Performance Monitoring in an HP SFS Environment

Roland Laifer

Computing Centre (SSCK)
University of Karlsruhe

Laifer@rz.uni-karlsruhe.de

Outline

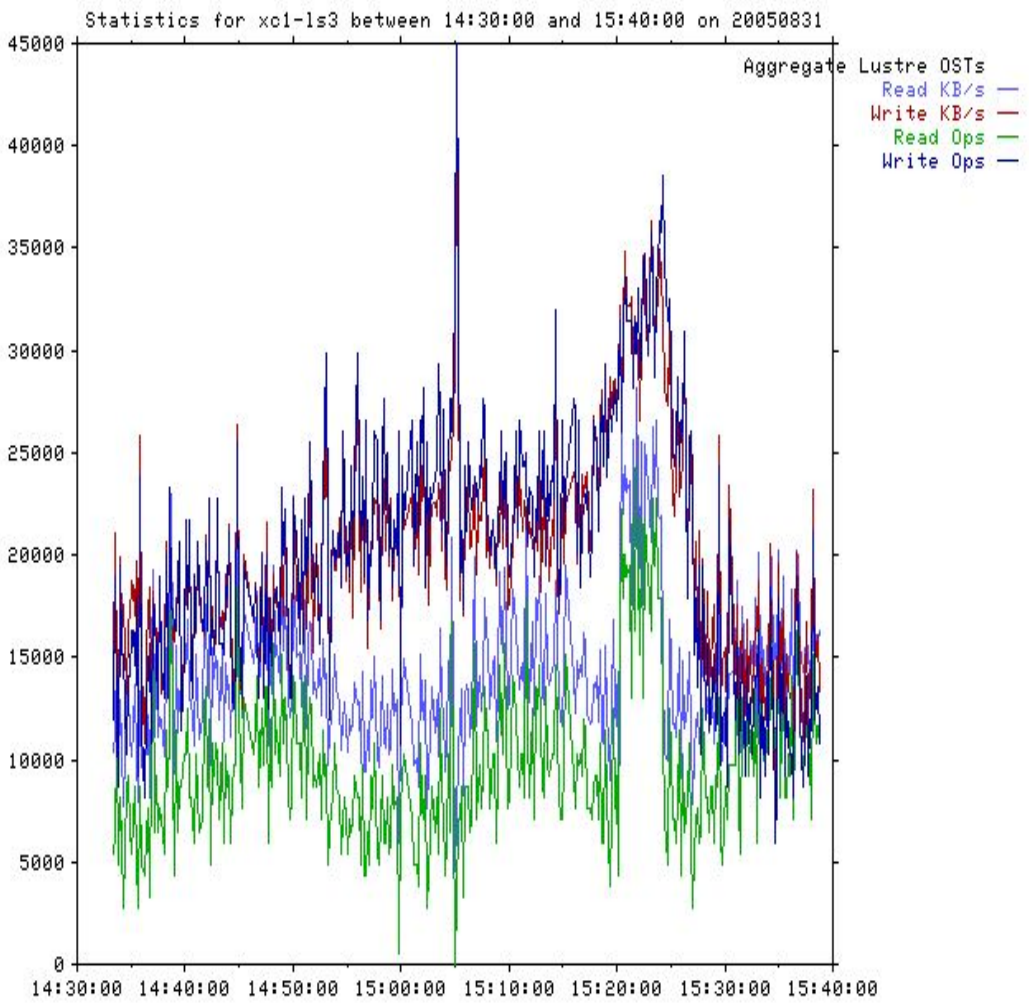
- **Motivation**
- **Performance monitoring on different layers**
- **Examples**

Why performance monitoring?

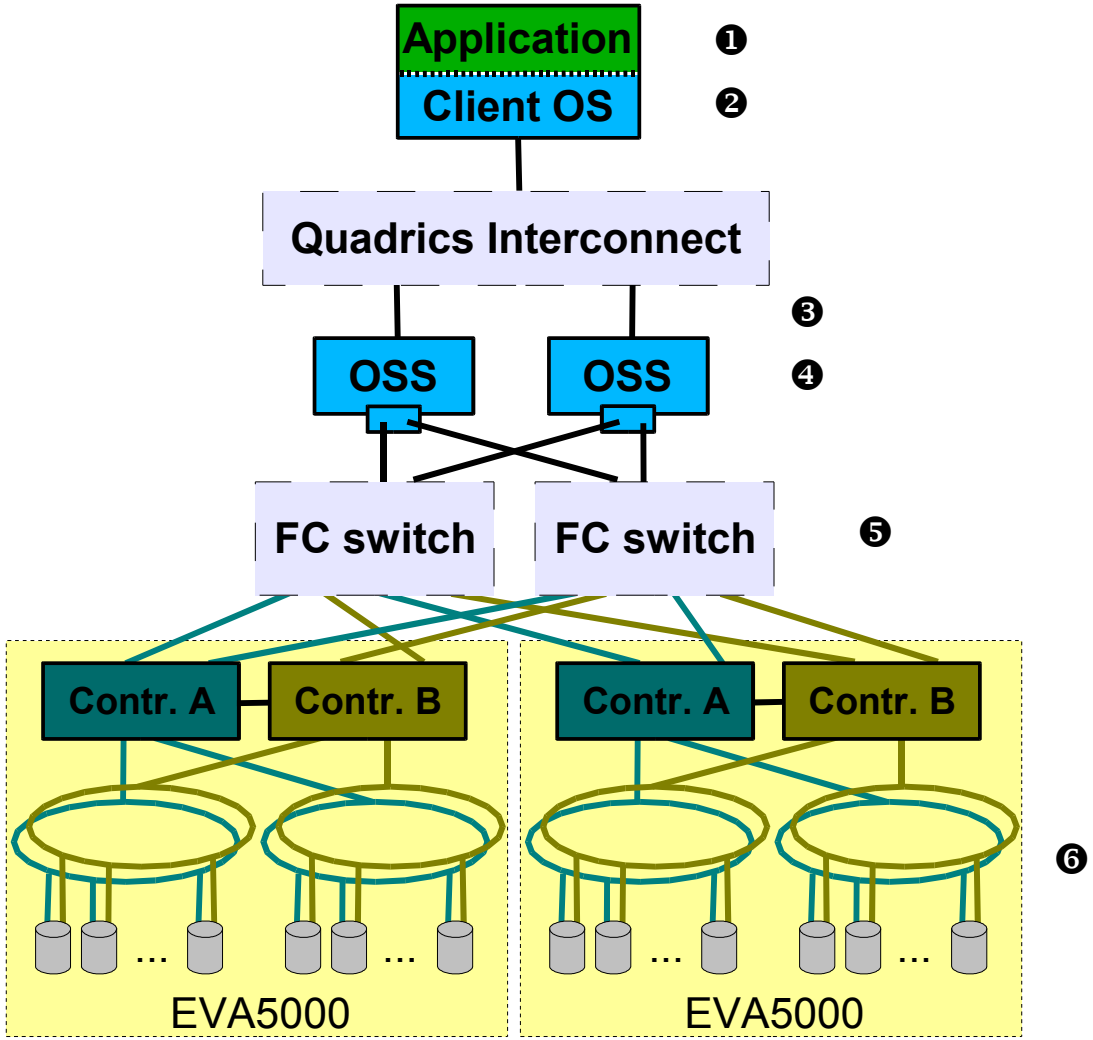
- Identify bottlenecks
- Investigate possible throughput
  - Is unused bandwidth left for additional applications?
- Identify applications with high IO usage
  - Try to optimize the IO behaviour of these applications
- Identify possible software or hardware problems

First example

- **Typical IO on an OSS in production:**
  - **See picture on right**
  - **Created by hpls_plot.sh**
- **But: Which applications are producing most IO?**
  - **About 20 apps are running concurrently**
- **Use collectl to find nodes with high IO usage**
  - pdsh -a collectl -sl -odHx –l LusKBS:1000 -i1 -c100
    - **This shows clients with throughput > 1 MB/s**
- **Use batch system to identify users on these nodes**



Performance monitoring on different layers



Possible tools:

❶ **bonnie++, dd, or ost_perf_check.bash**

❷ **collectl**

❸ **collectl, or qselantest**

❹ **collectl**

❺ **PortPerfShow**

❻ **EVAPerf**

Performance monitoring on the application layer

» Applications for performance measurement

- **bonnie++ -d /lustre/work**

```
------Sequential Output------ --Sequential Input- --Random-
-Per Chr- --Block-- -Rewrite- -Per Chr- --Block-- --Seeks--
Size K/sec %CP K/sec %CP K/sec %CP K/sec %CP K/sec %CP /sec %CP
  8G 13473  99 116666 27 95041  40 12930  99 178616 43 944.6  2
```

- **/usr/opt/hpls/diags/bin/ost_perf_check.bash --parallel --mount-point /lustre/work --remote-shell ssh --clients "xc0n8 xc0n9"**

```
Max Write: 115.44 MiB/sec (121.05 MB/sec)
Max Read:  181.08 MiB/sec (189.87 MB/sec)
```

  - **Displayed units are wrong and should be exchanged**

- **time dd if=/dev/zero of=test1 bs=1M count=10000**

```
real    1m26.824s   (i.e. 115 MB/s)
```

Performance monitoring on the client OS layer

» Monitoring Lustre client performance on command line

➢ /usr/sbin/collectl -sl -oh

```
# Reads ReadKB  Writes WriteKB  Open Close GAttr SAttr  Seek ...
      0      0     310  318156     0     0     2     0     0 ...
     16   1845     316  323993    10    10   103     0     0 ...
```

- **Peaks might be lost because of 10 sec default time interval**

» Long term monitoring with collectl as daemon

➢ Example for collectl.conf file:

```
DaemonCommands = -f /tmp/ -r00:01,7 -m -F60 -scdmxl -oz
```

➢ Start collectl as daemon

- **service collectl start**

➢ Process collected raw file

- **collectl -p xc0n3-20050907-152640.raw -sd -odh**

Performance monitoring on the MDS or OSS

» Quadrics performance

➢ qselantest | grep bytes | grep MB

```
0:  1048576 bytes   1325.26 uSec   791.23 MB/s
```

- This shows possible Quadrics throughput
- Unit is wrong and should be MiB/s

➢ collectl -sx –oh

- This shows current Quadrics throughput
- MB-Out shows always „0“ because Lustre uses DMA for writes

» Lustre performance on MDS or OSS

➢ ssh xc1-ls4 collectl -sl -oh -c2 -i1

#READ OPS	READ KB	WRITE OPS	WRITE KB
0	0	164	**82473**
0	0	169	**84517**

Performance monitoring on fibre channel

- **FC switch performance**
  - **xc1san1:admin> PortPerfShow**

```
... 5    6    7    8    9    10   11   12   13   14   15   Total
... ------------------------------------------------------------
... 0    86m  85m  43m  43m  42m  42m  504  0    3.0k 1.8k 343m
... 22k  84m  85m  43m  40m  43m  41m  22k  0    136  136  339m
```

  - **Identify ports of OSS and EVA controllers**

Performance monitoring on EVA storage systems (1)

- **What is EVAPerf?**
  - **Allows monitoring of all EVA components**
    - **Storage arrays, virtual and physical disks, and FC ports**
  - **Automatically installed with command view EVA 4.x**
    - **Runs on the Storage Management Appliance**
  - **For initial documentation see command view EVA user guide**
    - **For detailed description of displayed data see white paper**
  - **Command evaperf for command line monitoring**
    - **Below C:\Program Files\Hewlett-Packard\EVA Performance Monitor**
  - **Windows Perfmon for graphical monitoring**
- **Save all current component statistics to a file**
  - **evaperf all -KB -fo E:\evaperf_all.log**
    - **MB/s values are based on 1 MB = 1,000,000 bytes**

Performance monitoring on EVA storage systems (2)

» Display current performance on storage arrays

- **evaperf as**

```
Req/s    MB/s
  991 121.56 5000-1FE1-5002-74D0
```

» Display physical disk activity

- **evaperf pda**

```
Enc. Bay__1 Bay__2 Bay__3 Bay__4 Bay__5 ... Node
   5  12.56  13.60  11.64  14.39  11.27 ... 5000-1FE1-5002-74D0
   4  10.59  10.86  11.90   9.94  12.98 ... 5000-1FE1-5002-74D0
```

» Display virtual disk statistics

- **evaperf vd**

```
... Write  Write   Write  Flush Mirror Prefetch ... Ctlr ...
... Req/s   MB/s Latency   MB/s   MB/s     MB/s ...
...   467  59.33    19.1  60.15  66.84     0.00 ... Y09P ...
...   502  60.68    17.5  59.23  67.90     0.00 ... Y07M ...
```

Second example: Identify hardware problems

- EVA controller had rebooted
  - WSEA reported this via email
- Performance monitoring actions
  - dd showed a small performance degradation
  - collectl showed that one OSS had only half throughput
  - PortPerfShow showed that rebooted controller was unused
- Further troubleshooting
  - lfs getstripe showed that only 7 of 8 OSTs were used
    - Also users complained that they could not read some files
  - Reboot of the corresponding OSS solved the problem
  - Underlying reason: EVA controller failover did not work
    - A new FC driver repaired this bug