Lightning Talks

Platform Lab Students
Stanford University
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
MC Switch: Application-Layer Load Balancing on a Switch
MC Switch: Application-Layer Load Balancing on a Switch

```c
header_type memcached_binary_t {
    fields {
        magic : 8;
        opcode : 8;
        keyLen : 16;
        extraLen : 8;
        dataType : 8;
        status : 16;
        totalBodyLen : 32;
        opaque : 32;
        CAS : 64;
    }
}
```
MC Switch: Application-Layer Load Balancing on a Switch
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don’t wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
ReFlex: Remote Flash ≈ Local Flash

Why remote Flash?

- Independently scale storage from compute
- Share Flash over the network to improve utilization

What about performance? →

ReFlex achieves 850K IOPS/core
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashfield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don’t wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Flashield: a Key-value Cache that Minimizes Writes to Flash

• Widely-used key-value caches (Memcached, Redis) based on DRAM
• **Opportunity:** SSD is 10X cheaper than DRAM
• **Challenge:** Cache workload quickly degrades SSD durability
• **Solution:** Flashield is first general-purpose SSD key-value cache with same lifetime as DRAM
  • Same performance as DRAM caches
  • Leverage machine learning to reduce writes to SSD
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Homa: A Receiver-Driven Low-Latency Transport Protocol Using Network Priorities

Goals:

- Round-trip latency < 10µs for short messages
- Support high network load

Simulation results:

- 99th percentile tail latency for short messages within 2x of ideal, at 80% load
- Can support 90% bandwidth utilization for variety of workloads
- Outperforms both pFabric and pHost
Homa: A Receiver-Driven Low-Latency Transport Protocol Using Network Priorities

Key ideas:

- Manage congestion from receiver: receiver schedules incoming packets
  - Minimize buffer occupancy
- Small amount of unscheduled packets per message
- Use a few network priorities to prioritize small messages and resolve bipartite matching problem
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Distributing Simulations on Adaptive Grids

- OpenVDB is a hierarchical, sparse, geometric data structure, for representing 3D volumes with varying levels of details.
  - Resembles B+ trees.
  - Store data at multiple levels.
  - Uses geometric information for fast accessors.

- Partitioning dynamic geometric structures and computations over them is challenging:
  - Varying topology
  - Varying partition sizes
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Building Blocks of a Modular Notification System

Collin Lee and John Ousterhout

Motivation

- Popular notification systems are full-featured…
- … but that means applications pay for features they don’t need.
- New notification system focused on modularity, flexibility, and composability.
- Avoids unnecessary overhead in a unified system to improve performance for today’s applications or enable “granular computing.”

Want to chat?

Notifications, Applications, Granular Computing, Drones, and much much more!
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Parallelizing Computational Applications Requires High Task Throughput

Parallelize an application to more cores:

1. More tasks to keep cores busy.
2. Individual tasks are shorter.
3. More than 1 task per core gives scheduling flexibility.

A limited task throughput hurts parallelization.
Existing Solutions

Spark (centralized)

TensorFlow (distributed, but static)
Our Solution

A scheduling abstraction that enables nodes to spawn and executed tasks in parallel.

A lightweight control plane which pushes scheduling decisions to nodes, while nodes redistribute tasks accordingly.
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashfield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
NumFabric: Fast and Flexible Bandwidth Allocation in Datacenters

- Several proposals for bandwidth allocation frameworks in last 5 years
- Each to achieve a specific flow-level objective
- Which one does the operator chose if he wants flexibility?

NUMFabric provides a **flexible** data center fabric that is also **fast**
NumFabric: Flexibility

NUMFabric can flexibly implement several operator objective functions like minimize average flow completion time, weighted max-min, MPTCP objective, etc.
Why are existing NUM approaches not fast?

Overshooting might cause bloated queues and packet drops.

Larger steps to optimal

Smaller steps to optimal
NUMFabric: Fast

- In NUMFabric, sources give up direct control over rates.
- The sources specify "weights" and the Weighted Max-Min fabric allocates relative rates proportional to the weights of all flows.
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Grazelle

Hardware-Optimized

In-Memory

Graph Processing
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need ProactiveCongestion Control – Lavanya Jose and Steve Ibanez
Don't Wait for Sync to Achieve Strong Consistency!

- Primary-backup is widely used because of its simplicity
- For strong consistency, updates had to wait for replication to backup

<table>
<thead>
<tr>
<th>Primary-Backup</th>
<th>Async Replication (eg. Redis, TAO)</th>
<th>Strong consistency (eg. MySQL, RAMCloud)</th>
</tr>
</thead>
</table>

- Use "RPC retry" + async backup for consistent recovery

<table>
<thead>
<tr>
<th>Result:</th>
<th>Consistency</th>
<th>Strongest (linearizability)</th>
<th>Durability</th>
<th>Same as synchronous backup</th>
<th>RAMCloud write</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Latency</td>
<td>No delay by backup</td>
<td></td>
<td>14.6 μs → 7.8 μs</td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>Maximum</td>
<td>batching for backup</td>
<td></td>
<td>138 kops → 500 kops (~3x)</td>
<td></td>
</tr>
</tbody>
</table>
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don’t wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
NanoLog: A Nanosecond Scale Logging System

- 10-100x faster than its competitors such as Log4j2 or spdlog
  - Achieves a throughput of over **60M log messages/second** at a **12.5ns** 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

- Talk later today

<table>
<thead>
<tr>
<th>Zero Arguments</th>
<th>Boost v1.55</th>
<th>Log4j2</th>
<th>spdlog</th>
<th>NanoLog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput (Log/s)</td>
<td>0.82M</td>
<td>1.43M</td>
<td>1.50M</td>
<td>60.1M</td>
</tr>
<tr>
<td>Average Latency (ns)</td>
<td>1110 ns</td>
<td>697ns</td>
<td>668 ns</td>
<td>16.5ns</td>
</tr>
<tr>
<td>Median Latency (ns)</td>
<td>1080 ns</td>
<td>207ns</td>
<td>603 ns</td>
<td>13ns</td>
</tr>
</tbody>
</table>
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – ZI Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
TorcDB: TinkerPop on RAMCloud
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don’t wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Arachne: Make Multi-Core Great Again

It is time for applications to take scheduling back from the kernel!

Schedule threads in user land.

Allocate cores in user land.
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashfield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don’t wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
Self Driving Networks

SILC: a System for Sensing, Inference, Learning and Control for the Cloud Infrastructure
Lightning Talks

1. MC Switch: Application-Layer Load Balancing on a Switch – Eyal Cidon and Sean Choi
2. ReFlex: Remote Flash == Local Flash – Ana Klimovic
3. Flashield: a key-value cache that minimizes write to flash – Assaf Eisenman
5. Distributing Simulations on Adaptive Grids – Chinmayee Shah
6. Building Blocks of a Modular Notification System – Collin Lee
7. Dynamic, High-throughput Task Scheduling with Task Recipes – Hang Qu
8. NumFabric - Fast and Flexible Bandwidth Allocation in Datacenters – Kanthi Nagaraj
10. Don't wait for sync to achieve strong consistency – Seo Jin Park
12. TorcDB: TinkerPop on RAMCloud – Jonathan Ellithorpe
13. Arachne: Make Multi-Core Great Again – Henry Qin and Jacqueline Speiser
14. SILC: A Sensing, Inference, Learning and Control System for a Self Driving Network – Zi Yin and Shiyu Liu
15. High Speed Networks Need Proactive Congestion Control – Lavanya Jose and Steve Ibanez
High Speed Networks Need Proactive Congestion Control
High Speed Networks Need Proactive Congestion Control

Synchronous Scheme in NetFPGA SUME
No Per Flow State. Static Flows.

Holy Grail??
- Asynchronous Scheme
- Simple to implement
- No Per Flow State
- Dynamic Flows.