Red Hat Performance Briefs

Optimizing Fusion ioMemory on Red Hat Enterprise Linux 6 for Database Performance Acceleration

Sanjay Rao, Principal Software Engineer

Version 1.0

August 2011

redhat.

1801 Varsity Drive™
Raleigh NC 27606-2072 USA
Phone: +1 919 754 3700
Phone: 888 733 4281
Fax: +1 919 754 3701
PO Box 13588
Research Triangle Park NC 27709 USA

The GPG fingerprint of the security@redhat.com key is:
CA 20 86 86 2B D6 9D FC 65 F6 EC C4 21 91 80 CD DB 42 A6 0E

Send feedback to refarch-feedback@redhat.com

Table of Contents

1 Executive Summary

This paper examines the performance characteristics of Fusion ioMemory modules on Red Hat Enterprise Linux 6. Testing was done in two phases. Phase one was running different transfer sizes and queue depth to test the limits of the drive. In Phase two, the drives were used for running database workloads with different database engines and different types of database workloads. The workload results obtained on the Fusion ioMemory modules were compared with results obtained with 4G Fibre Channel storage.

It is important to note that given enough hardware, fibre channel storage is capable of delivering the throughput results of Fusion ioMemory drives. This comparison is based on the storage capacity.

Red Hat partnered with Fusion ioMemory for this effort. Fusion ioMemory provided the hardware required for the testing and also provided tuning guidelines. With this collaboration, Red Hat was able to show different use cases for the Fusion ioMemory drives to obtain significant performance gains.

2 Test Configuration

2.1 Hardware

Server	4 Socket – 16 Cores (with Hypertheads) Intel – Xeon X7650 @ 2.26 GHz 128 GB RAM (32 GB per NUMA node)
Fibre Channel Storage	2G controller cache / 2G Data cache Total storage - 3.5 TB 28 Physical Disks – 10000 RPM
Fusion ioMemory	4 - Fusion ioDrive Duo 1.28TB drives 4 - Fusion ioDrive Duo 320GB Total storage – 6.4 TB

Table 1: Hardware configuration

2.2 Software

Operating System	Red Hat Enterprise Linux 6.1 (kernel-2.6.32-131.0.15.el6.x86_64.rpm)
Fusion ioMemory	Firmware v5.0.6, rev 101583 Fusion ioMemory driver version: 2.3.0 build 281

Table 2 : Software configuration

3 Testing – Phase 1

3.1 I/O Characterization

The goal of the first phase was to understand the I/O characteristics of the Fusion ioMemory. The memory modules were configured and the firmware and drivers were updated to the versions listed in Table 2. The following options were added to the driver module configuration file, /etc/modprobe.d/iomemory-vsl.conf. These options were added based on the recommendations of the Fusion ioMemory tuning guide.

The following option coalesces interrupts by waiting before sending an interrupt.

```
options iomemory-vsl tintr_hw_wait=50
```

The following option enables MSI improving cpu efficiency.

```
options iomemory-vsl disable_msi=0
```

The linux devices created for the Fusion ioDrives were formatted using ext4, the default file system for Red Hat Enterprise Linux 6.1. The I/O tests used DirectIO and asynchronous IO. DirectIO was used to bypass the file system cache and therefore run at disk speed and asynchronous IO was used to queue multiple IOs for each device. Sequential and Random IO with different transfer sizes and queue depths were run for this test.

3.2 Results

The transfer size of the I/Os was increased from 2k to 1M while the peak throughput was measured during each run. Table 3 lists the peak throughput in MB/sec and table 4 lists the corresponding I/O Operations per Seconds (IOPs) for the different transfer sizes.

Transfer Size	Sequential Writes	Sequential Reads	Random Writes	Random Reads
2k	4777	6764	3412	2639
4k	7717	8385	6334	5783
8k	10027	11250	8670	6540
16k	10060	11982	9996	9139
32k	9909	12433	9934	9391
64k	9699	12129	9691	12016
128k	9667	12873	9611	12390
256k	9711	12872	9616	12614
512k	9651	12871	9684	12620
1024K	9662	12854	9607	12630

Table 3 : Peak Throughput for Different I/O Transfer Sizes

For transfer sizes of 8K or more, the drives can achieve peak throughput of ~10 GB/s for sequential and random writes and peak throughput of ~12GB/sec for sequential and random reads. This is highlighted in gray in Table 3.

I/O size	Sequential Writes	Sequential Reads	Random Writes	Random Reads
2k	2388488	3382185	1706205	1319743
4k	1929280	2096151	1583569	1445821
8k	1253313	1406300	1083795	817511
16k	628748	748876	624777	571215
32k	309649	388523	310432	293460
64k	151548	379047	151428	187753
128k	75520	100569	75086	96796
256k	75869	100560	75127	98544
512k	75395	100558	75657	98594
1024K	75483	100423	75055	98674

Table 4: IOPs Corresponding to Peak Throughput

The data in table 4 proves that that with a 8K transfer size, the drives can do in million IOPs or greater depending on the IO type. For smaller transfer sizes, the IOPs count exceeds 2 million for specific IO types as highlighted in the gray portion of the table.

3.3 What does it all mean?

It simply means that Fusion ioDrives perform remarkably well for small as well as large transfer sizes. Accordingly, it means Fusion ioDrives can be deployed into any environment and provide significant performance improvements to its applications

Another performance plus of these drives are their random I/O characteristics. Most I/O subsystems including traditional hard drives, deliver good throughput with sequential I/O but are incapable of delivering the same rates when it comes to random I/O due to hardware limitations presented by spinning disks and read/write head on the drives. Tables 3 and 4 highlight how Fusion ioDrives deliver similar throughput and IOPs with random IO as they do with Sequential I/O. Most applications have random data access patterns and traditionally the only way to ensure good performance was to buy more hardware thus adding to the cost of

- the acquisition of additional hardware
- building large data centers
- increasing lab power and cooling requirements
- increasing man hours in hardware maintenance.

Fusion ioMemory delivers the same throughput and storage capacity with a much smaller data center footprint providing substantial savings.

4 Testing Phase 2 (Workload Testing)

In this phase, the Fusion ioDrives were used to store files for database applications. The data collected in the first phase highlighted the low latency high bandwidth of the memory modules. Phase demonstrates how these drives can be deployed in database applications using various methods to take advantage of these characteristics and benefit from performance improvements.

4.1 Testing Methodology

4.1.1 Test 1 – Entire Database

Database applications were configured and executed on the hardware with fibre channel storage. In the case of the online transactions processing (OLTP) application, the transaction rates were collected as user counts increased until the server reached a saturation point and could not scale any further. The server is considered saturated when the I/O sub-system can no longer process more I/O resulting in I/O waits. The same database application was then configured and executed on the Fusion ioDrives.

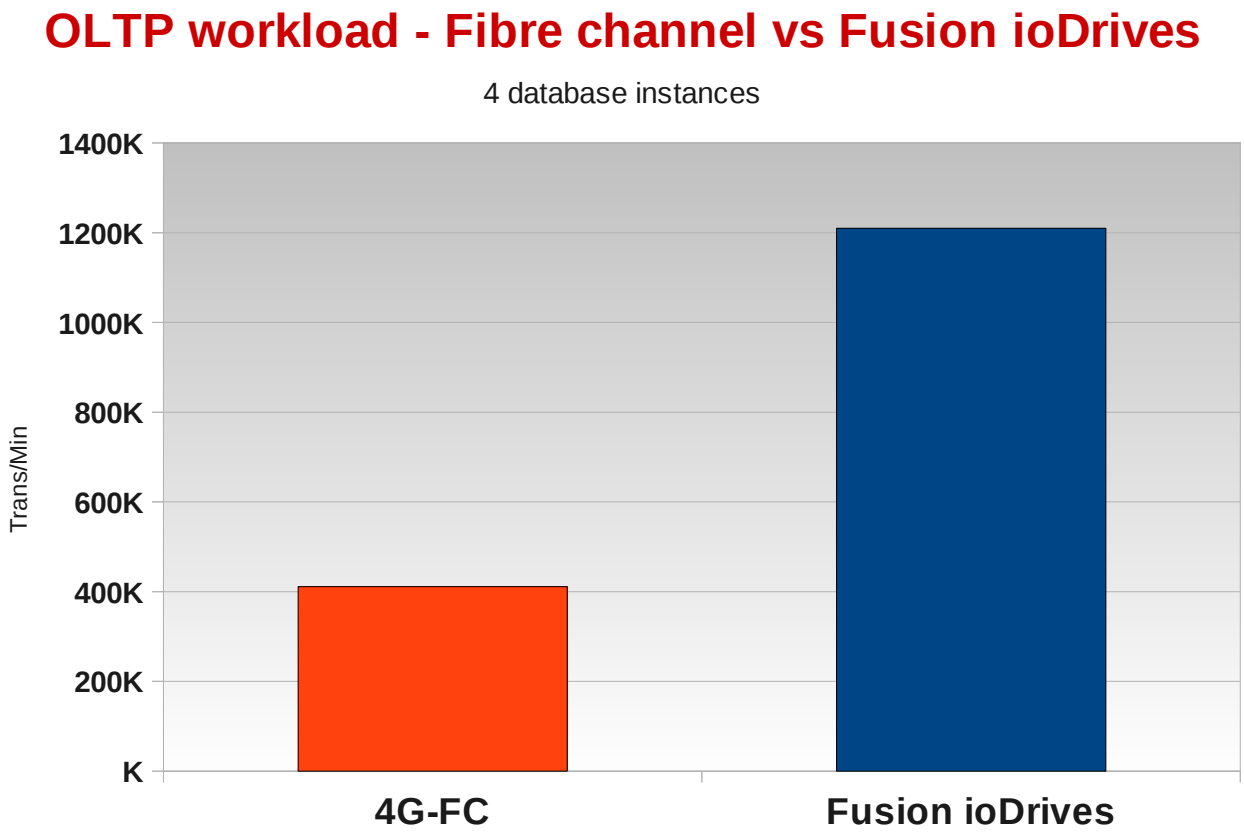

Figure 1

Figure 1 compares the peak transaction rate collected on Fibre channel to those collected on the Fusion ioDrives. The same database workload performed on a Fusion ioDrives shows a 3x improvement in transaction rate compared to the Fibre channel storage. Analyzing the system statistics for the runs reveals how the Fusion ioDrives help in greatly improving the transaction rate.

Vmstat output of Fibre channel run

r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	id	wa	st
62	19	5092	44894312	130704	76267048	0	0	8530	144255	35350	113257	43	4	45	9	0
3	20	5092	43670800	131216	77248544	0	0	6146	152650	29368	93373	33	3	53	11	0
21	27	5092	42975532	131620	77808736	0	0	2973	147526	20886	66140	20	2	65	13	0
7	20	5092	42555764	132012	78158840	0	0	2206	136012	19526	61452	17	2	69	12	0
25	18	5092	42002368	132536	78647472	0	0	2466	144191	20255	63366	19	2	67	11	0
4	21	5092	41469552	132944	79111672	0	0	2581	144470	21125	66029	22	2	65	11	0
1	32	5092	40814696	133368	79699200	0	0	2608	151518	21967	69841	23	2	64	11	0
4	17	5092	40046620	133804	80385232	0	0	2638	151933	23044	70294	24	2	64	10	0
17	14	5092	39499580	134204	80894864	0	0	2377	152805	23663	72655	25	2	62	10	0
35	14	5092	38910024	134596	81436952	0	0	2278	152864	24944	74231	27	2	61	9	0
20	13	5092	38313900	135032	81978544	0	0	2091	156207	24257	72968	26	2	62	10	0
1	14	5092	37831076	135528	82389120	0	0	1332	155549	19798	58195	20	2	67	11	0
23	24	5092	37430772	135936	82749040	0	0	1955	145791	19557	56133	18	2	66	14	0
34	12	5092	36864500	136396	83297184	0	0	1546	141385	19957	56894	19	2	67	13	0

Examining the I/O and cpu utilization in the `vmstat` output during the Fibre Channel run, the I/O is highlighted with the yellow block while the CPU utilization is highlighted in blue. The “bo” column (blocks out) in yellow shows peak write rates of 150 MB/s to the Fibre Channel storage. As the peak rates are achieved, the CPU utilization block shows I/O waits under the “wa” column and consequently resulting in “id” (idle) times when cpus spin in user space.

Now, lets review the vmstat data from the Fusion ioMemory run.

Vmstat output of Fusion ioMemory run

r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	id	wa	st
77	0	6604	55179876	358888	66226960	0	0	7325	266895	70185	149686	90	7	4	0	0
77	1	6604	50630092	359288	70476248	0	0	6873	306900	70166	149804	88	7	5	0	0
76	3	6604	46031168	360132	74444776	0	0	5818	574286	77388	177454	88	8	4	0	0
81	1	6604	41510608	360512	78641480	0	0	4970	452939	75322	168464	89	7	3	0	0
82	3	6604	35358836	361012	84466256	0	0	4011	441042	74022	162443	88	7	4	0	0
81	3	6604	34991452	361892	84740008	0	0	2126	440876	73702	161618	88	7	5	0	0
79	1	6604	34939792	362296	84747016	0	0	2323	400324	73091	161592	90	6	3	0	0
79	0	6604	34879644	362992	84754016	0	0	2275	412631	73271	160766	89	6	4	0	0
76	2	6604	34844616	363396	84760976	0	0	2275	415777	73019	158614	89	6	4	0	0
61	3	6604	34808680	363828	84768016	0	0	2209	401522	72367	159100	89	6	4	0	0
77	1	6604	34781944	364180	84774992	0	0	2172	401966	73253	159064	90	6	4	0	0
54	4	6604	34724948	364772	84803456	0	0	3031	421299	72990	156224	89	6	4	0	0
80	2	6604	34701500	365500	84809072	0	0	2216	573246	76404	175922	88	7	5	1	0

Comparing the `vmstat` output from both runs, the Fusion ioDrives write throughput peaks at 400-500 MB/s. However in this case, the I/O is not the limiting factor. Examining the CPU utilization block in the Fusion ioDrives `vmstat` output, there are no I/O waits in the “wa”

column and very little idle time in the “id” column so the system is running at its full potential. Using Fusion ioDrives in this situation helped remove the I/O bottleneck presented by the fibre channel storage and allowed the application to fully utilize the processing power of the server.

4.1.2 Test 2 – Database Logs

The data in Test 1 emphasizes the advantage of using Fusion ioDrives to replace traditional Fibre Channel storage for a database application. That test used a database approximately 300GB in size, but in many production environments databases can be significantly larger and it may not be feasible to replace all of the storage with the Fusion ioMemory devices. Test 2 was designed and performed to better address these situations. The database statistics collected during Test 1 revealed that the I/O hot spots, or bottlenecks, were primarily occurring during database logging. The more traditional storage could not keep up with the rate at which the logs were being flushed by the databases, resulting in I/O waits and CPU idle time.

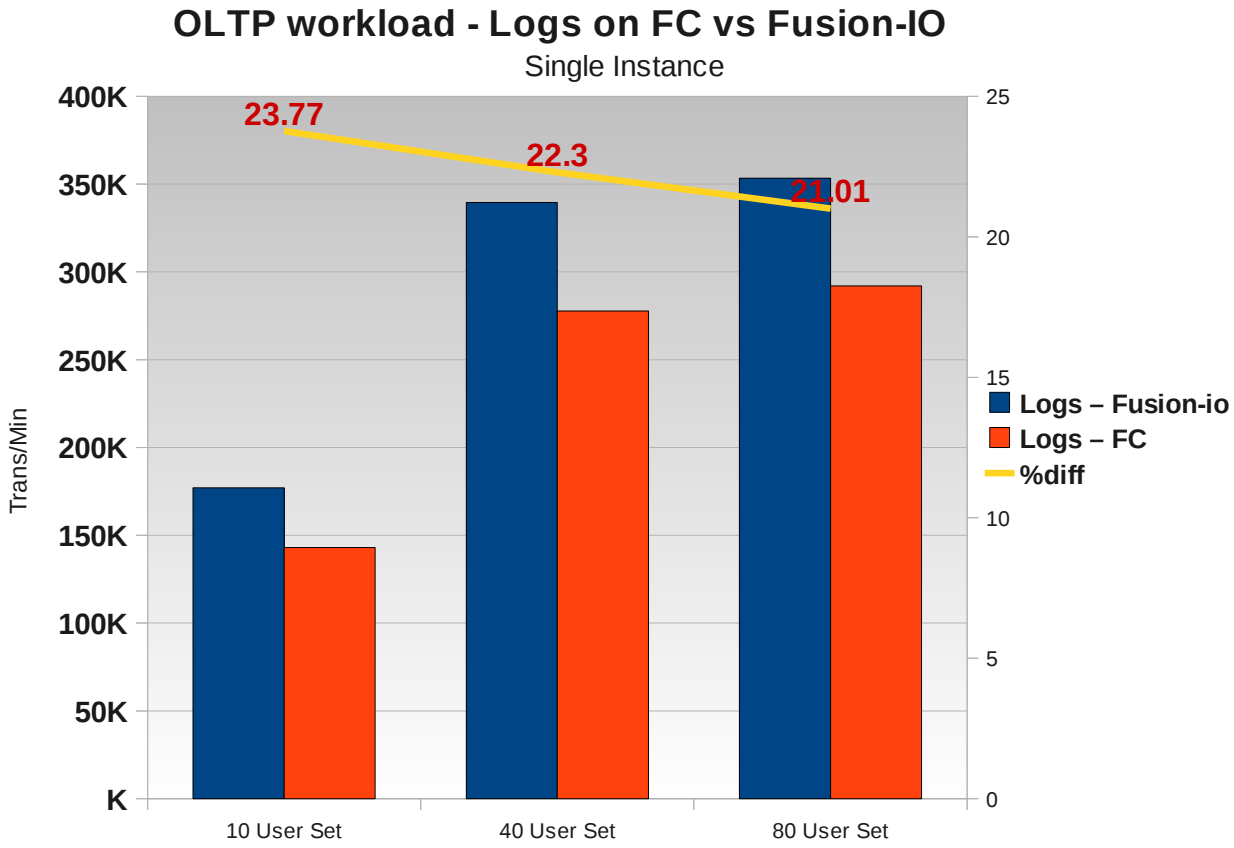

Figure 2

By moving the log files to Fusion ioDrives and repeating the test, performance improved over 20% as shown in Figure 2, obtained by a simple layout reconfiguration of the database log

storage. This demonstrates that the addition of a single ioMemory module to handle the logs of a database system, with the rest of the database on conventional storage, can still have a significant benefit to overall system performance.

4.1.3 Test 3 – Temporary segments

Both Test1 and Test2 showcased how Fusion ioDrives were used to improve the performance of databases running OLTP application. Test 3 demonstrates how to use the drives in a Business Intelligence (BI) database application. BI applications databases are typically much larger than OLTP applications as large amounts of data are sorted and merged based on various analytical formulas to extract information from data warehouses. In these applications, the analytical execution makes heavy use of temporary segments. By moving the temporary segments from traditional Fibre Channel storage to Fusion ioDrives, large performance gains are achieved.

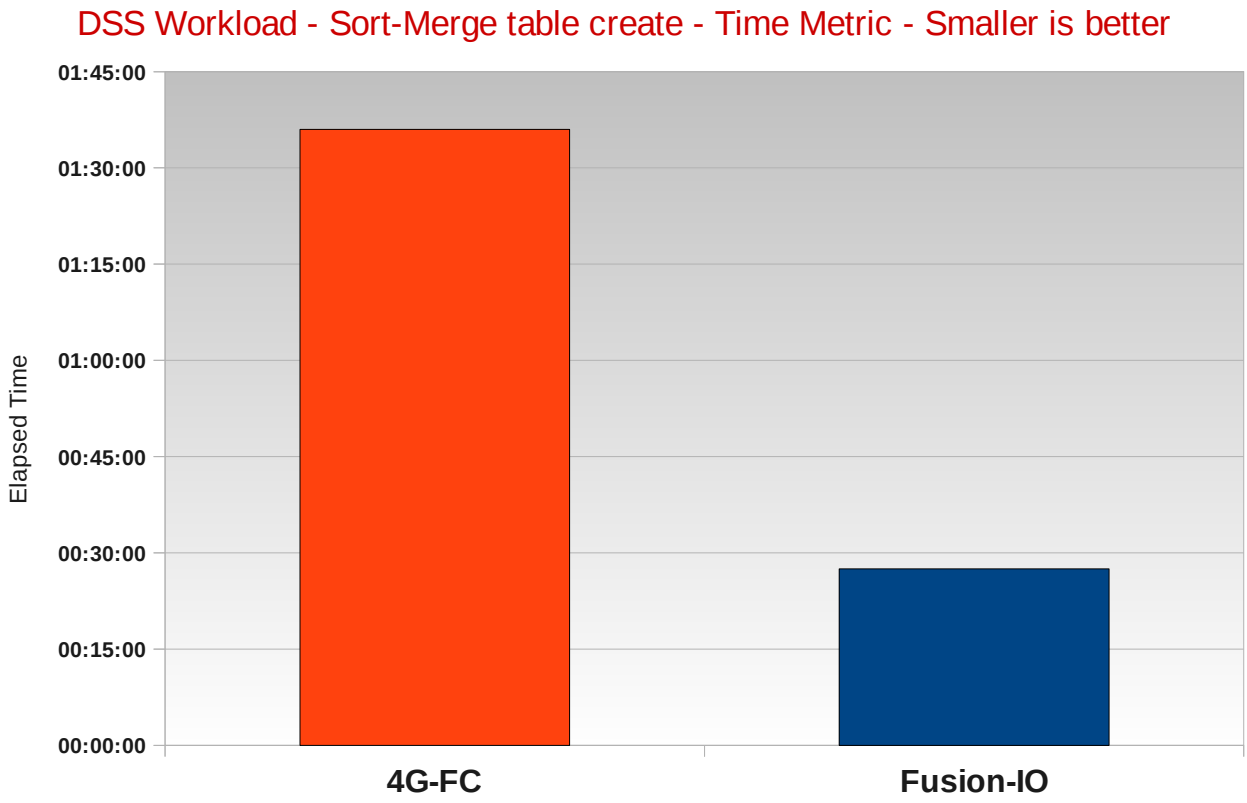

Figure 3: Merge Operation Completion Times

Figure 3 compares completion times for an analytics query using temporary segments where the merge operation is three times faster using a Fusion ioDrives.

5 Conclusion

Data collected during both phases of testing with Fusion ioDrives show the high throughput and low latency characteristics of the Fusion ioDrives and how it can be utilized in different production situations to dramatically improve performance over or in conjunction with traditional storage.

Appendix A: Revision History

Revision 1.0	Tuesday August 23 2011	Sanjay Rao
Initial Release		