Congestion Management for Ethernet-based Lossless DataCenter Networks

Pedro Javier Garcia\textsuperscript{1}, Jesus Escudero-Sahuquillo\textsuperscript{1}, Francisco J. Quiles\textsuperscript{1} and Jose Duato\textsuperscript{2}

\textsuperscript{1}: University of Castilla-La Mancha (UCLM)
\textsuperscript{2}: Technical University València (UPV)

NENDICA
DCN: 1-19-0012-00-ICne
Abstract

This paper describes congestion phenomena in lossless data center networks and its negative consequences. It explores proposed solutions, analyzing their pros and cons to determine which are suited to the requirements of modern data centers. Conclusions identify important issues that should be addressed in the future.
Agenda

Introduction
Congestion Dynamics in DCNs
Reducing In-Network and Incast Congestion
Combining Congestion Management Mechanisms
Conclusions
Agenda

Introduction
Congestion Dynamics in DCNs
Reducing In-Network and Incast Congestion
Combining Congestion Management Mechanisms
Conclusions
Introduction

On-Line Data Intensive (OLDI) Services [Congdon18]

• Require **immediate answers to requests** that are coming in at a high rate.

• **End-user experience** is highly dependent upon the system responsiveness.

• The **network becomes a significant component** of overall DC latency when congestion occurs in the network.
Introduction
Data-Center Networks (DCNs)

• Todays DCNs require a **flexible fabric** for carrying in a convergent way traffic from different types of applications, storage of control.

• **Latency is a concern:** Fabric design for DCNs must minimize or **eliminate packet loss**, provide **high throughput** and maintain **low latency**.

• These **goals** are crucial for applications of OLDI, Deep Learning, NVMe over Fabrics and the Cloudified Central Offices.

• However, **congestion** threatens these applications.
Introduction

Why congestion isolation is needed?

- **HoL-blocking** dramatically degrades the network performance (e.g. PFC has not enough granularity and there is no congested flow identification) [Garcia05].

- **Classical e2e congestion control** for lossless networks is difficult to tune, reacts slowly, and may introduce oscillations and instability [Escudero11].
Introduction
Why congestion isolation is needed?
Introduction

Why congestion isolation is needed?

• We need a congestion isolation (CI) mechanism that **reacts quickly** when transient congestion situations appear, preventing network performance degradation caused by the HoL blocking.

• We want a CI mechanism that **complements other technologies** available in the DCNs, so that CI improves their performance, while the others reduce the CI complexity.
Agenda

Introduction

**Congestion Dynamics in DCNs**
Reducing In-Network and Incast Congestion
Combining Congestion Management Mechanisms
Conclusions
Congestion Dynamics in DCNs

Appearance of Congestion

Congestion (t0)

Congestion (t0+T)

Speedup = 2

Speedup = 1.5

Injection rate at 100% of the link bandwidth (full rate)
Congestion Dynamics in DCNs

Growth of Congestion Trees (from root to leaves)

Switch 1

Switch 2

Switch 3

Switch 4

Switch 5

Switch speedup = 1.5
Packet flows
Congestion point
Congestion Dynamics in DCNs
Growth of Congestion Trees (from leaves to root)
Congestion Dynamics in DCNs

Growth of Congestion Trees (Roots movement)

Switch speedup = 1.5
- Packet flows (start)
- Packet flows (after)
- Congestion point
Congestion Dynamics in DCNs
Growth of Congestion Trees (in-network roots)
Congestion Dynamics in DCNs

Growth of Congestion Trees (Overlapping)

Switch speedup = 1.5

- Packet flows addressed to X
- Packet flows addressed to Y
- Congestion point
Congestion Dynamics in DCNs

Growth of Congestion Trees (Vanishing)

Switch speedup = 1.5
- Red: Permanent packet flows
- Blue: Packet flows disappearing first
- Red dot: Congestion point first appeared in the switch
Agenda

Introduction

Congestion Dynamics in DCNs

Reducing In-Network and Incast Congestion

Combining Congestion Management Mechanisms

Conclusions
Reducing Congestion

Incast congestion reduction - ECMP
Reducing Congestion

In-network congestion reduction - ECN

Switch 1
Switch 2
Switch 3
Switch 4
Switch 5
Switch 6
Switch 7
Switch 8
Switch 9

Switch speedup = 1.5
- Packet flows addressed to X
- Packet flows addressed to Y
- Victim flow
- Congestion point

X
Y
Reducing Congestion

Limitations of current technologies [Escudero19]

- These technologies may work together to eliminate loss in the cloud data center network.
- **Load-balancing and destination scheduling** are end-to-end solutions incurring in the RTT delays when congestion appear.
- However, there is no time for loss in the network due to congestion and congestion trees grow very quickly.
- **Transient congestion may still produce HoL blocking** that leads to increase latency, lower throughput and buffers overflow, significantly degrading performance.
- Even using these mechanisms, **we still need something to deal with HOL Blocking locally and fast.**
Agenda

Introduction
Congestion Dynamics in DCNs
Reducing In-Network and Incast Congestion
Combining Congestion Management Mechanisms
Conclusions
Combining Congestion Management Mechanisms

- CI is needed to react locally and very fast to immediately eliminate HoL blocking.
- Previous technologies reduce the use of PFC and ECN, but their closed- and open-loop approach cause delays still happening.
- Congestion trees appear suddenly, are difficult to predict (even worse when load balancing is applied) and grow quickly.
- New techniques can be applied in combination to the previous technologies, improving their behavior.
Combining Congestion Management Mechanisms

Dynamic Virtual Lanes (DVL)

Switch A

Switch B

Legend
- Output port requested by the packet on top.
- Congestion root.
- Congestion Isolation Packets (CIP).
- Packets from congested flows.
- Packets from non-congested flows.
Agenda

Introduction
Congestion Dynamics in DCNs
Reducing In-Network and Incast Congestion
Combining Congestion Management Mechanisms

Conclusions
References


[Escudero11] Jesús Escudero-Sahuquillo, Ernst Gunnar Gran, Pedro Javier García, Jose Flich, Tor Skeie, Olav Lysne, Francisco J. Quiles, José Duato: Combining Congested-Flow Isolation and Injection Throttling in HPC Interconnection Networks. ICPP 2011: 662-672