Congestion Spreading

The Dark Side of Link-Based Congestion Control

Pat Thaler
Agilent Technologies

Agilent Technologies

Modeling congestion and congestion control

Previous presentations to the CMSG have shown that flow control provides advantages

- **e.g. gupta_1_0704.pdf**
- **the modeled system was one switch with three source nodes and one sink node**
- **advantage shown was higher throughput and lower latency from head of transmit queue to receipt.**

The problem with focusing on such a limited scenario is that the same mechanism, flow control, provides different results in a slightly more complex network

The Congestion Spreading behavior of flow control reduces throughput for innocent flows when congestion occurs.

What is Link Flow Control?

A method of flow control where the receiver on a link controls its input by causing the transmitter on that link to stop sending all traffic when the receiver lacks receive resources and allowing the transmitter to send when the receiver has resources.

Link flow control usually disables or enables all traffic on a link. In some cases it may disable some classes of traffic while enabling others, but it never applies control based on the end-to-end path of the traffic.

Link flow control may be XON-XOFF (e.g. IEEE 802.3x) or credit based; that is, it may enable and disable the transmitter or it may inform the transmitter of the available receive resources.

What is Congestion Spreading?

Congestion spreading occurs when congestion (slowing of traffic due to lack of resources) on one path in the network slows traffic traveling on paths which have adequate resources.

Because congestion spreading allows a receiver to disrupt traffic entirely unrelated to the receiver it produces poor network performance which can be very erratic and hard to trace.

Congestion spreading is caused by the use of link flow control on switched networks.

Single-switch single-source scenario

A is device transmitting to multiple devices.

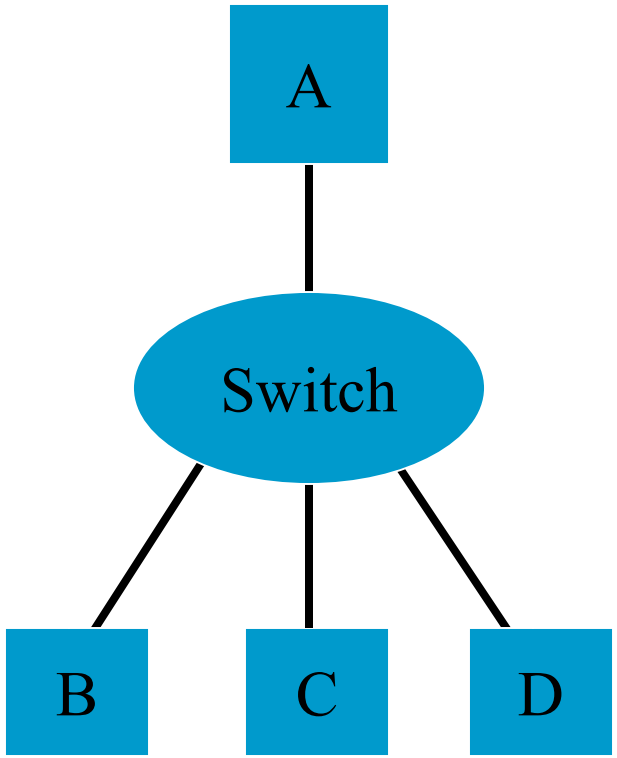
For instance, it may be a server on an Ethernet

A, B, and C are on 10 Gbit/s links and are capable of transmitting and receiving at link rate.

D is on a 1 Gbit/s link and can therefore operate at 10% of A's link rate.

Consider the case where A is transmitting to B and C. B and C will not exert flow control or do so only briefly and rarely. Therefore, traffic will flow through the switch without filling its buffers and A will be able to transmit at link rate.

A will transmit at link rate while B and C each receive traffic at about 50% link rate.



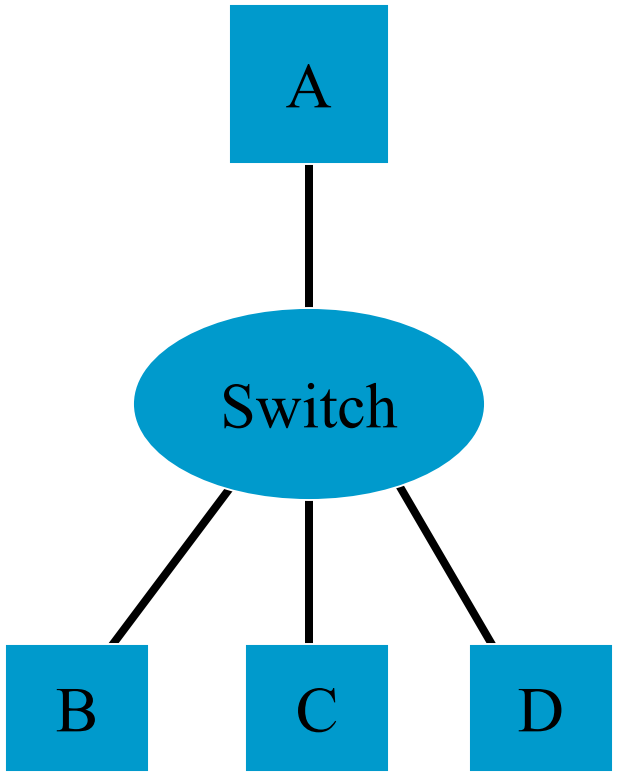
Single-switch source-server scenario

Now A begins to transmit to D as well. It sends approximately the same amount of traffic for each node.

Since frames for D flow into the switch at about 33% link rate and flow out at 10%, they consume switch buffers and soon the receive buffers for A's switch port fill.

The switch exerts flow control to A and traffic from A to all ports will stop.

As D deasserts and asserts flow control, the switch will control the traffic from A and transactions to B and C will be slowed to about 10% link rate.

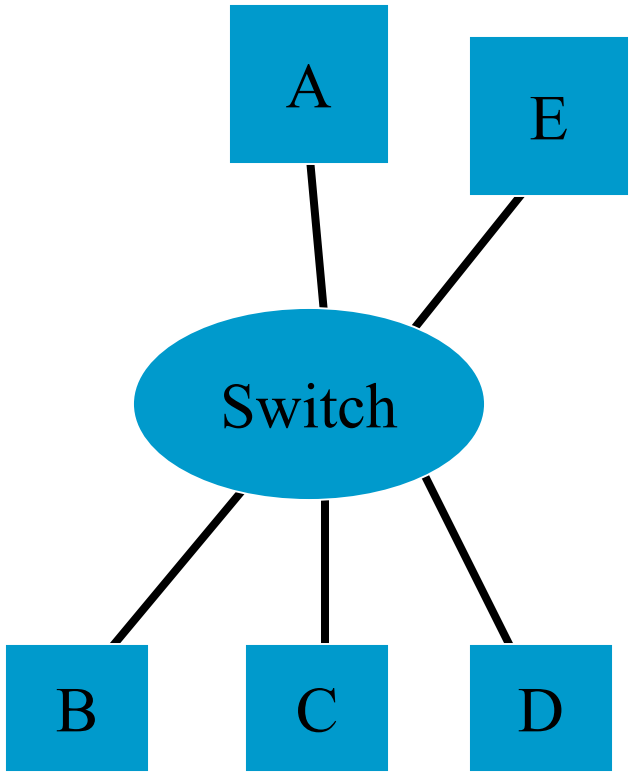


Single-switch multi-source scenario

We might make A or D a bit smarter. D might realize it is a slow device and negotiate shorter transactions to reduce the chance of filling the switch buffers. Or A might realize that D is a slow device type and reduce its transaction size.

However, multiple processes on A transmitting to D or transmissions to D from other nodes such as E may still congest D resulting in flow control to A and throughput reduction to B and C.

D may not even be a slow device. It may be a device receiving many data streams.



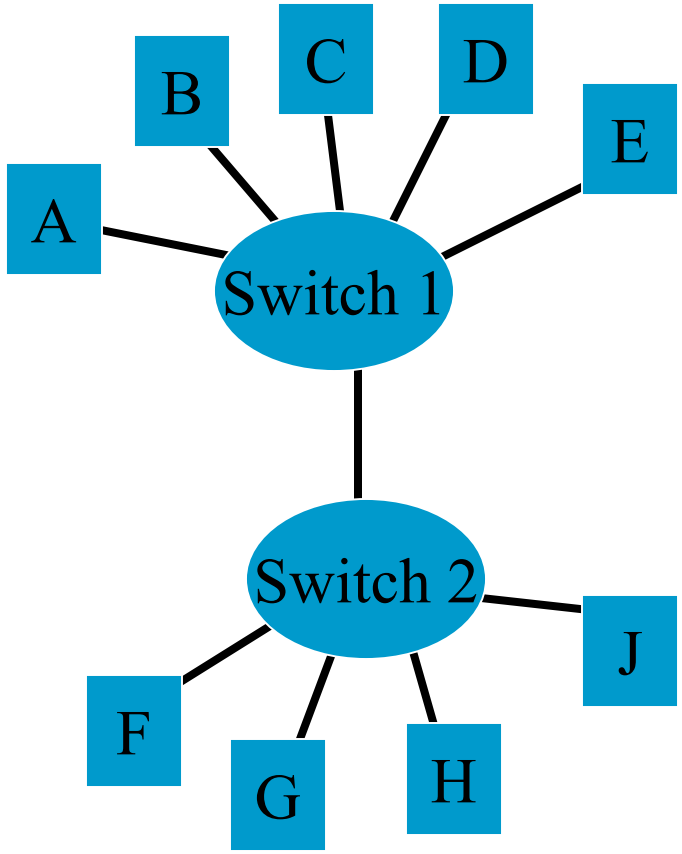
Multi-switch scenario - fabric congestion

As our networks grow, the potential for congestion spreading increases.

The source of congestion spreading may be an overloaded fabric element rather than an end node.

A is sending to E and to H. Nodes A through E are sending more traffic to nodes F through J than the inter-switch link can handle so switch 1's buffers fill and flow control is exerted to A.

Traffic from A to E is now slowed even though adequate bandwidth is available.



Agilent Technologies

Multi-switch scenario

The multi-switch scenario also demonstrates the ability of link flow control to spread congestion to an unrelated source.

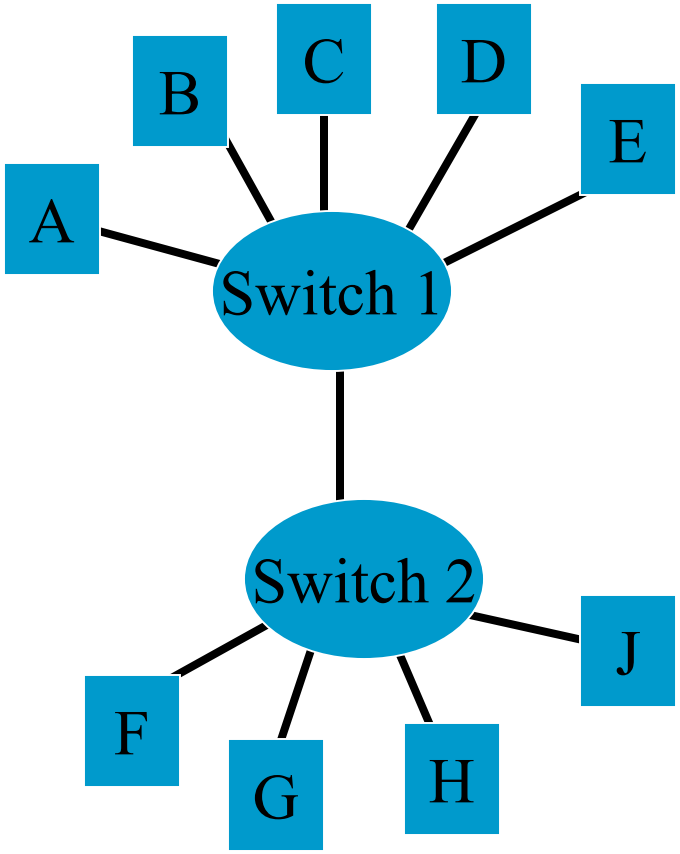
A is transmitting to F and both are able to transmit and receive at the same speed.

B is transmitting to at a higher rate than G can receive.

G exerts link flow control (or G's link rate limits flow), Switch 2's buffers for the inter-switch link fill. Switch 2 exerts flow control to switch 1.

Switch 1's buffers fill and it exerts flow control to users of the inter-switch link, A and B.

Traffic from A to F is slowed despite the presence of ample bandwidth on every path to which A transmits.



What does this mean for CMSG?

The draft 5 Criteria for CSMG states:

- **Simulations have shown reduced latency and increased throughput, which improves the overall performance of Ethernet.**

This statement is only true for an extremely narrow case. A broader simulation would show decreased throughput and increased latency for some traffic flows.

Therefore, current CSMG work has not shown been broad enough to demonstrate technical and economic feasibility of congestion management at the Link layer.

More work is necessary to demonstrate that the 5 Criteria have been met.