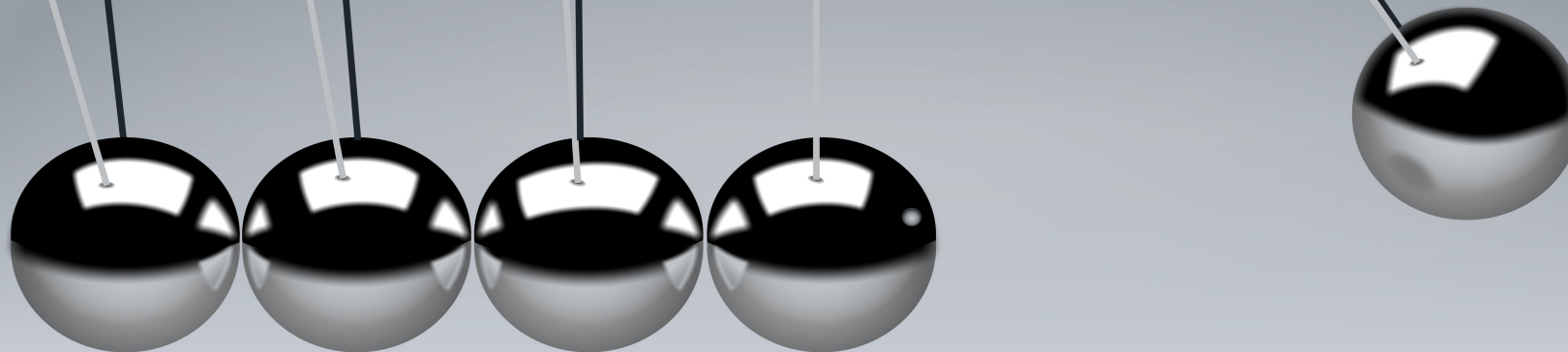


¿Es posible la convergencia HPC y Big Data?

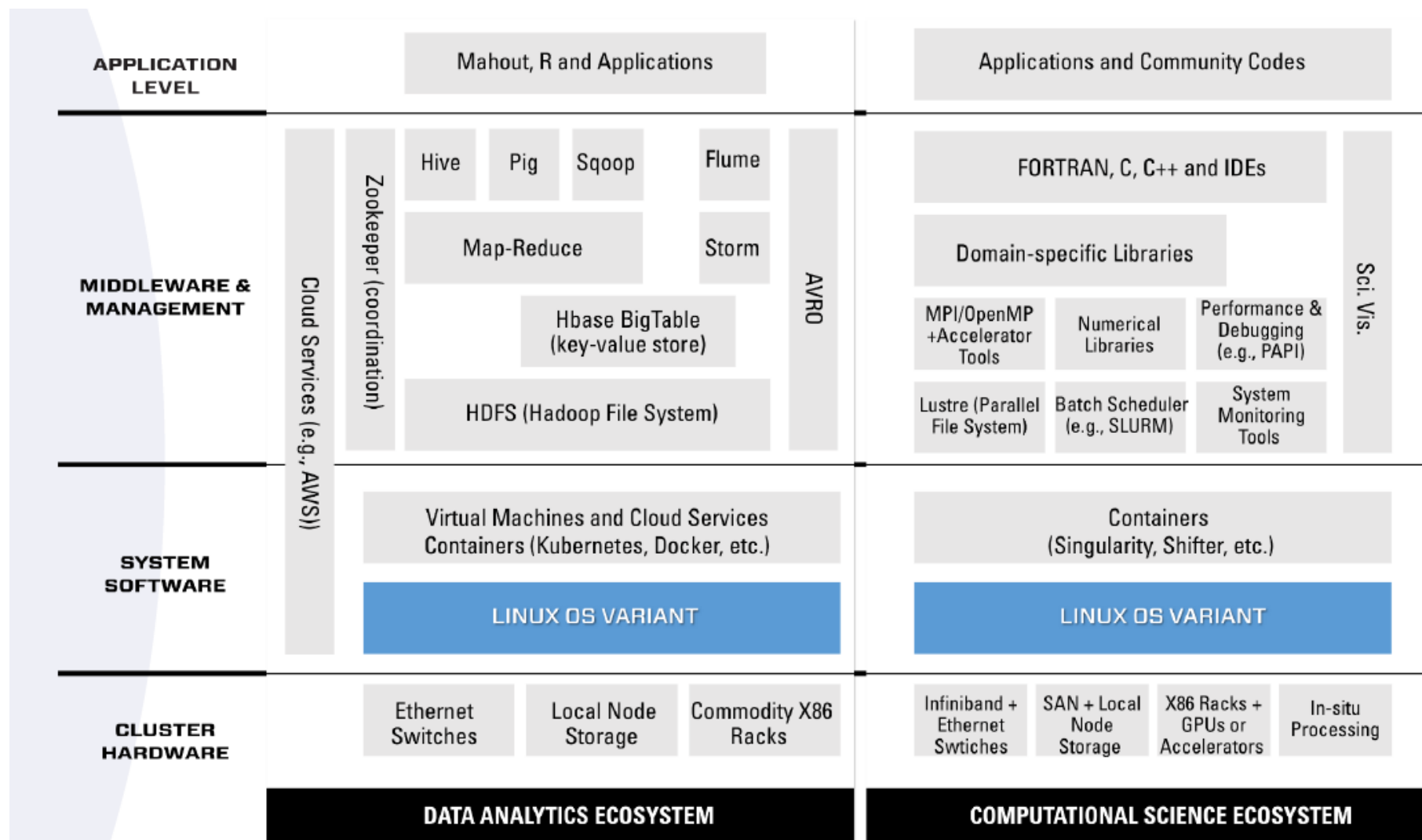
Lecciones aprendidas de los sistemas de E/S



María S. Pérez
mperez@fi.upm.es

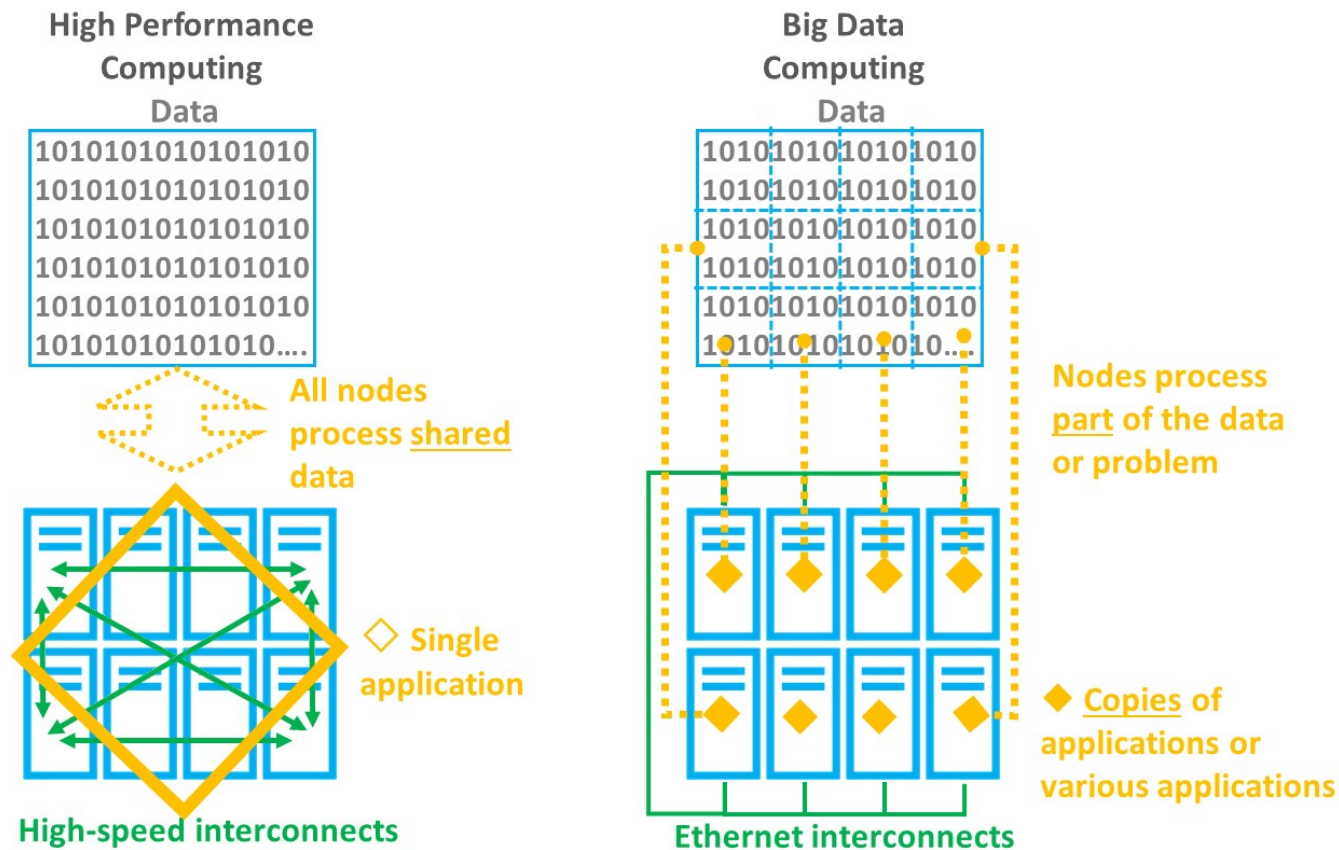
- **Introducción**
- Contexto
- Problema general de convergencia HPC-Big Data
- Convergencia HPC-Big Data a nivel de almacenamiento
- Estudio y evaluación
- Conclusiones y líneas de trabajo futuro

Divergencia DA-CS



Source: Big Data and Extreme-Scale Computing, BDEC

Divergencia a nivel de hardware y arquitectura



Source: "Creating synergies across HPC & Big Data platforms", BDVA-ETP4HPC White Paper

Diferencias históricas entre BD y HPC

	Carga de trabajo típica	Principios de diseño
Big Data	Aplicaciones intensivas en datos La mayoría del tiempo de ejecución es dedicado a E/S y manipulación de datos	Optimizado para coste Menor prioridad al rendimiento (IOPS)
HPC	Aplicaciones intensivas en cómputo La mayoría del tiempo de ejecución es dedicado a computación	Optimizado para rendimiento Menor prioridad al coste (FLOPS)



[European Commission](#) > [Strategy](#) > [Digital Single Market](#) > [Policies](#) >

Digital Single Market

POLICY

The European High-Performance Computing Joint Undertaking - EuroHPC

PAGE CONTENTS

Useful links

The European High-Performance Computing Joint Undertaking (EuroHPC JU) will pool European resources to develop top-of-the-range exascale supercomputers for processing big data, based on competitive European technology.

About HPC

[Policies](#) [+](#)

[Blog posts](#)

[News](#)

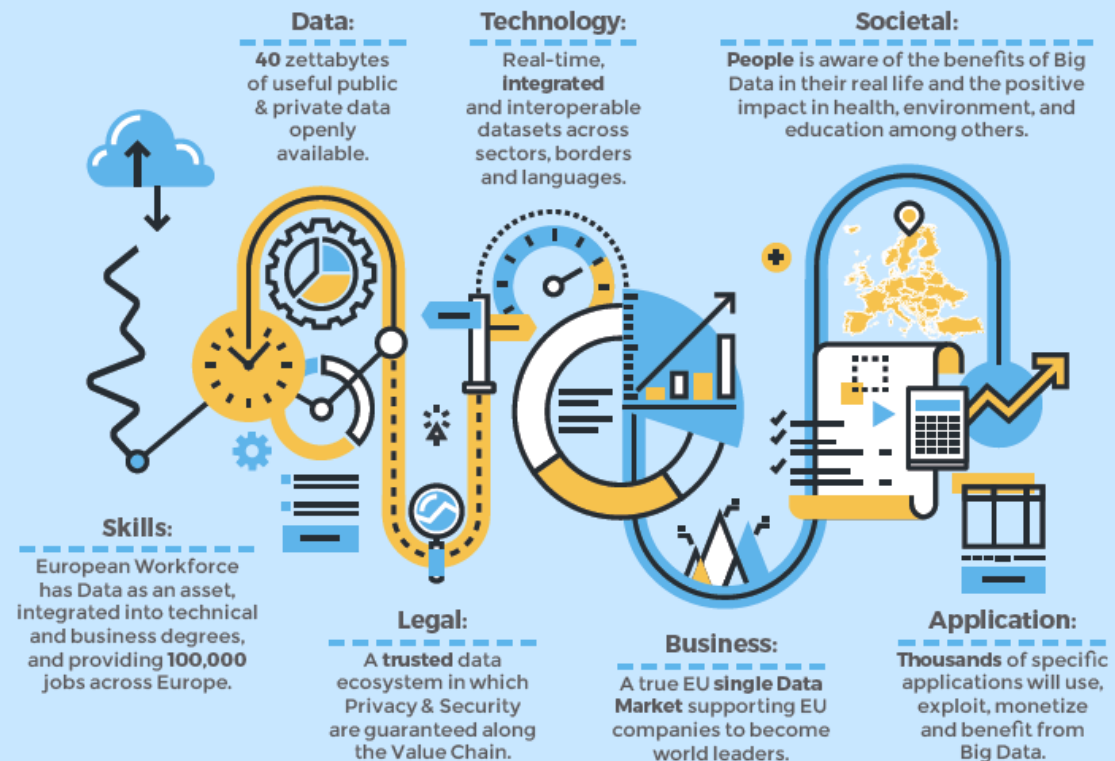
- Introducción
- **Contexto**
- Problema general de convergencia HPC-Big Data
- Convergencia HPC-Big Data a nivel de almacenamiento
- Estudio y evaluación
- Conclusiones y líneas de trabajo futuro



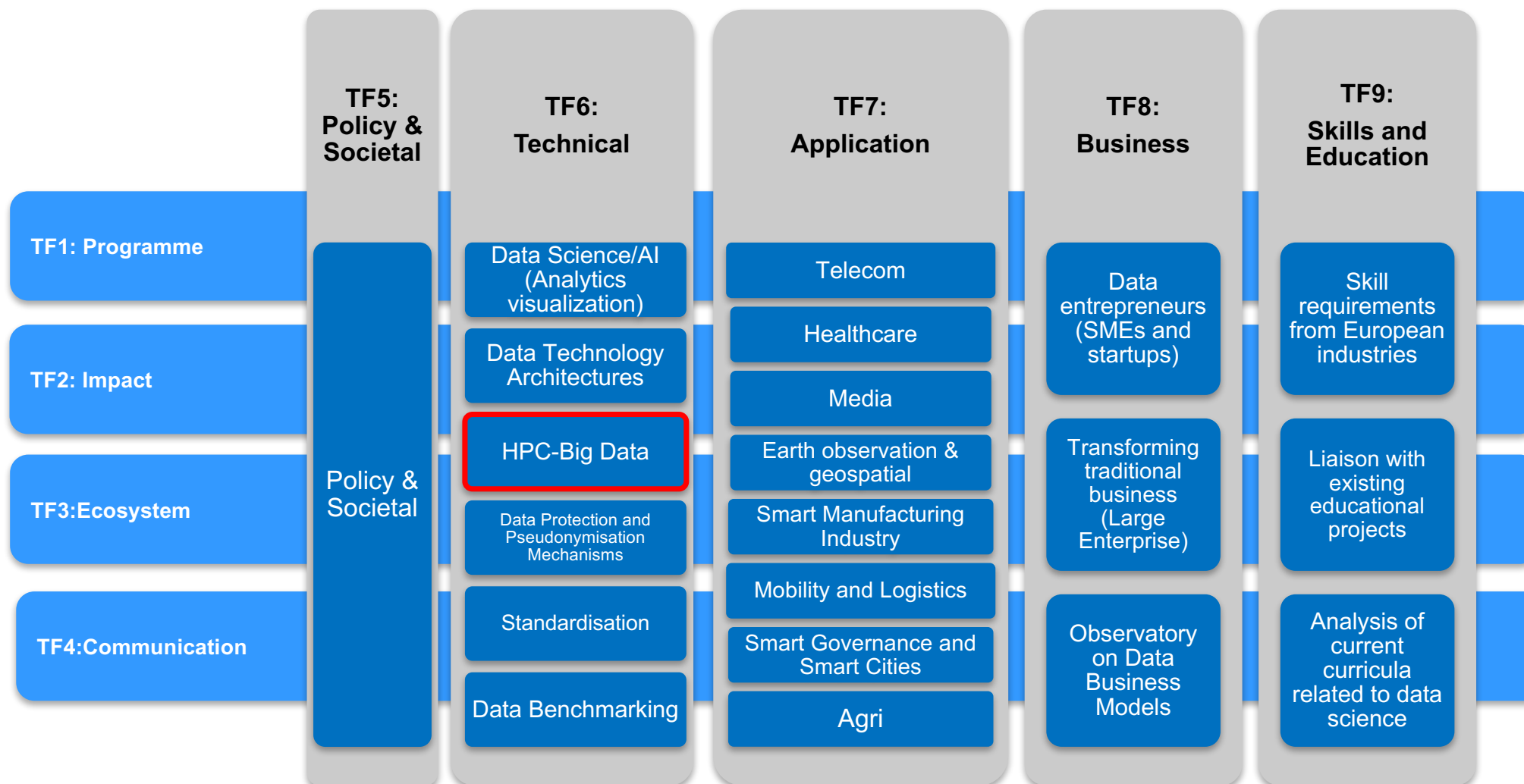
BDVA (~200) members include large industries, SMEs, research organisations and data users and providers to support the development and deployment of the EU Big Data Value Public-Private Partnership with the European Commission

BDVA focuses its activities on updating the multi-annual roadmap and on providing regular advice to enable the European Commission to prepare, draft and adopt the periodic Work Programmes, as well as on delivering Data Innovation Recommendations, developing Big Data Value Ecosystem, guiding Standards, and, facilitating Know-how exchange.

Big Data Value Vision for 2020



BDVA Task Forces



EuroHPC Joint Undertaking

<https://eurohpc-ju.europa.eu>

EuroHPC – Activities



EuroHPC
Joint Undertaking

Infrastructure & Operations		R&I, Applications & Skills		JU Admin/Running costs	
HPC Ecosystem					
~270	min 180	10	486 m€	EC	
~290	~186	10	486 m€	Participating States	
560	392	20	972 m€	Total	
0	~420 (in kind)	2	422 m€	Private Members	

■ Infrastructure + Operations

Acquisition of 2 pre-exascale machines and several (tbd) mid-range machines

■ Applications & Skills + R&I

R&I, exascale technologies and systems (incl. low-power processor); applications

■ JU Admin/running costs

■ JU Operation: 2019 to 2026

CABAHILA-CM: Convergencia BD-HPC: de los sensores a las aplicaciones

- Proyecto financiado por la Comunidad de Madrid, ayudas para la realización de programas de actividades de I+D entre grupos de investigación de la Comunidad de Madrid en tecnologías y en biomedicina (2019-2022)
- 4 grupos:
 - ArTeCS, Universidad Complutense de Madrid
 - ARCOS, Universidad Carlos III de Madrid
 - SciTrack, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
 - OEG, Universidad Politécnica de Madrid
- Objetivo: Mejorar la integración de los paradigmas de HPC y Big Data
 - Plataforma intensiva en computación y datos
 - Dos casos de uso: captación y modelado de datos de sensores para la predicción de radiación solar con alta resolución espacio-temporal y procesamiento de datos masivos en imagen médica del cerebro

- Introducción
- Contexto
- **Problema general de convergencia
HPC-Big Data**
- Convergencia HPC-Big Data a nivel de
almacenamiento
- Estudio y evaluación
- Conclusiones y líneas de trabajo futuro

Lower <-----> higher

Data Intensive

Traditional Big Data



Data-intensive workloads

[Example] Inferring new insights from big data-sets e.g. pattern recognition across suppliers, consumers, etc for data-driven insights and innovation

Extreme Data Analytics



Compute- and Data intensive workloads:

[Example] Reshaping healthcare through advanced analytics and artificial intelligence – leading to predictive and personalized medicine

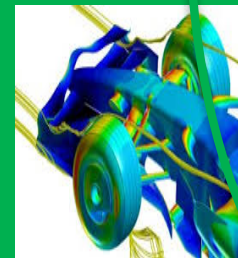
Enterprise IT



'Regular' workloads

[Example] Running the enterprise – HR, Legal, Payroll, finance, etc.

HPC



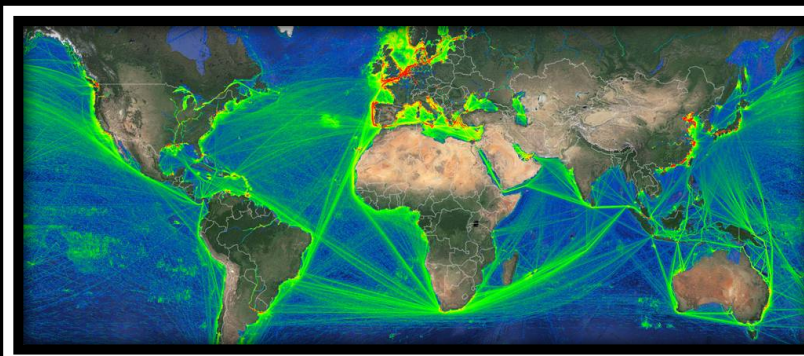
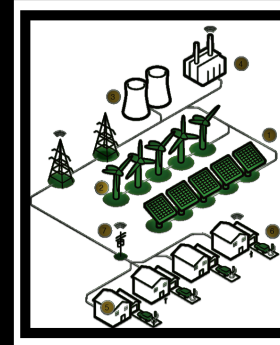
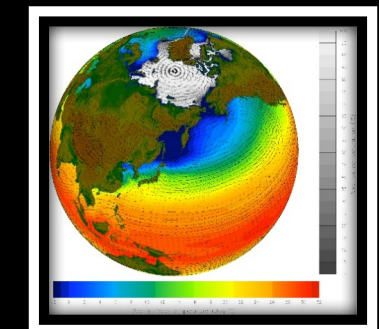
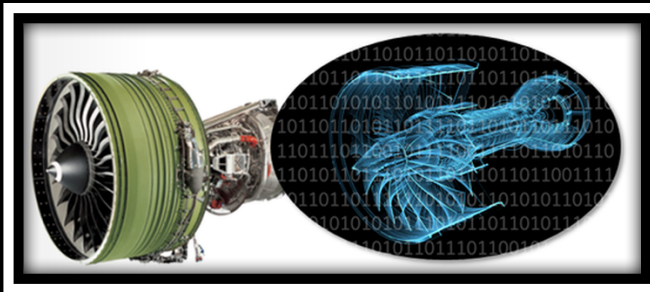
Compute-intensive workloads

[Example] Modelling and simulating focusing on interaction amongst parts of a system and the system as a whole e.g. product design

Compute Intensive
Lower <-----> higher

Source: Subgroup HPC-BD BDVA

Estudio de aplicaciones/casos de uso



Source: Subgroup HPC-BD BDVA

HPC, Big Data y Deep Learning

	<u>Supercomputing (SC)</u>	<u>Deep Learning (DL)</u>	<u>Big Data (BD)</u>
Apps...	Boundary Interaction Services In-house, commercial & OSS applications [e.g. Paraview], Remote desktop [e.g. Virtual Network Computing (VNC)], Secure Sockets Layer [e.g. SSL certificates]	Framework-dependent applications [e.g. NLP, voice, image], Web mechanisms [e.g. Google & Amazon Web Services], Secure Sockets Layer [e.g. SSL certificates]	Framework-dependent applications [e.g. 2/3/4-D], Secure Sockets Layer [e.g. SSL certificates]
Middleware & MGMT	Processing Services Domain specific frameworks [e.g. PETSc], Batch processing of large tightly coordinated parallel jobs [100s – 10000s of processes communicating frequently with each other]	DNN training & inference frameworks [e.g. Caffe, Tensorflow, Theano, Torch], DNN numerical libraries [e.g. dense LA], DNNs, statistics, diagnostics [e.g. ?]	Machine Learning (traditional) [e.g. Mahout, Scikit-learn], Analytics / Statistics [e.g. Python, ROOT, R, Matlab, SAS, SPSS, Sci-Py], Iterative [e.g. Apache Hama], Interactive [e.g. Dremel, Drill, Tez, Impala, Shark, Presto, BlinkDB], Batch / Map-Reduce [e.g. MapReduce, YARN, Sqoop, Spark], Real-time/streaming [e.g. Flint, YARN, Druid, Pinot, Storm, Samza, Spark streaming]
Middleware & MGMT	Model / Information Management Services Data Storage: Parallel File Systems [e.g. Lustre, GPFS, BeeGFS, PanFS, PVFS], I/O libraries [e.g. HDF5, PnetCDF, ADIOS]	Data Storage [e.g. HDFS, Hbase, Amazon S3, GlusterFS, Cassandra, MongoDB, Hana, Vora]	Serialization [e.g. Avro], Meta Data [e.g. HCatalog], Data Ingestion & Integration [e.g. Flume, Sqoop, Apache Nifi, Elastic Logstash, Kafka, Talend, Pentaho], Data Storage [e.g. HDFS, Hbase, Amazon S3, GlusterFS, Cassandra, MongoDB, Hana, Vora], Cluster Mgmt [e.g. YARN, MESOI]
Middleware & MGMT	Communication Services Messaging & Coordination [e.g. MPI/PGAS, direct fabric access], Threading [e.g. OpenMP, task-based models]	Messaging & Coordination [e.g. Machine Learning Scaling Library (MLSL)]	Messaging [e.g. Apache Kafka (streaming)]
Middleware & MGMT	Workflow / Task Services Conventional compiled languages [e.g. C/C++/Fortran], Scripting languages [e.g. Python, Julia,]	Scripting languages [e.g. Python]	Workflow & Scheduling [e.g. Oozie], Scripting languages [e.g. Keras, Mocha, Pig, Hive, JAQL, Python, Java, Scala]
System SW	System Management & Security Services Domain numerical libraries [e.g. PETSc, ScaLAPACK, BLAS, FFTW], Performance & debugging [e.g. DDT, Vtune, Vampir], Accelerator APIs [e.g. CUDA, OpenCL, OpenACC], Data Protection [e.g. System AAA, OS/PFS file access control], Batch scheduling [e.g. SLURM], Cluster management [e.g. OpenHPC], Container Virtualization [e.g. Docker], Operating System [e.g. Linux OS Variant]	Batching for training [built into DL frameworks], Reduced precision [e.g. Inference engines], Load distribution layer [e.g. Round robin/load balancing for inference], Accelerator APIs [e.g. CUDA, OpenCL], LA numerical libraries [e.g. BLAS, LAPACK, ...], Virtualisation [e.g. Docker, Kubernetes, VMware, Xen, KVM, HyperX], Operating System [e.g. Linux (RedHat, Ubuntu, etc.), (Windows?)]	Distributed Coordination [e.g. ZooKeeper, Chubby, Paxos], Provisioning, Managing & Monitoring [e.g. Ambari, Whirr, BigTop, Chukwa], SVM systems [e.g. Google Sofia, libSVM, svm-py, ...], Hardware Optimization Libraries [e.g. cuDNN, MKL-DNN, etc.], Virtualisation [e.g. Docker, Kubernetes, VMware, Xen, KVM, HyperX], Operating System [e.g. Linux (RedHat, Ubuntu, etc.), windows]
Hardware	Infrastructure Local storage [e.g. Storage & I/O nodes, NAS], Servers [e.g. CPU & Memory] [Gen Purpose CPU nodes, GPUs, FPGAs], Network [e.g. Infiniband & OPA fabrics]	Local storage [e.g. Local storage or NAS/SAN], Servers [e.g. CPU & Memory] [Gen Purpose CPU + GPU/FPGA, TPU], Network [e.g. Ethernet?]	Local storage [e.g. Direct attached Storage], Servers [e.g. CPU & Memory, [Gen Purpose CPU hyper-convergent nodes]], Network [e.g. Ethernet fabrics]

Source: “Creating synergies across HPC & Big Data platforms”, BDVA-ETP4HPC White Paper

- Introducción
- Contexto
- Problema general de convergencia HPC-Big Data
- **Convergencia HPC-Big Data a nivel de almacenamiento**
- Estudio y evaluación
- Conclusiones y líneas de trabajo futuro

HPC a nivel de almacenamiento

Aplicación HPC

Sistema de ficheros POSIX

- Lecturas y escrituras aleatorias en fichero
- Jerarquía (directorios)
- Permisos
- Renombrado atómico
- Protección multiusuario

POSIX

Lecturas y escrituras aleatorias en fichero
Jerarquía (directorios)
Permisos
Renombrado atómico
Protección multiusuario

POSIX

Lecturas y escrituras aleatorias en fichero

POSIX

Lecturas y escrituras aleatorias en fichero **Objetos**

HPC a nivel de almacenamiento

Aplicación HPC

Sistema de almacenamiento
basado en objetos

HPC y Big Data

Aplicación HPC

Sistema de almacenamiento
basado en objetos

Aplicación Big Data

HPC y Big Data

Aplicación HPC

Sistema de almacenamiento
basado en objetos

Aplicación Big Data

Sistema
K/V

BBDD

Sist. de
ficheros

HPC y Big Data

Aplicación HPC

Sistema de almacenamiento
basado en objetos

Aplicación Big Data

Sistema
K/V

BBDD

Sistema de ficheros

HPC y Big Data

Aplicación HPC

Sistema de almacenamiento
basado en objetos

Aplicación Big Data

Sistema
K/V

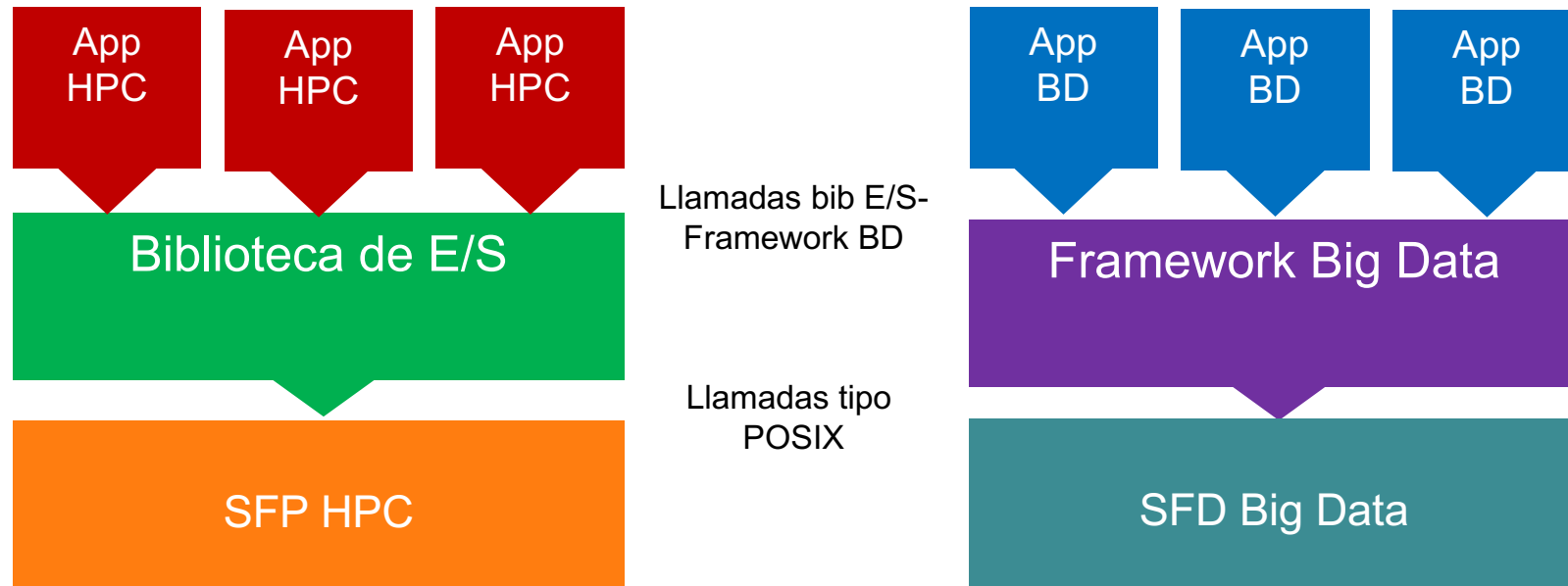
BBDD

Sistema de almacenamiento
basado en objetos

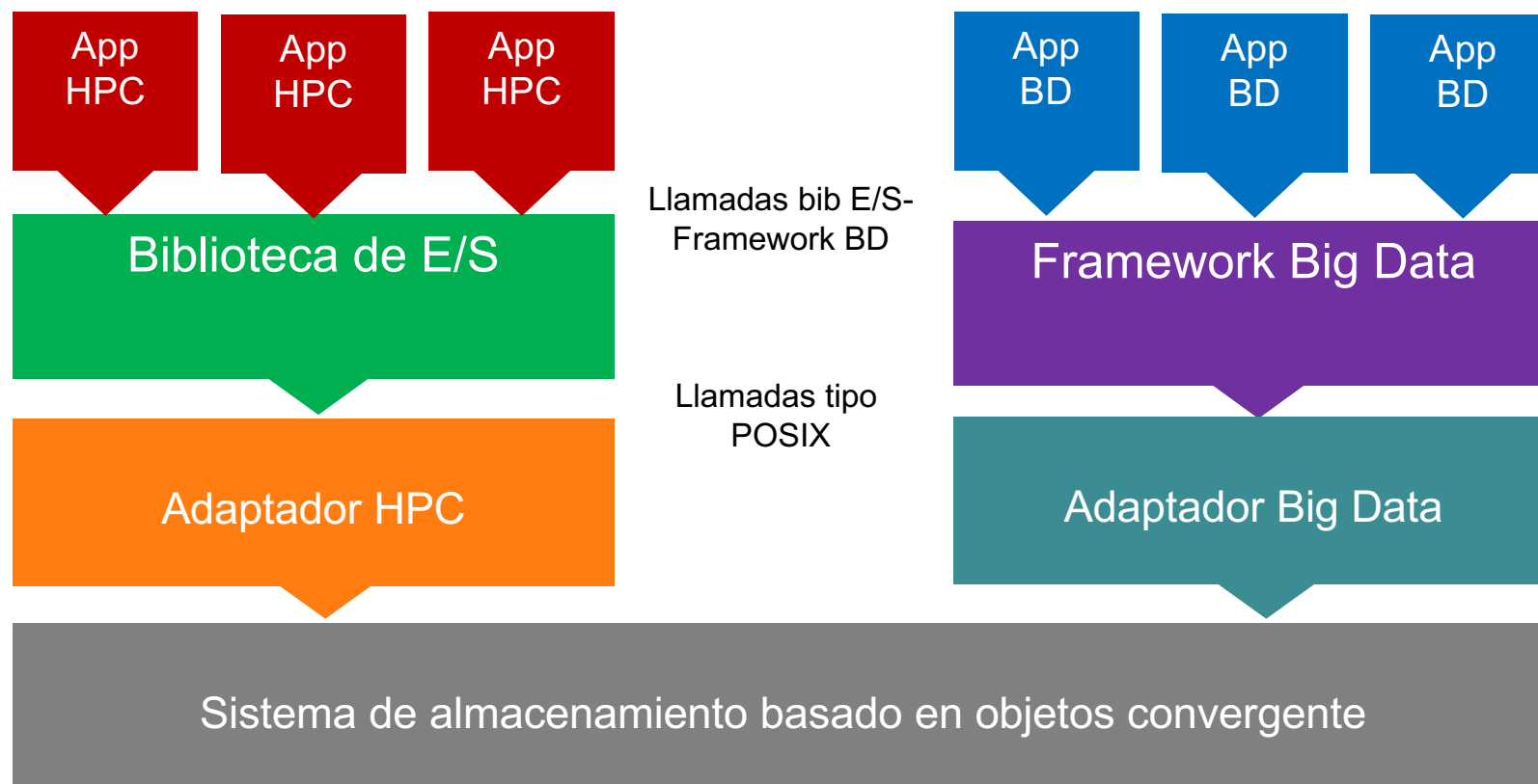
HPC y Big Data



Pila de almacenamiento actual



Pila de almacenamiento actual



- Introducción
- Contexto
- Problema general de convergencia HPC-Big Data
- Convergencia HPC-Big Data a nivel de almacenamiento
- **Estudio y evaluación**
- Conclusiones y líneas de trabajo futuro

Primitivas orientadas a objetos

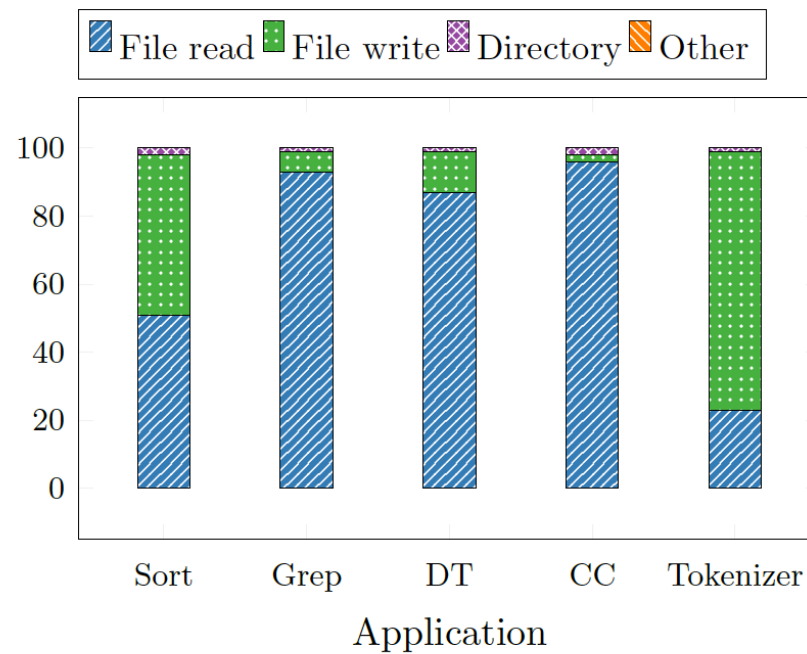
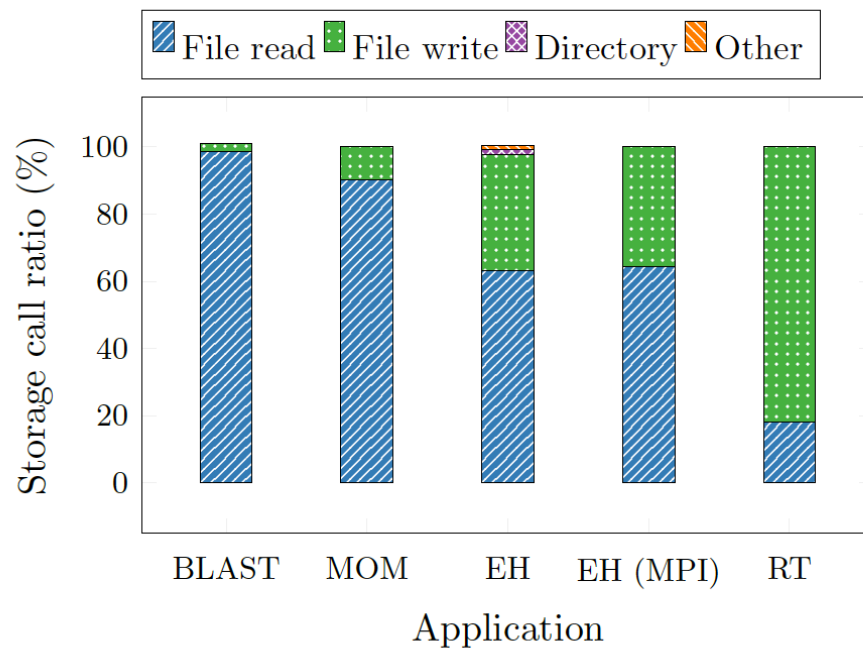
- Acceso al objeto: lectura (read) aleatoria del objeto, tamaño (size) del objeto
- Manipulación del objeto: escritura (write) aleatoria del objeto, truncado (truncate)
- Gestión del objeto: crear objetos (create), borrar objetos (delete)
- Estas operaciones son similares a las operaciones POSIX básicas sobre un fichero
- Las operaciones a nivel de directorio no tienen contrapartida a nivel de objeto:
 - Bajo número de este tipo de operaciones
 - Se pueden emular mediante la operación “scan” (operación poco eficiente, pero compensado por el hecho de utilizar un espacio de nombres plano y una semántica más sencilla)

Estudio sobre aplicaciones HPC/BD

Plataforma	Aplicación	Uso	Lecturas	Escrituras	Ratio R/W	Perfil
HPC/MPI	mpiBLAST	Protein docking	27.7 GB	12.8 MB	$2.2 \cdot 10^3$	Read-intensive
	MOM	Oceanic model	19.5 GB	3.2 GB	6.09	Read-intensive
	ECOHAM	Sediment propagation	67.4 GB	71.2 GB	0.94	Balanced
	Ray Tracing	Video processing	0.4 GB	9.7 GB	$4.1 \cdot 10^{-2}$	Write-intensive
Cloud/Spark	Sort	Text processing	5.8 GB	5.8 GB	1.00	Balanced
	Connected Component	Graph processing	13.1 GB	71.2 MB	$1.8 \cdot 10^2$	Read-intensive
	Grep	Text processing	55.8 GB	863.8 MB	66.14	Read-intensive
	Decision Tree	Machine Learning	59.1 GB	4.7 GB	12.57	Read-intensive
	Tokenizer	Text processing	55.8 GB	235.7 GB	0.23	Write-intensive

Pierre Matri, Yevhen Alforov, Álvaro Brandón, María S. Pérez et al. Mission possible: Unify HPC and Big Data stacks towards application-defined blobs at the storage layer. Future Generation Computer Systems, In press.

Distribución de operaciones

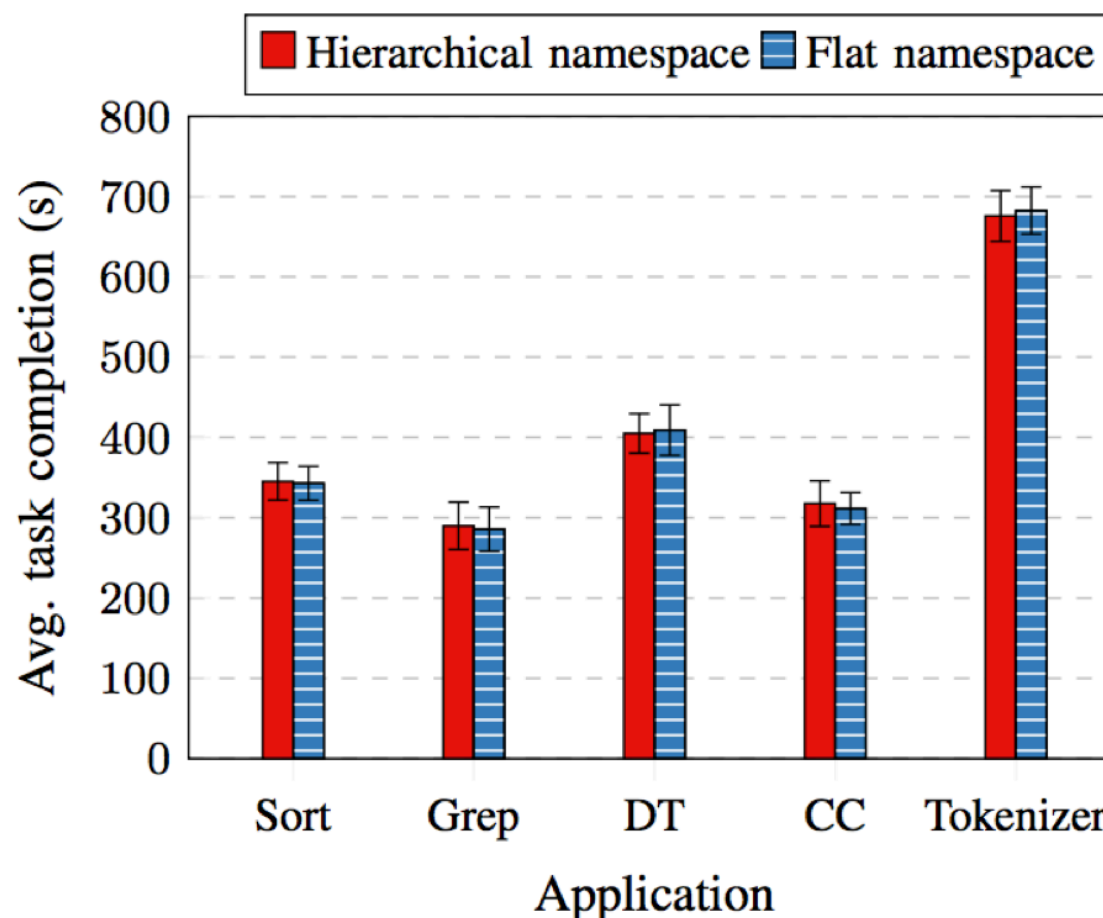
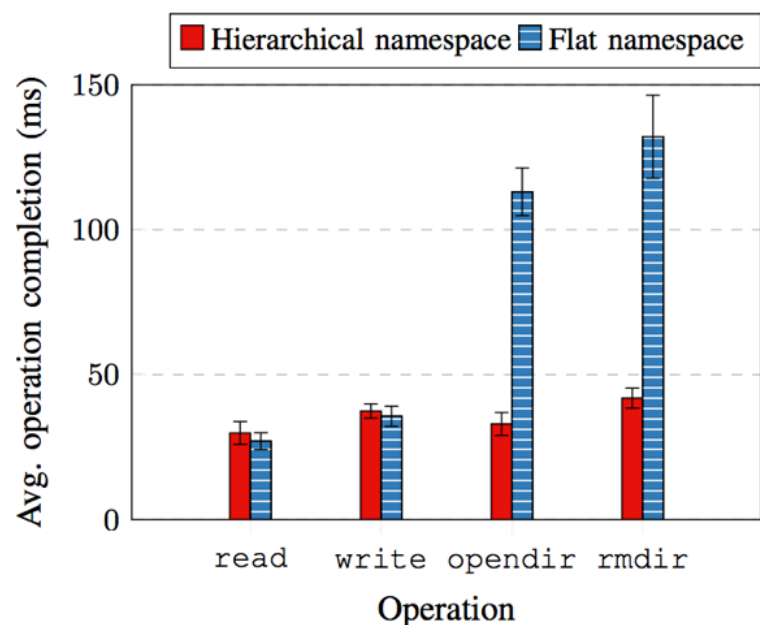


Operaciones sobre directorios (App BD)

Operación	Acción	Núm. Oper.
mkdir	Crear directorio	43
rmdir	Borrar directorio	43
opendir (directorio datos)	Abrir directorio	5
Opendir (otros directorios)	Abrir directorio	0

Operación original	Operación reescrita
create(/foo/bar)	create(/foo__bar)
open(/foo/bar)	open(/foo__bar)
read(fd)	read(bd)
write(fd)	write(bd)
mkdir(/foo)	-----
opendir(/foo)	scan(/), devolver todos los ficheros que encajan con foo__*
rmdir(/foo)	scan(/), borrar todos los ficheros que encajan con foo__*

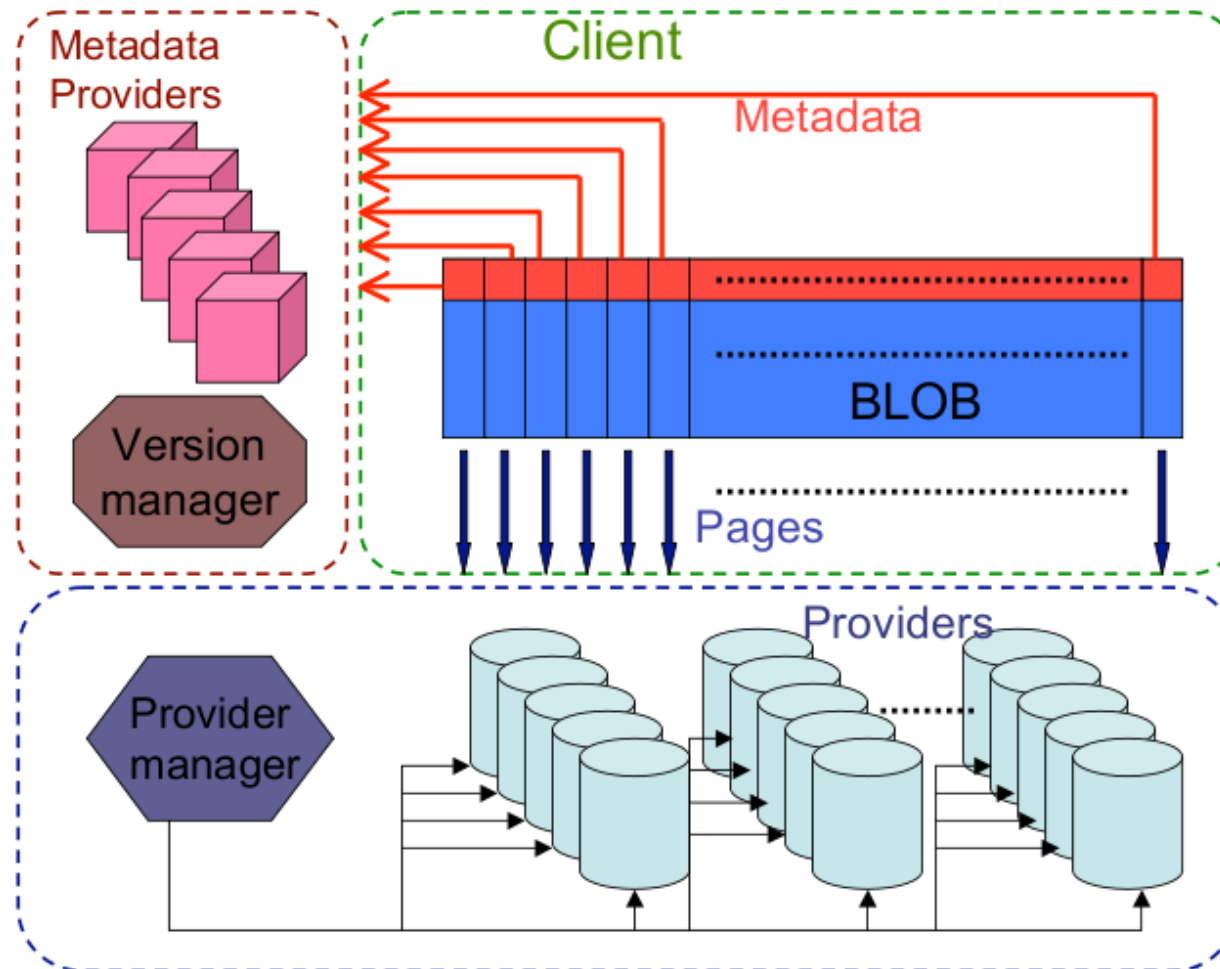
Influencia de las operaciones sobre directorios



BlobSeer/RADOS vs Lustre (HPC) and HDFS/Ceph (BD)

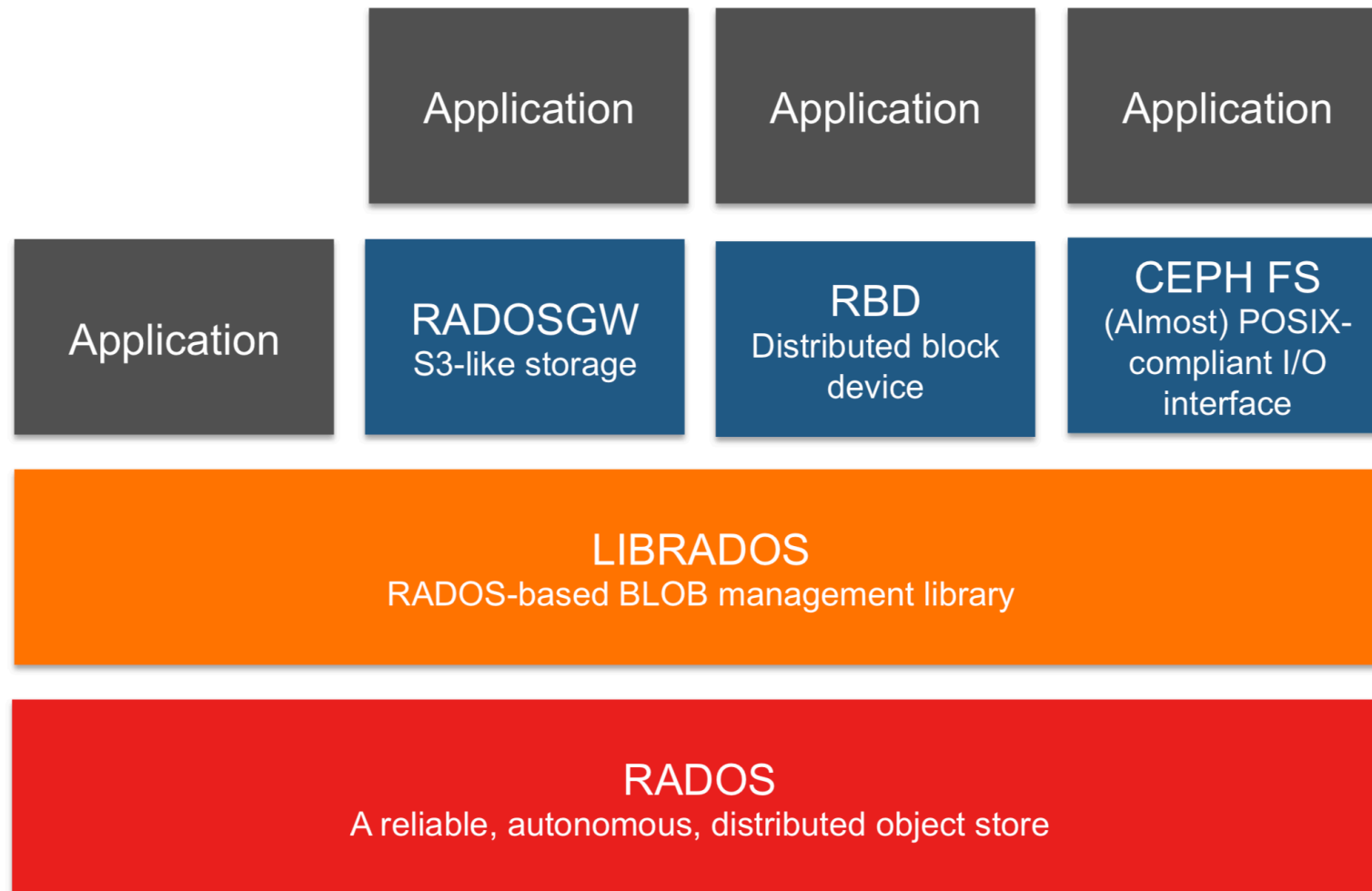
- Grid'5000 testbed distribuido en 11 sitios en Francia y Luxemburgo (parapluie cluster, Rennes)
- Cada nodo: 2 x 12-core 1.7 Ghz 6164 HE, 48 GB of RAM y 250 GB HDD.
- Apps HPC: Lustre 2.9.0 y MPICH 3.2 [67], en un cluster de 32 nodos (InfiniBand)
- Apps BD: Spark 2.1.0, Hadoop / HDFS 2.7.3 and Ceph Kraken, en un cluster de 32 nodos (Gigabit Ethernet)

BlobSeer

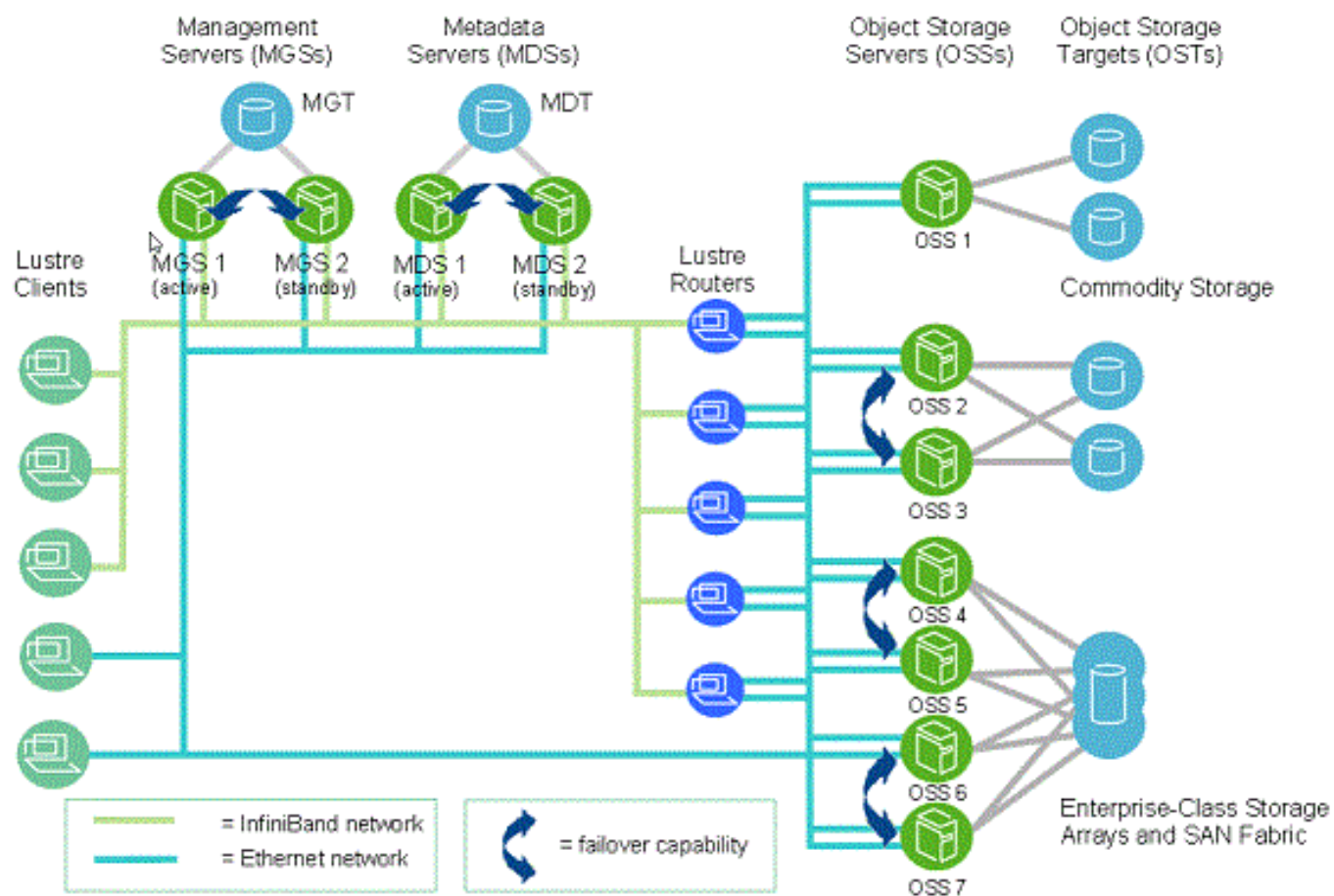


Bogdan Nicolae; Gabriel Antoniu; Luc Bougé; Diana Moise; Alexandra Carpen-Amarie. 2011. BlobSeer: Next-generation data management for large scale infrastructures. J. Parallel Distrib. Comput. 71, 2 (February 2011), 169-184.

RADOS/Ceph



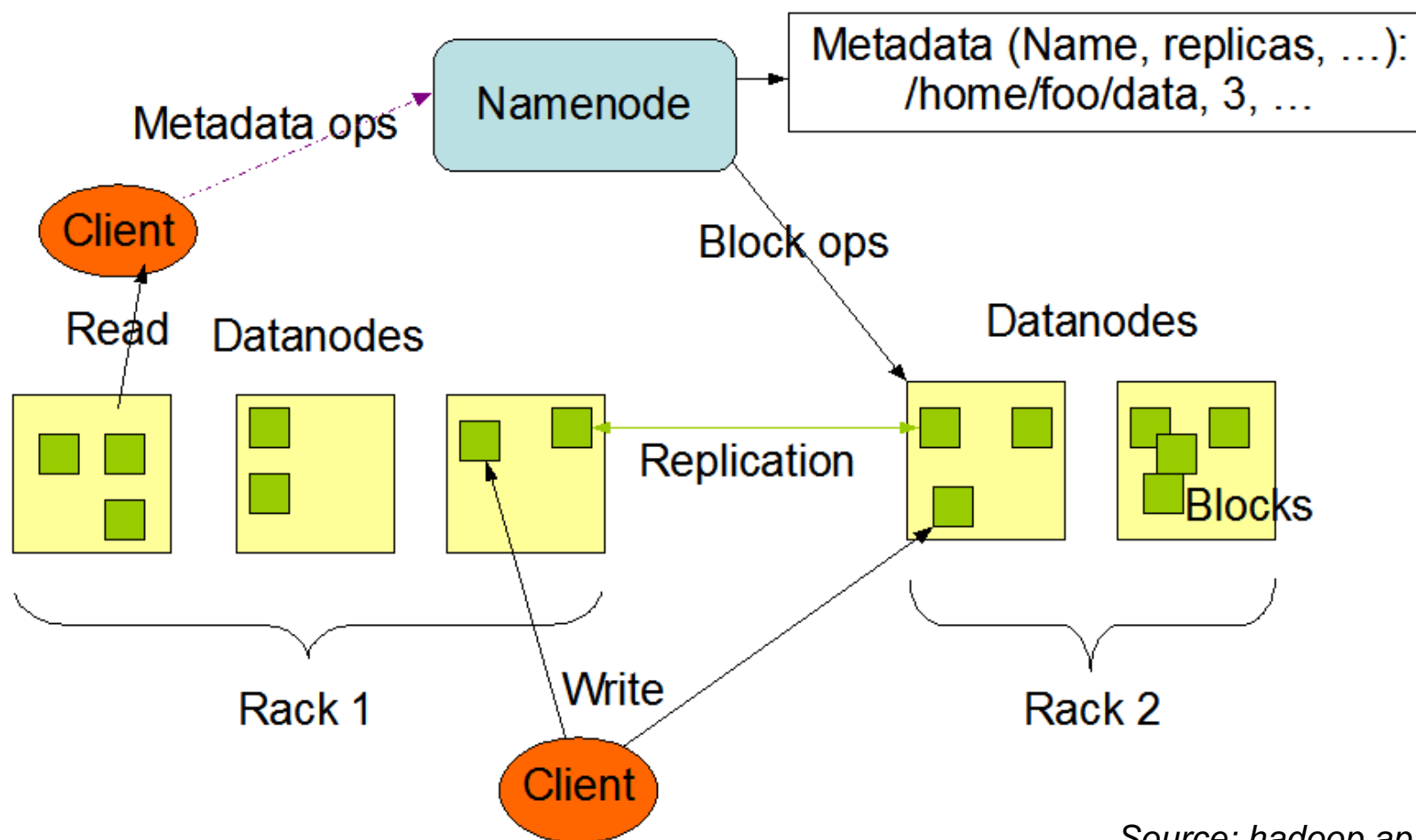
Lustre



Source: lustre.org

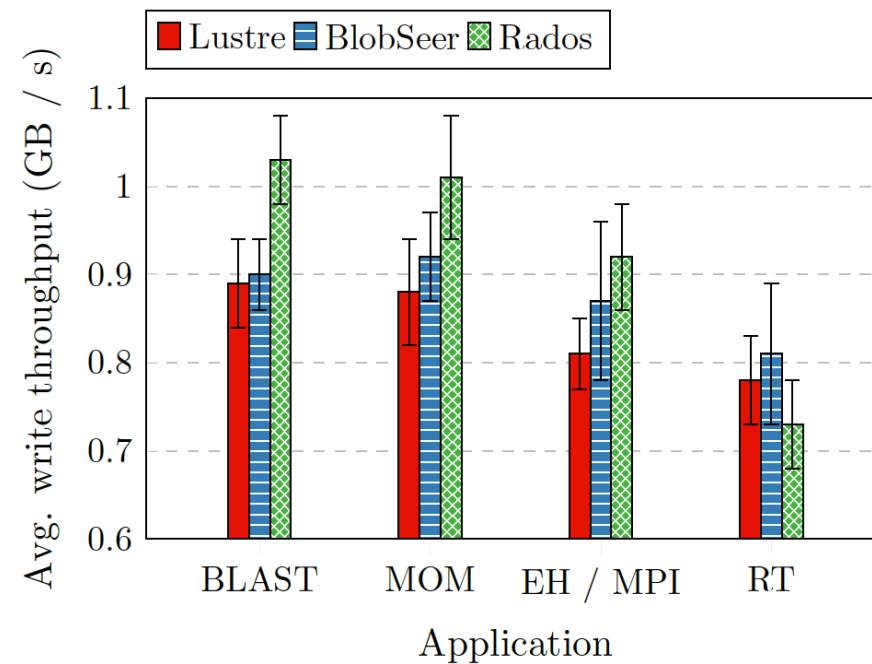
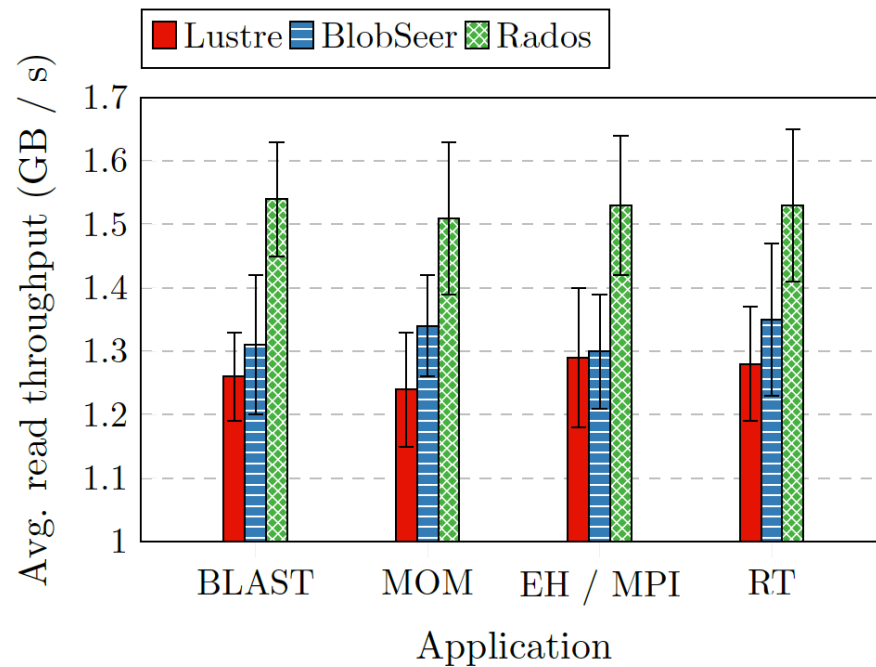
HDFS

HDFS Architecture

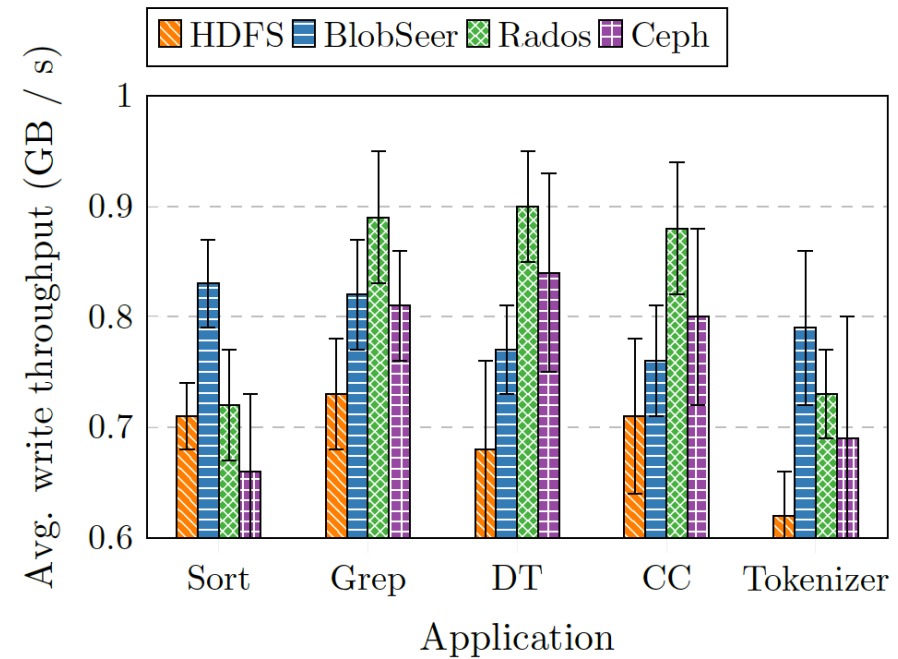
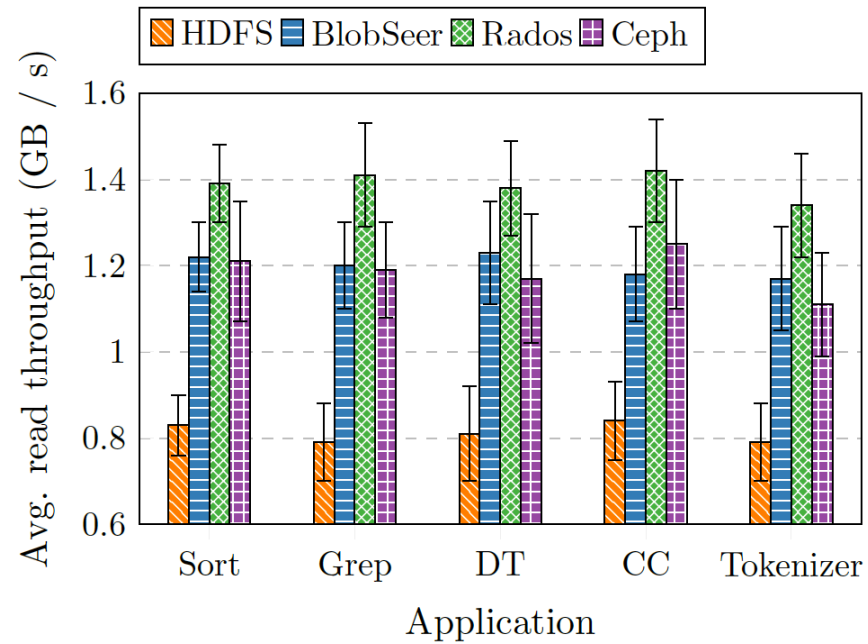


Source: hadoop.apache.org

Aplicaciones HPC



Aplicaciones BD



Análisis del estudio

- La **convergencia a nivel de almacenamiento es posible mediante el uso de almacenamiento basado en objetos**, llevando a una mejora en el rendimiento para ambos tipos de plataformas (HPC y Cloud)
- Utilizando objetos, se consigue una **mejora de hasta el 32%**
 - Principalmente por el **espacio de nombres plano**
 - Rados: capacidad de lectura directa y esquema sencillo y descentralizado de gestión de metadatos (gran rendimiento para lecturas)
 - BlobSeer: control de concurrencia multi-versión da soporte a gran velocidad de escritura para cargas de trabajo altamente concurrentes (gran rendimiento para escrituras)
- **Problemas de ambos sistemas:**
 - Aunque el rendimiento de **Rados** es excelente cuando hay baja contención de escritura, su **control de concurrencia basado en cerrojos** limita el rendimiento de casos de uso altamente concurrentes.
 - El control de concurrencia multi-versión de BlobSeer da buen soporte de escritura, pero el **árbol de metadatos distribuidos de BlobSeer** supone una latencia de lectura significativa

Týr

- ¿Se pueden lograr los beneficios de ambos sistemas?
- Además de combinar las ventajas de Rados y BlobSeer, hay un conjunto significativo de casos de uso que requieren semánticas de consistencia más estrictas
 - Indexación y agregación de datos (Ejemplo: ALICE CERN LHC experiment)
 - Distributed shared logs (Ejemplo: Computational steering + in-situ visualization)

Týr

One problem...

A scientific monitoring service, monitoring the
ALICE CERN LHC experiment:

- Ingests events at a rate of up to 10⁹ events per second
- Produces more than 10⁹ data points per second

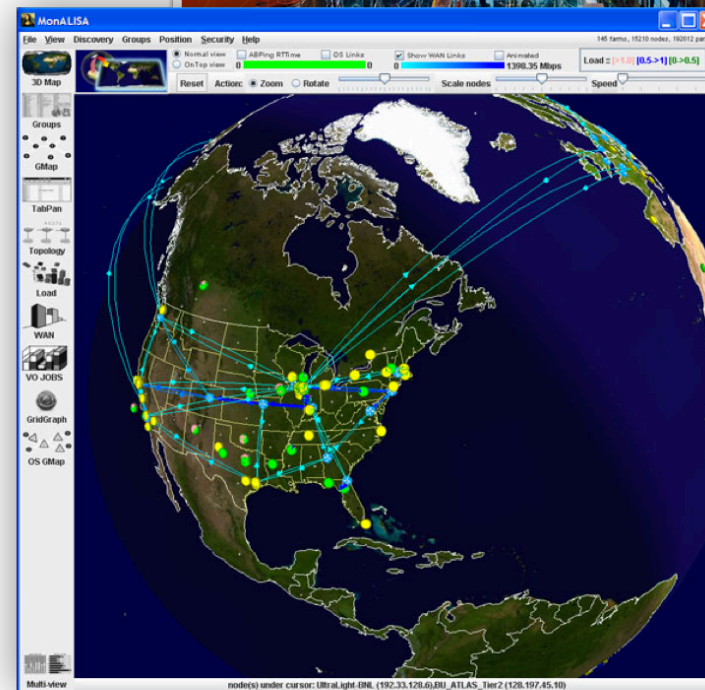
Computes 35.000+ events per second

Current load

tyr
a (converged) blob store with built-in transactions

Visualization support

Leads
Stability



Pierre Matri; Alexandru Costan; Gabriel Antoniu; Jesús Montes; María S. Pérez. "Týr: Blob Storage Systems Meet Built-In Transactions". SC '16 Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Article n. 49, Best student paper award finalist

Týr

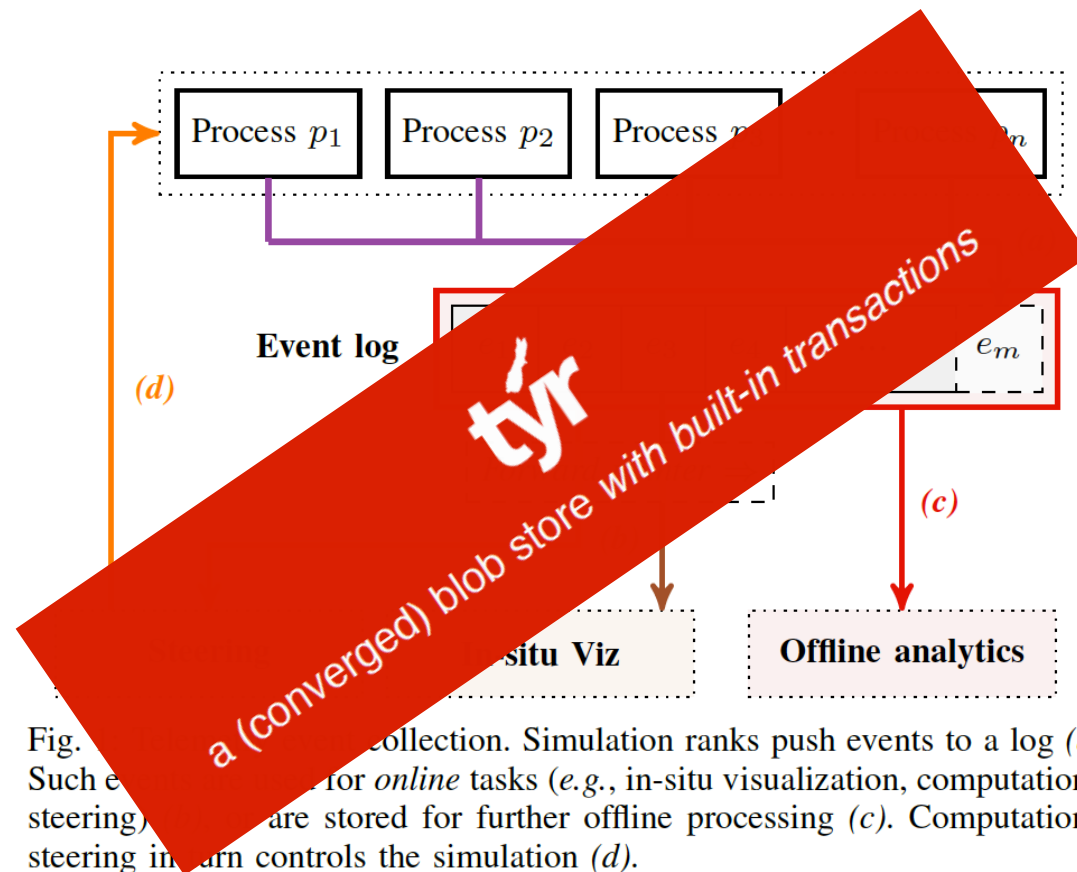


Fig. 1. Týr architecture. Simulation ranks push events to a log (a). Such events are then processed for *online* tasks (e.g., in-situ visualization, computational steering) (b) or stored for further *offline* processing (c). Computational steering in turn controls the simulation (d).

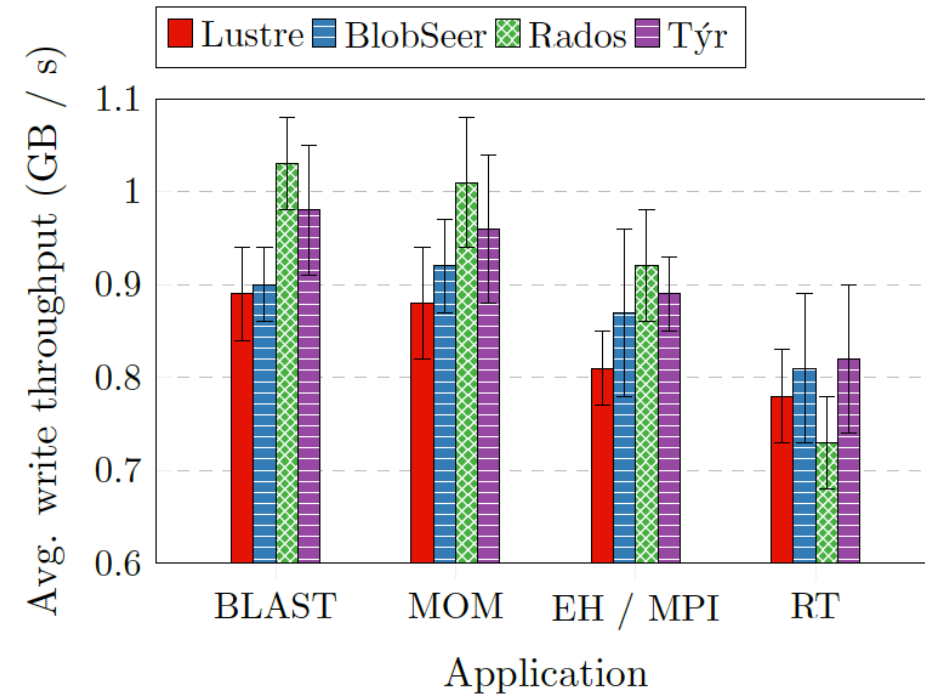
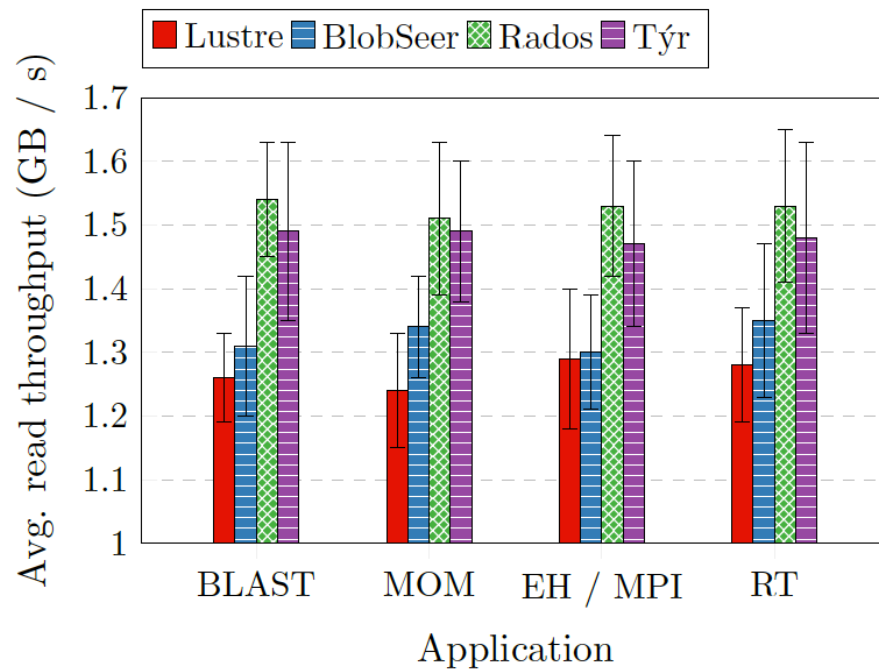
Pierre Matri; Philip Carns; Robert Ross; Alexandru Costan; María S. Pérez; Gabriel Antoniu;. "SLoG: A large-scale Logging Middleware for HPC and Big Data convergence". ICDCS'2018. pp. 1507-1512, 2018.

Diseño de Týr

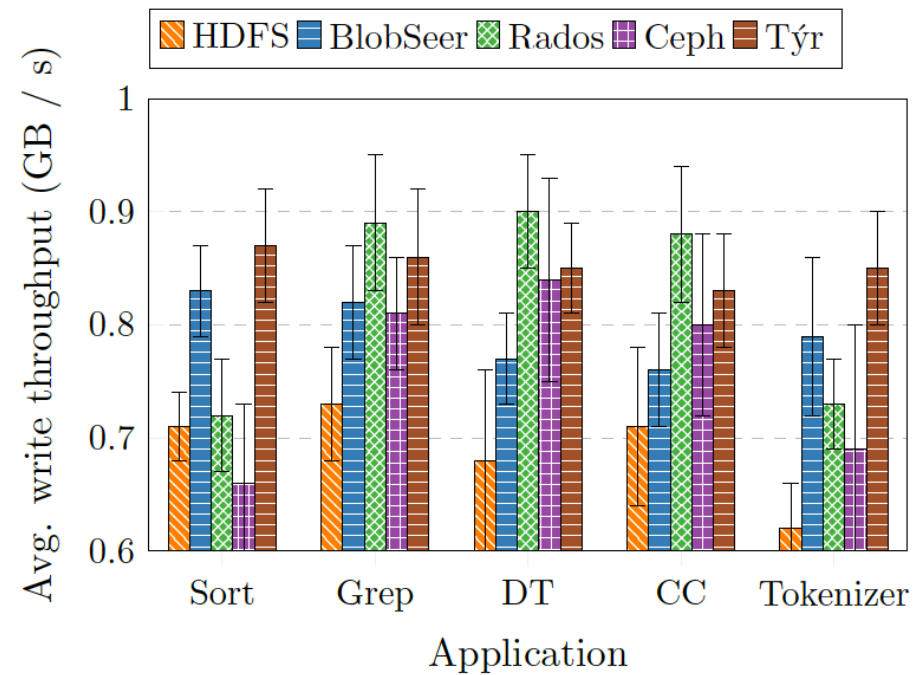
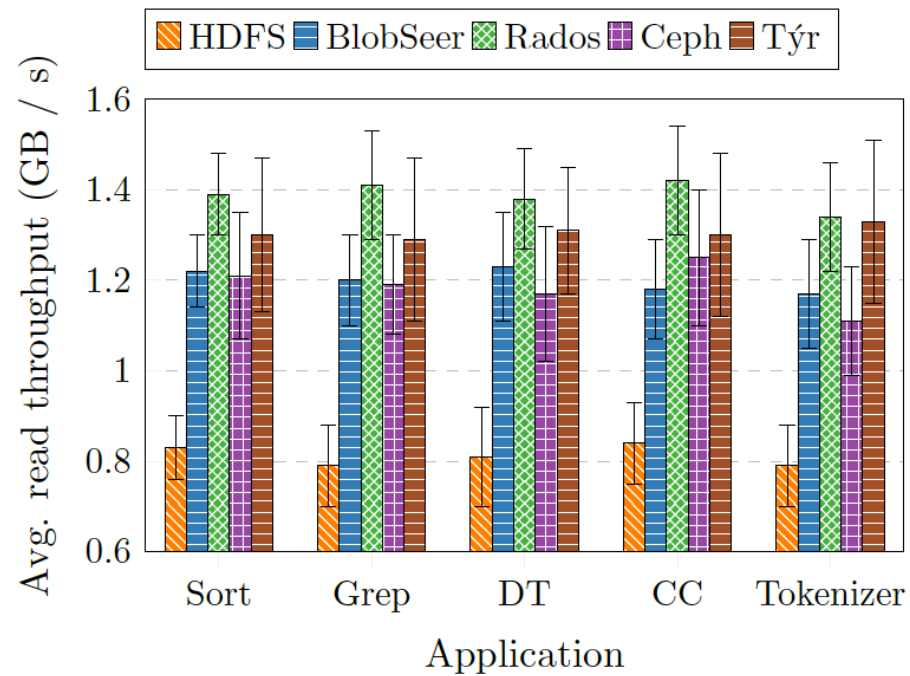
- Distribución de datos predecible
 - Combinación de técnicas de *data striping* y *consistent hashing*
 - Elimina la necesidad de un servidor de metadatos centralizado
- Control de concurrencia multi-versión transparente
 - Se utiliza *versioning* a nivel de *chunk* y números no secuenciales para el versionado
- Semántica transaccional ACID
 - Utilizando el protocolo transaccional Warp*
 - Utilizan cadenas de los servidores afectados por las transacciones y grafos de dependencias
- Operaciones de transformación atómicas
 - Eficientes operaciones *read-modify-write*
 - Particularmente interesante cuando las operaciones de transformación son sencillas (aritméticas, a nivel de bit)
 - El cliente no comunica el nuevo dato a escribir, sino la modificación a aplicar, evitando *two-round trips*
- Prototipo software con aproximadamente 25,000 líneas de código Rust y GNU C

* R. Escriva, B. Wong and E. Sirer. Warp: Lightweight multi-key transactions for key-value stores. arXiv preprint arXiv:1509.07815, 2015.

Aplicaciones HPC



Aplicaciones BD



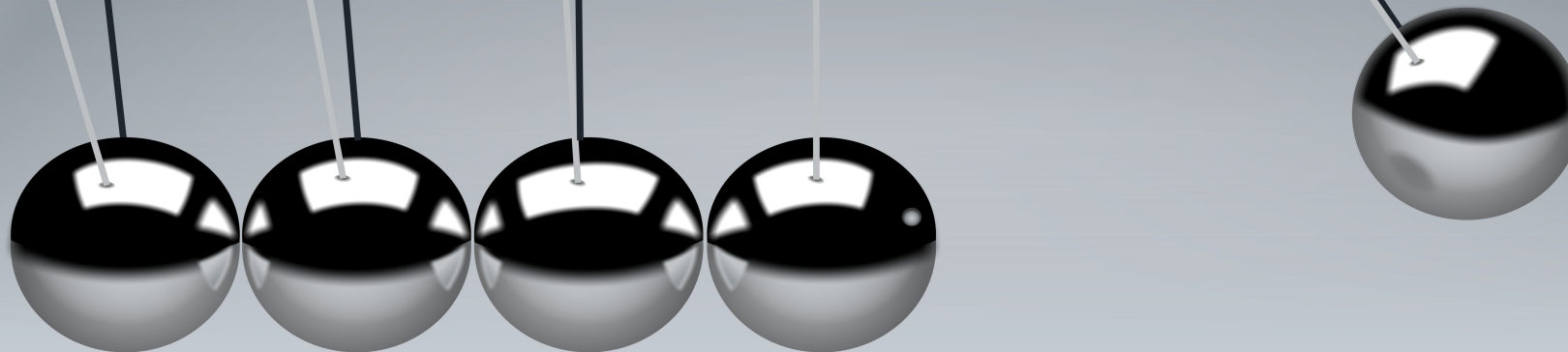
- Introducción
- Contexto
- Problema general de convergencia HPC-Big Data
- Convergencia HPC-Big Data a nivel de almacenamiento
- Estudio y evaluación
- **Conclusiones y líneas de trabajo futuro**

Conclusiones y líneas de futuro

- **El diseño de Týr supera significativamente BlobSeer y sistemas de ficheros tradicionales**, tanto para aplicaciones HPC como BDA
 - Escrituras no bloqueantes, mediante el uso de control de concurrencia multi-versión
 - Escrituras directas utilizando técnicas de *consistent hashing*
- **Týr sufre una pequeña penalización respecto a Rados**
 - **Excepto para aplicaciones intensivas en escritura**, debido a la eficiencia del control de concurrencia multi-versión
 - **Resultado de unas garantías de consistencia más fuerte** (transacciones)
- Este es un **primer paso sobre el que se puede trabajar para la convergencia HPC y BDA** a otros niveles, más allá del almacenamiento

¿Es posible la convergencia HPC y Big Data?

Lecciones aprendidas de los sistemas de E/S



María S. Pérez
mperez@fi.upm.es