MultiPath TCP in OpenFlow Networks

Michael Bredel, Caltech@CERN

Outline

Motivation

MultiPath TCP

- Basics and Design Objectives
- Connection Setup
- Congestion Control and Fairness

OpenFlow Link-Layer MultiPath Switching

- OLiMPS - OpenFlow Link Layer MultiPath Switching
- Floodlight/OLiMPS OpenFlow Controller
- Path Calculation Engine

Preliminary Results

- International MultiPath OpenFlow Network

Multiple Paths?

Why do we need multiple paths?

- Data sets are growing exponentially
- Copying these data sets in reasonable time between sites requires a lot of bandwidth

Multiple Paths?

Why do we need multiple paths?

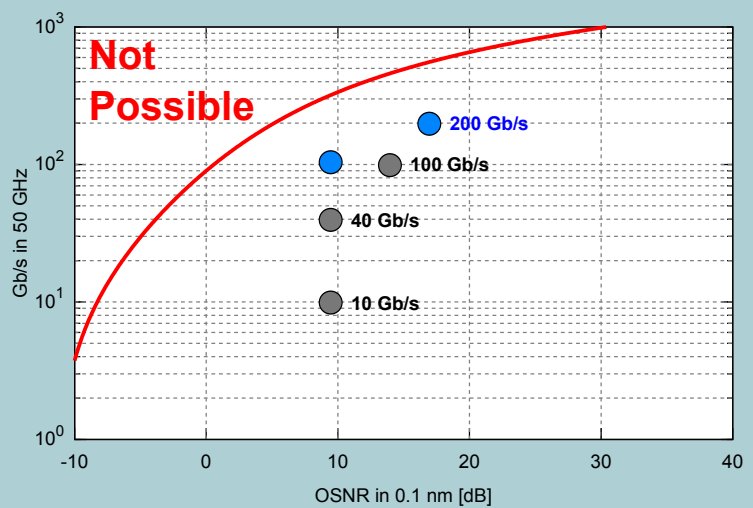
- Data sets are growing exponentially
- Copying these data sets in reasonable time between sites requires a lot of bandwidth

A single sperm has 37.5 MB of DNA information in it. That means a normal ejaculation represents a data transfer of arround 1.6 GB in about 3 seconds ... and you though 4G was fast.

Multiple Paths?

Why do we need multiple paths?

- Data sets are growing exponentially
- Copying these data sets in reasonable time between sites requires a lot of bandwidth
- 40 Gbit/s or 100 Gbit/s end-to-end is not always available (e.g. transatlantic) or to costly
- We are approaching the theoretical limit of fibre capacity



Multiple Paths?

Why do we need multiple paths?

- Data sets are growing exponentially
- Copying these data sets in reasonable time between sites requires a lot of bandwidth
- 40 Gbit/s or 100 Gbit/s end-to-end is not always available (e.g. transatlantic) or to costly
- We are approaching the theoretical limit of fibre capacity
- Probabilistic backlog and delay bounds [5]

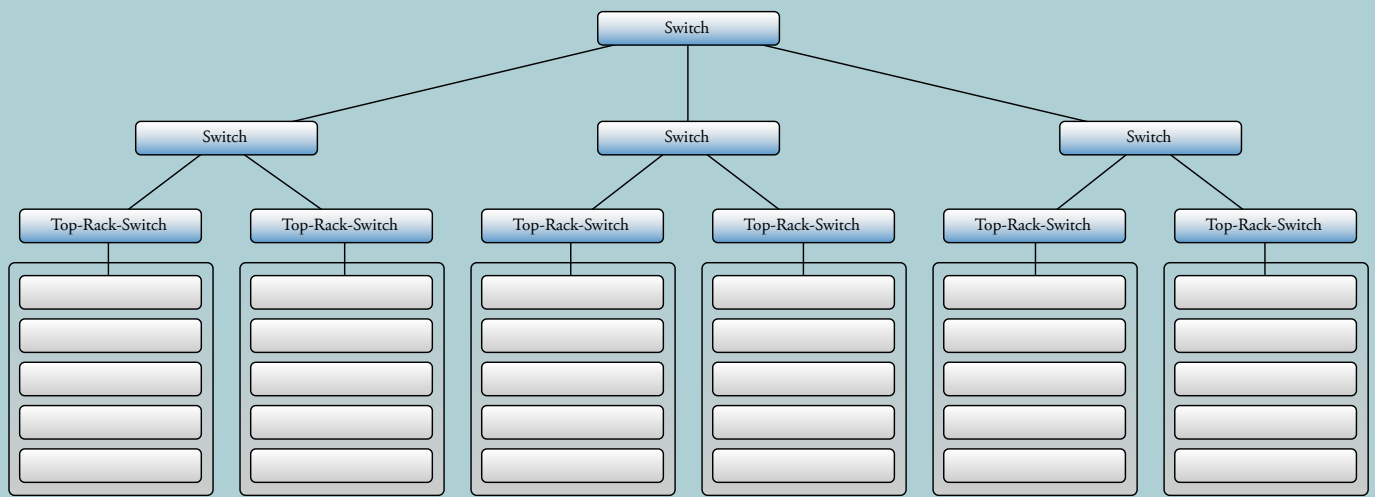
$$P[B \geq b] \leq \epsilon_s = \frac{\Gamma(\frac{1}{2\beta})}{2\beta(-\log \eta)^{\frac{1}{2\beta}}}$$

$$\eta = \exp\left(-\frac{1}{2\sigma^2}\left(\frac{C-\lambda}{H+\beta}\right)^{2(H+\beta)}\left(\frac{b}{1-(H+\beta)}\right)^{2-2(H+\beta)}\right)$$

Network Structure - Local Area Networks

Evolution of data center networks

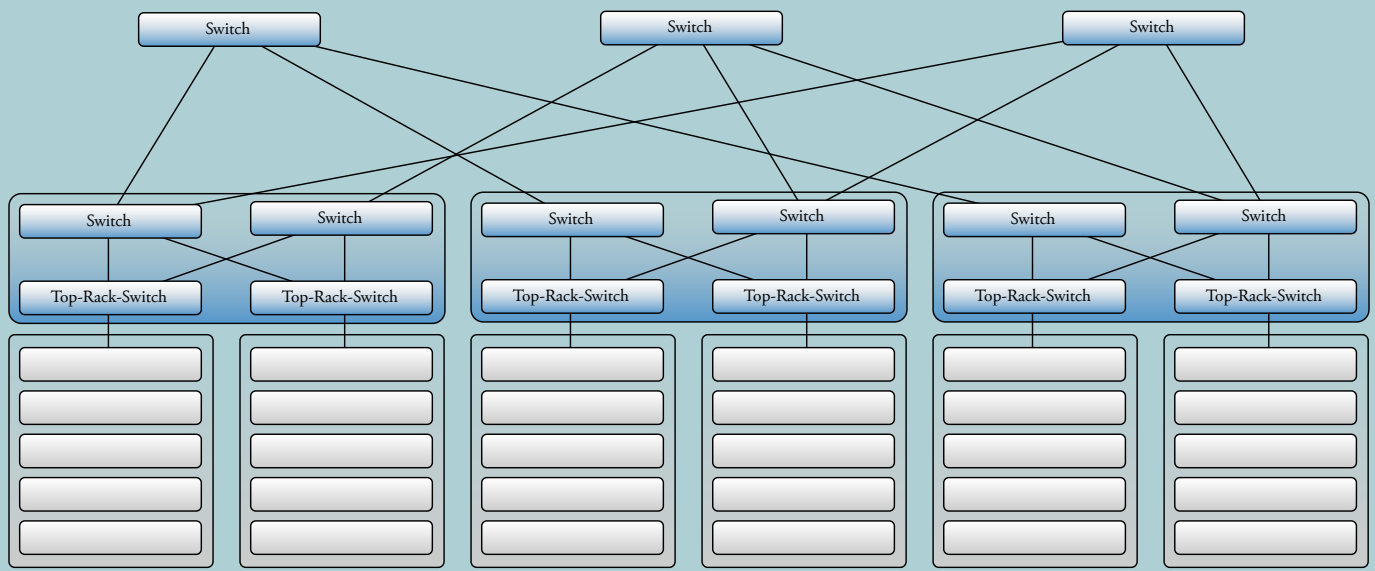
- Traditional topologies are tree based
  - Poor performance
  - Not fault tolerant
- Shift towards multipath topologies
  - FatTree [1], BCube [2], EC2



Network Structure - Local Area Networks

Evolution of data center networks

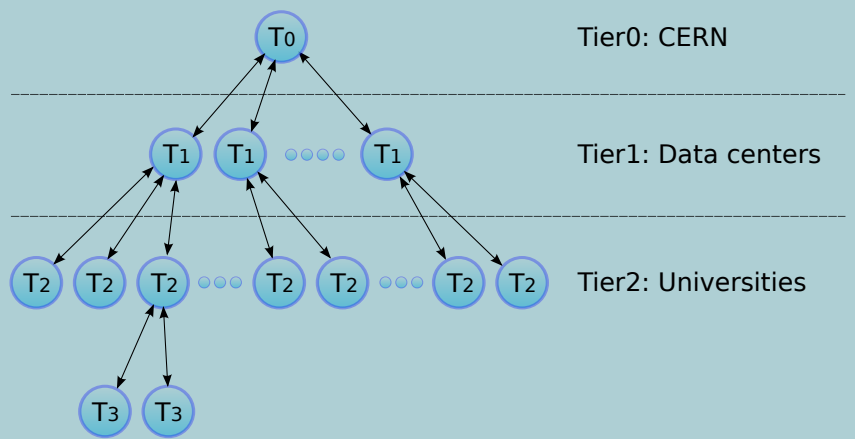
- Traditional topologies are tree based
  - Poor performance
  - Not fault tolerant
- Shift towards multipath topologies
  - FatTree [1], BCube [2], EC2



Network Structure - Wide Area Networks

LHC experiments and computing resources

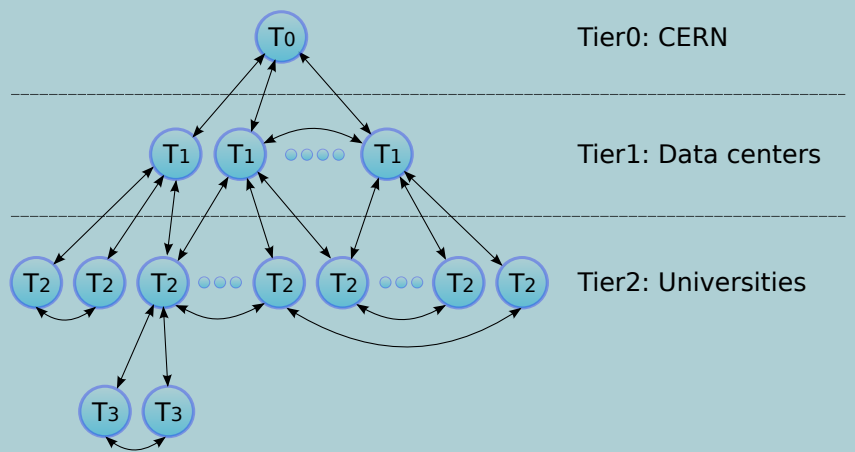
- Aims at allowing physicists to test the predictions of different theories, e.g. searching for the Higgs boson
- Hosts 4 big experiments
- Produce approx. 15-25 petabytes data per year
- The LHC Computing Grid connects 170 computer centres in 36 countries
- Challenges: Moving from a strict hierarchic model to a mashed grid



Network Structure - Wide Area Networks

LHC experiments and computing resources

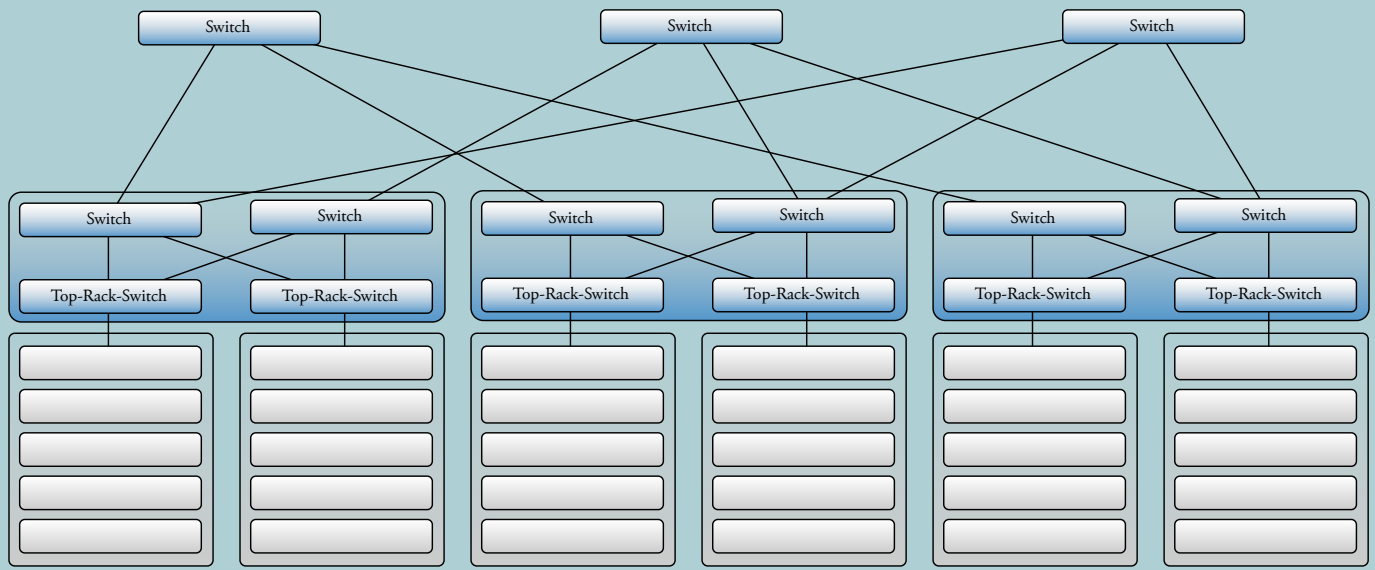
- Aims at allowing physicists to test the predictions of different theories, e.g. searching for the Higgs boson
- Hosts 4 big experiments
- Produce approx. 15-25 petabytes data per year
- The LHC Computing Grid connects 170 computer centres in 36 countries
- Challenges: Moving from a strict hierarchic model to a mashed grid



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

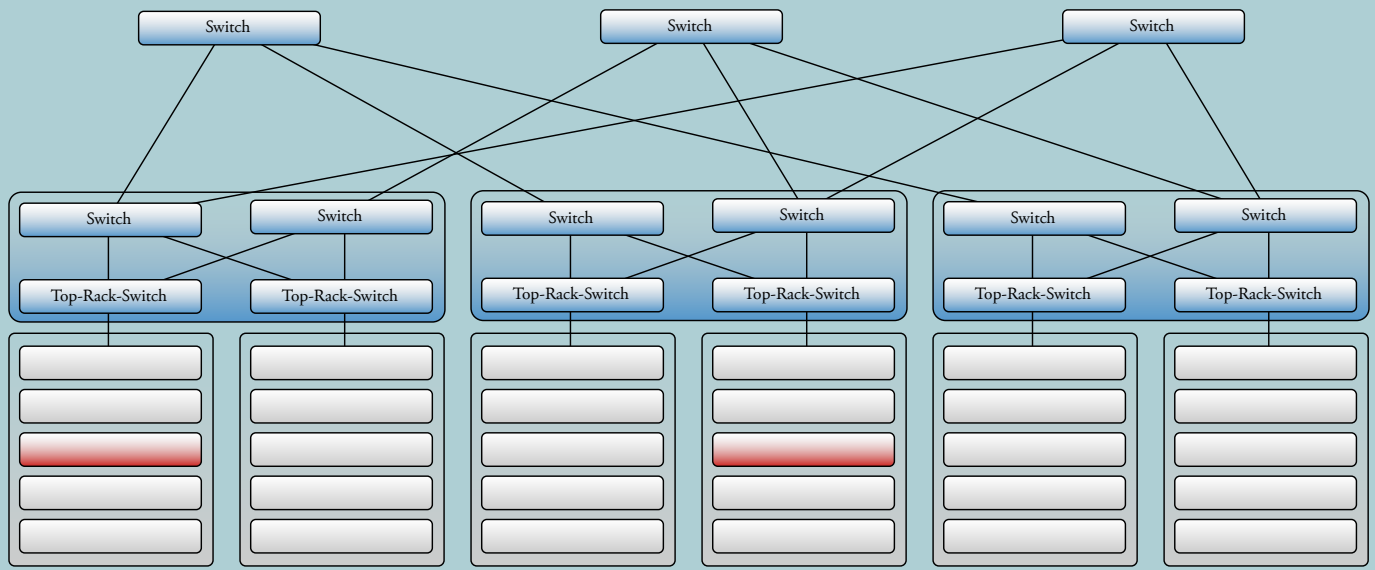
- Paths are chosen randomly
- Deploying an (unknown) hash function



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

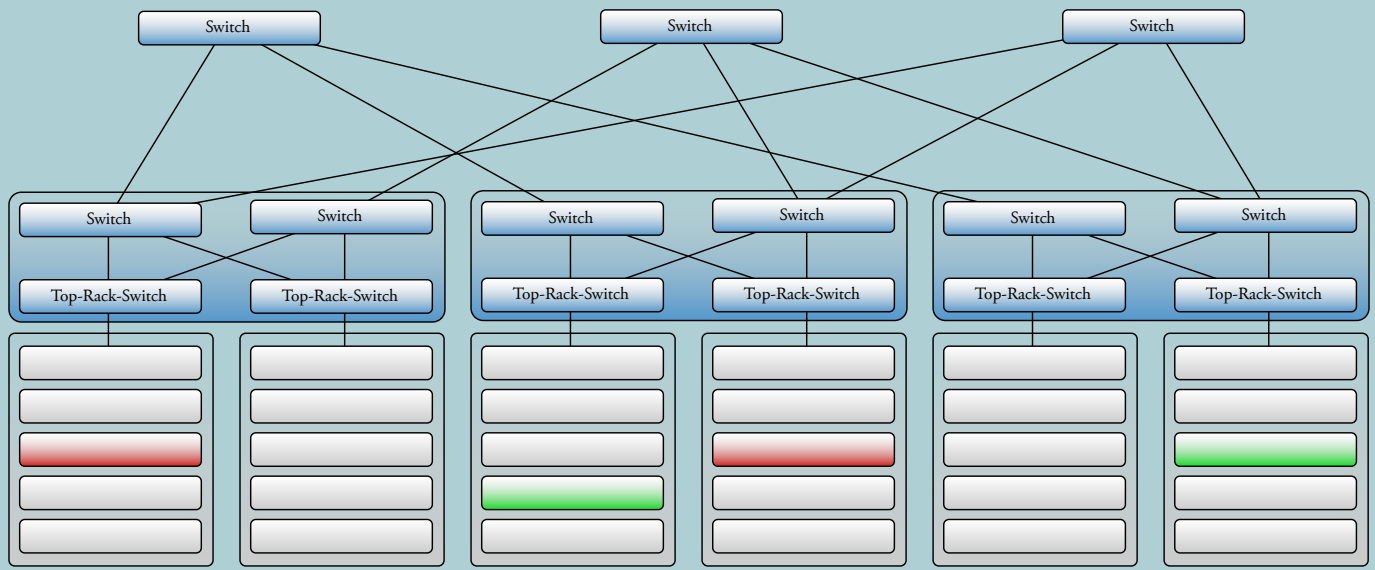
- Paths are chosen randomly
- Deploying an (unknown) hash function



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

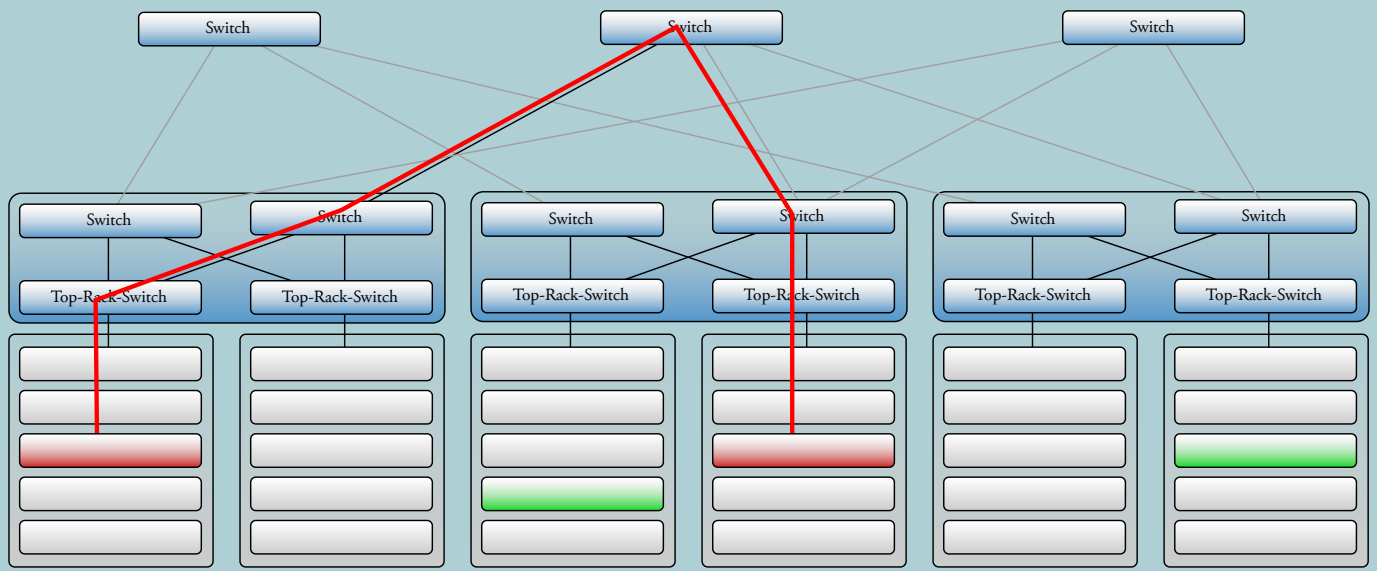
- Paths are chosen randomly
- Deploying an (unknown) hash function



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

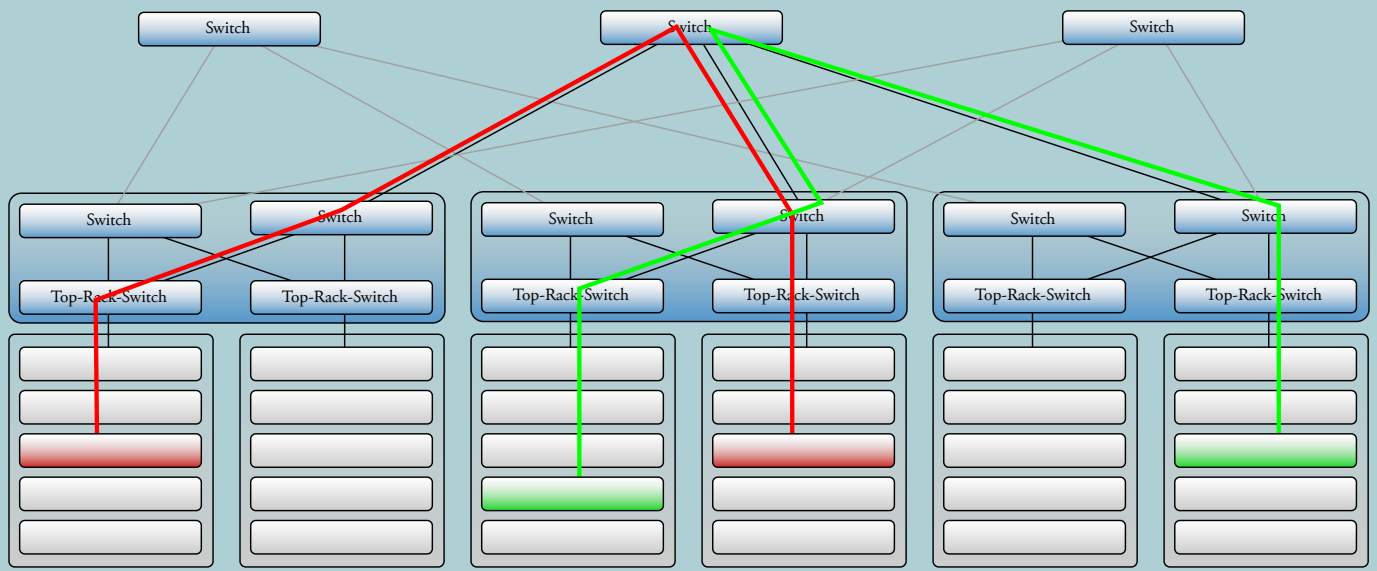
- Paths are chosen randomly
- Deploying an (unknown) hash function



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

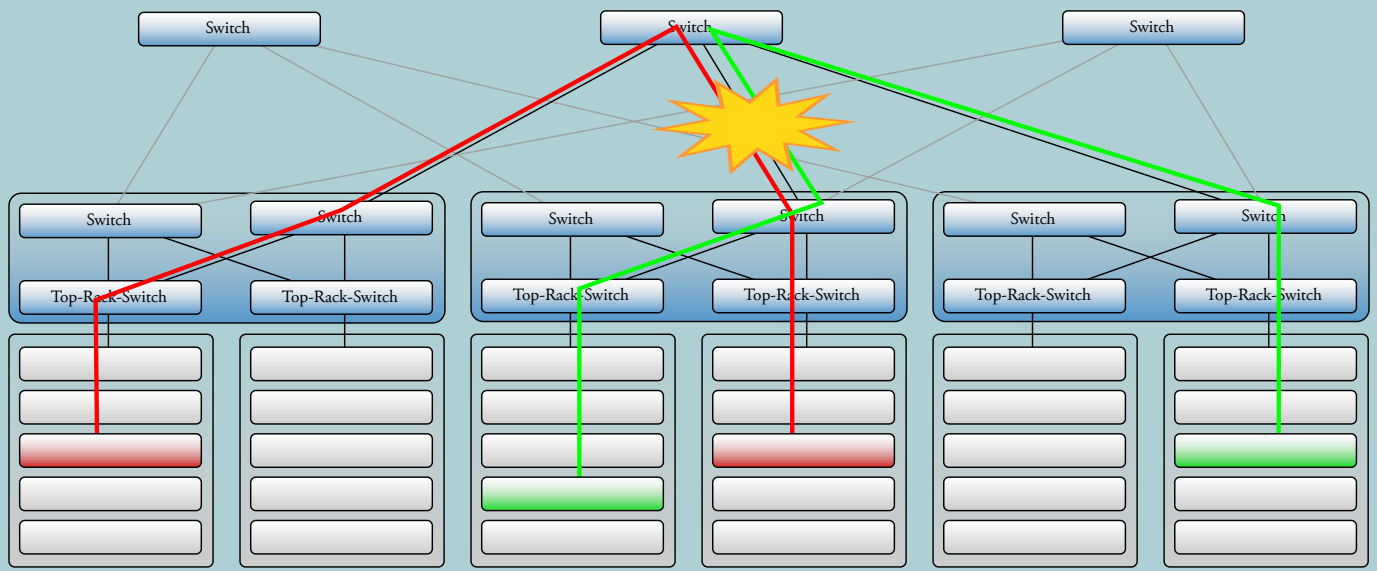
- Paths are chosen randomly
- Deploying an (unknown) hash function



Multipathing - Collisions in (Data Center) Networks

Multipathing based on ECMP

- Paths are chosen randomly
- Deploying an (unknown) hash function

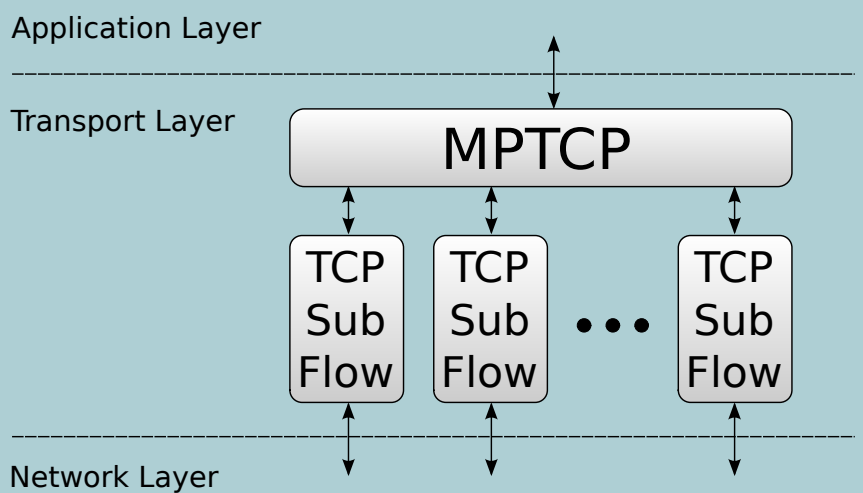


MultiPath TCP

MultiPath TCP - Design Objectives

MultiPath TCP (MPTCP) is an evolution of TCP that can effectively use multiple paths between a single transport connection. [3]

- It supports unmodified applications, since MPTCP looks like standard TCP.
- It works in today's networks.
- It is standardized at the IETF



MultiPath TCP - Connection Setup

MPTCP Connection Setup (simplified)

- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP

MultiPath TCP - Connection Setup

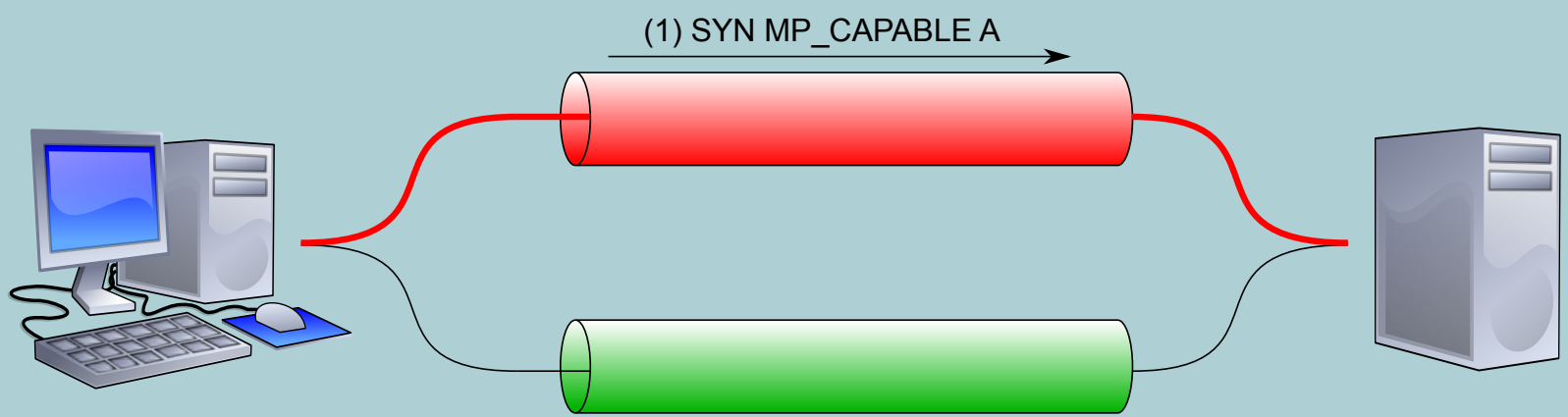
MPTCP Connection Setup (simplified)

- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP

MultiPath TCP - Connection Setup

MPTCP Connection Setup (simplified)

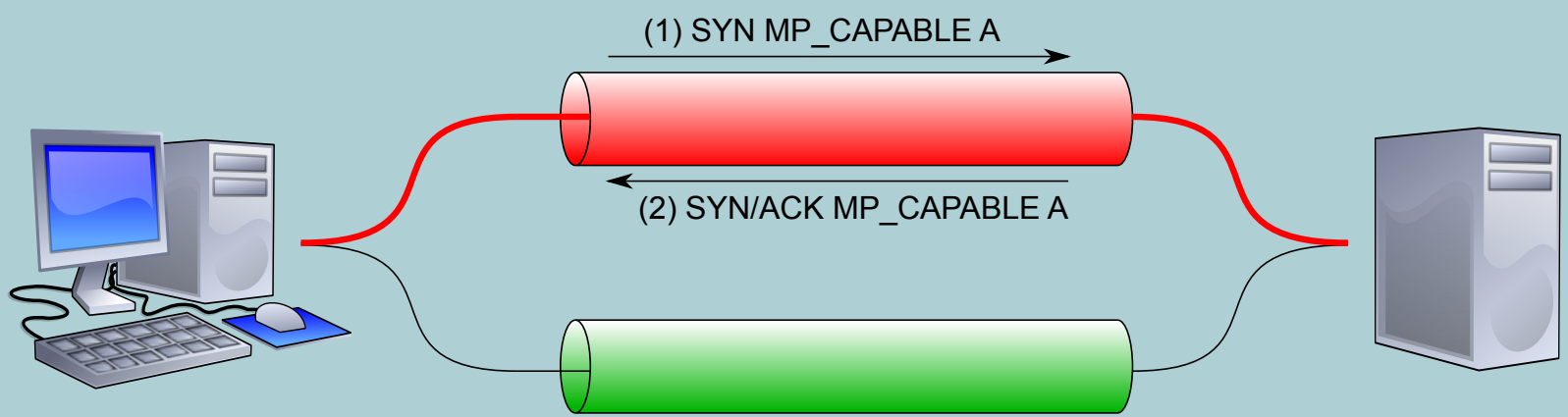
- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP



MultiPath TCP - Connection Setup

MPTCP Connection Setup (simplified)

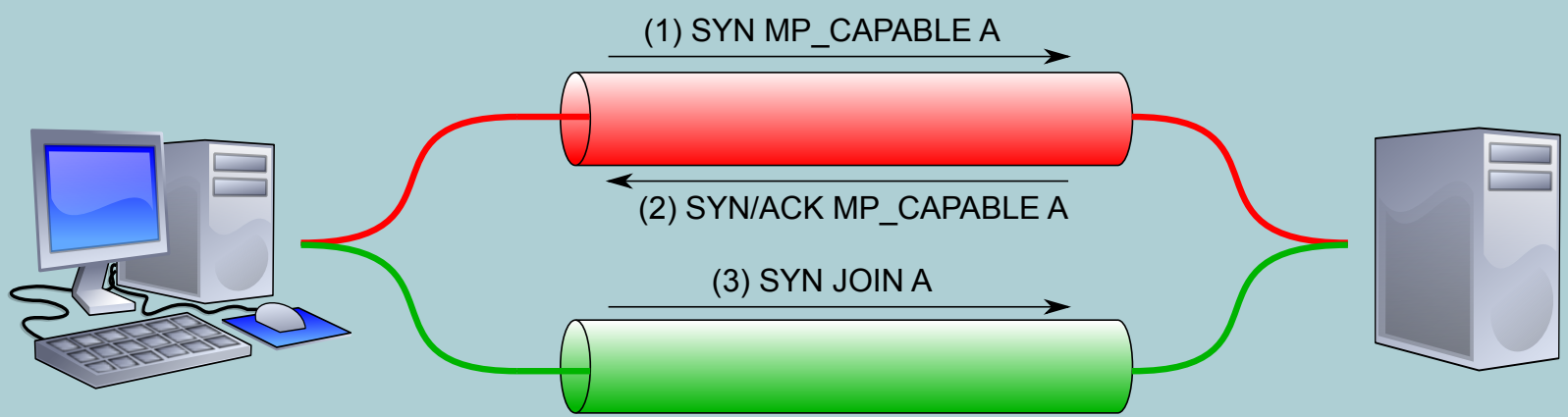
- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP



MultiPath TCP - Connection Setup

MPTCP Connection Setup (simplified)

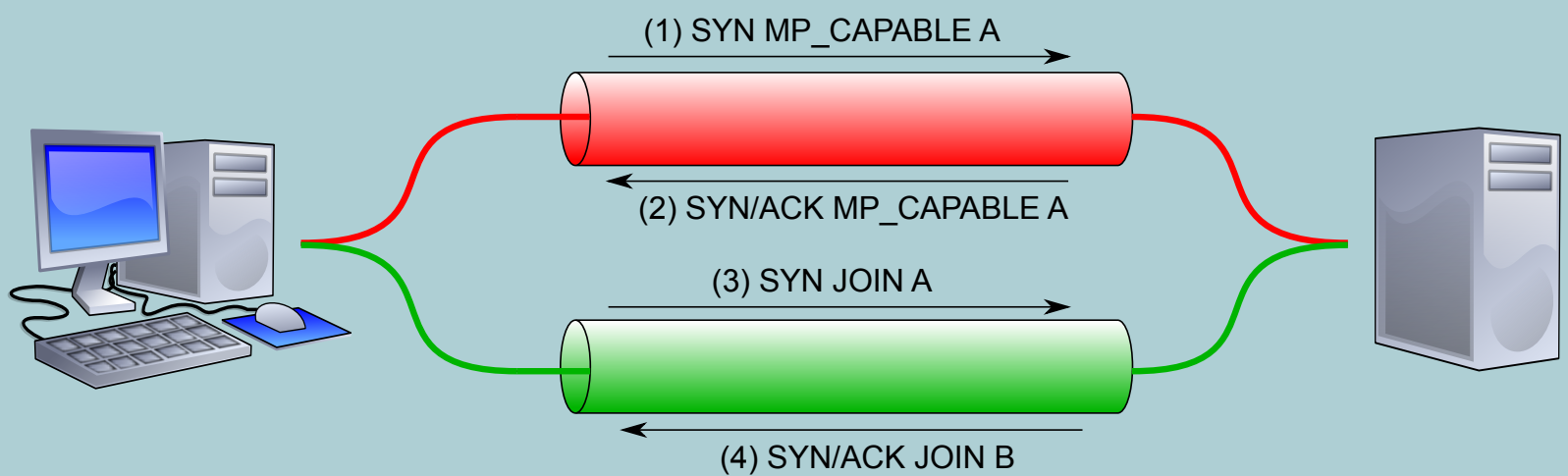
- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP



MultiPath TCP - Connection Setup

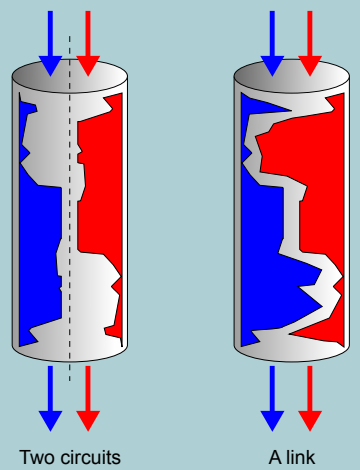
MPTCP Connection Setup (simplified)

- Deploying new TCP options to indicate MPTCP and to join subflows
- For subflows, the server keeps the same state variables as for regular TCP



A little bit of history:

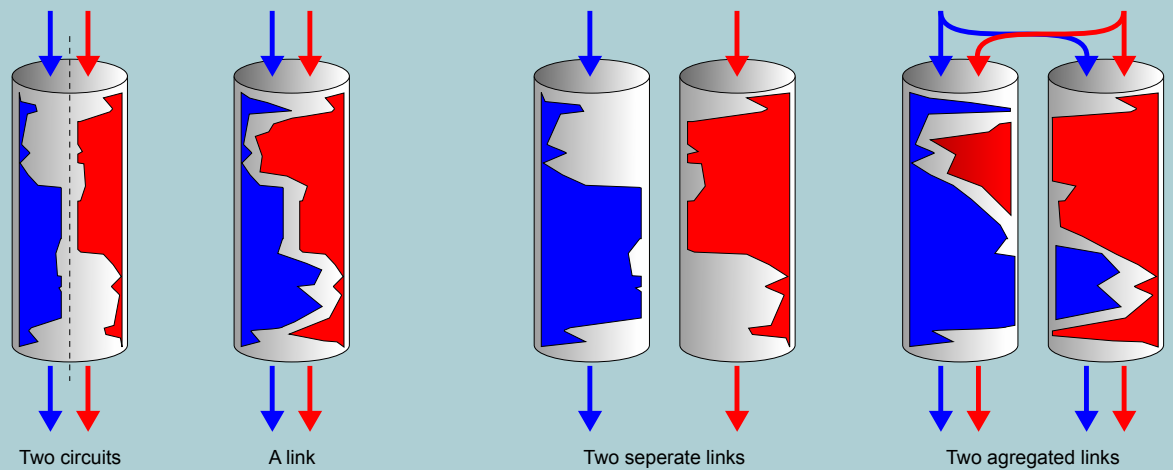
- Packet switching pools circuits



MultiPath TCP - Congestion Control

A little bit of history:

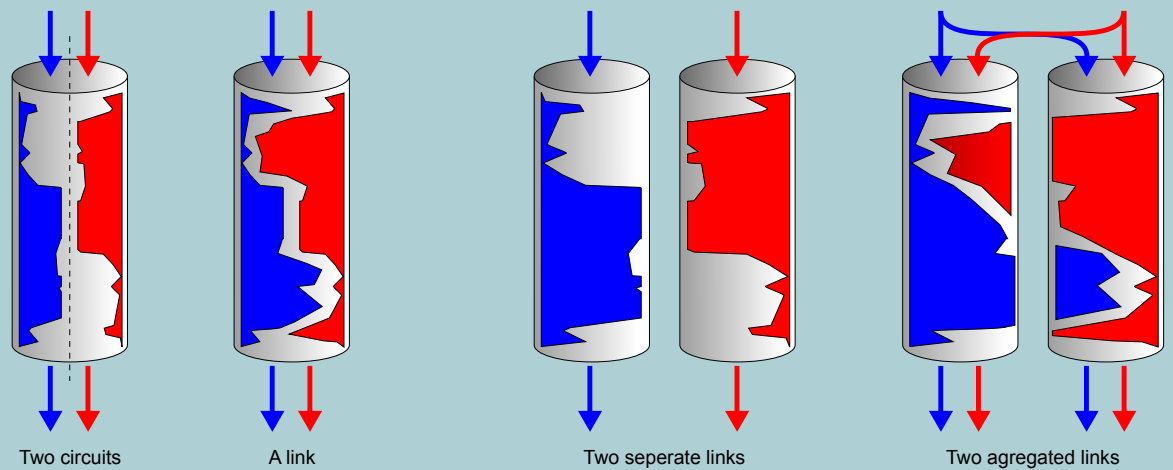
- Packet switching pools circuits
- Multipath pools links



MultiPath TCP - Congestion Control

A little bit of history:

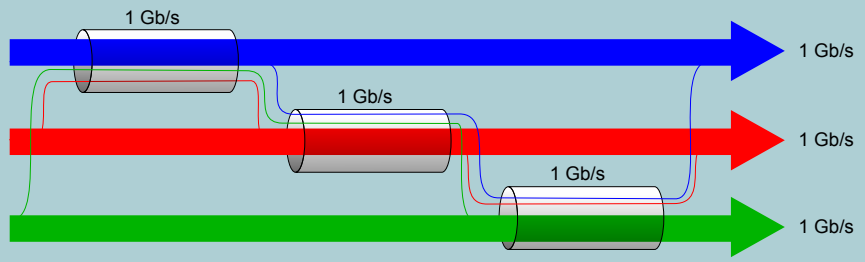
- Packet switching pools circuits
- Multipath pools links



- How should a link pool be shared?

MPTCP Congestion Control Design Goals

- MPTCP should be fair to regular TCP at shared links
  To this end, MPTCP should take as much capacity as regular TCP on a bottleneck link, no matter how may subflows are present.
- MPTCP should use efficient paths



- MPTCP should get at least as much throughput as TCP on the best path
  To this end, MPTCP should take congestion as well as RTTs into account

How does MPTCP congestion control work? (simplified)

- Maintain a congestion window w_r, for each subflow, where $r \in R$ ranges over the set of available paths.
- Increase w_r for each ACK on path r by

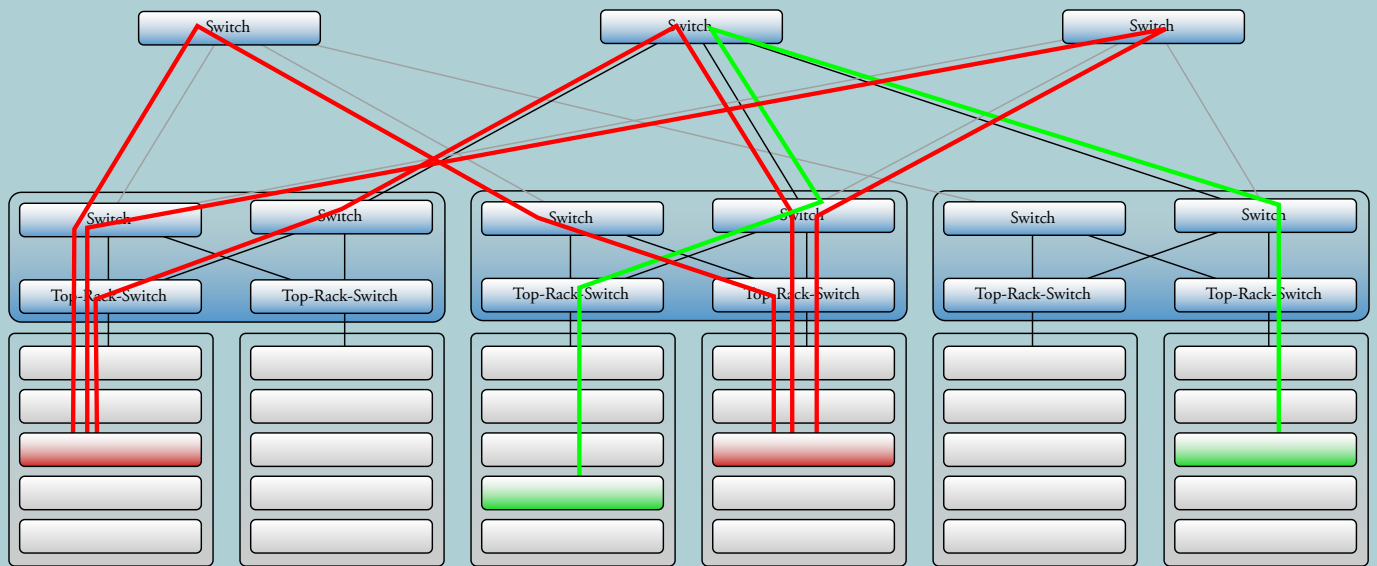
$$\frac{\alpha}{\sum_r w_r}$$

- Decrease w_r for each packet drop in subflow r by $w_r/2$

MultiPath TCP - Congestion Control

MPTCP ...

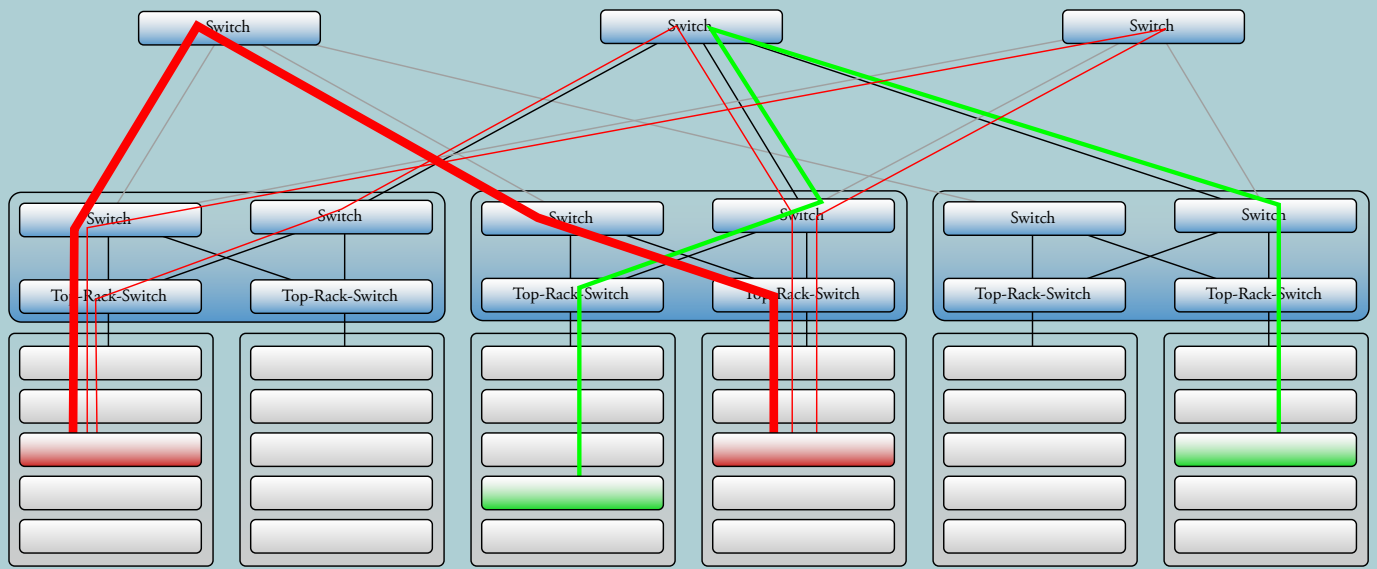
- uses all available paths
- moves data to least congested paths



MultiPath TCP - Congestion Control

MPTCP ...

- uses all available paths
- moves data to least congested paths



OpenFlow Link-Layer MultiPath Switching

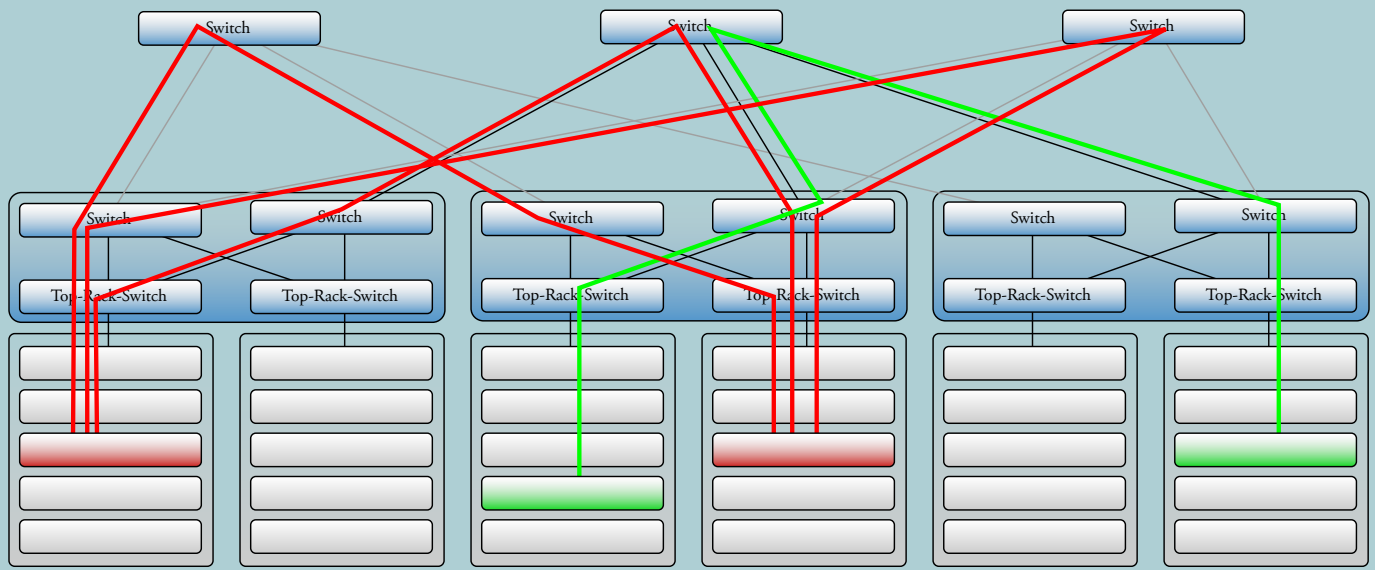
OLiMPS - OpenFlow Link-layer MultiPath Switching

- Addresses the problem of topology limitations in large-scale layer 2 networks
- Remove the necessity of a tree structure in the topology achieved though the use of Spanning Tree Protocol
- Allow for per-flow multipath switching and increase the robustness and efficiency of layer 2 network resources
- Integrate dynamic circuit provisioning systems like OSCARS and OpenFlow

OLiMPS - Use Case

Multipathing based on OpenFlow

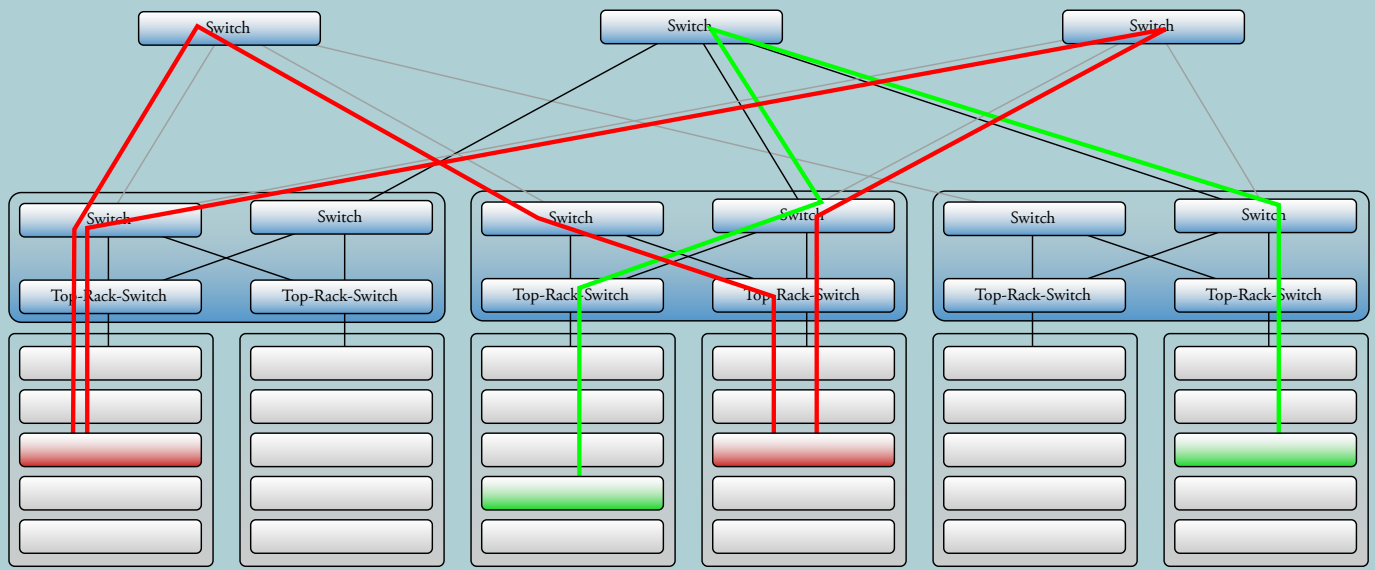
- Full control, thus, paths can be chosen deterministically
- Applicable to a variety of flow definitions.
- Works also for a small number of flows



OLiMPS - Use Case

Multipathing based on OpenFlow

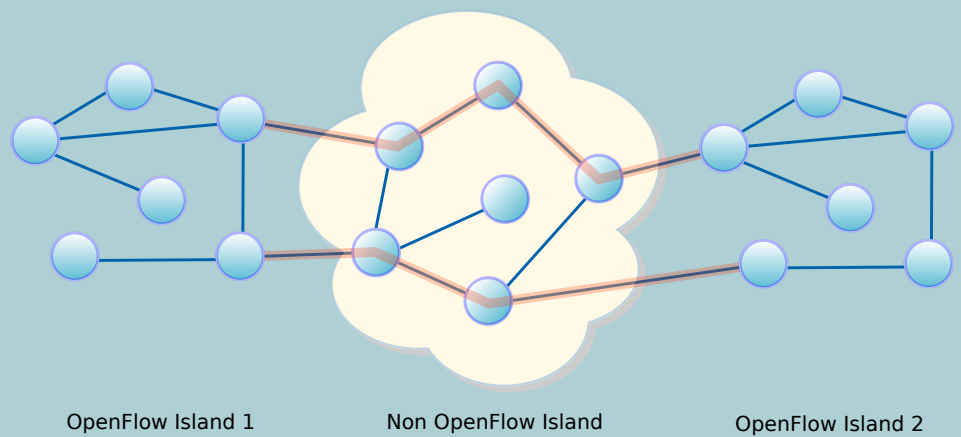
- Full control, thus, paths can be chosen deterministically
- Applicable to a variety of flow definitions.
- Works also for a small number of flows



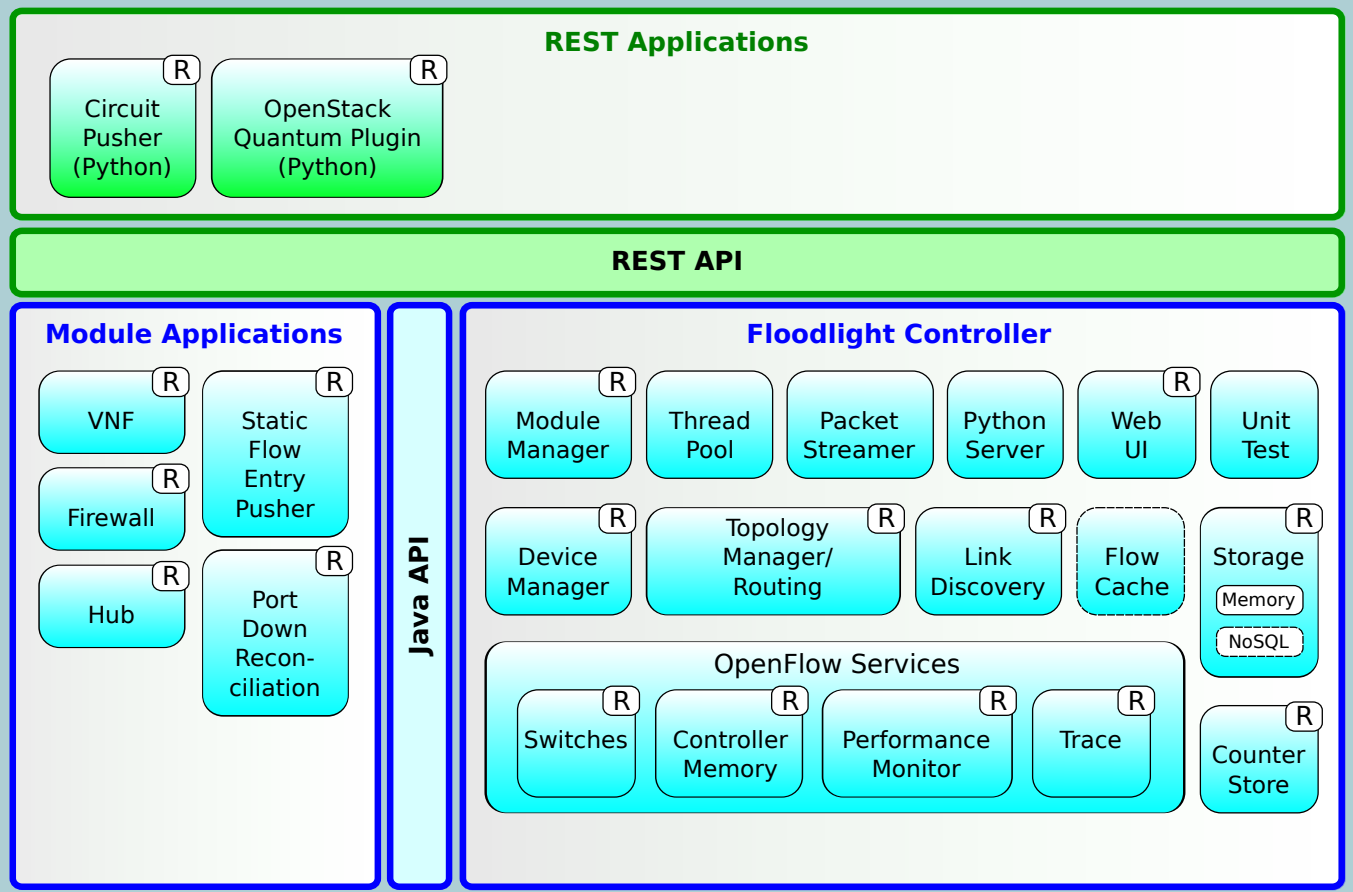
OLiMPS - OpenFlow Controller

OLiMPS OpenFlow Controller

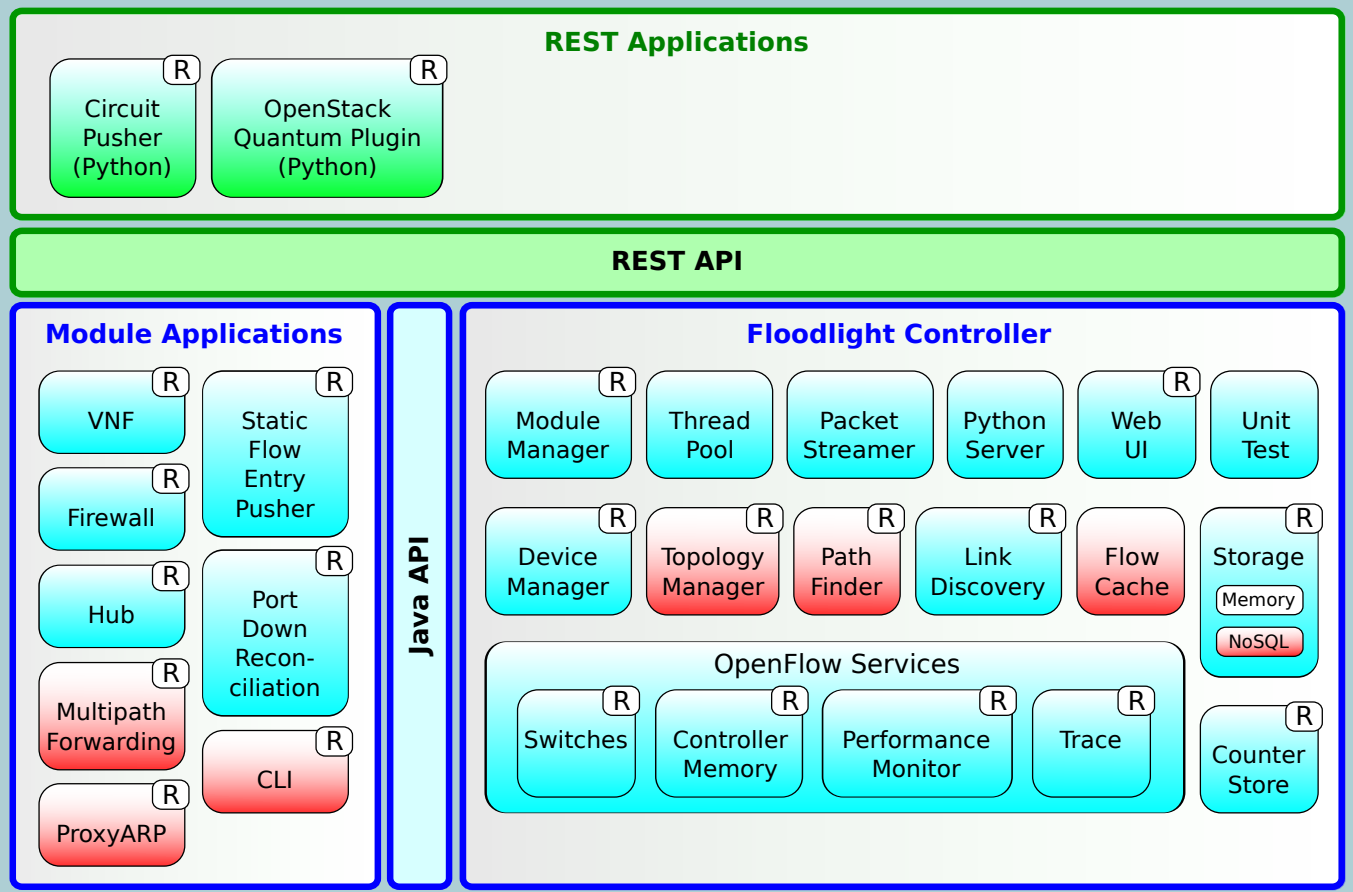
- Based on Floodlight [4]
  - Written in Java
  - Supports OpenFlow 1.0
- Implements a set of OpenFlow applications
  - ProxyARP
  - Pathfinder
  - Multipath Forwarding
- Allows for multiple paths between OpenFlow islands



Floodlight/OLiMPS controller architecture



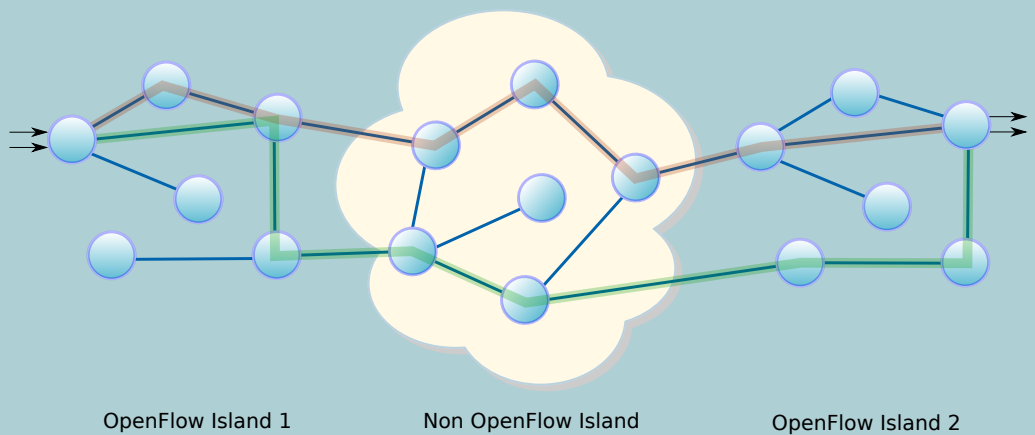
Floodlight/OLiMPS controller architecture



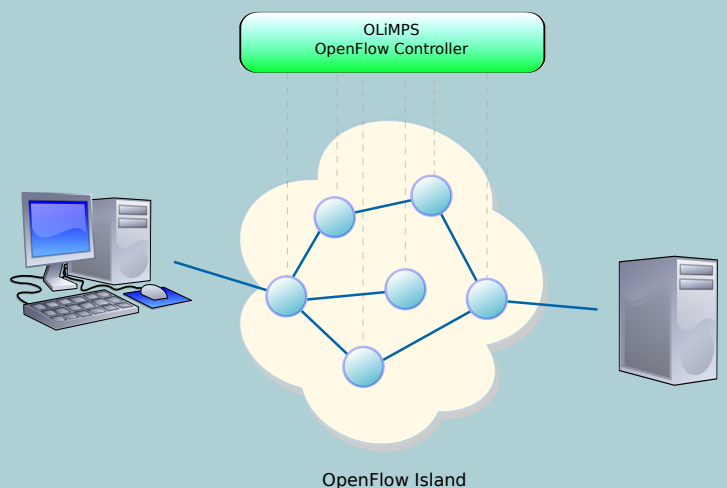
OLiMPS - OpenFlow Controller

OLiMPS Pathfinder and Multipath Forwarding

- Two modules (in contrast to the original Floodlight) implementing IRoutingService and extending ForwardingBase
- Calculate multiple link-disjoint paths from source to destination
- Per flow multi-pathing
- Reactive flow handling
  - New paths are calculated whenever a new flow appears at an edge switch
  - Flows are mapped to paths in a (capacity weighted) round robin manner
  - Flow rules are pushed to all switches of a paths

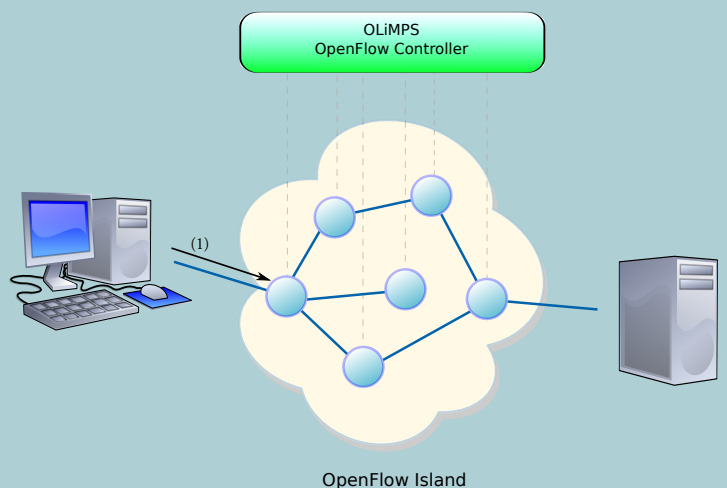


Path setup



Path setup

(1) First packet of a new flow arrives at OpenFlow switch

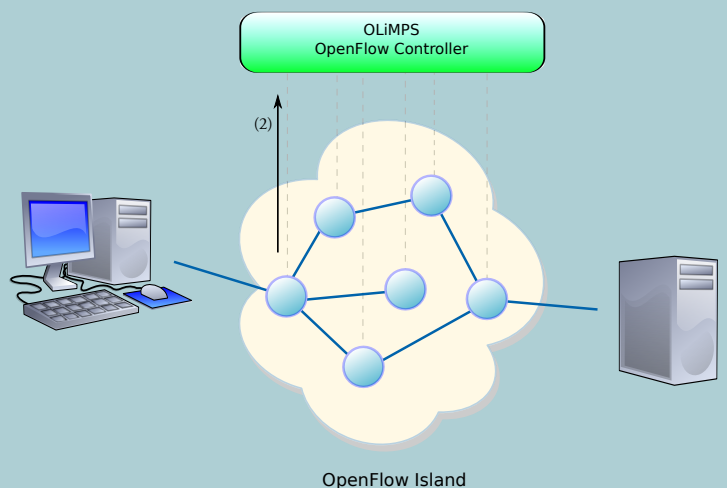


OLiMPS - OpenFlow Controller

Path setup

(1) First packet of a new flow arrives at OpenFlow switch

(2) Packet is forwarded to OpenFlow controller



OLiMPS - OpenFlow Controller

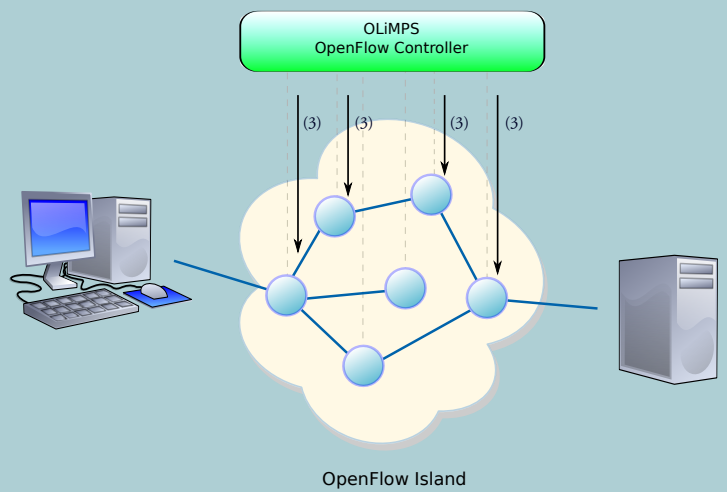
Path setup

(1) First packet of a new flow arrives at OpenFlow switch

(2) Packet is forwarded to OpenFlow controller

(3a) The controller calculates all paths between source and destination switch

(3b) The controller installs the flow mods for one path for the new flow



OLiMPS - OpenFlow Controller

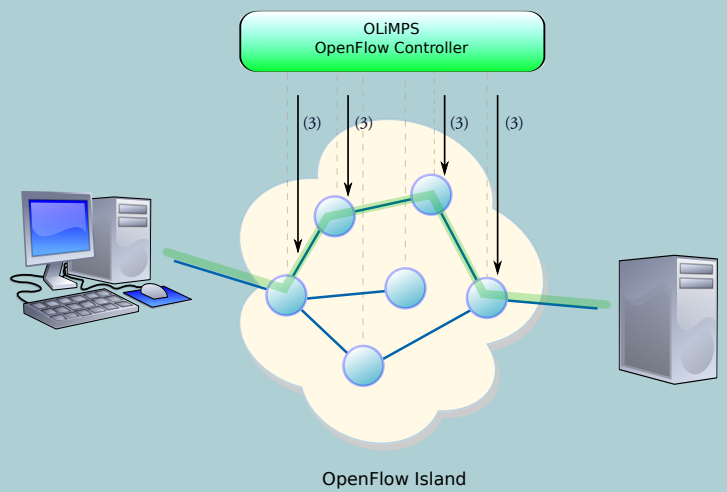
Path setup

(1) First packet of a new flow arrives at OpenFlow switch

(2) Packet is forwarded to OpenFlow controller

(3a) The controller calculates all paths between source and destination switch

(3b) The controller installs the flow mods for one path for the new flow



OLiMPS - OpenFlow Controller

Path setup

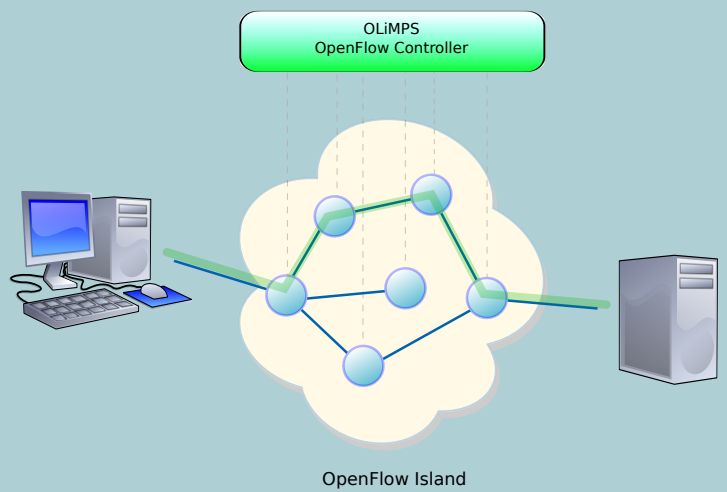
(1) First packet of a new flow arrives at OpenFlow switch

(2) Packet is forwarded to OpenFlow controller

(3a) The controller calculates all paths between source and destination switch

(3b) The controller installs the flow mods for one path for the new flow

(4) Packets are forwarded on the newly installed path



OLiMPS - OpenFlow Controller

Path setup

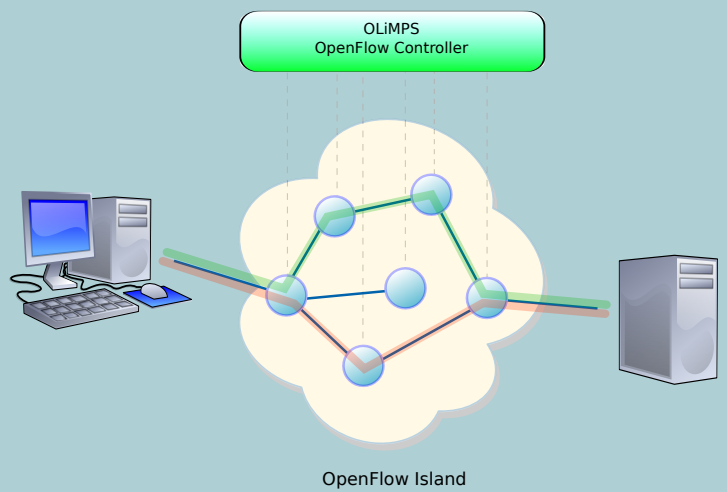
(1) First packet of a new flow arrives at OpenFlow switch

(2) Packet is forwarded to OpenFlow controller

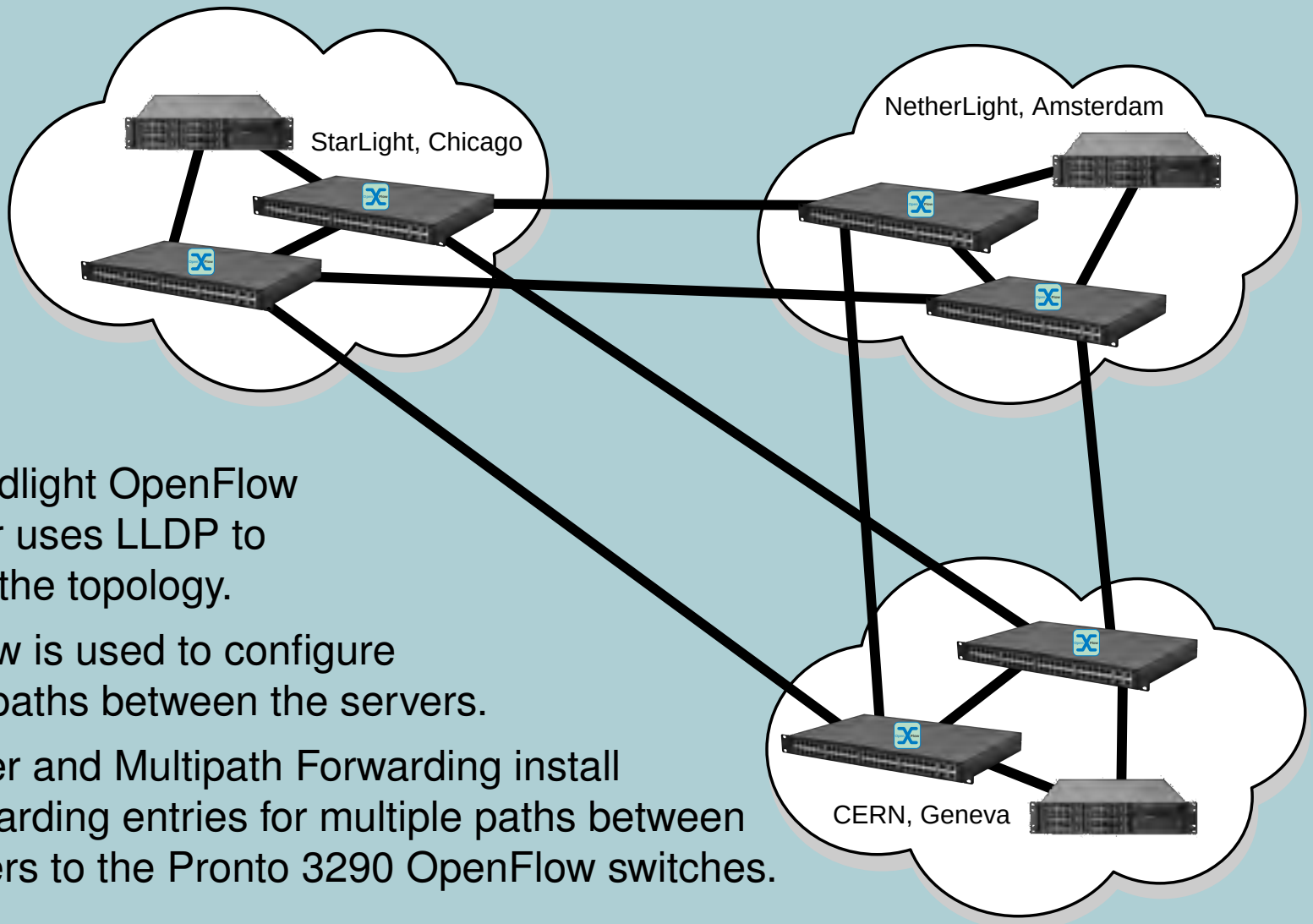
(3a) The controller calculates all paths between source and destination switch

(3b) The controller installs the flow mods for one path for the new flow

(4) Packets are forwarded on the newly installed path



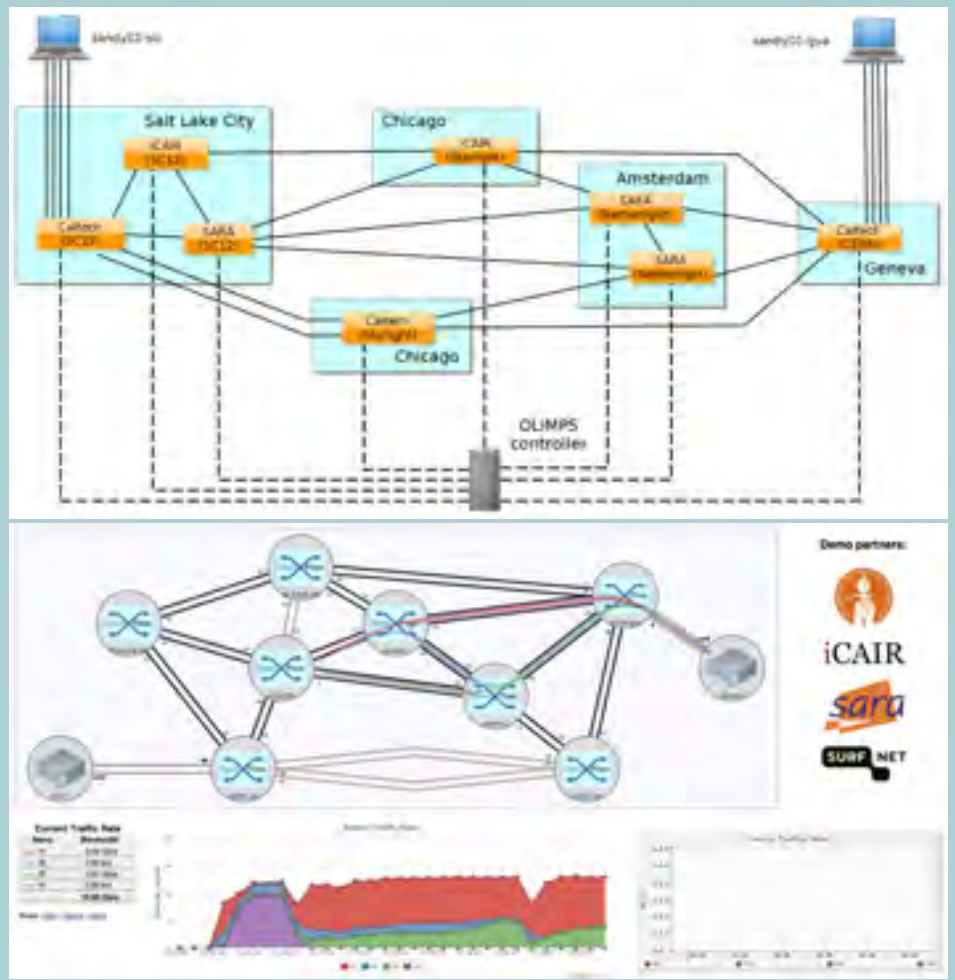
OLiMPS - International Multipath OpenFlow Network



- The Floodlight OpenFlow controller uses LLDP to discover the topology.
- OpenFlow is used to configure multiple paths between the servers.
- Pathfinder and Multipath Forwarding install flow forwarding entries for multiple paths between the servers to the Pronto 3290 OpenFlow switches.

OLiMPS - International Multipath OpenFlow Network

SuperComputing 2012: Streaming from GVA to CHI



OLiMPS - New ideas and next steps

OLiMPS Roadmap

- Implement intelligent path selection, e.g. based on measurements
- Implement in-network load balancing
- Integrate QoS policies, e.g. rate limits per path
- Extend the error handling, e.g. seamless flow redirection
- Move to OpenFlow version 1.2/1.3

Some open (research) questions remain

- Where to do traffic load balancing: In the end hosts or in the network?
- Is the system still stable or can it oscillate?
- What is the overall performance of such a system in terms of resource efficiency, throughput, fairness, etc.

Summary & Conclusion

MultiPath TCP

- ... is an evolution of TCP that uses multiple paths between a single transport connection
- ... supports unmodified applications and works in today's networks
- ... implementations work fine for moderate fast datacenter networks
- There is room for improvement on high speed networks, i.e. $\geq$ 10 Gb/s and WANs

OpenFlow Link-Layer MultiPath Switching

- ... removes some limitations in large-scale layer 2 networks
- ... allows for an effective calculation of multiple paths between source and destination
- There is room for improvement towards a production ready system

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat. *A Scalable, Commodity Data Center Network Architecture*, In Proc. of SIGCOMM 2008

[2] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu. *Bcube: A high Performance, Server-Centric Network Architecture for Modular Data Centers*, In Proc. of SIGCOMM 2009

[3] C. Raiciu and C. Paasch. *MultiPath TCP*, Google TechTalk, Apr. 2012

[4] BigSwitch. *Floodlight OpenFlow Controller*, http://floodlight.openflowhub.org

[5] A. Rizk and M. Fidler. *Sample Path Bounds for Long Memory FBM Traffic*, In Proc. of INFOCOM 2010