# TODO

Caching?

### Lustre

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

Proseminar "Ein-/Ausgabe - Stand der Wissenschaft"

2013-06-10

- Introduction
- 2 The Project
  - Goals and Priorities
  - History
  - Involved Companies
- 3 Lustre Architecture
  - Network Architecture
  - Data Storage and Access
  - Software Architecture
- 4 Performance
  - Throughput Examples
  - Scalability
  - ???
- 5 Conclusion
- 6 References

#### What is Lustre

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

The Project 0000 Lustre Architecture 000000000000 Performance 000 Conclusion References

# The Project

- 1 Introduction
- 2 The Project
  - Goals and Priorities
  - History
  - Involved Companies
- 3 Lustre Architecture
  - Network Architecture
  - Data Storage and Access
  - Software Architecture
- 4 Performance
  - Throughput Examples
  - Scalability
  - ???
- 5 Conclusion
- 6 References

The Project ●000 Lustre Architecture 0000000000000 Performance 000 Conclusion References

Goals and Priorities

### Goals

2007: performance >features >stability

"its a science project"

2010: stability >performance >features

used in high-performance production environments

# History

History

-

## **Involved Companies**

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

# Supercomputers

Titan & Co. use it!

roduction The Project 0000 Lustre Architecture 000000000000 Performance 000 Conclusion References

#### **Lustre Architecture**

- 1 Introduction
- 2 The Project
  - Goals and Priorities
  - History
  - Involved Companies
- 3 Lustre Architecture
  - Network Architecture
  - Data Storage and Access
  - Software Architecture
- 4 Performance
  - Throughput Examples
  - Scalability
  - 7??
- 5 Conclusion
- 6 References

Network 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)

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

ntroduction The Project 0000 **Lustre Architecture 0000**00000000 Performance 000 Conclusion Reference

Network Architecture

#### Failover

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

Data Storage and Access

#### Excursion: INodes

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

ntroduction The Project 0000 **Lustre Architecture** 00000**00**000000 Performance 000 Conclusion Reference

Data Storage and Access

#### 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

### Software Architecture

what software is running where?

Software Architecture

## Interversion Compatibility

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

#### Idiskfs - Customized ext3

why we need a customized filesystem to work ON TOP of

Software Architecture

# Kernel patching (serverside)

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

Software Architecture

#### Patchfree Client

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

roduction The Project 0000 Lustre Architecture 00000000000 Performance 000 Conclusion Reference

Software Architecture

### 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

tion The Project 0000 Lustre Architecture 0000000000000 Performance 000 Conclusion References

### **Performance**

- 1 Introduction
- 2 The Project
  - Goals and Priorities
  - History
  - Involved Companies
- 3 Lustre Architecture
  - Network Architecture
  - Data Storage and Access
  - Software Architecture
- 4 Performance
  - Throughput Examples
  - Scalability
  - ???
- 5 Conclusion
- 6 References

The Project 0000 Lustre Architecture 000000000000 Performance ●00 Conclusion References

### 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

# 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

...