WHITE PAPER

QoS & Traffic Management

Advanced Features for Managing Application Performance and Achieving End-to-End Quality of Service in Data Center and Cloud Computing Environments using Chelsio T4 Adapters

Chelsio Terminator 4 (T4)-based network adapters have sophisticated features for both Traffic Management and Quality of Service (QoS). These features can be used to achieve required levels of performance by controlling how certain application I/O is handled within the Chelsio T4 NIC, allowing you to effectively control end-to-end QoS for various host applications using a shared 10GbE network resource. Any latency sensitive application can be assigned a high priority which allows it to bypass application traffic that is not latency sensitive. Two examples of high priority data include inter-node communication by distributed lock managers in clustered database systems and streaming data traffic from stock market feed servers to high frequency trading systems.

End-to-End Application Quality of Service (QoS)

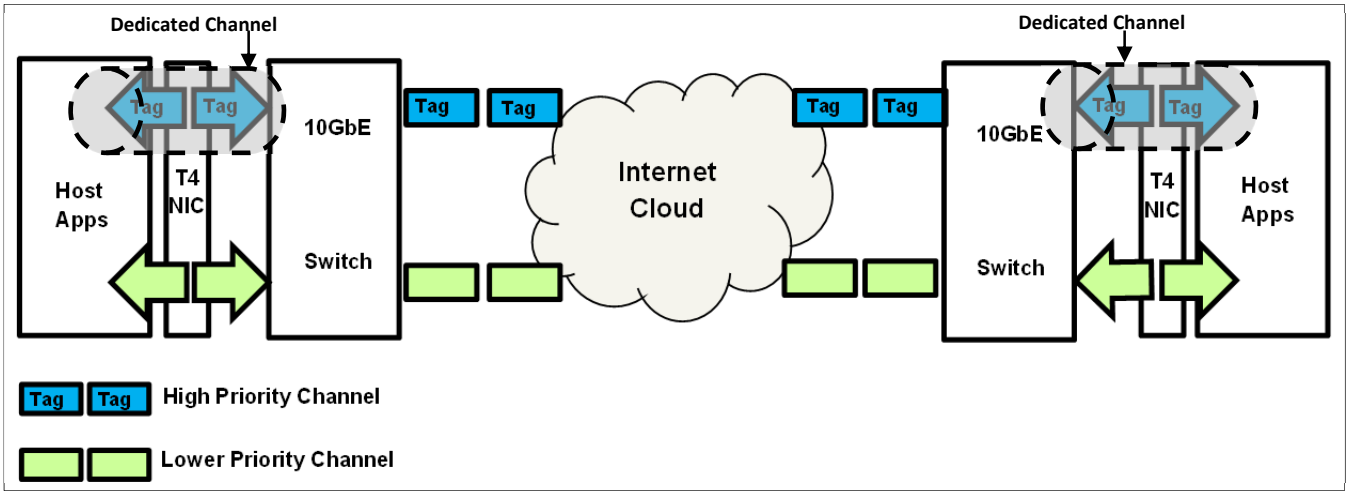

Figure 1: Quality of Service in Chelsio T4

End-to-end QoS for application I/O is achieved by the Chelsio T4 NIC using dedicated and independent high priority hardware channels for receive and transmit. The desired QoS can be extended through the network using VLAN Tags to designate high priority packets, such as those needed for inter-node communication in clustered databases. Chelsio's T4 NIC diverts high priority packets to a separate hardware channel allowing them to bypass lower priority packets. Tagged packets then pass through the intervening switches and are treated as high priority packets by the T4 NIC on the host at the other end, where the Chelsio T4 NIC again diverts high priority tagged packets to a dedicated channel to the host. This ensures end-to-end quality of service to latency sensitive traffic, even in the presence of high bandwidth traffic.

Advanced Traffic Management

Traffic Management capabilities in the Chelsio T4 NIC can shape transmit data traffic through the use of sophisticated queuing and scheduling algorithms built-in to the T4 ASIC hardware, which provide fine-grained control over packet rate and byte rate. These features can be used in a variety of data center application environments to solve traffic management problems. Examples include configuring video servers to deliver predictable bandwidth and jitter, and implementing bandwidth provisioning for virtual interfaces carved out of a shared network resource in a virtualized host.

Traffic flows from various applications are rate controlled, then prioritized and provisioned using a weighted round-robin scheduling algorithm. For example, in Figure 2, App3 could be a cloud based SaaS application with a set of specified SLAs (service level agreements) that place demands on certain required levels of performance and QoS. Traffic from multiple flows can be load-balanced, with programmable parameters (bandwidth and latency), across host CPUs to deliver predictable latency and bandwidth thus ensuring predictable host application performance (see Figures 8 & 10 for Chelsio T4 Traffic Management performance test results).

In Summary, Traffic Management features in Chelsio's T4 adapters allow you to control three main functions:

- **Guarantee low latency in the presence of high bandwidth (data mover) traffic**
- **Control maximum bandwidth that a connection, flow or class (group of flows) can use**
- **Allocate available bandwidth to several connections or flows based on desired levels of performance**

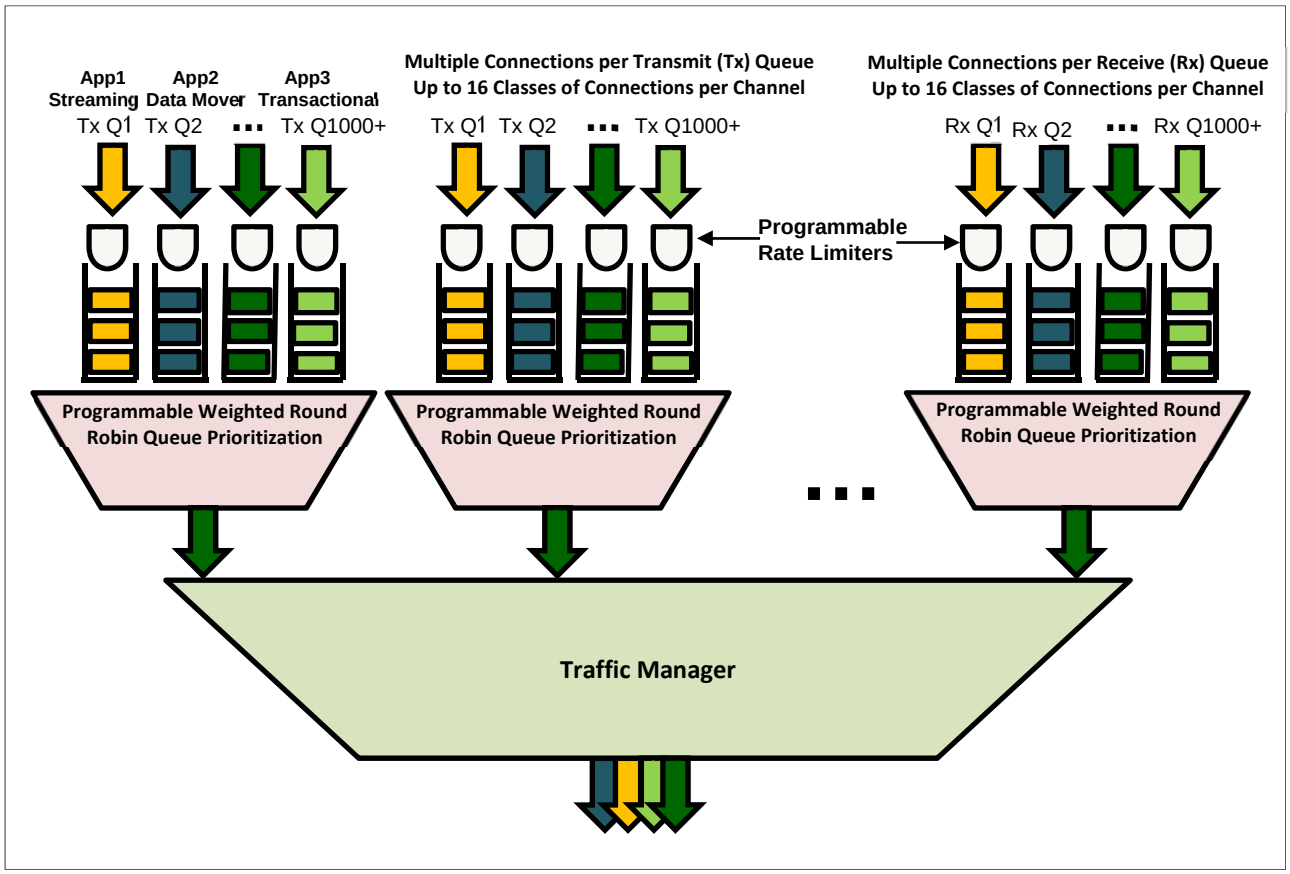

Figure 2: Traffic Management in Chelsio T4

Applications for Traffic Management Features in Chelsio T4

Listed below are some of the key applications for Chelsio's Traffic Management Features:

Messaging Traffic – Cluster Interconnects

Oracle Real Application Clusters (RAC) - Shared DB Scale-Out Architecture

Oracle Real Applications Clusters (RAC) is an N+1 scale-out shared data database architecture. A typical 4-Node RAC cluster is shown below. The RAC nodes are clustered using standard 1Gb or 10Gb Ethernet. Inter-node Traffic (Red) (Figure 3) consisting of Cluster Synchronization and Distributed Lock Manager (DLM) Traffic is prioritized over the NAS or SAN storage Data Mover Traffic (Blue) allowing Inter-node Traffic to be delivered with the lowest latency possible without affecting the bandwidth of the Data Mover Traffic.

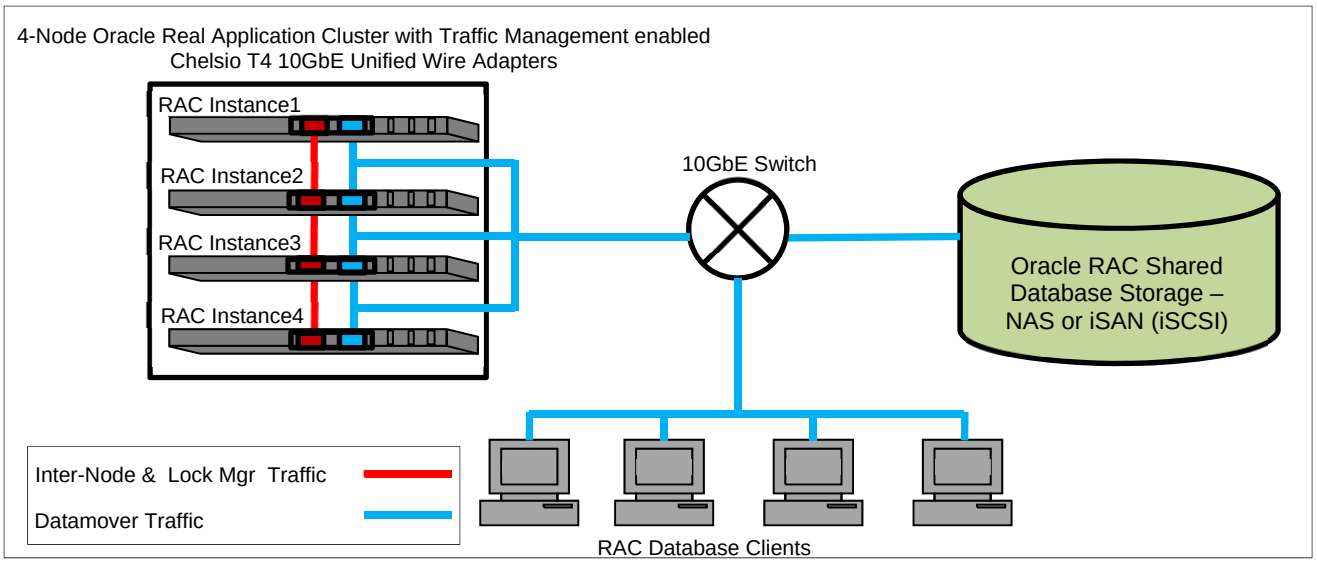

Figure 3: Traffic Management with Oracle Real Application Cluster Interconnects

Landmark Seismic Simulation High Performance Computing (HPC) Scale-Out Clusters

High Performance Computing (HPC) Clusters use an N+1 scale-out computing architecture for numerical processing workloads in various vertical industry applications such as Energy, Biotech, Media and National Labs. The HPC nodes are clustered using standard 1Gb or 10Gb Ethernet. Messaging Traffic (Red) (Figure 4) consisting of HPC inter-node communication & cluster synchronization traffic is prioritized over NAS Data Mover Traffic (Blue) allowing Messaging Traffic to be delivered with the lowest latency possible without affecting the bandwidth of the Data Mover Traffic.

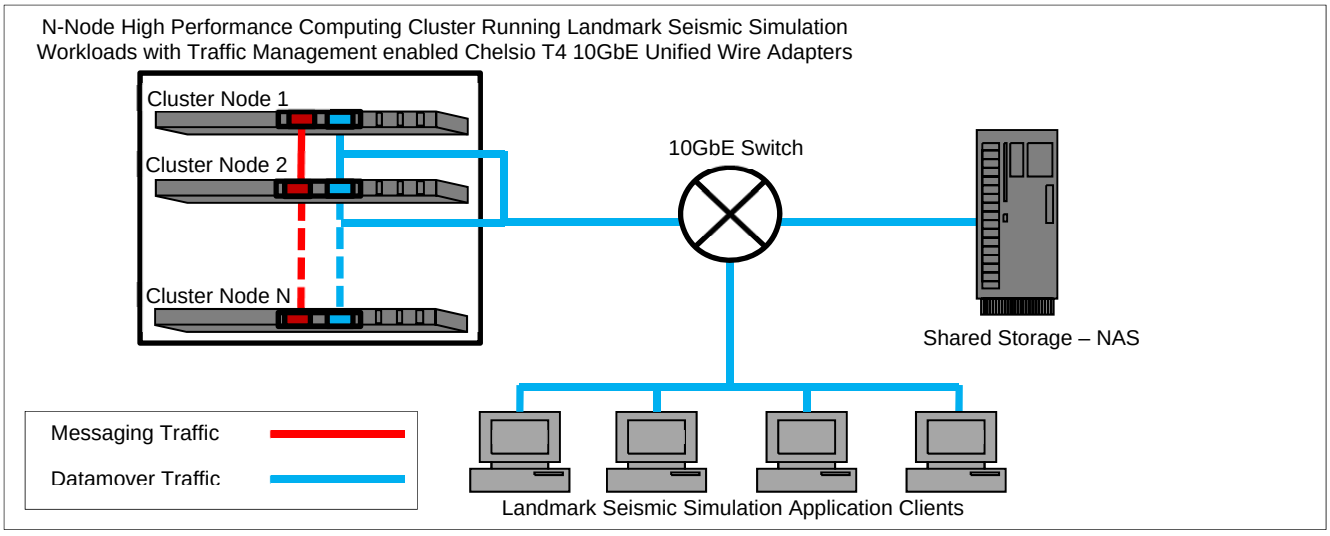

Figure 4: Traffic Management with Landmark Seismic Processing HPC Cluster Interconnects

Storage QoS

Implementing Storage QoS is an excellent use of Chelsio T4's advanced Traffic Management functionality. In this example (Figure 5), all client traffic to LUN1 (show in Red) is prioritized over all other LUNs and Volumes in the unified storage server shown. LUN1 Client is running a demanding latency sensitive application that requires its traffic to be delivered with minimal latency and acceptable bandwidth. Bypassing all traffic to LUN1 over a dedicated channel allows LUN1 Client to meet the desired level of storage I/O QoS measured in terms of IOPs and latency.

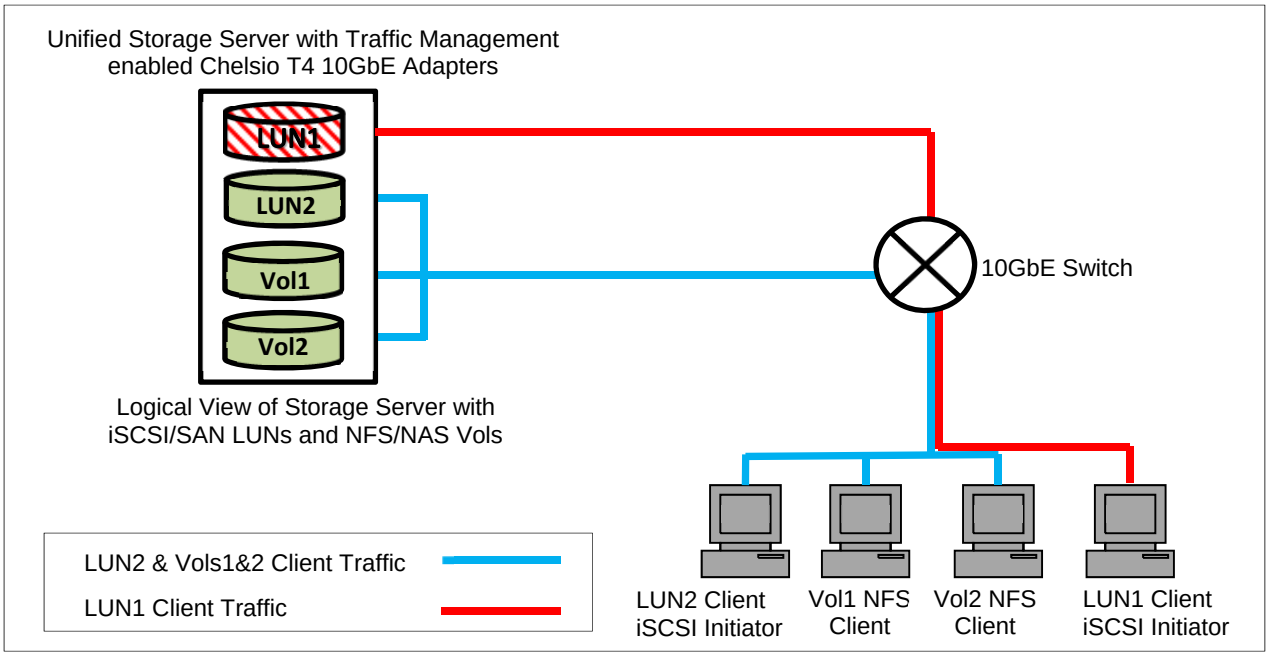

Figure 5: Traffic Management in Storage Servers

Streaming Data – Video

Video streaming presents a very compelling application for Traffic Management. In this example (Figure 6) T4 traffic manages tens of thousands of IPTV and Web/HTTP video streams at the MPEG2/4 and AVC/H.264 rates ex. 200kbps, 500kbps, 1Mbps, 2.5Mbps, 4Mbps, 12Mbps and 18Mpbs. The traffic manager can handle multiple media rates simultaneously and manage each of the flows with low jitter. The media rate for a particular flow can be adjusted dynamically by moving the flow from one traffic class to another, e.g. from the traffic class for AVC/H.264 at 4Mbps to the class at 12Mbps, etc...

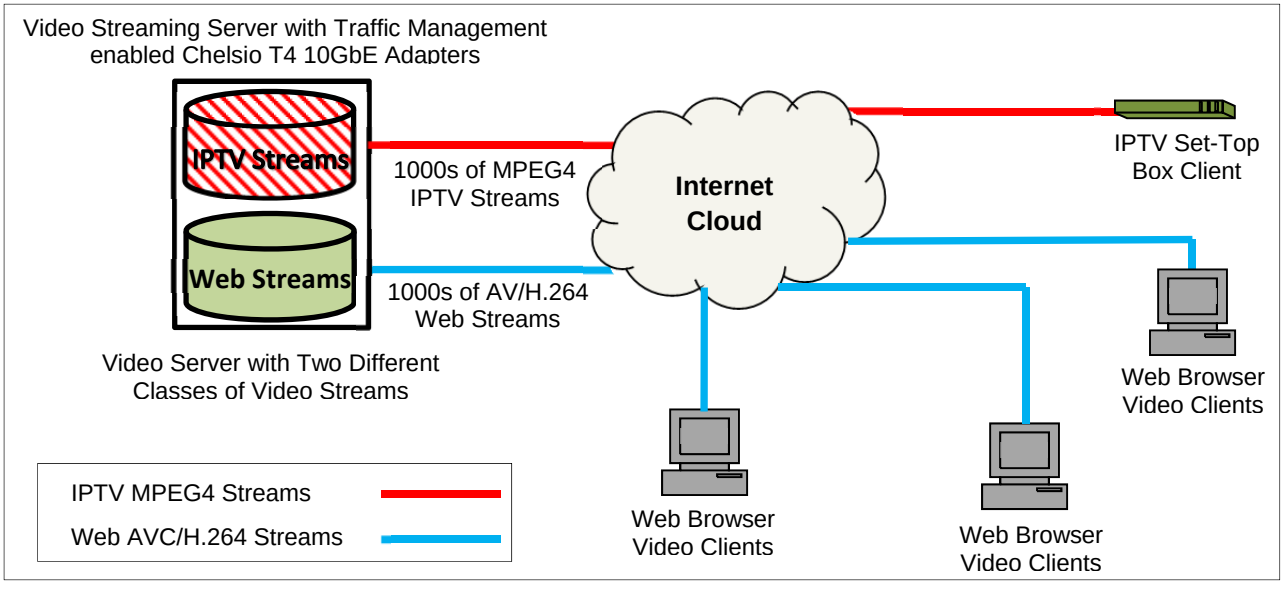

Figure 6: Traffic Management in Video Servers

Cloud Computing – Multi-Tenant Software as a Service (SaaS) Applications

In SaaS based Cloud Computing environments, computing resources are carved up based on customer SLAs for their application. These SLAs specify availability and response time (example, query response time) requirements among other things. Based on these SLAs and the type of hosted SaaS application, multiple customers sharing the SaaS application and hardware infrastructure can be assigned different classes of network resources such as bandwidth and latency.

A typical cloud based SaaS application configuration (see Figure 7) uses server, storage and network virtualization technologies to partition hardware resources to be shared among a set of SaaS customers. In Figure 7, Customer4 has a higher level of SLA (service level agreement) shown in Red. For example, the customer may require a more bandwidth or lower latency or both relative to the other customers (Customer 1-3). Each customer's hosted application environment is siloed within a Virtual Machine (VM) with well-defined ACL security and SLA policies associated with each customer/VM container. Configuring and enabling Chelsio T4 Traffic Management for allocating shared network resources on a per-customer access basis is an extremely cost effective use of this powerful feature to meet or exceed customer SLAs.

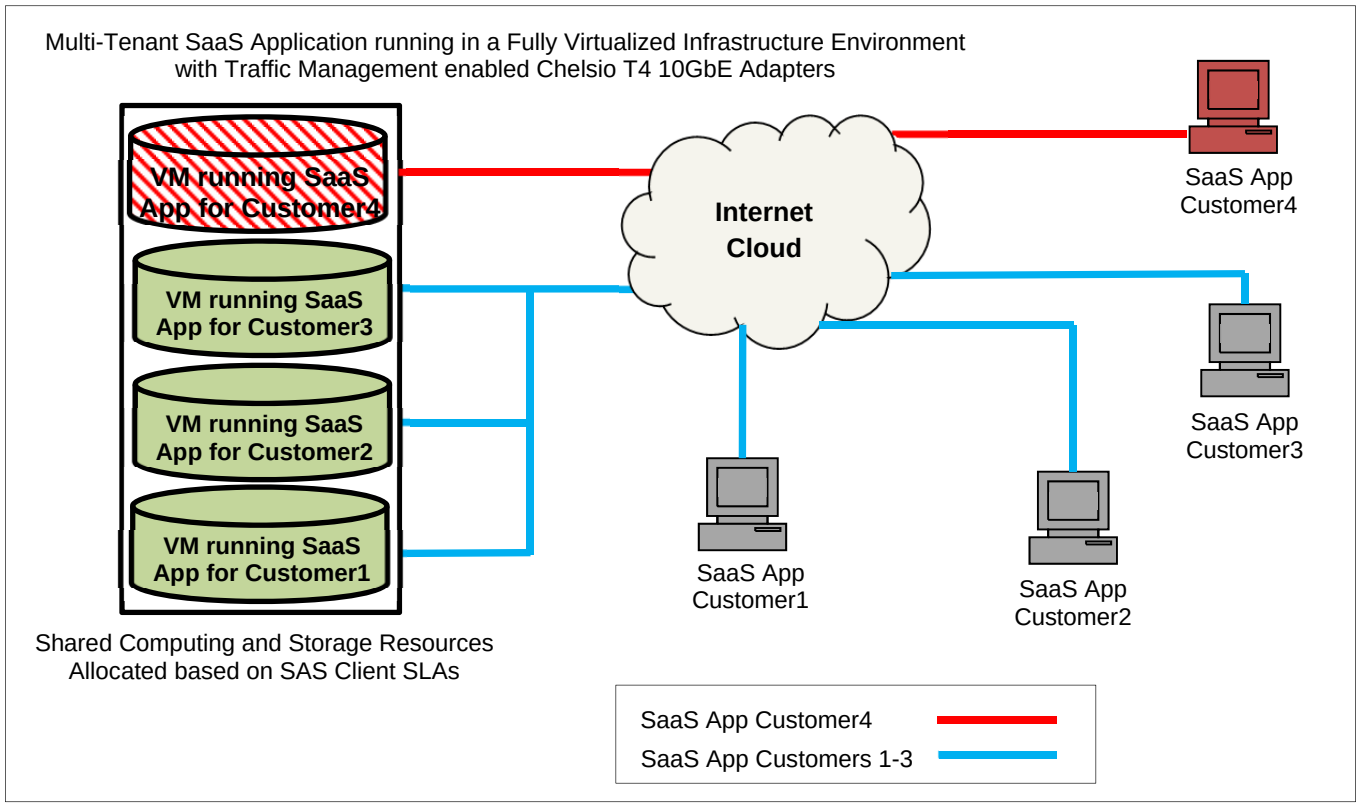

Figure 7: Traffic Management in Cloud Computing

The following graphs illustrate the benefits of the traffic prioritization provided by Chelsio's T4 based NICs, highlighting high priority traffic can be delivered with low latency even in the presence of high bandwidth traffic (large TCP sends – see details in following section). In particular, Figure 8 shows that traffic management maintains latency at near baseline numbers, whereas a regular NIC would have shown an order of magnitude deterioration in latency performance.

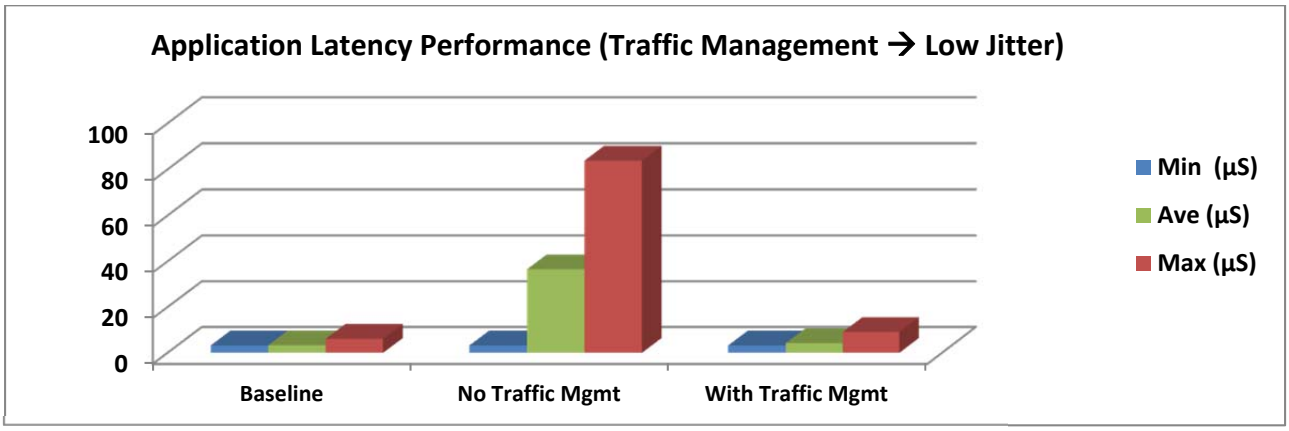

Figure 8: Application QoS - Latency With Traffic Management is Shielded From Data Mover Traffic

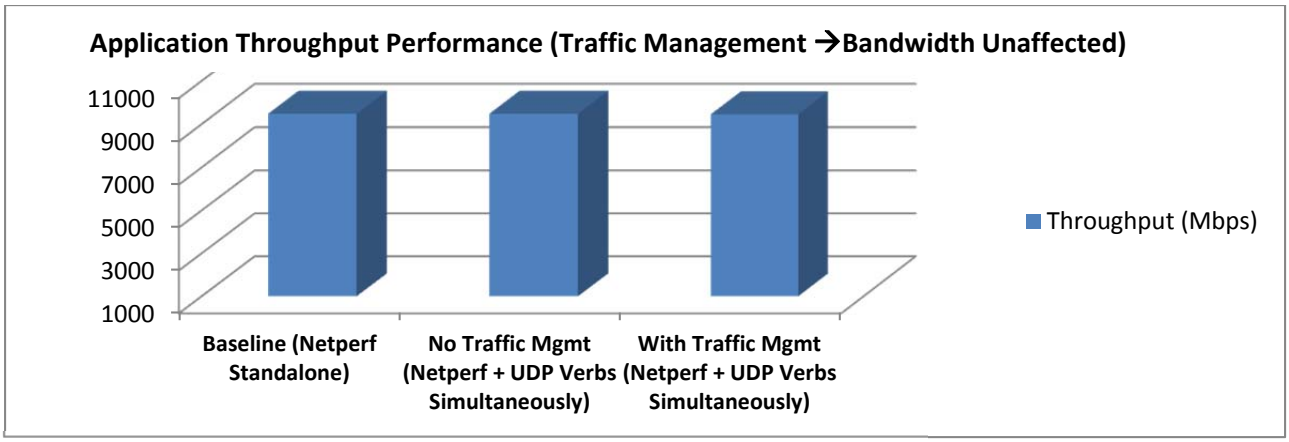

Figure 9: Application QoS – Throughput Aggregate is Preserved with Traffic Management

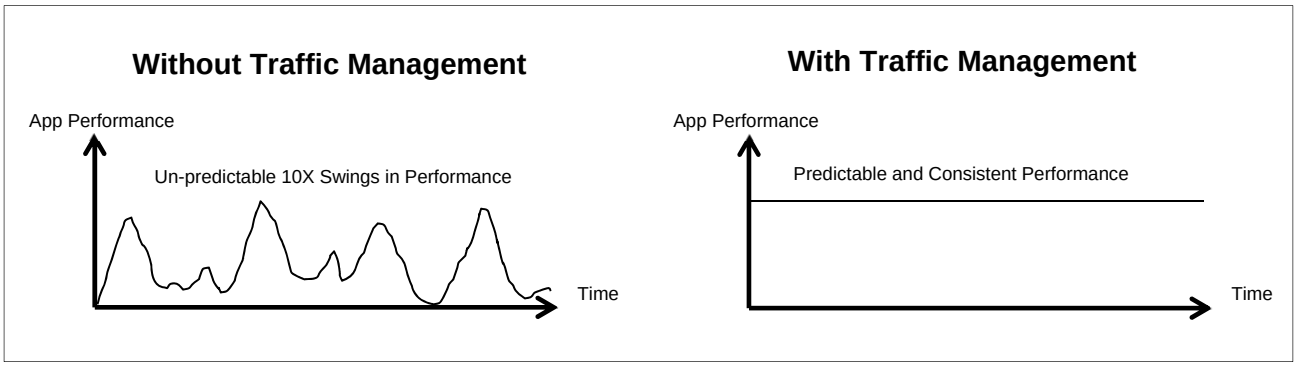

Figure 10: Application Performance is Predictable and Consistent with Traffic Management

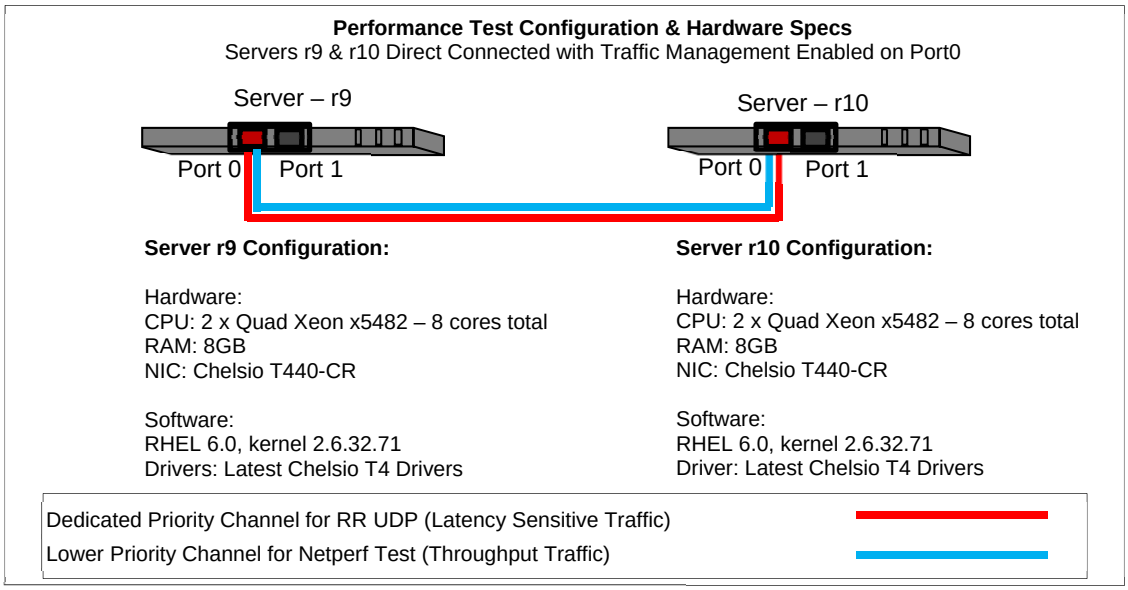

Figure 11: Traffic Management Performance Test Configuration

First, UDP low latency traffic and netperf throughput test run separately to baseline performance:

```
[root@r9 ~]# udp_test -t rr -T 5 -I
192.168.42.111 192.168.42.112
Average RTT: 6.551 usec, RTT/2: 3.276 usec
RTT/2: calls 763228 samples 3072
```

Ave 3.30113 µS, Min 3.21038 µS, Max 6.13572 µS

```
[root@r9 ~]# netperf -H 192.168.42.112
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port
0 AF_INET to 192.168.42.112 (192.168.42.112)
port 0 AF_INET
Recv   Send    Send
Socket Socket  Message  Elapsed
Size   Size    Size     Time     Throughput
bytes  bytes   bytes    secs.    10^6bits/s
87380  65536   65536    10.00    9478.53
```

RR UDP & netperf tests run simultaneously without Traffic Management (single channel):

```
[root@r9 ~]# netperf -H 192.168.42.112
&udp_test -t rr -T 5 -I 192.168.42.111
192.168.42.112
[1] 6419
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port
0 AF_INET to 192.168.42.112 (192.168.42.112)
port 0 AF_INET
Average RTT: 258.428 usec, RTT/2: 129.214
usec
RTT/2: calls 19348 samples 3072
```

Ave 36.5201 µS, Min 3.24538 µS, Max 83.806 µS

```
[root@r9 ~]# Recv   Send    Send
Socket Socket  Message  Elapsed
Size   Size    Size     Time     Throughput
bytes  bytes   bytes    secs.    10^6bits/s
87380  65536   65536    10.00    9482.35
```

Average latency increases tenfold from 3.3usecs to 36.5usecs without traffic management.

Finally, RR UDP &netperf tests simultaneously with traffic management (two channels):

```
[root@r9 linux_t4_build]# netperf -H
192.168.42.112 &udp_test -t rr -T 5 -I
192.168.42.111 192.168.42.112
[1] 7979
TCP STREAM TEST from 0.0.0.0 (0.0.0.0) port
0 AF_INET to 192.168.42.112 (192.168.42.112)
port 0 AF_INET
Average RTT: 8.508 usec, RTT/2: 4.254 usec
RTT/2: calls 587716 samples 3072
```

Ave 4.20943 µS, Min 3.24626 µS, Max 9.08127 µS

```
[root@r9 linux_t4_build]#
Recv   Send    Send
Socket Socket  Message  Elapsed
Size   Size    Size     Time     Throughput
bytes  bytes   bytes    secs.    10^6bits/s
87380  65536   65536    10.00    9454.00
```

When the RR UDP traffic is mapped to an independent channel, average RR latency drops from 36.5 µS to 4.2 µS, which is near baseline. Note that due to the non-preemptive sharing of the PCI bus and Ethernet MAC port, the maximum latency slightly increases due to wait for partial frame transmission.