Lustre Centre of Excellence

LEADERSHIP
COMPUTING FACILITY
NATIONAL CENTER FOR COMPUTATIONAL SCIENCES

presented by

Sarp Oral & David Vasil

Lustre User Group Meeting
April 23rd, 2007

Oak Ridge National Laboratory
U.S. Department of Energy

ORNL LCF Lustre efforts

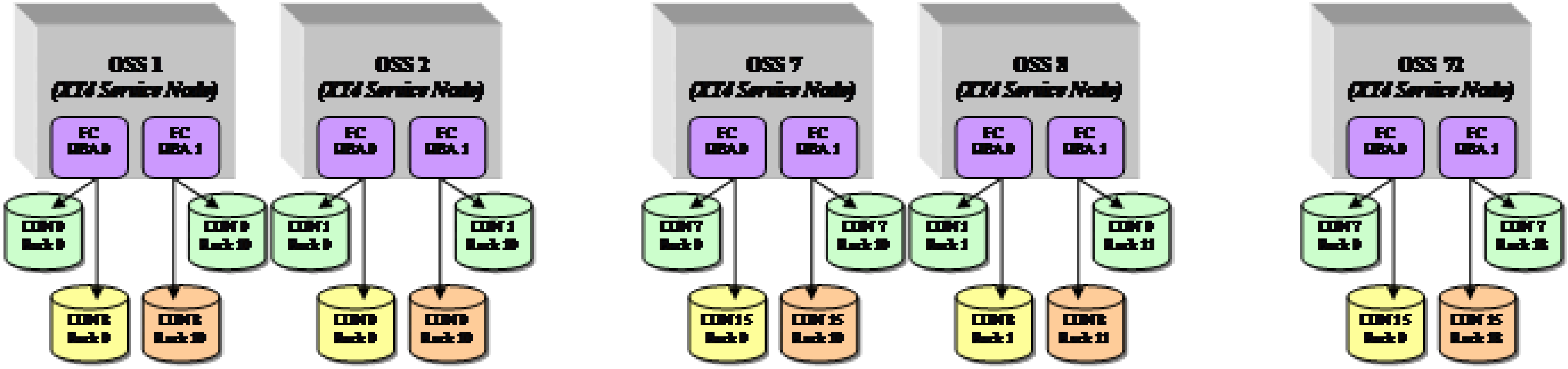
- Lustre tool development
  - Parallel Lustre copy tool
  - Portals I/O function shipping tool
  - Text-based top-like LMT tool
  - Web-based LMT tool
- Lustre and HPSS integration
- Server side client statistics
- File Joins
- High speed storage options for MDS
  - TCP and IB based DDN 9500 vs. ViON's TMS
- Lustre 1.6 on Cray UNICOS
- ORNL LCF Lustre FS in production
  - Cray XT4/XT3
  - End-to-end cluster
- Center-wide Lustre cluster

ORNL LCF Lustre efforts

- Cray XT4/XT3 (Jaguar) Lustre
  - 119 TF Cray XT4/XT3, 3-D tori
  - 3 Lustre file systems for production runs: 2×150 TB, 1×300 TB
    - Uses 3 MDS service nodes out of 8 available
    - 72 XT4 service nodes as OSSs
      - 4 OSTs/OSS
      - 2 OSTS for the 300 TB FS
      - 1 OST per each remaining 150 TB FS
    - 2 1-port 4 Gb FC HBAs per OSS
    - 45 GB/s block I/O for the 300 TB FS
  - DDN 9550s
    - 18 racks/couplets
    - Write-back cache is 1MB on each controller
    - 36 TB per couplet w/ Fibre Channel drives
    - Each LUN has a capacity of 2 TB and 4 KB block size

ORNL LCF Lustre efforts

Cray XT4/XT3 LUN configuration



LUN	Label	Owner	Tiers	Tier list
0	LUN0	1	2	1 2
1	LUN1	1	2	3 4
2	LUN2	1	2	5 6
3	LUN3	1	2	7 8
4	LUN4	2	2	9 10
5	LUN5	2	2	11 12
6	LUN6	2	2	13 14
7	LUN7	2	2	15 16

LUN	Label	Owner	Tiers	Tier list
8	LUN8	1	2	1 2
9	LUN9	1	2	3 4
10	LUN10	1	2	5 6
11	LUN11	1	2	7 8
12	LUN12	2	2	9 10
13	LUN13	2	2	11 12
14	LUN14	2	2	13 14
15	LUN15	2	2	15 16

ORNL LCF Lustre efforts

- End-to-end cluster (Ewok) Lustre
  - Lustre 1.4.9 for production runs
    - 6 OSS, 2 OST/OSS, OFED 1.1 IB, 81 clients
- Center-wide Lustre cluster (Spider)
  - To serve all NCCS resources
    - Jaguar, 1 PF Cray Baker, Viz, and end-to-end clusters by the end of 2008
    - Commissioned CFS to develop Lustre routers
      - Tests reveal satisfying results
      - ~450 MB/s per XT4 service node as a router over TCP/Cray Portals
      - Parallel copy from XT3 to XT4 Lustre FS over 3 routers @ ~1-2GByte/s
      - Encountered bug #11706: Instability on routers (details in bugzilla)
  - Phase 0
    - Proof of concept is in acceptance
    - 20 OSS, 80 OSTs, 4 OST/OSS, 10Ge & 4xSDR IB
    - 10 couplets of DDN 8500s, FC 2 Gb direct links w/ failover configured
    - Issues encountered so far
  - Phase 1: additional 20 GB/s by the end of 2007
    - Various OSS/MDS architectures are under investigation
  - Phase 2: total 200 GB/s by the end of 2008

Lustre Centre of Excellence at ORNL

Goals, metrics, and progress

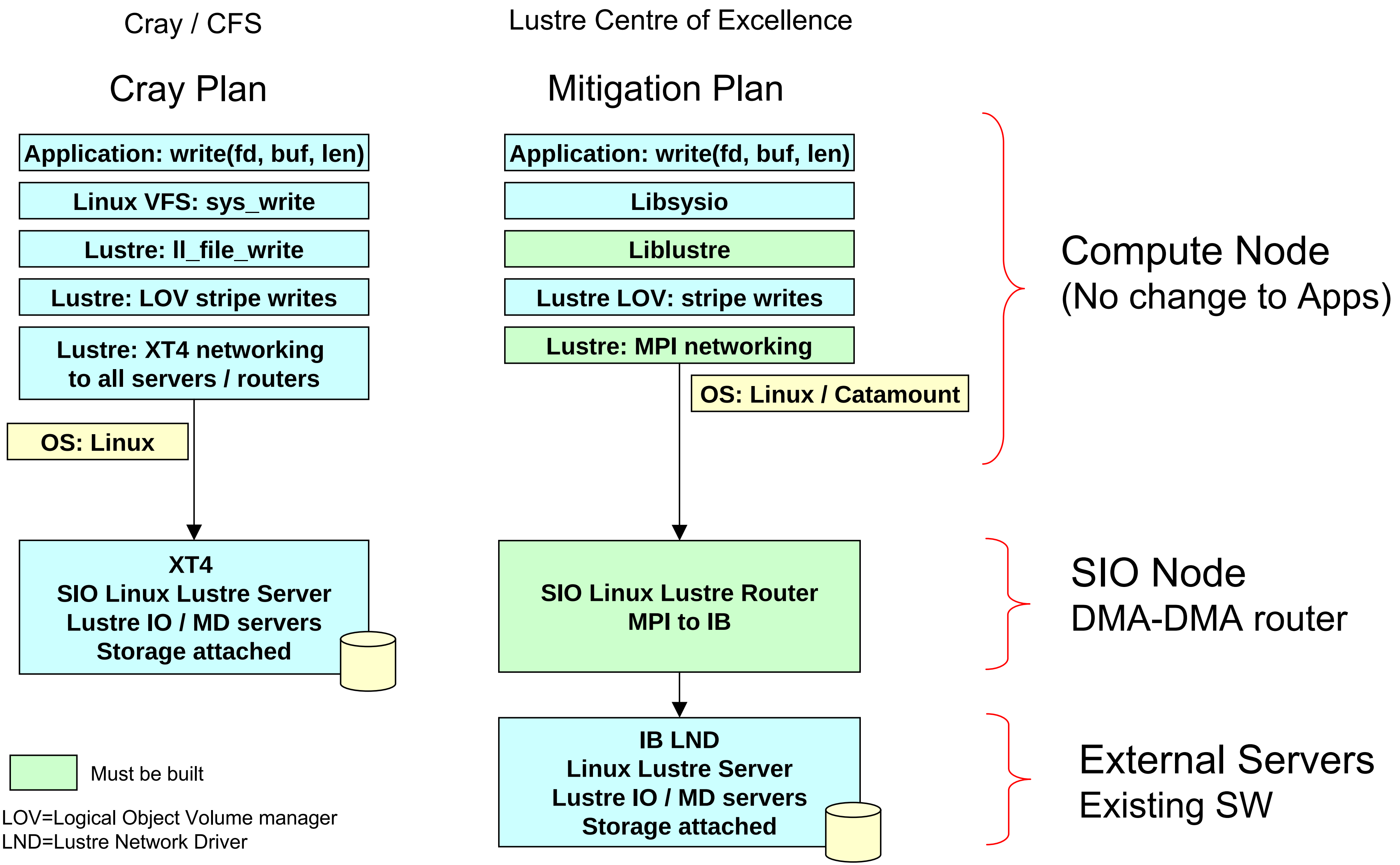
Lustre Centre of Excellence established in December 2006.

- Create an on-site presence at ORNL (1st floor back hall)
  - Two on-site staff, rotating additional
  - Oleg Drokin first hire at Lustre Centre of Excellence
- Develop a risk mitigation Lustre package for ORNL
  - A single lowest risk, scalable implementation to 1PF
  - In out-years explore possible 1 TB/s solutions
- Train ORNL staff in Lustre Source
  - Develop local expertise to reduce dependence on CFS and Cray
  - Peter Braam gave a 3 day tutorial on Lustre Internals in January
  - A sys admin training is being planned
- Assist Science teams in tuning their application I/O
  - Focus on 2-3 key apps initially and document results (Second Centre hire will focus on this goal)
  - On-site Lustre workshops for application teams

Choosing Lowest Risk Mitigation Strategy
Risk Reduction points

- Success not dependent on Cray software efforts
  - Independent of Cray network API's
  - Independent of Cray SW delays in OS or FS
  - Much SW can be developed outside XT4
- Works with Linux & Catamount
  - Cray and other solutions require Linux LWK
- Uses proven external servers
  - Known performance and leverage existing SW
- Timely
  - Available for 250 TF system
  - Larger scale, higher performance solution in time for 1 PF
- Less complexity
  - Simple, well defined plan
  - Many pieces exist today
- Benefits over other solutions
  - HA
  - Parallel I/O

Comparison of two Lustre Approaches on Cray



Cray / CFS

Cray Plan

Application: write(fd, buf, len)

Linux VFS: sys_write

Lustre: ll_file_write

Lustre: LOV stripe writes

Lustre: XT4 networking to all servers / routers

OS: Linux

XT4 SIO Linux Lustre Server Lustre IO / MD servers Storage attached

Lustre Centre of Excellence

Mitigation Plan

Application: write(fd, buf, len)

Libsysio

Liblustre

Lustre LOV: stripe writes

Lustre: MPI networking

OS: Linux / Catamount

SIO Linux Lustre Router MPI to IB

IB LND Linux Lustre Server Lustre IO / MD servers Storage attached

Compute Node
(No change to Apps)

SIO Node
DMA-DMA router

External Servers
Existing SW

🟩 Must be built

LOV=Logical Object Volume manager
LND=Lustre Network Driver