Cartel: A System for Collaborative Transfer Learning at the Edge

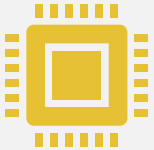
HARSHIT DAGA, PATRICK K. NICHOLSON, ADA GAVRILOVSKA, DIEGO LUGONES

Authors

- Harshit Daga
  - Computer Science Ph.D. Candidate Georgia Tech
- Patrick K. Nicholson
  - Works at NOKIA Bell Labs
  - Received his Ph.D. in 2013 from the University of Waterloo
- Ada Gavrilovska
  - Associate Professor at the College of Computing and the Center for Experimental Research in Computer Systems (CERCS) at Georgia Tech
- Diego Lugones
  - Global R&D Manager at NOKIA Bell Labs




What is Transfer Learning



"Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task."

Use of knowledge gained while solving one problem and applying it to a different related problem

Problem

- Introduction of Multi-access Edge Computing + 5G has created a 47% increase in network traffic since 2016
  - 7 – 49 exabytes per month
- Deploying Machine Learning on this amount of data has become increasingly difficult
  - Network traffic concerns
  - Training speed and accuracy concerns

Multi-access Edge Computing (MEC)

"Cloud services, running at the edge"

"Mini Data Centers, deployed near communication points"

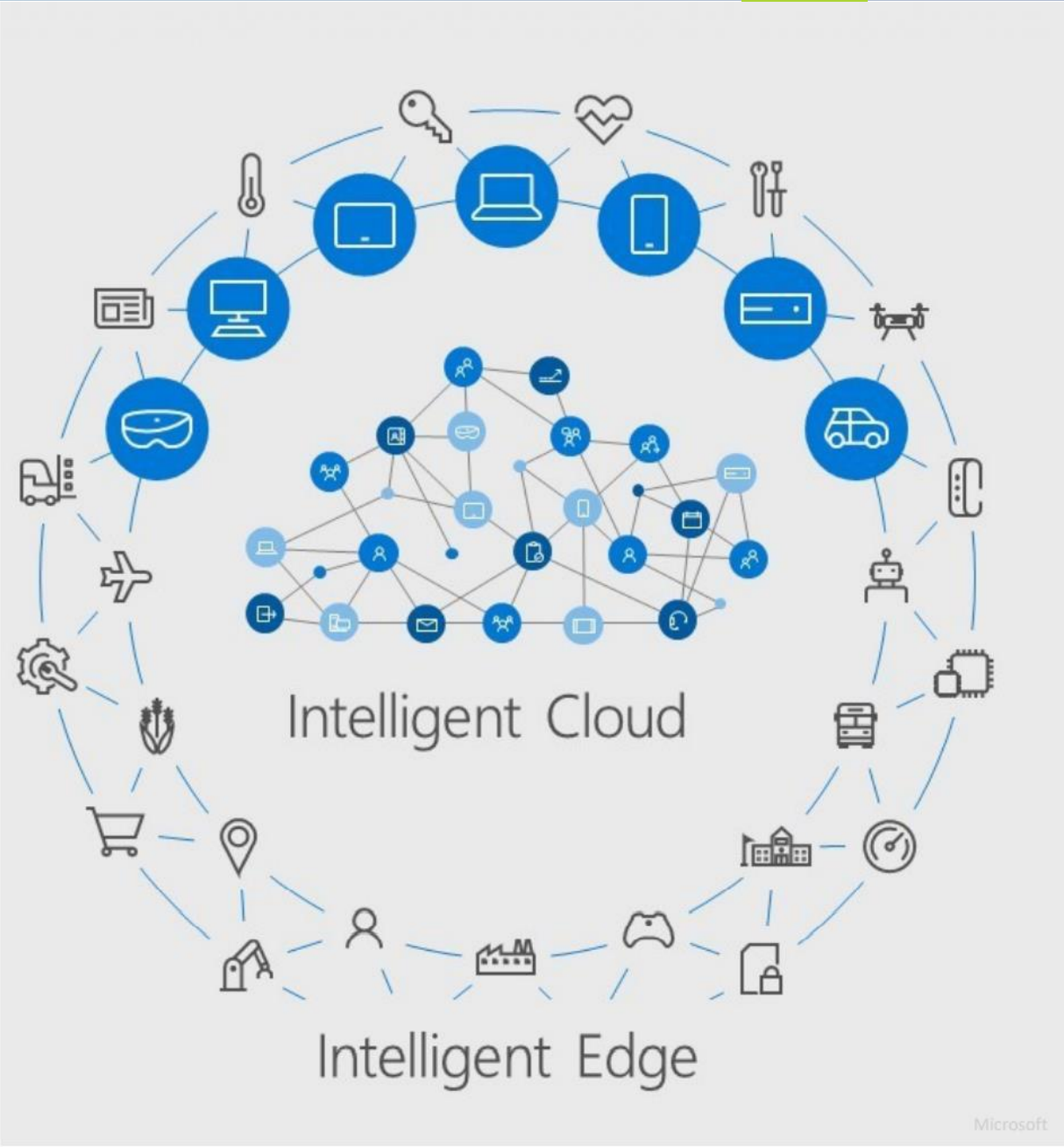
- Compute & Storage closer to the end user
- Provides ultra-low latency

NOKIA Bell Labs Georgia Tech

Machine Learning@ Edge

- There is tremendous growth of data generated at the edge from end-user devices and IoT.
- We explore machine learning in the context of MEC:

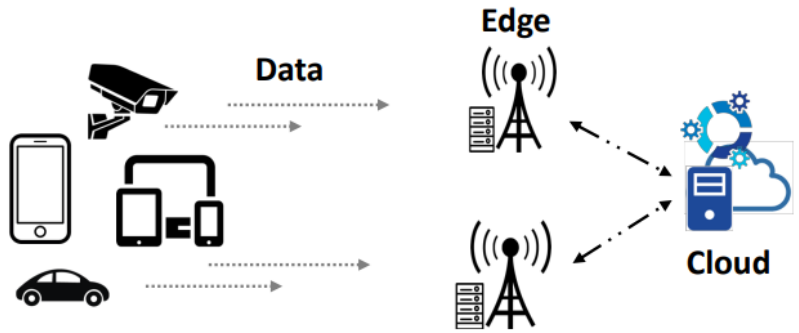
  - Results are only needed locally
  - Latency is critical
  - Data volume must be reduced



NOKIA Bell Labs Georgia Tech

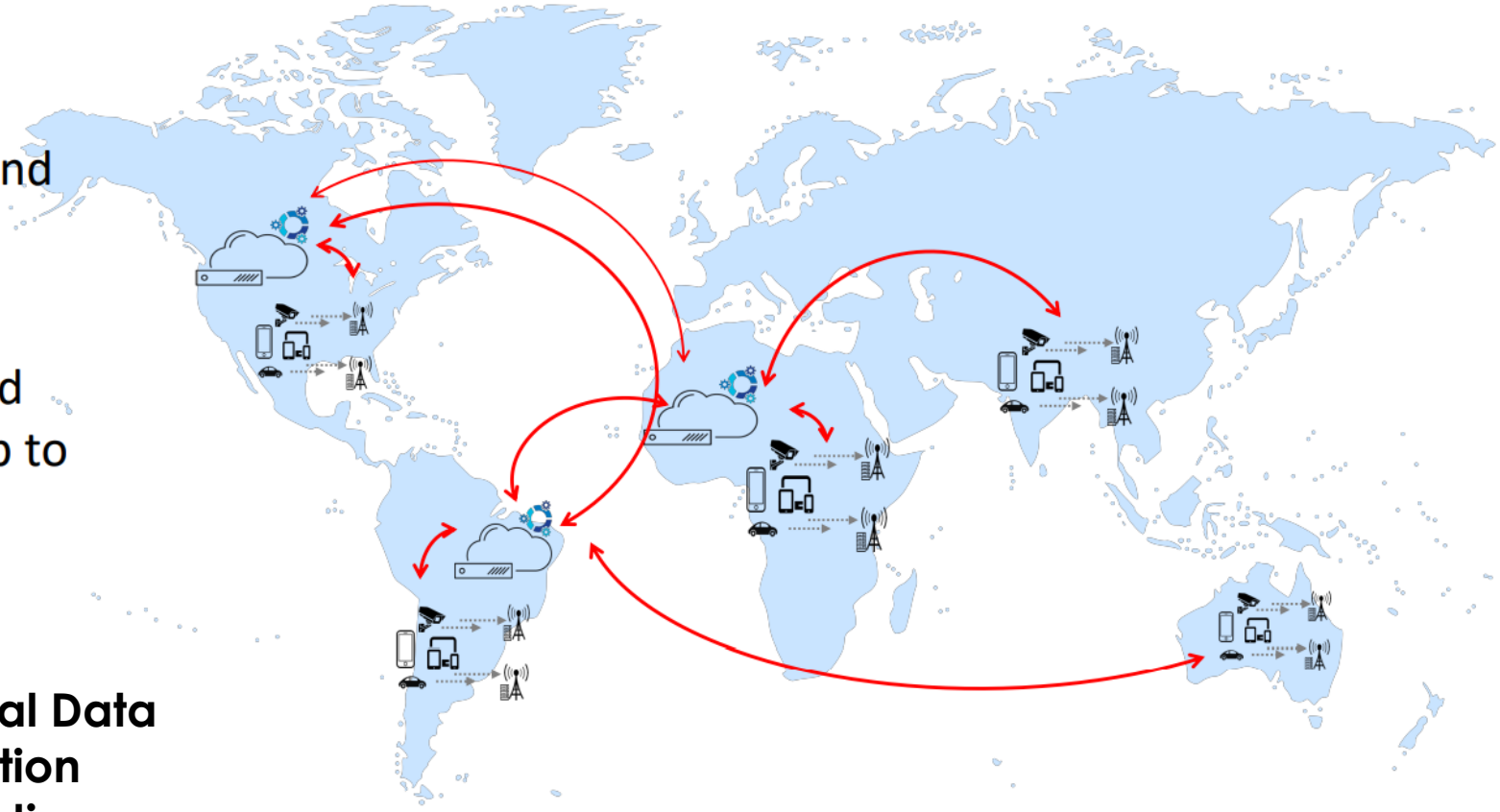
Existing Solution

Centralized System



Problems

- **Data movement** is time consuming and uses a lot of backhaul network bandwidth.
- **Distributed ML** across geo-distributed data can **slow down** the execution up to 53X[1].
- **Regulatory constraints** (GDPR)



General Data Protection Regulation

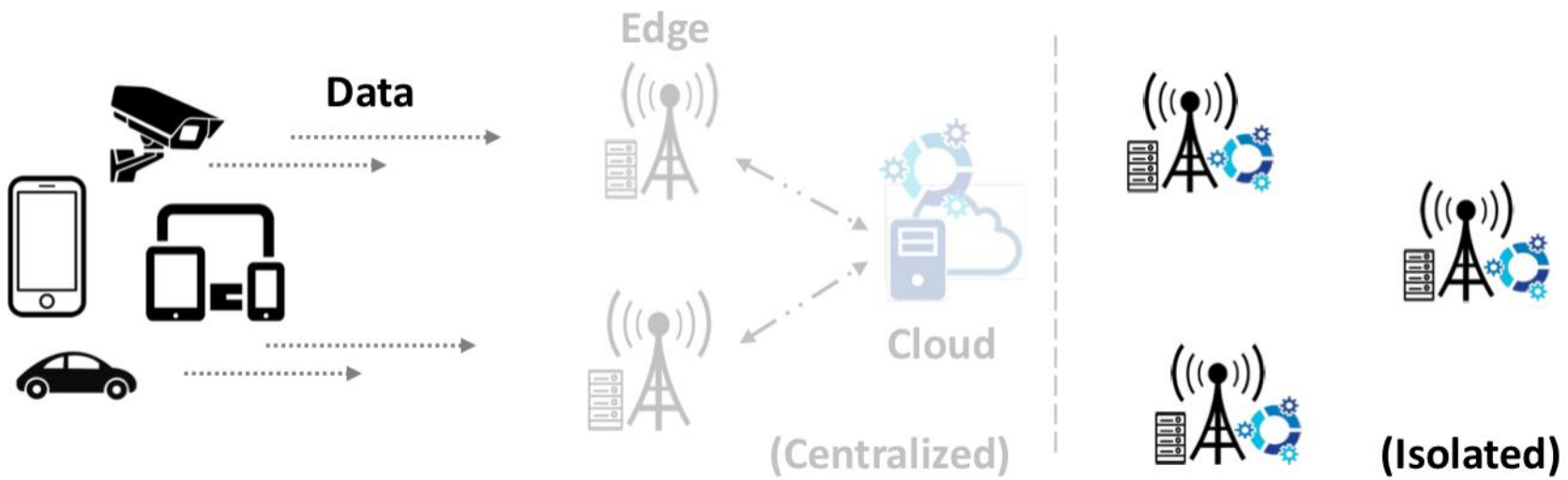
[1] Kevin et al. Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds.

NOKIA Bell Labs Georgia Tech

An Alternative Approach

Isolated System

- Train machine learning models independently at each edge, in isolation from other edge nodes.
- The isolated model performance gets heavily impacted in scenarios where there is a need to **adapt** to changing workload.



Motivation

Can we achieve a balance between centralized and isolated system?

Leverage the resource-constrained edge nodes to train **customized (smaller)** machine learning **models** in a manner that **reduces training time** and **backhaul data transfer** while **keeping the performance closer to a centralized system**?

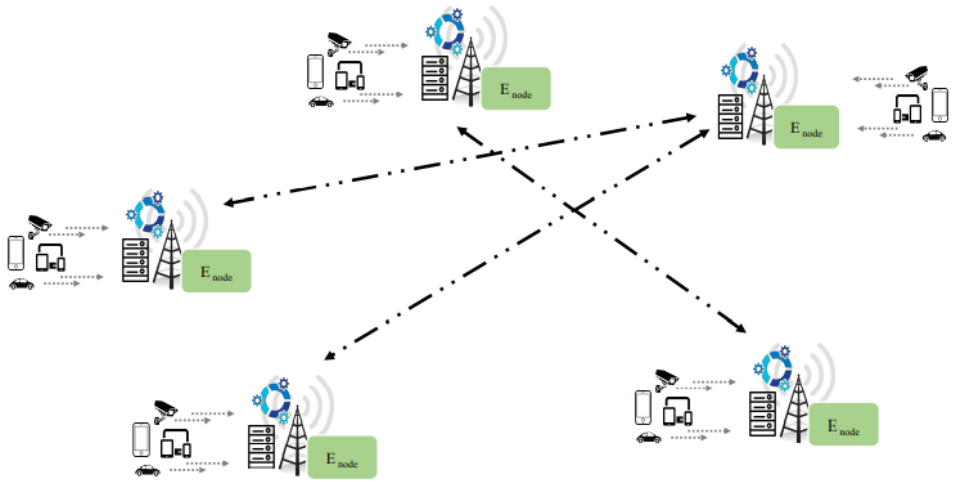
Backhaul: The sending of data to the backbone of the network… Kinda like sending all the way to the "cloud"

Opportunity

- Each edge node has its own attributes / characteristics → **a full generic model** trained on broad variety of data **may not be required at an edge node.**

Solution Overview

Cartel : A System for Collaborative Transfer Learning at the Edge



	Centralized	Isolated	Cartel
Light Weight Models	x	✓	✓
Data Transfer	↑	↓	↓
Online Training Time	↑	↓	↓
High Model accuracy	✓	x	✓

- Cartel maintains **small customized models** at each edge node.
- When there is change in the environment or variations in workload patterns, Cartel provides a jump start to **adapt** to these changes by transferring knowledge from other edge(s) where similar patterns have been observed.

NOKIA Bell Labs Georgia Tech

Key Challenges

C1 : When to request for model transfer?

C2 : Which node (logical neighbor) to contact?

C3 : How to transfer knowledge to the target edge node?

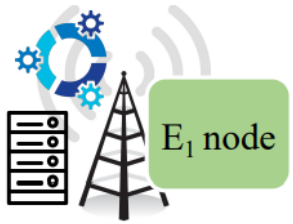
When... Where... How

Solution Design

~~Raw data~~ v/s Metadata

- Do not share raw data between any edge nodes or with the cloud.
- Use Metadata
  - Statistics about the network
  - Software configuration
  - Active user distribution by segments
  - Estimates of class priors (probability of certain classes), etc.

Metadata Server (MdS)



Cartel maintains and aggregates metadata locally and in the metadata server (MdS).

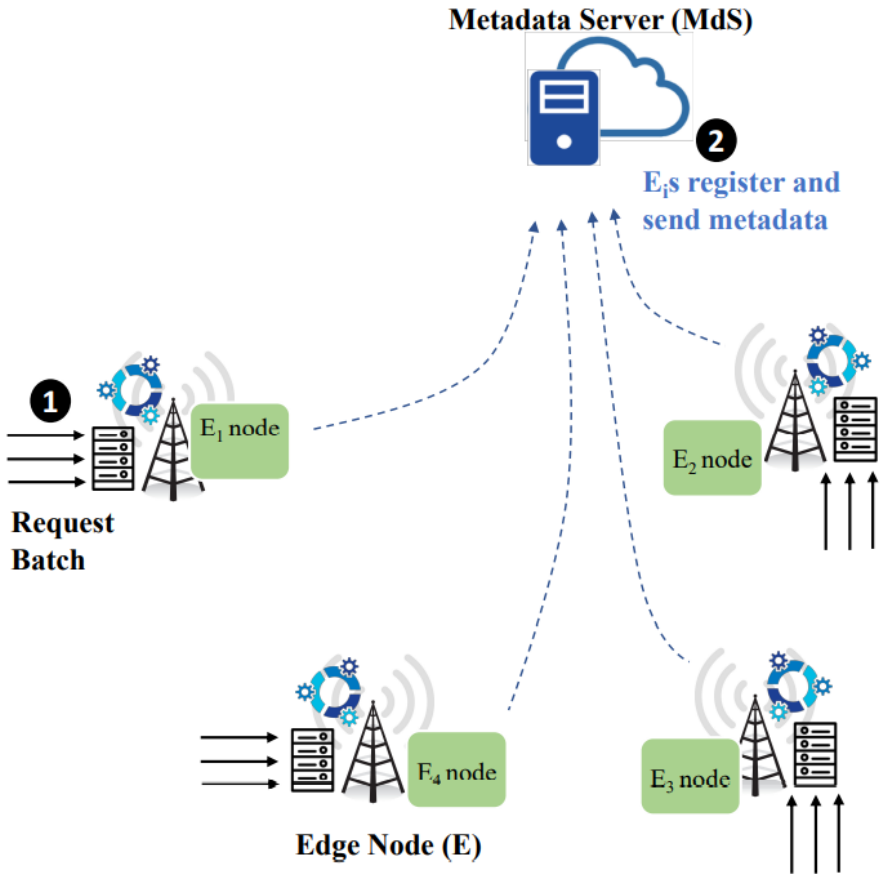
NOKIA Bell Labs Georgia Tech

C1: When to request for model transfer?

Drift Detection

- Determine when to send a request to collaborate with edge nodes for a model transfer.
- In our prototype we use a threshold-based drift detection mechanism.

When data drift reaches a certain percentage



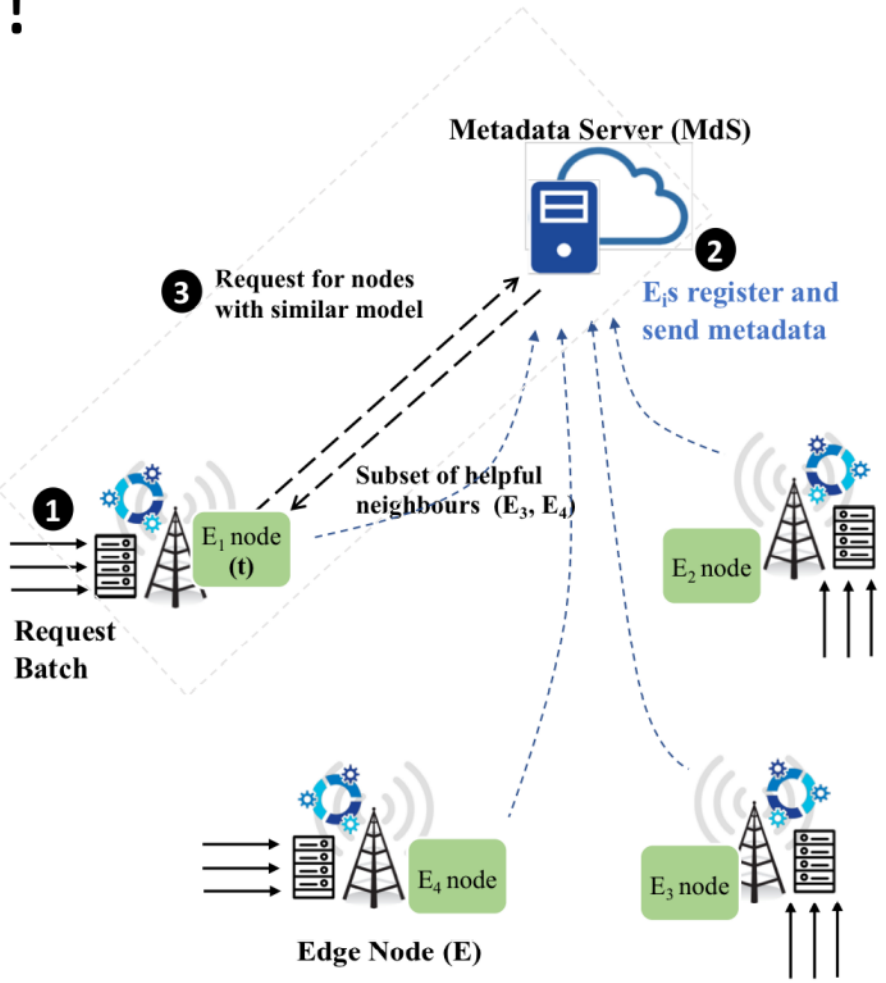
NOKIA Bell Labs Georgia Tech

Use the Meta Data Service to check which neighors have data most similar to themselves

C2: Which neighbors to contact?

Logical Neighbor

- Find the neighbor that has similar class priors to the target node.
- We call them as “logical neighbors” as they can be from anywhere in the network.
- In our prototype class priors are undergoing some shift, the empirical distributions from the target node is compared with those from the other nodes at the MdS to determine which subset of edge nodes are logical neighbors of the target node.



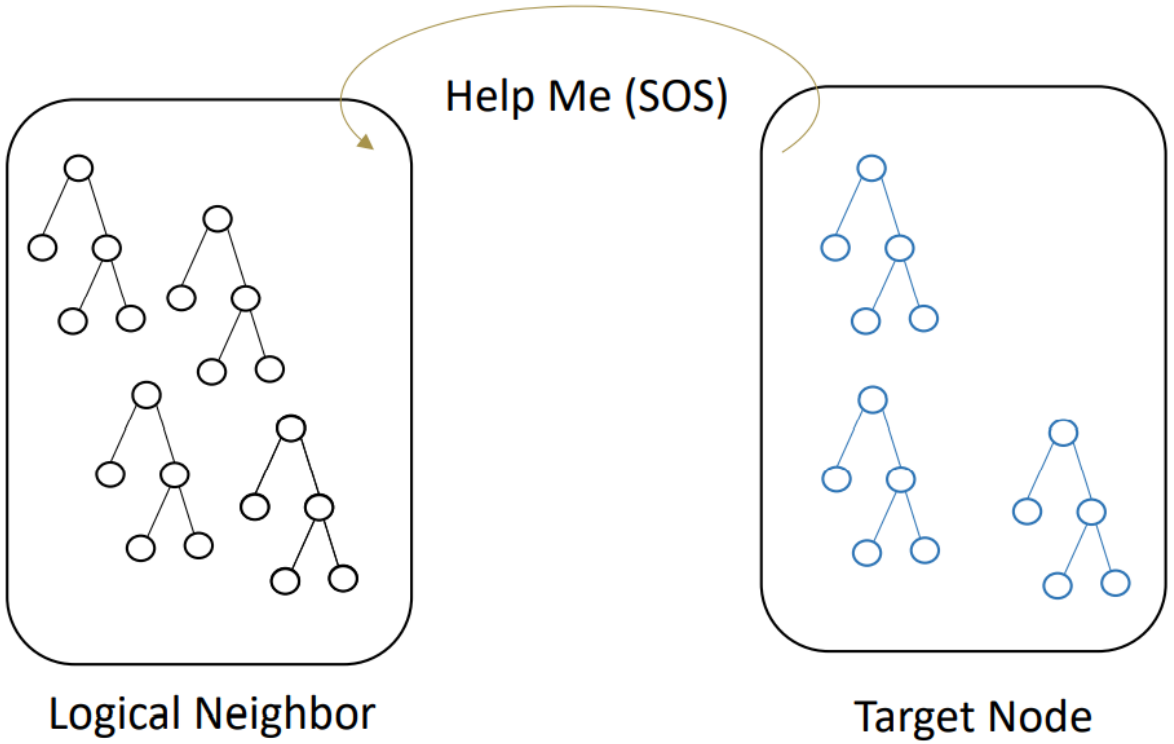
NOKIA Bell Labs Georgia Tech

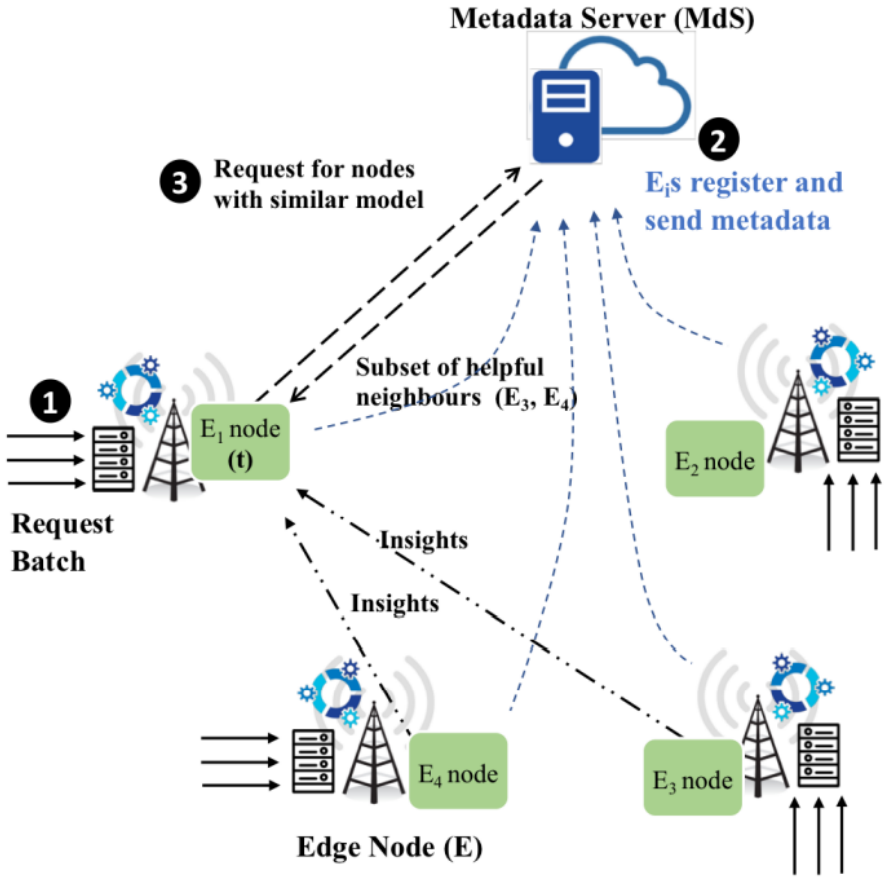
Find which classifications have the least confidence and request help in classification

C3: How to transfer knowledge to the target?

Knowledge Transfer

- Two steps process
  1. Partitioning
  2. Merging





NOKIA Bell Labs Georgia Tech

Transfer Learning: Partition + Merge

Partition + Merge is good for adaptability to many ML algorithms

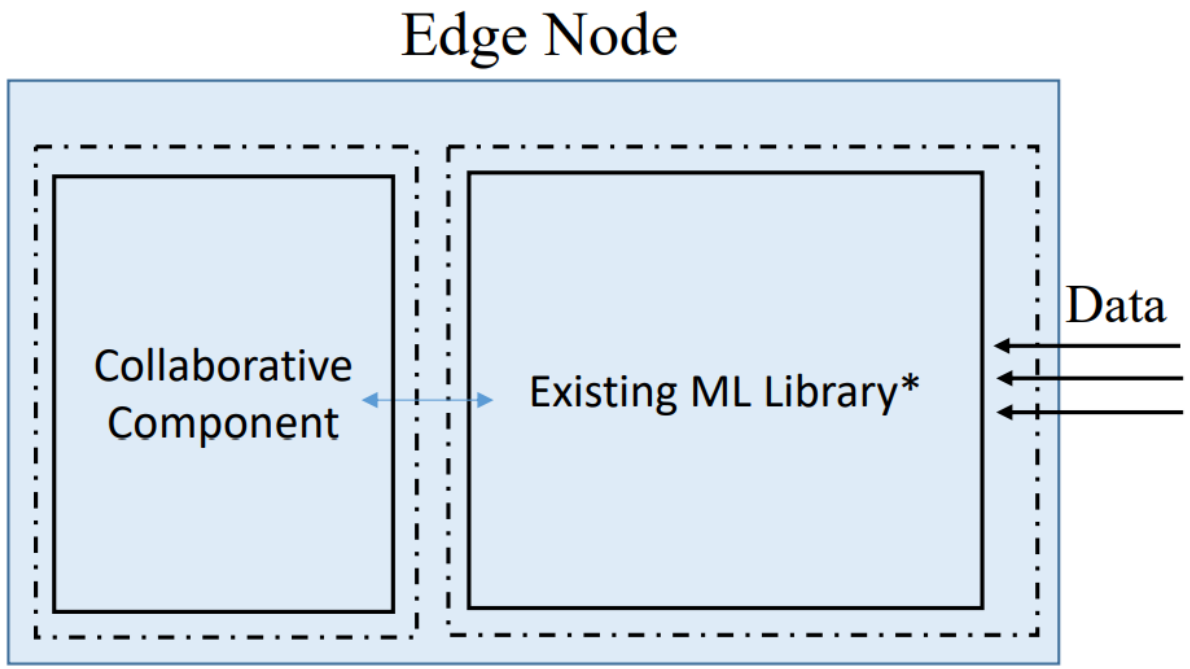
Online Random Forest (ORF)

- Bagging Approach: Remove a percentage of your existing tree structure and merge in a percentage of a neighbor's tree structure to your random forest

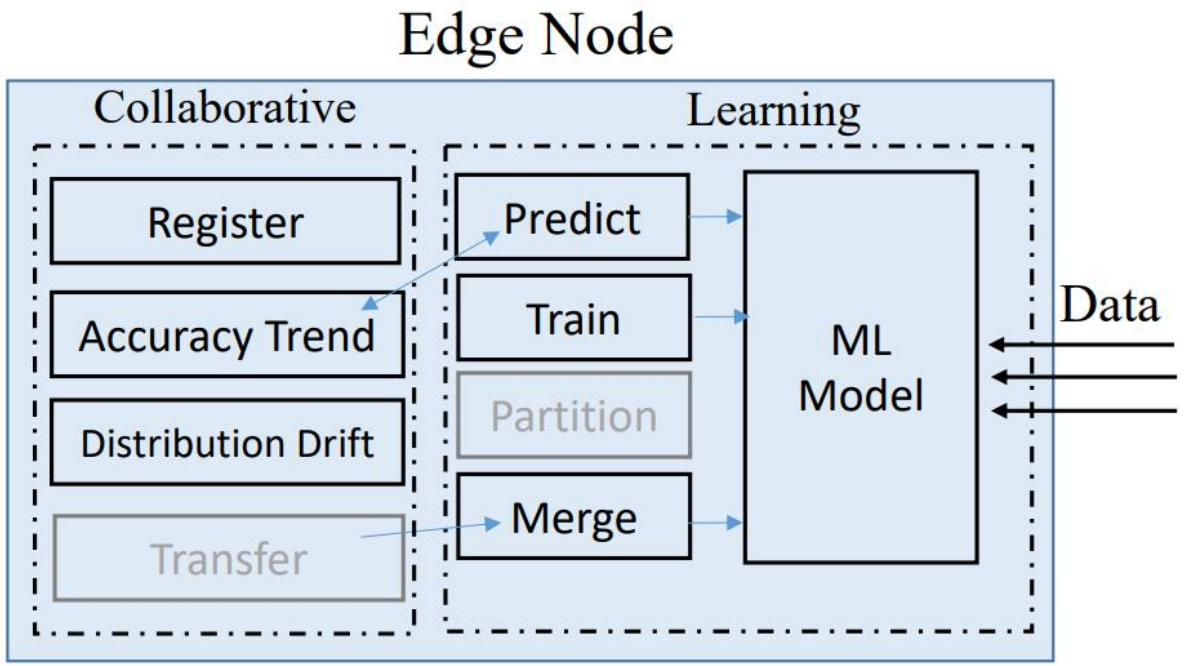
Online Support Vector Machine (OSVM)

- One-Versus-Rest: Drop the weight vectors from your data set that are proving to be inseparable in N-dimensions and inherit the weight vectors from your neighor

Solution Overview



Solution Overview



Experiment

5 edge nodes, 1 central node

Intel Xeon's w/ 48 GB ram

Tested ORF and OSVM algorithms

Tested image classification and network monitoring (instruction classification)

Evaluation

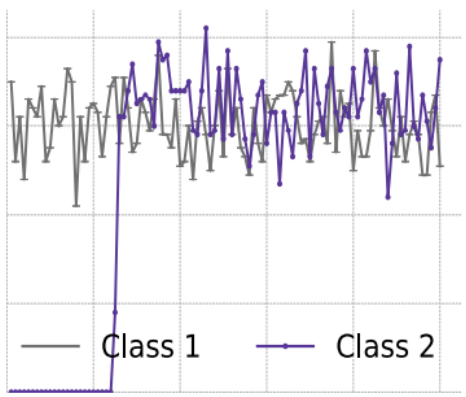
Shows how they introduced varying workloads over time

Goals

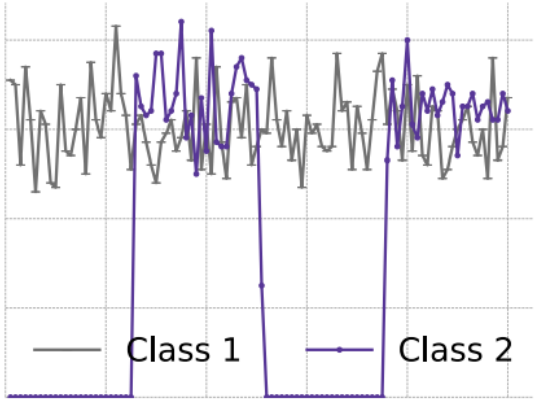
- How effectively system **adapts to the change in workload**?
- How effective is Cartel in **reducing data transfer costs**, while providing **lightweight** and accurate models?
- What are the costs in the mechanisms of Cartel and the design choices?
- How does Cartel perform in a real-world scenario?

Methodology

- Workload



Introduction Workload



Fluctuation Workload

- Machine Learning Model – ORF & OSVM
- Datasets used - MNIST & CICIDS2017

NOKIA Bell Labs Georgia Tech

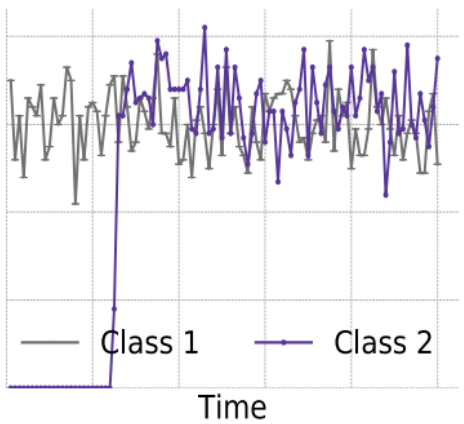
Evaluation

Goals

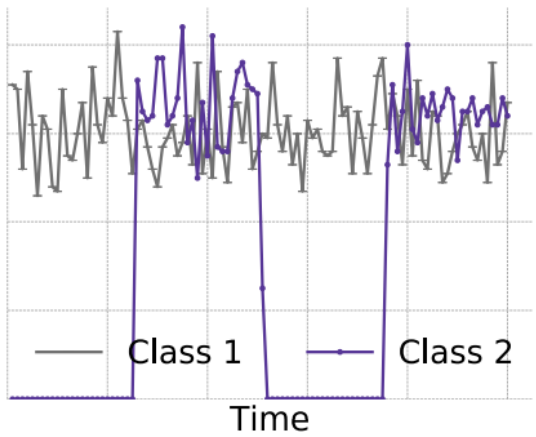
- How effectively system **adapts to the change in workload**?
- How effective is Cartel in **reducing data transfer costs**, while providing **lightweight** and accurate models?
- What are the costs in the mechanisms of Cartel and the design choices?
- How does Cartel perform in a real-world scenario?

Methodology

- Workload



Introduction Workload



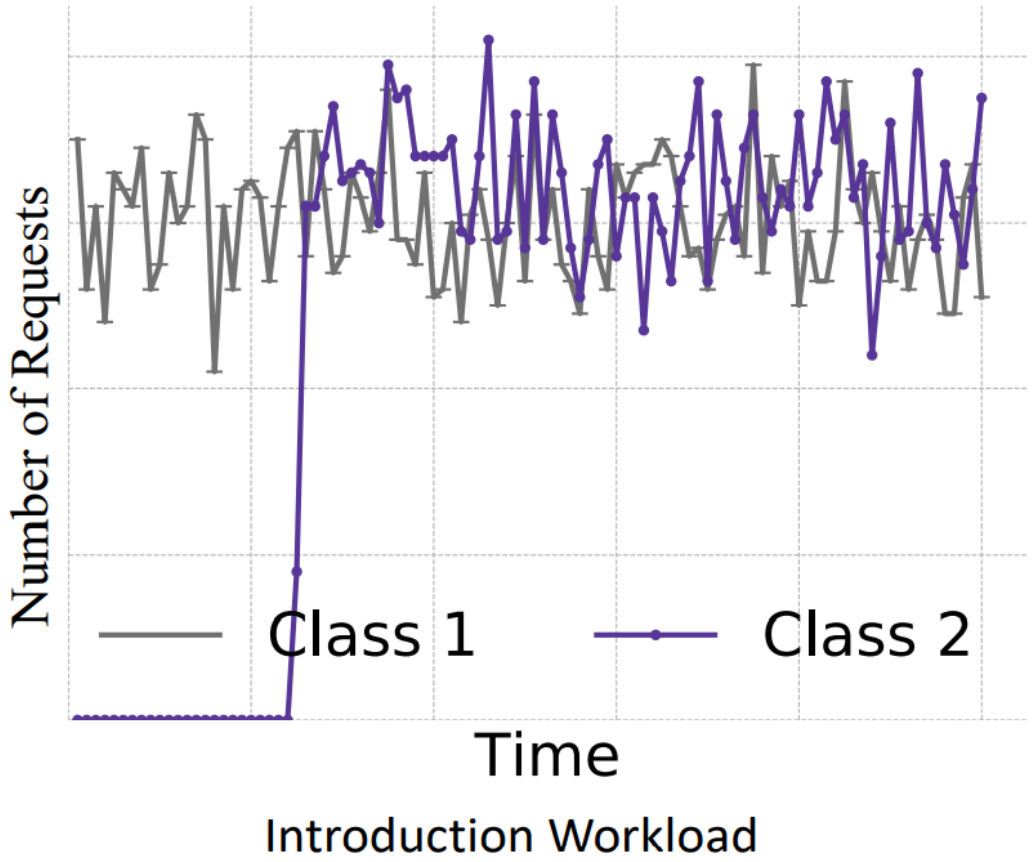
Fluctuation Workload

- Machine Learning Model – ORF & OSVM
- Datasets used - MNIST & CICIDS2017

NOKIA Bell Labs Georgia Tech

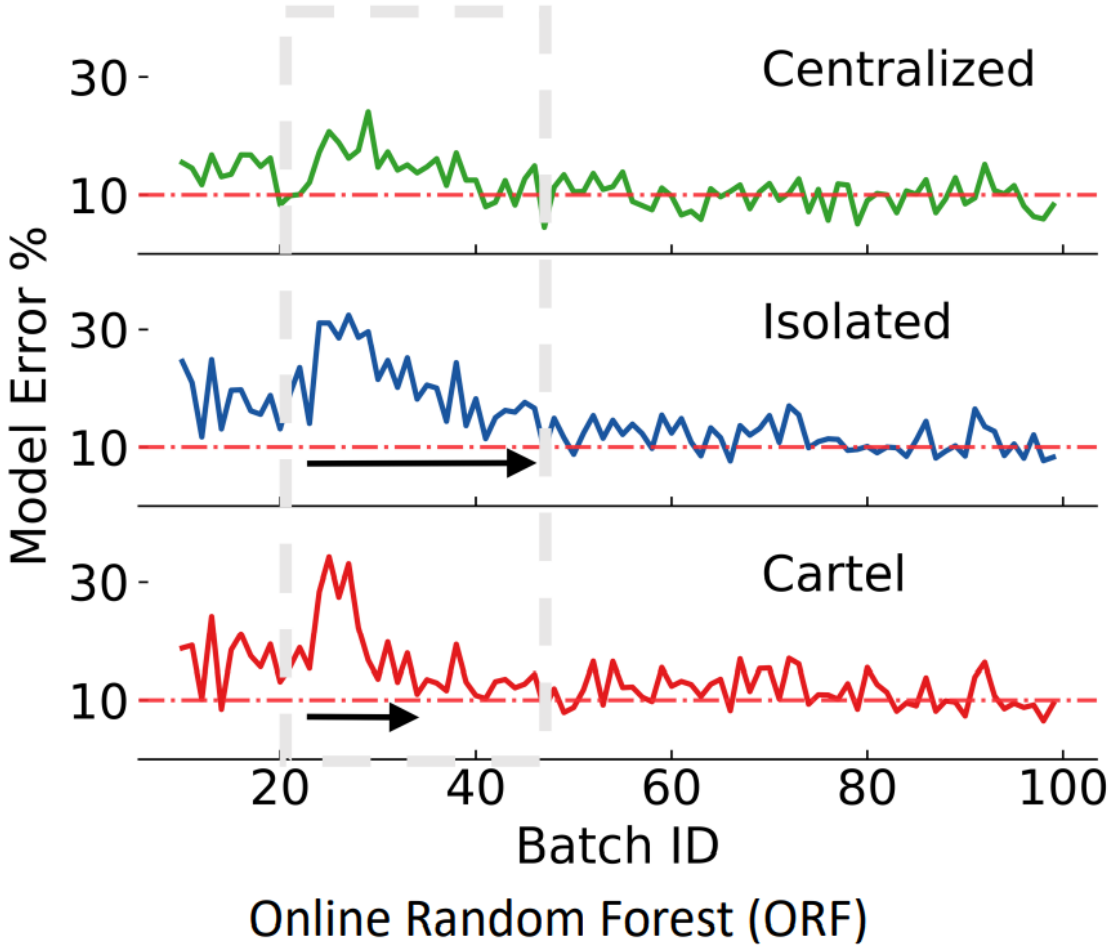
Evaluation

Adaptability to Change in the Workload



Introduction Workload

Able to achieve normalized error rates as fast as centralized training and far better than isolated

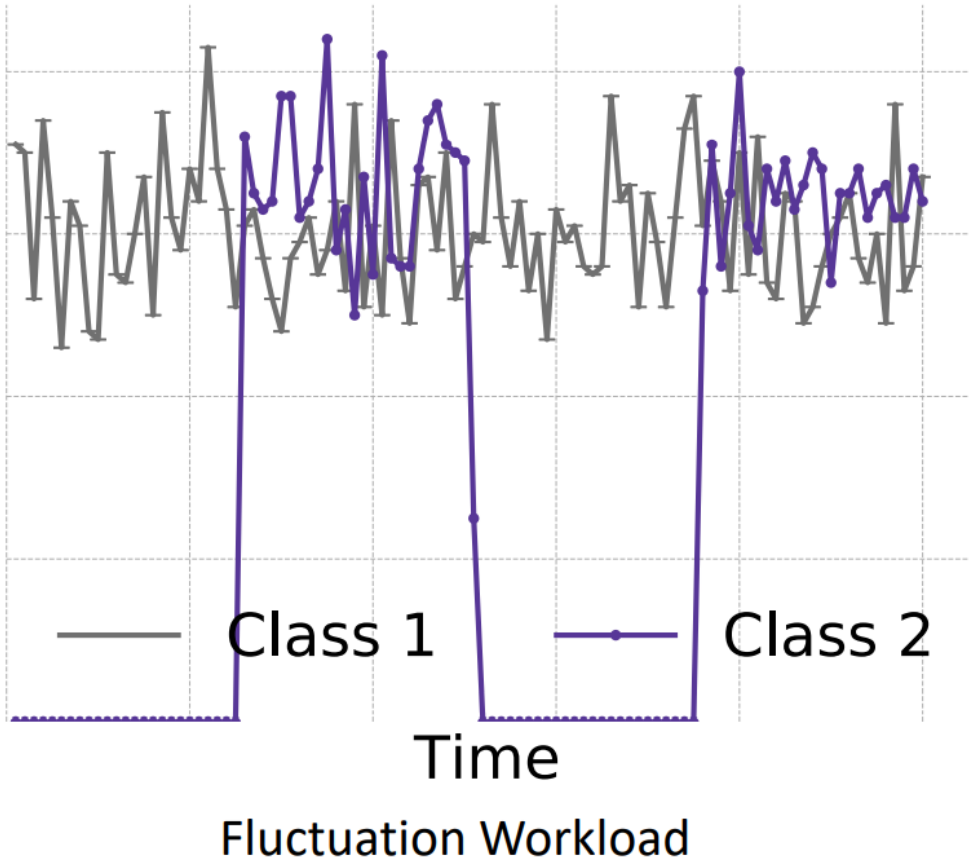


Online Random Forest (ORF)

NOKIA Bell Labs Georgia Tech

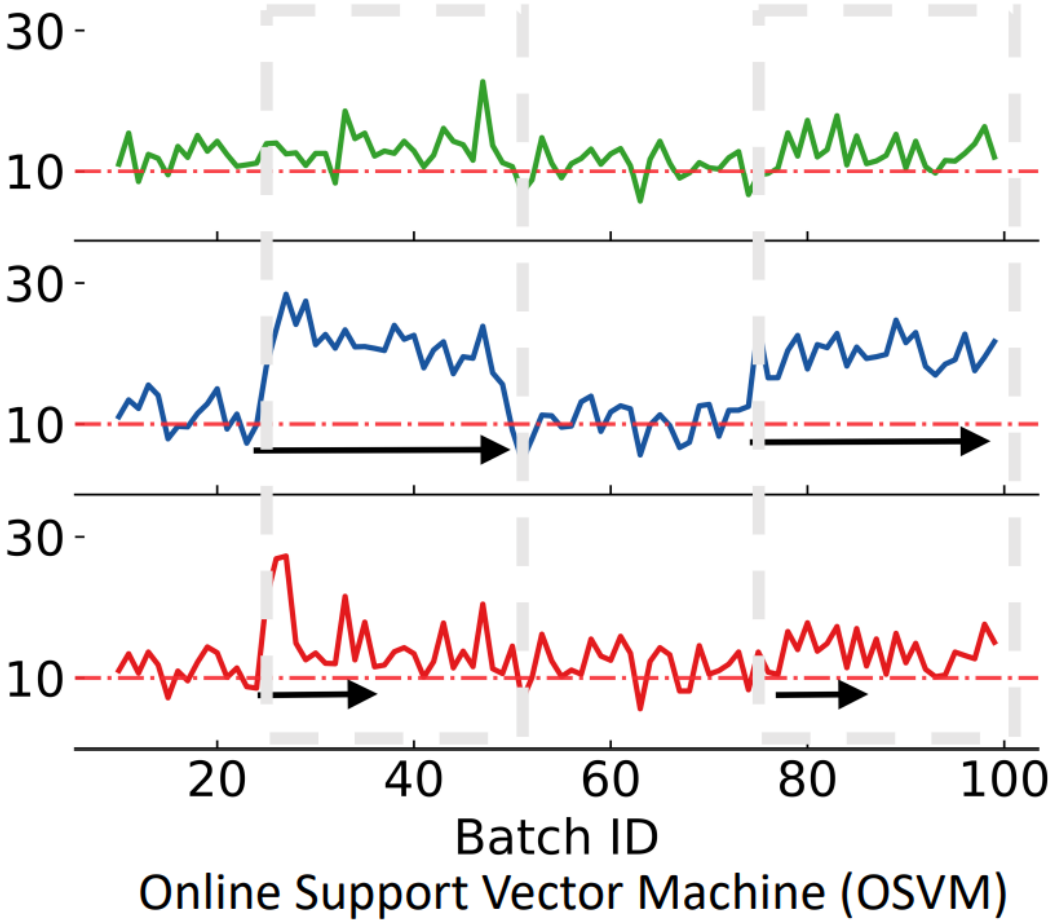
Evaluation

Adaptability to Change in the Workload



Fluctuation Workload

Able to achieve normalized error rates as fast as centralized training and far better than isolated

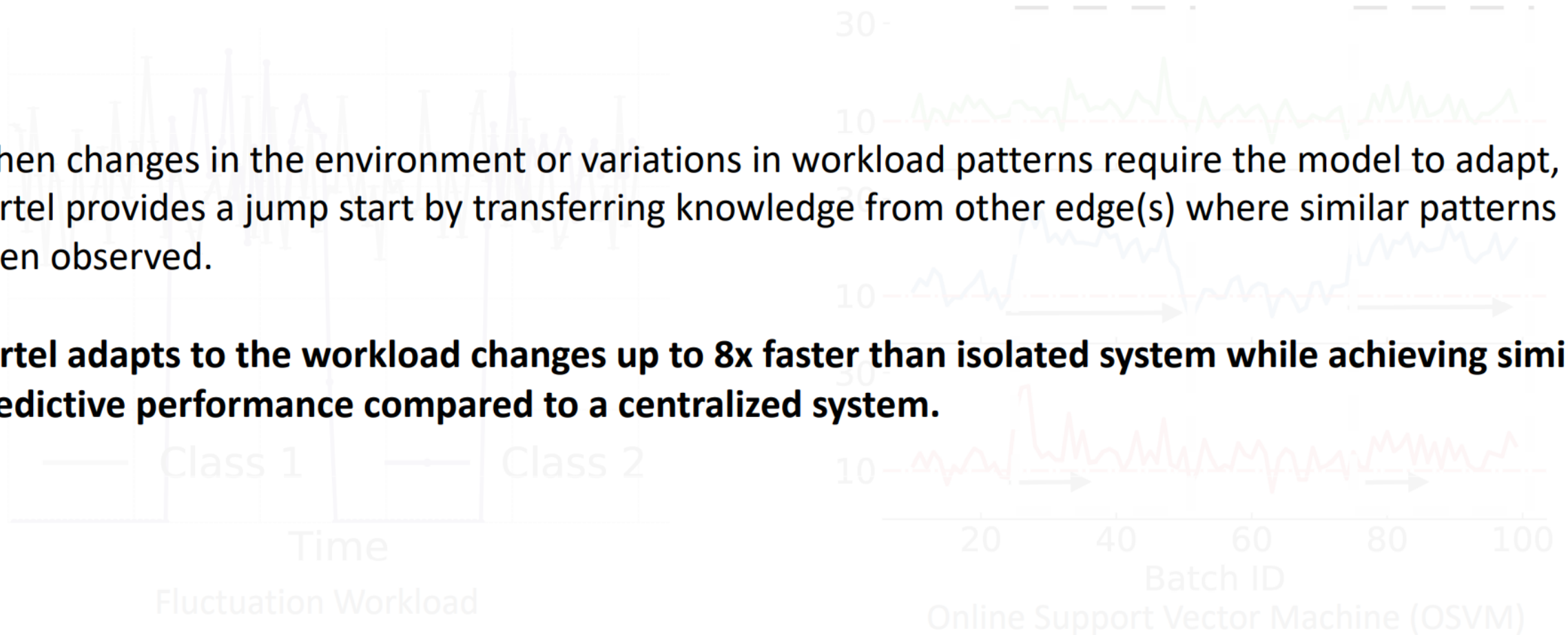


Online Support Vector Machine (OSVM)

NOKIA Bell Labs Georgia Tech

Evaluation

Adaptability to Change in the Workload



- When changes in the environment or variations in workload patterns require the model to adapt, Cartel provides a jump start by transferring knowledge from other edge(s) where similar patterns have been observed.

- **Cartel adapts to the workload changes up to 8x faster than isolated system while achieving similar predictive performance compared to a centralized system.**

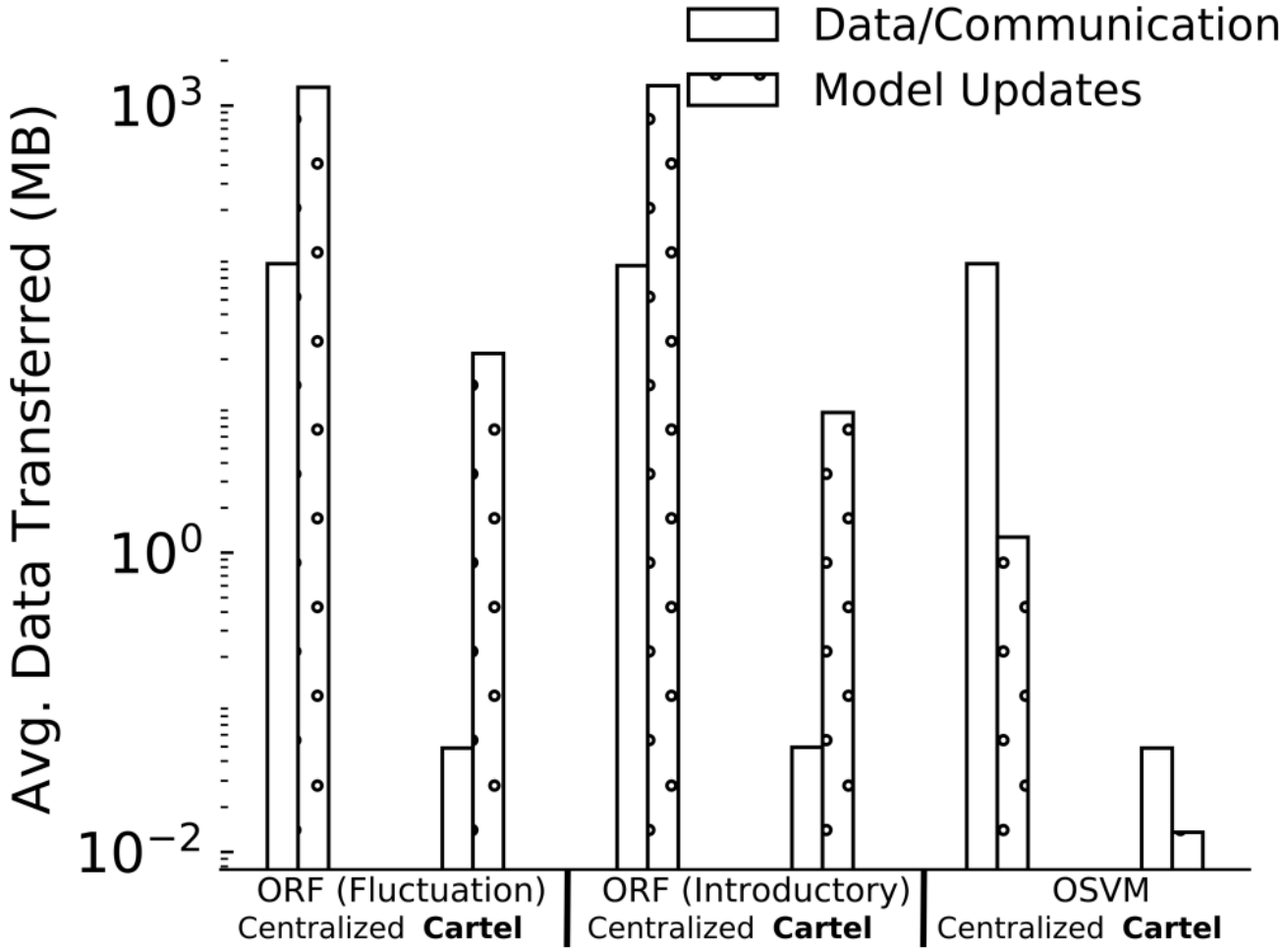
NOKIA Bell Labs Georgia Tech

SIGNIFICANTLY lower transfer cost

Evaluation

Data Transfer Cost

- Data/Communication cost includes the transfer of raw data or metadata updates.
- Model transfer cost captures the amount of data transferred during model updates to the edge (periodically in case of centralized system or partial model request from a logical neighbor in Cartel).
- **Cartel reduces the total data transfer cost up to 1500x when compared to a centralized system.**

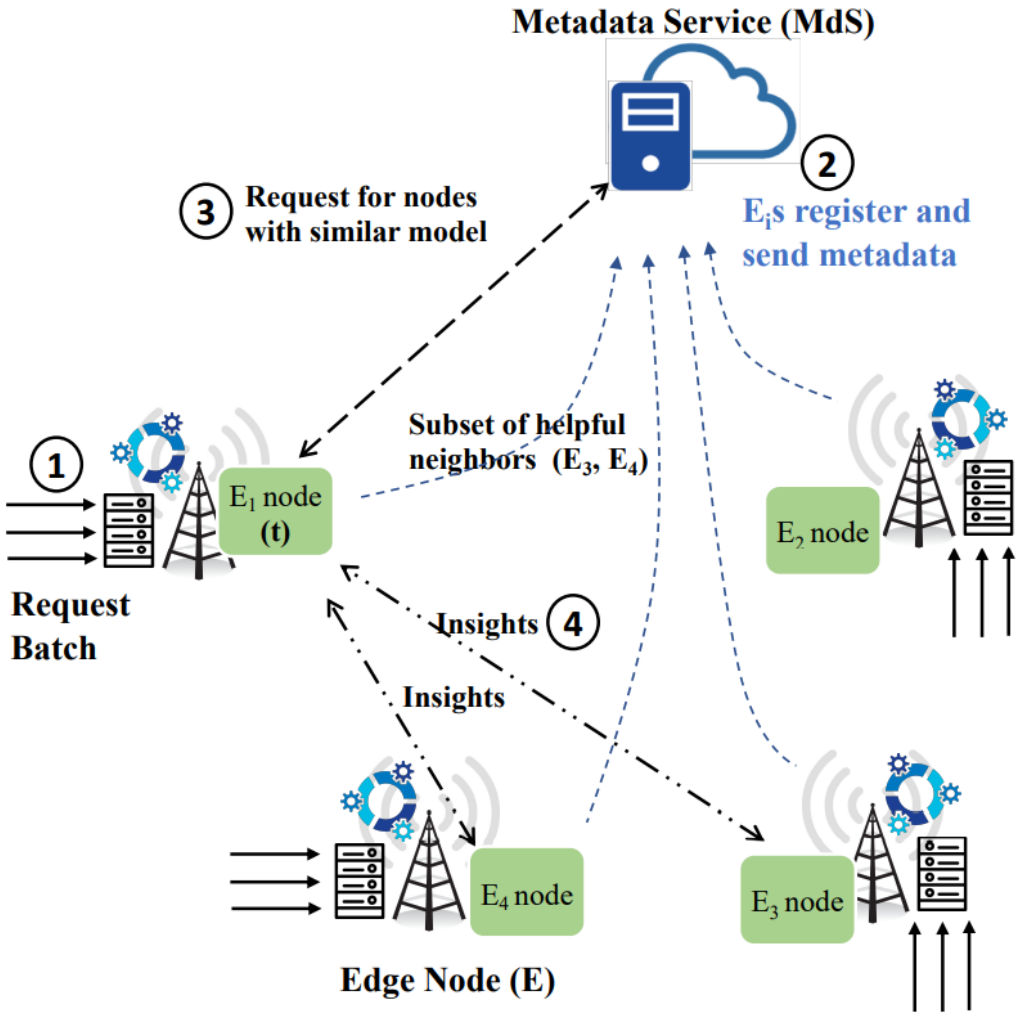


NOKIA Bell Labs Georgia Tech

Summary

- We introduce **Cartel**, a system for sharing customized machine learning models between edge nodes.

- Benefits of Cartel include:
  - Adapts quickly to changes in workload (up to 8x faster compared to an isolated system).
  - Reduces total data transfer costs significantly (1500x ↓ compared to a centralized system).
  - Enables use of smaller models (3x ↓) at an edge node leading to faster training (5.7x ↓) when compared to a centralized system.



NOKIA Bell Labs Georgia Tech

Questions

- What is the added cost of aggregating the meta data?
  - They leave this as a black box but it could be significant
- What are some real-world scenarios where we would actually have MEC's that are training on related data?