CS 91: Cloud Systems & Datacenter Networks

Networks Background

CAN YOU NAME ALL THE DWARFS FROM SNOW WHITE?

SURE, THERE'S, UM...

SNEEZY PHYLUM EUROPE SLOTH GUACAMOLE DATA LINK COLOSSUS OF RHODES

I HAVE THIS PROBLEM WHERE ALL SETS OF SEVEN THINGS ARE INDISTINGUISHABLE TO ME.
Walrus / Bucket
Agenda

• Overview of traditional network topologies

• Introduction to software-defined networks

• Layering and terminology
Topology

• Shape and structure of the network

“Bus”

“Star” or
“Hub and spoke”
Topology

• Shape and structure of the network

“Tree”

Most common in datacenter. Why?
Growth

“Star” or “Hub and spoke”
Growth
“Traditional” Datacenter
“Traditional” Datacenter
“Traditional” Datacenter
Recall: Oversubscription

- Core
- Aggregation
- Edge

40 Servers
40 Servers
40 Servers
40 Servers
Oversubscription

All hosts connected to same switch. Full speed between all of them. (1:1) (e.g., 1 Gbps)
Oversubscription

If these links are also 1 Gbps, we now have $1/40^{th}$ of link capacity for each host. (40:1)

What might we do to help alleviate this?
Oversubscription

10 Gbps gives us a ratio of 4:1. Good enough?

Highlighted sub-section *might* be OK if:
- Application has good locality
- VMs assigned intelligently
Oversubscription

Might wants hosts to have 10 Gbps too!

Still problematic. Can’t increase speed forever.
Oversubscription

What else can we do to give our network more capacity?
Redundancy (Multi-rooted Tree)

Pro: More capacity!

Con: Decision must be made (complexity, reordering)
Redundancy (Multi-rooted Tree)

Pro: More capacity!
Con: Decision must be made (complexity, reordering)
Equal-Cost Multi Path (ECMP)

• If there’s a decision to be made, hash packet

• All packets of flow (conversation) use same path
• Hash collisions lead to poor utilization
Paper Preview

• “A Scalable, Commodity Data Center Network Architecture”
  – Alternative tree redundant tree topology design

• “Hedera: Dynamic Flow Scheduling for Data Center Networks”
  – Use live traffic information to decide path choice

• “Augmenting Data Center Networks with Multi-Gigabit Wireless Links”
Network Hardware

Traditional Hardware

SDN Hardware

This sounds great, why mess with a good thing?

Datacenter operator might want more control...
Software-Defined Networking (SDN)

Traditional Hardware

SDN Hardware

Left @ 400 Mbps
Right @ 200 Mbps

Controller
Centralized Network Control

• Not so great on the Internet
  – Failures common
  – Long distances
  – Who would be in charge?

• Powerful in controlled environment
  – Can express and enforce fine-grained policies
Centralized Network Control

- Treat device as a blank forwarding table.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Src</th>
<th>Dst</th>
<th>Src_Pt</th>
<th>Dst_Pt</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>10.1.1.1</td>
<td>10.1.1.2</td>
<td>54321</td>
<td>80</td>
<td>1</td>
</tr>
</tbody>
</table>
Centralized Network Control

- Table built from TCAM memory.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Src</th>
<th>Dst</th>
<th>Src_Pt</th>
<th>Dst_Pt</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>10.1.1.1</td>
<td>10.1.1.2</td>
<td>54321</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>*</td>
<td>10.1.3.3</td>
<td>10.1.3.4</td>
<td>*</td>
<td>*</td>
<td>2</td>
</tr>
</tbody>
</table>

Wildcard / “don’t care” entries allowed.
Centralized Network Control

• Table built from TCAM memory.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Src</th>
<th>Dst</th>
<th>Src_Pt</th>
<th>Dst_Pt</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>10.1.1.1</td>
<td>10.1.1.2</td>
<td>54321</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>*</td>
<td>10.1.3.3</td>
<td>10.1.3.4</td>
<td>*</td>
<td>*</td>
<td>2</td>
</tr>
</tbody>
</table>

Every row checked in parallel.
Paper (& Lab) Preview

• “Ethane: Taking Control of the Enterprise”
• “OpenFlow: Enabling Innovation in Campus Networks”
• “Extending Networking into the Virtualization Layer”

• Lab: Using SDN (OpenFlow) to implement multi-path routing in simulated network
Layering

• Primary network abstraction: stack of layers

• Each layer has distinct responsibilities

• Each layer could be replace independently
OSI Seven-Layer Model

Application: the application (e.g., the Web, Email)

Presentation: formatting, encoding, encryption

Session: sockets, remote procedure call

Transport: end-to-end connections, reliability

Network: routing

Link (data-link): framing, error detection

Physical: 1’s and 0’s/bits across a medium (copper, the air, fiber)
<table>
<thead>
<tr>
<th>Layer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Layer 7”</td>
<td>Application: the application (e.g., the Web, Email)</td>
</tr>
<tr>
<td>“Layer 3”</td>
<td>Network: IP</td>
</tr>
<tr>
<td>“Layer 2”</td>
<td>Link (data-link): Ethernet</td>
</tr>
<tr>
<td></td>
<td>Physical: Wires</td>
</tr>
</tbody>
</table>
Header Encapsulation

- Application Data
  - Transport: TCP
    - (payload)
  - Network: IP
    - (payload)
  - Link: Ethernet
    - (payload)

This whole chunk of bytes is what goes out on the wire.
## Header Encapsulation

<table>
<thead>
<tr>
<th>Link: Ethernet</th>
<th>Network: IP</th>
<th>Transport: TCP</th>
<th>Application Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(payload)</td>
<td>(payload)</td>
</tr>
</tbody>
</table>

This whole chunk of bytes is what goes out on the wire.

Adds reliability information. Adds port numbers to distinguish between apps.
This whole chunk of bytes is what goes out on the wire.

**Header Encapsulation**

- Adds addresses.
- Network: IP address
- Link: MAC address

<table>
<thead>
<tr>
<th>Application Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport: TCP</td>
</tr>
<tr>
<td>(payload)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network: IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(payload)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link: Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>(payload)</td>
</tr>
</tbody>
</table>
Addressing

• Layer 3 / Network / IP addresses:
  – Hierarchical routing
  – The numerical values say something about the location of the node in the network

• Layer 2 / Data Link / MAC addresses:
  – Just a flat, unique identifier
  – Plug-and-play behavior, no location info
“Middlebox”

• (Usually) transparent device that looks at traffic passing through, sometimes modifies it.

• Examples:
  – Firewall: Look at traffic headers. If the “wrong” machines are communicating, block it.
  – Load balancer, intrusion detection, many more.
Other terms

• Please post on Piazza or ask in class if you come across other terms that aren’t clear!