Lightning Talks
Platform Lab Students
Stanford University
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
CURP: Exploiting Commutativity for Practical Fast Replication

- Problem: replication doubles latency (1 $\rightarrow$ 2 RTTs)! Let's bring it back to 1 RTT.
- Why? Consistent replication must guarantee:
  - Consistent Ordering
  - Durability

Traditional protocols combined the two
CURP: Exploiting Commutativity for Practical Fast Replication

● Separate durability from ordering
  ○ Achieve **durability without ordering**
  ○ Defer ordering if **commutative**
    => Updates complete in 1 RTT.

● Result
  ○ RAMCloud’s performance improvement
    ■ 0.5x latency (13.8 µs → 7.3 µs)
      ● + 0.4 µs to no replication
    ■ 3.8x Throughput
  ○ Redis cache is now fault-tolerant with small cost (~12%)
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
14. Graph Databases and Graph Database Benchmarking – Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training – Song Han
16. Towards Intelligent Voice Assistants for SPNs – Vin Sachidananda
INITIAL DESIGN THOUGHTS FOR A GRANULAR COMPUTING PLATFORM
REQUEST FOR FEEDBACK

INITIAL DESIGN THOUGHTS FOR A GRANULAR COMPUTING PLATFORM

What is a granular computing platform?

- **Platform** for applications composed of a **large number of very small tasks** that **rapidly scale up/down**.
- 10 - 100µs granules
- Bursty workloads
- Dynamic code paths

Goal of poster: stimulate discussion.

Assertions:

- All code needs to be pre-loaded and ready to run.
- A shared polling infrastructure is needed.
- Trade memory protection for low latency.
- Scheduling must be decentralized.
- Fast resource preemption is necessary.
- Granules can only communicate via invocation.
- Durability of granules must be batched.
- Need reliable networks for low latency (geo-)replication.
- Storage system must expose data location.
- All feedback welcome.
- Poster session later today.
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
3. **Nanolog: A Nanosecond Scale Logging System** – Stephen Yang
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
NanoLog: A Nanosecond Scale Logging System

- 10-100x faster than existing solutions such as Log4j2 or spdlog
  - Achieves a throughput of over **80M log messages/second** at a **7-19ns** median latency

- Key Concept: Shift work out of Runtime
  - Extract static log information at Compile-Time
  - Only log dynamic information in binary format at Runtime
  - Postpone message formatting until Post-Execution

### Latency / System

<table>
<thead>
<tr>
<th>Latency / System</th>
<th>NanoLog</th>
<th>spdlog</th>
<th>Log4j2</th>
<th>Boost</th>
<th>glog</th>
<th>ETW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Latency (ns)</td>
<td>7</td>
<td>214</td>
<td>174</td>
<td>1764</td>
<td>1198</td>
<td>161</td>
</tr>
<tr>
<td>99.9th Latency (ns)</td>
<td>37</td>
<td>2546</td>
<td>3364</td>
<td>3772</td>
<td>5969</td>
<td>2967</td>
</tr>
</tbody>
</table>

[GitHub]: https://github.com/PlatformLab/NanoLog
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at µs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
DIY Hosting for Online Privacy

When using the web, users sacrifice privacy for high availability and low cost by relying on centralized services such as Gmail, Office 365, etc.

Strawman: Host Your Own VM! But too expensive esp. With high availability.

Deploy It Yourself (DIY) - Use serverless platforms to deploy personal web applications with high availability, low cost, and better privacy*

*Come see the poster to learn more!
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. **Cloud-configured backscatter cameras** – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Cloud-configured backscatter cameras

A core problem in resource-scarce IoT is making sure devices provide the QoS necessary for a given application.

Solution: use cloud edge devices to observe network conditions and process images to detect important events, and use that information to optimize the camera configuration for current/predicted conditions.
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. **Centralized Scheduling of Mixed Workloads at μs Granularity** – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Centralized Scheduling of Mixed Workloads at μs Granularity

- Data centers' leaf node servers have strict tail latency targets at μs-scale.
- Typically a server runs conservatively a single latency-critical application in order to satisfy tail latency constraints, which leads to underutilization.

---

Centralized Scheduling of Mixed Workloads at μs Granularity

Project Objective:

Develop scheduling policies and mechanisms that allow multiple tenants of different services to run on the same machine while maintaining each tenant's service level objective.
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Position tracking for virtual reality using commodity WiFi
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Homa: Low-Latency Transport For Datacenters

**Key ideas:**

- Minimize buffer occupancy
  - Manage congestion from receiver by scheduling incoming packets
- Minimize Short Messages Latency:
  - Small amount of unscheduled packets per message
  - Use a few network priorities to prioritize small messages
- High Network Bandwidth Utilization:
  - Receiver Controlled Overcommitment
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at µs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
An Implementation of the Homa Transport Protocol in RAMCloud

● **RAMCloud::HomaTransport: A Working Implementation**
  ○ Unusual features: message-oriented, connectionless, no ACKs, etc.

● **Excellent Performance**
  ○ Extreme network condition: **80%** network load on 10 Gbps network
  ○ Slowdown of 99%-tile latency of almost all message sizes within **2-3.5x**
  ○ 99%-tile round-trip latencies for small messages < **15 µs**
  ○ Nearly **100x** faster than best published result

● **Talk Later Today**
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
**Arachne: We made multi-core great again**

<table>
<thead>
<tr>
<th>Operation</th>
<th>Arachne</th>
<th>std::thread</th>
<th>Go</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thread Creation</td>
<td>160</td>
<td>5760</td>
<td>261</td>
</tr>
<tr>
<td>Condition Variable Notify</td>
<td>195</td>
<td>4137</td>
<td>317</td>
</tr>
<tr>
<td>Yield</td>
<td>103</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Null Yield</td>
<td>16</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Schedule threads in user land.

Allocate cores in user land.

### Diagram:

- **Before core request**: Blocked threads
- **Allocate dedicated core**: Cores
- **Unblock thread**: Cores
Arachne: Making memcached mightier

Classical Memcached

Arachne Memcached

Better load balancing
Arachne: Making memcached mightier

**Classical Memcached**

Fixed number of kernel threads, determined at startup.

**Arachne Memcached**

Variable number of dedicated kernel threads, determined by load. 😊 *Better resource utilization*
Arachne: Giving cores back to applications

- Giving applications knowledge of core topology and fine-grained control over scheduling.

**Application**

- User Threads
- Core Manager

**Cores**

- Request Cores
- Assign Cores
1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Salsify: Low-Latency Network Video Through Tighter Integration Between Video Codec and Transport Protocol

Salsify is a real-time video system with 3.9× lower delay and 2.7 dB higher visual quality, on average, compared to FaceTime, Hangouts, Skype, and WebRTC.

- Three major design improvements:
  - Video-Aware Transport Protocol
  - Functional Video Codec
  - Unified Control Loop

To appear in NSDI’18
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li

12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads

Many interesting pipelines on video:

- Transformations, filters, machine learning algorithms, etc.

"Remake Star Wars Episode I without Jar Jar."
The Problem
Currently, running such pipelines on videos takes hours and hours, even for a short video.

The Question
Can we achieve interactive collaborative video editing by using massive parallelism?
The challenges

- Low-latency video processing would need thousands of threads, running in parallel, with instant startup.

- However, the finer-grained the parallelism, the worse the compression efficiency.
Enter *ExCamera*

- We made two contributions:
  - Framework to run **5,000-way parallel jobs** with IPC on a commercial "cloud function" service.
  - Purely functional video codec for **massive fine-grained parallelism**.
- We call the whole system *ExCamera*. 
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Users often need to delegate their SSH credentials to remote partially-trusted hosts (e.g., to `git clone` from GitHub to AWS).

The popular `ssh-agent forwarding` offers minimal assurances—putting user’s credentials at risk.

**Guardian Agent** allows the user to verify and enforce security policies: **who** can do **what to whom** with the user’s credentials.

Dmitry Kogan, Henri Stern, Ashley Tolbert, David Mazières, Keith Winstein
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan

14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Graph Databases and Benchmarking

github.com/PlatformLab/TorcDB.git
1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Deep Gradient Compression

saving the communication bandwidth of distributed training by 500x
# Deep Gradient Compression

<table>
<thead>
<tr>
<th>Model</th>
<th>Task</th>
<th>Baseline</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ResNet-50 On ImageNet</strong></td>
<td>Top-1 Accuracy</td>
<td>75.96%</td>
<td>76.15% (+0.19%)</td>
</tr>
<tr>
<td></td>
<td>Top-5 Accuracy</td>
<td>92.91%</td>
<td>92.97% (+0.06%)</td>
</tr>
<tr>
<td></td>
<td>Compression Ratio</td>
<td>1 ×</td>
<td>277 ×</td>
</tr>
<tr>
<td><strong>5-Layer GRU On LibriSpeech Corpus</strong></td>
<td>Word Error Rate (WER) On test-clean</td>
<td>9.45%</td>
<td>9.06% (-0.39%)</td>
</tr>
<tr>
<td></td>
<td>Word Error Rate (WER) On test-other</td>
<td>27.07%</td>
<td>27.04% (-0.03%)</td>
</tr>
<tr>
<td></td>
<td>Compression Ratio</td>
<td>1 ×</td>
<td>608 ×</td>
</tr>
<tr>
<td><strong>2-Layer LSTM Language Model On Penn Treebank</strong></td>
<td>Perplexity</td>
<td>72.30</td>
<td>72.24 (-0.06)</td>
</tr>
<tr>
<td></td>
<td>Compression Ratio</td>
<td>1 ×</td>
<td>462 ×</td>
</tr>
</tbody>
</table>
Training Speedup / Scalability

Image Classification: no loss of accuracy

Training Speedup on GPU cluster with 1Gbps Ethernet

- AlexNet Baseline
- AlexNet TernGrad
- AlexNet Deep Gradient Compression
- DeepSpeech Baseline
- DeepSpeech TernGrad
- DeepSpeech Deep Gradient Compression

Perplexity of 2-Layer LSTM on PTB Dataset

- Baseline
- Deep Gradient Compression
Trevor: Automatic configuration and scaling of stream processing pipelines

- Please come to the full talk by Manu Bansal

- Many stream processing problems require throughput guarantees
- Provisioning for target throughput requires difficult manual tuning
Trevor: Automatic configuration and scaling of stream processing pipelines
Trevor: Automatic configuration and scaling of stream processing pipelines
Lightning Talks

1. Exploiting Commutativity for Practical Fast Replication – Seo Jin Park
2. Initial Design Thoughts for a Granular Computing Platform – Collin Lee
4. DIY Hosting for Online Privacy – Shoumik Palkar
5. Cloud-configured backscatter cameras – Colleen Josephson
6. Centralized Scheduling of Mixed Workloads at μs Granularity – Kostis Kaffes & Timothy Chong
7. Position tracking for virtual reality using commodity WiFi – Manikanta Kotaru
10. Arachne: Core Aware Thread Management – Henry Qin & Qian Li
12. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads – Sadjad Fouladi
13. Guardian Agent: secure ssh-agent forwarding - Dima Kogan
14. Graph Databases and Graph Database Benchmarking - Jonathan Ellithorpe
15. Deep Gradient Compression for distributed training - Song Han
16. Towards Intelligent Voice Assistants for SPNs - Vin Sachidananda
Towards Intelligent Voice Assistants for SPNs

- Thanks for listening to my talk, please come by our poster!