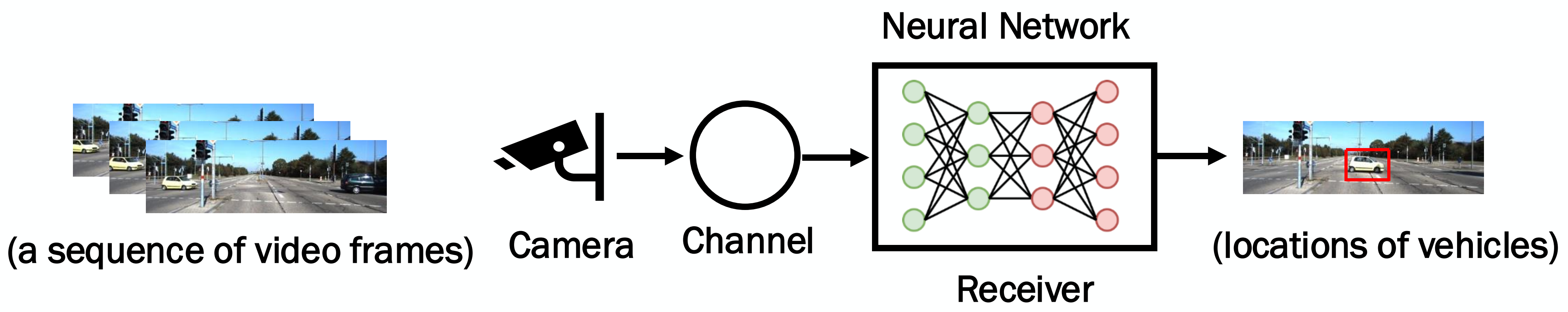
Computation and Communication Co-scheduling for Timely Multi-Task Inference at the Wireless Edge

Md Kamran Chowdhury Shisher, Adam Piaseczny, Yin Sun* and Christopher G. Brinton

Dept. of ECE, Purdue University, USA
Dept. of ECE, Auburn University, USA*

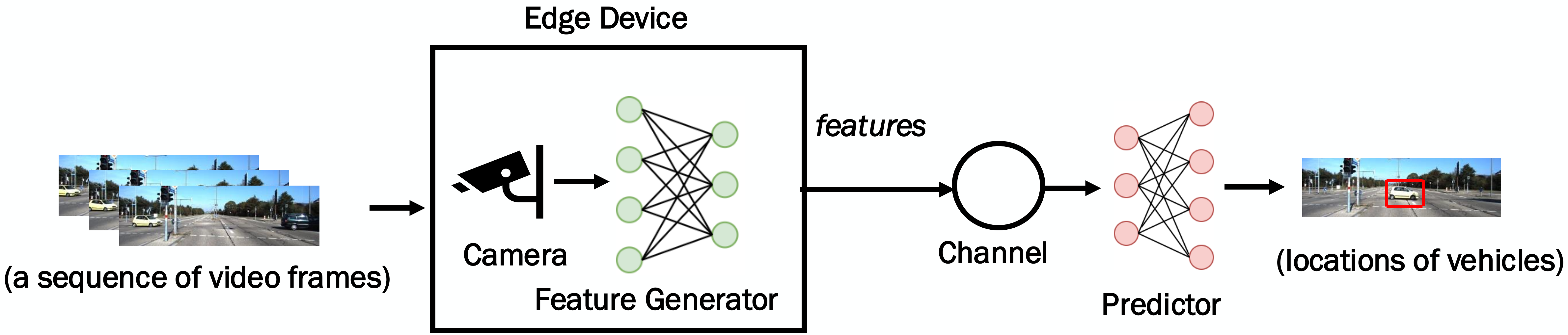
PURDUE UNIVERSITY

Remote Inference



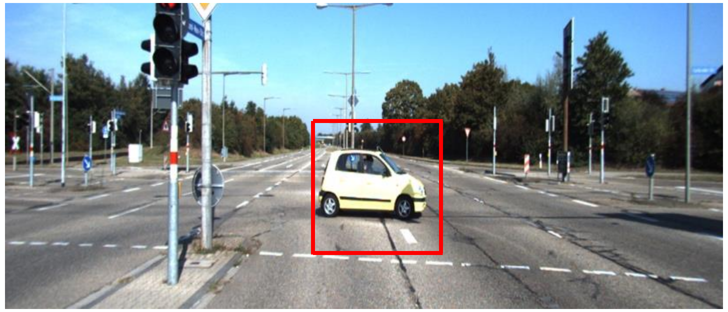
- A sensor or a camera sends observed signals (e.g., a sequence of video frames) to a receiver that predicts a time-varying target (e.g., location of vehicles) by using a neural network.
- Due to limited communication resources, sending high dimensional sensor observations to a remote receiver is not efficient. The observations can get delayed, and information can be lost.
- This will yield inaccurate inference that can affect real-time decision for critical applications.

Remote Inference



- We *split* the neural network to **feature generator** and **predictor.**
- Use feature generator in the edge device which takes the signal observation and generates low dimensional features, then sends them to the receiver.

Multi-Task Inference



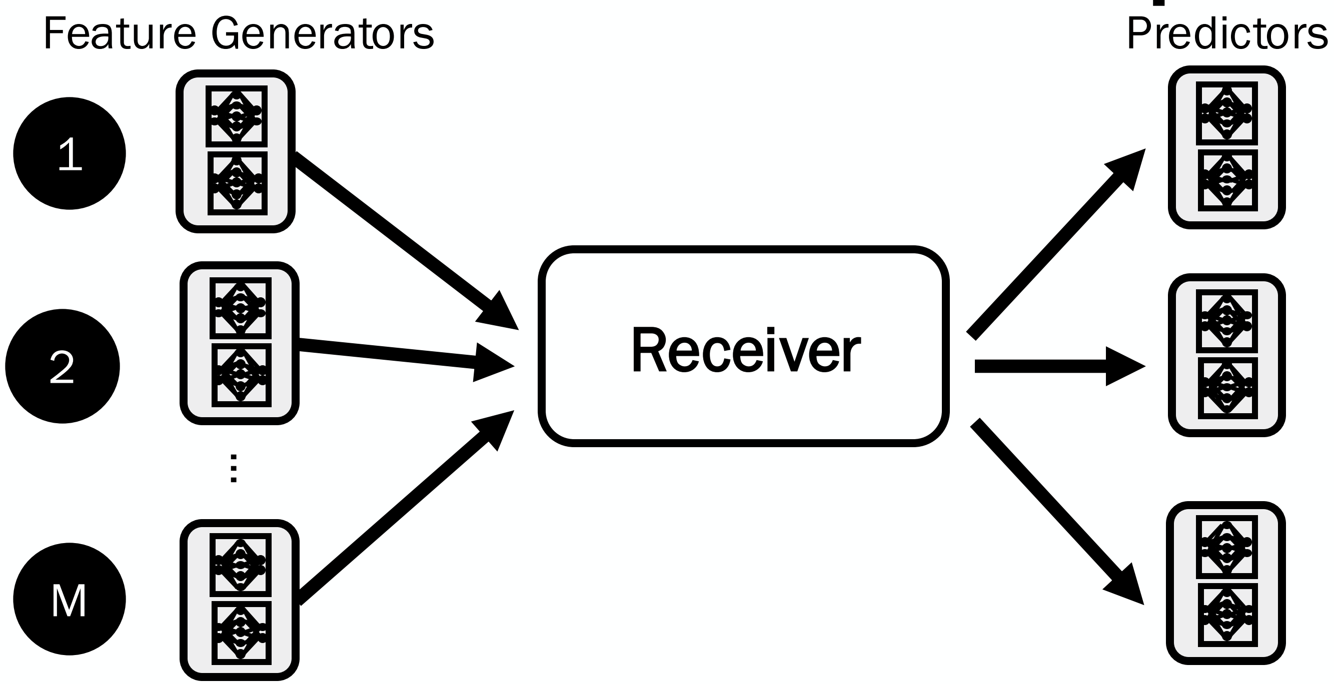
- Predict locations of vehicles
- Classify road signs

- Classify friendly Vs. hostile agents
- Predict position of own agents

- Classify customer reaction
- Predict current inventory

- From **autonomous vehicle, military, smart retail to Digital Twin**, edge device may need to perform **multiple inference tasks.**
- Edge device may have **limited computation resources** to **generate features** for all tasks at the same time slot.

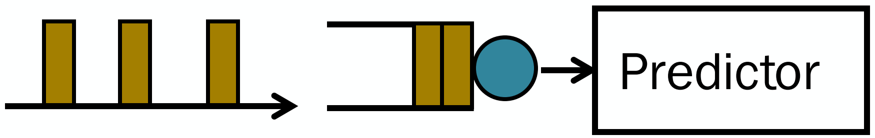
Multi-Task Remote Inference with Multiple Sources

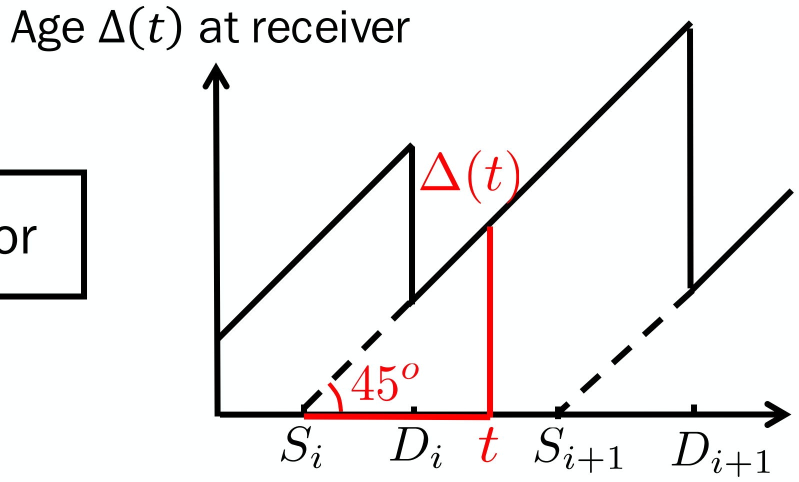


- M edge devices are connected to a receiver
- Receiver predicts K_m targets for each device m based on the most recently delivered features
- Delivered features may not be fresh due to limited computation and communication resources.
- We use Age of Information (AoI) to measure freshness

How can we develop a computation and communication co-scheduling methodology to minimize the inference errors across tasks while adhering to network resource constraints?

Age of Information





Definition: At time t , the **Age of Information (AoI)** $\Delta(\mathrm{t})$ is time difference between the current time t and the generation time $t - \Delta(t)$ of the freshest received feature

If feature i is generated at S_i and delivered at D_i

$$\Delta(t) = \underbrace{t - \max\{S_i : D_i \leq t\}}_{\text{Time difference between data generation and usage}}$$

Inference Error

Definition: At time t, the Inference error for j-th inference task of source m can be expressed as a function of AoI $\Delta_{m,j}(t) = \delta$ [Shisher and Sun, MobiHoc' 22, Shisher et al, ToN' 24]

$$p_{m,j}(\delta) = \mathbb{E}_{Y,X \sim P_{Y_{m,j,t}, X_{m,t-\delta}}} \left[L_{m,j}(Y, \psi_{m,j}(\phi_{m,j}(X), \delta)) \right]$$

- Notations:
  - $X_{m,t-\delta}$ denote the signal value generated δ time slots ago from m-th edge devices
  - $\phi_{m,j}(\cdot)$ and $\psi_{m,j}(\cdot)$ are feature generator and predictor functions, respectively
  - $L(y, \hat{y})$ measures the incurred loss when the actual target is y and the predicted value is $\hat{y}$

Our results can be applied to any loss function $L(y, \hat{y})$. Some examples are: 0-1 loss, quadratic loss, and log loss

Problem Formulation

- Our goal is to **minimize infinite horizon discounted sum of inference errors** for all inference tasks subject to computation and communication resource constraints:

$$\bar{p}_{opt} = \inf_{\pi \in \Pi} \sum_{t=0}^{\infty} \frac{\gamma^t}{K} \sum_{m=1}^{M} \sum_{j=1}^{k_m} \mathbb{E}_\pi \left[p_{m,j}(\Delta_{m,j}(t)) \right],$$

$$\text{s.t.} \underbrace{\sum_{j=1}^{k_m} \pi_{m,j}(t) \leq C_m, t = 0, 1, \ldots, m = 1, \ldots, M,}_{\text{Computation Resource Constraints}}$$

$$\underbrace{\sum_{m=1}^{M} \sum_{j=1}^{k_m} \pi_{m,j}(t) \leq N, \quad t = 0, 1, 2, \ldots,}_{\text{Communication Resource Constraint}}$$

Combinatorial Decision Problem

- $M + 1$ constraints
- M computation constraints, 1 for each device m
- 1 communication constraint shared by all devices
- **Weakly Coupled MDP (PSPACE-Hard)**

- $\pi_{m,j}(t) = 1$: feature for j-th inference task of m-th device is generated and transmitted
- At most, N features can be sent at one time slot
- Edge device m can generate features for at most C_m tasks

Related Works

- AoI-based Scheduling:
  - Prior works [Kadota et al, ToN' 18, Shisher and Sun, MobiHoc' 22, Shisher et al, ToN' 24, Tripathi and Modiano ToN'24, Ornee and Sun MobiHoc' 23] considers only one communication constraint
  - Prior works are modeled as RMAB, a special case of weakly coupled MDP
  - Whittle Index Policy are used in RMAB provided that the problem is indexable

- Systematic introduction of the Remote Inference problem [Shisher and Sun, MobiHoc' 22, Shisher et al, ToN' 24]
- Learning and Communications Co-design for Remote Inference [Shisher et al, JSAIT' 23]

- Interpretation of Information Aging on Remote Inference
  - Information-theoretic interpretation of information aging for Markov signals [Sun and Cyr, SPAWC' 18, JCN' 19]
  - Information-theoretic interpretation of information aging for general non-Markov signals [Shisher et al, INFOCOM AoI Workshop'21, Shisher and Sun, MobiHoc' 22, Shisher et al, ToN' 24]
  - AR-model-based analysis/interpretation of information aging [Shisher and Sun, INFOCOM ASoI workshop' 24]
  - Experimental results of remote inference [Shisher and Sun, MobiHoc' 22, Shisher et al, ToN' 24, JSAIT' 23]

Q. How to design Computation and Communication Co-scheduling?

Lagrangian Primal Dual

- **Primal Problem: (Reoptimized) At every time $\boldsymbol{\tau}$** given AoI value $\Delta_{m,j}(\tau)$,
  we truncate the problem to T time slots and apply Lagrange multipliers to constraints

$$\bar{p}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \tau : T) = \inf_{\pi \in \Pi} \sum_{t=\tau}^{T} \sum_{m=1}^{M} \sum_{j=1}^{k_m} \frac{\gamma^t \mathbb{E}_\pi \left[p_{m,j}(\Delta_{m,j}(t))\right]}{K}$$
$$+ \sum_{t=\tau}^{T} \sum_{m=1}^{M} \lambda_{m,t} \frac{\gamma^t}{K} \left(\left(\sum_{j=1}^{k_m} \pi_{m,j}(t) \right) - C_m \right)$$
$$+ \sum_{t=\tau}^{T} \mu_t \frac{\gamma^t}{K} \left(\left(\sum_{m=1}^{M} \sum_{j=1}^{k_m} \pi_{m,j}(t) n_{m,j} \right) - N \right),$$

- **Dual Problem:** We obtain optimal Lagrange Multipliers after solving the dual problem

$$\max_{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \geq 0} \bar{p}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \tau : T)$$

PURDUE UNIVERSITY

Solution of Primal and Dual

- **Solution of Primal Problem**:

Decompose the Primal problem into **per-inference task problem**:

$$\bar{p}_{m,j}(\boldsymbol{\lambda}_m, \boldsymbol{\mu}; \tau : T) = \inf_{\pi_{m,j} \in \Pi_{m,j}} \sum_{t=\tau}^{T} \gamma^t \mathbb{E}_{\pi_{m,j}} \left[p_{m,j}(\Delta_{m,j}(t)) + \lambda_{m,t}\pi_{m,j}(t) + \mu_t \pi_{m,j}(t) n_{m,j} \right]$$

We solve the problem by dynamic programming:

$$\min_{\pi_{m,j}(t) \in \{0,1\}} \underbrace{Q^{\boldsymbol{\lambda}_m, \boldsymbol{\mu}}_{m,j,t}(\Delta_{m,j}(t), \pi_{m,j}(t))}_{\text{Action Value Function}}$$

- **Solution of Dual Problem**:

$$\max_{\boldsymbol{\lambda} \geq 0, \boldsymbol{\mu} \geq 0} \sum_{t=\tau}^{T} \sum_{m=1}^{M} \sum_{j=1}^{k_m} V^{\boldsymbol{\lambda}_m, \boldsymbol{\mu}}_{m,j,t}(\Delta_{m,j}(t))$$

Value Function

$$- \sum_{t=\tau}^{T} \sum_{m=1}^{M} \gamma^{t-\tau} \lambda_{m,t} C_m + \sum_{t=\tau}^{T} \gamma^{t-\tau} \mu_t N$$

Maximum Gain First Policy (Reoptimized)

- **Gain Index:**

$$\alpha_{m,j,t}(\delta) = Q_{m,j,t}^{\boldsymbol{\lambda}^*_m,\boldsymbol{\mu}^*}(\delta, 0) - Q_{m,j,t}^{\boldsymbol{\lambda}^*_m,\boldsymbol{\mu}^*}(\delta, 1)$$

- At time t, maximize sum of Gain Indices of all inference tasks subject to the resource constraints

At time t,

- We iterate through all tasks, ordered by their maximum gain indices.
- If constraints satisfies, we schedule the task

Algorithm 1: Maximum Gain First (MGF) Policy

for $t = 0, 1, \ldots$ **do**
    Update $\Delta_{m,j}(t)$ *for all* (m, j)
    Initialize $\pi_{m,j}(t) \leftarrow 0$ for all (m, j)
    Get $\boldsymbol{\lambda}^*$ *and* μ^* *that maximizes* $\bar{p}(\boldsymbol{\lambda}, \boldsymbol{\mu}; t : T)$
    $\alpha_{m,j} \leftarrow \alpha_{m,j,t}(\Delta_{m,j}(t))$ for all (m, j)
    $C_{m,\text{curr}} \leftarrow 0$ and $N_{\text{curr}} \leftarrow 0$
    $A(t) \leftarrow \{(m, j) : \alpha_{m,j} > 0)\}$
    while $A(t)$ *is not empty* **do**
        $(m^*, j^*) \leftarrow \arg\max_{m,j} \alpha_{m,j}$
        $c \leftarrow C_{m^*,\text{curr}} + 1$ and $n \leftarrow N_{\text{curr}} + n_{m^*,j^*}$
        if $c \leq C_{m^*}$ *and* $n \leq N$ **then**
            Update $\pi_{m^*,j^*}(t) \leftarrow 1$
            Update $C_{m^*,\text{curr}} \leftarrow c$ and $N_{\text{curr}} \leftarrow n$
        $A(t) = A(t) \setminus (m^*, j^*)$

Maximum Gain First Policy (Reoptimized)

Theorem: If all AoI functions $p_{m,j}(\delta)$ are bounded and the following holds

$$T \geq \log_{\frac{1}{\gamma}}\left(\sum_{m=1}^{M}\sqrt{k_m}\right),$$

then the MGF policy is asymptotically optimal as the number of inference tasks per source increases, i.e.,

$$\bar{p}_{\text{MGF}} - \bar{p}_{\text{opt}} = O\left(\frac{1}{\sum_{m=1}^{M}\sqrt{k_m}}\right)$$

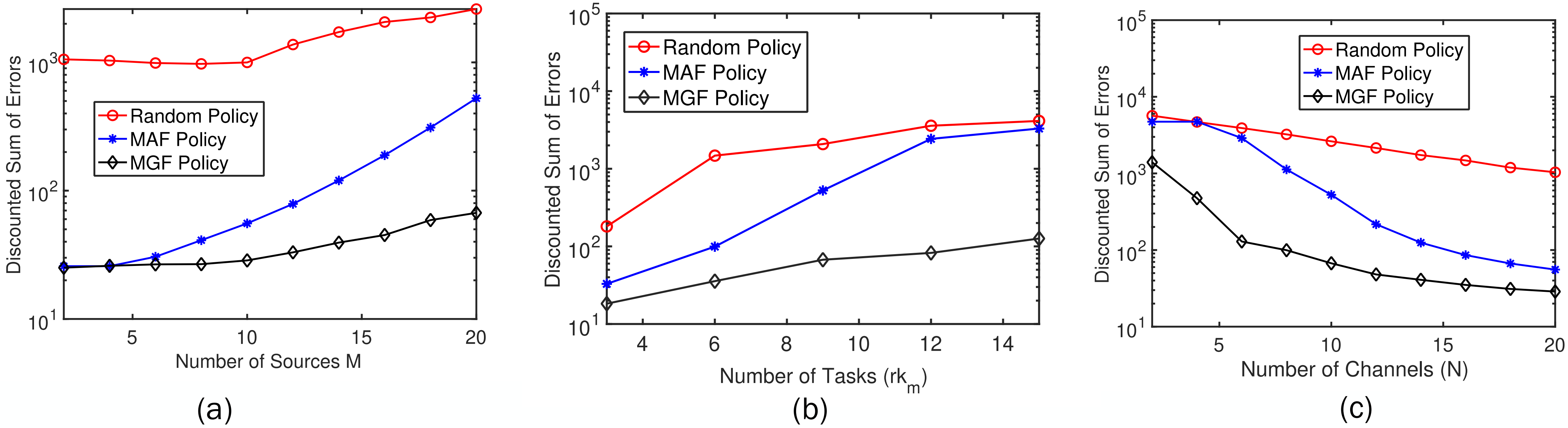
- [Brown and Zhang, *Operation Research*' 23] introduced Reoptimized Fluid Policy using LP and Occupancy measures. The paper considers all constraints are shared by all sources
- The optimality gap provided in our paper is tighter than

$$\bar{p}_{\text{MGF}} - \bar{p}_{\text{opt}} = O\left(\frac{1}{\sqrt{\sum_{m=1}^{M} k_m}}\right)$$

  of Reoptimized Fluid Policy provided in [Brown and Zhang, *Operation Research*' 23]
- This is because we use the structural information that M computation constraints are not shared by all devices M

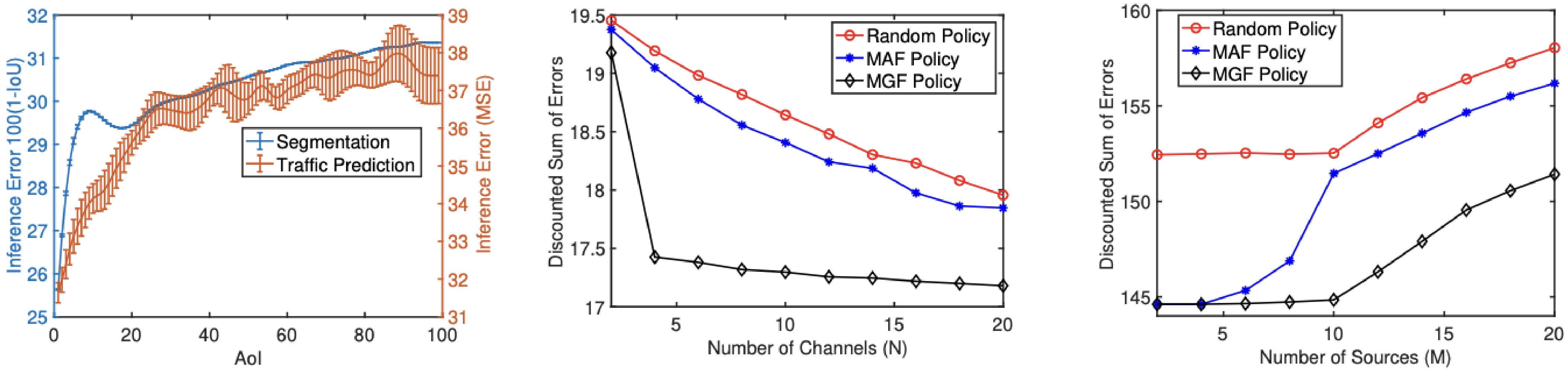
PURDUE UNIVERSITY

Simulation Results (Synthetic Evaluations)



- We consider linear AoI function, logarithmic AoI function, and exponential AoI function [Tripathi and Modiano ToN'24]
- (a) $N = 10, k_m = 3$, (b)$M = 20$, N $= 10$, (c) $M = 20, k_m = 3$
- Our policy 26 times better compared to Maximum AoI First (**MAF)** and 32 times better compared to **Random Policy**

Simulation Results (Real World Evaluations)



(a) Inference Error vs AoI (b) Dis. Sum of Errors vs. N (c) Dis. Sum of Errors vs. M

- We consider traffic prediction and segmentation using dataset collected from Next-generation Simulation Program of US Department of Transportation Federal Highway Administration

Summary

Use Inference Error as a metric to design Computation and Communication Co-scheduling

Inference Error = f(AoI)

Modeled as Weakly Coupled MDP

Proposed Lagrangian-based Reoptimized **(MGF)** Policy to solve the problem

- Computation + Communication Policy for optimizing AoI and Remote Inference
  - Remote Inference problem
  - AoI-based Scheduling

Future Works:

- How to optimally design feature generators and predictors?
- Jointly optimize learning, computation, and communication?
- System implementation

PURDUE UNIVERSITY

Thank You

Email: mshisher@purde.edu

PURDUE
UNIVERSITY®