A triad-based architecture for a multipurpose Lustre filesystem at /rdlab
The research and development Lab (context)

- Founded in 2010 at the Computer Science department
- IT support for research groups only
- National and European Projects (FP7, H2020…)
- Technology transfer

The research and development Lab (Infrastructure)

- 160 researchers, 18 research groups
- HPC and Cloud services for research projects
- 400TBytes Lustre (2.12.5 + ZFS) storage
SQUARING THE CIRCLE

• **Why not Lustre?**

  - Well-known project
  - Using Lustre since 2010 (HPC service)
  - Most of our data was already in Lustre
  - Lustre provides a flexible architecture to play

• **OK, but...**

  - Misconceptions (expensive, difficult to understand...)
  - Compatibility issues (vendors and technologies)
  - Who is using Lustre as a general purpose filesystem? (Early adopter panic)
  - Undocumented experiences and good practices
• **Classical Lustre setups**

- **Type A**: Several n-disk volumes OST governed by a single dedicated OSS

- **Type B (HA)**: A multiple-disk OST pair attached to a couple of OSS
THE SCIENTIFIC METHOD

• Cooking the idea

- Identify the main ingredients, goal(s) and constraints
- Set metrics and baselines
- Play: Combine, test and “taste”
THE SCIENTIFIC METHOD II

• Milestones

2015
Just an idea

2018
Lustre ZFS
Deployment

2021
LUG2021

https://rdlab.cs.upc.edu/
• **Ingredients for a triad based recipe**

- 3 physical dedicated disk servers (different model/vendors?)
- Same disk technology layout
- Dedicated high-speed low-latency network (IB + iSER)

---

High speed Local Area Network

- Point to Point dedicated low latency network (IB) for iSER device export

ZFS mirror

ZFS mirror

ZFS mirror

ZFS mirror

ZFS mirror

ZFS mirror

---

LUG2021

https://rdlab.cs.upc.edu/
• **Spicing the triad**

- Group alike ZFS mirrors into ZFS Stripes
- Group ZFS stripes into a 3 OST setup
- “Serve” every OST with HA and a Zpool cache disk

---

**High speed Local Area Network**

- OSS 1
- OSS 2
- OSS 3

- **OST 0**
  - ZFS mirror (1+1'+1'") A
  - ZFS mirror (2+2'+2") B
  - ZFS mirror (3+3'+3") C
  - ZFS mirror (4+4'+4") D
  - ZFS mirror (5+5'+5") E
  - ZFS mirror (6+6'+6") F
  - ZFS mirror (7+7'+7") G
  - ZFS mirror (8+8'+8") H
  - ZFS mirror (9+9'+9") I
  - ZFS mirror (10+10'+10") J
  - ZFS mirror (11+11'+11") K
  - ZFS mirror (12+12'+12") L

- **OST 1**

- **OST 2**

- **OST 3**

---

Zpool cache (active OST only)

Point to Point dedicated low latency network (IB) for iSER device export

Point to Point dedicated low latency network (IB) for iSER device export

---

ZFS Stripes (A,B,C,D)

ZFS Stripes (E,F,G,H)

ZFS Stripes (I,J,K,L)
WHY A TRIAD-BASED ARCHITECTURE?

• Features and flavors

- Customization

- Performance (dedicated network + I/O split)

- Data cost vs redundancy

- Reliability (data CRC + Quorum)

- Isolation (maintenance, disaster)

- Rebuild impact

- Big File support (in Lustre, size matters)

- ZFS benefits (compression, deduplication, cache…)

LUG2021
https://rdlab.cs.upc.edu/
50 anys sense limits
REFERENCES


[10] https://www.upc.edu


[12] https://rdlab.cs.upc.edu

LUG2021