uc3m | Universidad Carlos III de Madrid

ARCOS

Network accelerated in-memory ad-hoc file system for data-centric and high-performance applications

Javier Garcia-Blas, Genaro Sanchez-Gallegos, Cosmin Petre and Jesus Carretero

University Carlos III of Madrid

fjblas@inf.uc3m.es

BDCSA2023: Big Data Convergence: from Sensors to Applications
PDP 2023
Naples, Italy

Motivation (I)

- I/O-intensive HPC-based applications have been primarily based on distributed object-based file systems
  - **Separate data** from **metadata** management
  - Enable each client to **communicate in parallel** with multiple storage servers.
- Exascale I/O raises the throughput and storage capacity requirements by several orders of magnitude.
- Current challenges
  - Systems already developed for data analytics are not directly applicable to HPC due to the **fine-granularity** involved in scientific applications.
  - Semantic gap between the application requests and the way they are managed by the storage back-end at the block level.

Motivation (and II)

- **Alluxio** conforms a storage solution located between computation frameworks and persistent data stores that aims to reduce the complexity of storage APIs while taking advantage of memory speed
- **Hermes** focuses on the implementation of a MRAM based storage system improving file system performance through the effective use of MRAM devices.
- **WekaIO** provides a high-performance storage architecture
- However, they lack of:
  - Data locality mechanisms.
  - Ad-hoc storage characteristics.

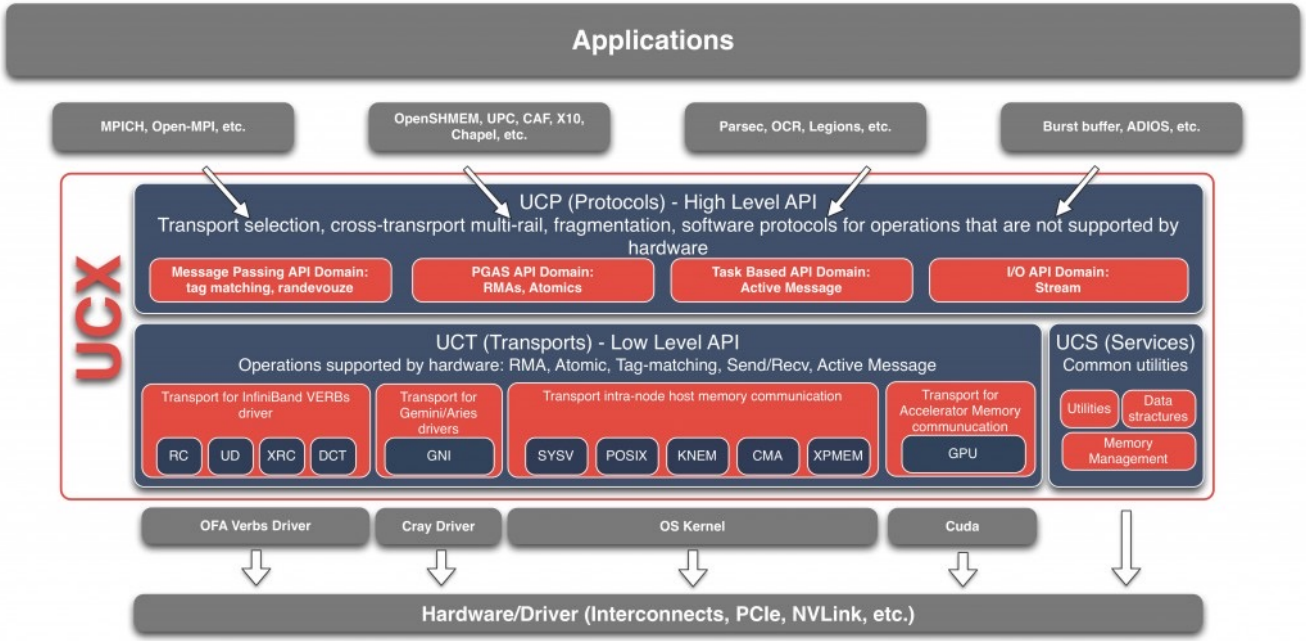
Hercules

- Ad-hoc/in-memory storage solution.
- Distributed key-value store.
- Provides a flexible API.
- Makes use of main memory as the storage device.
- Provides multiple data distribution policies.
- Exposes a non-POSIX interface.
- Open source project.

https://gitlab.arcos.inf.uc3m.es/admire/imss

Unified Communication X (UCX)

- Generic abstraction of the network layer
- Supported devices: Infiniband, Omni-path, TCP, shared memory
- Zero-copy
- MPICH, OpenMPI, Dash , Spark, Charm++, …



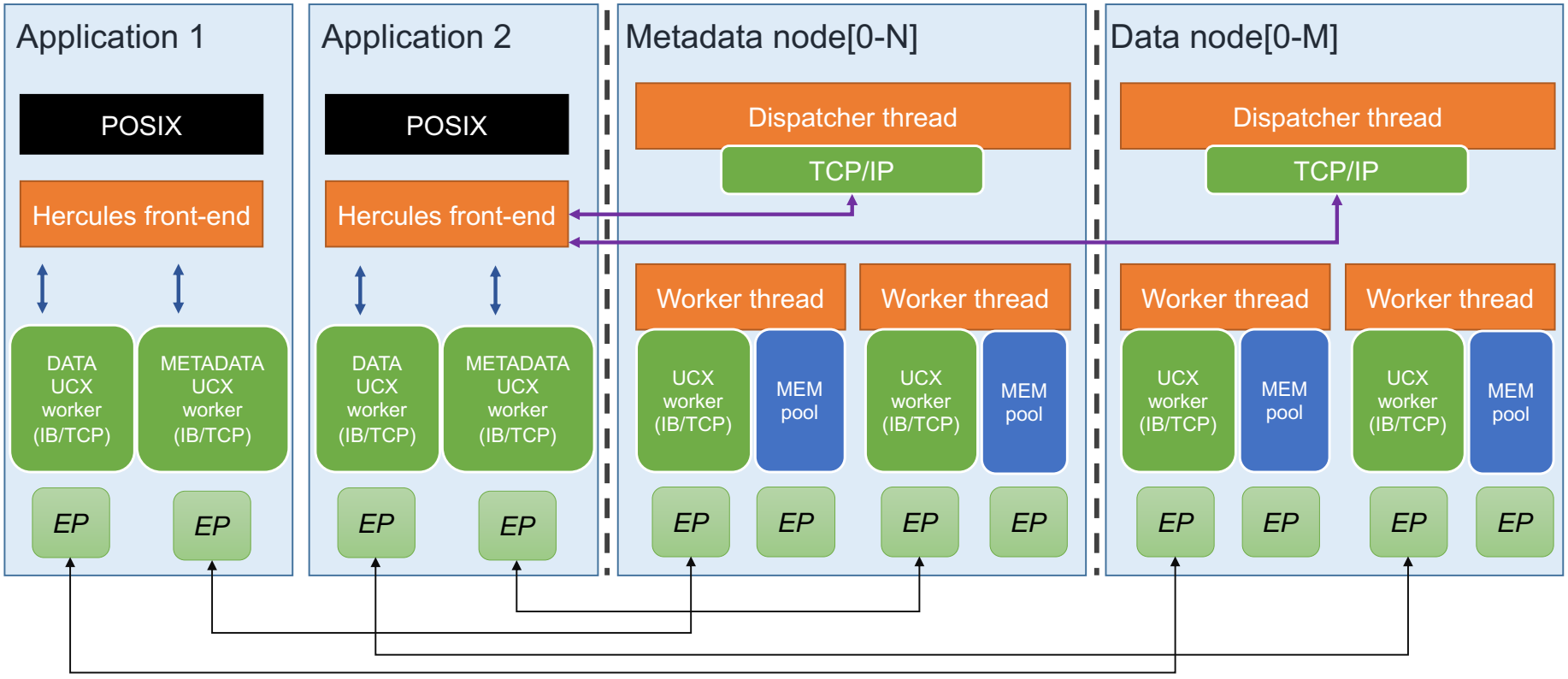
Features

- Migrated from ZeroMQ to UCX.
- Benefits of using UCX inside Hercules:
  - Multiple network interfaces/protocols available (TCP/IP, Omnipath, Infiniband supported).
  - Zero-copy message transfers of large data packages (>= 1 Mbytes).
  - Eliminated internal copies from application to network layer.
  - Asynchronous communication between peers.
  - RDMA QoS isolation.
  - End-point/two-sided-based communication.
- Fully implemented POSIX support (passed full IO500 benchmark).

New communication layer developed

- Non-blocking/tag-based communication (MPI style)
- Low-level communication schema (in contrast to Margo RPC)
- Client-side
  - Data and metadata UCX's workers enables **communication overlap**.
  - Malleability
    - Client nodes store a list of current available workers.
    - This list can be adapted during runtime.
  - QoS
    - Interfaces and protocols can be enabled/disabled to adapt **network requirements**.
    - Communication can be upgraded/downgraded (Infiniband to TCP).
  - Communication parameters configured by using environment variables.
- Server-side
  - One single listener per worker thread.
  - Stores a pool of active end-points (two-sided communication).

Architecture

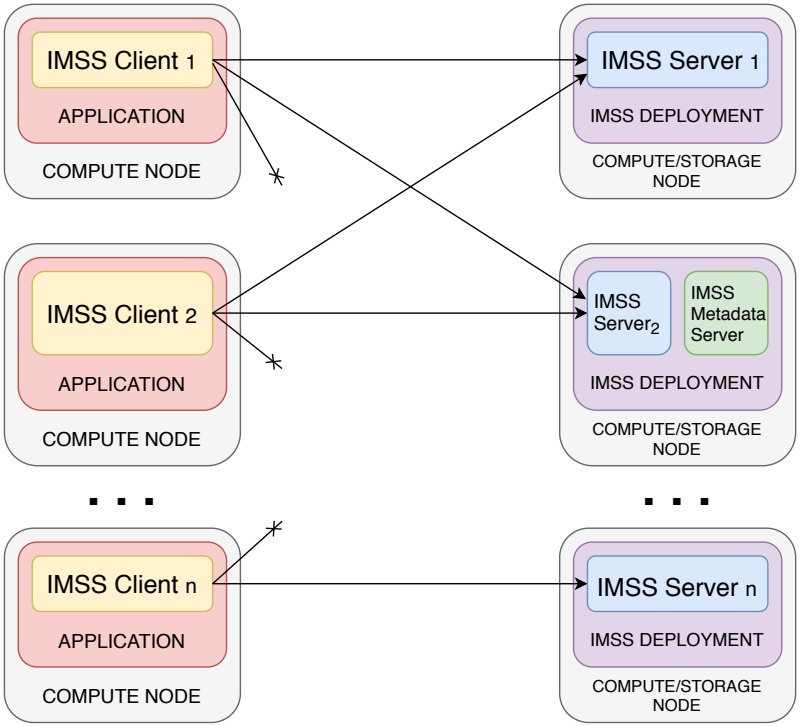


Data distribution policies

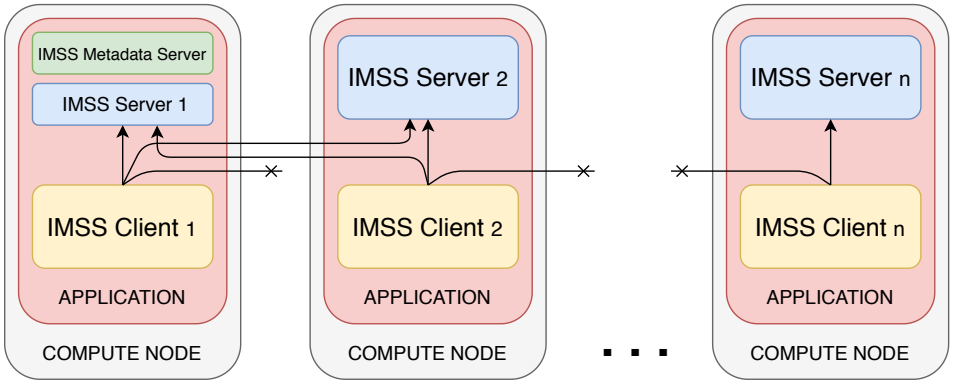
- **ROUND ROBIN**: data blocks are distributed among the Hercules servers.
- **BUCKETS**: each dataset is divided into the same number of chunks as number of servers. Each chunk is composed by a consecutive number of data blocks, equally distributed. Then, each chunk is assigned to a unique server.
- **HASHED**: a hash operation is applied over each data block key to discover the mapped server.
- **CRC16bits** & **CRC64bits**: similar to HASHED policy, but a sixteen/sixty four bits CRC operation is applied over the data block key.
- **LOCAL**: each data block is handled by the Hercules server running in the same node that the client.

Deployment strategies

application-dettached



application-attached



Access to the storage infrastructure

- API library
- FUSE
- LD_PRELOAD by overriding symbols

LD_PRELOAD

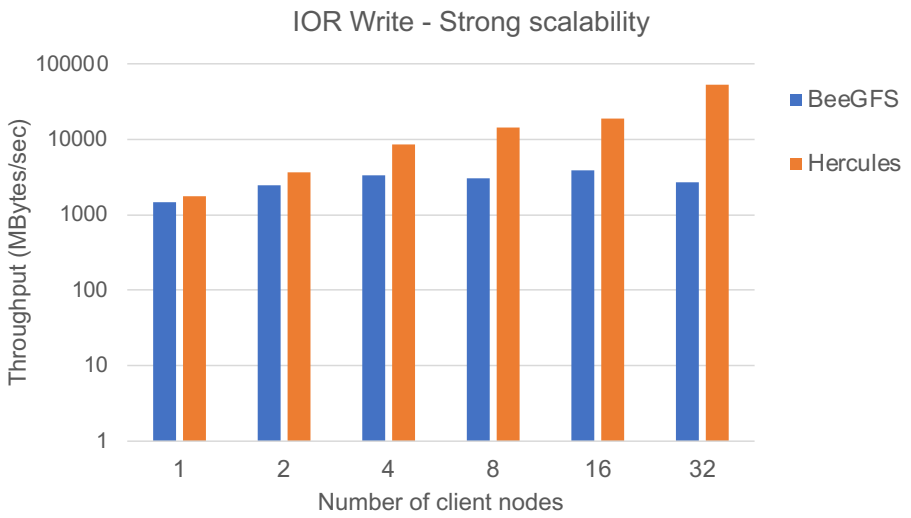
- Facilitates to integrated with existing applications.
- Works on booth attached and detached deployment strategies.
- Passed IO500 benchmark succesfully.

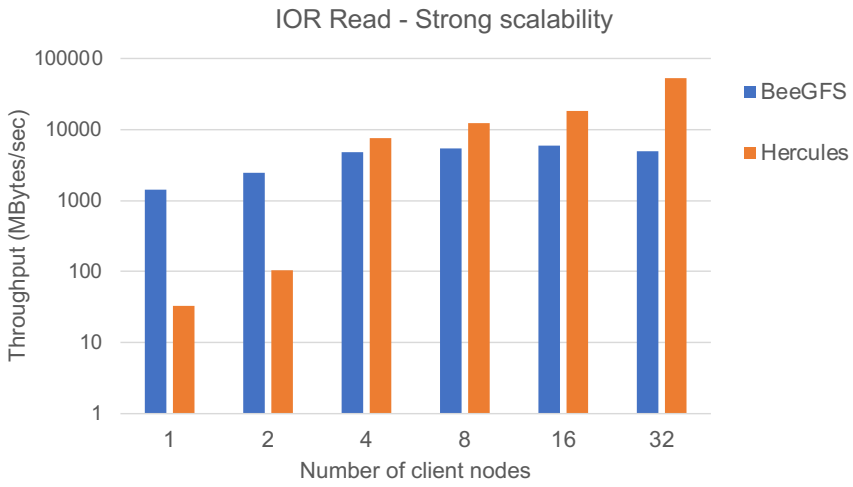
Evaluation

- University of Torino cluster
- 64 Broadwell compute nodes
- Intel Onmi-path running at 100 Gbps
- UCX 1.15
- OpenMPI 4.1

Evaluation (Scalability)

- IOR.
- Strong scalability, single shared file accesses.
- 512 Kbytes block size.





Future work

- Malleability:
  - Current efforts by modifying existing pools for controlling data location.
  - Missed API connector.
- Monitoring
  - Performance metrics already gathered (i.e., memory bandwidth, network bandwidth).
  - Missed connector
- QoS
  - Working progress

uc3m | Universidad Carlos III de Madrid

ARCOS

Network accelerated in-memory ad-hoc file system for data-centric and high-performance applications

Javier Garcia-Blas, Genaro Sanchez-Gallegos, Cosmin Petre and Jesus Carretero

University Carlos III of Madrid

fjblas@inf.uc3m.es

BDCSA2023: Big Data Convergence: from Sensors to Applications
PDP 2023
Naples, Italy