Red Hat Linux v4 I/O Performance Case Study:

Linux Performance with software iSCSI initiator and EqualLogic storage arrays approaches 2 Gbit Fiber-channel arrays running an Oracle 10G OLTP industry standard workload.

Abstract:

Red Hat Enterprise Linux v4 (RHEL v4) supports iSCSI storage using the open source developed iSCSI software initiator for the 2.6.9 kernel. The RHEL v4 iSCSI driver is optimized for high performance iSCSI storage arrays like EqualLogic's PS 3800XV. This paper evaluates the performance of iSCSI compared to 2 Gbit FiberChannel using low level benchmarks (IOzone and aio-stress) as well as an OLTP application workload running Oracle 10G. Two comparable storage environments were configured, each with 14 disk RAID 0+1 hardware stripe. The RHEL4 v4 results show that a server with 2, 1-Gbit NICs running the Linux iSCSI software initiator connected to a pair of EqualLogic PS 3800XV arrays can achieve within 2% the throughput and within 6% of the system overhead of traditional 2Gbit FiberChannel HBAs even when running complex application workloads under Oracle 10G.

Methodology:

With the I/O benchmarks we measured the performance differences between 1 to 4 file systems varying the file size from 1MB to 4GB. We then tested 1KB-to-1MB transfer sizes doubling each parameter separately. The results measure throughput as either MB/sec or IO/sec but do not demonstrate the CPU utilization differences between iSCSI and FiberChannel configurations. To measure system/driver overhead, we choose to use a single instance of Oracle 10G R2 running against a 15 GB database. This part of the evaluation drove the system close to CPU saturation by driving the Oracle 10G database application with an appropriate number of clients (80-100 in this case.) We then tuned Oracle's memory and I/O parameters to approach CPU saturation. Any differences in throughput and utilization should be visible in higher system or idle time.

Configuration:

The iSCSI configuration included two EqualLogic PS 3800XV arrays configured with 16, 15k RPM SAS disk drives. On each or the two arrays 14 physical drives were configured as RAID 0+1 leaving 2 drives as spares. The FiberChannel configuration used a QLogic 2300 connected to a pair of Hewlett Packard MSA 1000 controllers with 14, 15k RPM ultraSCSI drives. On each MSA 2, 14-disk SCSI shelves were connected and configured with the same RAID 0+1 stripe to closely match the iSCSI setup. The server under test was the latest Intel Woodcrest 3 Ghz, 64-bit dual-core Servers with 4 CPUs and 16 Gbytes of memory. The hardware setup included:

- 2-socket dual-core 3Ghz Woodcrest, 16 GB memory, 2 PCIX for both dual Fiber and dual NICs.

- 1 QLogic QLA2400 2Gbit FC, 2 MSA1000s with 28 RAID 0+1 15k rpm disks.
- 2 Intel PRO/100 MT NICs with 2 EqualLogic SAS array's with 28 (active) RAID 0+1 15k rpm ultraSCSI disks.
- All three ethernet interfaces on the arrays were configured.

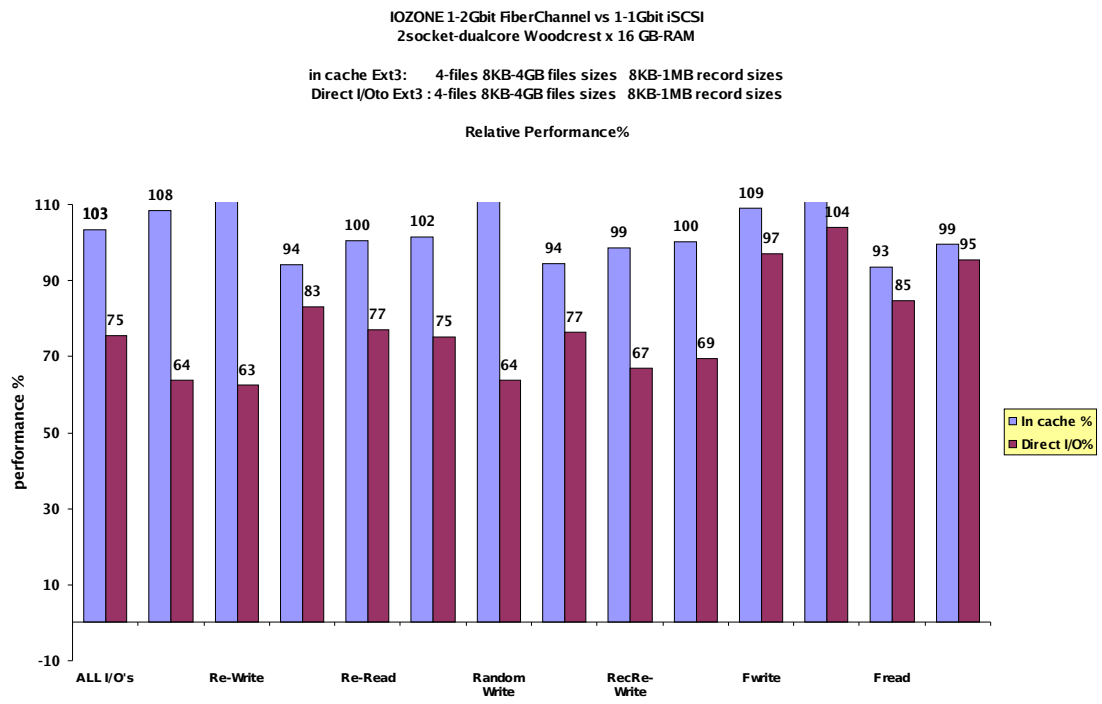
The software setup included:

- Red Hat Enterprise Linux 4, update 4 (RHEL4 U4) using routing for the multiple Intel PRO/1000 MT NICs.
- 2.6.9-42.ELsmp #1 SMP Wed Jul 12 23:32:02 EDT 2006 x86_64 GNU/Linux)
- iSCSI software initiator from RHEL4 - iSCSI-initiator-utils-4.0.3.0-4
- Oracle 10G R2 10.0.1.2 for 64-bit x86_64 architecture

RAID 0+1 configurations were used to maximize performance and provide very good high availability. Two 50GB volumes were created; one for the Oracle log and one for data. The Oracle log and data volumes were bound to the two arrays in the group. Four separate 100GB LUNs(volumes) were allocated for the IOzone tests.

Results:

Chart 1 shows the comparison of a single 1-Gbit Intel PRO/1000 MTNIC versus QLogic FiberChannel 2-Gbit to a single 14 disk array. The chart shows the relative performance of iSCSI as a ratio of the performance of FiberChannel, based on a FiberChannel performance at 100%.



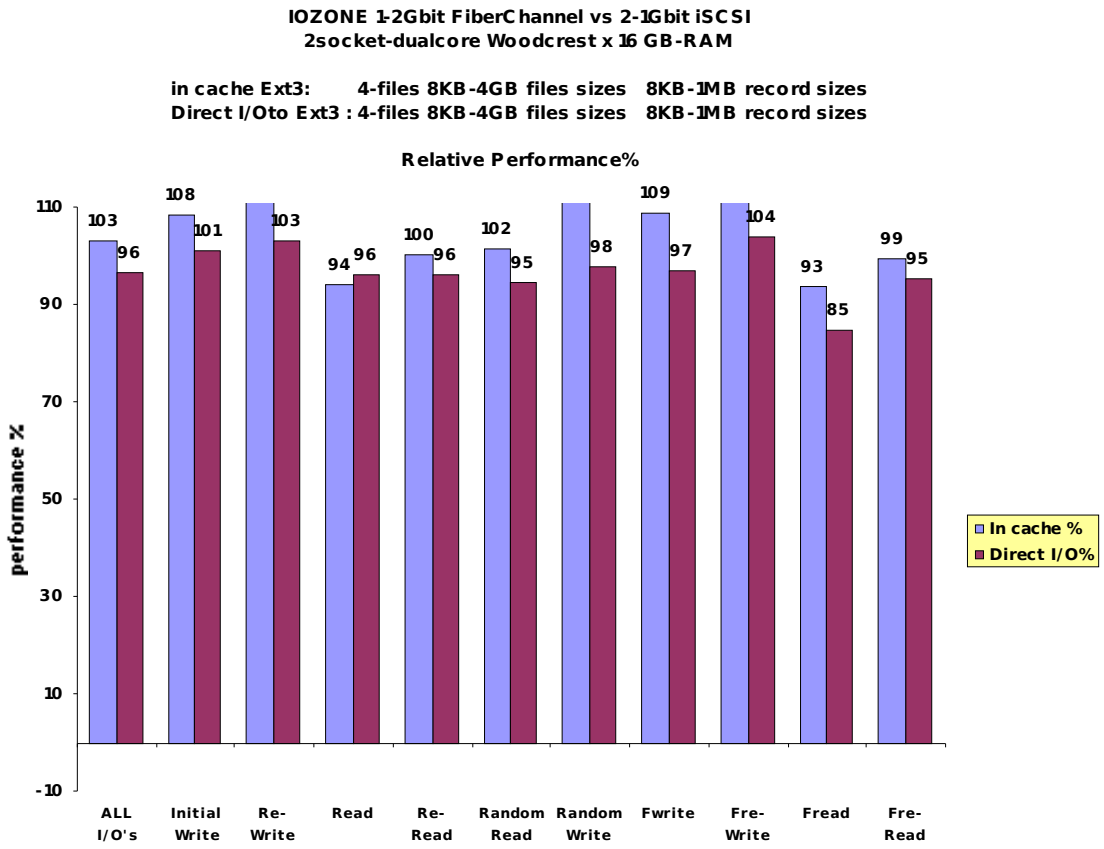
These results show that in throughput tests a single Gb network link can become a bottleneck as compared to a 2Gb FC HBA.

To match 2-Gbit FiberChannel, 2, 1Gbit NIC cards are needed. This is still a cost effective trade-off since most servers today come with 2, 1Gbit cards. Also each EqualLogic array provides up to 3 independent 1-Gbit storage connections helping the server and workload optimize multiple outstanding requests at the lowest latency.

To spread I/Os from the server simple routing information was added into the /etc/rc5.d/K89iscsi file. With this we were able to effectively use the 3 ethernet ports available on each of the EqualLogic storage arrays as follow;

```
route add -host 10.10.10.1 eth0
route add -host 10.10.10.2 eth0
                :
                :
route add -host 10.10.10.8 eth1
```

Chart 2 shows the results of running the benchmark with 2-NICs with IOzone using AIO against 4 LUNs on the iSCSI array and against 4 LUNs on the FiberChannel array improved throughput of iSCSI. Future results will investigate the use of MPIO instead of routing and using QLogic's iSCSI HBA drivers.



Oracle 10G Application Performance:

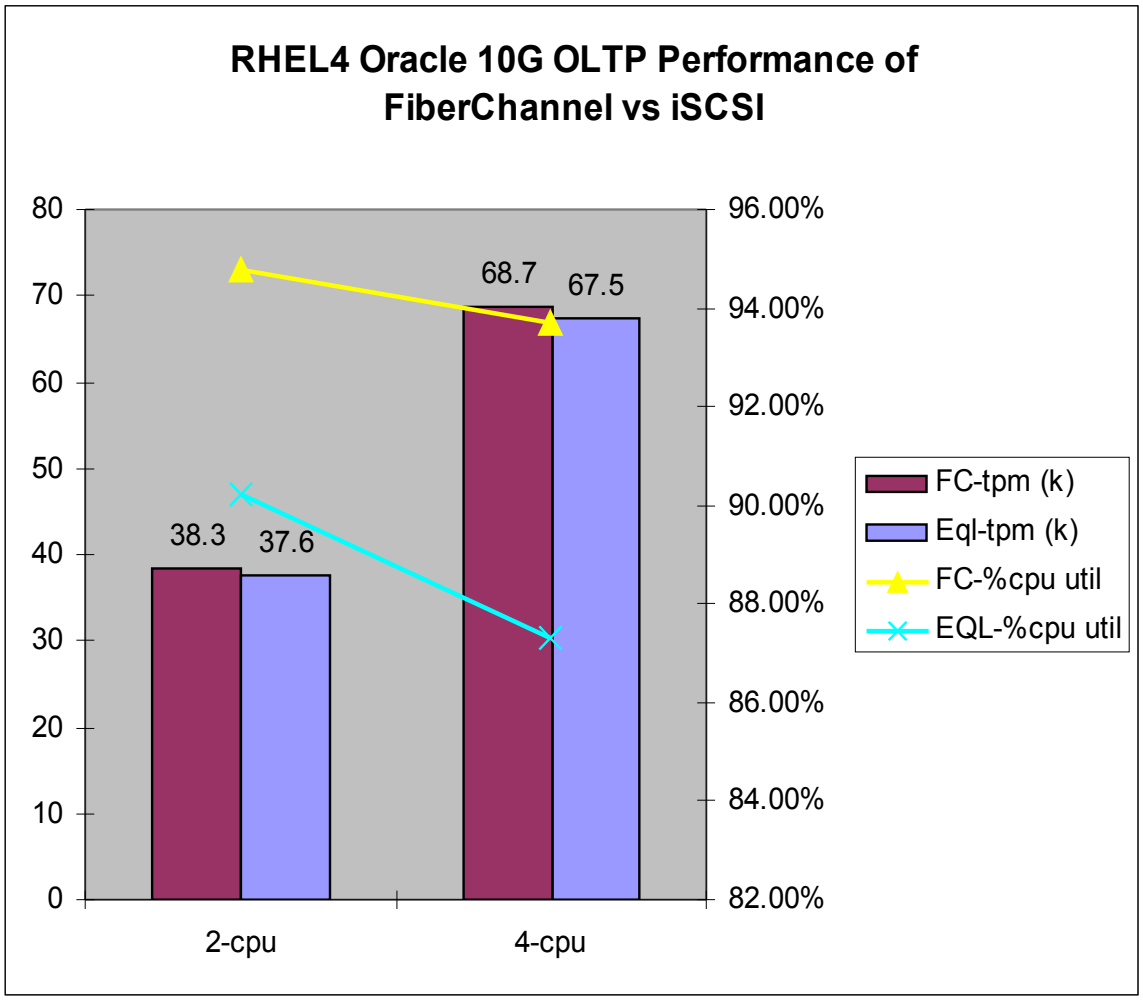
The performance on the IO benchmarks (like IOzone) do not accurately simulate application performance. Therefore, we built an Oracle 10G OLTP workload on 4 of the LUNs tested above using the EXT3 filesystem. We then tuned Oracle 10G support to use 7.8 GB of SGA, roughly ½ the 16 GB of total memory on the server and added clients to attempt to saturate the CPU, memory and I/O capacities to mimic a real application environment.

In Red Hat Enterprise Linux v4, EXT3 now supports both Direct I/O (DIO) and Asynchronous I/O (AIO) simultaneously, giving it the full capabilities of raw, while maintaining file system semantics. Asynchronous I/O is implemented in the kernel and used by Oracle 9i and 10G to optimize the concurrency of queuing multiple I/O requests to the storage device, allowing the application code to continue processing until the point where it simply must wait for the I/O requests to complete. Asynchronous I/O (AIO) will guarantee integrity at I/O completion.

To take advantage of these optimization under Red Hat Enterprise Linux, the following tuning was added to the init.ora file;

set filesystemio_options=directio
or
set filesystemio_options=async
or both
set filesystemio_options=setall

Chart #3 shows an Oracle 10G database running OLTP transactions to EXT3.



The result above shows that Oracle 10G OLTP performance on the iSCSI software initiator with EqualLogic PS3800XV arrays, performs within 2% of 2Gbit Fiber Channel as measured in transactions per minute (tpm.) The sole difference in system utilization with the same 80 clients driving the server, was within 4.6% for 2-CPUs and 6.4% for 4-CPUs. This performance was achieved on a difficult OLTP workload running greater than 90% CPU saturation on the most recent Intel Woodcrest based 4-CPU server and is therefore a good indicator of worst case overhead for a real database application.

Conclusion:

Red Hat Enterprise Linux v4 ships with the improved Open Source iSCSI driver which can be used with iSCSI storage arrays like EqualLogic's PS 3800XV to match 2-Gbit FiberChannel performance. The results show that 2, 1Gbit NICs, when configured with IP routing can achieve equal or better IOzone throughput for 8 different types of IO streams (Write, Rewrite, Read, Re-Read, sequential and random). After establishing a baseline using the IOzone, the same I/O subsystem was able to run Oracle 10G using an industry standard OLTP workload on a 15 GB database to mimic a real customer load. By using standard Oracle tuning (SGA memory sizes and filesystem tuning parameters), we were able to drive both FiberChannel and iSCSI backend systems to greater than 90% of cpu saturation. The Oracle 10G R2 results, in transaction/second, show the combination of Linux software iSCSI initiator with EqualLogic PS 3800XV arrays can achieve within 2% of the throughput and within 6% of the system overhead of traditional 2Gbit FiberChannel.

Future work:

We plan to measure the performance effects due to enhancements in RHEL4.5 and RHEL 5.0 versions of the iSCSI software initiator. RHEL4.5 and RHEL5 also support both the QLogic iSCSI driver and the Intel IOAT technology. Both are expected to perform network based disk I/O with less system overhead and higher throughput. Tests with multi-path IO using QLogic iSCSI HBAs will also be performed.

Acknowledgements:

This was a joint project between Red Hat and EqualLogic. Primary project lead for Red Hat is John Shakshober with network setup and analysis from Barry Marson.