



Load Balancing System Design

Using Advanced Features in FocalPoint

White Paper

June, 2008

Introduction

Data Centers are evolving into virtualized environments where resource utilization can be maximized, which provides other advantages such as reducing cost and power dissipation. To optimize the use of these virtualized environments, advanced load balancing techniques can be employed. This paper will give a brief overview of trends in the data center, how load balancing is employed and how the advanced features in FocalPoint can be utilized to improve performance and reduce cost in these load-balancing systems.

Data Center Requirements

This section will discuss current data center requirements for enterprises that deploy a significant number of servers. In some cases, these servers will be deployed in multiple locations for reasons of security and disaster recovery.

Resource Optimization

As server power dissipation and densities increase, many data centers are reaching power delivery and thermal management limits. Server virtualization helps to alleviate this problem by optimizing hardware utilization. But the allocation of these virtual machines cannot be adjusted quickly enough to react to the varying usage requirements coming from the network. Load balancing addresses this problem by intelligently spreading the workload across multiple virtual machines. This mechanism will be described further in the next section.

Performance

Data center bandwidth requirements are escalating as new services such as video streaming are offered. In addition, storage network bandwidth is increasing from 4Gb/s to 8Gb/s in many data center installations. These factors along with the trend to converge data, storage and HPC traffic into one Ethernet fabric, mean that standard 1GbE switches are no longer adequate for data center networks. Emerging data center switches must not only support 10GbE, they must provide scalability, low latency and guaranteed bandwidth for storage and HPC traffic.

Security

Data center security continues to be a challenge as outside threats have become more sophisticated while, at the same time, the workforce has become more mobile. Devices at the edge of the network such as load balancers must provide security against outside threats such as denial of service attacks while authenticating valid users.

Failure Recovery

Enterprises must be able to quickly recover from equipment failure. This is especially true for applications such as e-commerce where a significant amount of money can be lost during down time. Because of this, both data and server resources can be replicated at local or remote sites. But idle resources waste money. Load balancing can solve this problem by not only spreading the workload across distance resources, but by also quickly redirecting traffic away from a resource during failure.

Load Balancing

As discussed above, load balancing can be used to optimize data center resources while also providing fast failover and security. This section will give a brief overview of how load balancing is typically implemented in the data center.

Virtual Services

Hardware based virtualization using virtual machines (VMs) have become popular in data center deployments enabled by products such as Microsoft Virtual Server, VMware ESX Server and Xen. Each VM will have an IP address and each service within the VM will have a port number. This IP address, port number combination is commonly known as a “virtual service”. Load balancing must be done across a pool of these virtual services.

Load Balancing Methods

Load balancing is done using network address translation (NAT). Requests come in using the load balancers IP address. When the load balancer receives a service request, it determines the optimal virtual service to send this request to. It then modifies the destination IP address and sends it to the correct VM. When the result of the request is sent back to the load balancer, it modifies the source IP address to its own IP address. In this way, the client sees the load balancer as a single monolithic server.

The load balancing decision algorithm is where system designers can differentiate their products. Simple round-robin service distribution may be far from optimal depending on processing requirements. Also, the health of each service must be monitored so that failed virtual services can be quickly removed from the pool. In some disaster scenarios, the service requests must be re-routed to remote locations.

Connection Persistence

For some requests, it is desirable not to load balance. For example, a long-lived connection such as an FTP request must stay connected to the same virtual service for a period of time. Other cases, such as multiple e-commerce requests from the same client, new requests must be directed to the same virtual service due to transaction state information. Because of the widespread use of proxy servers, the load balancer may also need to read packet payload information in order to identify unique clients.

Data Center Fabric Virtualization

Data center hardware and software are evolving to meet ever-growing resource demands while administrators are striving to reducing the deployment and operating cost per resource. Virtualization has become key to achieving these goals. This section will give a brief overview of data center fabric virtualization and how this impacts load balancing.

IO Virtualization

PCI-Express is the dominant host interface in servers today. The PCI software model employed by most operating systems assumes that each OS will ‘own’ any components such as IO adapter cards connected to its PCI bus. First generation PCI-Express (PCIe) switches are really not switches in the true sense, but actually virtual PCI busses used to connect multiple peripheral devices to a single host. Although some host fail-over

White Paper: Load Balancing System Design

mechanisms are available, two or more hosts cannot share a single IO adapter card. This is also true for server virtualization where two or more VMs cannot share a single IO adapter card.

The PCI-SIG has addressed this server virtualization problem with the release of the PCI-Express Single Root Complex IO Virtualization specification (the multi-root complex spec is not complete at the time of this writing). This specification defines a method whereby multiple virtual machines within a host can share a single peripheral device such as a 10G Ethernet interface adapter. It does this by implementing several virtual Ethernet adapters and converging them into a single 10GbE data flow. Each VM understands that it owns a (virtual) Ethernet adapter, but has no concept of bandwidth sharing and is completely independent from the other virtual machines. This allows the use of standard PCIe drivers in each VM. Significant cost savings is achieved by sharing a high bandwidth IO adapter across multiple VMs. In the example shown in figure 1, multiple virtual path connections are shown from each server to the data center switch through 10GbE data pipes.

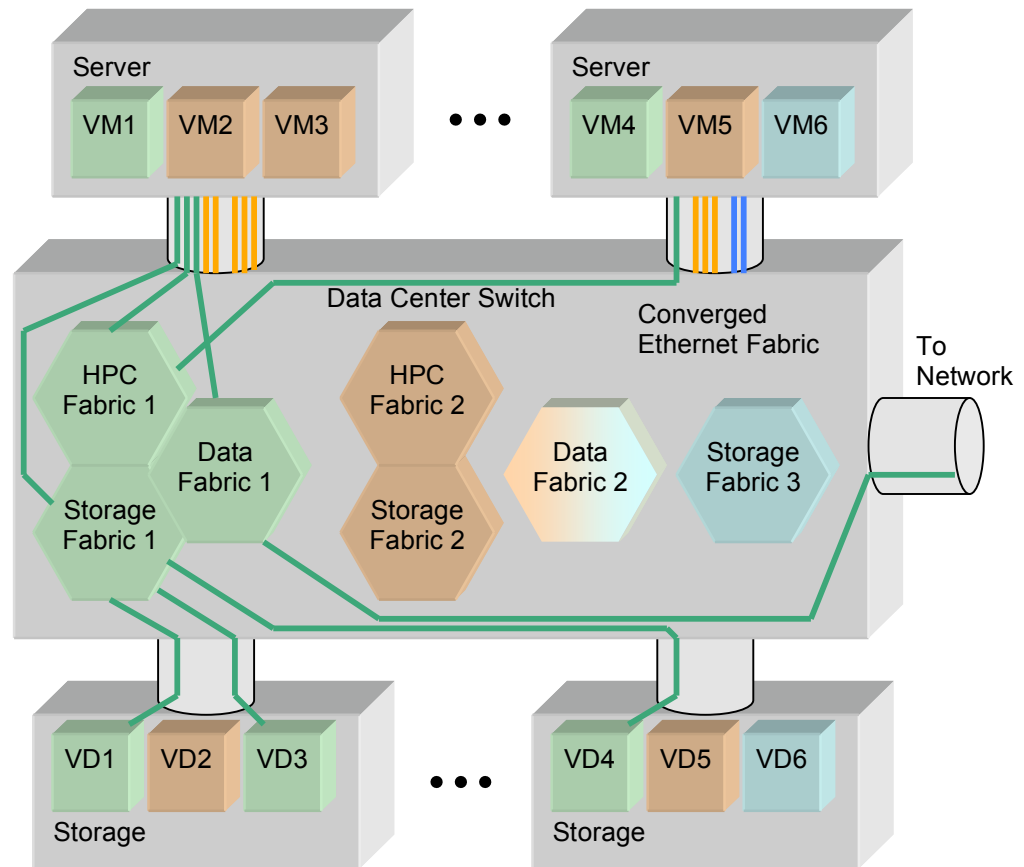


Figure 1: Virtualized data center example

The FocalPoint Advantage

The FocalPoint family of switch devices contain up to 24 10GbE ports and utilize a high performance, low latency Ethernet switch architecture including a comprehensive suite of Layer 3 features and advanced security, congestion management and system management capabilities. These switches contain all the features necessary to enable Ethernet as the single, converged datacenter fabric, transporting inter-processor, storage and networking traffic. Traffic classification, class-based congestion management, and sophisticated traffic policing and queue management mechanisms allow multiple traffic types to be classified and effectively processed on a prioritized basis. Advanced tag switching, L2/3/4 access control lists and other security features, and in-band management capabilities -- all while preserving extremely low 300ns latency -- contribute to provide a resilient, highly-scalable datacenter switch platform.

Virtual Fabric Technology

The figure below shows how a single FocalPoint switch can be configured as two virtual fabrics for load balancing applications. The redundant switch is not shown in this figure. To the right of the figure, there are multiple servers, each containing multiple VMs. Several virtual services can be running on each VM. Assuming that PCIe IO virtualization NICs are used, a single 10G cable can transport both data and iSCSI or FCoE traffic for each VM. A combination of class-based pause frames, separate memory partitions and class-of-service queues within the FocalPoint switch can be used to guarantee maximum latency for storage traffic through the switch and isolate it from data fabric congestion.

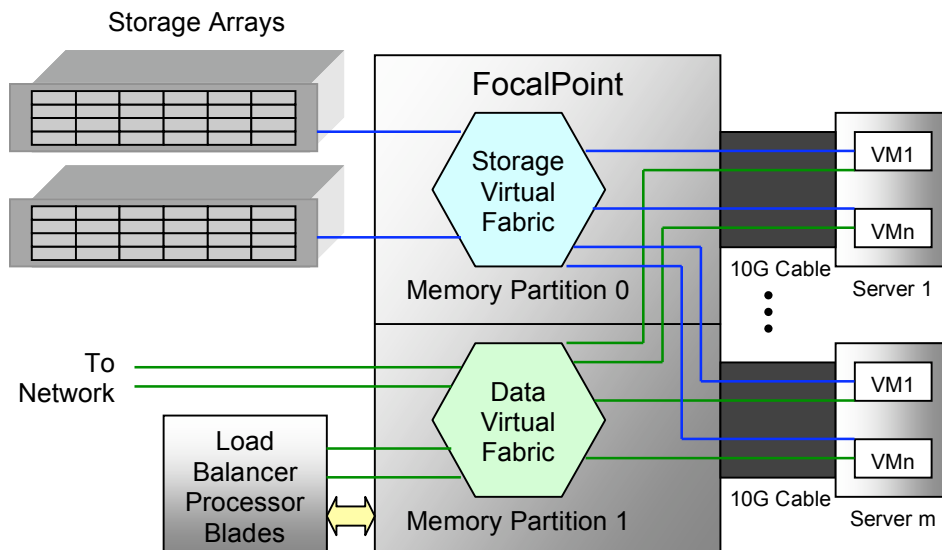


Figure 2: Virtualized load balancing example

In this example, all incoming requests first go to the FocalPoint switch. If necessary, FocalPoint can distribute this traffic across multiple processor blades that are running load-balancing algorithms. The requests are processed by these blades and then load balanced across multiple virtual services. The result of the request is sent back to the load

White Paper: Load Balancing System Design

balancer before it is sent out on the network. Some flows do not need load balancing and can bypass the load balancer.

The load balancer blades can connect to the CPU port on FocalPoint or use in-band management frames to configure FocalPoint to assist with packet forwarding. For example, it may be desirable to drop or redirect some IP source addresses before the load balancer blades ever see them. These blades can also configure and monitor the fabric through this interface.

Frame Filtering and Forwarding Unit

The frame filtering and forwarding unit (FFU) inside FocalPoint can be used for both network security, and to assist in frame forwarding. It can process up to 360M packets per second allowing full line rate processing on all ports. The FFU can be configured using bit masks to read any part of the L2/L3/L4 header. If there is a match, several actions can be taken such as route, deny, modify, count, log, change VLAN, change priority etc. Multiple actions can be associated with a given match. These features can help offload functions such a security from the load balancer blades. See the FM3000 Policy Engine White Paper for more information.

Load Distribution

FocalPoint can be configured to distribute incoming network traffic across multiple hosts in order to reduce the bandwidth requirement on each host. This provides a very scalable solution for applications such as load balancing where deep packet inspection needs to be performed at line rate. At the network ingress in figure 2, FocalPoint can distribute packets to the load balancing processor blades. This can also be used for scaling as described below. The load distribution uses a modified Persons hash function based on information in the packet header that defines a flow. FocalPoint can be configured to use various L2/L3/L4 header fields. A given flow will always use the same path through the fabric. In addition, the hash function is symmetric so that service requests and service responses will always be directed to the same load balancing processor blade.

Scalability

FocalPoint can be used to create low latency multi-stage fabric topologies for the data center. The figure below shows how a load balancing solution can be scaled to support a large number of servers. Here, the Fulcrum Top of Rack (TOR) switches connect to the network and also connect to the load balancing systems and storage arrays. Stacking topologies can also be used by connecting the TOR switches in a ring or mesh configuration.

In this example, the load balancers also contain FocalPoint switches as shown in figure 2. Incoming network traffic is distributed to one of the load balancing systems as described above. The load balancer sends the request result back to the network connection that it arrived on. Since FocalPoint has cut-through latency less than 300nS, the total fabric latency from server-to-server or server-to-storage will be less than 1uS. This can be critical for applications that require low-latency storage access, plus it allows more flexible storage implementations.

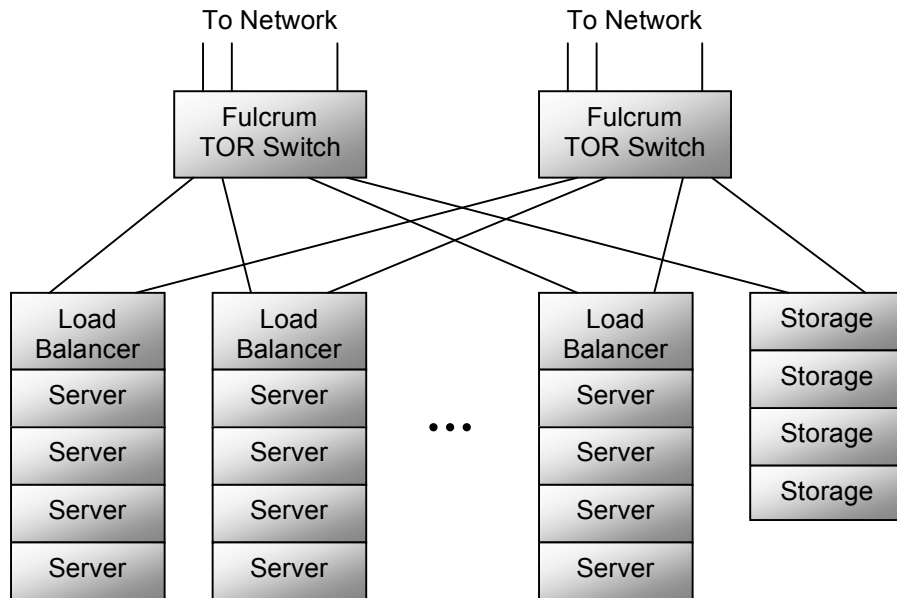


Figure 3: Load balancing example

Fail Over

FocalPoint uses a link mask table to determine how to distribute the load across multiple egress ports. It also contains several mechanisms to detect link failure such as loss-of-signal (LOS) or CRC errors. As the packet header is processed, the forwarding unit resolves to the address of a pointer, which points to an entry in the mask table. If a link or connected device fails, this pointer can be quickly changed by software so that the failing link is no longer part of the load distribution group. Since distribution is flow based, only flows to the failed device will be affected.

Security

FocalPoint has the ability to operate using Access Control Lists (ACLs) to prevent security violations such as denial of service attacks. A set of rules and a set of actions for each rule can be defined as described in the FFU section above. Counters can help identify misbehaving flows or they can be trapped to the local CPU. The local CPU can then adjust the ACL rules to deny access to these flows. If FocalPoint is between the network and the load balancers, it can offer a first line of defense and also off-load some of these security functions from the load balancing processors.

Conclusions

Load balancing systems help optimize server workloads in virtual data center environments. FocalPoint can enhance this function by offering virtual fabric memory partitions for storage and data as well as providing low latency storage access. FocalPoint can also help distribute the processing load across multiple load balancing cards or systems, providing scalable solutions. Finally, FocalPoint offers advanced frame forwarding and security features ideally suited to advanced load balancing systems.

Fulcrum Microsystems, Inc.
26630 Agoura Road
Calabasas Hills, CA 91301
818.871.8100
www.fulcrummicro.com