Improve Operations of Data Center Networks with Physical-Layer Programmability

Yiting Xia
Research Scientist, Facebook
Network is a static graph

Ring

Mesh

Star

Fully Connected

Line

Tree

Bus

Common Network Topologies
Network is a static graph

Common Network Topologies

Ring
Mesh
Star
Fully Connected
Line
Tree
Bus

Basic assumption of the networking world
Cloud Data Center Network

Google Data Center in London
Topology of Data Center Network

Clos (Multi-Rooted Tree) Topology
Network operation is hard
Network operation is hard

• Failures
Network operation is hard

• Failures

Google Cloud Outage Triggered By Networking Issue

Google's Tuesday afternoon outage brought down popular services, including Spotify and Snapchat.

By Gina Narcisi

July 17, 2018, 04:50 PM EDT

Google Cloud suffered an outage that slowed down or stopped several popular services on Tuesday afternoon, including Spotify and Snapchat.

Google confirmed via its cloud status dashboard that it became aware of a networking issue impacting its load balancers just after noon PT on Tuesday.
Network operation is hard

Google Cloud Outage Triggered

Google Cloud suffered an outage that slowed popular services on Tuesday afternoon, including

Google confirmed via its cloud status dashboard.

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Cloudflare and Google dealt with issues that affected countless sites and users on Tuesday.

By David Yaffe-Bellany

July 2, 2019

When a website won’t load, many internet users turn to DownDetector, a site that keeps track of online disruptions, providing frequent updates on the status of the world’s digital infrastructure.

But the site, which calls itself the “weatherman of the digital world,” was no help on Tuesday when thousands of major websites showed the same so-called 502 error message for part of the morning. In a twist, DownDetector had also gone down.
Network operation is hard

• Failures

Google Cloud Outage Triggered

Google’s Tuesday afternoon

By Gina Narcisi

Google Cloud suffered an outage involving popular services on Tuesday.

Google confirmed via its cloud networking issue impacting it.

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

We’re aware that some people are currently having trouble accessing the Facebook family of apps. We’re working to resolve the issue as soon as possible.

47K 9:49 AM - Mar 13, 2019

29.4K people are talking about this

showed the same so-called 502 error message for part of the morning. In a twist, DownDetector had also gone down.
Network operation is hard

• Failures

Google Cloud Outage Triggers

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Google's Tuesday afternoon

By Gina Narcisi

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft's commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.
Network operation is hard

• Failures

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft’s commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.
Network operation is hard

• Failures

Google Cloud Outage Trigger

Google's Tuesday afternoon
By Gina Narcisi

A Failure Here, Damaged Fiber There and a Day of Internet Glitches

Microsoft

Even Microsoft faced its share of cloud outages this year affecting Azure, Microsoft 365, Dynamics, and DevOps. In May, Microsoft had to face an outage that lasted for more than an hour showing network connectivity errors in Microsoft Azure that deeply affected its cloud services including Office 365, Microsoft Teams, Xbox Live, and several others which are widely used by Microsoft’s commercial customers. Engineers identified the root cause to be an incorrect name server delegation issue that affected DNS resolution, network connectivity, and downstream impact. While the services were recovered, no customer DNS records were impacted during this incident.

• Median case of failures: 10% less traffic delivered
• Worst 20% cases of failures: 40% less traffic delivered

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011
Network operation is hard

- Failures
  - Failures are disruptive
  - Fixed topology: have to live with a crippled network

- Median case of failures: 10% less traffic delivered
- Worst 20% cases of failures: 40% less traffic delivered

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011
Network operation is hard

- Service provisioning
Network operation is hard

• Service provisioning
  - *Public cloud*: VM clusters
  - *Private cloud*: sub-systems supporting the service
Network operation is hard

- Service provisioning
Network operation is hard

- Service provisioning

![Network traffic locality graphs](image-url)

Table 3 further breaks down the locality of traffic generated by all of Facebook's clusters. Each entry represents 1% of packets. Total traffic is 100%. The clear majority of traffic is intra-rack, with a significant amount of inter-datacenter traffic. The locality patterns vary significantly by system type.

### Table 3: Different clusters have different localities

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>Inter-DC</th>
<th>Intra-Datacenter</th>
<th>Intra-Cluster</th>
<th>Intra-Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop</td>
<td>17.7%</td>
<td>23.7%</td>
<td>57.5%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Web server</td>
<td>11.9%</td>
<td>18.0%</td>
<td>34.8%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Cache follower</td>
<td>12.9%</td>
<td>16.1%</td>
<td>34.5%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Cache leader</td>
<td>5.2%</td>
<td>10.2%</td>
<td>30.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td>DB</td>
<td>2.5%</td>
<td>5.2%</td>
<td>23.7%</td>
<td>7.3%</td>
</tr>
</tbody>
</table>

*Roy et al., Inside the Social Network’s (Datacenter) Network, SIGCOMM 2015*
Network operation is hard

- Service provisioning

Roy et al., Inside the Social Network’s (Datacenter) Network, SIGCOMM 2015
Network operation is hard

- **Service provisioning**

---

Table 3: Different clusters have different localities; last row

<table>
<thead>
<tr>
<th>Locality</th>
<th>Percentage</th>
<th>Hadoop</th>
<th>Web server</th>
<th>Cache follower</th>
<th>Cache leader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter-Datacenter</td>
<td>17.7</td>
<td>57.5</td>
<td>21.5</td>
<td>10.2</td>
<td>10.2</td>
</tr>
<tr>
<td>Intra-Datacenter</td>
<td>11.9</td>
<td>80.9</td>
<td>18.0</td>
<td>16.1</td>
<td>16.1</td>
</tr>
<tr>
<td>Intra-Cluster</td>
<td>57.5</td>
<td>17.7</td>
<td>5.2</td>
<td>40.7</td>
<td>40.7</td>
</tr>
<tr>
<td>Intra-Rack</td>
<td>2.5</td>
<td>2.5</td>
<td>15.9</td>
<td>15.7</td>
<td>15.7</td>
</tr>
</tbody>
</table>

---

Roy et al., *Inside the Social Network’s (Datacenter) Network*, SIGCOMM 2015
Network operation is hard

- Service provisioning

\[\begin{array}{c|c|c|c|c}
\text{Cluster Type} & \text{Inter-DC} & \text{Intra-Datacenter} & \text{Intra-Cluster} & \text{Intra-Rack} \\
\hline
\text{Hadoop} & \text{Green} & \text{Red} & \text{Blue} & \text{Turquoise} \\
\text{Web server} & \text{Green} & \text{Red} & \text{Blue} & \text{Turquoise} \\
\end{array}\]

- Different clusters have different traffic localities
- Hard to fit into the same network topology

Roy et al., Inside the Social Network’s (Datacenter) Network, SIGCOMM 2015
Network operation is hard

- Maintenance
Network operation is hard

- Maintenance

Drain traffic  Maintenance  Undrain traffic
Network operation is hard

• Maintenance

Drain traffic  Maintenance  Undrain traffic
Network operation is hard

- Maintenance

Drain traffic ➔ Maintenance ➔ Undrain traffic

Misconfiguration

Misconfiguration
Network operation is hard

- Maintenance

Drain traffic

Maintenance

Undrain traffic

- Important source of oncall problems
- Change network states to “fake” loss of connectivity
Network operation is hard

- Wiring
Network operation is hard

• Wiring

Facebook FRC Data Center
Network operation is hard

- Wiring
  - Time-consuming and error-prone
  - Rewiring inevitable: expansion, device upgrade
Network Operation  Topology Change
## Network Operation  Topology Change

<table>
<thead>
<tr>
<th>Operational Problems</th>
<th>Topology Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure</td>
<td>Bypass or fix failures</td>
</tr>
<tr>
<td>Service Provisioning</td>
<td>Change into the right topology</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Partition the graph</td>
</tr>
<tr>
<td>Wiring</td>
<td>Automatic wiring with software</td>
</tr>
</tbody>
</table>
Network Operation  Topology Change

- If fast enough, the change should be hidden from upper layers of the network stack

<table>
<thead>
<tr>
<th>Operational Problems</th>
<th>Topology Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure</td>
<td>Bypass or fix failures</td>
</tr>
<tr>
<td>Service Provisioning</td>
<td>Change into the right topology</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Partition the graph</td>
</tr>
<tr>
<td>Wiring</td>
<td>Automatic wiring with software</td>
</tr>
</tbody>
</table>
Physical-Layer Programmability
Physical-Layer Programmability

• The network topology is configurable
Physical-Layer Programmability

• The network topology is configurable

• Circuit switching
  - optical or wireless
  - reconfigure internal connections
Physical-Layer Programmability

• The network topology is configurable

• Circuit switching
  - optical or wireless
  - reconfigure internal connections
Physical-Layer Programmability

• The network topology is configurable

• Circuit switching
  - optical or wireless
  - reconfigure internal connections
Physical-Layer Programmability

• The network topology is configurable

• Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*

• Fast topology change
  - *ms or us*
Physical-Layer Programmability

- The network topology is configurable

- Circuit switching
  - *optical or wireless*
  - *reconfigure internal connections*

- Fast topology change
  - *ms or us*

- Controlled by software
High-Level Idea: New Network Model
High-Level Idea: New Network Model

Today's Data Center

Top-of-Rack (ToR) switch

Aggregation switch

Core switch
High-Level Idea: New Network Model

- **Core switch**
- **Aggregation switch**
- **Top-of-Rack (ToR) switch**
High-Level Idea: New Network Model

- Interleave Circuit Switch & Ethernet Switch
- Restructure the network == uncable + recable
High-Level Idea: New Network Model

- **Core switch**
- **Aggregation switch**
- **Circuit switch**
- **Top-of-Rack (ToR) switch**
High-Level Idea: New Network Model

- Small distributed circuit switches \(\rightarrow\) local change
- 600x cost reduction & scalability
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]
Failure Recovery

Flat-tree
[HotNets’16, SIGCOMM’17]
Service Provisioning

OmniSwitch
[HotCloud’15]
Wiring & Maintenance

Lighthouse
(In submission)
Physical-Layer Programmability in WAN
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]
Failure Recovery

Flat-tree
[HotNets’16, SIGCOMM’17]
Service Provisioning

OmniSwitch
[HotCloud’15]
Wiring & Maintenance

Lighthouse
(In submission)
Physical-Layer Programmability in WAN
Default Failure Recovery: Rerouting
Default Failure Recovery: Rerouting
Default Failure Recovery: Rerouting
Default Failure Recovery: Rerouting

- Fast local rerouting
  - *Inflated path length*
Default Failure Recovery: Rerouting

- Fast local rerouting
  - Inflated path length

- Global optimal rerouting
  - High latency
Default Failure Recovery: Rerouting

- Fast local rerouting
  - *Inflated path length*
- Global optimal rerouting
  - *High latency*
- Impact other flows
  - *Degraded performance*
Default Failure Recovery: Rerouting

- Fast local rerouting
  - Inflated path length
- Global optimal rerouting
  - High latency
- Impact other flows
  - Degraded performance

Restore bandwidth immediately!
Shareable Backup

No backup  1:1 backup
Shareable Backup

No backup  Different backup ratios  1:1 backup
Shareable Backup

No backup  Different backup ratios  1:1 backup

Circuit Switches  \(\rightarrow\)  a pool of backup switches
Shareable Backup

- No backup
- 1. Failures are rare
- 2. Failed switch replaced in ms
- Circuit Switches → a pool of backup switches

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011
Shareable Backup

1. Failures are rare
2. Failed switch replaced in ms

No backup

1:1 backup

Circuit Switches → a pool of backup switches

ShareBackup: Data Center with Shareable Backup

Gill et al., Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications, SIGCOMM 2011
Architecture Design
Architecture Design
Architecture Design

Edge switches

Backup switch

Circuit switches

Hosts
Architecture Design

edge switches

Backup switch

Hosts

Circuit switches
Architecture Design

Aggregation switches

Backup switch

Circuit switches

Edge switches

Backup switch
Architecture Design
Architecture Design

Aggregation switches

Backup switch

Circuit switches

Edge switches

Backup switch
Architecture Design
Architecture Design

Core switches

Aggregation switches

Circuit switches

Backup switch
Challenge: Fast Failure Recovery
Challenge: Fast Failure Recovery

• Distributed controllers
  - Configure circuit switches quickly
Challenge: Fast Failure Recovery

- Distributed controllers
  - *Configure circuit switches quickly*

- Live impersonation
  - *Backup switch as hot standby*
Challenge: Fast Failure Recovery

• Distributed controllers
  - *Configure circuit switches quickly*

• Live impersonation
  - *Backup switch as hot standby*

• Routing table in place
  - *Save the time of setting forwarding rules*
Live Impersonation

Routing Table of Every Edge Switch

<table>
<thead>
<tr>
<th>Routing Table 0</th>
<th>VLAN 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routing Table 1</td>
<td>VLAN 1</td>
</tr>
<tr>
<td>Routing Table 2</td>
<td>VLAN 2</td>
</tr>
</tbody>
</table>

Edge switches

Backup switch

Hosts
Live Impersonation

Routing Table of Every Edge Switch

- Routing Table 0  VLAN 0
- Routing Table 1  VLAN 1
- Routing Table 2  VLAN 2

Edge switches

Backup switch

Hosts
Live Impersonation

Routing Table of Every Edge Switch

- Routing Table 0, VLAN 0
- Routing Table 1, VLAN 1
- Routing Table 2, VLAN 2

Edge switches
- Backup switch

Hosts
Simulation

• Facebook prod traffic trace
• Microsoft prod failure distribution
• Near-zero slow down during failures
Testbed

• 24 hosts, 12 regular switches, 6 backup switches
• Hadoop & Spark applications
Testbed

- MapReduce Sort w/ 100GB data
Testbed

- MapReduce Sort w/ 100GB data
- Same as the no-failure case
Testbed

- MapReduce Sort w/ 100GB data
- Same as the no-failure case
Testbed

- MapReduce Sort w/ 100GB data
- Same as the no-failure case
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]
Failure Recovery

Flat-tree
[HotNets’16, SIGCOMM’17]
Service Provisioning

OmniSwitch
[HotCloud’15]
Wiring & Maintenance

Lighthouse
(In submission)
Physical-Layer Programmability in WAN
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]

Flat-tree
[HotNets’16, SIGCOMM’17]

Omniswitch
[HotCloud’15]

Failure
Recovery

Service
Provisioning

Wiring &
Maintenance

Lighthouse
(In submission)

Physical-Layer Programmability in WAN
Different Topologies Needed

• Public cloud
  - VM clusters have different traffic characteristics
  - Cloud providers should meet SLA
Different Topologies Needed

• Public cloud
  - VM clusters have different traffic characteristics
  - Cloud providers should meet SLA

• Private cloud
  - Sub-systems of the service create different clustering features
  - Content providers should ensure service availability
Different Topologies Needed

• Public cloud
  - VM clusters have different traffic characteristics
  - Cloud providers should meet SLA

• Private cloud
  - Sub-systems of the service create different clustering features
  - Content providers should ensure service availability

• Network vulnerable during service provisioning
  - Utilization increases
Clos Topology

Core switch

Aggregation switch

Edge switch
Clos Topology

• Implementation friendly
  - Central wiring
  - Flexible scale and oversubscription
  - Pod modular design
Clos Topology

- Implementation friendly
  - Central wiring
  - Flexible scale and oversubscription
  - Pod modular design

- Good rack-level performance
  - Affluent intra-rack bandwidth
  - Congested network core
Random Graph

[Jellyfish NSDI’12]
Random Graph

• Good connectivity
  - Low average path length
  - Rich bandwidth
  - Near optimal throughput for uniform traffic

[Jellyfish NSDI’12]
Random Graph

• Good connectivity
  - Low average path length
  - Rich bandwidth
  - Near optimal throughput for uniform traffic

• Hard to implement
  - Neighbor-to-neighbor wiring complicated

[Jellyfish NSDI’12]
Flat-tree

Tree Network vs. Flat Network
Flat-tree

Tree Network vs. Flat Network

Easy implementation vs. Good connectivity
Flat-tree

Tree Network vs. Flat Network

Easy implementation
Clustered traffic

Good connectivity
Uniform traffic
Flat-tree
Flat-tree

- Start from Clos
- Flatten tree structure
- Approximate random graphs
Flatten the Tree

• How to flatten the tree structure?
## Flatten the Tree

- How to flatten the tree structure?

<table>
<thead>
<tr>
<th>Difference</th>
<th>Clos</th>
<th>Random graph</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server distribution</td>
<td>Edge switches</td>
<td>All switches</td>
<td>Relocate servers</td>
</tr>
<tr>
<td>Wiring</td>
<td>Central</td>
<td>Neighbor-to-neighbor</td>
<td>Diversify connections</td>
</tr>
</tbody>
</table>
Circuit Switch Configurations

C: core switch
A: aggregation switch
E: edge switch
S: server

6-port Circuit Switch
Circuit Switch Configurations

C: core switch
A: aggregation switch
E: edge switch
S: server

6-port Circuit Switch
Flat-tree Example

- Core Switch
- Aggregation Switch
- Edge Switch
- Server

Clos Pod
Flat-tree Example

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server

Flat-tree Pod
Clos Network

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Approximate Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Approximate Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Approximate Random Graph

Core Switch
Aggregation Switch
Edge Switch
Circuit Switch
Server
Approximate Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Approximate Local Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Approximate Local Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Server
- Circuit Switch
Approximate Local Random Graph

- Core Switch
- Aggregation Switch
- Edge Switch
- Circuit Switch
- Server
Routing Challenges

• Server mobility
  - Server relocated to different switches
  - Assign IP addresses to support prefix matching
Routing Challenges

• Server mobility
  - Server relocated to different switches
  - Assign IP addresses to support prefix matching

• Different routing schemes per topology
  - Clos: ECMP, two-level routing, SDN
  - Random graph: k-shortest-path routing + MPTCP
Routing Challenges

• Server mobility
  - *Server relocated to different switches*
  - *Assign IP addresses to support prefix matching*

• Different routing schemes per topology
  - *Clos: ECMP, two-level routing, SDN*
  - *Random graph: k-shortest-path routing + MPTCP*

• k-shortest-path routing
  - *k paths for every server pairs*
  - *Enormous number of states → exceed switch capacity*
  - *No solution from random graph networks*
Routing Challenges

• Server mobility
  - Customized addressing scheme
  - Different sets of IP addresses per topology

• Different routing schemes per topology
  - Clos: ECMP, two-level routing, SDN
  - Random graph: k-shortest-path routing + MPTCP

• k-shortest-path routing
  - k paths for every server pairs
  - Enormous number of states $\rightarrow$ exceed switch capacity
  - No solution from random graph networks
Routing Challenges

• Server mobility
  • Customized addressing scheme
  • Different sets of IP addresses per topology

• Different routing schemes per topology
  • k-shortest-path routing for all topologies
  • ECMP/two-level routing/SDN encoded as k paths

• k-shortest-path routing
  - *k paths for every server pairs*
  - *Enormous number of states → exceed switch capacity*
  - *No solution from random graph networks*
Routing Challenges

• Server mobility
  - Customized addressing scheme
  - Different sets of IP addresses per topology

• Different routing schemes per topology
  - k-shortest-path routing for all topologies
  - ECMP/two-level routing/SDN encoded as k paths

• k-shortest-path routing
  - Addressing: server-level $\rightarrow$ switch-level k paths
  - Source routing: further reduce network states

- No solution from random graph networks
Transmission Performance

• Packet-level simulation

• Traffic traces from 4 Facebook data centers
  - Hadoop-1: no locality
  - Hadoop-2: rack-level locality
  - Web: Pod-level locality
  - Cache: Pod-level locality
Network-wide Traffic

- Hadoop-1: no locality
Rack-level Locality

- Hadoop-2: rack-level locality

![Graph showing CDF of flow completion time for different modes]
Pod-level Locality

- Web: Pod-level locality

![Graph showing CDF of Flow completion time for different modes](image)
Pod-level Locality

- Cache: Pod-level locality

![Graph showing CDF of Flow completion time (ms) for Global mode, Local mode, and Clos mode.]

CDF

Flow completion time (ms)
Testbed

- Implementation of motivating example
  - Hadoop & Spark
  - 27.6% more bandwidth
  - 10% less data read time
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]
Failure Recovery

Flat-tree
[HotNets’16, SIGCOMM’17]
Service Provisioning

OmniSwitch
[HotCloud’15]
Wiring & Maintenance

Lighthouse
(In submission)
Physical-Layer Programmability in WAN
OmniSwitch

• Universal Building Block
OmniSwitch

- Universal Building Block

![Diagram of OmniSwitch Panel]

- 8 Multilink Connectors
- 16 x 25G Multilink
- Front View of OmniSwitch Panel

![Diagram of OmniSwitch Stack]

- Ethernet Switch
- Circuit Switch
- Front Panel

Universal Building Block
OmniSwitch

• Universal Building Block
Automatic Wiring
Automatic Wiring

4 Ethernet Switches (128 ports each)
Automatic Wiring

4 Ethernet Switches (128 ports each)
Automatic Wiring

4 Ethernet Switches (128 ports each)

32 Multilink Connectors
(16 individual links each)
Automatic Wiring

4 Ethernet Switches (128 ports each)

Wiring Software

32 Multilink Connectors
(16 individual links each)
Automatic Wiring

4 Ethernet Switches (128 ports each)

Wiring Software

32 Multilink Connectors
(16 individual links each)
Easy Maintenance

4 Ethernet Switches (128 ports each)

32 Multilink Connectors
(16 individual links each)
Easy Maintenance

4 Ethernet Switches (128 ports each)

Wiring Software

32 Multilink Connectors
(16 individual links each)
Easy Maintenance

4 Ethernet Switches (128 ports each)

32 Multilink Connectors (16 individual links each)

Wiring Software
Easy Maintenance

4 Ethernet Switches (128 ports each)

32 Multilink Connectors
(16 individual links each)

Wiring Software
Outline

ShareBackup
[HotNets’17, SIGCOMM’18]
Failure Recovery

Flat-tree
[HotNets’16, SIGCOMM’17]
Service Provisioning

Omniswitch
[HotCloud’15]
Wiring & Maintenance

Lighthouse
(In submission)
Physical-Layer Programmability in WAN
Outline

- **ShareBackup**
  - [HotNets'17, SIGCOMM'18]
  - Failure Recovery

- **Flat-tree**
  - [HotNets'16, SIGCOMM'17]
  - Service Provisioning

- **OmniSwitch**
  - [HotCloud’15]
  - Wiring & Maintenance

- **Lighthouse**
  - (In submission)
  - Physical-Layer Programmability in WAN
Large-scale study of an operational WAN

- 50 optical cross connects
- 100 fiber segments
- 1000 Amplifiers

Optical Cross Connect (ROADM)

Amplifier

Fiber

Wide Area Network (WAN)

Level 3’s North America Internet Backbone
Fast Wavelength Shifting

- Wavelength shifting is slow: ~10min
- Model power profile with Virtual Amplifier
Testbed Demo

• Wavelength shifting in 8 seconds

20+ engineers, 1 month time
## Summary

- Physical-layer programmability for network operation
- Four example architectures

<table>
<thead>
<tr>
<th>Design purpose</th>
<th>ShareBackup</th>
<th>Flat-tree</th>
<th>OmniSwitch</th>
<th>Lighthouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuition</td>
<td>Shareable backup</td>
<td>Topology conversion</td>
<td>Universal building block</td>
<td>Wavelength shifting</td>
</tr>
</tbody>
</table>
| Key ideas      | Failure group | 1. Server mobility  
2. Link diversification | Wiring software | Model power profile |
Social Impact
Social Impact

• Connect the world
  - 1st week of social distancing: 15% increase of FB utilization
  - Reliability and availability at highest priority
  - More oncall efforts to guard our infrastructures
Social Impact

• Connect the world
  - 1st week of social distancing: 15% increase of FB utilization
  - Reliability and availability at highest priority
  - More oncall efforts to guard our infrastructures

• Simplify & automate network management
Social Impact

• Connect the world
  - 1st week of social distancing: 15% increase of FB utilization
  - Reliability and availability at highest priority
  - More oncall efforts to guard our infrastructures

• Simplify & automate network management

• Provide high-quality service
  - Netflix and Amazon ceased HD content streaming
  - Zhihu (Chinese Quora) down for overload
Social Impact

• Connect the world
  - 1st week of social distancing: 15% increase of FB utilization
  - Reliability and availability at highest priority
  - More oncall efforts to guard our infrastructures

• Simplify & automate network management

• Provide high-quality service
  - Netflix and Amazon ceased HD content streaming
  - Zhihu (Chinese Quora) down for overload

• Make elastic capacity of hardware possible
Bright Future: Truly Flexible Cloud

Virtualization
Bright Future: Truly Flexible Cloud
Bright Future: Truly Flexible Cloud
Bright Future: Truly Flexible Cloud
Bright Future: Truly Flexible Cloud
Direction 1: Network Verification
Direction 1: Network Verification

Validate connectivity
Direction 1: Network Verification

Validate routing

Validate connectivity
Direction 1: Network Verification

- Validate routing
- Validate connectivity
Direction 1: Network Verification

Validate routing

Validate connectivity

Inconsistency
Direction 2: End-to-End / Cross-Layer Programmability

Internet

Edge Network

Backbone

Backbone Network

Data Center Network
Direction 2: End-to-End / Cross-Layer Programmability

- Internet
- Edge Network
- Backbone
- Backbone Network
- Data Center Network

Programmable
Direction 2: End-to-End / Cross-Layer Programmability

Internet

Edge Network

Backbone Network

Data Center Network

Programmable

Programmable
Direction 2: End-to-End / Cross-Layer Programmability
Direction 2: End-to-End / Cross-Layer Programmability

Smart NIC

Internet

Edge Network

Backbone

Backbone Network

Data Center Network

Smart NIC
Direction 2: End-to-End / Cross-Layer Programmability

- SDN
- P4
- Physical-Layer Programmability

Smart NIC

Internet

Backbone

Edge Network

Backbone Network

Data Center Network
Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology
Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Fit network topology to traffic
Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology  

Fit network topology to traffic  

Formulation
Direction 3: Joint Optimization of Traffic and Network Topology

Fit traffic to network topology

Formulation

Optimization

Fit network topology to traffic