Responses to Qcz analysis in March

Jesus Escudero-Sahuquillo (UCLM)
Pedro Javier Garcia (UCLM)
Francisco J. Quiles (UCLM)
Jose Duato (UPV)

Presentation

- **17 years of experience** in congestion management (from 2001).
- Authored **tens of research papers** on congestion management for lossless networks.
- We have also expertise in **network topologies, routing algorithms and quality of service** for lossless interconnects (QoS).
- Important references:

José Duato, Ian Johnson, Jose Flich, Finbar Naven, Pedro Javier García, Teresa Nachiondo Frinós: ***A New Scalable and Cost-Effective Congestion Management Strategy for Lossless Multistage Interconnection Networks****. HPCA 2005: 108-119*

Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato, Ian Johnson, Finbar Naven: ***Efficient, Scalable Congestion Management for Interconnection Networks****. IEEE Micro 26(5): 52-66 (2006)*

Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato: ***An Effective and Feasible Congestion Management Technique for High-Performance MINs with Tag-Based Distributed Routing.***
IEEE Trans. Parallel Distrib. Syst. 24(10): 1918-1929 (2013)

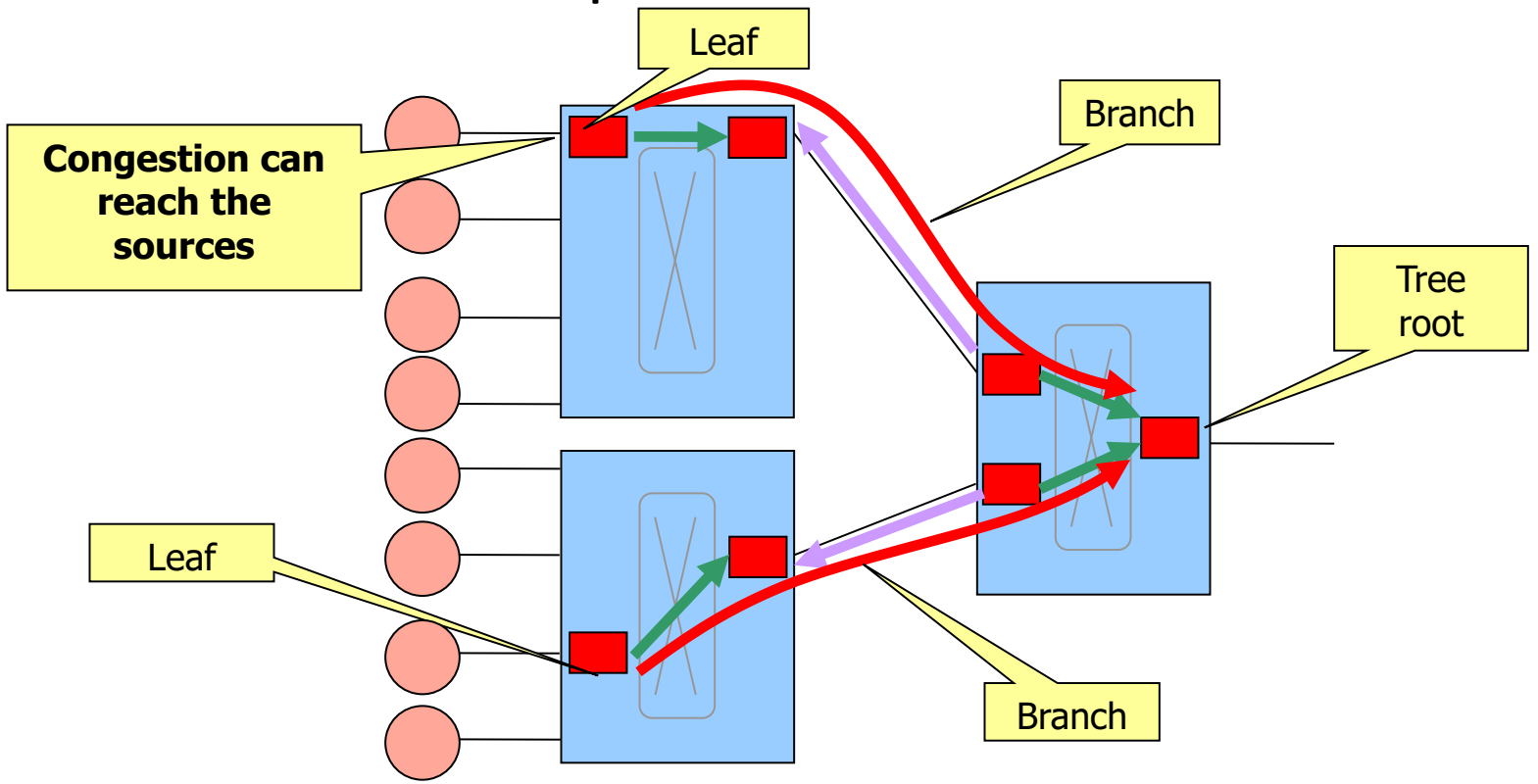
Jesús Escudero-Sahuquillo, Ernst Gunnar Gran, Pedro Javier García, Jose Flich, Tor Skeie, Olav Lysne, Francisco J. Quiles, José Duato: ***Efficient and Cost-Effective Hybrid Congestion Control for HPC Interconnection Networks****.*
IEEE Trans. Parallel Distrib. Syst. 26(1): 107-119(2015)

Agenda

- Initial comments
- Sampling based congested flow detection
- Timeout based transition from congested to uncongested flows
- PFC ingress thresholds need to be larger than CI egress thresholds
- Congested flow packets can be in noncongested queue when PFC is triggered
- Conclusions

Initial comments

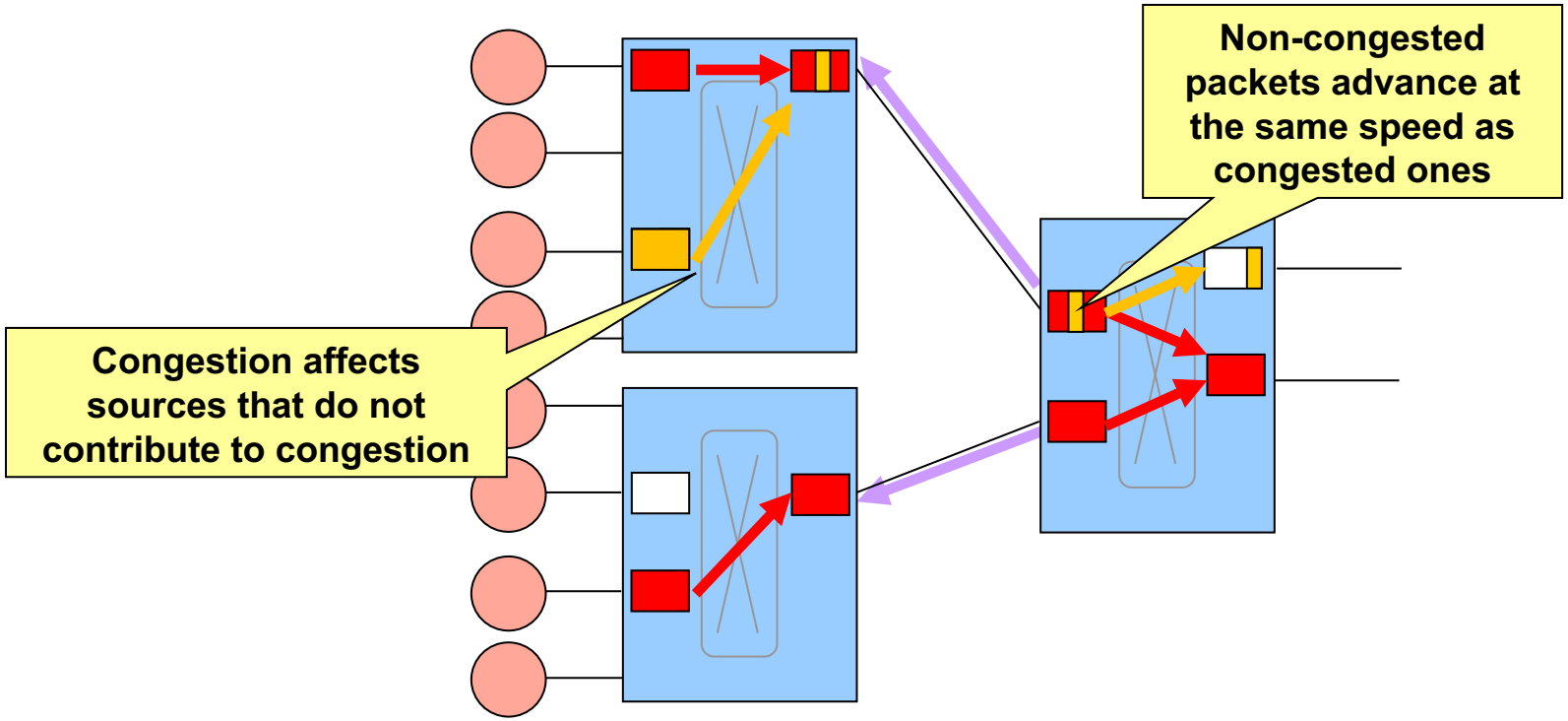
- **Congestion tree dynamics** need to be understood to design better congestion isolation techniques



Pedro Javier García, Jose Flich, José Duato, Ian Johnson, Francisco J. Quiles, Finbar Naven: ***Dynamic Evolution of Congestion Trees: Analysis and Impact on Switch Architecture****. HiPEAC 2005: 266-285*

Initial comments

- **Congestion trees** may cause Head-of-Line (HoL) blocking, the main negative effect of the congestion



Initial comments

- When appropriately designed, **CI strategies** are the most effective approach to **quickly set congested flows aside and eliminate HoL blocking**. Thus, they are very suitable for handling sudden bursts of short-lived flows.
- Additional queues are required to isolate congested flows. Eliminating congestion trees by means of **e2e congestion management is an effective way to deallocate congestion queues as soon as possible**, making them available to handle newly appearing congestion trees.
- There are several studies [1] on lossless networks reporting that **fairness can be achieved by combining e2e congestion management and CI.**

[1] Ernst Gunnar Gran, Eitan Zahavi, Sven-Arne Reinemo, Tor Skeie, Gilad Shainer, Olav Lysne: ***On the Relation between Congestion Control, Switch Arbitration and Fairness****. CCGRID 2011: 342-351*

Sampling based congested flow detection

Short-lived flows are not stopped effectively

- **Local isolation occurs immediately after detection**. CI is started faster than the closed-loop ECN congestion management reacts when a sudden burst of short-lived flows arrives (since it uses local resources).
- **CI is equally useful for long- and short-lived flows** since local isolation and notification propagation is faster than ECN-based congestion management, preventing HoL blocking and leaving congestion harmless.

Sampling based congested flow detection

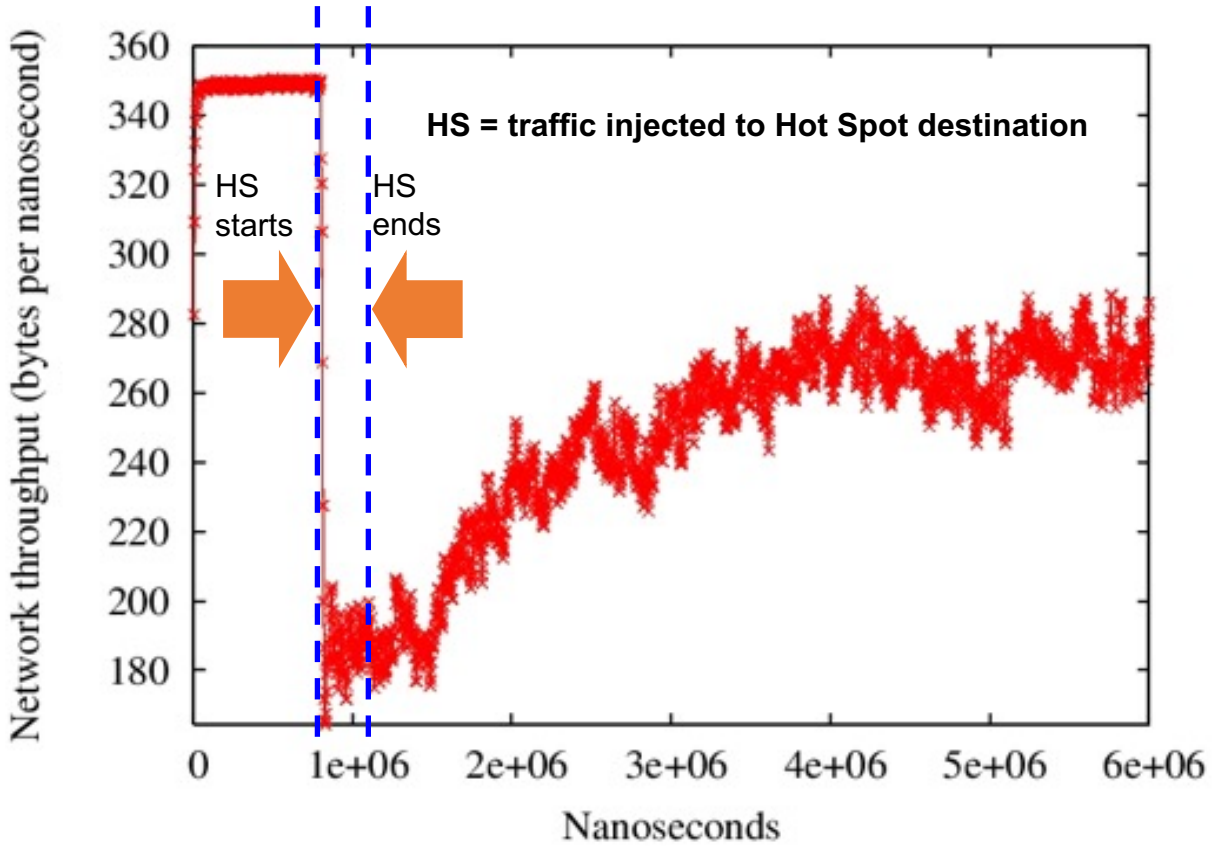
Lag in detecting flows that cause congestion

- If the detected flow is not the right one, the contributors will be identified later (in a fast way) by the detection mechanism.
- **False-positive detection** (i.e. non-congested flows in steady-state are impacted by congestion but chosen as contributors) is solved later with congested flows deallocation.
- However if **detection mechanisms at 802.1Qau** are assumed, we will use the available mechanisms for congestion detection.

Sampling based congested flow detection

Poor tail latency and FCT

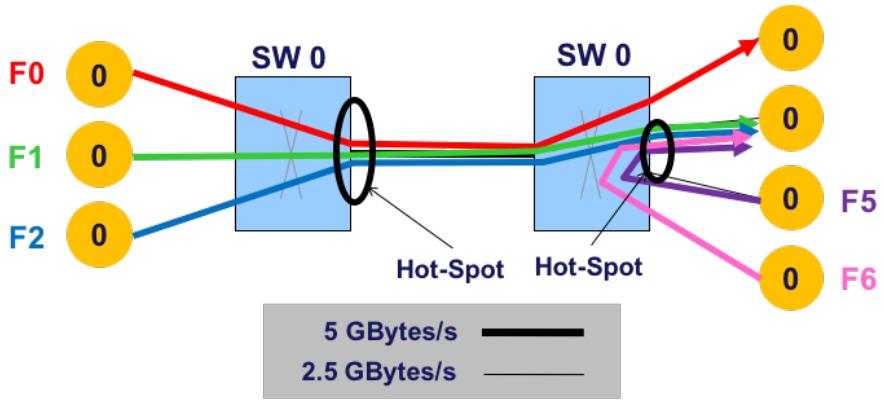
- **HoL-blocking dramatically** degrades the network performance (i.e. throughput and latency), according to simulations and real measurements, due to PFC has not enough granularity and no congested flow identification.

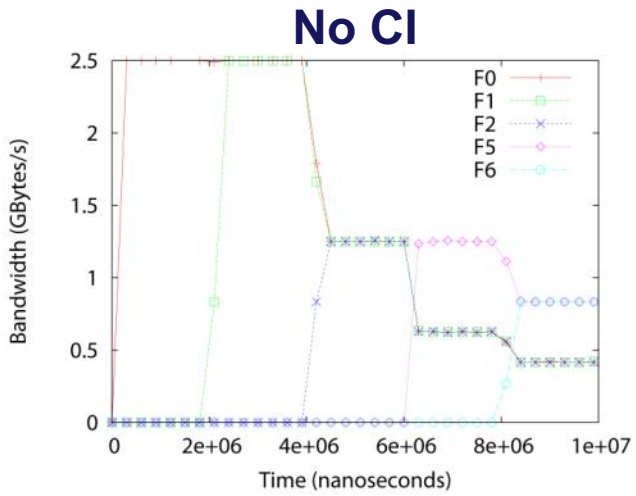


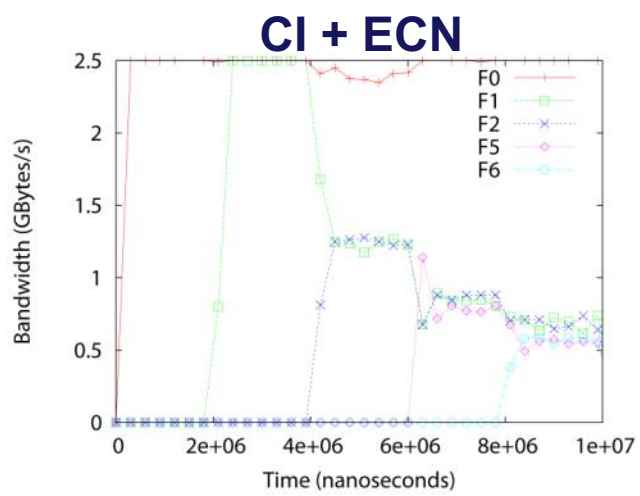
Sampling-based congested flow detection

Unfairness across flows

- **Congested packets** in congested queues **advance much slower** than non-congested packets (assuming a lot of congestion at destination endpoints). Is this unfairness or is it simply that no network with the same link bandwidth can do better?
- **E2e congestion management has been reported to improve the fairness** of elephant flows in lossless networks when local congestion isolation is used [1].







[1] 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

Timeout based transition from congested to uncongested flows

- Once e2e congestion management reduces injection rate for a congested flow and you start using round-robin among both queues, the **congested flow will eventually be drained and could be deallocated**.
- The described approach **does not introduce out-of-order delivery**. A suitable marking mechanism can be used to achieve this.
- It is **useful to deallocate the congested flow** because the associated congested flow table can be flushed and made available for future congestion scenarios.
- There are **several solutions** proposed to perform the deallocation.

Buffer Requirements

PFC ingress thresholds need to be larger than CI egress thresholds

- With a larger switch radix there will be more contention on the internal crossbar to reach the congested egress queue. Thus, **congestion queue allocation at the ingress side will trigger more frequently**, but this does not imply larger buffers.

Buffer Requirements

Congested flow packets can be in noncongested queue when PFC is triggered

- CI upstream switch **can send congested flow packets** after its congested queue is stopped.
- This situation is very transient **not affecting too much neither to latency nor buffer requirements.**
- There are **several solutions** to solve this issue that **guarantee that no packets are dropped**.

Final Remarks

- **Congestion Isolation (CI) mechanism quickly reacts locally**, avoiding short-lived flows to be delayed by long-lived flows that may likely be contributing to generate congestion.
- CI also **propagates congestion information** (CIMs) to upstream neighbors, who can also quickly isolate the congested flows.
- The **e2e mechanism drains the congested queues and adjusts the injection of long lived flows**, so that resources used by CI mechanism can be deallocated faster since congestion vanishes.

Responses to Qcz analysis in March

Jesus Escudero-Sahuquillo (UCLM)
Pedro Javier Garcia (UCLM)
Francisco J. Quiles (UCLM)
Jose Duato (UPV)