Data center Networking: New advances and Challenges (Ethernet)

Anupam Jagdish Chomal

Principal Software Engineer

DellEMC Isilon

Bitcoin mining – Contd

- Main reason for bitcoin mines at Iceland is the natural cooling for servers and cheap energy due to Iceland's abundance of renewable energy from geothermal and hydroelectric power plants.
- Data centers are specially designed to utilize the constant wind on the bare peninsula.
- Walls are only partial on each side, allowing a draft of cold air to cool down the equipment and move out from the other end
- Example – http://www.businessinsider.com/photos-iceland-bitcoin-ethereum-mine-genesis-mining-cloud-2016-6?r=UK&IR=T

Agenda

- Typical Datacenters
- New class and existing TCP issues
- TCP Variants
- Google’s BBR
- Facebook’s Open Compute Project

Why Ethernet?

- InfiniBand has low and predictable latency, flatter topology, and less computing power on the CPU
- Many of the top500 supercomputers(HPC) use Infiniband
- However, InfiniBand itself makes up just a small part of data-center networking
- A small number (about 5%) percent of all server controllers and adapters shipping these days use InfiniBand, with most of the rest using Ethernet
- Ethernet offers more connectivity across the market for networking equipment.

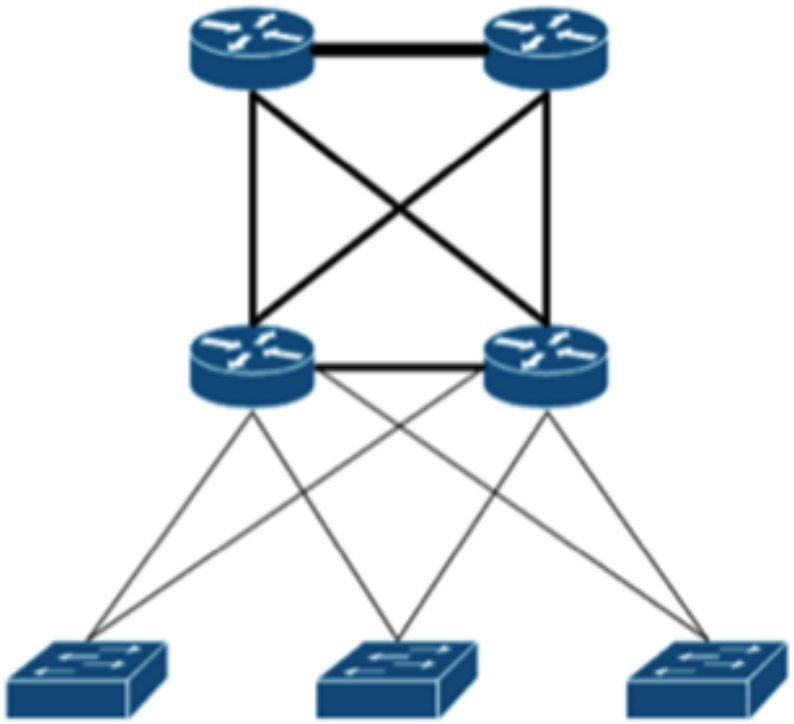
A Typical Datacenter

- Switch Placement
  - Top Of Rack (TOR)
  - End of Row (EOR)
- Traffic Patterns
  - North-South & East-west traffic
- Architectures
  - Core-Access-Edge Architecture
  - Leaf Spine Architecture
- Different organizations and different class of applications share cloud racks/infrastructure. Its easier to strictly share CPU and memory between then but tough to get a fair sharing of Network resource

TOR Vs EOR

- A point of delivery, or PoD, is "a module of network, compute, **storage**, and application components that work together to deliver networking services
- TOR (Top of Rack)
  - The edge/access switch in placed at the top of a server rack
  - Servers in the rack are directly connected to this switch
  - Each rack would have one or two such switches
  - All edge switches then connect to the aggregation layer
- EOR (End of Row)
  - Every server directly connects to a aggregation switch. Switch from the rack is removed
  - Reduces the number of networking switches and improves the port utilization
- Example – https://blog.gigamon.com/2016/10/04/visibility-is-the-best-disinfectant-for-ransomware/

Core – Aggregation – Access Architecture



CORE

AGGREGATION

ACCESS

Core – Aggregation – Access Contd

- The aggregation layer establishes the Layer 2 domain size and manages it with a spanning tree protocol
- Common application or departmental servers are kept together in a common VLAN or IP Subnet
- Since the layer2 topology is looped, a loop protection mechanism like Spanning tree is used
- The aggregation layer does the work of Spanning tree processing
- STP cannot use parallel forwarding paths, and it always blocks redundant paths in a VLAN.

Leaf Spine Network Topology

- Also called CLOS after its architect – Charles Clos
- Servers are connected to "leaf" switches. These are often arranged as "top-of-rack" or TOR switches. In a redundant setup, each server connects to *two* leaf switches.
- Each leaf switch has connections to all "spine" switches in a full-mesh topology.
- The spine layer is the “Backbone” of the Network and is responsible for interconnecting all leaf switches.
- The spine switches aren't connected directly to each other. Any packet from a given server to another server in another rack goes through the sending server's leaf, then one of the spine switches, then the receiving server's leaf switch.
- Equal-Cost multipath routing is used to distribute traffic across the set of spine switches.
- Example – https://kb.pert.geant.net/PERTKB/LeafSpineArchitecture

Leaf Spine Network Topology

- A spine-leaf design scales horizontally through the addition of spine switches which add availability and bandwidth, which a spanning tree network cannot do.
- Spine-leaf also uses routing with equal-cost multipathing to allow for all links to be active with higher availability during link failures.
- No matter which leaf switch to which a server is connected, its traffic always has to cross the same number of devices to get to another server.
- Latency is at a predictable level because a payload only has to hop to a spine switch and another leaf switch to reach its destination.

New Class and Existing TCP issues

- TCP Out-of-order
- TCP Incast
- TCP Outcast
- TCP Unfairness
- Long queue completion time

Some TCP Terms

- TCP uses a retransmission timer to ensure data delivery in the absence of any feedback from the remote data receiver. The duration of this timer is referred to as RTO (retransmission timeout)
- Round Trip Time (RTT): It measures the time sending a packet to getting the acknowledgment packet from the target host.
- Congestion Window: TCP uses a congestion window in the sender side to do congestion avoidance. The congestion window indicates the maximum amount of data that can be sent out on a connection without being acknowledged.

TCP Retransmission Timeout (RTO)

- TCP starts a **retransmission timer** when an outbound segment is handed down to IP. If there is no acknowledgment for the data in a given segment before the timer expires, then the segment is retransmitted.
- On the initial packet sequence, there is a timer called **Retransmission Timeout** (RTO) that has an initial value of three seconds. After each retransmission the value of the RTO is doubled and the computer will retry up to three times
- If the sender does not receive the acknowledgement after three seconds it will resend the packet. At this point the sender will wait for six seconds to get the acknowledgement. If the sender still does not get the acknowledgement, it will retransmit the packet for a third time and wait for 12 seconds, at which point it will give up

TCP Incast

- TCP Incast is a catastrophic TCP throughput collapse that occurs as the number of storage servers sending data to a client increases past the ability of an Ethernet switch to buffer packets.
- In a clustered file system, for example, a client application requests a data block striped across several storage servers, issuing the next data block request only when all servers have responded with their portion.
- This synchronized request workload can result in packets overfilling the buffers on the client's port on the switch, resulting in many losses.
- Under severe packet loss, TCP can experience a timeout that lasts a minimum of 200ms, determined by the TCP minimum retransmission timeout (RTO_{min}).

TCP Incast

- When a server involved in a synchronized request experiences a timeout, other servers can finish sending their responses, but the client must wait a minimum of 200ms before receiving the remaining parts of the response, during which the client's link may be completely idle.
- The resulting throughput seen by the application may be as low as 1-10\% of the client's bandwidth capacity, and the per-request latency will be higher than 200ms

TCP Incast Mitigation

- Larger switch buffers can delay the onset of Incast (doubling the buffer size doubles the number of servers that can be contacted).
- Reducing TCP's minimum RTO allows nodes to maintain high throughput with several times as many nodes.
- Example: How reduced RTO improves goodput – Source: http://www.pdl.cmu.edu/Incast/

TCP Outcast

- The unfairness caused by bandwidth sharing via TCP in data center networks is called TCP Outcast problem.
- Throughput of a flow with small Round Trip Time (RTT) turn out to be less than that with large RTTT
- The Outcast problem is caused by port blackout in data center

TCP Outcast

- In a multi rooted tree topology, when many flows and a few flows arrive on two ports of a switch destined to one common output port, the small set of flows lose out on their throughput share significantly.
- This occurs mainly in taildrop queues that commodity switches use. These taildrop queues exhibit a phenomenon known as ***port blackout*** where a series of packets from one port are dropped.
- Port blackout affects the fewer flows more significantly, as they lose more consecutive packets leading to TCP timeouts.

TCP Outcast

- When different flows with different RTTs share a given bottleneck link, TCP throughput is inversely proportional to RTT.
- Low RTT flows will get a higher share of the bandwidth than high RTT flows.
- This problem occurs when two conditions are met:
  - Network comprises of commodity switches that emply the simple taildrop queuing discipline
  - When many flows and a few flows arrive on two ports of a switch destined to one common output port

TCP Outcast Mitigation

- Random Early Detection (RED)
  - RED monitors the average queue size and drops packets based on statistical probabilities.
  - If the buffer is almost empty, then all incoming packets are accepted. As the queue grows, the probability for dropping an incoming packet grows too. When the buffer is full, the probability has reached 1 and all incoming packets are dropped
- Stochastic Fair Queue (SFQ)
  - Output buffers are divided into buckets, and flows sharing a bucket get their share of throughput corresponding to the bucket size
- Minimize buffer occupancy at the switches

Types of DataCenter Applications

- Time Sensitive Applications
- Online Data Intensive (OLDI) applications, includes Web search, online retail, and advertisement. Operate under Soft Real Time Constraints (e.g., 300 ms latency)

TCP Variants

Sr.No	Algorithm	Acronym	Details
1	Adaptive Data Transmission in the Cloud	Adaptive TCP (ATCP)	TCP's fairness leads to poor outcomes for time-sensitive applications who should be allocated. The basic idea is to modify the congestion control behavior in TCP(additive increase behavior of TCP congestion control) and perform adaptive weighted fairness sharing among flows. In order to distinguish flows with different timing targets, count how many bytes a flow has delivered already. Then, dynamically tune a flow's weight such that it decreases as a flow transfers more data. In effect, prioritize small flows' bandwidth allocation and get them to complete faster than the larger flows that they are contending with.
2	Deadline Aware Datacenter TCP	D2TCP	Handles bursts, is deadline-aware, and is readily deployable. It uses a distributed and reactive approach for bandwidth allocation which fundamentally enables D2 TCP's properties. D2 TCP employs a novel congestion avoidance algorithm, which uses ECN feedback and deadlines to modulate the congestion window via a gamma-correction function
3	Data Center TCP	DCTCP	Is deadline-agnostic. TCP congestion control scheme for data-center traffic. Cannot be deployed over the public internet.
4	Adaptive-Acceleration Data Center TCP	A2DTCP	A^{2}DTCP can co-exist with conventional TCP as well without requiring more changes in switch hardware than D^{2}TCP and DCTCP Takes into account both network congestion and latency requirement of application service reduces the missed deadline ratio compared to D^{2}TCP and DCTCP.
5	BBR: Congestion-Based Congestion Control		BBR runs purely on the sender and does not require changes to the protocol, receiver, or network, making it incrementally deployable. It depends only on RTT and packet-delivery acknowledgment, so can be implemented for most Internet transport protocols. It's a three-year quest to create a congestion control based on measuring the two parameters that characterize a path: bottleneck bandwidth and round-trip propagation time, or BBR TCP BBR has significantly increased throughput and reduced latency for connections on Google's internal backbone networks and google.com and YouTube Web servers throughput by 4 percent on average globally – and by more than 14 percent in some countries. The TCP BBR patch needs to be applied to the Linux kernel. Use linux kernel 4.9 or above From <https://www.cyberciti.biz/cloud-computing/increase-your-linux-server-internet-speed-with-tcp-bbr-congestion-control/>

DCTCP

- Uses a kind of active Queue Management (AQM), with explicit feedback from congested switches
- An arriving packet(and all subsequent packets) is marked as soon as the queue occupancy is greater than K
- As soon as the sender gets the ECN(Explicit Congestion Notification) it reduces that data flow as per an equation till the ECN bit is cleared
- TCP and DCTCP cannot co-exist
- Not all switches support ECN and can be purely tail-drop

Google’s congestion control algorithm - BBR

- BBR - Bottleneck Bandwidth and Round-Trip Propagation Time
- It use two parameters: the estimated round-trip time for a packet and the maximum bandwidth available for the connection
- Using these two parameters, BBR seek an optimal operating point for the connection with high throughput and low delay which is more adapted to the buffers used by the various types of network equipment in today’s access and core Internet networks
- Replacing CUBIC with BBR experimentally on google.com and youtube.com has resulted in significant improvement in network latency and application metrics

BBR

- Older TCP congestion control systems lead to bottlenecks in Internet traffic because these algorithms were built around the idea of detecting a congestion after it happened, which would be too late to re-route some users.
- BBR was designed to prevent bottlenecks before they happen
- Calculates the congestion window size by measuring the bottleneck bandwidth and round-trip propagation time and sends packets at a paced rate
- BBR is neither delay-based nor loss-based and it ignores packet loss as congestion signal. It also does not explicitly react to congestion, whereas congestion window-based approaches often use a multiplicative decrease strategy

Facebook’s Open Compute Project (OCP)

- The focus of Open Compute is efficient server, storage, and data center hardware designs for scalable computing
- Means to share good server and data center designs
- Is helping create a unified standard that will go from top to bottom, even the way racks are designed and built

Facebook’s Open Compute Project (OCP)

- Power Usage Effectiveness (PUE)
  - **PUE** is the ratio of total amount of energy used by a computer **data center** facility to the energy delivered to computing equipment. **PUE** was originally developed by a consortium called The Green Grid
- Industry standard for PUE stands at 1.9 while OCP claims to provide 1.07
- Came up with designs like OpenRack and OpenVault

OCP Contd

- Facebook's extraordinary Open Compute Project is doing for hardware what Linux, Android, and many other popular products did for software: making it free and "open source."
- That means that anyone can look at, use, or modify the designs of the hugely expensive computers that big companies use to run their operations. All for free. Contract manufacturers are standing by to build custom designs and to build, in bulk, standard designs agreed upon by the group.
- Google’s top hardware infrastructure guy Urs Hölzle says – “It will be relevant only for the very, very large companies -- for the Facebooks, the Ebays, the Microsofts”

References

- **Handbook on Data Centers -** edited by Samee U. Khan, Albert Y. Zomaya
- https://kb.pert.geant.net/PERTKB/LeafSpineArchitecture
- DCTCP https://tools.ietf.org/html/rfc8257
- Attaining the Promise and Avoiding the Pitfalls of TCP in the Datacenter
- Transport protocols for data center networks: a survey of issues, solutions and challenges