High Throughput Data Center Topology Design

Ankit Singla, P. Brighten Godfrey, Alexandra Kolla
“How long must we wait until our *pigeon system* rivals those of the Continental Powers?”

- The Nineteenth Century, 1899
The need for throughput

Bandwidth Consumption

- Ingress (from Internet)
- Egress (to Internet)
- Inter-cluster

Inter-cluster traffic more than doubled in the last 7 months (and is accelerating)

March, 2011

[Facebook, via Wired]

May, 2012
Many topology options ...
How do we design throughput optimal network topologies?
How do we design throughput optimal network topologies?
How close can we get to optimal network capacity?
1. How close can we get to optimal network capacity?

2. How do we handle heterogeneity?
Jellyfish: Networking Data Centers Randomly

[NSDI 2012: Singla, Hong, Popa, Godfrey]
Jellyfish: Networking Data Centers Randomly

[NSDI 2012: Singla, Hong, Popa, Godfrey]
Jellyfish: Networking Data Centers Randomly

- High capacity
- Beat fat-trees by 25%+

[NSDI 2012: Singla, Hong, Popa, Godfrey]
Jellyfish: Networking Data Centers Randomly

- High capacity
  - Beat fat-trees by 25%+

- Easier to expand
  - 60% cheaper expansion

[NSDI 2012: Singla, Hong, Popa, Godfrey]
Jellyfish: Networking Data Centers Randomly

- High capacity
  - Beat fat-trees by 25%+

- Easier to expand
  - 60% cheaper expansion

- Routing and cabling are solvable problems

[NSDI 2012: Singla, Hong, Popa, Godfrey]
How close can we get to optimal network capacity?
1. How close can we get to optimal network capacity?

2. How do we handle heterogeneity?
How do we measure throughput?
How do we measure throughput?

Maximize the minimum flow
How do we measure throughput?

Maximize the minimum flow
under random permutation traffic
How do we measure throughput?

Maximize the minimum flow under random permutation traffic

Measuring and Understanding Throughput of Network Topologies
Sangeetha AJ, A. Singla, P. B. Godfrey, A. Kolla
SIGMETRICS 2014 (Short paper)
http://arxiv.org/pdf/1402.2531
How do we measure throughput?

Measuring and Understanding Throughput of Network Topologies

Sangeetha AJ, A. Singla, P. B. Godfrey, A. Kolla
SIGMETRICS 2014 (Short paper)
http://arxiv.org/pdf/1402.2531
How do we measure throughput?

• Bisection bandwidth ≠ throughput
How do we measure throughput?

- Bisection bandwidth ≠ throughput
- Near-worst case traffic patterns
How close can we get to optimal network capacity?
A simple upper bound
A simple upper bound

# flows
A simple upper bound

# flows • capacity used per flow
A simple upper bound

# flows \cdot capacity used per flow

\leq \text{total capacity}
A simple upper bound

\[ \# \text{ flows} \cdot \text{capacity used per flow} \leq \text{total capacity} \]
A simple upper bound

\# flows \cdot \text{throughput per flow} \cdot \text{mean path length} \leq \text{total capacity}
A simple upper bound

\[
\text{throughput per flow} \leq \frac{\text{total capacity}}{\# \text{ flows} \cdot \text{mean path length}}
\]
A simple upper bound

\[
\text{throughput per flow} \leq \frac{\sum_{\text{links}} \text{capacity}(\text{link})}{\# \text{flows} \cdot \text{mean path length}}
\]
A simple upper bound

\[
\text{throughput per flow} \leq \frac{\sum_{\text{links}} \text{capacity}(\text{link})}{\# \text{flows} \cdot \text{mean path length}}
\]
Lower bound on mean path length
Lower bound on mean path length
Lower bound on mean path length

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Lower bound on mean path length

<table>
<thead>
<tr>
<th>Distance</th>
<th># Nodes</th>
</tr>
</thead>
</table>

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Lower bound on mean path length

<table>
<thead>
<tr>
<th>Distance</th>
<th># Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Lower bound on mean path length

Distance | # Nodes
---------|---------
1        | 6
2        | 6

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Lower bound on mean path length

<table>
<thead>
<tr>
<th>Distance</th>
<th># Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

(Ugliness omitted)

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Lower bound on mean path length

<table>
<thead>
<tr>
<th>Distance</th>
<th># Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

(Ugliness omitted)

[Cerf et al., “A lower bound on the average shortest path length in regular graphs”, 1974]
Random graphs vs. bound

Throughput (Ratio to Upper-bound)

Network Size
Random graphs vs. bound

Throughput (Ratio to Upper-bound)

Network Size

5 servers per switch, random permutation traffic
Random graphs vs. bound

Throughput (Ratio to Upper-bound)

Network Size

all-to-all

10 servers

5 servers per switch, random permutation traffic
Random graphs vs. bound

Random graphs within a few percent of optimal!
Random graphs vs. bound

Measuring and Understanding Throughput of Network Topologies
Sangeetha AJ, A. Singla, P. B. Godfrey, A. Kolla
SIGMETRICS 2014 (Short paper)
http://arxiv.org/pdf/1402.2531

Random graphs within a few percent of optimal!
Random graphs exceed throughput of other topologies
How close can we get to optimal network capacity?

Very close!!
How do we handle heterogeneity?

Image credit: Legolizer (www.drububu.com)
Heterogeneity
Heterogeneity
Random graphs as a building block
Random graphs as a building block

- Low-degree switches
- High-degree switches
- Servers
Random graphs as a building block

Low-degree switches

?  

Servers

?  

High-degree switches

How should we distribute servers?
Random graphs as a building block

1. How should we distribute servers?

2. How should we interconnect switches?

- Low-degree switches
- High-degree switches

Servers
Random graphs as a building block

1. How should we distribute servers?

2. How should we interconnect switches?

Low-degree switches

Servers

High-degree switches
Distributing servers
Distributing servers
Distributing servers
Distributing servers
Distributing servers

Normalized Throughput

Number of Servers at Large Switches
(Ratio to Expected Under Random Distribution)
Distributing servers

Normalized Throughput

Number of Servers at Large Switches
(Ratio to Expected Under Random Distribution)
Distributing servers

Distributing servers in proportion to switch port-counts

Number of Servers at Large Switches
(Ratio to Expected Under Random Distribution)
Distributing servers

Distributing servers in proportion to switch port-counts

Number of Servers at Large Switches
(Ratio to Expected Under Random Distribution)
Distributing servers

Distributing servers in proportion to switch port-counts

Networks aren’t built like this today!
Random graphs as a building block

1. How should we distribute servers?

2. How should we interconnect switches?

- Low-degree switches
- High-degree switches
- Servers
Interconnecting switches
Interconnecting switches
Interconnecting switches
Interconnecting switches
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)

?
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)

Questions mark
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Interconnecting switches

Vanilla random interconnect

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)
Intuition
Intuition
Intuition
Intuition
Intuition
Intuition
Intuition
Intuition

Still need one crossing!
Intuition

Still need one crossing!

$\Theta \left( \frac{1}{APL} \right)$
Intuition

Throughput should drop when less than $\Theta\left(\frac{1}{APL}\right)$ of total capacity crosses the cut!

Still need one crossing!
Explaining throughput

Upper bound

Empirical value

And constant-factor matching lower bounds in special case
Two regimes of throughput

Normalized Throughput
Cross-cluster Links
(Ratio to Expected Under Random Connection)
Two regimes of throughput

Normalized Throughput

Cross-cluster Links (Ratio to Expected Under Random Connection)
Two regimes of throughput

Normalized Throughput

Cross-cluster Links
(Ratio to Expected Under Random Connection)

sparsest cut

“plateau”:
(total cap) / APL
Implications

A wide range of connectivity options
Implications

A wide range of connectivity options
Implications

A wide range of connectivity options

Bisection bandwidth ≠ throughput
Implications

A wide range of connectivity options

Bisection bandwidth ≠ throughput

Greater freedom in cabling
Quick recap!
How close can we get to optimal network capacity?
How close can we get to optimal network capacity?

How should we distribute servers?
0. How close can we get to optimal network capacity?

1. How should we distribute servers?

2. How should we interconnect switches?
Improving a REAL heterogeneous topology
The VL2 topology

[Greenburg, Hamilton, Jain, Kandula, Kim, Lahiri, Maltz, Patel, Sengupta, SIGCOMM’09]
The VL2 topology

Link-state network carrying only LAs (e.g., 10/8)

Internet

D₁x10G

Dₐ/2 x Intermediate Switches

Dₐ/2 x 10G

D₁ x Aggregate Switches

DₐD₁/4 x ToR Switches

2x10G

Dₐ/2 x 10G

D₁ x Aggregate Switches

20(DₐD₁/4) x Servers

Fungible pool of servers owning AAs (e.g., 20/8)

20 Servers

The VL2 topology
Building on proven networking technology, VL2 uses a scalable, reliable directory system to maintain the map of client requests. The VL2 topology, as shown in Figure 1, is a Clos network for which there is no over-subscription. The VL2 topology is a variation of the Clos network, where the interconnection pattern of the switches is designed to provide a high degree of connectivity and scalability.

The VL2 topology is a two-tiered network, with an aggregate level and a server level. The aggregate level consists of a set of aggregate switches, which connect to a set of intermediate switches. The intermediate switches connect to the Internet, and the server level consists of a set of server switches, which connect to the servers.

The VL2 topology is designed to meet the requirements of large-scale data centers, which are characterized by high volume of traffic, high degree of connectivity, and high degree of scalability. The VL2 topology is able to meet these requirements by providing a high degree of connectivity and scalability, while also providing a high degree of fault tolerance.

The VL2 topology is also designed to be easily manageable and to be able to scale to thousands of servers. The VL2 topology is able to scale by adding more aggregate switches, intermediate switches, and server switches as needed. The VL2 topology is also designed to be easily monitored and to be able to detect and recover from failures.

The VL2 topology is a good choice for large-scale data centers, as it is able to meet the requirements of these environments while also being easy to manage and scale.
The VL2 topology

Link-state network carrying only LAs (e.g., 10/8)

Internet

High-degree switches

Low-degree switches

ToRs ≈ Servers

Fungible pool of servers owning AAs (e.g., 20/8)

Aggr

Int

D_A/2 x Intermediate Switches

D_A/2 x 10G

D_1 x Aggregate Switches

D_A/2 x 10G

D_1 x 10G

2 x 10G

20(D_A D/4) x Servers

ToR

Servers

Relation to network tiers

Relation to network tiers

Relation to network tiers

Relation to network tiers

Relation to network tiers

Relation to network tiers

Relation to network tiers

Relation to network tiers
Rewiring VL2

Link-state network carrying only LAs (e.g., 10/8)

Internet

D₁ x10G

D₁ x Aggregate Switches

2 x10G

D₁ x Aggregate Switches

D₂/2 x Intermediate Switches

D₂/2 x 10G

D₄/2 x Intermediate Switches

D₄/2 x 10G

D₂ x Intermediate Switches

D₄ x Intermediate Switches

20(D₄D₄/4) x Servers

20(D₄D₄/4) x Servers

Fungible pool of servers owning AAs (e.g., 20/8)

Connect ToRs proportional to Intermediate/Agg degree
Rewiring VL2

- Link-state network carrying only LAs (e.g., 10/8)
- Internet
- Intermediate Switches
- Aggregate Switches
- ToR Switches
- Servers

Uniform-random interconnection
Connect ToRs proportional to Intermediate/Agg degree

Fungible pool of servers owning AAs (e.g., 20/8)
Rewiring VL2

40% more servers with server-to-server random permutation traffic

Servers at Full Throughput (Ratio Over VL2)

Aggregation Switch Degree
Rewiring VL2

40% more servers with server-to-server random permutation traffic

Servers at Full Throughput (Ratio Over VL2)

Aggregation Switch Degree

6 8 10 12 14 16 18 20

0.95 1 1.05 1.1 1.15 1.2 1.25 1.3 1.35 1.4
Rewiring VL2

40% more servers with server-to-server random permutation traffic

Servers at Full Throughput (Ratio Over VL2)

Aggregation Switch Degree

all-to-all

rack-to-rack
How do we design throughput optimal network topologies?

https://github.com/ankitsingla/topobench