Dynamic Buffer Pool

White Paper

October 2007

Website: www.plxtech.com
Technical Support: www.plxtech.com/support

Dynamic Buffer Pool Description

Available on-chip memory utilization and architecture are critical in determining the performance and throughput of a switch. When architecting a switch, one can choose between fixed and dynamic buffer memory allocation. PLX's doubly bifurcatable port architecture (x16 to two x8 or four x4s) both provides highly flexible port configurations and lends itself to dynamic buffer allocation. Dynamic buffer allocation makes best use of available switch memory, leading to higher performance and/or lower cost.

In our switches, the initial flow control advertisement is designed to sufficiently maintain full wire speed throughput with full peer-to-peer traffic patterns of the usage model. The rest of the memory and corresponding flow control credits are kept in a reserve pool for use in case of congestion. Typically, between 3 and 4 maximum payloads of payload credit are required to maintain flow-through, more or less according to the DLL latencies of the link partner. The initial credit advertisement may represent only half the buffer memory available to the ports that share the buffer pool.

One of the advantages of dynamic buffer allocation is faster credit update, lowering the amount of credits required to maintain flow-through, in turn lowering overall switch buffer memory requirements. The reserve buffer pool is drawn down only when a backlog, the first sign of congestion, develops in the switch. So long as it isn't empty, a credit update may be returned to a link partner as soon as the header of a packet is seen at the ingress of the switch. If the policy were to hold no credits in reserve, the credit consumed by a packet ingress to the switch could only be restored to the link partner after an ACK was received from some other link partner of the switch to which the packet had been forwarded. Since ACK latencies are themselves several packets long at a minimum this policy at least doubles the amount of outstanding flow control credit required to maintain flow-through without providing any flexibility towards buffering transient congestion.

In well behaved applications, all the traffic is uniformly distributed with continuous rather than bursty flows. Switches for such applications don't need congestion buffers. However, these well behaved applications are hard to find. In the real world of PCIe fabrics, transient congestion is a frequent occurrence. Switches must be judged on their ability to maintain high throughput despite non-uniform, bursty traffic flows that frequently cause backlogs to develop in the switch.

To deal with this condition, switches provide extra buffer memory beyond that needed to support their flow control policies. In the fixed allocation architecture, that excess memory is evenly distributed across all the ports. In PLX's Dynamic Buffer Pool architecture, a sharable pool is provided for each group of 4 ports. Thus, for the same amount of ***excess*** buffer memory, a port in the PLX architecture has four times as much memory available for transiently buffering packets from a single ingress port to be forwarded towards the backlogged port. While this advantage evens out if all ingress ports have traffic simultaneously for backlogged ports, the non-uniformity and burstiness of real world traffic suggests that this almost never occurs. Remember that the fixed

allocation architecture typically requires twice as much memory before any more counts as ***excess***.

Conclusion

Because of the aforementioned reasons, and because of buffer sharing, one should always expect the Dynamic Buffer Pool architecture to have a significant advantage in this regard, given an equal amount of on-chip buffer memory. This has been demonstrated in numerous customer benchmarks with our Gen 1 8500 ExpressLane™ family of switches and will be demonstrated again with our Gen 2 8600 family of switches.