Cloud Computing Applications and Services
(Aplicações e Serviços de Computação em Nuvem)

Distributed Systems Research @HASLab

INESC TEC and U. Minho

2022/2023

Topics

- Efficient data management
  - Key-value / File-System / Block-based storage
    https://dsr-haslab.github.io
  - SQL / NoSQL databases
    https://dbr-haslab.github.io

- Large scale data storage and processing
  - Scalability and Dependability
  - Distributed protocols

- Distributed systems monitoring and benchmarking

- Secure data storage and processing
  - Collaboration with cryptography group @ HASLab

Targets

- Cloud computing
- High-Performance computing
- Big Data applications
  - Data analytics
  - Machine / Deep learning
- Blockchain

Team

- Researchers and Professors (14)
- Phd students (>10)
- Msc students (>20)
- Bsc students (>8)

Distributed Storage Research

Vision and Goal

- Storage systems are critical for different research and industrial topics
  - Cloud Computing, HPC, IoT, Data Analytics, AI

- Our group aims at building new storage solutions that
  - Are efficient, scalable, resilient and secure
  - Can handle the exponential growth of digital information
  - Are adaptable to heterogeneous applications and infrastructures

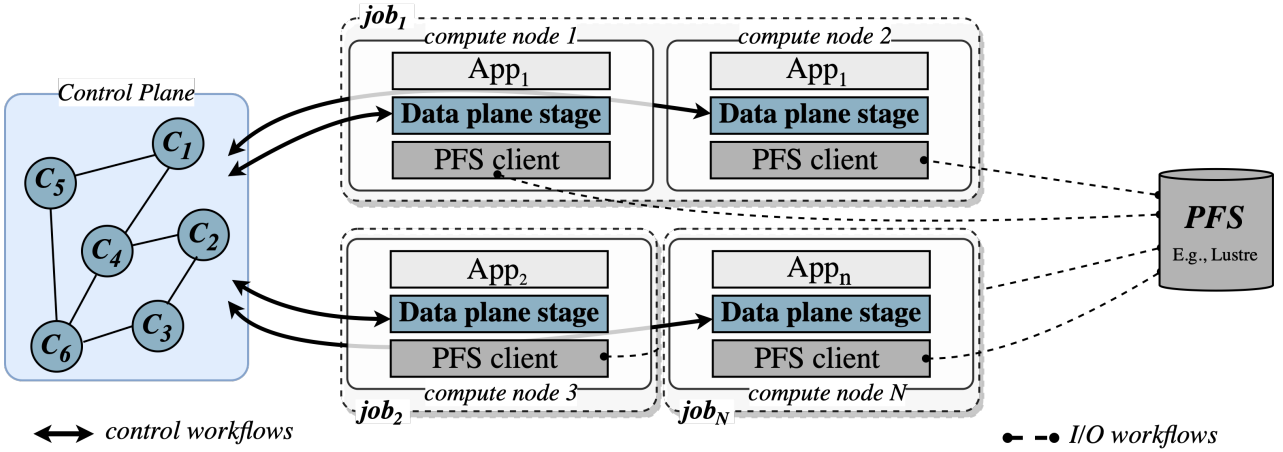
- Research divided into three main topics
  - Software-Defined Storage
  - Storage Optimizations
  - Storage Benchmarking and Monitoring

Software-Defined Storage

- Adaptable and programmable distributed storage solutions for HPC and Cloud

- Data plane
  - User-space storage optimizations (e.g., caching, rate-limiting)
  - Programmable and generally applicable frameworks

- Control Plane
  - QoS policies (e.g., I/O fairness, priority)
  - Control algorithms
  - Scalability
  - Dependability



- *PAIO: General, Portable I/O Optimizations With Minor Application Modifications*. Macedo et al. FAST'22
- *A Survey and Classification of Software-Defined Storage Systems*. Macedo et al. ACM CSUR 2020

BIG HPC

- Improve storage performance for HPC services
  - Alleviate I/O pressure at the shared parallel file system
  - Improve Quality of Service

- Improve the monitoring of HPC infrastructures
  - Large-scale setup
  - Unified framework and metrics

- Improve the deployment of Big Data applications and the management of HPC computational resources
  - Containerization technologies (e.g., singularity, charlie cloud)

https://bighpc.wavecom.pt

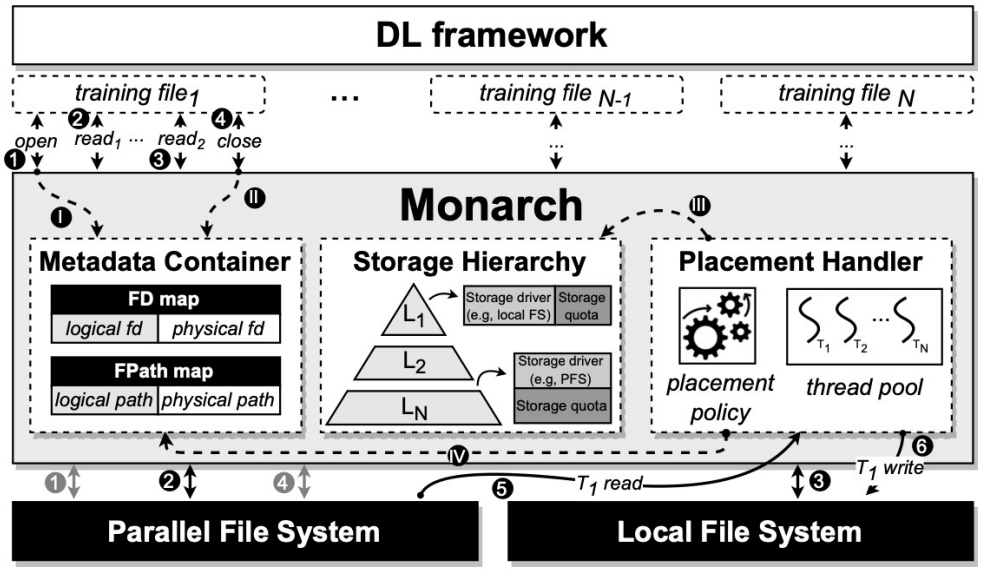
BIG HPC

Partners:

Funding:

Storage Optimizations

- Storage for AI
  - Data tiering and pre-fetching
- Data reduction techniques
  - Deduplication
- Dependability and Security
  - Very large-scale storage
  - Secure storage



- *Accelerating Deep Learning Training Through Transparent Storage Tiering.* Dantas et al. CCGrid'22
- *S2Dedup: SGX-enabled Secure Deduplication.* Miranda et al. SYSTOR'21

PASTOR

- Improve storage performance for AI frameworks
  - E.g., TensorFlow, PyTorch, …
- Novel Software-Defined Storage solution providing
  - Reusable optimizations for AI applications (e.g., caching, tiering, QoS)
  - Holistic visibility and automatic configuration of storage resources
  - Easy integration with existing HPC software and hardware

https://pastor-project.github.io

Partners

INESCTEC

TACC

MACC Minho Advanced Computing Center

HOOD COLLEGE

Funding

FCT Fundação para a Ciência e a Tecnologia

UTAustin Portugal

GOVERNO DE PORTUGAL

- CENTRA - Collaborations to Enable Transnational Cyberinfrastructure Applications

- Partners from Europe, US and Asia

- Efficient and Secure Data Management for HPC and Cloud Computing
  - Optimize the performance and dependability of data-centric applications (e.g., databases, data analytics, ML)
  - Privacy-by-design approach for storing and processing data at third-party infrastructures

https://www.globalcentra.org/projects/#prv

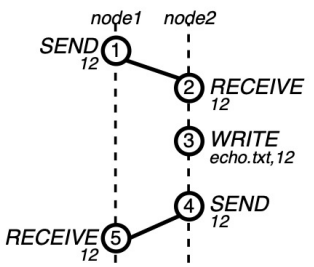
Storage Benchmarking and Monitoring

- Benchmarking tools for storage systems
  - File system and block device
  - Realistic content generation (e.g., duplication / compression ratio)
  - Fault-injection

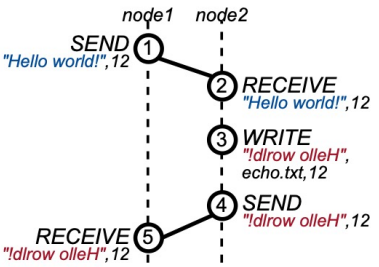
LazyFS

INESC TEC and **Jepsen, LLC** collaboration

- I/O Monitoring
  - Black-box tracing of distributed systems
  - Content-aware tracing and analysis
  - I/O diagnosis and vizualization



(a) Context-based tracing



(b) Content-based tracing

- *LazyFS repository: https://github.com/dsrhaslab/lazyfs*
- *CaT: Content-aware Tracing and Analysis for Distributed Systems. Esteves et al. Middleware'21*

Meet the Team

Find more about us at https://dsr-haslab.github.io