XCo: Explicit Coordination to Prevent Network Fabric Congestion in Cloud Computing Cluster Platforms

Presented by Wei Dai
Reasons for Congestion in Cloud

• Cloud operators use virtualization to consolidate thousands of VMs on shared hardware platforms due to cost concerns.

• Most VMs host service-oriented applications that are inherently communication intensive.
Reasons for Congestion in Cloud

- Cloud computing infrastructures consist of large data center clusters using commodity servers and networking hardware.

Pros:
- Cheap
- Easy to install and manage
- Can be shared by a wide range of network services and protocols
Reasons for Congestion in Cloud

• Cloud computing infrastructures consist of large data center clusters using commodity servers and networking hardware.

  Cons:
  ▪ Higher latency
  ▪ Smaller / lower-performance packet buffers

• Switch buffers can easily become overwhelmed by high-throughput traffic that can be bursty and synchronized, leading to significant packet losses.
Types of Congestion

• TCP throughput collapse (also known as Incast)
  
  *Well-known example of congestion experienced by barrier-synchronized traffic, e.g. synchronous reads in networked storage*

• Congestion caused by non-TCP traffic, e.g. UDP

• Congestion caused by traffic not TCP-friendly
  
  *voice/video over IP, and peer-to-peer traffic*

• Congestion caused by large number of short TCP sessions
How to Solve the Problem

• Root cause: transient overload of buffers within switches

• Hardware and software mechanisms are hard to deploy at scale.

• Ethernet flow control in IEEE 802.3x helps in low-end edge switches, but is counter-productive in backbone switches.
How to Solve the Problem

• Current industry practice:
  ▪ Add higher capacity network switches
  ▪ Multi-port network cards
  ▪ Physically separate networks for data and control traffic

• Drawback: increase cost and complexity without addressing the root cause
XCo – Explicit Coordination

- Coordinate network transmissions from multiple VMs to avoid throughput collapse and increase network utilization

- Advantages: simple, effective, feasible, and independent of switch-level hardware support, transparent implementation without modifying any applications, standard protocols, network switches or VMs
Figure 8: High-level architecture of XCo.
Central Controller

- Resides in the same switched network as other nodes

- Takes as input:
  - Switch interconnection topology and link capacities
  - Location of VMs on physical nodes
  - Current traffic matrix of the network
  - Administrative policies

- Whenever detects congestion buildup at any link, computes and sends transmission directives to local coordinators at each end-host that is contributing to the congestion
Local Coordinator

- Intercepts and regulates the outgoing traffic aggregates (VM-to-VM flows) from all VMs within the corresponding end-host according to transmission directives.

- Provides traffic feedback to the central controller.

- The specific regulation pattern is dictated by transmission directives.
Transmission Directives

• Explicit instructions for transmission

• Various forms:
  ▪ Explicit timeslice scheduling
    *which V2V flow transmits when and for how long*
  ▪ Explicit rate limiting
    *at what rate a V2V flow should transmit for the next N ms*
  ▪ Combination of the above two or other forms
Explicit Timeslice Scheduling

(a) TIMESLICE SCHEDULING FOR FIGURE 1(a)

(b) TIMESLICE SCHEDULING FOR FIGURE 1(b)

Figure 9: Timeslice scheduling for Figure 1 setup. Value of 1 indicates that the corresponding sender is allowed to transmit during the timeslice.
Work Conservation

• Some nodes may finish early with their timeslice.

• Local coordinators return the remaining part of timeslice back to central controller.

• Central controller then permits another node to transmit.

• Local coordinators introduce a small hysteresis delay before returning the timeslice, in case that more packets might arrive during the delay.
Figure 1: Experimental setups: Multiple senders transmit to (a) one receiver via 1Gbps link, (b) different receivers via 10Gbps uplink.
Impact of Ethernet Congestion

Figure 2: Incast problem with iSCSI setup.
Performance Evaluation of XCo

Addressing the Incast Problem with iSCSI Setup
dd Of Striped 1.1 GB Data With Block Size 1MB Over 1Gbps Link

Received Throughput (Mbps)

Without Coordination
Rate Limiting
XCo Timeslice Scheduling

Number of iSCSI Servers

Figure 12: Addressing Incast for iSCSI in Fig.1(a).
Impact of Ethernet Congestion

Network Congestion Over 1Gbps Link
Five Senders -- K UDP Senders, (5-K) TCP Senders

Figure 3: Collapse at 1Gbps link in Fig 1(a).
Performance Evaluation of XCo

Network Congestion Over 1Gbps Link
Five Senders -- K UDP Senders, (5-K) TCP Senders, 1.5KB MTU

Received Throughput (Mbps)

Number of UDP Senders (K) (with (5-K) TCP senders)

- Without Coordination (Aggregate)
- Rate Limiting (Aggregate)
- XCo Timeslice Scheduling (Aggregate)
- Avg Per-flow TCP Throughput Without Coordination
- Avg Per-flow TCP Throughput With XCo Timeslice

Figure 13: Addressing collapse in Fig 1(a).
Impact of Ethernet Congestion

Network Congestion Over 10Gbps Link
Thirteen Senders -- K UDP Senders, (13-K) TCP Senders

Received Throughput (Mbps)

Number of UDP Senders (K)

- Total Received Throughput
- Avg. Per-flow TCP Throughput

Figure 4: Collapse at 10Gbps link in Fig 1(b).
Performance Evaluation of XCo

Network Congestion over 10Gbps Link
Thirteen senders - K UDP Senders, (13-K) TCP senders, 1.5KB MTU

Figure 14: Addressing collapse in Fig 1(b)
Experimental Setup

Figure 5: Topology for short TCP flow experiment.
Impact of Ethernet Congestion

Throughput Collapse Due to Short TCP Flows

Figure 6: Throughput collapse for a long-lived TCP flow due to multiple short-lived TCP flows.
Performance Evaluation of XCo

Improving Throughput With Short TCP Flows

Figure 15: Addressing collapse due to short TCP.
Live VM Migration

Live VM Migration Time During Congestion

Figure 17: Improving live VM migration time.
Fairness among V2V Flows

Fairness Across Competing V2V Flows
1Gbps Bottleneck Link, 3TCP and 2UDP Netperf Senders

Figure 18: Fairness among competing flows.
Work Conservation

Figure 19: Throughput of five competing UDP flows, with and without work conservation. Three UDP flows transmit at 50Mbps and two UDP flows transmit at the maximum available bandwidth.
Reference

• **XCo**: explicit coordination to prevent network fabric congestion in cloud computing cluster platforms

  Vijay Shankar Rajanna, Smit Shah, Anand Jahagirdar, Christopher Lemoine, and Kartik Gopalan

  *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, June 2010*