Improving Datacenter Performance and Robustness with Multipath TCP

Costin Raiciu, Sebastien Barre, Christopher Pluntke, Adam Greenhalgh, Damon Wischik, Mark Handley

SIGCOMM 2011
“Putting Things in Perspective”

• High performing network crucial for today’s datacenters

• Many takes…
  – How to **build** better performing networks
    • VL2, PortLand, c-Through
  – How to **manage** these architectures
    • Maximize link capacity utilization, Improve performance
    • Hedera, Orchestra, DCTCP, **MPTCP**
Modern datacenters provide many parallel paths...

- Traditional topologies are tree-based
  - Poor performance
  - Not fault tolerant

- Shift towards multipath topologies: FatTree, BCube, VL2, Cisco, EC2
Fat Tree Topology

(Fares et al., 2008; Clos, 1953)
Fat Tree Topology
(Fares et al., 2008; Clos, 1953)

K=4

Aggregation Switches

How to efficiently utilize the capacity?

Racks of servers

K Pods with K Switches each
State of the Art
(as discussed in Hedera)

Statically stripe flows across available paths using ECMP

Collision
How about mapping each flow to a different path?
How about mapping each flow to a different path?
How about mapping each flow to a different path?

Not fair
How about mapping each flow to a different path?

Not fair

Mapping each flow to a path is the wrong approach!
Instead, pool capacity from links
Use Multipath Transport

– Instead of using one path for each flow, use many random paths

– Don’t worry about collisions

– Just don’t send (much) traffic on colliding paths
Multipath TCP Primer
(IETF MPTCP WG)

- A drop in replacement for TCP
- Spreads application data over multiple sub flows

- For each ACK on sub-flow $r$, increase the window $w_r$ by $\min(\alpha/w_{total}, 1/w_r)$
- For each loss on sub-flow $r$, decrease the window $w_r$ by $w_r/2$
MPTCP better utilizes the Fat Tree network.
Understanding Gains

1. How many sub-flows are needed?

1. How does the topology affect results?

1. How does the traffic matrix affect results?
At most 8 sub-flows are needed

Total Throughput

Throughput (% of optimal)

TCP 2 3 4 5 6 7 8

Multipath TCP
MPTCP improves fairness in VL2

![Graph showing throughput vs rank of flow for Single Path TCP, MPTCP, 2 subflows, and MPTCP, 4 subflows in VL2.]
MPTCP improves throughput and fairness in BCube
Performance improvements depend on traffic matrix
MPTCP enables better topologies

In single homed topologies:

- Host links are often bottlenecks
- ToR switch failures wipe out tens of hosts for days

Multi-homing is necessary
MPTCP enables better topologies
MPTCP enables better topologies

Fat Tree Topology

Upper Pod Switch
ToR Switch
Servers
MPTCP enables better topologies

Dual Homed Fat Tree Topology
DHFT provides significant improvements when core is not overloaded.
EC2 Experiment

![Graph showing throughput vs. flow rank for TCP and MPTCP with different subflows. The graph indicates that MPTCP with 4 subflows outperforms both TCP and MPTCP with 2 subflows, especially when flow rank is high. The shaded area represents the same rack condition.]
Conclusion

• Multipath topologies need multipath transport

• Multipath transport enables better topologies
Thoughts (1)

• Old idea applied to datacenters
  – First suggested in 1995, then 2000s
  – Not very nice for middleboxes
• Works on a wide variety of topologies (as long as there are multiple paths)
• Number of advantages
  – Fairness
  – Balanced congestion
  – Robustness (hotspots)
  – Backward compatible with normal TCP
  – Can build optimized topologies
Thoughts (2)

• However…
  – Needs changes at all end-hosts
  – Benefits heavily depend on traffic matrix, congestion control
  – What’s the right number of sub flows?
  – No evaluation “in the wild”
  – No benefits for in-rack or many-to-one traffic
  – Prioritization of flows might be hard

How much benefit in practice?
Understanding Datacenter Traffic

• A few papers that analyzed datacenter traffic:
  – “Network Traffic Characteristics of Data Centers in the Wild” – IMC 2010
    • 3 US universities - distributed file servers, email server
    • 2 private enterprises - custom line-of-business apps
    • 5 commercial cloud data centers - MR, search, advertising, datamining etc.
Understanding Datacenter Traffic

“Most flows in the data centers are small in size (<10KB)…”. In other words, elephants are a very small fraction.
Understanding Datacenter Traffic

Majority of the traffic in cloud datacenters stay within the rack.
Understanding Datacenter Traffic

- Only a fraction of the existing bisection capacity is likely to be utilized at any given time $\Rightarrow$ no need for more bandwidth
- 25% of core links are hot spots at any time $\Rightarrow$ Load balancing mechanisms for spreading traffic across the existing links in the network’s core helpful
Understanding Datacenter Traffic

- Centralized controllers:
  - Significant amount of flows (~20%) arrive within 20us.
    - Parallelization important.
  - Most flows last less than 100ms
    - Reactive controllers add ~10ms overhead
    - This overhead might not be acceptable.

MPTCP would be useful, but totally depends on traffic
Backups
# Hedera vs MPTCP

<table>
<thead>
<tr>
<th></th>
<th>Hedera</th>
<th>MPTCP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Load balancing</strong></td>
<td>Centralized</td>
<td>Distributed</td>
</tr>
<tr>
<td><strong>Overhead</strong></td>
<td>Flow measurement, running scheduler</td>
<td>Creating sub-flows</td>
</tr>
<tr>
<td><strong>Deployment</strong></td>
<td>Open flow support in switches, central scheduler</td>
<td>Replace TCP stack</td>
</tr>
<tr>
<td><strong>Traffic differentiation</strong></td>
<td>Easier compared to a distributed solution</td>
<td>Hard(?)</td>
</tr>
<tr>
<td><strong>Optimality</strong></td>
<td>Centralized, so better optimized solution</td>
<td>Doesn’t have global view so might not be the most optimal</td>
</tr>
</tbody>
</table>
# DCTCP vs MPTCP

<table>
<thead>
<tr>
<th></th>
<th>DCTPC</th>
<th>MPTCP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deployment</strong></td>
<td>ECN support in switches, TCP stack changes</td>
<td>Replace TCP stack</td>
</tr>
<tr>
<td><strong>Coexistence</strong></td>
<td>Might throttle regular TCP flows due to the difference in congestion control</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Multi-homed topologies</strong></td>
<td>Cannot fully utilize since it still is single flow</td>
<td>Can fully utilize multi-homed topologies</td>
</tr>
</tbody>
</table>

Perhaps MP-DCTCP?