Performance Analysis of Resilient Packet Rings with Single Transit Buffer

Fengjie Yuan, Changcheng Huang
Department of Systems and Computer Engineering, Carleton University
fj_yuan@hotmail.com, huang@sce.carleton.ca,

Harry Peng, John Hawkins
Nortel Networks
{hpeng, jhawkins}@nortelnetworks.com

Abstract¾ Resilient Packet Ring (RPR) is a new technology to be standardized as IEEE 802.17 and several proposals have been made to the IEEE 802.17 working group. In this paper we will investigate the performance of one of the proposals, which features a single transit buffer, utilization-based congestion detection, and congestion notification with explicit rates. The key benefit of this proposal, known as 1TB-RPR (1 transit buffer RPR), is that it is extremely simple to implement in hardware. In addition to its simplicity, we will show that 1TB-RPR also performs extremely well, and can be considered as an excellent candidate for 802.17 standard. The two key performance metrics for this ring technology are throughput (or utilization) and ring access delay. In this paper we will show that 1TB-RPR can achieve as high as 95% utilization with a very low ring access delay.

Index Terms¾ Resilient Packet Ring (RPR), IEEE 802.17, Metropolitan Area Networks (MAN), Synchronous Optical Network (SONET) /Synchronous Digital Hierarchy (SDH).

1. Introduction

Resilient Packet Ring (RPR) [1, 2] is a new data transport technology designed to meet the new requirements of Metropolitan Area Networks (MAN). Traditionally, MANs have been designed for voice traffic using Synchronous Optical Network (SONET)/Synchronous Digital Hierarchy (SDH). With the growth of Internet applications, network services have become increasingly data-centric. The traditional solution that maps data packets into circuit-switched networks is costly and inefficient from both the carrier and end-user points of view. In order to optimize transport networks to handle the increase in data traffic, a new Media Access Control (MAC) layer protocol – Resilient Packet Ring (RPR) is under development. RPR shares SONET's ability to provide fast recovery from link and node failures, but also benefits from Ethernet's cost and simplicity. Furthermore RPR provides a fairness and congestion control mechanism that has not been addressed by either SONET or Ethernet, which allows RPR to be significantly more efficient than either technology. Several vendors are shipping and continue to develop RPR products. The Institute of Electrical & Electronic Engineers, Inc. (IEEE) has set up the 802.17 RPR working group to standardize this technology.

A RPR network, like SONET/SDH, is a ring-based architecture that consists of two counter-rotating rings with each station connecting to two adjacent stations over a link pair. RPR has also generalized the spatial reuse concept that can be found in SONET BLSR to allow low priority traffic to use the redundant capacity otherwise reserved for protection. A typical RPR node architecture is shown as in Fig 1. An RPR node has three kinds of traffic, namely add-in traffic, drop-off traffic and pass-through traffic. It is well known that statistical multiplexing can increase the overall efficiency of a transport system by accommodating multiple bursty traffic streams. The key difference between SONET/SDH and RPR is that an RPR node employs statistic multiplexing at the packet level. As shown in Fig.1 the add-in traffic stream is statistically multiplexed with the pass-through traffic stream yielding a higher utilization of the fiber link.

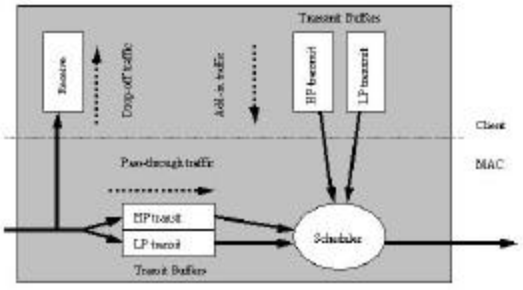


Fig.1 A typical RPR node

Since add-in traffic may compete with pass-through traffic for access to the ring's bandwidth when a downstream link becomes congested, buffers are needed to avoid temporal overflow. A RPR node uses the so-called transmit buffer and transit buffer to hold the bursts of add-in traffic and pass-through traffic respectively when congestion happens. Both transmit

buffer and transit buffer can be further partitioned to hold traffic flows with different priorities, e.g., one for high priority (HP) traffic and one for low priority (LP) traffic. A scheduler is required to decide which traffic stream can claim the bandwidth at a particular moment. When the pass-through traffic flows are given priority over the add-in traffic flows, add-in traffic flows will be temporally blocked and therefore suffer the so-called ring access delay. This is analogous to a collision in Ethernet terms. But unlike Ethernet, a simple yet intelligent bandwidth and congestion management scheme within the RPR MAC results in significantly more efficiency than Ethernet. Several proposals have been made to the IEEE 802.17 WG on how the bandwidth and congestion management scheme should be implemented. The proposals differ in the number of transit buffers being implemented, the way congestion is detected and controlled [3, 4, 5]. In this paper we will investigate the performance of one of the proposals originally proposed by Nortel Networks [6]. This proposal features a single transit buffer, utilization-based congestion detection, and congestion notification with explicit rates. The key benefit of this proposal, known as 1TB-RPR (1 transit buffer RPR), is that it is extremely simple to implement in hardware. In addition to its simplicity, we will show that 1TB-RPR also performs extremely well, and can be considered as an excellent candidate for 802.17 standard. The two key performance metrics for a ring technology are throughput (or utilization) and ring access delay. In this paper we will show that 1TB-RPR can achieve as high as 95% utilization with a very low ring access delay.

A significant amount of research has been done on the issues of access delay, utilization and fairness for legacy MAC protocols such as high-speed bus networks of the past [7, 8, 9, 10]. But since RPR is a brand-new concept, it has been rarely studied. We believe that this paper is one of the earliest research works to be conducted on RPR.

The paper is organized as follows. We begin with an introduction to the 1TB-RPR proposal in section 2. We then move on to simulation results in section 3. We will examine the access delay, utilization, and fairness issues through these simulations. Section 4 concludes the paper.

2. The 1TB-RPR proposal

A typical 1TB-RPR node is shown in Fig.2. The 1TD-RPR scheduling algorithm is based on Buffer Insertion Ring (BIR) technology [11] where pass-through packets always have priority over add-in packets from the transmit buffers. When the transit buffer has traffic to send, add-in packets will be queued in the transmit buffer until the transit buffer is emptied. Because the pass-through traffic has absolute priority over the add-in traffic, only a very small transit buffer (2 or 3 packet sizes) is required. This significantly simplifies the hardware implementation of the MAC. To reduce the ring access delay, a fairness algorithm based on feedback control is designed to control the access of the total bandwidth for all nodes during periods of congestion. The 1TB-RPR-fairness algorithm uses explicit congestion notification to manage bandwidth on the ring so that a weighted bandwidth fairness is achieved. The weight assigned to a node represents how much bandwidth the node requires for low priority traffic during periods of congestion. A topology discovery process ensures that each station knows the weights of all other stations on the ring.

To detect congestion, the fairness algorithm uses two trigger conditions: one triggered by high utilization, and one triggered by high ring access delay. The utilization is estimated using a weighted sliding-window estimator. The equation for the estimated throughput is as follows:

$$ESTIMATEDrate(t) = ESTIMATEDrate(t-1) - \frac{ESTIMATEDrate(t-1)}{WEIGHT1} + \frac{CURRENTrate}{WEIGHT2}$$

where *ESTIMATEDrate(t)* is the current estimated "rate" in bytes, *WEIGHT* is an integer value of: 2, 4, 8, 16, ……128, and *CURRENTrate* is a sampled rate measured in bytes. It should be noted that *WEIGHT* here decides how fast the *ESTIMATEDrate(t)* is updated and is totally different from the weights assigned to each node for fairness purposes.

Fig.2 shows a typical 1TB-RPR-node structure in which scheduler chooses data packets from five queues:

- Packets from transit buffer.
- Packets from expedited forwarding (EF) class transmit buffer.
- Packets from assured forwarding (AF) class transmit buffer.
- Packets from in-span best effort (BE) class transmit buffer.
- Packets from out-of-span best effort (BE) class transmit buffer.

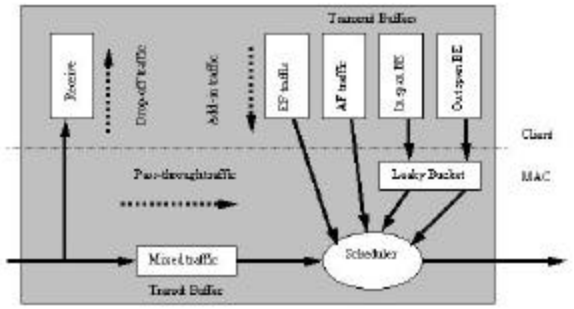


Fig.2 A typical 1TB-RPR node

EF class and AF class traffic is engineered based to a Committed Information Rate (CIR) and therefore their performance is guaranteed. BE traffic flow is regulated by a leaky bucket and the advertised rate received from the down stream node during periods of congestion.

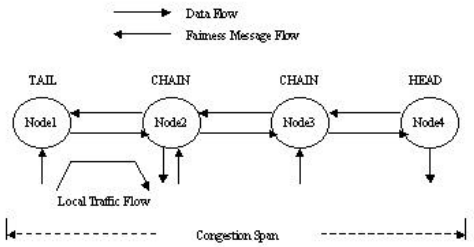


Fig.3 An example of a congestion span with local traffic in the span

Fig.3 gives an example of a congestion span, which is defined as the span of all nodes contributing to the congestion on a link. A congestion span typically consists of a head node, several chain nodes and a tail node. In Fig.3, for example, node 4 is a head node, node 2 and node 3 are chain nodes and node1 is a tail node. A node that detects a congested outgoing link is called head node. The head node knows the whole congestion span because each node tracks the IDs of all the source nodes with traffic passing through them. Based on the utilization of its downstream link, the head node calculates a fair rate and then advertises it to the upstream nodes. The initial advertised rate is normalized by the sum of all the weights assigned to the nodes within the congestion span. Having received the normalized advertised rate from the downstream node, each node calculates its target rate by multiplying the normalized advertised rate with its own weight and then

applying the rate to its leaky bucket thus controling the admission of its BE traffic flow. Using this scheme the 1TB-RPR-fairness algorithm distributes any spare capacity to all the nodes in the congestion span in a weighted fashion.

The example shown in Fig. 3 is a hub application where each node sends traffic to a hub node. Another important ring application is the distributed case where each node sends traffic to any other node on the ring. In distributed case, the congestion control scheme avoids adversely affecting the local traffic that does not contribute to the congestion on the congested link as shown in Fig.3. Otherwise the 1TB-RPR-fairness algorithm would be unfair and would result in loss of efficiency. The solution to this issue is to use Virtual Destination Queues (VDQ) that can be implemented outside the MAC layer. Packets are queued in different buffers based on their destinations. The leaky bucket only applies the advertised rate to the packets whose destinations are out of the congestion span. At least two VDQs are required (as shown in Fig.2), one in-span queue that buffers the in-span traffic flows and one out-of-span queue that buffers out-of-span traffic flows.

To let readers better understand 1TB-RPR fairness algorithm, we introduce the state machine that is the heart of the fairness scheme. As shown in Fig.4, the fairness state machine consists of 4 major states and 4 minor states. The minor states automatically exit after the required operation is completed. The major states are described as follows:

- Normal or Un-Congested state

This is the default state when a RPR node is powered on.

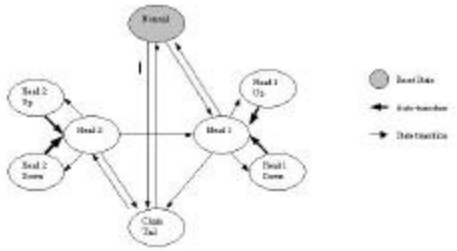


Fig.4 1TB-RPR fairness state machine

- HEAD 1 state

When a node in Normal state detects congestion, the node will change its state from Normal to HEAD 1. When congestion is released, the node will go back to Normal state if it is no longer in a congestion span. Otherwise it will change to CHAIN state.

While in HEAD 1 state, the node advertises the normalized rate to the upstream nodes and also calculates its own target rate applied it to its add-in traffic. Meanwhile, the node in HEAD 1 state continuously adjusts its advertised rate up or down depending on the output link utilization.

* Head Up state: This is a minor state to lower the advertised rate to the next lower value.
* Head Down state: This is a minor state to raise the advertised rate to the next high value.

- HEAD 2 state

If a node in CHAIN state also detects congestion in its outgoing link, the node will change its state to HEAD 2. This second head state is required to take into account congestion that may occur within the span. As the traffic pattern changes, not all traffic flows from the tail to the head. Congestion may occur within the span or multiple spans may merge into one. The HEAD2 state allows a single span to encompass other spans. HEAD1 and HEAD2 differ in their advertised rate calculations. The node in HEAD1 state is the true head node of the congested span. It controls

the target rate for all nodes in the span. The node in HEAD2 state will propagate the minimum target rate received from the node in HEAD1 state or from its own perspective.

- CHAIN state

A node can jump into CHAIN state from Normal state, HEAD1 state or HEAD2 state when it is in a congestion span but experiences no congestion on its outgoing link. A node in CHAIN state is identified as a node that receives a target rate lower than the maximum rate allowed. A chain node forwards the received target rate upstream while applying it to its leaky bucket.

- TAIL state

The tail state differs from a chain state in that it has very little pass-through traffic. A tail node does not forward its downstream rate to its upstream node, but simply applies the received rate to its own leaky bucket. The node will change back to Normal state from CHAIN or TAIL state if the congestion span is removed.

A simulation model using OPNET was built to investigate the performance of 1TB-RPR. In the next section we will discuss the simulation results.

3. Simulation results

The objective of the simulation is to examine the performance of 1TB-RPR as measured by ring access delay, utilization, and fairness. In order to study ring access delay and utilization, we designed a worst-case scenario to test the algorithm. A hub application where all the nodes on the ring send traffic to one hub node is clearly more stressful than distributed applications. Furthermore we loaded the ring in such a way that the traffic load of each node jumps at the same time so that the target load of the last link raises from 50% to 100% instantaneously. We refer to this case as the Step Response scenario. The ring model is set up as shown in Fig.5 where node 0 is the hub, and nodes 1 to 15 send traffic to node 0 along counter clockwise direction (inner ring).

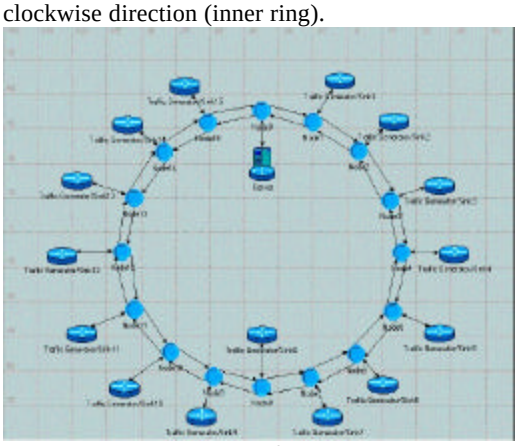


Fig.5 Ring model for hub application

Target Utilization	95%
HOL Delay Threshold	1,000 us
Link Rate	10 G bps
Propagation Delay of one link	70 us

Table 1 Common parameters set in the hub case

The traffic model uses three traffic classes: EF, AF, and BE. The packet inter-arrival distribution of AF and BE is exponential (Poisson traffic) while the packet inter-arrival distribution of EF is constant. While 444.4 bytes is the mean packet size, the packet size distribution for all classes is tri-modal, with sixty percent being 64 bytes, twenty percent 512 bytes, and twenty percent 1518 bytes. This tri-modal distribution is believed to have captured the characteristics of Internet traffic [12].

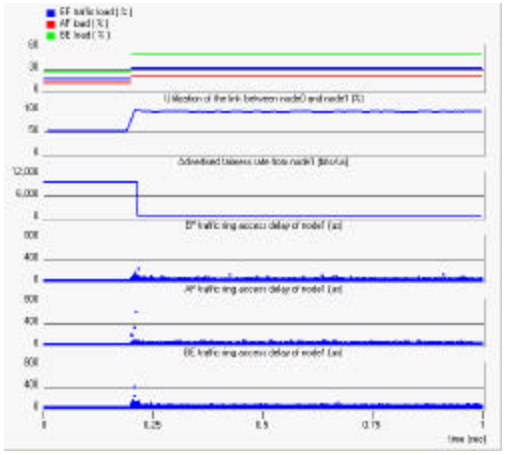


Fig.6 Step response results

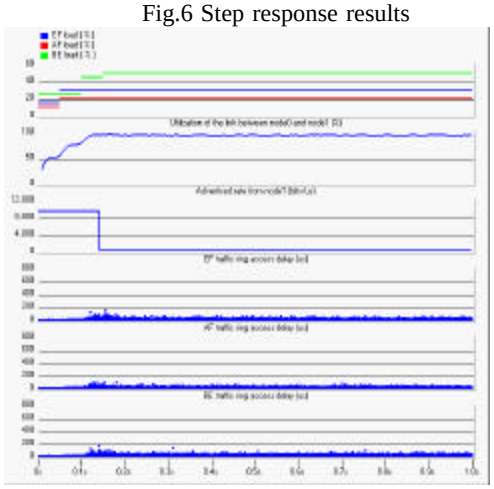


Fig.7 Piecewise linear results

The common parameters for the simulation setup are described in the Table1. We ran the Step Response scenario for 1.0 simulated second. Initially, all nodes are in Normal state, and BE traffic from each node is only limited by the node itself. When the total traffic load jumps from 50% to 100% at 0.1 simulated second, node1 detects congestion and advertises a fair rate to its upstream nodes to control BE traffic. From

the link utilization diagram shown in Fig.6, we note that 1TB-RPR can achieve 95% utilization and still work stably under 100% traffic load. The ring access delay diagram indicates that in the transient period when the traffic load suddenly jumps from 50% to 100%, the ring access delays of several packets can range around 200 us to 600 us while all other packets suffer less than 200 us access delays. In most the cases, it should be noted that this is still reasonably small and happens rarely during the transient period. In reality, this scenario cannot happen because the real data traffic flows can hardly be synchronized in such a way that the total load would spike from 50% to 100% instantaneously. Under realistic traffic scenarios the ring access delay can be significantly lower if the mean traffic load changes slowly. To test this conclusion, we designed a so-called Piecewise Linear scenario in which the traffic load changes from 50% to 100% gradually within 150 ms. From Fig.7 we can see that the maximum ring access delay of all the packets in the transient period of a Piecewise Linear scenario is three times lower than that in Step Response scenario.

We now move on to study the ring access delay characteristic. Given the statistical nature of ring access delays, we study this performance metric under steady state conditions. Fig.8 is the Cumulative Distribution Function (CDF) of EF packet ring access delay of Node1 at 95% target utilization and 100% traffic load. The CDF indicates that 98% of all ring access delays are lower than 100 us with either a 2000 bit or 150000 bits deed leaky bucket, and also shows that a smaller leaky bucket size can improve the performance in terms of ring access delay.

Since the 1TB-RPR supports weighted bandwidth fairness, we now examine this feature in our simulations. In Fig.9, diagrams A and C indicate the weight distribution pattern among the nodes on the ring, while diagrams B and D display the bandwidth each node gets when congestion happens. From the diagrams in Fig.9 we can see that the pattern of throughputs for all nodes matches the pattern of their corresponding weights very well.

As mentioned in section 2, 1TB-RPR uses multiple Virtual Destination Queues (VDQs) to prevent a congestion span from affecting the local traffic within the span in distributed applications. Therefore when simulating distributed applications, we focused on comparing local traffic throughput within the congestion span of one BE transmit buffer with the throughput of two BE transmit buffers (one in-span and another out-of-span).

Fig.10 shows the ring model we built for the distributed application having four hubs on the ring. In this model, all non-hub nodes send traffic along the counter clockwise direction (inner ring).

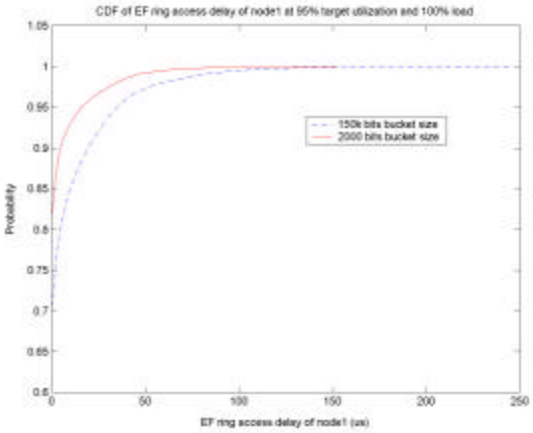
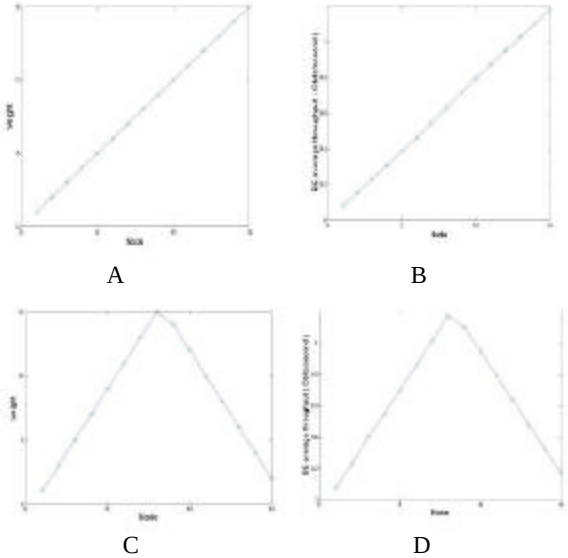


Fig.8 CDF of EF packet ring access delay of Node 1 at 95% target utilization and 100% load

The traffic configuration is described in Table2. From Table 2 we can see that the traffic to Server1 is over loaded to 105% utilization, but Server 2, Server3 and Server4 are 60% loaded. That means only the link from node 1 to node 0 on the ring is congested. By our earlier definition, node1 is the head node. Because all the nodes from node1 to node15 send traffic to server1, the congestion span spans from node1 to node15. The local traffic flows within the span are:

- Traffic flows from node 13, 14, 15, to Server 4.
- Traffic flows from node 9, 10, 11, 13, 14, 15, to Server 3.
- Traffic flows from node 5, 6, 7, 9, 10, 11, 13, 14, 15, to Server 2.



A B C D

Fig.9 1TB-RPR support for unequal bandwidth requirements

All those local traffic flows should not be affected by the congestion of node1. If the 1TB-RPR node is

implemented with only one transmit queue for local BE traffic, all the BE traffic flows will be regulated by the advertised rate from the head node. Therefore the in-span local traffic flows will be reduced unnecessarily. The consequent result is that the drop-off throughputs of server 2, server 3 and server 4 will become lower than the ideal values when the local flows are not affected by the congestion span. This is clearly unfair to local traffic.

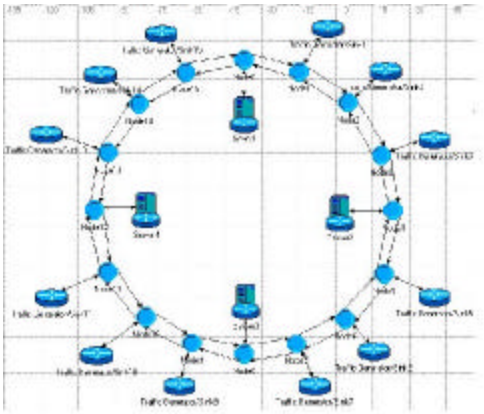


Fig.10 Ring model for the distributed application

Unit: Gbps

	Server 1	Server 2	Server 3	Server 4	Total
TG 1	0.8830	0.0500	0.0500	0.0500	1.0330
TG 2	0.8830	0.0500	0.0500	0.0500	1.0330
TG 3	0.8830	0.0500	0.0500	0.0500	1.0330
TG 5	0.8830	0.0500	0.0500	0.0500	1.0330
TG 6	0.8830	0.0500	0.0500	0.0500	1.0330
TG 7	0.8830	0.0500	0.0500	0.0500	1.0330
TG 9	0.8830	0.0500	0.0500	0.0500	1.0330
TG 10	0.8830	0.0500	0.0500	0.0500	1.0330
TG 11	0.8830	0.0500	0.0500	0.0500	1.0330
TG 13	0.8830	0.0500	0.0500	0.0500	1.0330
TG 14	0.8830	0.0500	0.0500	0.0500	1.0330
TG 15	0.8830	0.0500	0.0500	0.0500	1.0330
Total	10.5960	0.6000	0.6000	0.6000	

Table 2 Traffic configuration for the distributed application

Fig.11 demonstrates the significant improvement in local traffic throughput by using nodes having two transmit queues (one for in-span, one for out-of-span) for local BE traffic, versus the throughput of the single transmit queue scenario.

4. Conclusions

The simulation results demonstrate that an RPR ring with single transit buffer can achieve more than 95% utilization with extremely low ring access delay. The 1TB-RPR fairness scheme is stable and fair to all nodes under congestion. In addition, 1TB-RPR can support unequal bandwidth requirements and distributed applications effectively as predicted.

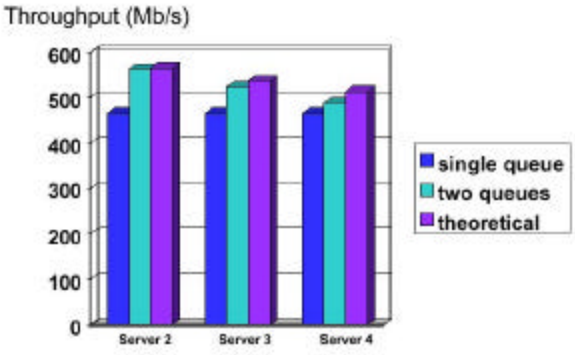


Fig.11 Single queue vs. multiple (two) queues

References

[1] Cole, N., Hawkins, J., Green, M., Sharma, R. and Vasani, K., "Resilient Packet Rings for Metro Networks," http://www.rpralliance.org/ , August 2001.

[2] RPR Alliance, "An Introduction to Resilient Packet Ring Technology,"http://www.rpralliance.org/, October 2001.

[3] IEEE 802.17 Working Group, "Proposed IEEE Standard 802.17– Darwin," Draft version, http://www.ieee802.org/17/documents/drafts/Darwin_v1_0.pdf, January 2002.

[4] IEEE 802.17 Working Group, "Proposed IEEE Standard 802.17 – Gandalf," Draft version. http://www.ieee802.org/17/documents/drafts/gandalf_04Plus.pdf, November 2001.

[5] IEEE 802.17 Working Group, "Proposed IEEE Standard 802.17 – Alladin,"Draft version. http://www.ieee802.org/17/documents/, November 2001.

[6] Nortel Networks, "OPTera Packet Edge System General Specification," Document release: 1.0, September 2000.

[7] Chiu, A. L., and Gallager, R. G., "Full Utilization Dynamic Fairness and Bounded Access Delay on High Speed Slotted Bus Network,", 1998 IEEE International Symposium on Information Theory, 1998.

[8] Chiu, A. L., and Gallager, R. G., "Full Utilization and Fairness on High Speed Bus Network," GLOBECOM'96. Vol.1. Page(s): 508 – 512, Vol.1, 1996

[9] Mukherjee, B. and Bisdikian, C., "A Journey through the DBDQ Network Literature," Performance Evaluation, 165:129-158, 1992.

[10] Hahne, E. L., Choudhury, A. K., and Maxemchuk, N. F., "Improving the Fairness of Distributed-Queue-Dual-Bus Networks," Infocom 1990, San Francisco, Ca., PP. 175 – 184, 1990.

[11] Peng, H., "iPT Fairness: Control Access Protocol (CAP)," 2001 Northern Telecom Limited, Issue number: 1.0, May 2001.

[12] IEEE 802.17 Working Group, "Guidance for IEEE 802.17 RPR Performance Simulations," http://www.ieee802.org/17/performance_committee.htm, November 2001.