Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

Analyses of the usefulness of Software Defined Storage Solutions for Web-based Digital Preservation Applications

Peter Tiernan
Systems and Storage Engineer
Digital Repository of Ireland
TCD

An Roinn Post, Fiontar agus Nuálaíochta
Department of Jobs, Enterprise and Innovation

Ireland's EU Structural Funds Programmes 2007 - 2013
Co-funded by the Irish Government and the European Union

HEA
Higher Education Authority
An tÚdarás um Ard-Oideachas

EUROPEAN REGIONAL DEVELOPMENT FUND

Outline

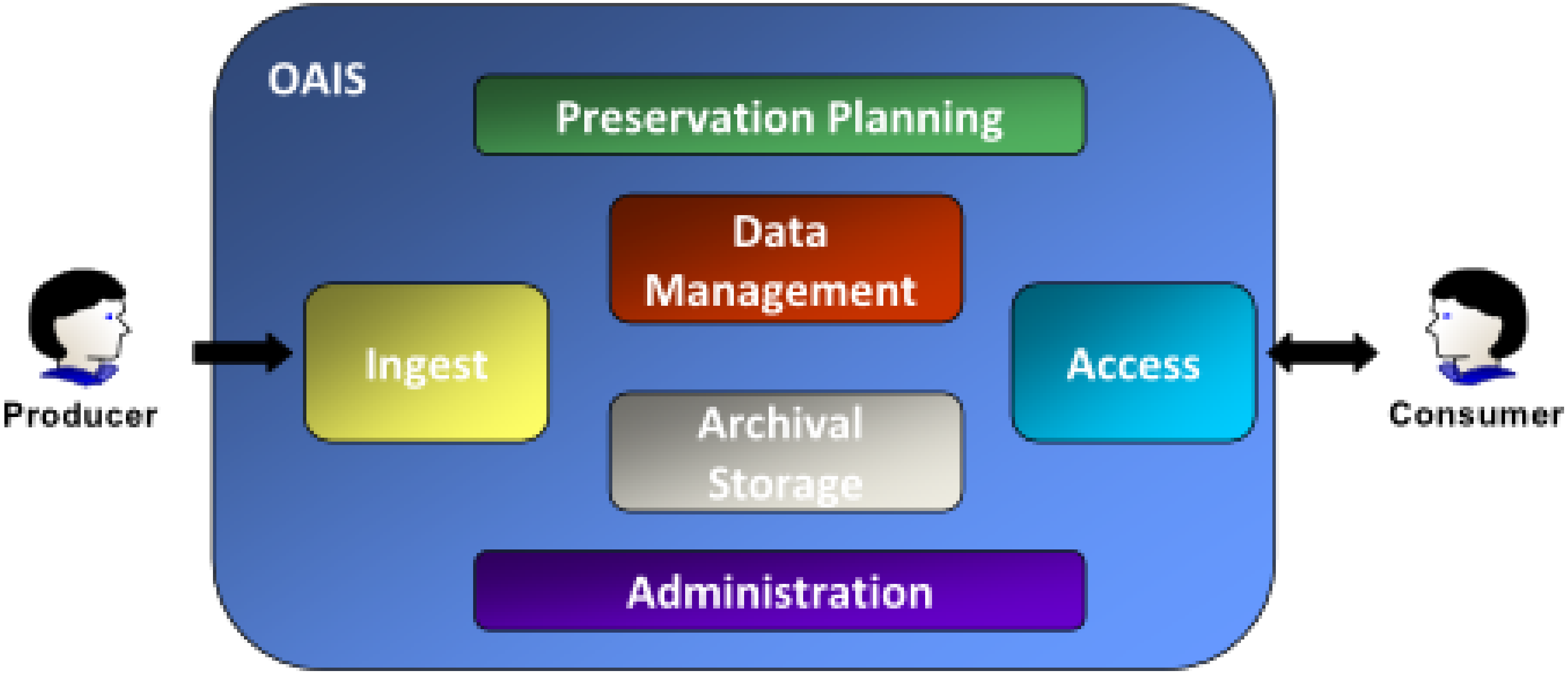
- Storage Requirements
- Storage solutions we tested
- Why we made our choice
- DRI Infrastructure
- DRI bit preservation

DRI:

The Digital Repository Of Ireland (DRI) is an interactive, national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions.

The DRI follows the Open Archival Information System (OAIS) ISO reference model and The Trusted Repository Audit Checklist (TRAC)

OAIS Model:



Source:www.digital-preservation.com

DRI Storage Requirements:

OAIS/TRAC requires the following from storage:

- Minimal conditions for performing long-term preservation of digital assets
- Long Term Preservation of digital assets, even if the OAIS (repository) itself is not permanent or present.

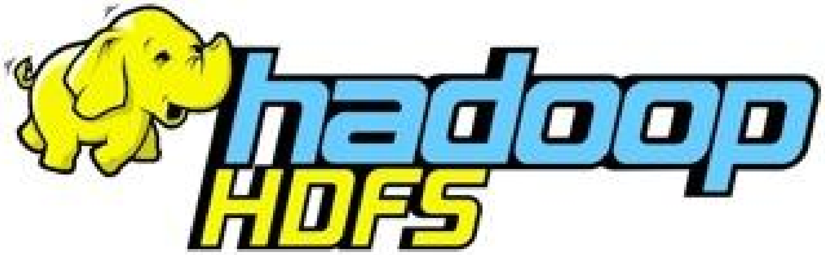
DRI Storage Requirements:

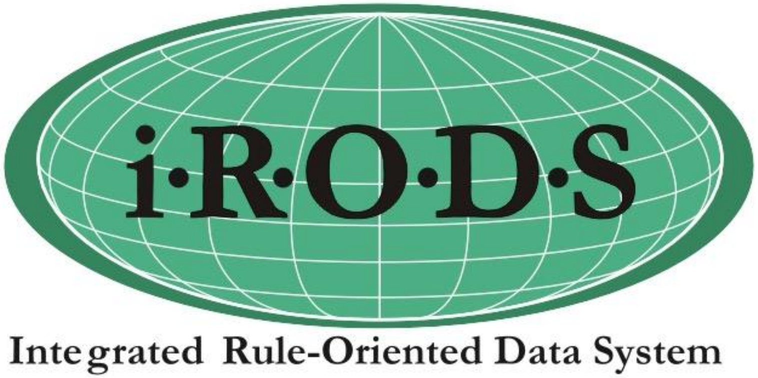
- Open Source/Open Standards
- Independence
- High Availability
- Dynamically Configurable
- Ease of Interoperability (Interfaces, APIs)
- Data Security/Placement (Replication, Erasure coding, Placement, Tiering, Federation)
- Self Contained
- Commodity Hardware

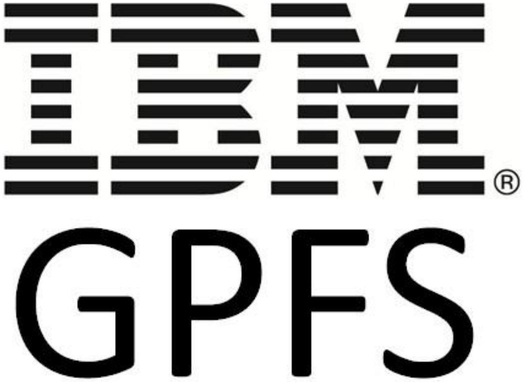
Software Defined Storage vs SAN:

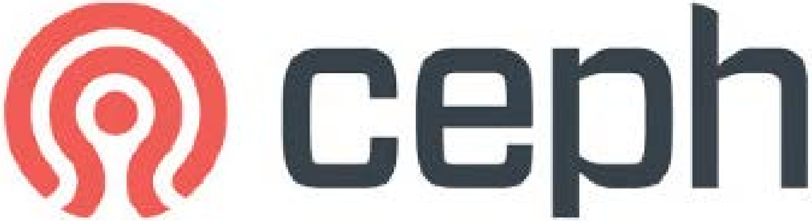
- Lower Cost (Open Source, Commodity hardware)
- No Vendor Lock-In
- Utilise old or existing servers/infrastructure
- Flexibility (IOPS or Space or Bandwidth)
- Incremental hardware upgrade path

Storage Solutions We Tested:

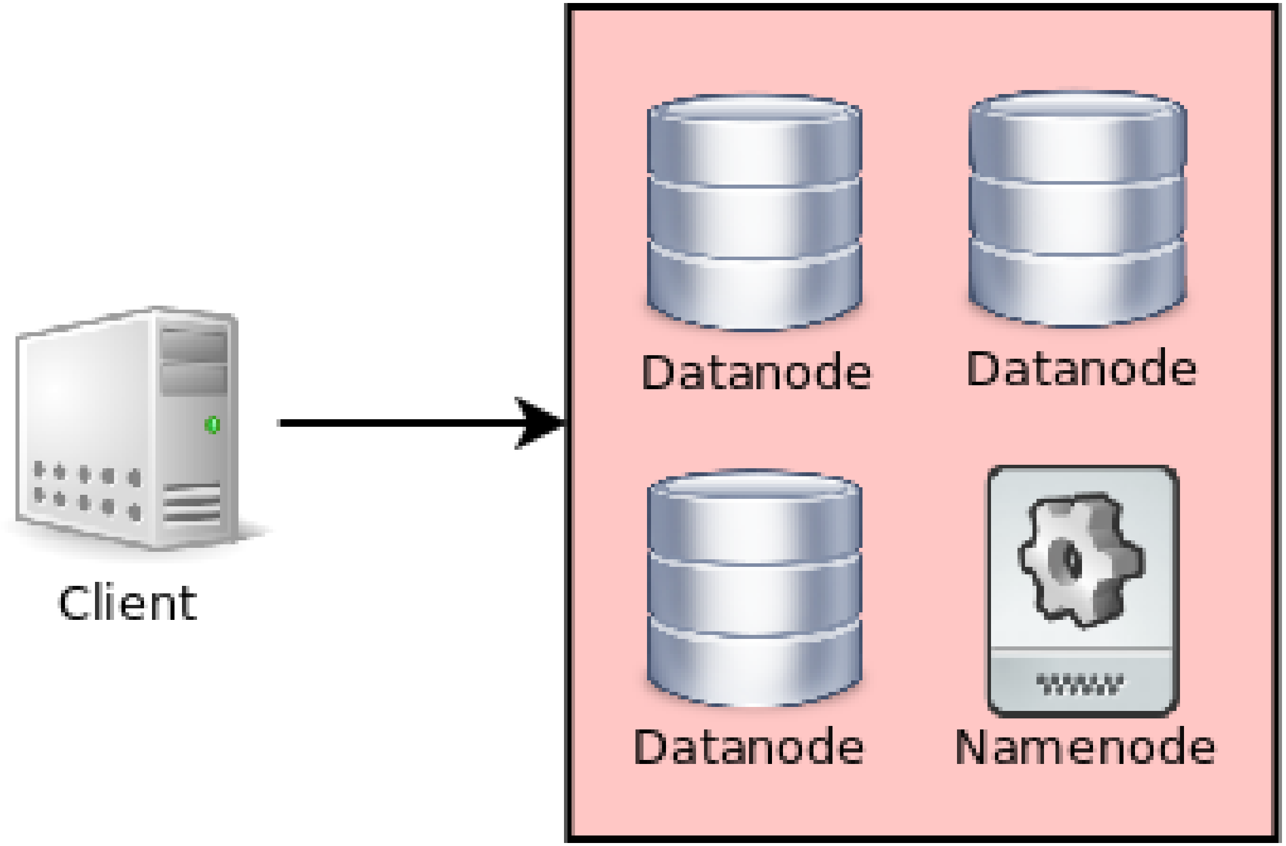








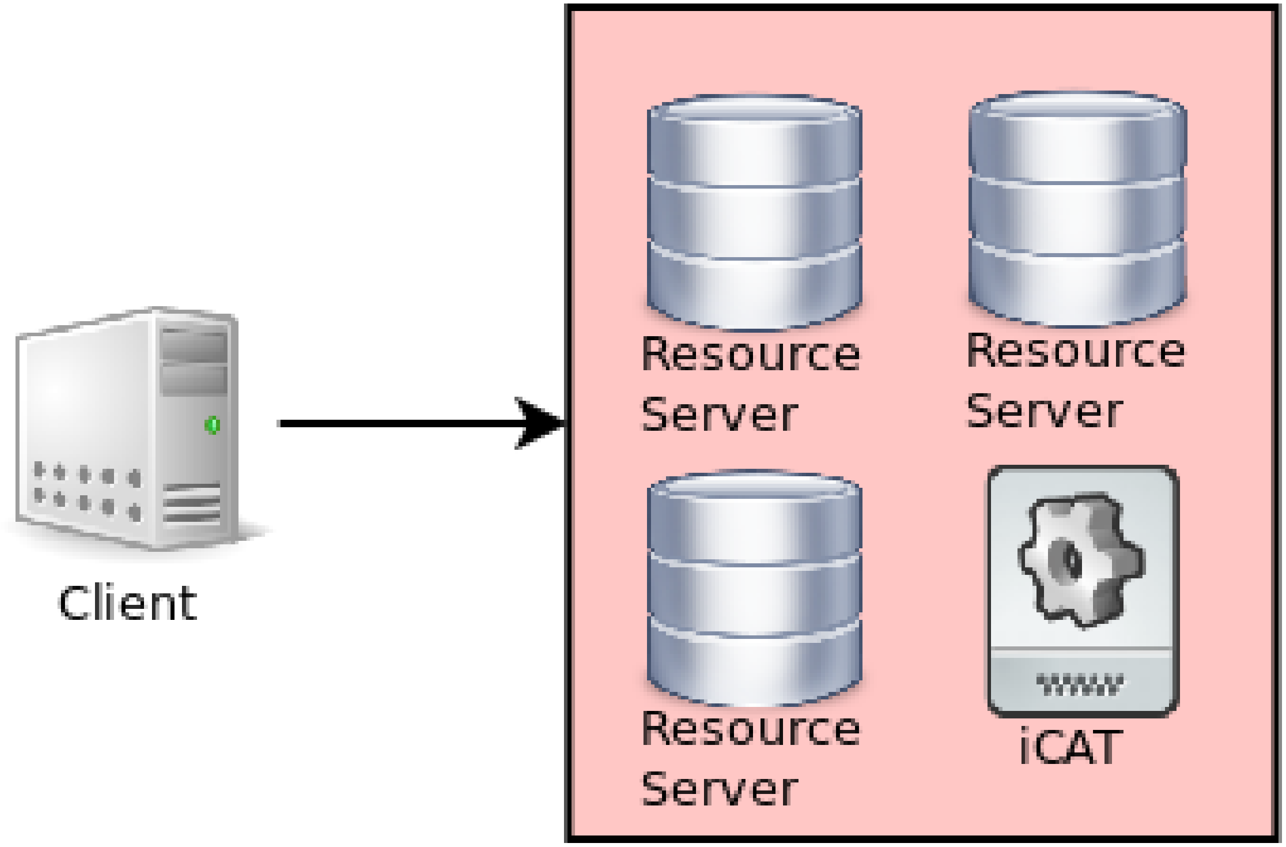
HDFS:



Why we didn't choose HDFS:

- Only provides RESTful API interface. No posix or RBD.
- Performance geared towards large data sets. I/O of many small files is poor.
- Single point of failure and bottleneck at its Namenode.
- Doesn’t provide any federation

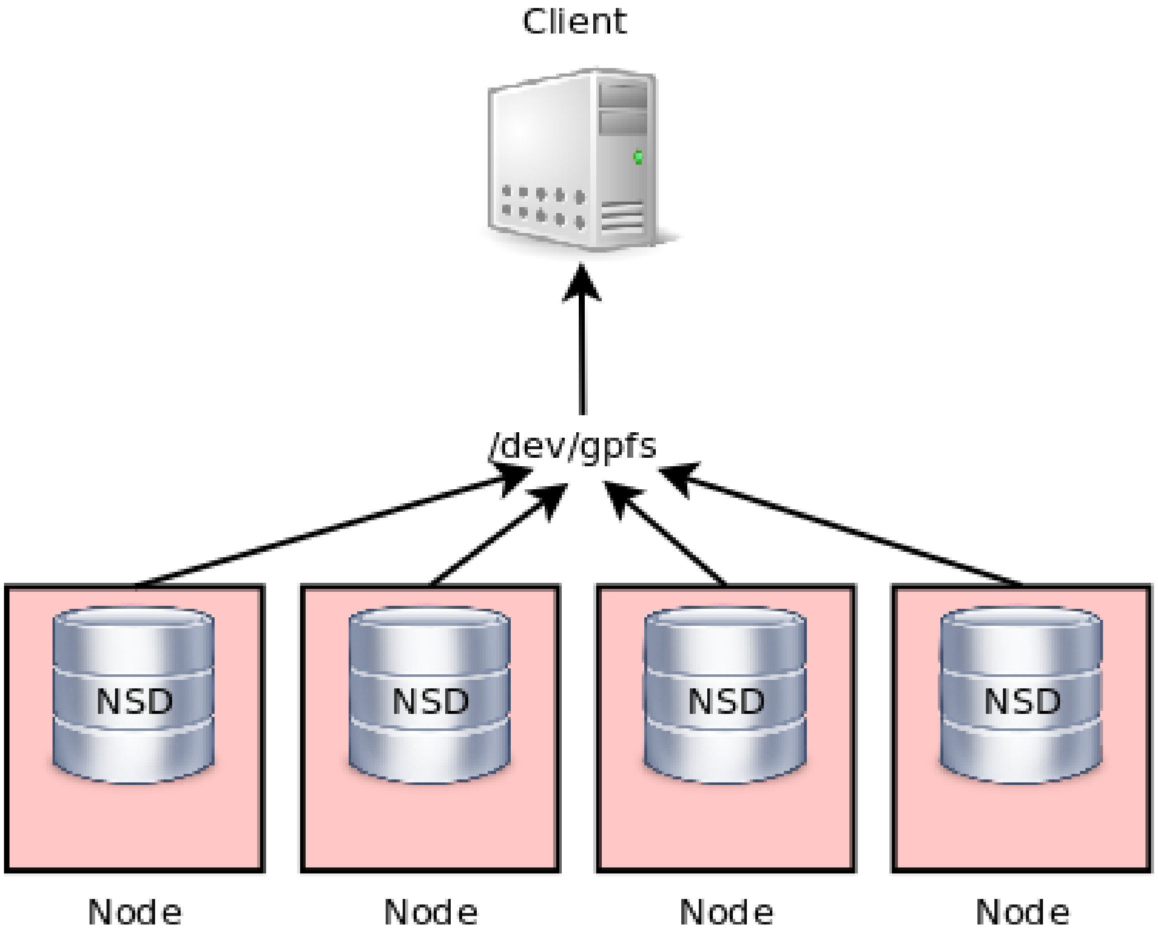
iRODS:


Client
Resource Server
Resource Server
Resource Server
iCAT
www

Why we didn't choose iRODS:

- Default Interfaces limited. No Restful, RBD.
- Single point of failure at its iCAT metadata server
- Overlapping functionality with Fedora Commons

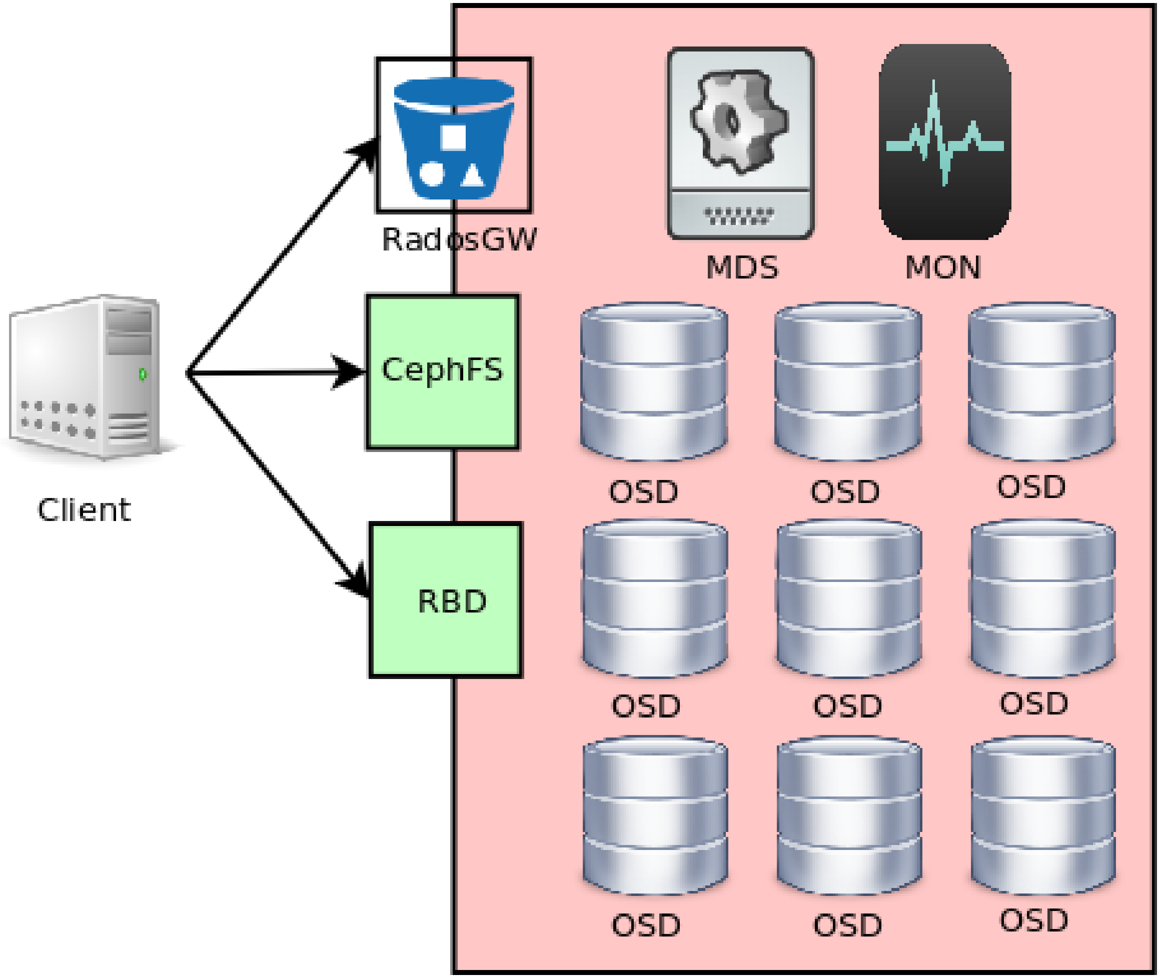
GPFS:



Why we didn't choose GPFS:

- Default Interfaces limited. No Restful, RBD.
- Data Replica limit of 2.
- Closed source

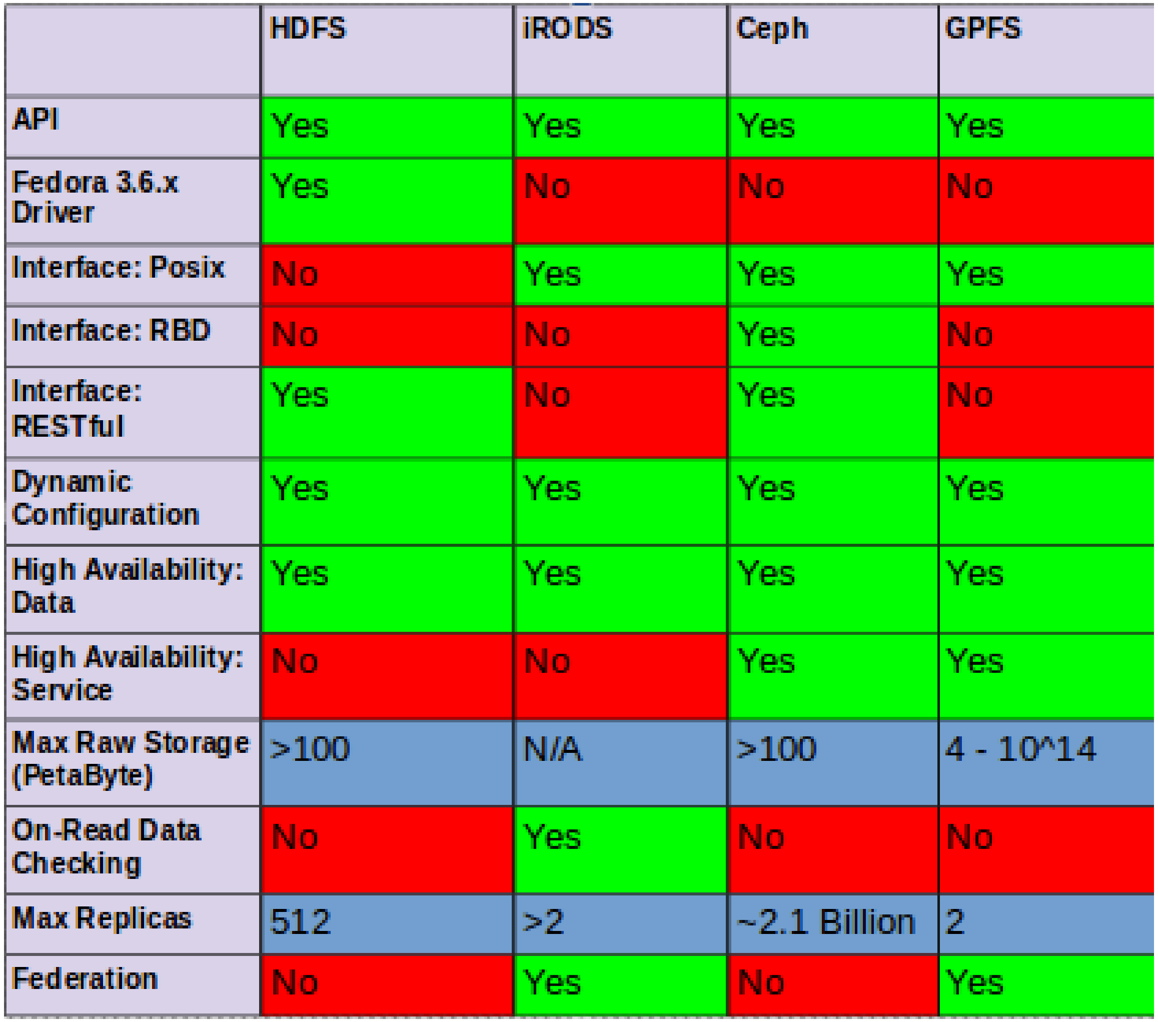
CEPH:


RadosGW
MDS
MON
Client
CephFS
RBD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD
OSD

Why we chose Ceph:

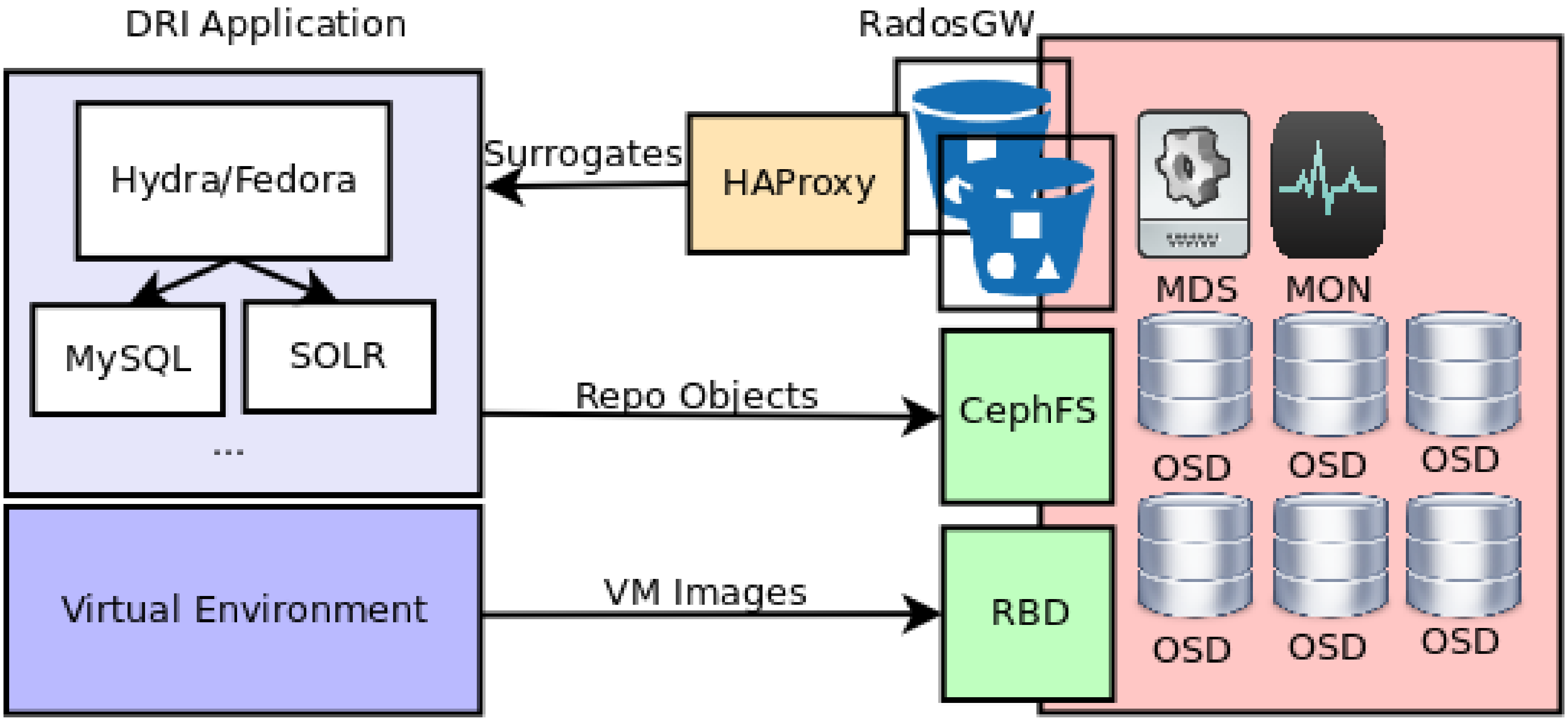
- We like its distributed, clustered architecture
- Provides complete high availability on install
- Scales out horizontally to massive levels
- Data Security/Placement: Distributed, Replicated
- Many interface options
- Rich, documented, multi-level APIs
- Dynamically configurable
- Good Performance for general use (many small file I/O)
- Solid release schedule, new features

Findings:

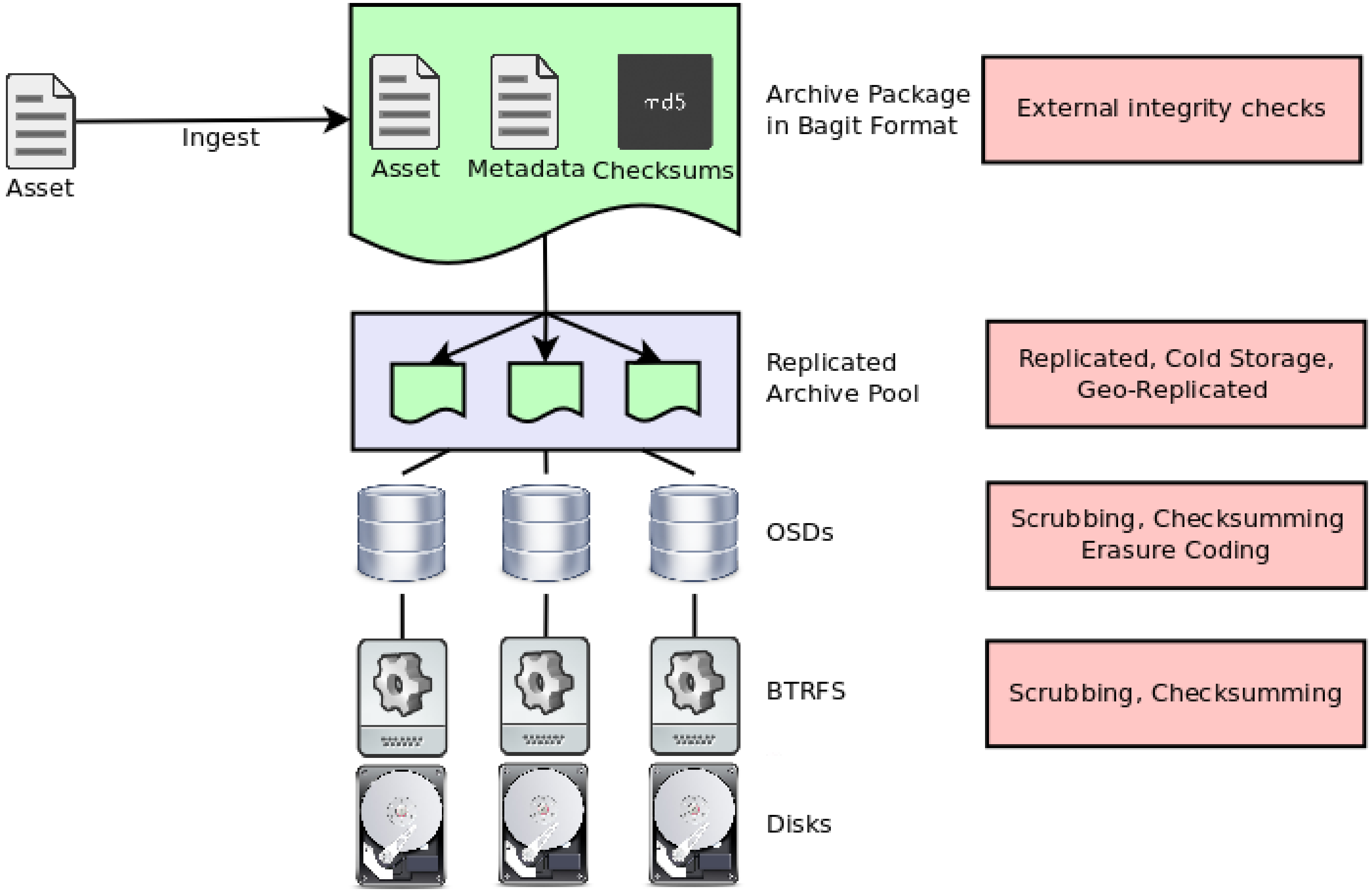


	HDFS	iRODS	Ceph	GPFS
API	Yes	Yes	Yes	Yes
Fedora 3.6.x Driver	Yes	No	No	No
Interface: Posix	No	Yes	Yes	Yes
Interface: RBD	No	No	Yes	No
Interface: RESTful	Yes	No	Yes	No
Dynamic Configuration	Yes	Yes	Yes	Yes
High Availability: Data	Yes	Yes	Yes	Yes
High Availability: Service	No	No	Yes	Yes
Max Raw Storage (PetaByte)	>100	N/A	>100	4 - 10^14
On-Read Data Checking	No	Yes	No	No
Max Replicas	512	>2	~2.1 Billion	2
Federation	No	Yes	No	Yes

DRI Infrastructure



DRI Bit Preservation



Asset

Ingest

Asset

Metadata

md5

Checksums

Archive Package
in Bagit Format

External integrity checks

Replicated
Archive Pool

Replicated, Cold Storage,
Geo-Replicated

OSDs

Scrubbing, Checksumming
Erasure Coding

BTRFS

Scrubbing, Checksumming

Disks

New Ceph Features:

- Asynchronous Geo-Replication
- Erasure Coding
- Tiering

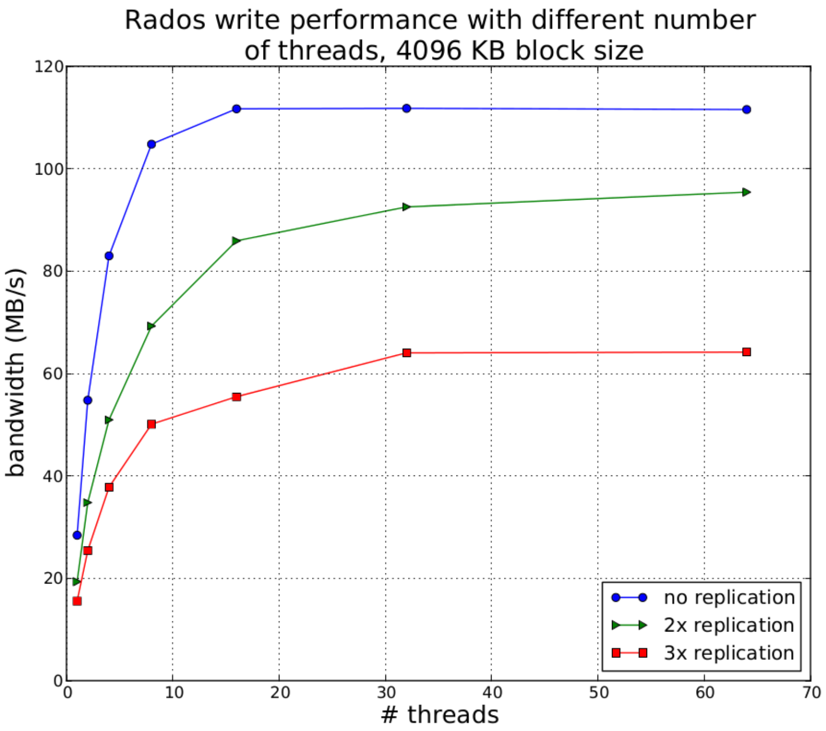
Questions?

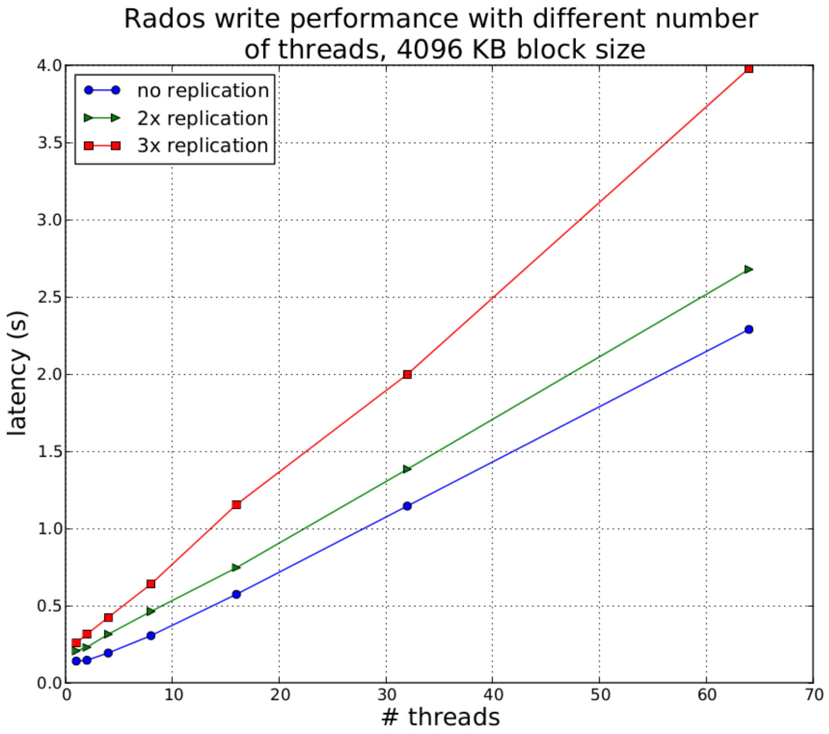
DRI: www.dri.ie
Trinity HPC: www.tchpc.tcd.ie
Trinity College Dublin: www.tcd.ie

Links:

Ceph:	www.ceph.com
HDFS:	hadoop.apache.org
IRODS:	www.irods.org
GPFS:	www.ibm.com/systems/software/gpfs/
Project Hydra:	projecthydra.org
Fedora Commons:	www.fedora-commons.org
Apache SOLR:	lucene.apache.org/solr/
HAProxy:	haproxy.1wt.eu

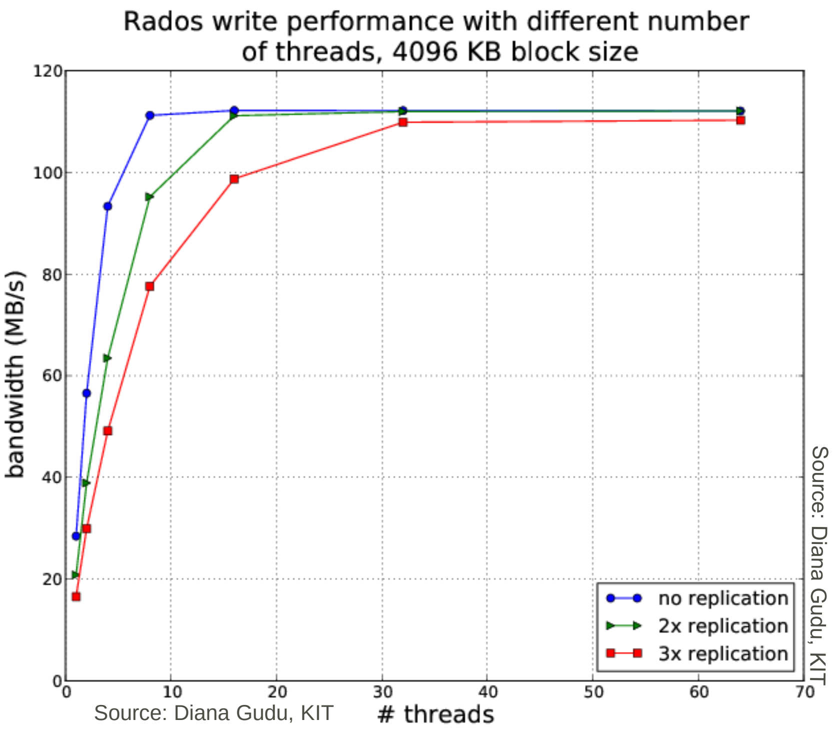
Performance





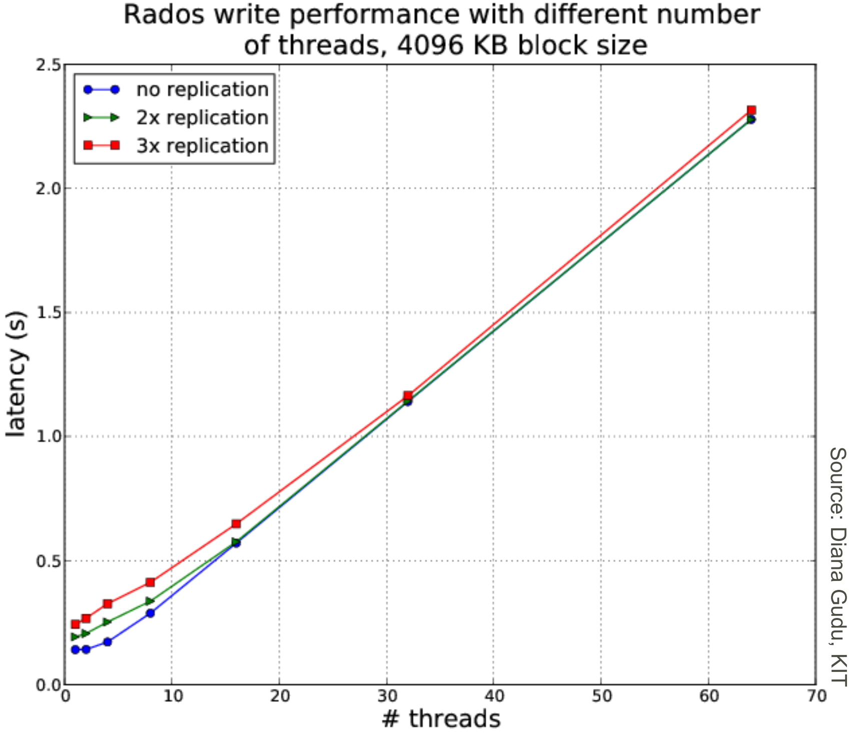
Poor performance with low number of OSDs (6) and replication.

Performance



Source: Diana Gudu, KIT

Source: Diana Gudu, KIT



Source: Diana Gudu, KIT

Adding OSDs (26) improves replicated performance