TODO

- Caching?

Lustre

Paul Bienkowski
2bienkow@informatik.uni-hamburg.de

Proseminar "Ein-/Ausgabe - Stand der Wissenschaft"

2013-06-10

Outline

What is Lustre

parallel, scaling, for clusters, based within linux kernel...

The Project

Goals

2007: performance >features >stability
"itś a science project"
2010: stability >performance >features
used in high-performance production environments

History

?

Involved Companies

Sun, Oracle, Cray, ... (what do they do, why do they do it?)

Supercomputers

Titan & Co. use it!

Lustre Architecture

Network Structure

What data is stored where? (graph)

Metadata Server

Where are they? How to access?

Object Storage Server

Contain OSTs (Object storage targets)

Network Capabilities

How is data transfered?
Protocol stack (TCP, ...)
Different network types (ethernet, infiniband, ...)

Failover

Failover mechanism and typical setups (graphs)
Why are is a failover mechanism cool? Live-Upgrades!

Excursion: INodes

... because MDS do something similar (metadata records)
(graph)
compare this on next slide

Metadata

how metadata is stored in the MDS
what metadata is stored?
how metadata is fetched from the MDS

Striping

[repeat] what is striping (RAID 0)
why do they use it in lustre – speed advantage

Software Architecture

what software is running where?

Interversion Compatibility

Sun "guarantees" [citation needed] compatibbility between minor versions
$\rightarrow$ on-line upgrade-ability using failover systems

ldiskfs - Customized ext3

why we need a customized filesystem to work ON TOP of

Kernel patching (serverside)

just tell them the kernel needs to be patched (2.6.*) and what that means

Patchfree Client

How can clients access the data? (lustre-fs, liblustre, NFS)
(kernel-independent)
even NFS, that works everywhere!

Limitations

Server
very platform dependent
needs compatible kernel
Client
all linux kernel >2.6 supported
NFS for Windows, MacOS
even FUSE support on the way

Performance

Throughput Examples

Yes, the speeds add up!
There are systems with 5000 OSS.
Up to 160 OSS / file.
16 OST/OSS is quite normal.
1 TiB/OST

Scalability

Just multiply. Works as long as your network supports it (e.g. InfiniBand >WiFi ...)

???

Maybe I can find some more data samples – where is the interesting stuff?

Conclusion

...

References

...