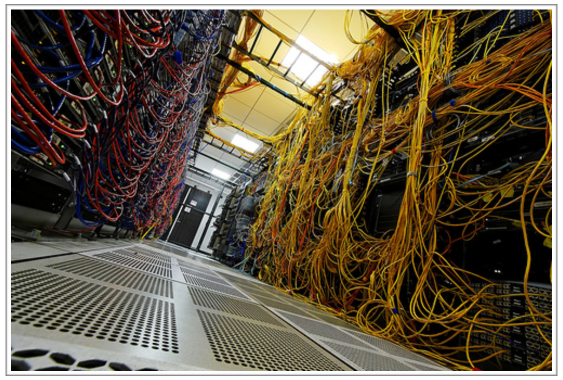
Datacenter Networking

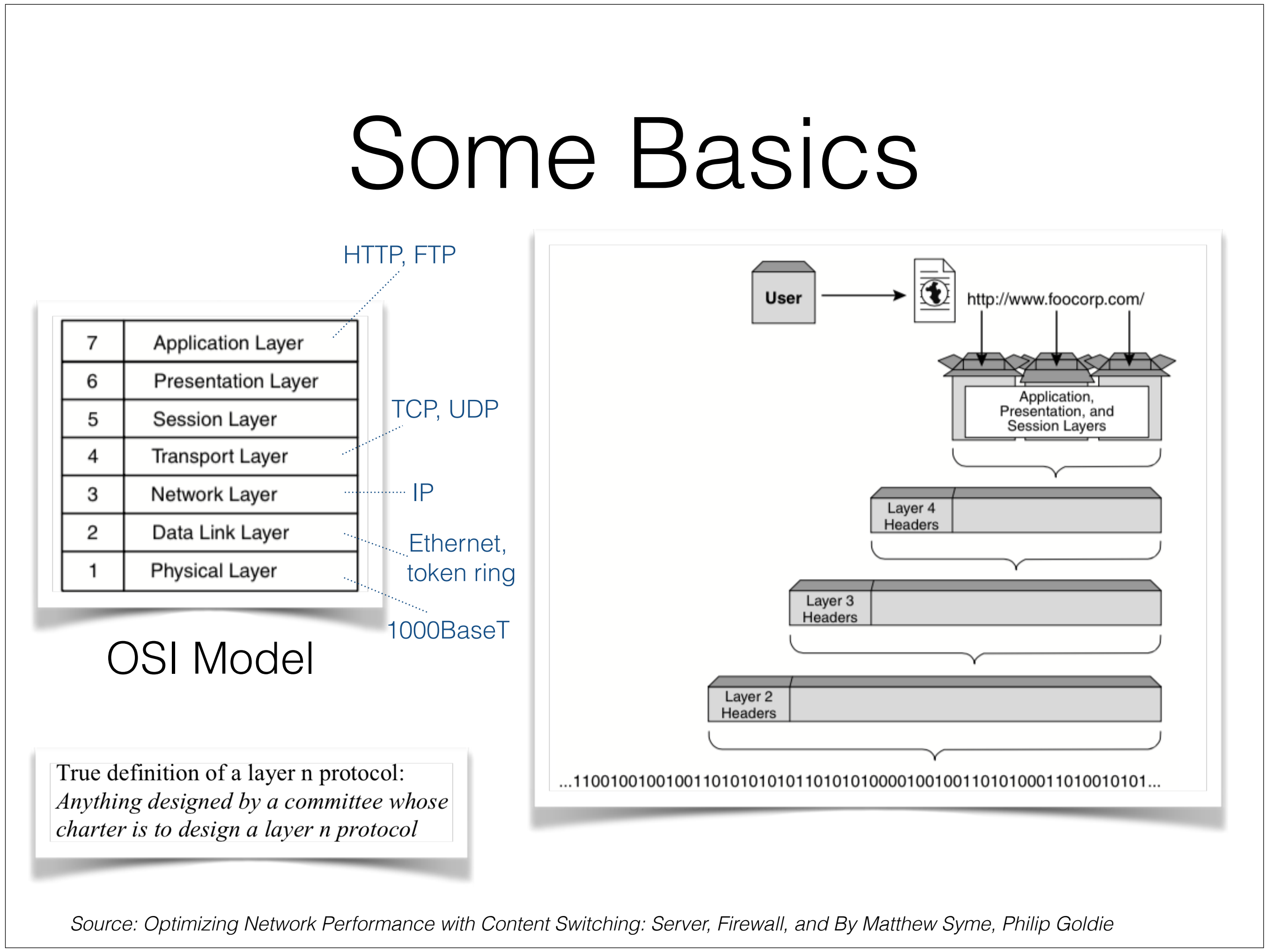
Torgny Holmberg



Outline

- Basic stuff
    - OSI and TCP/IP model
    - Transport - TCP
    - Network - Routing
    - Link - Bridging
- Some Network Technologies
    - VLAN
    - Tunneling
    - VXLAN
    - Open vSwitch
- Structure of the data center
- DC Network Traffic Characteristics

Some Basics



7	Application Layer
6	Presentation Layer
5	Session Layer
4	Transport Layer
3	Network Layer
2	Data Link Layer
1	Physical Layer

OSI Model

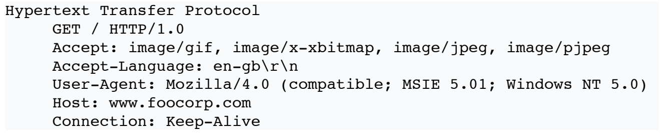
True definition of a layer n protocol:
Anything designed by a committee whose charter is to design a layer n protocol

Source: Optimizing Network Performance with Content Switching: Server, Firewall, and By Matthew Syme, Philip Goldie

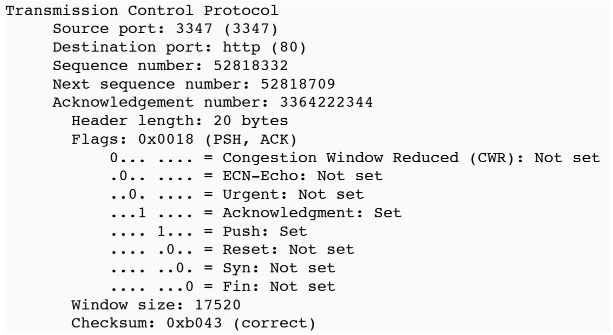
OSI - Open Systems Interconnection
ISO - International Standards Organization

presentation - data conversion, compression/decompresison, encryption/decryption
session - authentication, authorisation, session restoration

Example: www access

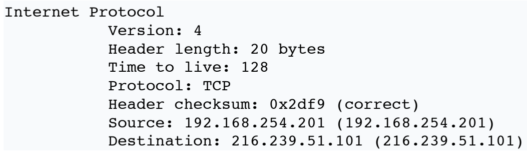


```
Hypertext Transfer Protocol
     GET / HTTP/1.0
     Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg
     Accept-Language: en-gb\r\n
     User-Agent: Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0)
     Host: www.foocorp.com
     Connection: Keep-Alive
```



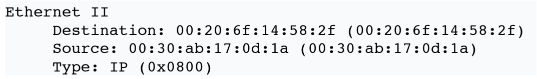
```
Transmission Control Protocol
     Source port: 3347 (3347)
     Destination port: http (80)
     Sequence number: 52818332
     Next sequence number: 52818709
     Acknowledgement number: 3364222344
       Header length: 20 bytes
       Flags: 0x0018 (PSH, ACK)
           0... .... = Congestion Window Reduced (CWR): Not set
           .0.. .... = ECN-Echo: Not set
           ..0. .... = Urgent: Not set
           ...1 .... = Acknowledgment: Set
           .... 1... = Push: Set
           .... .0.. = Reset: Not set
           .... ..0. = Syn: Not set
           .... ...0 = Fin: Not set
       Window size: 17520
       Checksum: 0xb043 (correct)
```

Transport at L4



```
Internet Protocol
           Version: 4
           Header length: 20 bytes
           Time to live: 128
           Protocol: TCP
           Header checksum: 0x2df9 (correct)
           Source: 192.168.254.201 (192.168.254.201)
           Destination: 216.239.51.101 (216.239.51.101)
```

Routers at L3



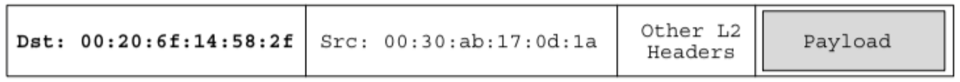
```
Ethernet II
     Destination: 00:20:6f:14:58:2f (00:20:6f:14:58:2f)
     Source: 00:30:ab:17:0d:1a (00:30:ab:17:0d:1a)
     Type: IP (0x0800)
```

Switches at L2

Source: Optimizing Network Performance with Content Switching: Server, Firewall, and By Matthew Syme, Philip Goldie

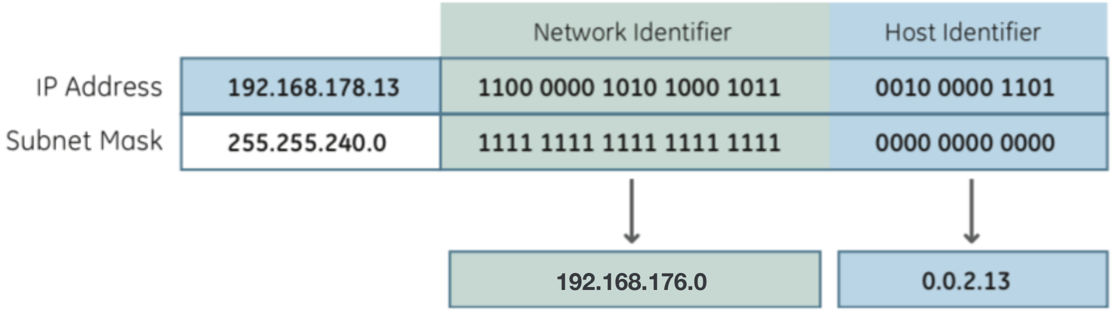
- Ethernet addresses are “flat”, which means that they have nothing to do with where a device is located/connected

- Fixed at time of manufacture



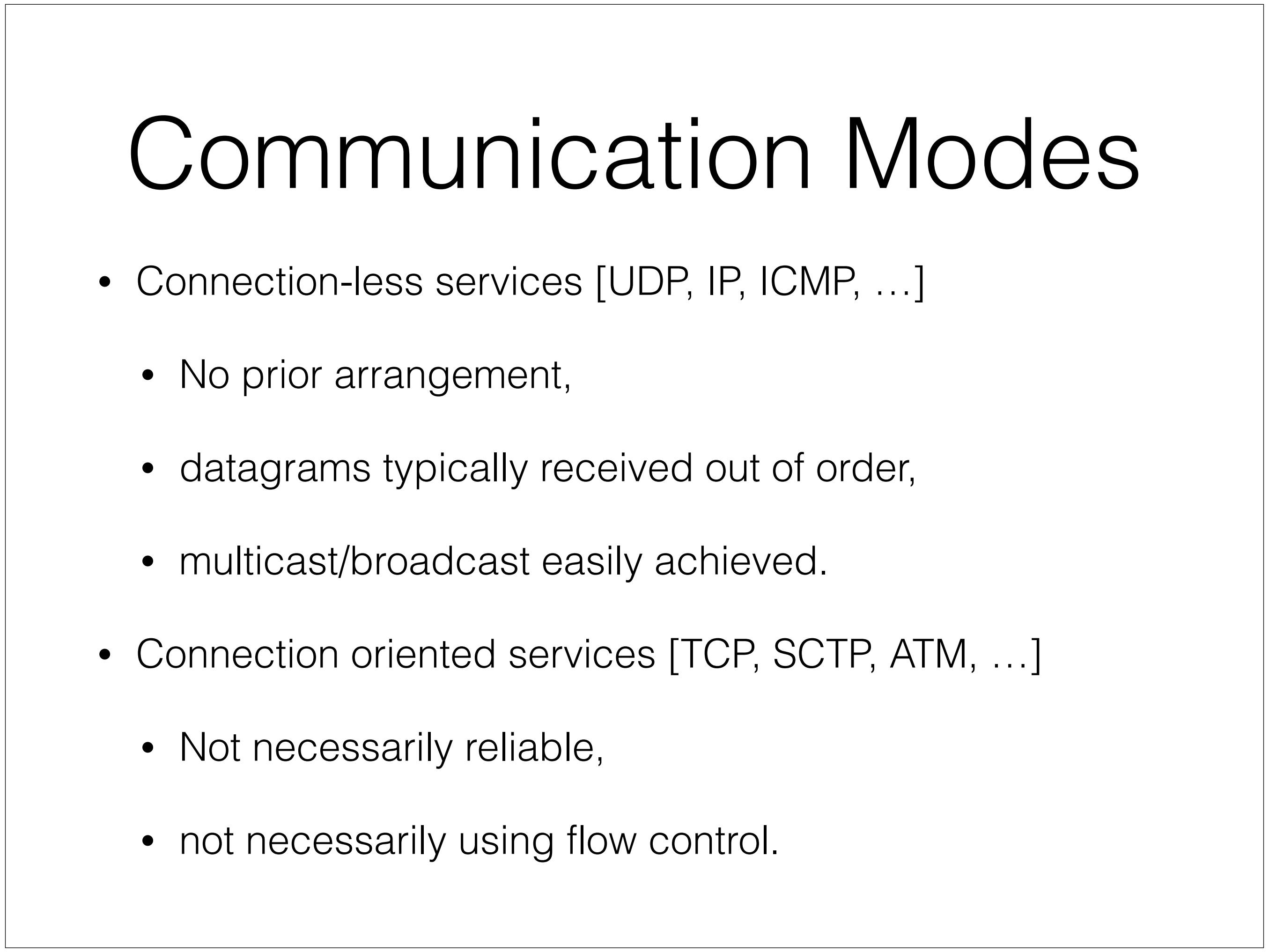
Dst: 00:20:6f:14:58:2f	Src: 00:30:ab:17:0d:1a	Other L2 Headers	Payload

- IP addresses consist of network id and host id



		Network Identifier	Host Identifier
IP Address	192.168.178.13	1100 0000 1010 1000 1011	0010 0000 1101
Subnet Mask	255.255.240.0	1111 1111 1111 1111 1111	0000 0000 0000
		192.168.176.0	0.0.2.13

Source: Radia Perlman

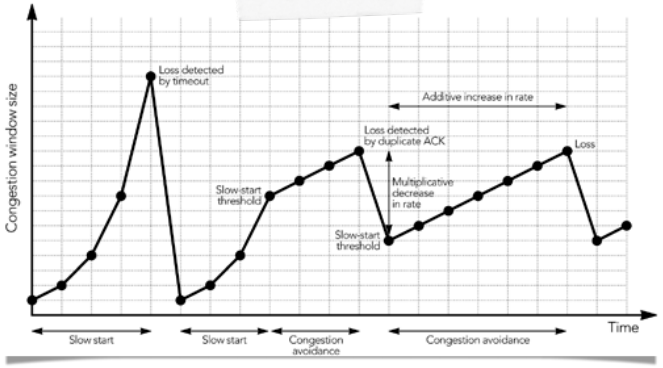


Communication Modes

- Connection-less services [UDP, IP, ICMP, …]
  - No prior arrangement,
  - datagrams typically received out of order,
  - multicast/broadcast easily achieved.
- Connection oriented services [TCP, SCTP, ATM, …]
  - Not necessarily reliable,
  - not necessarily using flow control.

Stream Control Transmission Protocol (SCTP)

TCP-Transmission Control Protocol



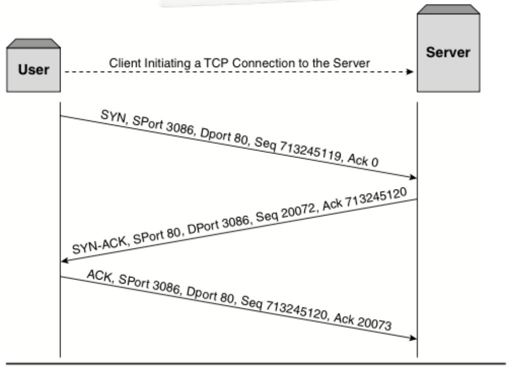
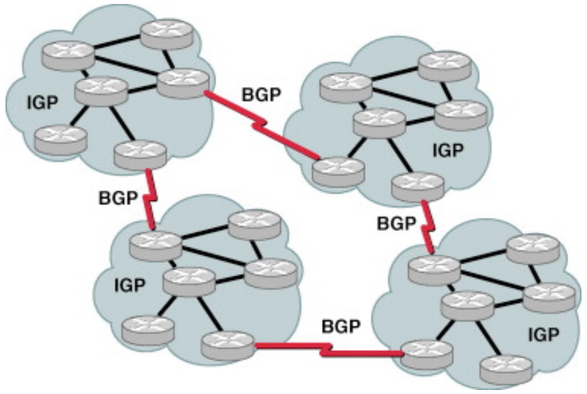

User
Server
Client Initiating a TCP Connection to the Server
SYN, SPort 3086, Dport 80, Seq 713245119, Ack 0
SYN-ACK, SPort 80, DPort 3086, Seq 20072, Ack 713245120
ACK, SPort 3086, Dport 80, Seq 713245120, Ack 20073

Figure 2–6 The TCP three-way handshake.

- TCP-Tahoe
- TCP-Reno
- TCP-New Reno
- TCP-Vegas
- TCP-SACK
- TCP-CUBIC
- **DCTCP (Data Center TCP)**

Cubic function of time since the last congestion event, with the inflection point set to the window prior to the event.

Routing

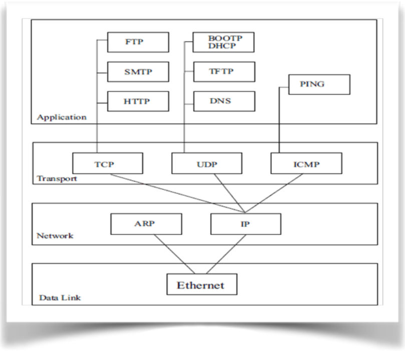


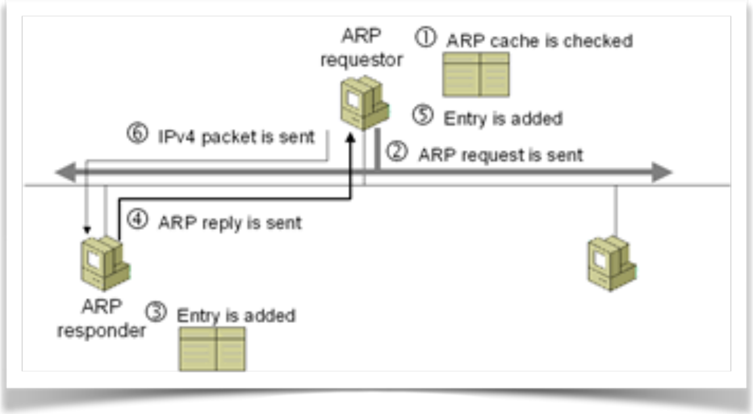
- **Interior Gateway Protocols** - Handle routing within an Autonomous System (one routing domain).
  - Distance Vector Protocols
    - Routing Information Protocol (RIP)
    - Interior Gateway Routing Protocol (IGRP)
  - Link State Protocols
    - Open Shortest Path First (OSPF)
    - Intermediate System to Intermediate System (IS-IS)

- **Exterior Gateway Protocols** - Routing outside AS. Finding best path traversing networks.
  - Border Gateway Protocol (BGP)
  - Exterior Gateway Protocol (EGP)

ARP - Address Resolution Protocol

- ARP is used to convert an IP address to a physical address such as an Ethernet address (also known as a MAC address).





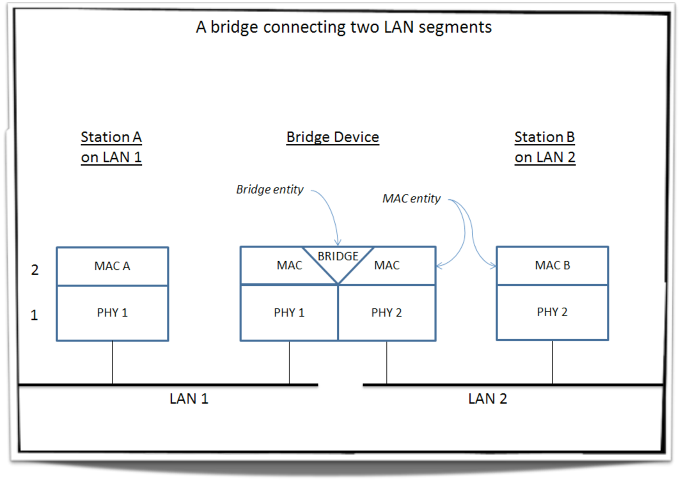
https://technet.microsoft.com/en-us/library/bb726993.aspx

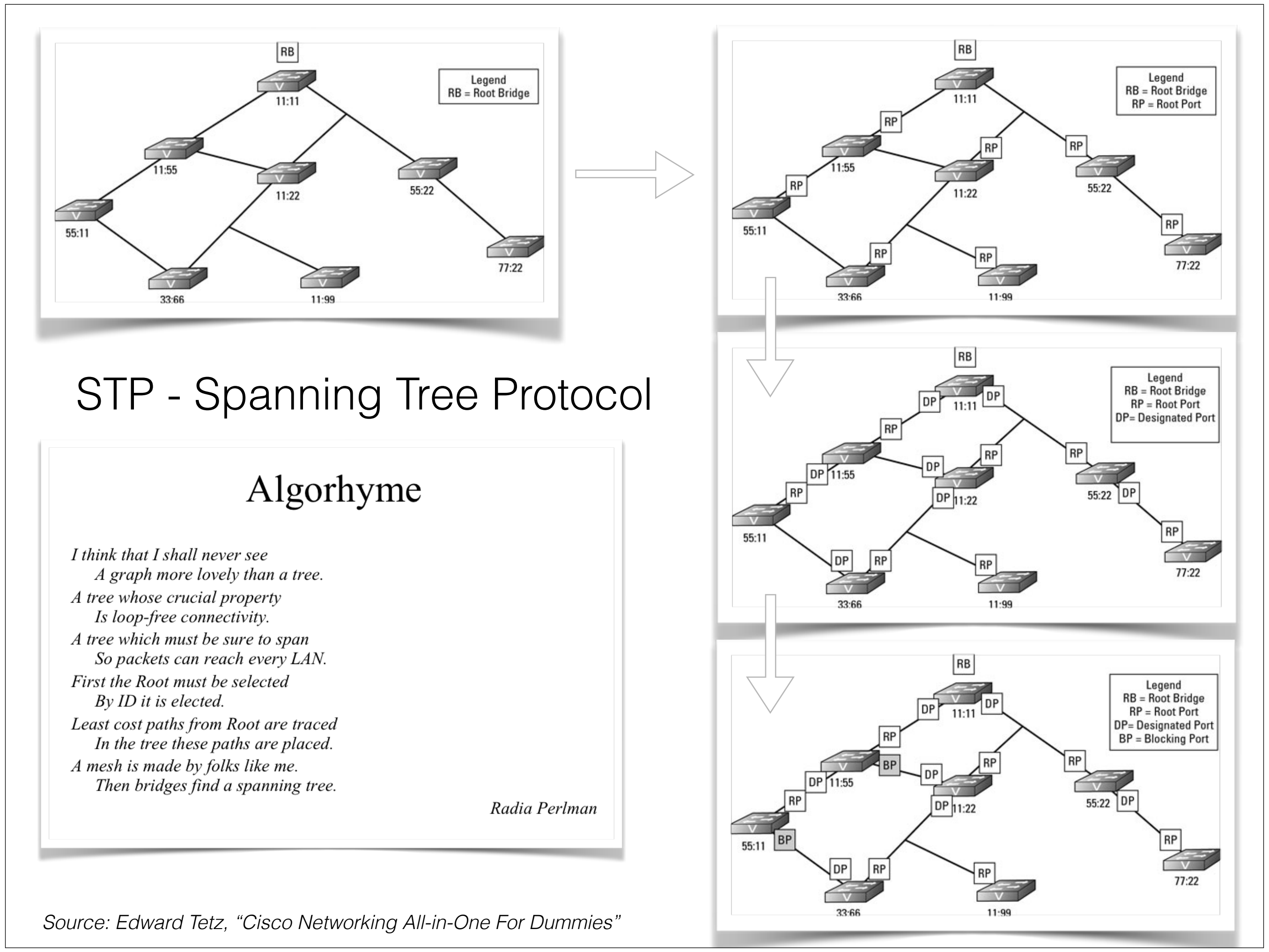
Bridging

- Layer-2 forwarding. Originates from a misconception that Ethernet´s
  - 48 bit addresses (no address collision, no configuration - sweet!), and
  - LANs are no longer point-to-point

  would solve all problems that people forgot about the networking layer, layer 3. But there was a need to reach farther beyond the LAN, so connecting LANs became the solution. A need for forwarding frames was invented - the *transparent* bridge.

- Partition subnet into collision domains.
- By learning addresses a forwarding table can be created. *Loops!*





STP - Spanning Tree Protocol

Algorhyme

I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree which must be sure to span
So packets can reach every LAN.
First the Root must be selected
By ID it is elected.
Least cost paths from Root are traced
In the tree these paths are placed.
A mesh is made by folks like me.
Then bridges find a spanning tree.

Radia Perlman

Source: Edward Tetz, "Cisco Networking All-in-One For Dummies"

Bridge Protocol Data Units (BPDUs)

1)Identify root bridge (RB)
lowest id (MAC+priority)
BPDU transmitted every 2 seconds

2) Identify root ports (RP)
shortest path to RB on switch
3) Identify designated ports (DP)
shortest path to RB on segment

—-
Blocking State - A port in the blocking state does not participate in frame forwarding and also discards frames received from the attached network segment. Only listens to BPDUs.

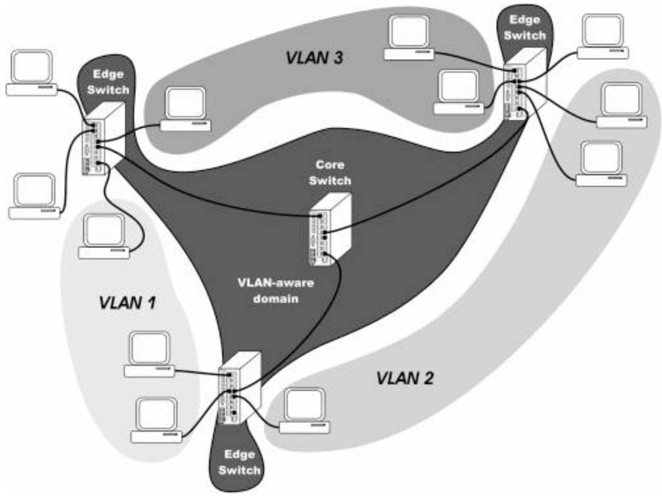
Listening State - After blocking state, a Root Port or a Designated Port will move to a listening state. All other ports will remain in a blocked state. At this state, the port receives BPDUs from the network segment and directs them to the switch system module for processing. After 15 seconds, the switch port moves from the listening state to the learning state.

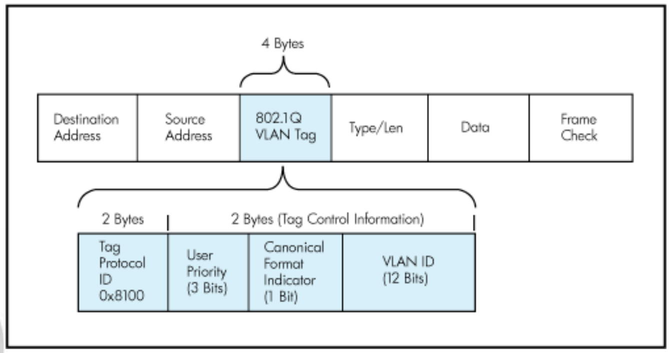
Learning State - A port changes to learning state after listening state. During the learning state, the port is listening for and processing BPDUs. In the listening state, the

VLAN- Virtual LAN

- Partition of a single layer-2 segment into several broadcast domains.
- A single VLAN can contain several physical segments connected with switches and routers.
    - VLAN-enabled switches cannot forward traffic across VLAN boundaries
    - Inter-VLAN communication requires a layer-3 switch (router): the VLAN identifies an IP subnet (and vice versa) => the router can direct the traffic.
- IEEE 802.1Q
- Provides isolation within the cloud. Loses isolation when traversing the Internet.

VLAN




4 Bytes

Destination Address	Source Address	802.1Q VLAN Tag	Type/Len	Data	Frame Check

2 Bytes	2 Bytes (Tag Control Information)		
Tag Protocol ID 0x8100	User Priority (3 Bits)	Canonical Format Indicator (1 Bit)	VLAN ID (12 Bits)

Limited number: 4096-2 = 4094. Static allocation.

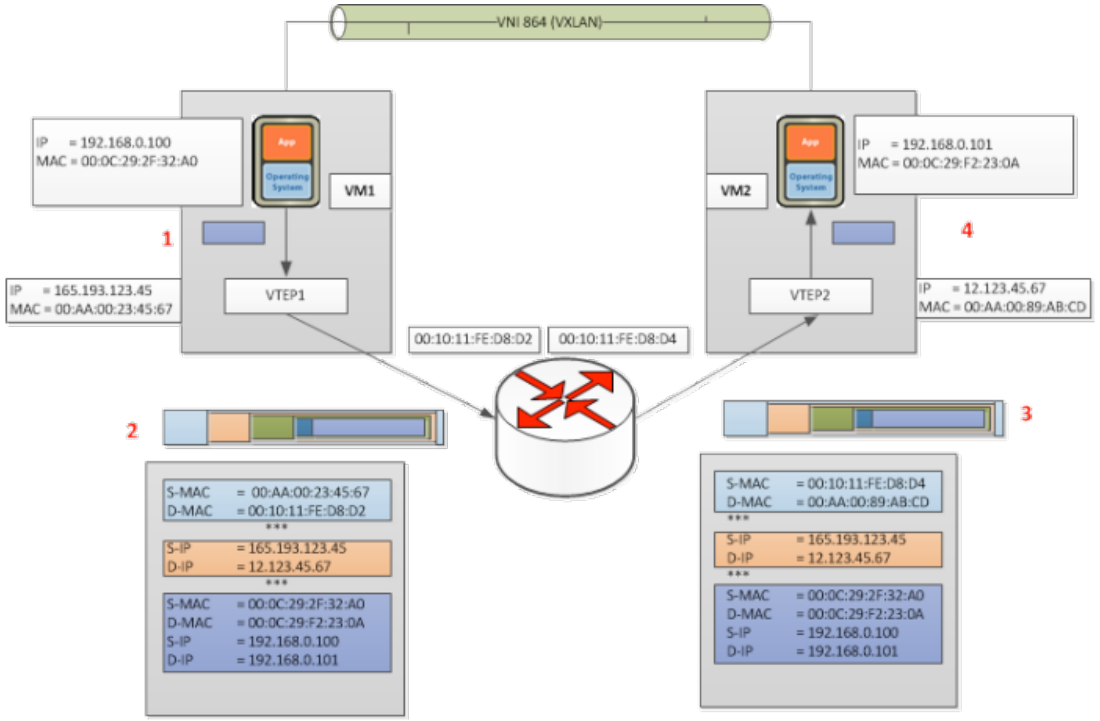
Tunneling

- Provides a network service that the underlying network cannot provide.
  - IPv6 over IPv4
  - VPN - Virtual Private Network, provide secure access to a network using non-secure networks. Uses IPSec “encrypt an IP datagram and put it in an IP datagram”
- Usually violates the OSI model, i.e., the layer m payload contains layer $n<m$ protocol data.
- Communication between data centers typically over tunnels.

VXLAN-
Virtual Extensible LAN

- VLAN on steroids.
- Addresses scalability problem of layer-2 networks.
- Allows 2^24 logical networks. Identified by VXLAN Network Identifier (VNI).
- Encapsulates layer-2 frame in UDP datagram. Layer 2 on top of layer 3!
- Connect separate layer-2 domains to create one domain.
- Machines are identified uniquely by the combination of their MAC address and VNI.
- VXLAN Tunnel End Points (VTEP) encapsulate/decapsulate layer-2 frames.

VXLAN



IP = 192.168.0.100
MAC = 00:0C:29:2F:32:A0

IP = 192.168.0.101
MAC = 00:0C:29:F2:23:0A

IP = 165.193.123.45
MAC = 00:AA:00:23:45:67

IP = 12.123.45.67
MAC = 00:AA:00:89:AB:CD

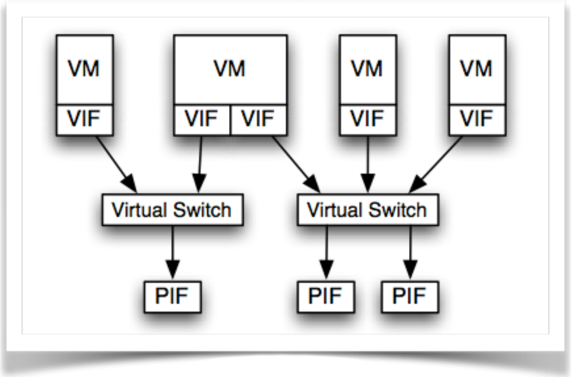
S-MAC = 00:AA:00:23:45:67
D-MAC = 00:10:11:FE:D8:D2
***
S-IP = 165.193.123.45
D-IP = 12.123.45.67
***
S-MAC = 00:0C:29:2F:32:A0
D-MAC = 00:0C:29:F2:23:0A
S-IP = 192.168.0.100
D-IP = 192.168.0.101

S-MAC = 00:10:11:FE:D8:D4
D-MAC = 00:AA:00:89:AB:CD
***
S-IP = 165.193.123.45
D-IP = 12.123.45.67
***
S-MAC = 00:0C:29:2F:32:A0
D-MAC = 00:0C:29:F2:23:0A
S-IP = 192.168.0.100
D-IP = 192.168.0.101

http://www.borgcube.com/blogs/2011/11/vxlan-primer-part-1/

Open vSwitch

- Virtualized bridge that provides VM to VM connectivity.
- Tight connection to the virtualization layer.



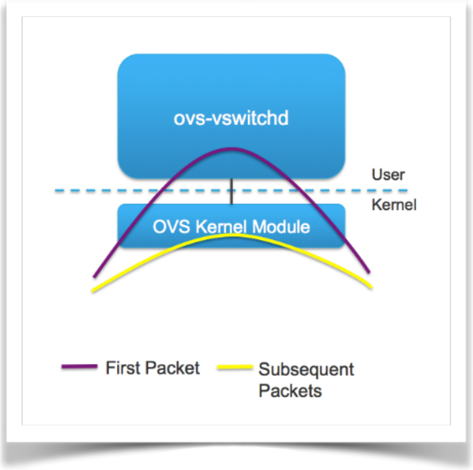
- Rich set of features and a clear configuration interface.
- Can use OpenFlow and OpenFlow controllers (support for complex actions, i.e., wildcards, priorities, QoS, multiple tables…)

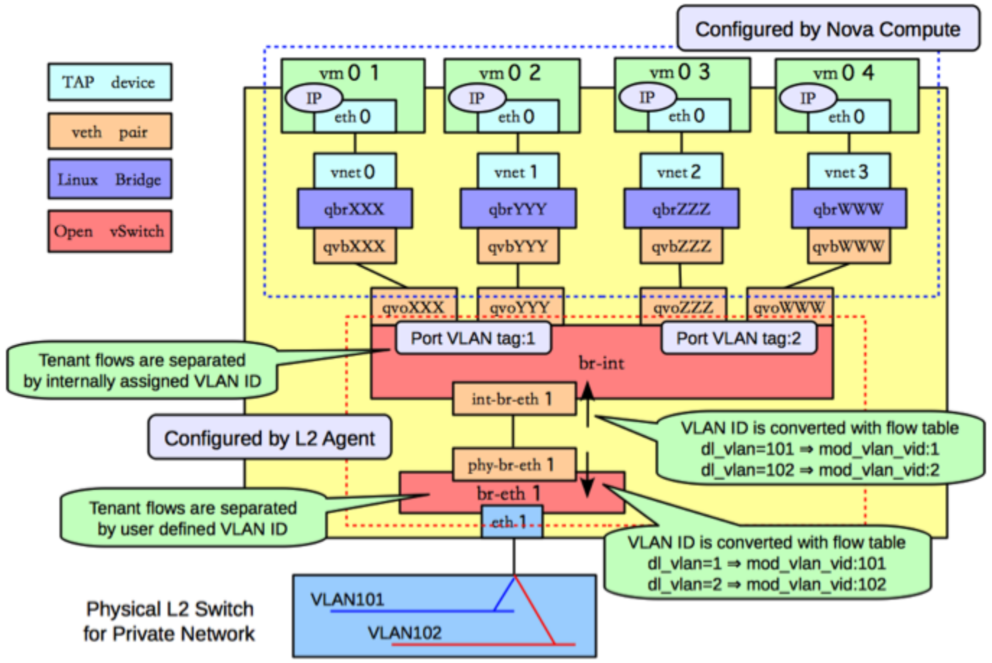
B. Pfaff, J. Pettit,T. Koponen, K. Amidon, M. Casado, S. Shenker. *Extending Networking into the Virtualization Layer*, ACM SIGCOMM Workshop on Hot Topics in Networking (HotNets), October 2009.

Open vSwitch

- Moves physical transmission and handling of packets onto the host.

- Multilayer switch - uses information from layer 2, layer 3 and layer 4 to direct flows.



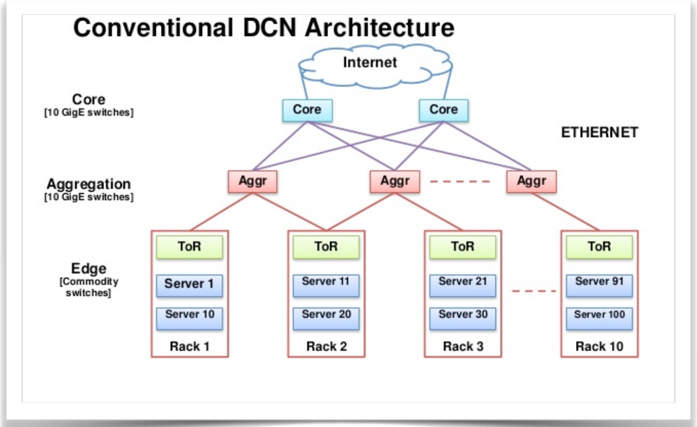

Configured by Nova Compute
TAP device
veth pair
Linux Bridge
Open vSwitch
vm01
vm02
vm03
vm04
IP
eth0
vnet0
vnet1
vnet2
vnet3
qbrXXX
qbrYYY
qbrZZZ
qbrWWW
qvbXXX
qvbYYY
qvbZZZ
qvbWWW
qvoXXX
qvoYYY
qvoZZZ
qvoWWW
Port VLAN tag:1
Port VLAN tag:2
br-int
Tenant flows are separated by internally assigned VLAN ID
int-br-eth1
Configured by L2 Agent
VLAN ID is converted with flow table dl_vlan=101 ⇒ mod_vlan_vid:1 dl_vlan=102 ⇒ mod_vlan_vid:2
phy-br-eth1
br-eth1
eth1
Tenant flows are separated by user defined VLAN ID
VLAN ID is converted with flow table dl_vlan=1 ⇒ mod_vlan_vid:101 dl_vlan=2 ⇒ mod_vlan_vid:102
Physical L2 Switch for Private Network
VLAN101
VLAN102

http://docs.openstack.org/admin-guide-cloud/content/under_the_hood_openvswitch.html#under_the_hood_openvswitch_scenario2

Linux bridge! OpenStack relies on Linux iptables firewalling. Currently tap/OvS does not implement iptables.

Cloud Networking

- Dynamics
  - mobility, migration of VMs
  - short lived services
  - on demand scaling
- Scaling
  - many VMs on many hosts
- Isolation
  - tenants sharing the same physical resource
- Traffic
  - North-south/East-west
  - Not always on physical links



http://www.datacenterknowledge.com/archives/2009/07/14/quality-tech-builds-its-data-center-network/
http://www.slideshare.net/AnkitaMahajan2/introduction-to-data-center-network-architecture

4-5 time as much east-west traffic compared to north-south.
Analysis of massive data sets is the driver for today’s data center.
NW dimensioning difficult problem: Traffic matrix/src-dest pairs is *n(n-1)* but the number of measurement points/links is *2n*. Lose information. Assume even distributed load?

Network Traffic Characteristics

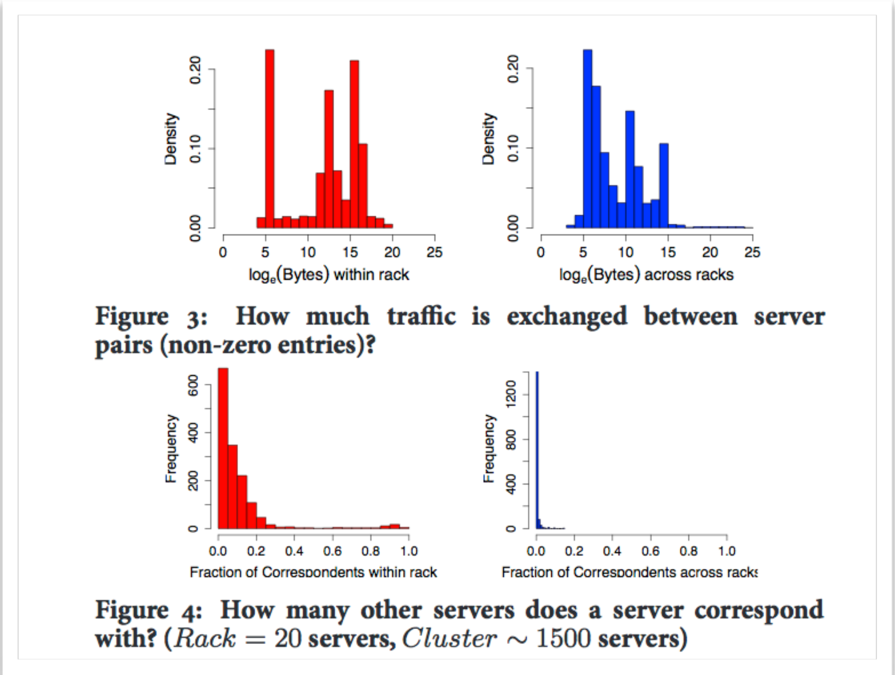


Figure 3: How much traffic is exchanged between server pairs (non-zero entries)?

Figure 4: How many other servers does a server correspond with? ($Rack = 20$ **servers,** $Cluster \sim 1500$ **servers)**

Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. 2009. The nature of data center traffic: measurements & analysis.
In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (IMC '09). ACM, New York, NY, USA, 202-208.
DOI=10.1145/1644893.1644918 http://doi.acm.org/10.1145/1644893.1644918

Mining data center. Map-reduce.
P(No traffic exchange between servers that exist in the same rack) = 0.89
P(No traffic exchange between servers in different racks) = 0.995
Either “all” or <= 25% of there servers are addressed within a rack.
Speaks to none or 1-10% of there servers outside the rack.

Engineering decision based on TCPs inability to recover from congestion when BW-delay product is low, e.g., when RTT is really short:
-limit number of contending flows: limiting number of simultaneously open connections
-separate flows; keep within rack or on separate VLANs (keep broadcast domain small)

Network Traffic Characteristics

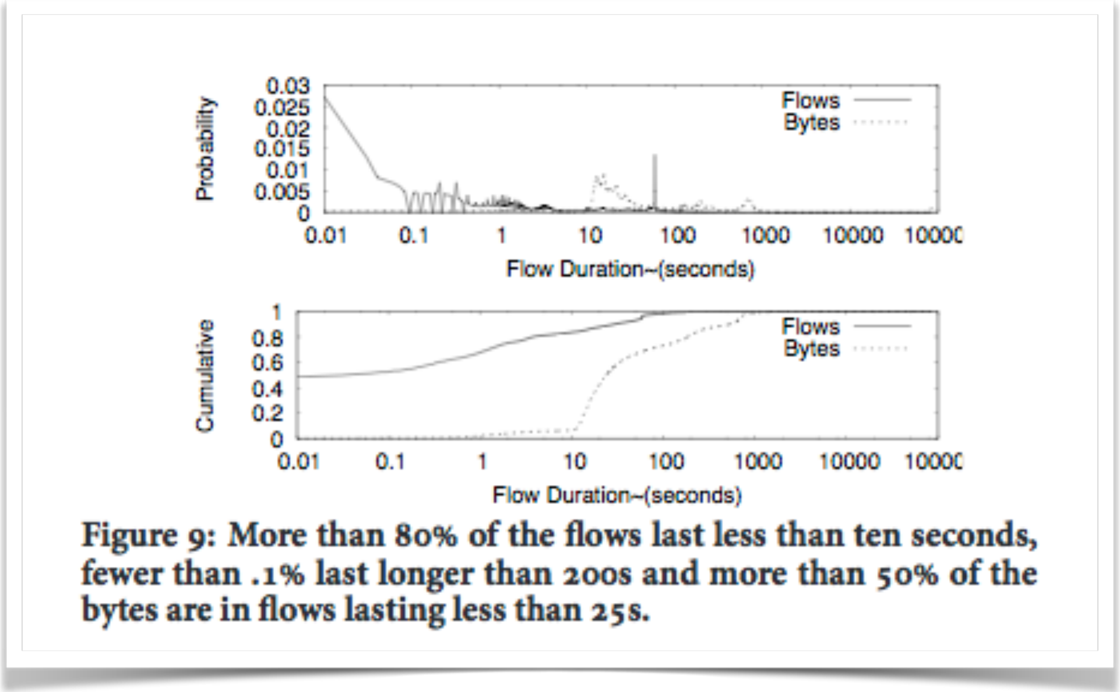


Figure 9: More than 80% of the flows last less than ten seconds, fewer than .1% last longer than 200s and more than 50% of the bytes are in flows lasting less than 25s.

Srikanth Kandula, Sudipta Sengupta, Albert Greenberg, Parveen Patel, and Ronnie Chaiken. 2009. The nature of data center traffic: measurements & analysis. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (IMC '09). ACM, New York, NY, USA, 202-208. DOI=10.1145/1644893.1644918 http://doi.acm.org/10.1145/1644893.1644918

DCTCP

- Observations:
  - ToR switches are typically also low-cost,
  - diverse mix of short and long flows,
  - flow requirements:
    - low latency for short flows,
    - high burst tolerance,
    - high utilization of long flows.
  - Actually not a large number of simultaneous traffic flows.

Partition/Aggregate flows

Data updates

Alizadeh M., Greenberg A. G., Maltz D. A., Padhye J., Patel P., Prabhakar B., Sengupta S., Sridharan M.: Data center TCP (DCTCP). SIGCOMM 2010: 63-74

Performance Impairment

- Incast
  - Large number of flows arriving simultaneously to the switch.

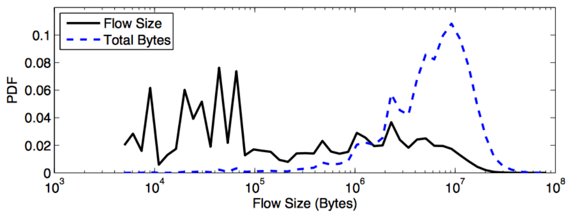


Figure 4: PDF of flow size distribution for background traffic. PDF of Total Bytes shows probability a randomly selected byte would come from a flow of given size.

- Queue buildup
  - Long flows penalise short flows. The short delay sensitive flow waits for the long flows in the switch buffer.
- Buffer pressure
  - Input traffic on one port is affected by traffic on other ports due to a shared memory design in the switches.

- DCTCP designed to handle diverse mix of short and long flows,
- keep low switch buffer occupancies (low RTT) yet maintaining hight TP for the long flows,
- address saw-tooth buffer buildup => variation in buffering => variation in packet delay.
- Solution:
  - Use TCP Explicit Congestion Control (ECN),
  - estimate fraction of marked packages (indicate level of congestion), and
  - adjust TCP congestion window accordingly.

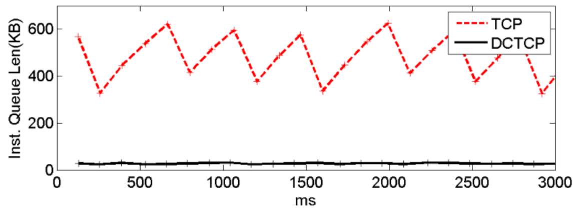
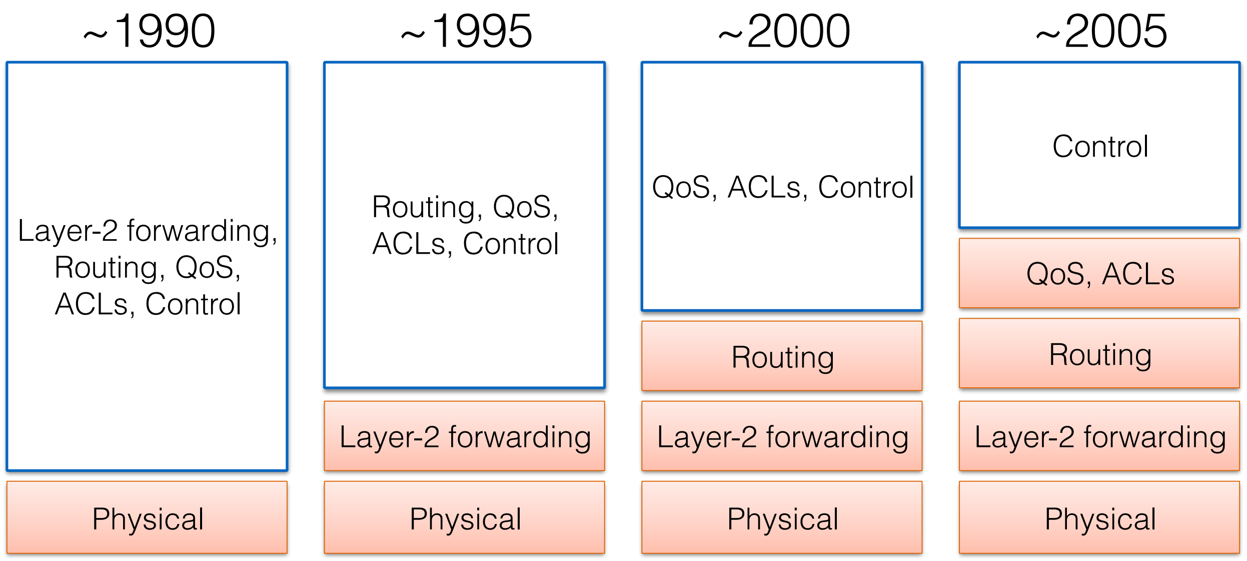


Figure 1: Queue length measured on a Broadcom Triumph switch. Two long flows are launched from distinct 1Gbps ports to a common 1Gbps port. Switch has dynamic memory management enabled, allowing flows to a common receiver to dynamically grab up to 700KB of buffer.

NW functionality migrating to HW



Software Defined Networks - A Comprehensive Approach, P. Göransson and C. Black