A Pause Link Configuration

Ariel Hendel

hendel@eng.sun.com

Sun Microsystems

IEEE 802.3z

28-January-1997

Scope of this presentation

- Propose how to configure Asymmetric Flow Control, considering:
  - **End Station**
  - **Infrastructure (Buffered Distributor/Switch side)**
- Assuming 802.3z adopts AFC
- this presentation does not judge the merits or need for Asymmetric Flow Control

Observations

- End station vs. infrastructure independence
  - **We (802.3z) look at the link end points generically**
  - **In practice there is a division between end stations and infrastructure that reflects:**

    a. different functional needs

    b. distinct administrative boundaries

    c. different cost tradeoffs
- End stations have a variety of performance profiles
  - **An End station network performance might be a function of:**

    a. H/W platform and OS

    b. Time in product life cycle

    c. Cost tradeoffs

    d. Variations in traffic models

Worthy Objectives

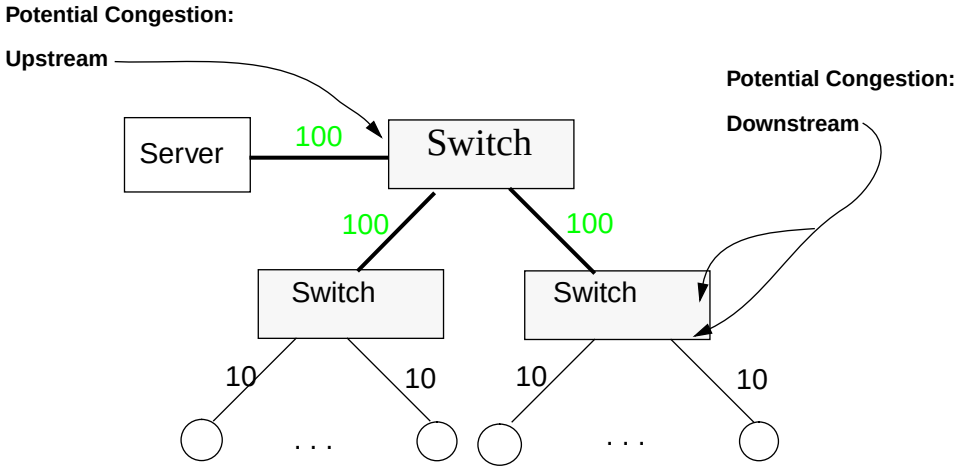
- Plug-and-Play link properties, while:
  - **Respecting the End station vs. Infrastructure independence**
  - **Accommodating different performance profiles**

For example, infrastructure deployment is typically uniform and independent of the specific End station (machine, OS, NIC architecture, etc.). Evolution and upgrades on either side of this divide should be independent, except for the nature of the link itself (100BASE-T, 100BASE-SX, etc.).

Link speed choice might be justified by latency, future-readiness, or uniformity, while the End station is not necessarily optimized, at every point in time, for worst case conditions. Do not impose an “all or nothing” restriction on End station performance.

802.3x Flow Control Uses

- Originally conceived as a (controversial) weapon to fight congestion



802.3x Flow Control Uses (cont.)

- And lately became an enabler for Buffered Distributors...
- Along came the concern about End stations blocking such Distributors

So, the question is, when does it makes sense to emit PAUSE XOFF, and how does a node know?

It certainly makes sense when the receiver buffers are smaller than the sender buffers (and neither the link nor the sender buffers are shared with other receivers).

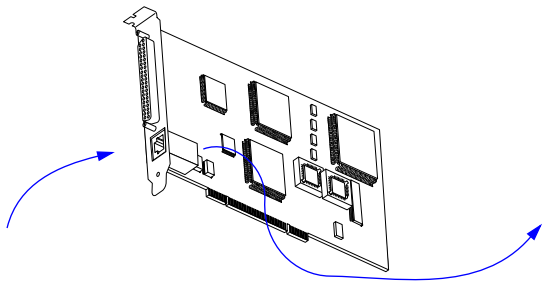
We claim that this is in many cases true for an End Station to Switch Gigabit link, but certainly not the case for an End Station to Buffered Distributor link.

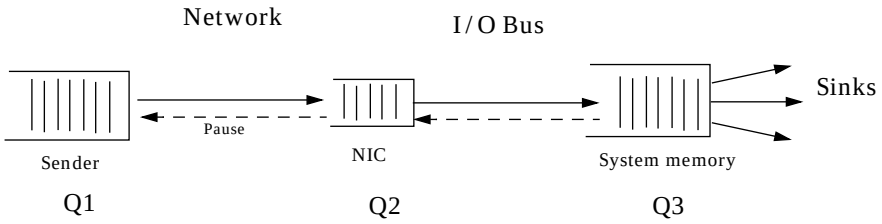
End Station PAUSEs

By Pausing the Sender, the receiver appears to have extended its receive queue into the Sender's queue.

No side effects as long as the Data path is dedicated and the sender buffers are dedicated to this link.

Pause is only defined for Full Duplex links, therefore the Data path is dedicated by definition.





How is a NIC to know?

Both sides know the other side can receive PAUSE because 802.3x is mandatory for Gigabit.

An End station interface knows whether it may benefit from sending PAUSE based on:

1. Its own architectural constraints
2. The system environment it is in

It just doesn't know who is on the other side.

The infrastructure side knows if receiving PAUSE makes sense based on:

1. Its own nature and architecture
2. Possibly configurable preferences

A single new bit seems appropriate to convey this information to the End station:

DP = Don't Pause.

The DP bit - Semantics and motivation

Exchanged during link configuration	**Plug-and-Play**
Not part of the arbitration function	**Capabilities do not need to match for link to come up**
Senders must comply with PAUSE regardless of their DP value	**Allows cascaded Buffered Distributors.** **Prevents End station blocking Buffered Distributors.**
Receipt of the DP bit set is interpreted as: Don't send PAUSE unless you really have to	
End Stations observe the link partner DP bit to determine whether to emit PAUSE	

DP bit usage examples:

End Stations advertise DP clear
Buffered Distributors advertise DP set
Buffered Distributors emit PAUSE regardless of the link partner DP
Some switches may set DP (switch dependent)

Summary of Proposal

- Define a single bit (DP) to convey the nature of the link partner as it pertains to PAUSE.
- Exchange DP during link configuration
- Sender must still respect PAUSE per 802.3x
- Use the DP bit to enable plug-and-play behavior when interconnecting Gigabit devices.
- Link does not need to be manually configured to advertise different capabilities when connected to different type of devices