Lightning Talks
Annual Review 2020

Platform Lab Students
Stanford University
Lightning Talks

1. **Group Communication for Granular Computing** – Yilong Li
2. **INFaaS: Model-less Inference Serving** – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. **Llama: Heterogeneous Serverless Video Analytics with SLAs** – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. **Multi-Agent Coordination for Flexible Manufacturing** – Oriana Claudia Peltzer
5. **Optimal Task Assignment and Route Planning for Robotic Assembly Planning** – Kyle Jordan Brown
6. **Exploring High-Performance Multi-Hop Communication in the Datacenter** – Collin Lee
7. **Edge Prefetching for Graph processing on reconfiguration hardware with HBM** – Timothy Chong
8. **Smart Harvesting of Spare CPU Cores in Cloud Servers** – Yawen Wang
9. **The Delay Issue for Video Object Detection - Problem and Solution** – Huizi Mao
10. **The (Almost) Self-Learning Camera** – Eyal Cidon
11. **GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks** – Kevin Nicholas Kininham
13. **On-Ramp: Using Precise Packet One-Way-Delays to Pause Flows at the Edge of the Network** – Shiyu Liu, Ahmad Ghalayini
14. **Building a Fair Exchange in the Cloud** – Vinay Sriram, Jinkun Geng
15. **AIoT Sensing framework over LPWAN** – Pan Hu
16. **WaDAR: Accelerating Data-Driven Agriculture with Wireless Soil Moisture Sensors** – Colleen Josephson
17. **FBOSS: Facebook Open Switching System** – Sean Choi
18. **The nanoPU: From Microservices to Nanoservices** – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Group Communication for Granular Computing
Yilong Li

- **Group communication is essential to granular computing**
  - MilliSort: shuffle, all-gather, etc.
  - MapReduce: shuffle
  - Matrix multiplication: all-gather
  - Graph analytics: reduce, broadcast

- **Efficient group communication is hard at scale**
  - Less data per message => RPC overhead dominates
  - No full-bisection bandwidth => network core congestion
  - Stragglers inevitable => huge impact on completion time

- **Group-Communication-as-a-Service (GCaaS)?**
  - Immediately useful to serverless workloads that use S3 to exchange ephemeral data
Lightning Talks

1. Group Communication for flash bursts – Yilong L
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
**Problem**

- **User:** hard to configure and manage
- **Provider:** waste resources
- **Cost-inefficient for both users and providers
- **Designing a good inference serving system is non-trivial**

**What We Need**

- **Managed:** Automatic model resource management and scaling
- **Model-less:** Select model variant to meet SLO-requirements

---

https://github.com/stanford-mast/INFaaS
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Video analytics is an immensely diverse workload

Existing solutions

- Require complicated management
- Are statically configured
- Do not support heterogeneous hardware
- Cannot handle complex pipelines
- Cannot respond to input-dependent behavior
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Multi-Agent Coordination for Flexible Manufacturing
Oriana Peltzer, Kyle Brown

Combinatorial search with deterministic travel time
Stochastic travel time model

Probabilistic collision-free routing
Collaborative object transport
1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
We study the multi-agent precedence-constrained task assignment and path-finding (PC-TAPF) problem. Robots in a factory environment must transport objects between manufacturing stations in a factory, where each object is the output of one manufacturing operation and the input to another.

### The Factory “Grid World” Environment

- **Environment**
  - \( n \): number of robots
  - \( m \): number of objects
  - \( s^i_t \): state of robot \( i \) at time \( t \)
  - \( o^j_t \): state of object \( j \) at time \( t \)
  - \( a^i_t \): action of robot \( i \) at time \( t \)

### Project Schedule

- **Project Schedule Graph**
  - \( G = (V, E) \): project schedule
  - \( v \in V \): schedule vertex
  - \( (v \rightarrow v') \in E \): schedule edge
  - \( v^0 \): start time of vertex \( v \)
  - \( v^t \): duration of vertex \( v \)
  - \( v^f \): end time of vertex \( v \)
  - \( T = \max_{v \in V} v^f \): project makespan

### Solving a PC-TAPF problem

- **Objective**: Minimize project makespan
- **Constraints**: No collisions, all precedence relations must be satisfied

### Our Algorithm

#### Level 1: Sequential Next-Best Assignment Search

- Optimal valid project schedule

#### Level 2: Conflict-Based Search

- Routing constraints

#### Level 3: Incremental Slack-Prioritized Search

- Partial route plan

#### Level 4: Slack-and-Collision-aware Tie-breaking A*

- Optimal path segment

### Results

Experimental results on a suite of 192 randomly generated test problems with varying numbers of robots and delivery tasks, as well as various levels of required collaboration between robots.

**Figure**: Runtime box plots are all on the same scale: the full algorithm (top); a single iteration of ISPS (bottom); and from top to bottom); a single iteration of ISPS (bottom). The total number of CBS iterations for the whole solution process is plotted 3rd from the top.

### Publications


<table>
<thead>
<tr>
<th>Lightning Talks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Group Communication for flash bursts – Yilong Li</td>
</tr>
<tr>
<td>2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero</td>
</tr>
<tr>
<td>3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero</td>
</tr>
<tr>
<td>4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer</td>
</tr>
<tr>
<td>5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown</td>
</tr>
<tr>
<td>6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee</td>
</tr>
<tr>
<td>7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong</td>
</tr>
<tr>
<td>8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang</td>
</tr>
<tr>
<td>9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao</td>
</tr>
<tr>
<td>10. The (Almost) Self-Learning Camera – Eyal Cidon</td>
</tr>
<tr>
<td>11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham</td>
</tr>
<tr>
<td>14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng</td>
</tr>
<tr>
<td>15. AoT Sensing framework over LPWAN – Pan Hu</td>
</tr>
<tr>
<td>17. FBOSS: Facebook Open Switching System – Sean Choi</td>
</tr>
<tr>
<td>18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown</td>
</tr>
</tbody>
</table>
EXPLORING HIGH-PERFORMANCE MULTI-HOP COMMUNICATION IN THE DATACENTER

COLLIN LEE

- Datacenter applications split into a larger number of smaller execution units
  - e.g. microservices, functions-as-a-service
- Communication between execution units can become a performance bottleneck
- Traditional RPC-style communication wasteful/inefficient for some communication patterns
- Multi-Hop communication pattern reduces latency and network load
- Improve latency and throughput of distributed applications
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee

7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Problem: Pull-based graph processing is more performant than push-based one but cannot easily take advantage of frontier optimization.

Objective: Develop a graph processing framework that exploits the parallelism of FPGA and high random bandwidth of HBM, and incorporates an edge frontier prefetching engine that allows high-throughput parallel frontier processing.
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on reconfiguration hardware with HBM – Timothy Chong

8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
On average, less than 50% of allocated CPU resources are used in a cluster from Google [1].

Harvest unused CPU cores from client VMs to run batch jobs to improve CPU utilization on VM host servers.

Use online cost-sensitive multiclass classification algorithm to predict peak CPU utilization of client VM.

1. Maximize harvested spare CPU resources while minimizing impact on co-located client VMs.
2. Quickly adapt to different CPU usage patterns.

[1] Reiss et al. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. SOCC’12
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang

9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
The Delay Issue for Video Object Detection - Problem and Solution

Huizi Mao

Latency = Computational latency + Algorithmic delay

Time to run an algorithm on specific platforms

Time for an algorithm to make valid predictions

- A new metric, Average Delay (AD), is proposed to measure algorithmic delay
  - The widely-used Average Precision (AP) metric only measures average accuracy over time
  - Existing AP-oriented works are found to deteriorate AD
- We study multiple methods to improve algorithmic delay
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
The (Almost) Self-Learning Camera
Eyal Cidon, Manabu Nakanoya, Sandeep Chinchali, Jenya Pergament

• Edge devices (drones, satellites, IoT) will be widely used for sensing tasks such as target tracking or automated surveillance.

• Often these devices will need to adapt quickly to their local environment.

• Due to high bandwidth and labeling costs collecting data for retraining in the cloud can be very expensive.

• Therefore we will need to retrain on the edge device itself.

=> Requires us to rethink the labeling pipeline
The (Almost) Self-Learning Camera
Eyal Cidon, Manabu Nakanoya, Sandeep Chinchali, Jenya Pergament
The (Almost) Self-Learning Camera
Eyal Cidon, Manabu Nakanoya, Sandeep Chinchali, Jenya Pergament
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks - Kevin Kinningham, Philip Levis, Chris Ré

- Graph Neural Networks = Deep Learning + Graphs
- Problem with existing DNN representations
  - Hard to represent irregular computation (e.g. message passing)
  - Limited support for sparse operations
  - Inefficient handling of large graphs
- Our solution: GReTA
  - **Simple**: Graph-framework based representation
  - **Flexible**: Wide variety of GNNs can be implemented
  - **Efficient**: Optimizations for execution on hardware accelerator using limited memory (e.g. dynamic partitioning)
1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham

14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Power Clocks: Dynamic Multi-Clock Management for Embedded Systems
Holly Chiang, Hudson Ayers, Daniel Giffin, Amit Levy, Philip Levis

Problem:
- Embedded systems operate on limited energy budgets, but manually specifying clock changes to save energy is difficult
- Current approaches to dynamic clock management are impossible in the case of multiprogrammed embedded systems

Solution:
- Move clock requests out of applications into peripheral drivers, and manage clock changes in the kernel
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
On-Ramp: Using Precise Packet One-Way Delays to Pause Flows at the Edge of the Network
Shiyu Liu, Ahmad Ghalayini

Motivation:
- Many recent congestion-control (CC) protocols have shown impressive performance in on-premise datacenters. However:
  - They require network or hardware support → cannot be deployed by cloud users
- Existing solutions in cloud suffer, especially during incast

Our Solution, On-Ramp:
- Can be deployed by cloud users in VMs without any network support
- A universal underlay: help any existing CC
  - by providing a forceful and quick reaction to transient overload

On Google Cloud, On-Ramp improves the 99th percentile RCT of incast by 2.8× with CUBIC and 5.6× with BBR.

More details and results in our talk and our poster!
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng


- Leverage Highly Synchronized Clock (Huygens) to Implement Fair Gaming Mechanism.

- CloudEx will be used in CS349F Course in Spring Quarter 2020!
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Srimam, Jinkun Geng
15. AIoT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
Image Compression for AIoT over LPWAN – Pan Hu

Uniqueness #1: IoT device with hardware AI support (AIoT)
- Resource constrained
- Less powerful

Uniqueness #2: Application Scenario
- Device has specific task
- Less versatile in images

Uniqueness #3: LPWAN
- Expensive
- Bi-directional traffic

Design #1: Carefully design DNN with quantization

Design #2: Capture prior information for each device

Design #3: Progressive update matching size of packet

AIoT compression framework
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
1. Average sensor setup is > $250
2. Not easy to install/maintain
3. Wireless data collection is expensive and unreliable

Cost for average US farm is more than $1 million!

Make the system cheaper and easier to install/maintain by putting RFID-like tags underground paired with a centralized radar reader.

Fast facts

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Power consumption</td>
<td>116uW</td>
</tr>
<tr>
<td>Battery life</td>
<td>15 years on 4xAA</td>
</tr>
<tr>
<td>Projected cost per tag</td>
<td>&lt; $9 USD</td>
</tr>
<tr>
<td>Sensing range</td>
<td>0-77cm</td>
</tr>
</tbody>
</table>

Backscatter sensing is within 2.4% of ground truth. Average error of commercial sensors is 1-5%.

more than $1 million!
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Srim, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu

17. **FBOSS: Facebook Open Switching System** – Sean Choi
18. **The nanoPU: From Microservices to Nanoservices** – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
FBOSS: Facebook Open Switching System

Networks are Growing Fast!

Growth of FBOSS Deployments

- 1x
- 5x
- 10x
- 15x
- 20x
- 30x

Months →

- External Software
- Protocols (BGP, ECMP)
- Network Configurator
- Monitoring Service

Switch Software

FBOSS

Switch Hardware

Switch ASIC
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Srim, Jinkun Geng
15. AloT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown
The nanoPU: From **Microservices** to **Nanoservices**

Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown, Changhoon Kim

---

**The lowest latency Ethernet NIC ever built**

**RISC-V Prototype**

**Nanoservice App** | **nanoPU Speedup**
--- | ---
NFV Streaming | 3.5-4x
NN Inference | 10x
State-Space Search | 4-6x

**HW Thread Scheduling Evaluations**

![ECDF Diagram](image)

- 20x lower std dev
Lightning Talks

1. Group Communication for flash bursts – Yilong Li
2. INFaaS: Model-less Inference Serving – Qian Li, Neeraja Yadwadkar, Francisco A Romero
3. Llama: Heterogeneous Serverless Video Analytics with SLAs – Mark Yuqi Zhao, Neeraja Yadwadkar, Francisco A Romero
4. Multi-Agent Coordination for Flexible Manufacturing – Oriana Claudia Peltzer
5. Optimal Task Assignment and Route Planning for Robotic Assembly Planning – Kyle Jordan Brown
6. Exploring High-Performance Multi-Hop Communication in the Datacenter – Collin Lee
7. Edge Prefetching for Graph processing on coarse-grain reconfiguration hardware with HBM – Timothy Chong
8. Smart Harvesting of Spare CPU Cores in Cloud Servers – Yawen Wang
9. The Delay Issue for Video Object Detection - Problem and Solution – Huizi Mao
10. The (Almost) Self-Learning Camera – Eyal Cidon
11. GReTA: A Hardware Optimized Graph Processing Framework for Graph Neural Networks – Kevin Nicholas Kininham
14. Building a Fair Exchange in the Cloud – Vinay Sriram, Jinkun Geng
15. AoIT Sensing framework over LPWAN – Pan Hu
17. FBOSS: Facebook Open Switching System – Sean Choi
18. The nanoPU: From Microservices to Nanoservices – Stephen Ibanez, Alex Mallery, Muhammad Shahbaz, Nick McKeown