Scaling SMB for Server 2012 with RDMA over Ethernet

The release of Windows Server 2012 is expected to be one of the largest, most feature rich Server Operating System releases for Microsoft. Besides a new user interface, **Windows Server 2012** brings to market a comprehensive set of solutions for virtualization, networking, and high availability storage.

iWARP & SMB

Noteworthy in the Windows Server 2012 release is the introduction to mainstream of technologies that to date have been confined to the High Performance Computing arena. One such technology is Remote Direct Memory Access (**RDMA**), which underlies the new **SMB Direct** protocol, part of SMB version 3.0. By using RDMA as a transport medium for SMB, unprecedented levels of performance and efficiency can be achieved. Chelsio worked closely with Microsoft to enable this functionality via its high performance implementation of RDMA over Ethernet – the Internet Wide Area RDMA Protocol (**iWARP**).

The iWARP protocol uses the tried and tested **TCP/IP** as its underlying transport protocol. As a result, iWARP enjoys all the benefits of familiar TCP/IP stack (routability across subnets, robustness, scalability and reliability). While iWARP can utilize the new Data Center Bridging Ethernet technologies, it does not require them. Hence it can function with standard switching technology and allows decoupling of the server and switch refresh cycles.

The Chelsio implementation of iWARP makes use of its high performance TCP Offload Engine, which fully handles the transport stack in hardware, with virtually no host CPU involvement. This frees up the CPU to do more of the real application work. In addition to RDMA, Chelsio's unified adapters support iSCSI, FCoE, TCP and UDP sockets offload as well as normal high performance networking and virtualization functions.

One of the main advantages of the new SMB 3.0 implementation is that once the network adapter driver is installed, all these features are automatically enabled. Furthermore, with the new multi-channel SMB technology, Windows can choose the best protocol to use at any time, as well as aggregate traffic over multiple different links using different protocols. The combination of Chelsio's T4 technology and Microsoft's SMB 3.0 therefore results in highly available solution that can move large amount of data at high speed with minimal CPU utilization.

Scalable Solution

Chelsio's iWARP functionality employed by SMB 3.0 provides the scalability of TCP to build first class solutions that work in multi-tiered switching in LAN environments, without the need for expensive "Converged Ethernet" switches where legacy switches can be used. Unlike IB and RoCE, iWARP is routable, robust against packet loss, and goes over long distances, so it can be used in routed LAN/WAN environments.

Comprehensive Solution

In addition to the iWARP functionality employed by SMB 3.0, Chelsio provides:

- Stateless offload for both IPv4 and IPv6 with full scalable networking support
- Hyper-V support
- TCP offload (Chimney) implemented in hardware
- Network Direct RDMA for HPC clustering
- iSCSI Initiator with Boot capabilities
- FCoE Initiator with Boot capabilities

Chelsio's complete solution enables the release of the full value and feature set of Server 2012. Chelsio further offers one of the most complete set of adapters in the industry. These include adapters with 2x10G ports, 4x10G ports, combination of 2x10G ports and 2x1G ports, and a 4x1G port card. Various physical interfaces are also supported such as SFP+, QSFP+, CX4, and RJ-45.

High Performance

Chelsio's solution is the highest performance and the most feature complete solution for the Windows Server 2012 platform today.

Windows SMB 3.0 Scaling

To illustrate the scalability and performance characteristic of this solution, a sample of benchmark results using a 48 node cluster test bed are presented below.

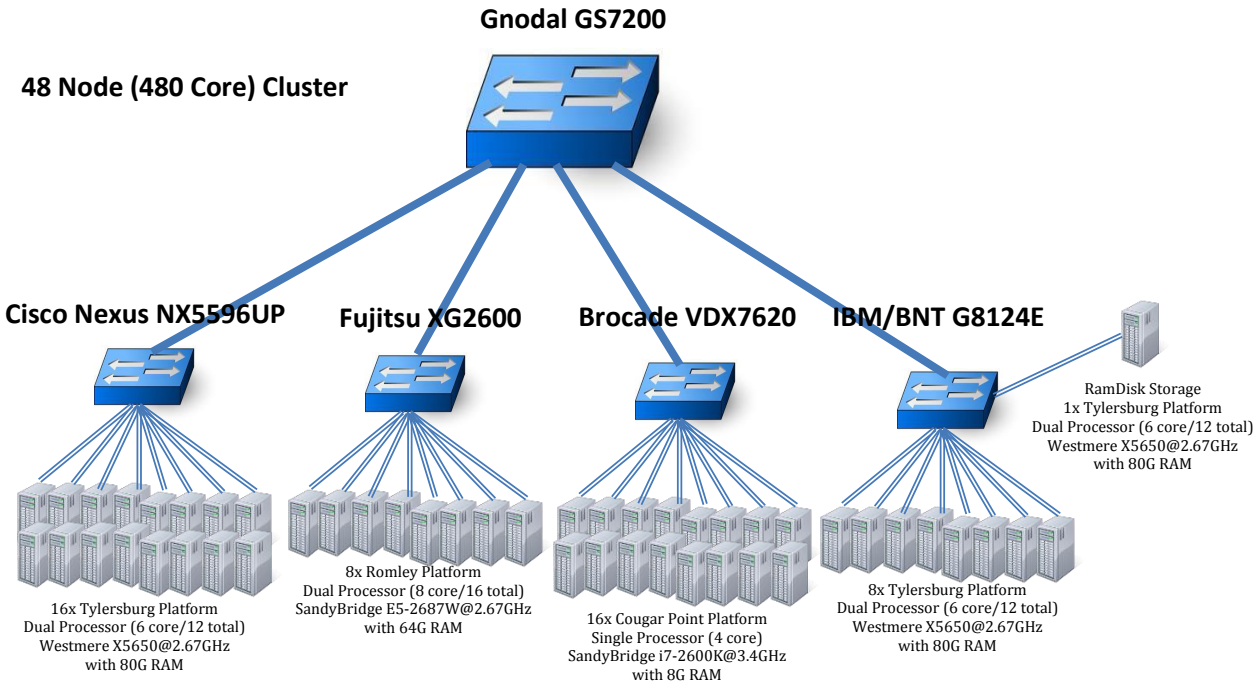

Figure 1 – Test Topology with 48 Nodes

The figure above shows the 10Gbps test bed setup. It consists of 48 server nodes connected through a multi-tier topology using Ethernet switches from different vendors.

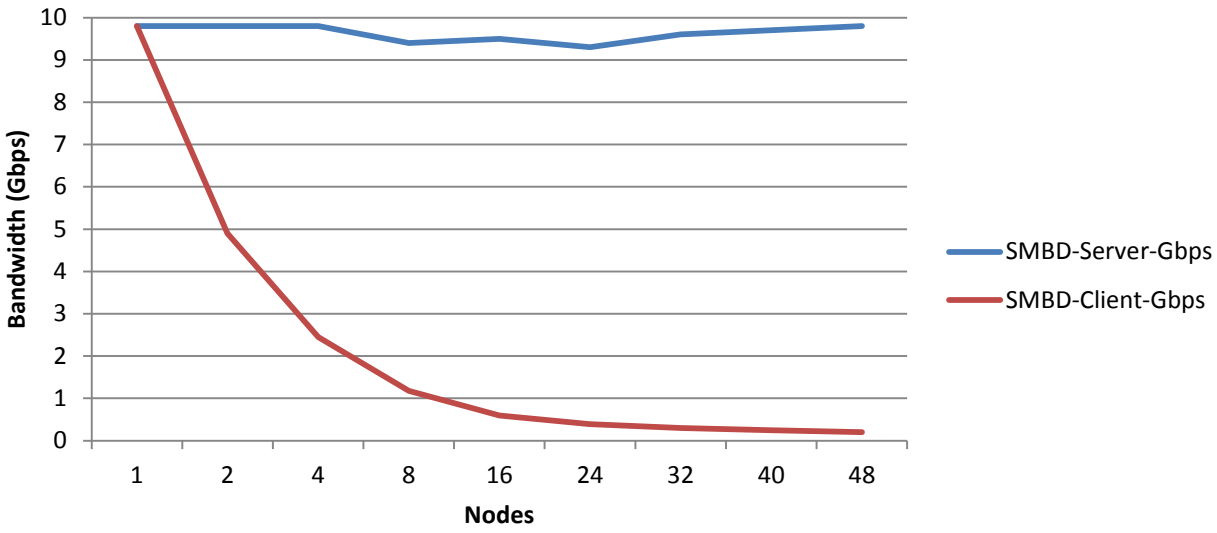

Figure 2 – Bandwidth at Server and bandwidth per Client vs. Number of Clients (8K Read Scaling)

The first graph shows how SMB performance is line rate with 1 client nodes, and scales at near line rate as the number of client nodes is increased (the per-client share drops as expected).

Windows SMB 3.0 Back-to-Back

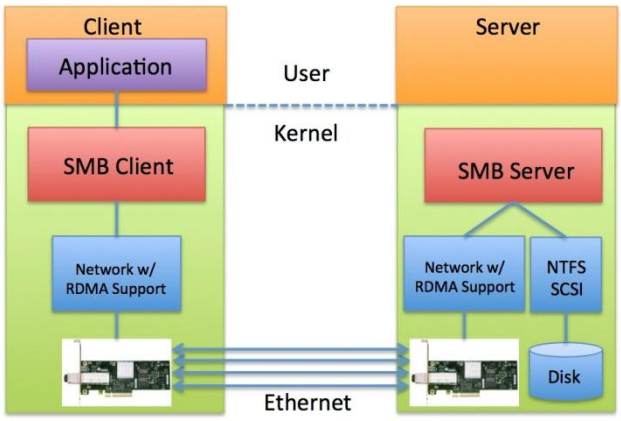

Figure 3 – Back-to-Back Test Setup

The back-to-back configuration above is used to demonstrate the SMB 3.0 multi-link aggregation using RDMA over Chelsio's T4 adapters.

Block Size	GB/sec	IOPS	Server CPU %
8K	2.3 GB/sec	288,890	72%
512K	3.8 GB/sec	7,700	20%

Table 1 – Back-to-Back Test Performance

The aggregate bandwidth is only limited by the Gen2 PCI-e bus bandwidth, while the CPU utilization and IOPS numbers are comparable to the published results for FDR InfiniBand.

Tests run with a Chelsio T420-LP-CR two port card. Test used was the Microsoft *sqlio* test program. Chelsio Driver 4.0.0.14. Disks in this test are four 3GB RAM Disks. Command line used was:

sqlio.txt.exe –s30 –T50 –t16 –o16 –b[8|512] –BN –LS –fsequential –duvxw testfile.da

Summary

The combination of Windows Server 2012 with SMB 3.0 and Chelsio's T4 based network adapters provides highly scalable and highly available storage performance. Leveraging the efficiency and power of RDMA, SMBDirect additionally reduces host CPU and memory subsystem utilization.