Lustre at DKRZ

Julian M. Kunkel

kunkel@dkrz.de

German Climate Computing Center (DKRZ)

21-06-2016

Lustre at DKRZ

- The Mistral supercomputer was shipped with Lustre
  - 3 PFLOP/s system
  - 52 PB Lustre storage
  - Roughly 6 M EURO
- System was procured in two phases
  - 2015: Phase 1 with 30 PB storage
  - 2016 (in production soon): Phase 2 with 22 PB storage
- Other systems/services at DKRZ use Mistral's Lustre storage
- Interesting aspects
  - RobinHood for QoS and policy management
  - Lustre 2.5 Seagate edition
- University of Hamburg is IPCC for Lustre
  - Researching file-system compression

I/O Architecture (Phase 1)

- 31 ClusterStor 9000 Scalable Storage Units (SSUs)
  - SSU: Active/Active failover server pair
- Single Object Storage Server (OSS)
  - 1 FDR uplink
  - GridRaid: (Object Storage Target (OST))
    - 41 HDDs, de-clustered RAID6 with 8+2(+2 spare blocks)
    - 1 SSD for the Log/Journal
  - 6 TByte disks
- 31 Extension units (JBODs)
  - Do not provide network connections
  - Storage by an extension is managed by the connected SSU
- Multiple metadata servers
  - Root MDS + 4 DNE MDS
  - Active/Active failover (DNEs, Root MDS with Mgmt)
  - DNE phase 1: Assign responsible MDS per directory

I/O Architecture (Phase 2)

- Additional file system (Now two file systems in total)
  - Mounted on all compute nodes
  - Characteristics: 11 k disks, 52 PB storage
- 34 ClusterStor L300 Scalable Storage Units (SSUs)
- 34 Extension units (JBODs)
- Storage hardware
  - Seagate Enterprise Capacity V5 (8 TB) disks
- Multiple metadata servers
  - Root MDS + 7 DNE MDS

Parallel File System

Lustre 2.5 (Seagate edition, some backports from 2.7+)

Filesystem

- We have two file systems: /mnt/lustre0[1,2]
- Symlinks: /work, /scratch, /home, ...
- For `mv`, each metadata server behaves like a file system

Assignment of MDTs to Directories

- In the current version, directories must be assigned to MDTs
  - /home/* on MDT0
  - /work/[projects] are distributed across MDT1-4
  - /scratch/[a,b,g,k,m,u] are distributed across MDT1-4
- Data transfer between MDTs is currently slow (mv becomes cp)
- We will transfer some projects to the phase 2 file system
  - New projects will be created on the phase 2 system

Peak Performance

Phase 1 + 2

- 65 SSUs · (2 OSS/SSU + 2 JBODs/SSU)
- 1 Infiniband FDR-14: 6 GiB/s ⇒ 780 GiB/s
- 1 ClusterStor9000 (CPU + 6 GBit SAS): 5.4 GiB/s
- L300 yield IB speed, still we consider 5.4 GiB/s ⇒ aggregated performance **704 GiB/s**
- Phase 2: `obd-filter` survey demonstrates that 480 GB/s and 580 GB/s can be delivered

Performance Results from Acceptance Tests

- Throughput in GB/s (% to peak) measured with IOR
  - Buffer size 2000000 (unaligned) on 42 OSS (Phase 1) and 64 (P 2)
  - In the phase 2 testing, the RAID of at least one OSS is rebuilding

	Phase 1		Phase 2	
Type	Read	Write	Read	Write
POSIX, independent[1]	160 (70%)	157 (69%)	215 (62%)	290 (84%)
MPI-IO, shared	52 (23%)	41 (18%)	65 (19%)	122 (35%)
PNetCDF, shared	81 (36%)	38 (17%)	63 (18%)	66 (19%)
HDF5,shared	23 (10%)	24 (11%)	62 (18%)	68 (20%)
POSIX, single stream	1.1 (5%)	1.05 (5%)	0.98 (5%)	1.08 (5%)

- Metadata measured with Parabench
  - Phase 1: 80 kOPs/s
    - 25 kOP/s for root MDS; 15 kOP/s for DNEs
  - Phase 2: 210 kOPs/s
    - 25 kOP/s for root MDS; 30-35 kOP/s for DNEs

[1] 1 stripe per file

Issues With Lustre

- Management tools have potential to be improved
  - RobinHood
  - Performance analysis
- Compatibility of Lustre clients
- DNE: data movement instead of metadata
- Defining the number of stripes (instead of automatically)
- But the situation improved over the last years