# Benchmarking LSI E2660 Storage with RHEL6

**Yemi Adesanya 2/24/2012**

We are now supporting a Linux-based storage server building block for various scientific computing production services including NFS, xrootd and Lustre. The storage unit includes one (or more) Dell 64-bit Intel servers running RHEL6 connected to an LSI E2660 array using 6Gb/sec SAS.

## Specifications

**Dell R610 1U server**
- Dual Intel 'Westmere' 6-core X5650 CPUs @ 2.66Ghz
- 48GB of RAM (6x8GB) @ 1333Mhz
- 6Gb/sec SAS HBA with two mini-SAS port connectors
- RHEL6.2 2.6.32-220.4.1.el6.x86_64 kernel, e2fsprogs-1.42-1, xfsprogs-3.1.1-6

**LSI E2660 4U array chassis**
- 60x2TB 7200RPM nearline-SAS Seagate drives
- Dual redundant RAID controllers each with 2GB cache
- Two mini-SAS ports per controller

## RHEL6

There have been performance and reliability issues with NFS under RHEL5. Improvements were made to the NFS stack in RHEL6. Another RHEL6 feature is support for the native multipath array driver. Each server connects to both of the dual-redundant controllers in the E2660. OS multipathing takes advantage of these multiple connections by seamlessly switching data paths in the event of a connection or controller failure. In the past, LSI-based redundant arrays required installation of additional drivers for RHEL. This functionality is now built into RHEL6 and supported by the array vendor. This eliminates the need to build driver modules for each kernel.

## Controller Configuration

Changes were made to the default array configuration to boost performance. First we increased the controller cache block size to the 32KB maximum. The 4KB default may be more suited to very small IO sizes using faster drives or SSDs. The E2660 arrays were also purchased with the High Performance Tier (A.K.A "Hyper" performance) feature. This option enables enhanced RAID controller algorithms that claim big performance gains across large numbers of drives. Initially we suspected that this option would only payoff if we added expansion trays but we saw a significant increase in throughput with our base 60x2TB configuration.

## RAID6 Configuration

60 Drives in a 4U has become a popular form factor for controller array and JBOD chassis designs. The question we are faced with is how many hotspares do we need and what is a good ratio of data drives to parity drives? For KIPAC NFS and Lustre applications, we chose to divide the 60 disks across four 12+2 RAID6 LUNs leaving 4 drives for global hotspares.

# Raw LUN I/O tests with sgpdd-survey

Before creating filesystems on the RAID6 LUNs, we wanted to get an idea of raw performance. The sgpdd-survey script attempts to simulate concurrent I/O on bare block devices by spawning parallel clients that read and write to a variable number of concurrent 'regions' on the device. These 'regions' represent files. As the the test progresses, the number of client threads (thr) and regions (crg) increases, resulting in more random I/O (seeks). See below for the results of sgpdd-survey using the RAID6 12+2 LUNs on the E2660. Each test scales up to 32 clients reading and writing to 32 regions. The clients transfer data using 1MB records (IO). We tested with the RAID stripe chunk (segment) set to 128k and 256k. The 256k chunk size performed better but a smaller chunk size may scale better for smaller transfer sizes. The tests show we can get 1.5GB/s writing and ~2.8GB/s reading using a single server, multiple 12+2 LUNs and mirrored controller caches. Throughput is even higher if controller cache mirroring is disabled. This is not recommended since controller failover could means cache loss and data corruption.

### single 12+2 LUN with 128k segment

```
total_size  8388608K rsz 1024 crg   1 thr   1 write  275.18 MB/s   1 x 275.22 =  275.22 MB/s read  437.38 MB/s   1 x 437.50 =  437.50 MB/s
total_size  8388608K rsz 1024 crg   1 thr   2 write  520.60 MB/s   1 x 520.75 =  520.75 MB/s read  773.17 MB/s   1 x 773.54 =  773.54 MB/s
total_size  8388608K rsz 1024 crg   1 thr   4 write  879.77 MB/s   1 x 880.25 =  880.25 MB/s read 1509.78 MB/s   1 x 1511.20 = 1511.20 MB/s
total_size  8388608K rsz 1024 crg   1 thr   8 write 1002.29 MB/s   1 x 1002.89 = 1002.89 MB/s read 1526.68 MB/s   1 x 1528.12 = 1528.12 MB/s
total_size  8388608K rsz 1024 crg   1 thr  16 write 1000.22 MB/s   1 x 1000.85 = 1000.85 MB/s read 1529.31 MB/s   1 x 1530.82 = 1530.82 MB/s
total_size  8388608K rsz 1024 crg   2 thr   2 write  557.14 MB/s   2 x 278.66 =  557.33 MB/s read  826.02 MB/s   2 x 413.22 =  826.44 MB/s
total_size  8388608K rsz 1024 crg   2 thr   4 write  880.87 MB/s   2 x 440.68 =  881.37 MB/s read 1148.91 MB/s   2 x 574.88 = 1149.77 MB/s
total_size  8388608K rsz 1024 crg   2 thr   8 write 1001.94 MB/s   2 x 501.28 = 1002.56 MB/s read 1222.52 MB/s   2 x 611.73 = 1223.47 MB/s
total_size  8388608K rsz 1024 crg   2 thr  16 write  997.28 MB/s   2 x 498.97 =  997.94 MB/s read 1297.57 MB/s   2 x 649.32 = 1298.64 MB/s
total_size  8388608K rsz 1024 crg   2 thr  32 write  977.27 MB/s   2 x 488.96 =  977.92 MB/s read 1281.75 MB/s   2 x 644.92 = 1289.84 MB/s
total_size  8388608K rsz 1024 crg   4 thr   4 write  876.71 MB/s   4 x 219.31 =  877.23 MB/s read 1091.34 MB/s   4 x 273.03 = 1092.11 MB/s
total_size  8388608K rsz 1024 crg   4 thr   8 write  999.18 MB/s   4 x 249.95 =  999.79 MB/s read 1149.93 MB/s   4 x 287.69 = 1150.74 MB/s
total_size  8388608K rsz 1024 crg   4 thr  16 write  994.56 MB/s   4 x 248.80 =  995.22 MB/s read 1183.87 MB/s   4 x 296.19 = 1184.77 MB/s
total_size  8388608K rsz 1024 crg   4 thr  32 write  977.03 MB/s   4 x 244.40 =  977.59 MB/s read 1215.53 MB/s   4 x 304.14 = 1216.55 MB/s
total_size  8388608K rsz 1024 crg   8 thr   8 write  992.52 MB/s   8 x 124.14 =  993.12 MB/s read  988.44 MB/s   8 x 123.63 =  989.07 MB/s
total_size  8388608K rsz 1024 crg   8 thr  16 write  992.66 MB/s   8 x 124.16 =  993.27 MB/s read 1056.20 MB/s   8 x 132.12 = 1056.98 MB/s
total_size  8388608K rsz 1024 crg   8 thr  32 write  974.41 MB/s   8 x 121.89 =  975.11 MB/s read 1159.49 MB/s   8 x 145.04 = 1160.35 MB/s
total_size  8388608K rsz 1024 crg  16 thr  16 write  968.93 MB/s  16 x  60.60 =  969.54 MB/s read  821.19 MB/s  16 x  51.36 =  821.69 MB/s
total_size  8388608K rsz 1024 crg  16 thr  32 write  963.68 MB/s  16 x  60.27 =  964.36 MB/s read  721.55 MB/s  16 x  45.12 =  721.89 MB/s
total_size  8388608K rsz 1024 crg  32 thr  32 write  829.42 MB/s  32 x  25.95 =  830.38 MB/s read  460.47 MB/s  32 x  14.40 =  460.82 MB/s
```

### single 12+2 LUN with 256k segment

```
total_size  8388608K rsz 1024 crg   1 thr   1 write  273.33 MB/s   1 x 273.38 =  273.38 MB/s read  440.43 MB/s   1 x 440.54 =  440.54 MB/s
total_size  8388608K rsz 1024 crg   1 thr   2 write  501.39 MB/s   1 x 501.55 =  501.55 MB/s read  829.95 MB/s   1 x 830.38 =  830.38 MB/s
total_size  8388608K rsz 1024 crg   1 thr   4 write  875.00 MB/s   1 x 875.47 =  875.47 MB/s read 1403.47 MB/s   1 x 1404.72 = 1404.72 MB/s
total_size  8388608K rsz 1024 crg   1 thr   8 write 1061.28 MB/s   1 x 1061.97 = 1061.97 MB/s read 1415.62 MB/s   1 x 1416.91 = 1416.91 MB/s
total_size  8388608K rsz 1024 crg   1 thr  16 write 1059.64 MB/s   1 x 1060.34 = 1060.34 MB/s read 1423.02 MB/s   1 x 1424.31 = 1424.31 MB/s
total_size  8388608K rsz 1024 crg   2 thr   2 write  538.35 MB/s   2 x 269.26 =  538.52 MB/s read  821.04 MB/s   2 x 410.86 =  821.72 MB/s
total_size  8388608K rsz 1024 crg   2 thr   4 write  881.38 MB/s   2 x 440.94 =  881.88 MB/s read 1097.05 MB/s   2 x 548.91 = 1097.81 MB/s
total_size  8388608K rsz 1024 crg   2 thr   8 write 1058.47 MB/s   2 x 529.58 = 1059.17 MB/s read 1186.84 MB/s   2 x 593.83 = 1187.67 MB/s
total_size  8388608K rsz 1024 crg   2 thr  16 write 1057.21 MB/s   2 x 528.99 = 1057.99 MB/s read 1244.78 MB/s   2 x 622.90 = 1245.80 MB/s
total_size  8388608K rsz 1024 crg   2 thr  32 write 1030.02 MB/s   2 x 515.34 = 1030.67 MB/s read 1302.11 MB/s   2 x 651.62 = 1303.23 MB/s
total_size  8388608K rsz 1024 crg   4 thr   4 write  877.53 MB/s   4 x 219.51 =  878.03 MB/s read 1019.26 MB/s   4 x 254.98 = 1019.94 MB/s
total_size  8388608K rsz 1024 crg   4 thr   8 write 1057.26 MB/s   4 x 264.49 = 1057.97 MB/s read 1089.49 MB/s   4 x 272.57 = 1090.28 MB/s
total_size  8388608K rsz 1024 crg   4 thr  16 write 1054.63 MB/s   4 x 263.97 = 1055.87 MB/s read 1157.77 MB/s   4 x 289.67 = 1158.68 MB/s
total_size  8388608K rsz 1024 crg   4 thr  32 write 1028.85 MB/s   4 x 257.38 = 1029.51 MB/s read 1190.00 MB/s   4 x 297.74 = 1190.95 MB/s
total_size  8388608K rsz 1024 crg   8 thr   8 write 1050.29 MB/s   8 x 131.38 = 1051.03 MB/s read  929.08 MB/s   8 x 116.20 =  929.57 MB/s
total_size  8388608K rsz 1024 crg   8 thr  16 write 1051.31 MB/s   8 x 131.50 = 1052.02 MB/s read 1057.70 MB/s   8 x 132.30 = 1058.43 MB/s
total_size  8388608K rsz 1024 crg   8 thr  32 write 1023.53 MB/s   8 x 128.02 = 1024.17 MB/s read 1073.88 MB/s   8 x 134.33 = 1074.60 MB/s
total_size  8388608K rsz 1024 crg  16 thr  16 write  954.56 MB/s  16 x  59.71 =  955.35 MB/s read  788.60 MB/s  16 x  49.32 =  789.18 MB/s
total_size  8388608K rsz 1024 crg  16 thr  32 write  922.84 MB/s  16 x  57.71 =  923.31 MB/s read  753.37 MB/s  16 x  47.11 =  753.78 MB/s
total_size  8388608K rsz 1024 crg  32 thr  32 write  762.61 MB/s  32 x  23.85 =  763.24 MB/s read  528.57 MB/s  32 x  16.53 =  528.87 MB/s
```

### Two 12+2 LUNs with 256k segment

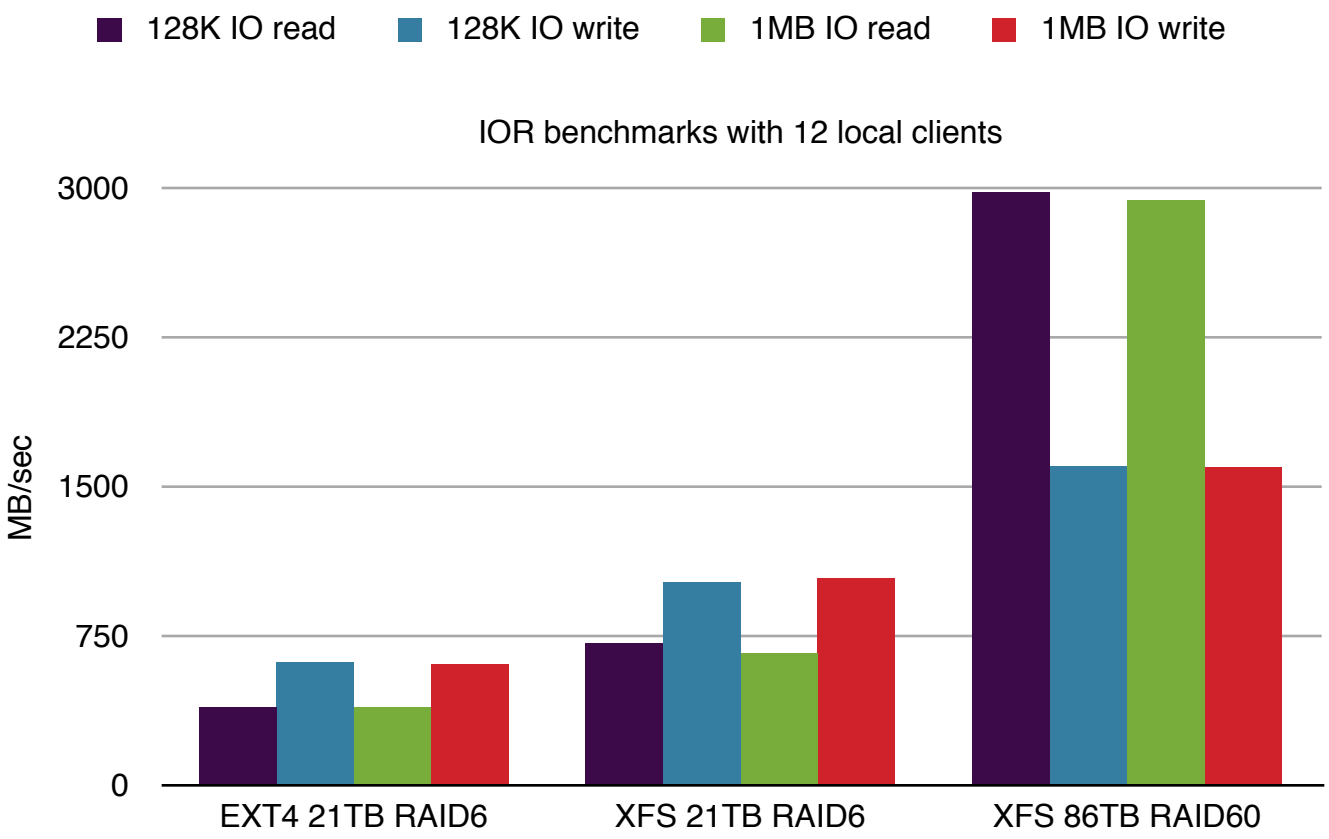
| | | | | |
|---|---|---|---|---|
| total_size 16777216K rsz 1024 crg | 2 thr | 2 write 512.19 MB/s | 2 x 256.25 = 512.50 MB/s read 803.35 MB/s | 2 x 401.76 = 803.53 MB/s |
| total_size 16777216K rsz 1024 crg | 2 thr | 4 write 877.52 MB/s | 2 x 438.88 = 877.76 MB/s read 1683.10 MB/s | 2 x 841.99 = 1683.98 MB/s |
| total_size 16777216K rsz 1024 crg | 2 thr | 8 write 1419.78 MB/s | 2 x 710.31 = 1420.61 MB/s read 2679.92 MB/s | 2 x 1341.05 = 2682.09 MB/s |
| total_size 16777216K rsz 1024 crg | 2 thr | 16 write 1543.15 MB/s | 2 x 771.93 = 1543.87 MB/s read 2839.13 MB/s | 2 x 1420.75 = 2841.49 MB/s |
| total_size 16777216K rsz 1024 crg | 2 thr | 32 write 1513.19 MB/s | 2 x 757.04 = 1514.07 MB/s read 2836.89 MB/s | 2 x 1419.68 = 2839.36 MB/s |
| total_size 16777216K rsz 1024 crg | 4 thr | 4 write 927.04 MB/s | 4 x 231.83 = 927.31 MB/s read 1679.61 MB/s | 4 x 420.12 = 1680.49 MB/s |
| total_size 16777216K rsz 1024 crg | 4 thr | 8 write 1380.82 MB/s | 4 x 345.35 = 1381.42 MB/s read 2260.51 MB/s | 4 x 565.55 = 2262.19 MB/s |
| total_size 16777216K rsz 1024 crg | 4 thr | 16 write 1536.58 MB/s | 4 x 384.32 = 1537.28 MB/s read 2307.82 MB/s | 4 x 577.54 = 2310.14 MB/s |
| total_size 16777216K rsz 1024 crg | 4 thr | 32 write 1522.05 MB/s | 4 x 380.70 = 1522.79 MB/s read 2639.27 MB/s | 4 x 660.36 = 2641.45 MB/s |
| total_size 16777216K rsz 1024 crg | 4 thr | 64 write 1572.73 MB/s | 4 x 393.47 = 1573.87 MB/s read 2769.18 MB/s | 4 x 692.93 = 2771.72 MB/s |
| total_size 16777216K rsz 1024 crg | 8 thr | 8 write 1375.34 MB/s | 8 x 172.02 = 1376.19 MB/s read 2129.23 MB/s | 8 x 266.41 = 2131.27 MB/s |
| total_size 16777216K rsz 1024 crg | 8 thr | 16 write 1530.67 MB/s | 8 x 191.43 = 1531.45 MB/s read 2366.29 MB/s | 8 x 296.01 = 2368.09 MB/s |
| total_size 16777216K rsz 1024 crg | 8 thr | 32 write 1538.67 MB/s | 8 x 192.45 = 1539.61 MB/s read 2481.99 MB/s | 8 x 310.50 = 2483.98 MB/s |
| total_size 16777216K rsz 1024 crg | 8 thr | 64 write 1568.56 MB/s | 8 x 196.21 = 1569.67 MB/s read 2582.01 MB/s | 8 x 323.13 = 2585.07 MB/s |
| total_size 16777216K rsz 1024 crg | 16 thr | 16 write 1529.77 MB/s | 16 x 95.68 = 1530.91 MB/s read 1947.63 MB/s | 16 x 121.90 = 1950.38 MB/s |
| total_size 16777216K rsz 1024 crg | 16 thr | 32 write 1547.21 MB/s | 16 x 96.77 = 1548.31 MB/s read 2279.76 MB/s | 16 x 142.58 = 2281.34 MB/s |
| total_size 16777216K rsz 1024 crg | 16 thr | 64 write 1505.48 MB/s | 16 x 94.17 = 1506.65 MB/s read 2393.49 MB/s | 16 x 149.77 = 2396.39 MB/s |
| total_size 16777216K rsz 1024 crg | 32 thr | 32 write 1565.88 MB/s | 32 x 48.96 = 1566.77 MB/s read 1719.05 MB/s | 32 x 53.76 = 1720.28 MB/s |
| total_size 16777216K rsz 1024 crg | 32 thr | 64 write 1465.38 MB/s | 32 x 45.83 = 1466.67 MB/s read 1687.25 MB/s | 32 x 52.78 = 1688.84 MB/s |
| total_size 16777216K rsz 1024 crg | 64 thr | 64 write 1405.99 MB/s | 64 x 21.99 = 1407.47 MB/s read 1073.51 MB/s | 64 x 16.78 = 1074.22 MB/s |

### Four 12+2 LUNs with 256k segment

| | | | | |
|---|---|---|---|---|
| total_size 33554432K rsz 1024 crg | 4 thr | 4 write 914.23 MB/s | 4 x 228.61 = 914.42 MB/s read 1681.28 MB/s | 4 x 420.43 = 1681.71 MB/s |
| total_size 33554432K rsz 1024 crg | 4 thr | 8 write 1306.76 MB/s | 4 x 326.76 = 1307.03 MB/s read 2629.53 MB/s | 4 x 657.64 = 2630.58 MB/s |
| total_size 33554432K rsz 1024 crg | 4 thr | 16 write 1531.95 MB/s | 4 x 383.11 = 1532.44 MB/s read 2836.78 MB/s | 4 x 709.51 = 2838.02 MB/s |
| total_size 33554432K rsz 1024 crg | 4 thr | 32 write 1523.96 MB/s | 4 x 381.12 = 1524.47 MB/s read 2859.12 MB/s | 4 x 715.10 = 2860.41 MB/s |
| total_size 33554432K rsz 1024 crg | 4 thr | 64 write 1539.49 MB/s | 4 x 385.07 = 1540.26 MB/s read 2872.94 MB/s | 4 x 718.77 = 2875.10 MB/s |
| total_size 33554432K rsz 1024 crg | 8 thr | 8 write 1340.61 MB/s | 8 x 167.67 = 1341.32 MB/s read 2586.23 MB/s | 8 x 323.41 = 2587.28 MB/s |
| total_size 33554432K rsz 1024 crg | 8 thr | 16 write 1517.24 MB/s | 8 x 189.70 = 1517.64 MB/s read 2853.88 MB/s | 8 x 357.01 = 2856.06 MB/s |
| total_size 33554432K rsz 1024 crg | 8 thr | 32 write 1539.37 MB/s | 8 x 192.48 = 1539.84 MB/s read 2871.74 MB/s | 8 x 359.13 = 2873.00 MB/s |
| total_size 33554432K rsz 1024 crg | 8 thr | 64 write 1553.62 MB/s | 8 x 194.32 = 1554.57 MB/s read 2877.55 MB/s | 8 x 359.96 = 2879.71 MB/s |
| total_size 33554432K rsz 1024 crg | 8 thr | 128 write 1553.71 MB/s | 8 x 194.28 = 1554.26 MB/s read 2869.69 MB/s | 8 x 358.89 = 2871.09 MB/s |
| total_size 33554432K rsz 1024 crg | 16 thr | 16 write 1544.27 MB/s | 16 x 96.58 = 1545.26 MB/s read 2845.55 MB/s | 16 x 177.99 = 2847.90 MB/s |
| total_size 33554432K rsz 1024 crg | 16 thr | 32 write 1546.97 MB/s | 16 x 96.74 = 1547.85 MB/s read 2860.91 MB/s | 16 x 178.96 = 2863.31 MB/s |
| total_size 33554432K rsz 1024 crg | 16 thr | 64 write 1553.40 MB/s | 16 x 97.15 = 1554.41 MB/s read 2821.48 MB/s | 16 x 176.50 = 2823.94 MB/s |
| total_size 33554432K rsz 1024 crg | 16 thr | 128 write 1556.51 MB/s | 16 x 97.35 = 1557.62 MB/s read 2869.16 MB/s | 16 x 179.47 = 2871.55 MB/s |
| total_size 33554432K rsz 1024 crg | 32 thr | 32 write 1552.55 MB/s | 32 x 48.55 = 1553.65 MB/s read 2813.40 MB/s | 32 x 88.02 = 2816.77 MB/s |
| total_size 33554432K rsz 1024 crg | 32 thr | 64 write 1558.10 MB/s | 32 x 48.73 = 1559.45 MB/s read 2844.32 MB/s | 32 x 88.92 = 2845.46 MB/s |
| total_size 33554432K rsz 1024 crg | 32 thr | 128 write 1556.11 MB/s | 32 x 48.67 = 1557.31 MB/s read 2828.04 MB/s | 32 x 88.42 = 2829.28 MB/s |
| total_size 33554432K rsz 1024 crg | 64 thr | 64 write 1551.85 MB/s | 64 x 24.26 = 1552.73 MB/s read 2094.57 MB/s | 64 x 32.74 = 2095.34 MB/s |
| total_size 33554432K rsz 1024 crg | 64 thr | 128 write 1457.44 MB/s | 64 x 22.80 = 1459.35 MB/s read 2179.24 MB/s | 64 x 34.07 = 2180.18 MB/s |
| total_size 33554432K rsz 1024 crg | 128 thr | 128 write 1425.46 MB/s | 128 x 11.16 = 1428.22 MB/s read 1565.44 MB/s | 128 x 12.24 = 1566.16 MB/s |

# Filesystems: ext4 vs XFS

Red Hat support ext4 and XFS with RHEL6. The official RHEL ext4 filesystem size limit remains at 16TB which is obviously a problem for our 12+2 LUNs (~21TB). The ext4 utilities (e2fsprogs) distributed with Fedora are capable of handling larger filesystem sizes but this is not a supported solution. XFS is Red Hat's large filesystem product and must be purchased separately. We created ext4 and XFS filesystems on top of two identical 12+2 LUNs (with 256k segment size) on the E2660. The large ext4 filesystem was created using Fedora e2fsprogs-1.42-1 and the XFS filesystem was created using xfsprogs-3.1.1-6. This time we used the IOR filesystem benchmarking utility to measure concurrent I/O. IOR uses MPI to coordinate client processes that write and read to separate files. We tested 1MB and 128k record sizes and XFS outperformed ext4 as shown in the following chart.




## RAID60 with LVM

The E2660 can hold four 21TB 12+2 RAID6 LUNs with 4 drives remaining which may be designated as hotspares. Creating separate volumes may be acceptable, but managing a single filespace would be simpler if we are setting large quotas for users and projects. RHEL6 quota limits are set on a per-filesystem basis. Using LVM, we can stripe the four RAID6 LUNs to create a single 86TB RAID60 logical volume. The IOR performance chart shows the increase in throughput using RAID60. We are now effectively using 56 drives for our filesystem. We are also using both of the redundant E2660 controllers for our RAID60. Each hardware RAID6 LUN is associated with just one controller at any time. By striping across multiple hardware LUNs, we are using both controllers concurrently.