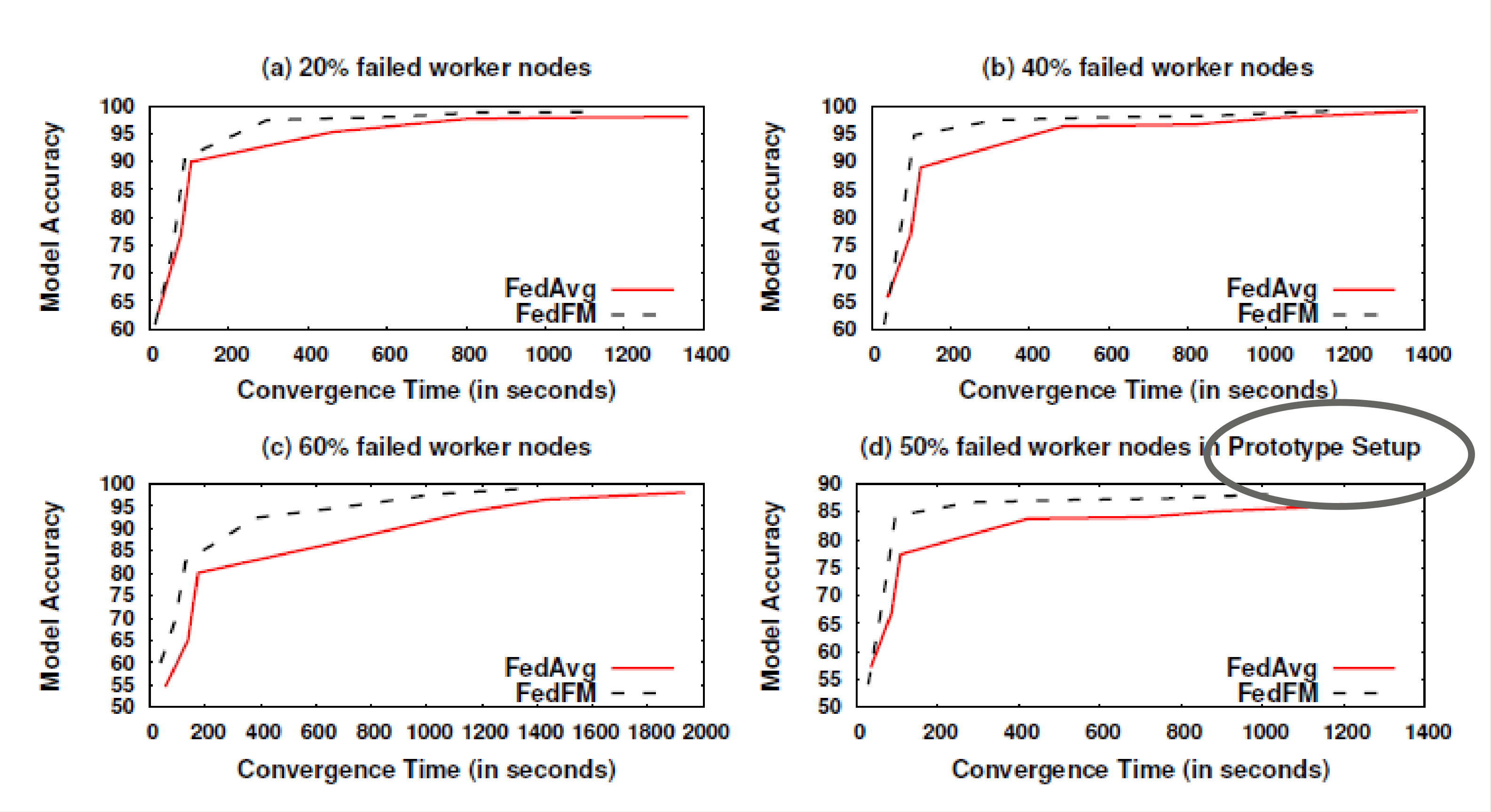
The Federated Fault Mitigation Algorithm (FedFM) Run on the Aggregator

Algorithm 1: Federated Fault Mitigation Algorithm (*FedFM*) run on the Aggregator. *ClientUpdate* (k, ω) [12] is the same function used by FedAvg.

Result: The Global Federated Learning Model with weight ω_{t+1}

$\omega_0 \leftarrow$ initialized model weights
$\mathcal{W} \leftarrow 0.6$ `// Fraction of total nodes to be selected`
$\mathcal{F} \leftarrow \{\}$
foreach *round* $t \in 1, 2, \ldots$ **do**
  $m \leftarrow max(\mathcal{W} * \mathcal{K}, 1)$;
  $\mathcal{N}_t \leftarrow$ Select top m workers based on $\mathcal{S}$.;
  foreach *client* $k \in \mathcal{N}_t$ *in parallel* **do**
    $\omega_{t+1}^k \leftarrow$ *ClientUpdate* (k, ω_t);
    if $\omega_{t+1}^k = null$ *after time* $\mathcal{T}$ **then**
      Append k to $\mathcal{F}$;
    end
  end
  if $|\mathcal{F}| > 0$ *and* $m > 1$ **then**
    $\mathcal{N}_t^f \leftarrow$ Select top $|\mathcal{F}|$ workers based on $\mathcal{S}$.;
    foreach *client* $k \in \mathcal{N}_t^f$ *in parallel* **do**
      $\omega_{t+1}^k \leftarrow$ *ClientUpdate* (k, ω_t);
    end
  end
  $\omega_{t+1}^k \leftarrow \sum_{k+1}^{m} \frac{n_k}{n} \omega_{t+1}^k$
end

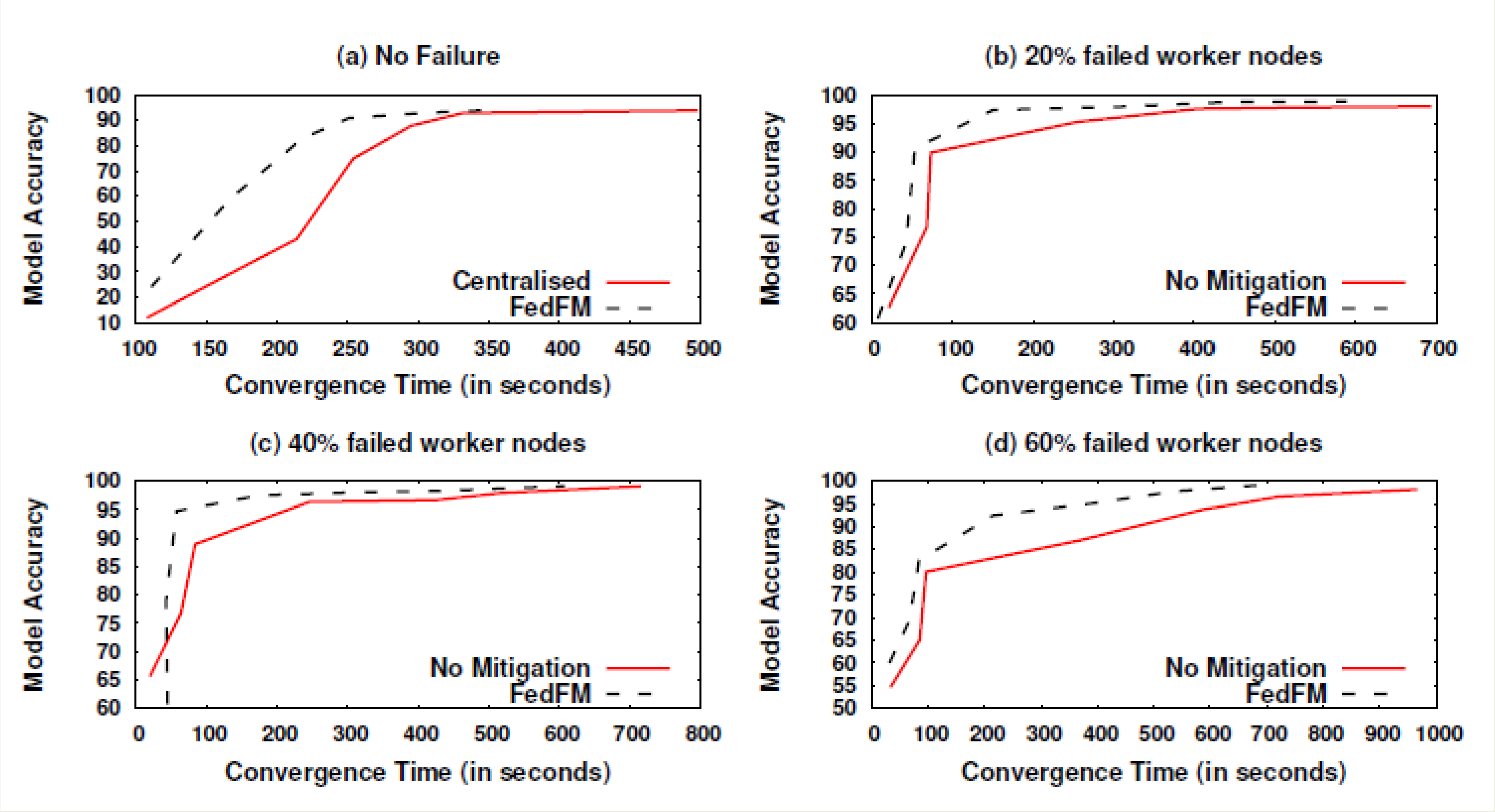
Convergence Time for FedAvg and FedFM in Different Scenarios



Key Takeaways:

- Fault mitigation is crucial for any Federated Learning Ecosystem
- With FedFM we are able to improve the Convergence Time and Model Accuracy for an FL technique

Convergence Time vs Accuracy Plots for Different Scenarios with and Without Failure



Key Takeaways:

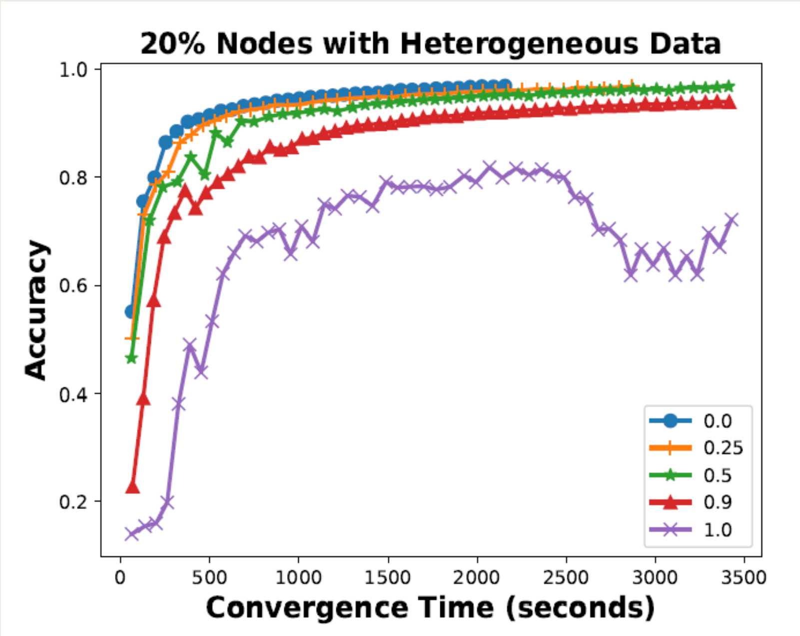
- The results highlight the utility of FedFM in IoT security applications
- Such utility is of utmost importance when there is a possibility of failure of nodes, which is true for any practical edge environment

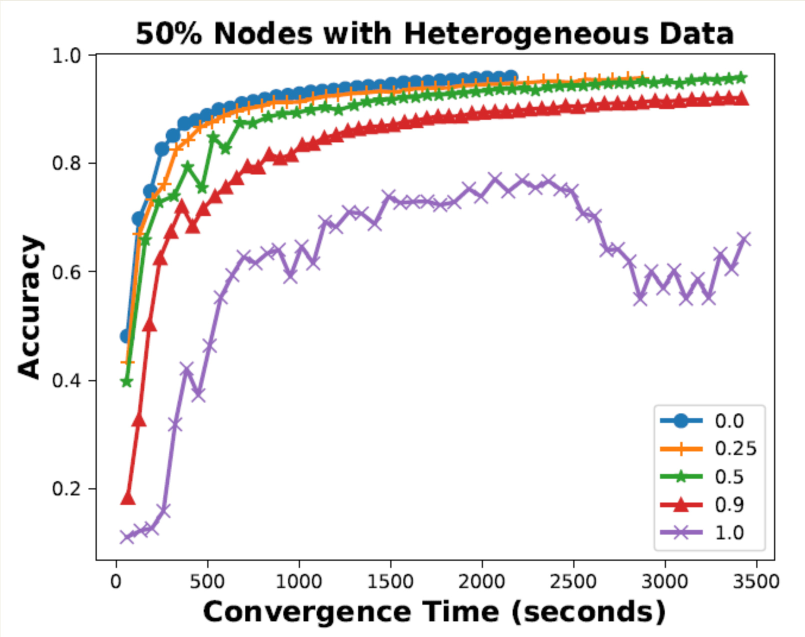
OPTIMAL NODE SELECTION FOR FEDERATED LEARNING WITH NON-IID DATA

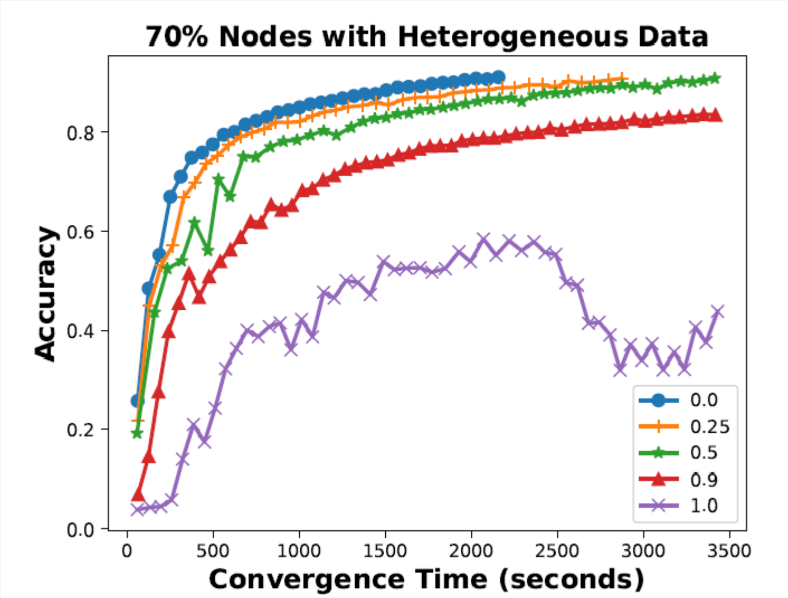
Defining Non-IITD

- There are different ways of defining a Non-IID data distribution
  - *Attribute skew*
  - *Label skew*
  - *Temporal skew*
  - *Quantity skew*
    - For every class, the quantity (i.e., size of data) is different
    - Not all classes have the same data size
- We work with quantity skewness which means that the training data can vary across all clients

Variation of Accuracy with Convergence Time for Different Levels of Skewness







Federated Node Selection with Entropy (FedNSE)

$$\eta(X) = \frac{H}{H_{max}} = -\sum_{i=1}^{n} \frac{p(x_i)log_b(p(x_i))}{log_b n}$$

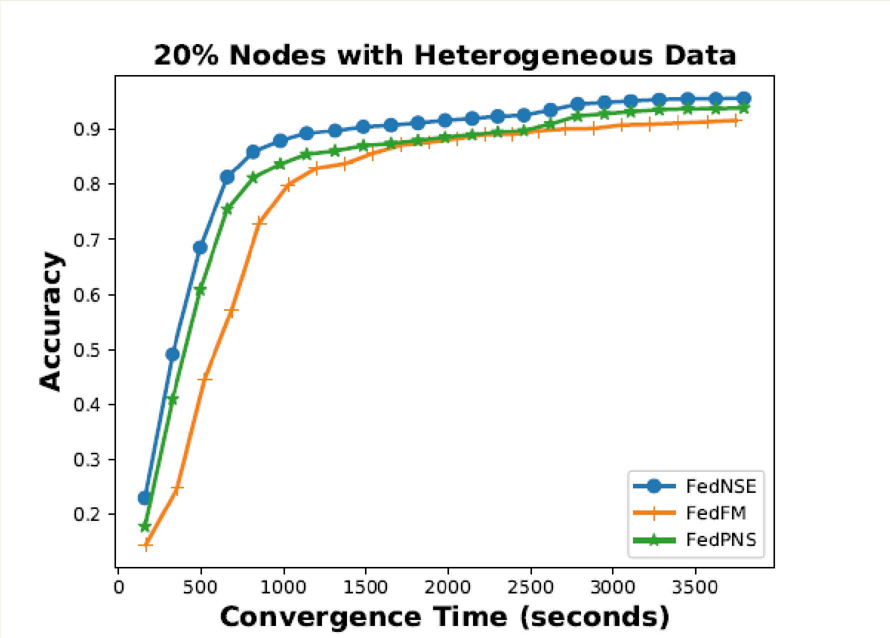
Naïve Selection Methodology

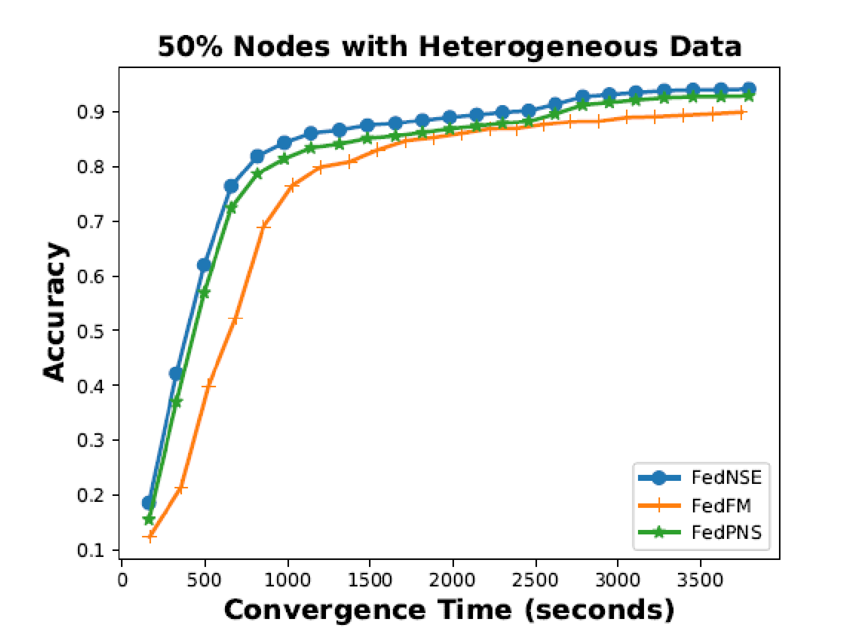
$$\mathcal{S} = \left(\frac{\alpha}{\mathcal{B}} + \frac{\kappa * \mathcal{V}}{\mathcal{P}}\right) * \frac{1}{\mathcal{V}}$$

New Selection Methodology

$$\mathcal{S}_\eta = \left(\frac{\alpha}{\mathcal{B} * \mathcal{V}} + \frac{\kappa}{\mathcal{P} * \mathcal{V}^{(\eta-1)}}\right) * \frac{1}{\eta}$$

Variation of Accuracy of the Competing Systems with Convergence Time for different levels of Skewness (x% of nodes have heterogeneous data distribution)





Key Takeaways

- The number of worker nodes plays an integral part in the efficiency of an FL technique and is dependent on the learning model's architecture
- Not all nodes in the network are required for an efficient FL model
  - *Empirically, 60% of the total nodes would perform as well as all the available nodes in a homogeneous setting*
- Having a specific number of working nodes in the network is not the same as having the same number of nodes post failure as the failed nodes could have exclusive data samples, thus hindering the model performance
- FedFM improves upon the existing FL techniques by employing fault mitigation strategies and has high utility in real world applications such as IoT security

Threats, Attacks and Defences in Federated Learning

Attack Vectors in Federated Learning

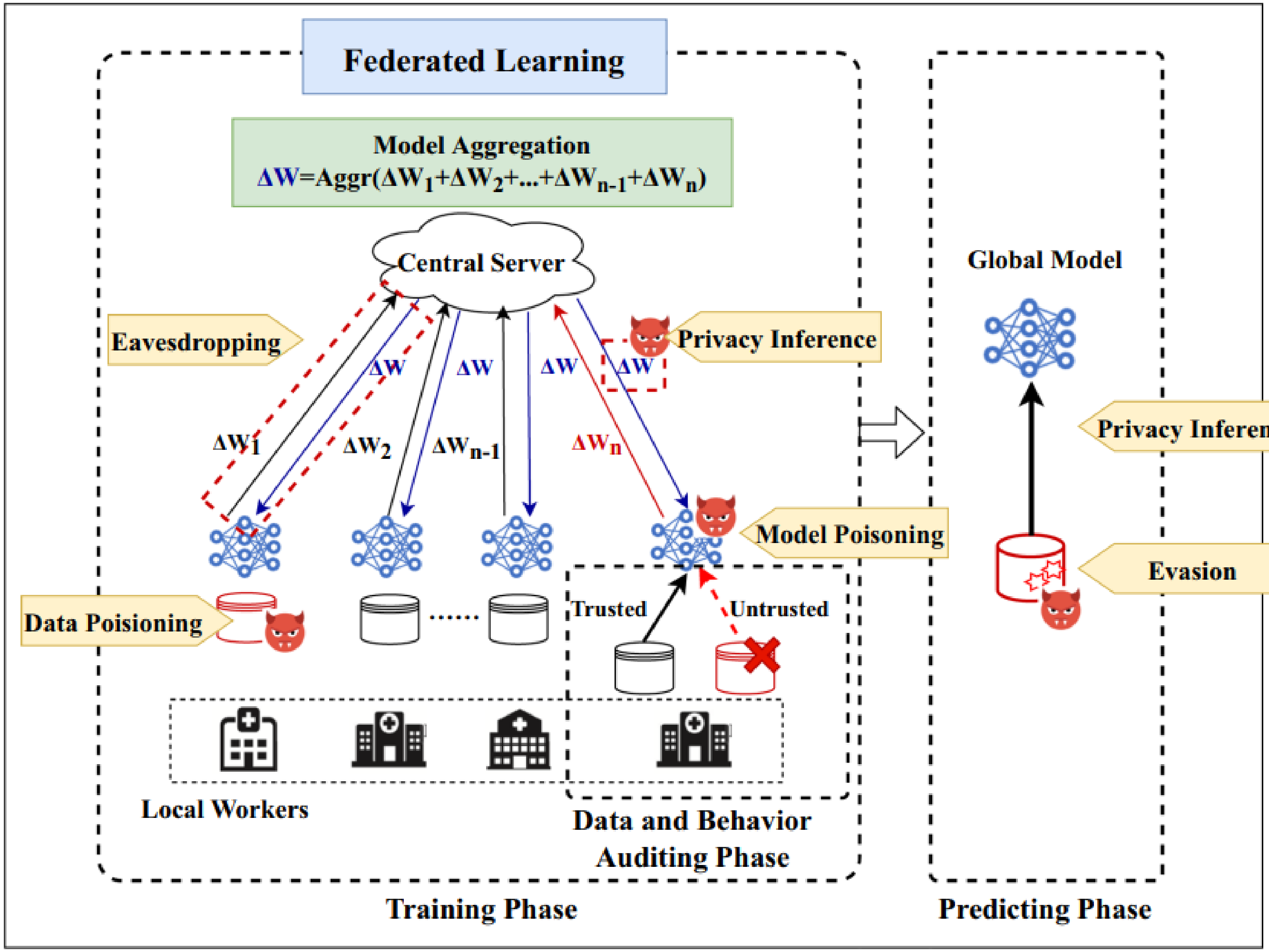
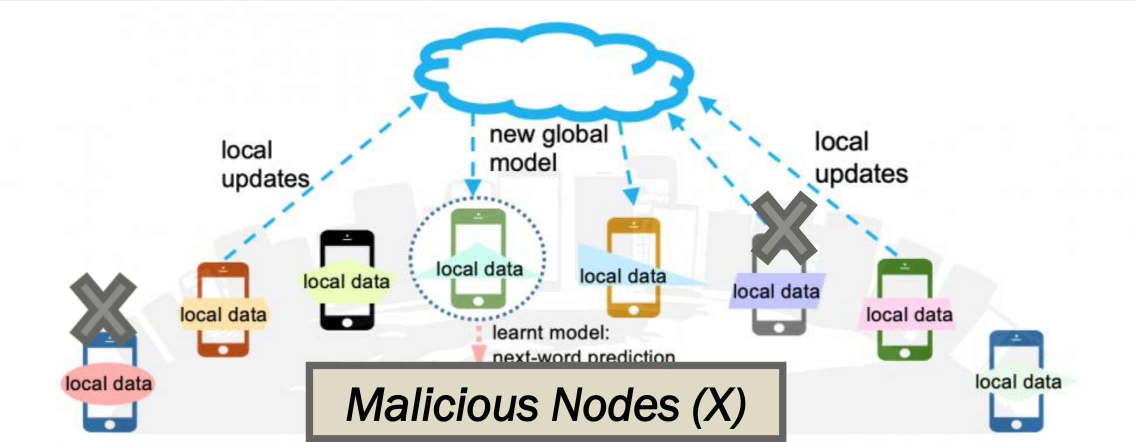


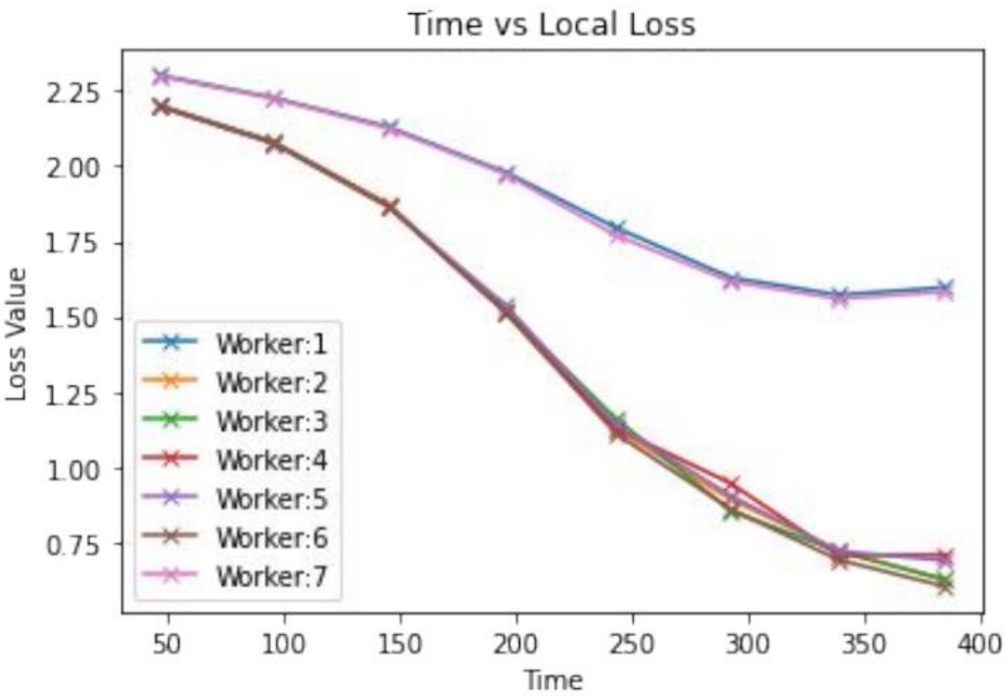
Fig. 1 The multi-phases framework of FL including data and behavior auditing, model training and model predicting

Maliciousness in Worker Nodes

- How do we detecting Maliciousness in Worker Nodes and incorporate the same in selection criteria?
- Malicious Nodes
  - *Nodes with wrongly labelled data*
- The extent of the malicious nodes could be varied
- The number of malicious nodes and the total number of nodes could be varied
- We can also test in a dynamic setting where the nodes may be initially benign and may start turning malicious after some internal of time
- Ignoring such nodes becomes quite important for the selection algorithm



Incorporating Maliciousness in Worker Nodes



Local Model Loss for Malicious Node Detection

Total Worker Nodes: 20
Malicious Nodes: 4 (Labels swapped)
Data Distribution: Homogeneous
Dataset: MNIST

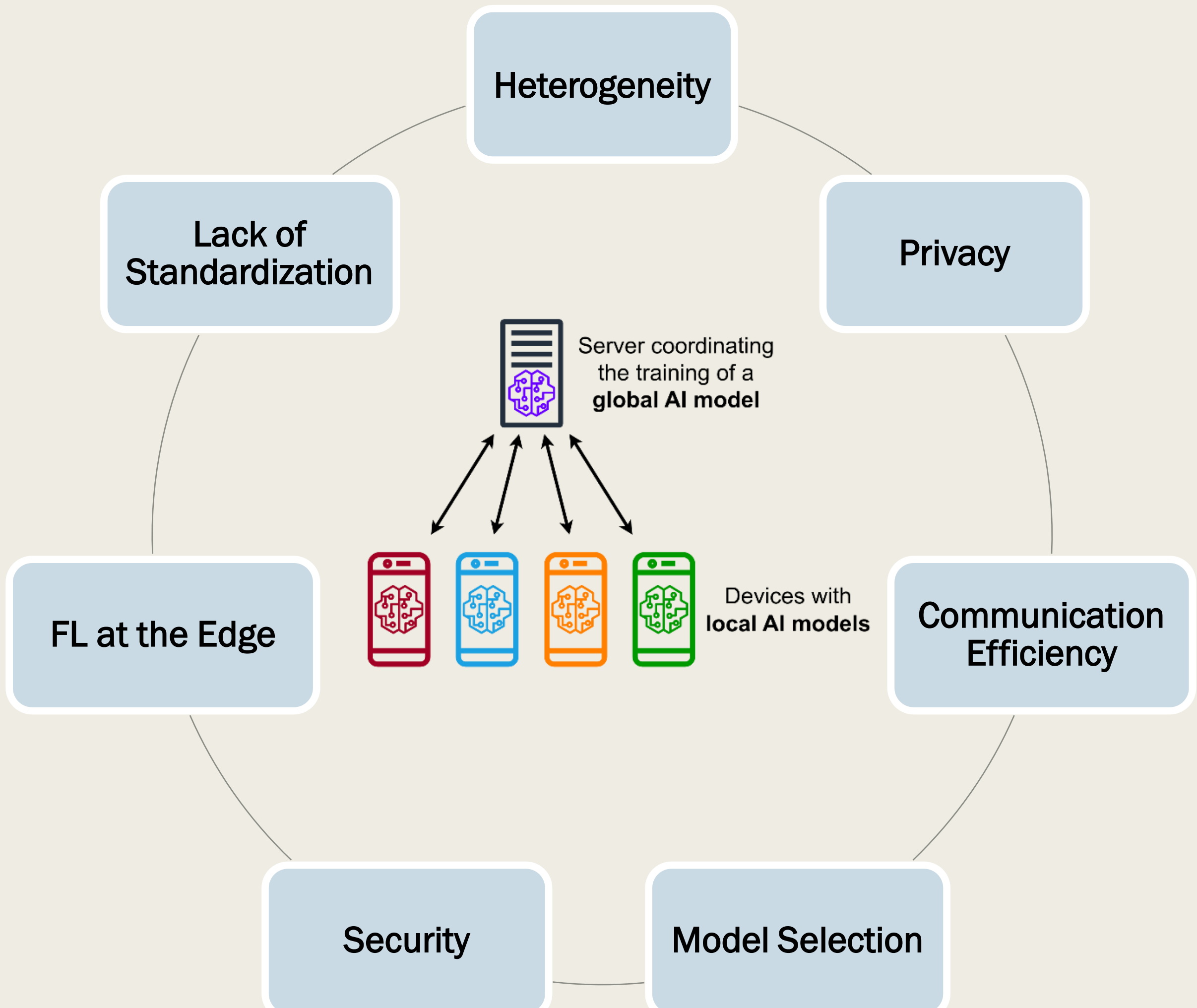
Considerably higher local model loss values for malicious nodes

Fairness in Federated Learning

Fairness in Federated Learning

- Client Selection
- FL Model Optimization
- FL Incentive Distribution
- …

Challenges of Federated Learning



Scope for Further Extensions

- Decentralized Federated Learning
- Dynamic Network Architecture
- Incorporating Fairness in Node Selection
- Investigating different definitions of Skewness
- Securing Federated Learning
  - *Additional Attack vectors*

THANK YOU

rajeevshorey@gmail.com