

# A Localized Congestion Control Mechanism for PCI Express Advanced Switching Fabrics

Venkata Krishnan and David Mayhew  
Stargen Inc.

Marlborough, MA 01752.

[krishnan, mayhew]@stargen.com

<http://www.stargen.com>

## Abstract

*Even though there is a commonality in their physical and link layers, Advanced Switching (AS) is more than merely an extension of PCI Express. The role of PCI Express is primarily as a chip-interconnect. In contrast, AS functions as a true system fabric interconnect encompassing domains that are vastly different and indeed more sophisticated than that of PCI Express'. In such a setup, congestion management is crucial for optimal utilization of the fabric bandwidth.*

*Congestion control in network fabrics has generally been based on end-to-end and link-by-link schemes for controlling packet injection. Though seemingly adequate, these schemes may not be effective in handling "transient" congestion arising from intermittent traffic – for transient congestion can indeed occur in fabrics operating well below their saturation limit. In such scenarios, a link-by-link scheme does not prevent congestion from spreading while an end-to-end scheme may result in under-utilization of the fabric bandwidth.*

*This paper details a congestion control mechanism called Status Based Flow Control (SBFC) that has been incorporated into AS and is specifically targeted for alleviating the transition congestion problem. The SBFC mechanism exploits the source-based path routing used in AS. It enables upstream switch nodes to modify the transmission of packets based on the congestion status of links in a downstream switch. Simulation studies show that the SBFC mechanism indeed permits optimal usage of the fabric bandwidth during periods of transient congestion and effectively complements the traditional end-to-end and link-by-link schemes for congestion control.*

## 1 Introduction

PCI Express, a low-latency, high-bandwidth switched serial interconnect, is poised to succeed PCI as the next generation intra-system interconnect[1]. The primary strength behind PCI Express is in its support for legacy PCI while addressing the inadequacies of PCI. Indeed, PCI Express is fundamentally nothing more than a serialization and packetization of PCI and is completely compatible with PCI. The additional features that PCI Express offers over PCI are in addition to and not a replacement of the mechanisms supported by PCI. Hence, the reams of existing software that expect PCI to underpin them will continue to function untouched in a PCI Express world. However, PCI Express' greatest advantage - its strict compatibility with PCI - is also one of its greatest liabilities. Since a PCI Express fabric spans a single global address space, there is no notion of a system boundary. In other words, a PCI Express fabric is, by definition, a *single* system - wherein multiple hosts cannot share a fabric. Since all communication is under the control of a single host, PCI Express is not well suited for an important application space that includes multiprocessing and peer-to-peer communication.

Advanced Switching (AS)[2, 11] builds upon PCI Express by using the physical and link layer while at the same time providing capabilities that include support for multiprocessing and peer-to-peer computing. It is our belief that the evolutionary path afforded by PCI Express from PCI and the layer commonality between Advanced Switching (AS) and PCI Express would indeed enable AS to become a dominant inter-system interconnect technology in the near future.

In a multi-stage interconnect such as AS, congestion avoidance and control[3] is essential for optimal usage of fabric bandwidth. Traditional congestion avoidance techniques include (a) link-by-link based and (b) end-to-end approaches. A common link-by-link scheme is the credit-based windowing approach[5]. Here, there is exchange of credit (or storage information) between nodes on both sides of the link. The sender must have the requisite credits before it can begin transmitting on its output link. Periodically, the receiver replenishes the credits. What this implies is that at least the specified buffer space (in terms of credits) is available on the input side of the link. Overall, this scheme is primarily used for avoiding packet drops typically seen in TCP/IP environments. This prevents unnecessary retransmission of packets thereby avoiding the infamous congestion collapse[6]. On the other hand, this scheme does not prevent congestion from spreading. If the incoming traffic exceeds the targeted output link bandwidth, the senders will soon lack the requisite transmission credits and would be blocked from sending additional packets. This blocking not only affects flows targeting the congested link but also penalizes flows targeting other links.

In an end-to-end scheme, end nodes control the rate of traffic entering the network. A common approach is to send an explicit congestion notification (ECN)[7] to the source node whenever an intermediate switch encounters congestion. ECNs may either be forward ECN (FECN) or backward ECN (BECN)[8]. A BECN requires the switch to generate a source notification packet on encountering congestion. Alternatively, the switch may play a passive role and simply mark the congested packets en-route to their destination. This is the FECN mode of congestion notification. It is the responsibility of the upper-level protocols on the destination node to take appropriate action on receipt of such marked packets. The action includes sending a notification to the source. Finally, the source node generally responds to a congestion notification by adjusting its packet injection rate by using the classic additive increase-multiplicative decrease (AIMD) algorithm or one of its variants[9]. Its complexity notwithstanding, the ECN approach enables the source node to adjust its injection rate in accordance with the offered load of the fabric. Nevertheless, ECN based mechanisms employing the AIMD approach result in under-utilization of the fabric bandwidth[20, 21] and may not be appropriate during periods of transient congestion.

Indeed, transient congestion will be a common occurrence even in systems that operate below the saturation bandwidth. This is due to network traffic being extremely bursty and not necessarily following a Poisson model[10]. Since the sender uses the same link











