LUG 2021

www.opensfs.org

UF UNIVERSITY FLORIDA

A triad-based architecture for a multipurpose Lustre filesystem at /rdlab

May 2021

Gabriel Verdejo Alvarez

https://rdlab.cs.upc.edu
rdlab@cs.upc.edu

50 ANYS
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
sense Limits

/rd|lab

Engaged with your research

/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)

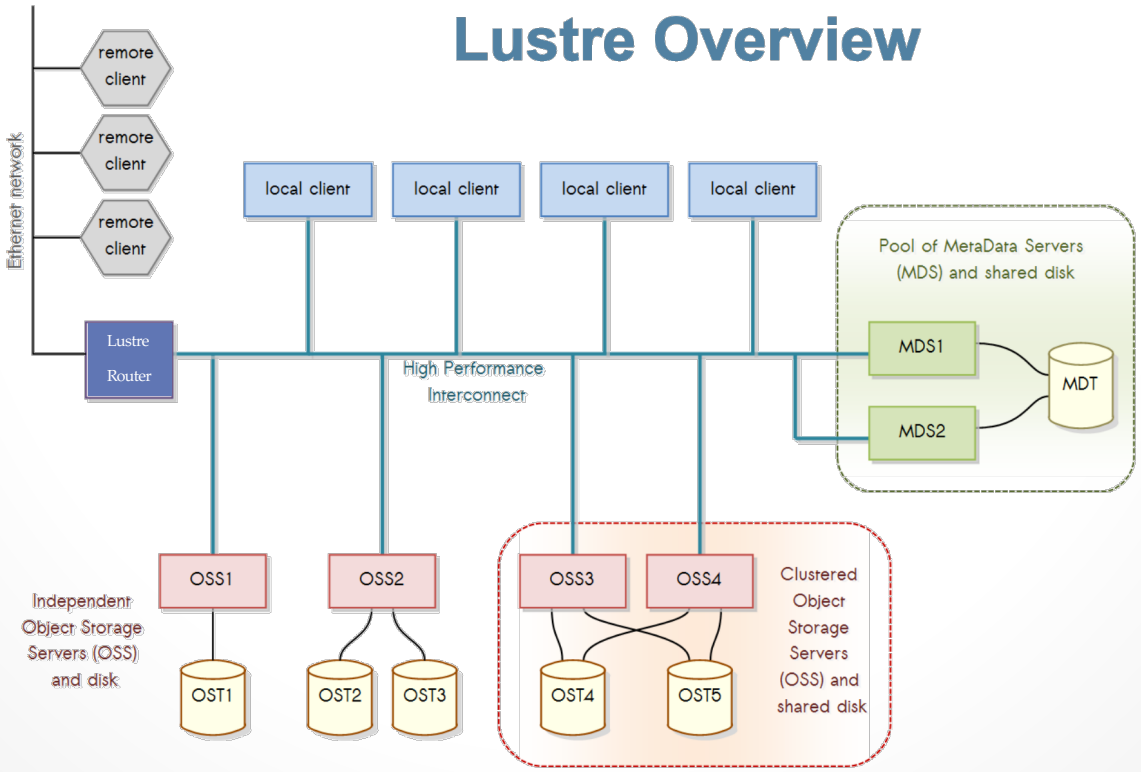
  - Undocumented experiences and good practices

STARTING POINT

- ***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



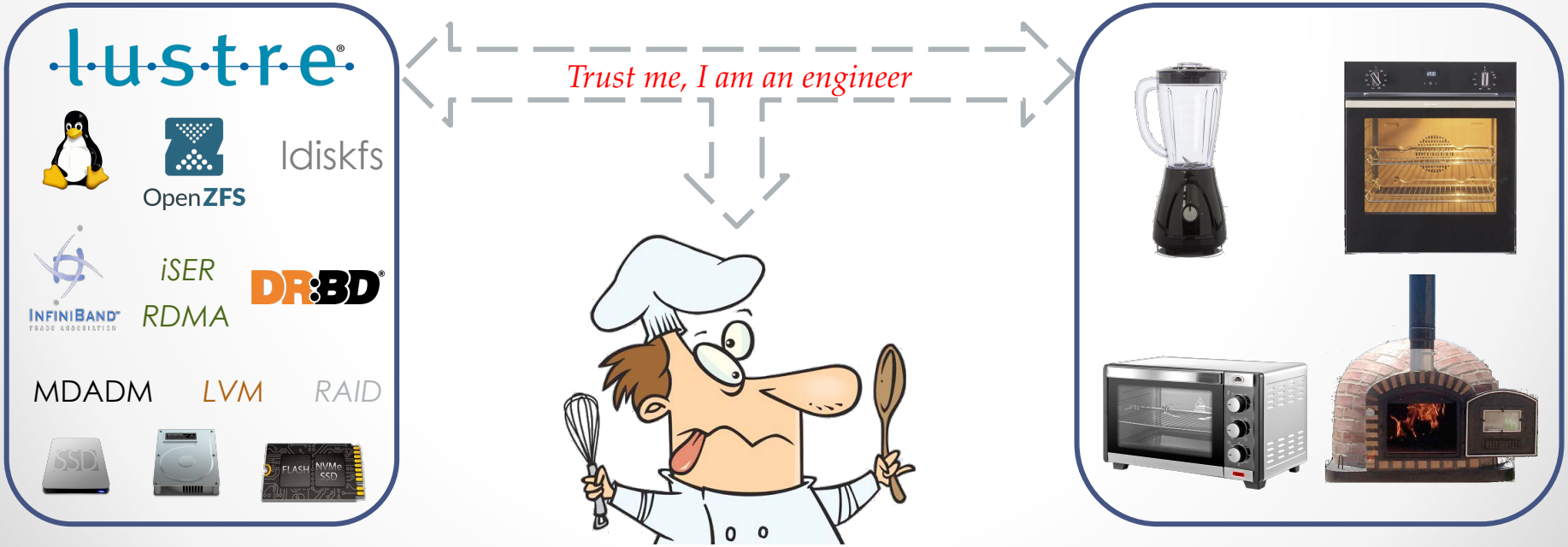
THE SCIENTIFIC METHOD

- ***Cooking the idea***

  - Identify the main ingredients, goal(s) and constraints

  - Set metrics and baselines

  - Play: Combine, test and "taste"



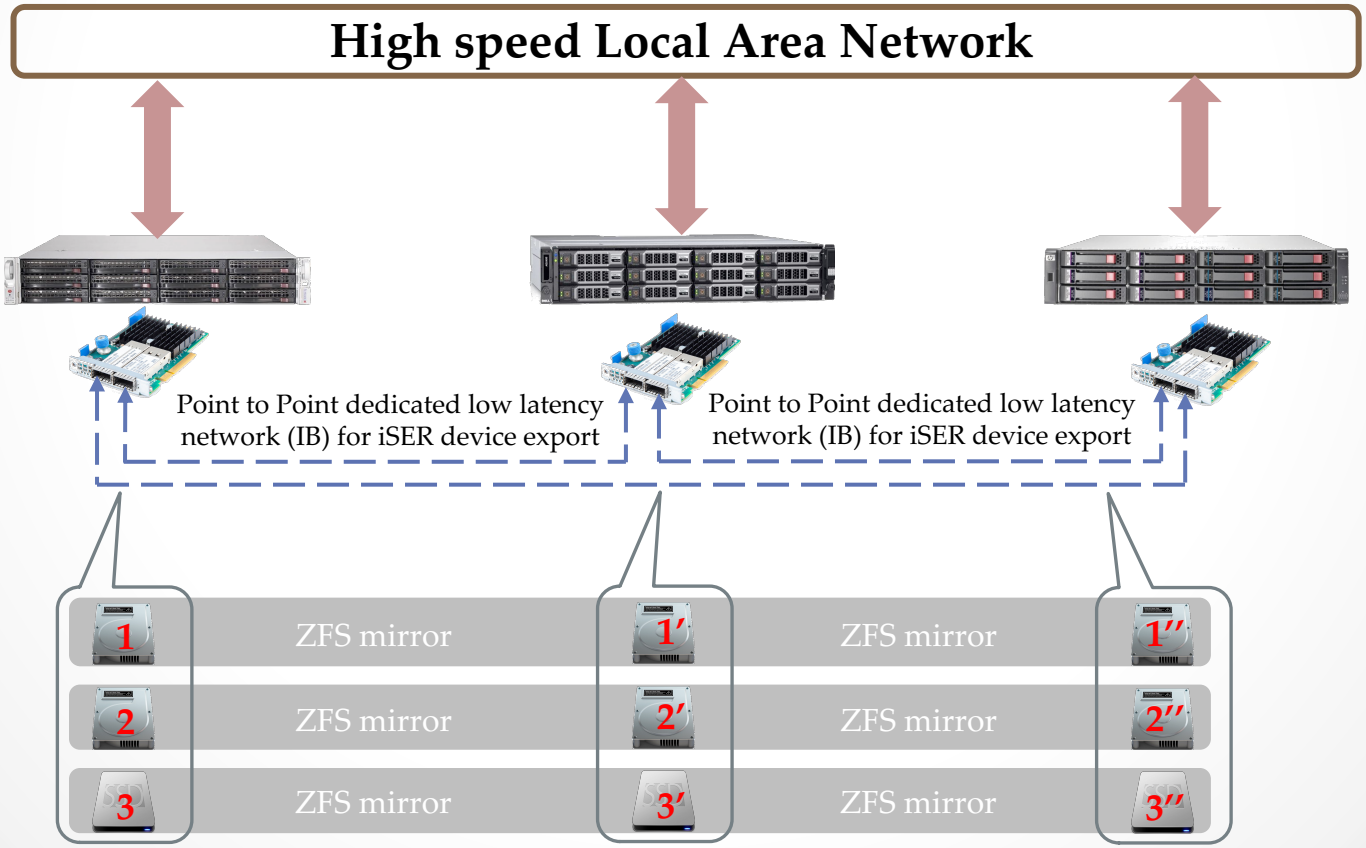
THE SCIENTIFIC METHOD II

- ***Milestones***

OFF THE BEATEN TRACK

- ***Ingredients for a triad based recipe***

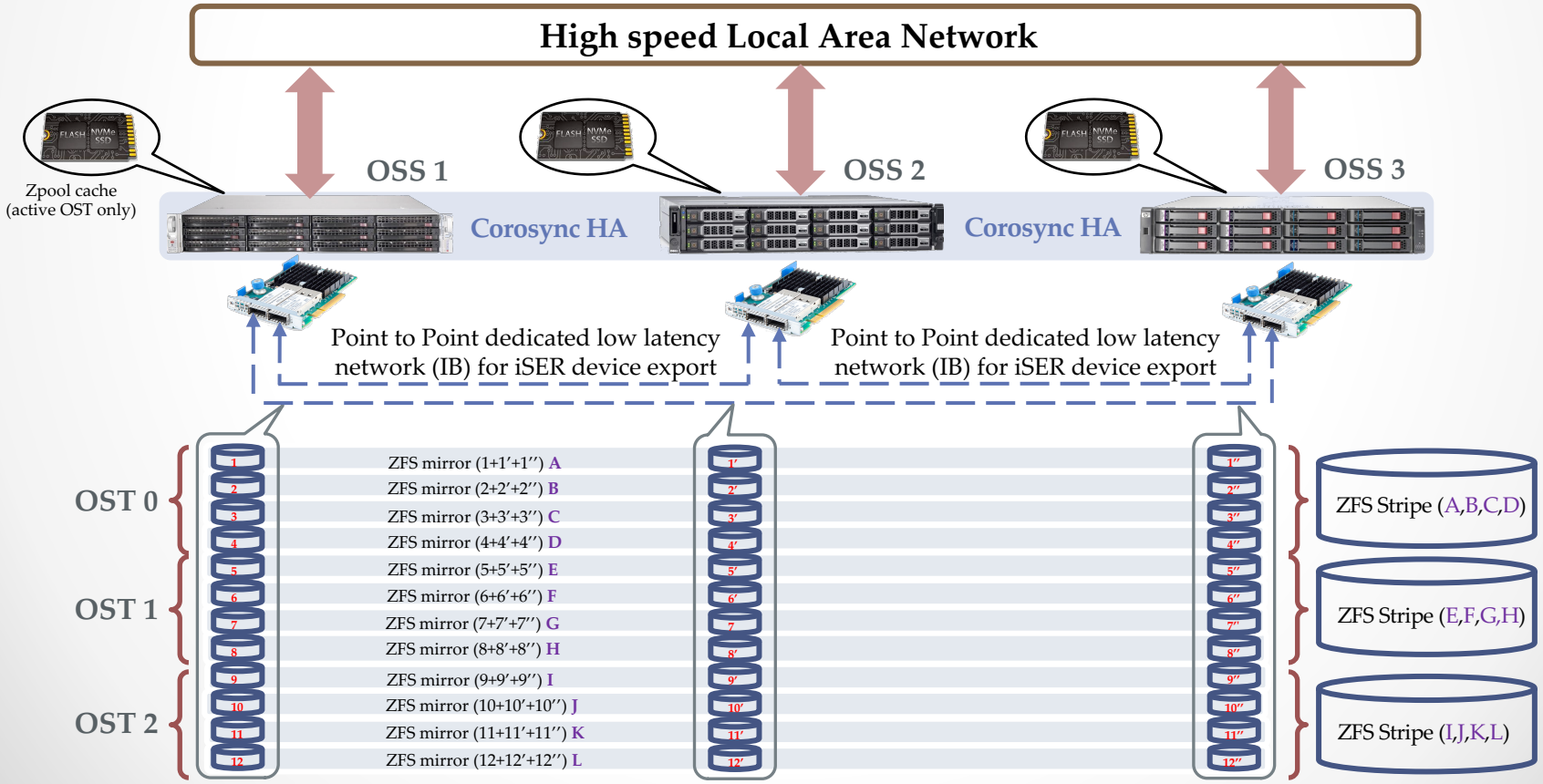
  - 3 physical dedicated disk servers (different model/vendors?)
  - Same disk technology layout
  - Dedicated high-speed low-latency network (IB + iSER)



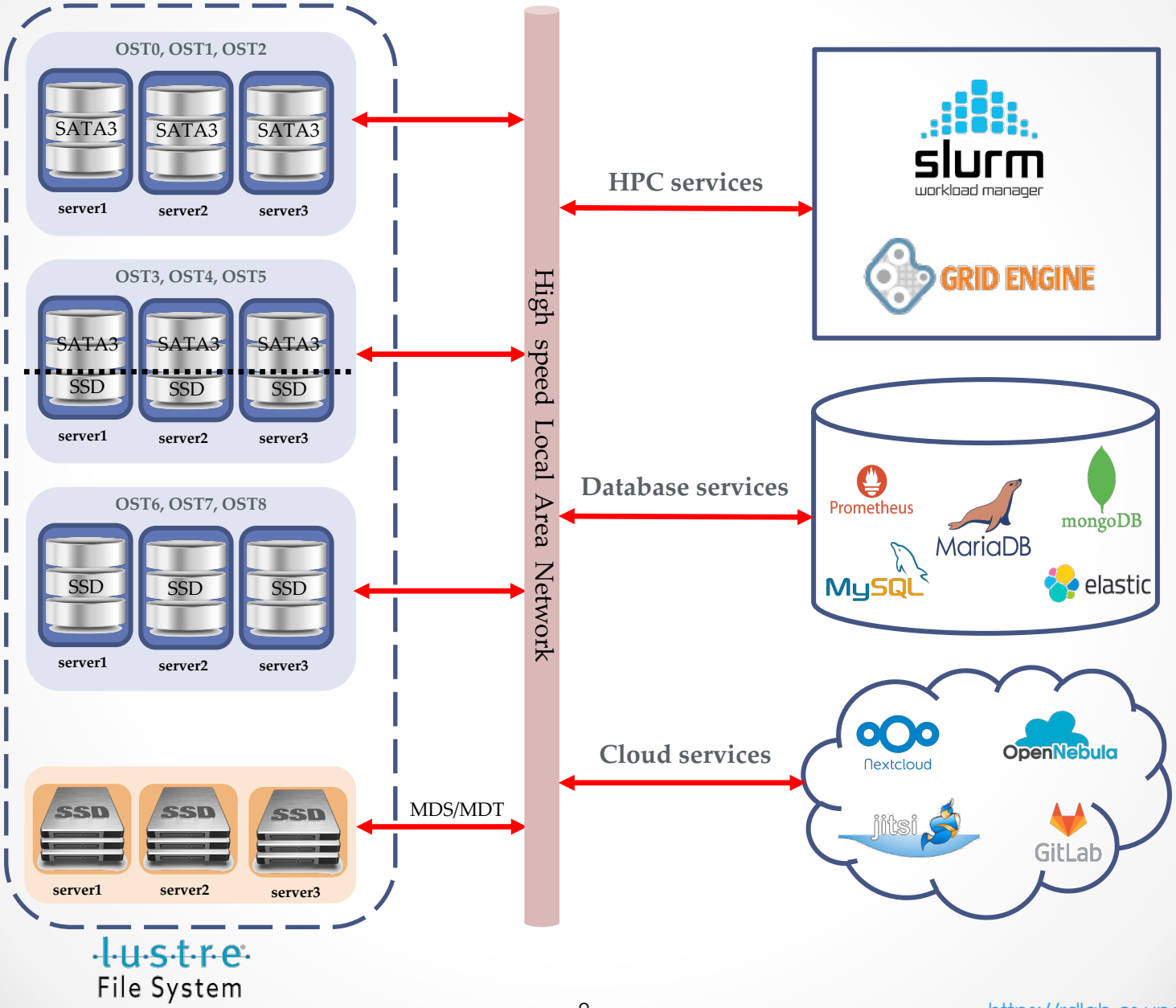
OFF THE BEATEN TRACK II

- ***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



THE FINAL FORM



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…)


50 ANYS
sense límits

REFERENCES

[1] https://www.jiscmail.ac.uk/cgi-bin/webadmin?A3=ind1012&N=DELL04.Handover_training.pdf

[2] https://wiki.lustre.org/Lustre_Object_Storage_Service_(OSS)

[3] https://en.wikipedia.org/wiki/ISCSI_Extensions_for_RDMA

[4] https://wiki.lustre.org/ZFS_OSD

[5] https://openzfs.org/wiki/

[6] https://www.digistor.com.au/the-latest/Whether-RAID-5-is-still-safe-in-2019/

[7] https://blog.westerndigital.com/hyperscale-why-raid-systems-are-dangerous/

[8] https://www.laverdad.es/gastronomia/clave-innovacion-estar-20201107011018-ntvo.html

[9] https://www.lavanguardia.com/comer/al-dia/20201005/33652/mejor-croissant-espana-encontraras-pasteleria-barcelona-brunells.html

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

[11] https://www.cs.upc.edu

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