Lookahead Packet Scheduling Algorithm for CIOQ DataCenter Switch Fabrics

EE384Y Packet Switch Architectures II

April 18 2003

Deepak Kakadia, Student ID: 4358289

Introduction

A Typical Datacenter N-Tier network architecture is shown below in Fig. 1. Current technology DataCenter Networking Equipment building blocks are composed of established high volume, optimized layer 2 and layer 3 packet switches and relatively new, mostly startup produced appliances for more complex ip services such as ssl, xml, url switching, nat etc. The appliances have evolved from functions that were previously performed on general purpose computers

The next logical evolution step is to converge and integrate these 2 product families to produce a device that is cost effective and optimized particularly for the data center traffic patterns . We know some things about the traffic patterns, which can be exploited to increase throughput by grooming traffic to avoid future contention on input and output resources in the switch fabric, keeping queues more evenly loaded over time. The Datacenter Edge traffic flows have some properties which can be used to allow us to make better scheduling decisions in order to increase overall throughput.

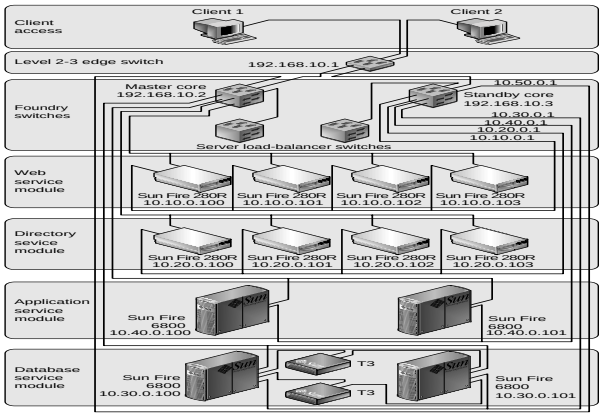

Fig.1 Typical Datacenter Edge Network Architecture

The distributed appliance and switch architecture requires islands of independent serial processing. For example, usually client traffic is first diverted to a NAT function, which then rewrites the packet and points to another ip device, such as a load balancer which again rewrites the packet. We can exploit this knowledge of the fixed set of services that are to be performed on this flow, to create one integrated device, perform 1 packet classification lookup, findout all the services to be performed, determine also in what sequence, then make more intelligent packetscheduling decisions, which proactively schedule the packet throught the fabric to prevent future congestion in the switch fabric. This is different from the first stage of the multistage switches of [7], where in this case we are making scheduling descisions now, which will impact the arrival traffic in future time slots. Fig. 2 below describes an example architecture, where the services to be performed are directly connected to the bidirectional ports of a switch fabric. In a practical example, we would have an SSL accelerator asic such as CAVIUM Nitrox asic which has a flow thru architecture, where packets are recieved via a SPI4.2 interface in one port, and either encrypt or decrypted traffic emerges out the other port which is connected to the same switch fabric port but of opposite direction.

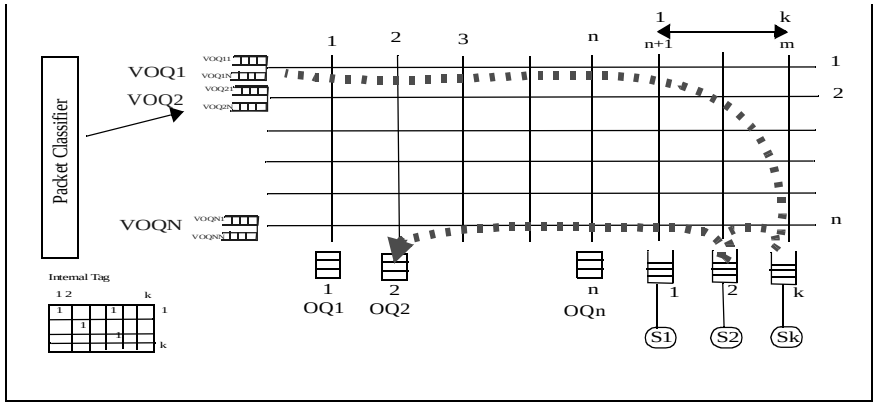

Fig. 2 Integrated Services and switch fabric example architecture. Packet classifier determines apriori all services that are to be performed on packet and hence has more information to make better scheduling decisions to groom traffic that re enters the fabric.

Fig. 3 below describes how the lookahead information, provided by the packetclassifier as a tag prepended to the packet header, can be used by the Lookahead based packet scheduler to make a better scheduling descision than Maximum Weight Matching. Suppose Packet 1 would require services at port 4,5, and Packet 2 would require services at port 4 and 6. Port 5 already has packets queued up to keep it busy. If packet 1 is chosen before packet 2, then there is a chance of an idle port 6, whereas if packet 2 was chosen, there would be less chance of an idle port, hence increasing throughput.

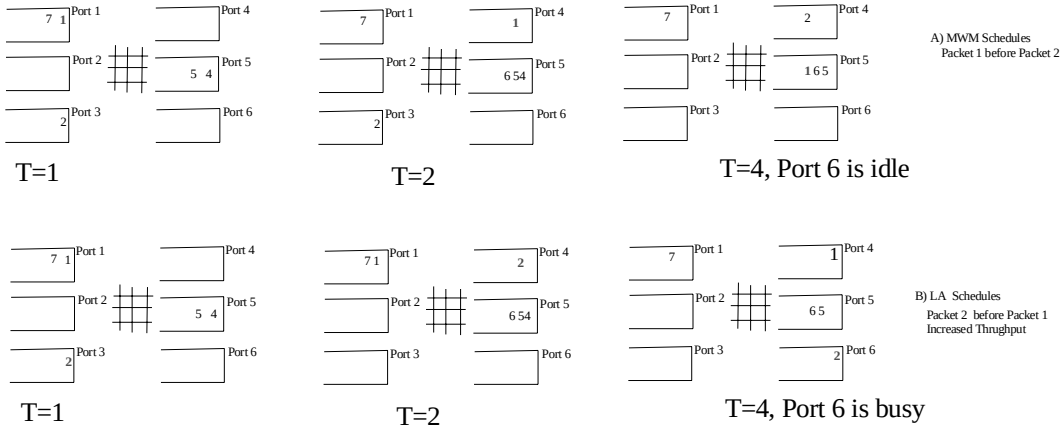

Fig 3 Lookahead Algorithm over Maximum Weight Matching, using Lookahed Tag

Proposed Problem

1- Describe the Lookahead packet scheduling algorithm and determine stability, throughput and average delay anal ysis

2 - Perform simulations and compare with MWM and other related algorithms.

References

[1] N.McKeown, M.Izzard, A. Mekkittikul, B.Ellersick and M.Horowitz," The Tiny Tera: A Packet Switch Core", Hot Interconnects V, Stanford University, August 1996

[2] N.McKeown, V.Anantharam and J.Walrand, 'Achieving 100% throughput in input queued switches", IEEE INFOCOM '98 p.792-799 1998.

[3] A.K Parek and R.G.Gallager, " A Generalized Processor Sharing Approach to Flow Control in Integrated Services Packet Networks:The Single Node Case, IEEE/ACM Transactions on Networking, vol.1, No.3, pp.344-357, June 1993.

[4] A.K Parek and R.G.Gallager, " A Generalized Processor Sharing Approach to Flow Control in Integrated Services Packet Networks:The Multiple Node Case, IEEE/ACM Transactions on Networking, vol.2, No.2, pp.137-150, April 1994.

[5] Shah D., Kopikare M. "Delay Bounds for approximate Maximum Weight Matching Algorithm for Input Queued Switches", IEEE INFOCOM 2002, New York, NY June 23-27 2002 http://www.ieee-infocom.org/2002/technical_programs.htm.

[6] Leonardi E., Mellia M., Neri F., Ajmone Marsan M., "Bounds on Average Delays and Queue Size Averages and Variances in Input Queued Cell Based Switches", IEEE INFOCOM 2001, Alaska April 2001, pp1095-1103.

[7] C.S Chang, e. al "Load balanced Birkhoff von Neuman switches, part II: Multistage buffering " http://www.ee.nthu.tw/~cschang/PartII.ps.