www.bsc.es


Barcelona Supercomputing Center
Centro Nacional de Supercomputación


# echofs: an NVRAM-based collaborative burst buffer with a POSIX interface

## 6th JLESC Workshop

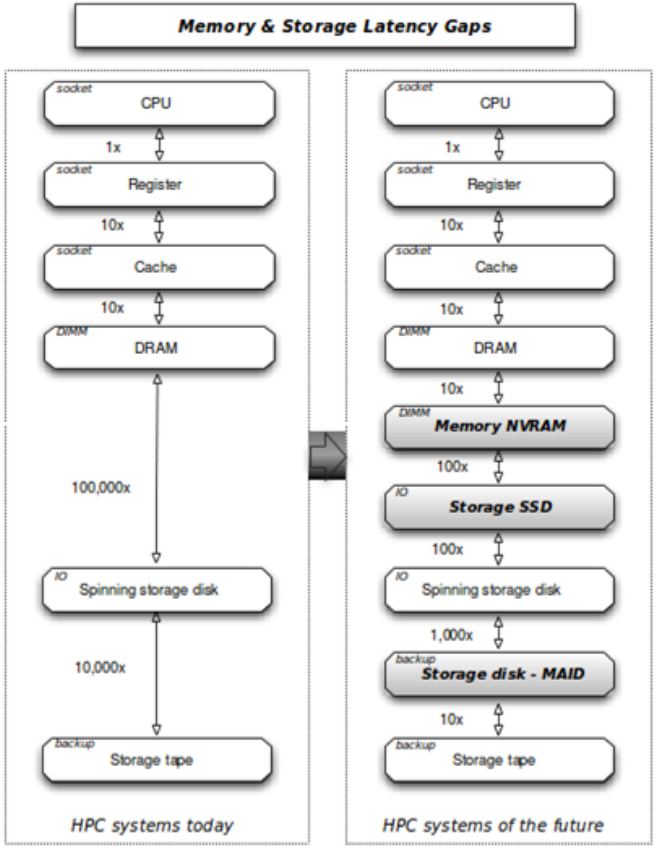
Ramon Nou

**The NEXTGenIO project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671951**

Kobe, December 2016

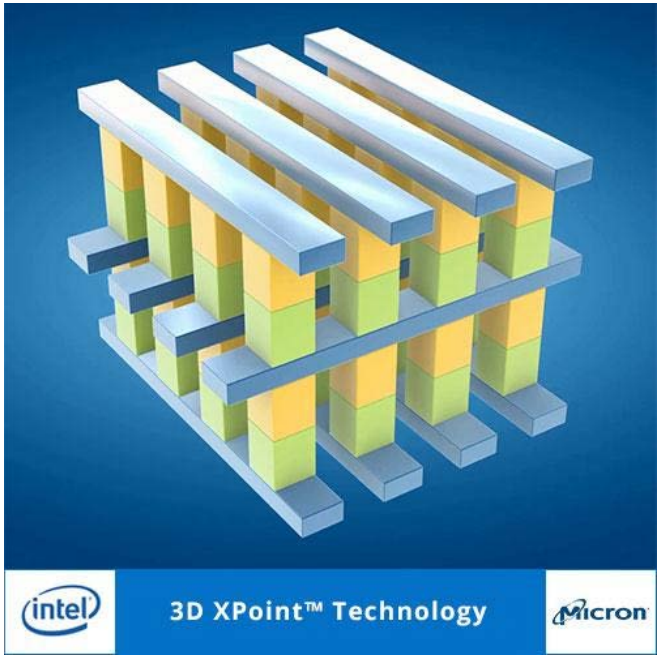
# I/O: a Fundamental Exascale Challenge!

- Next generation NVRAM technologies will change memory and storage hierarchies significantly
  - HPC systems and Data Intensive systems will merge



# The NEXTGenIO EU Project

- New server architecture based on Intel's 3D XPoint™ NVRAM technology
- NVRAM as a fundamental component of compute nodes
- Low-latency/high-throughput interconnection network



# echofs: Features and Goals

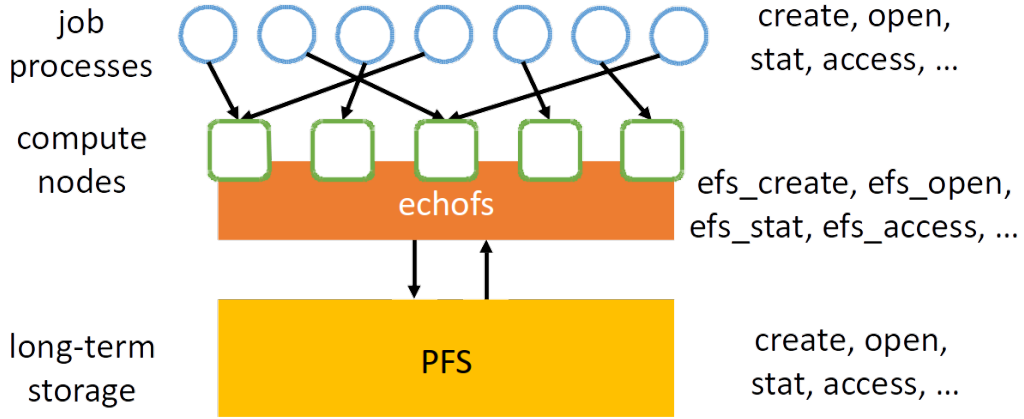
- New virtual layer on top of the Parallel Filesystem
  - Hide the complexity added to the I/O stack
  - echofs manages where data resides (e.g. NVRAM, Lustre, …)
- Unify NVRAM region's available to an HPC job to create a collaborative burst buffer
  - Allow HPC jobs to perform collaborative NVRAM I/O
- Support a POSIX-like interface
  - Allow legacy HPC applications to benefit from new architecture with minimal modifications
- Allow applications to offer hints through a service API to influence I/O

# echofs: Workflow

- User provides job I/O requirements through SLURM
  - Nodes required, files accessed, type of access (e.g. input/output/inout), expected lifetime of data (e.g. temporary/persistent), ...
- Compute nodes are allocated and echofs is mounted on them
  - The PFS's namespace is replicated → relative paths maintained
- Files specified by the job are copied from the PFS into NVRAM
- Job starts and I/O is redirected to NVRAM
- After job completes, future of data is decided by echofs
  - Persistent data *eventually* transferred back to PFS
  - Temporary data is deleted from NVRAM

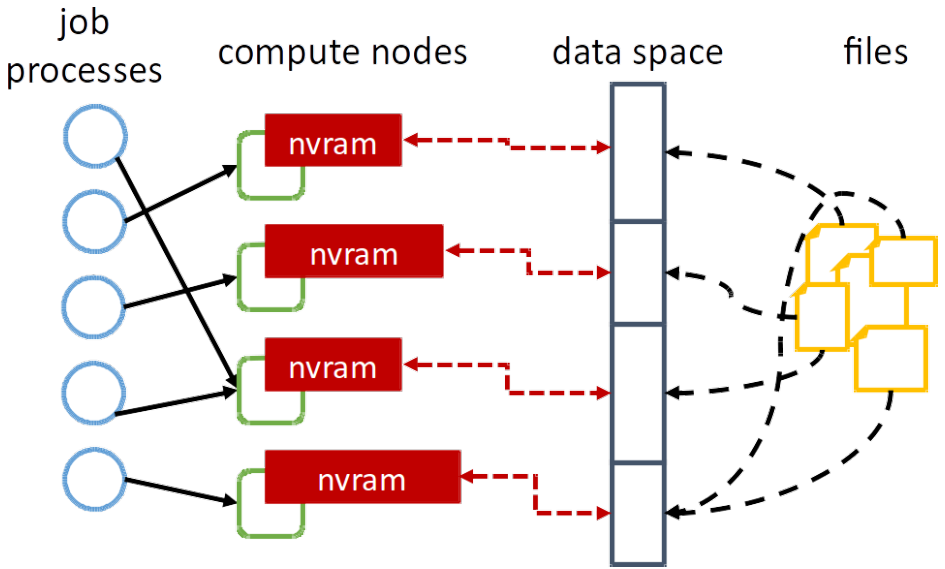
# echofs: Architecture Overview

- No centralized Metadata Servers
  - Metadata operations for persistent files forwarded to PFS
    - *Forwarding rate adjusted to avoid overwhelming the PFS (e.g. a job creating millions of files)*
  - Metadata operations for temporary files “absorbed” by echofs
  - PFS manages metadata concurrency/consistency



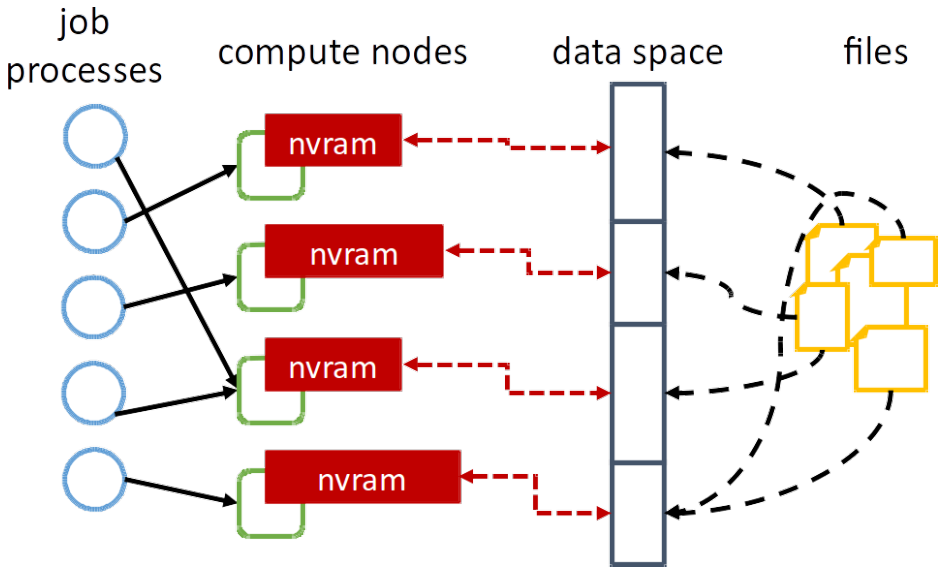
# echofs: Architecture Overview

- Distributed Data Servers
  - Data space is partitioned among compute nodes
  - Each node acts as a Data Server for its partition
  - Each node acts as a Data Client of other partitions
  - Low-latency network required for data transfers



# echofs: Architecture Overview

- Data Distribution
  - Static/pseudo-randomized data striping to collaborating nodes
- No replication ⇒ no distributed coherence mechanisms
  - Nodes act as a lock managers for their partitions to enforce POSIX semantics



# Future

- Looking for collaborations:
  - Use cases that may benefit from it
  - Hardware or environments to test it (even without NVRAM could be interesting)
- Include more hints, integration with SLURM

www.bsc.es


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Thank you!

For further information please contact

alberto.miranda@bsc.es / ramon.nou@bsc.es / toni.cortes@bsc.es