Integration of XRootD into the cloud infrastructure for ALICE data analysis

Kompaniets M., Shadura O., Yurchenko V., Zarochentsev A.

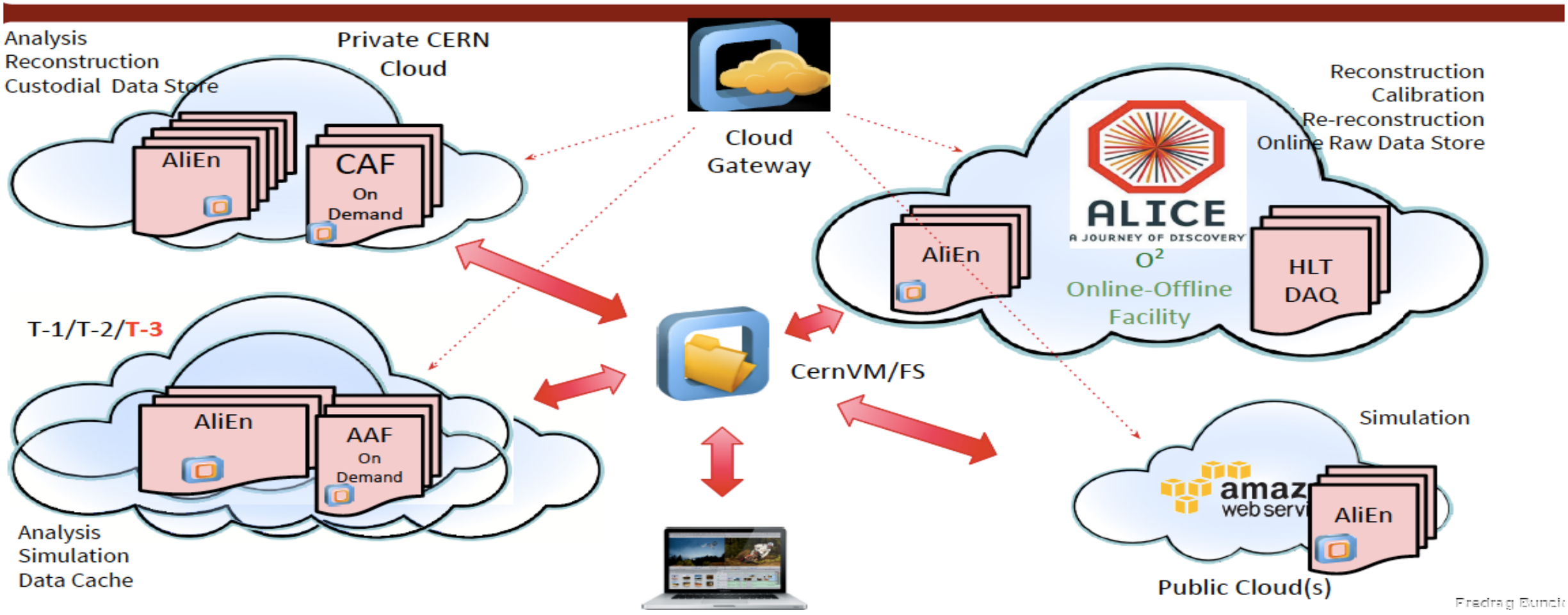
Motivation

- Our goal: to run ALICE data analysis on cloud infrastructures of small communities

- BITP&SPBSU cluster has limited computing capabilities → we need to organise our resources smartly

Part 1. Overall description

Tier3 cloud site as a part of overall scheme



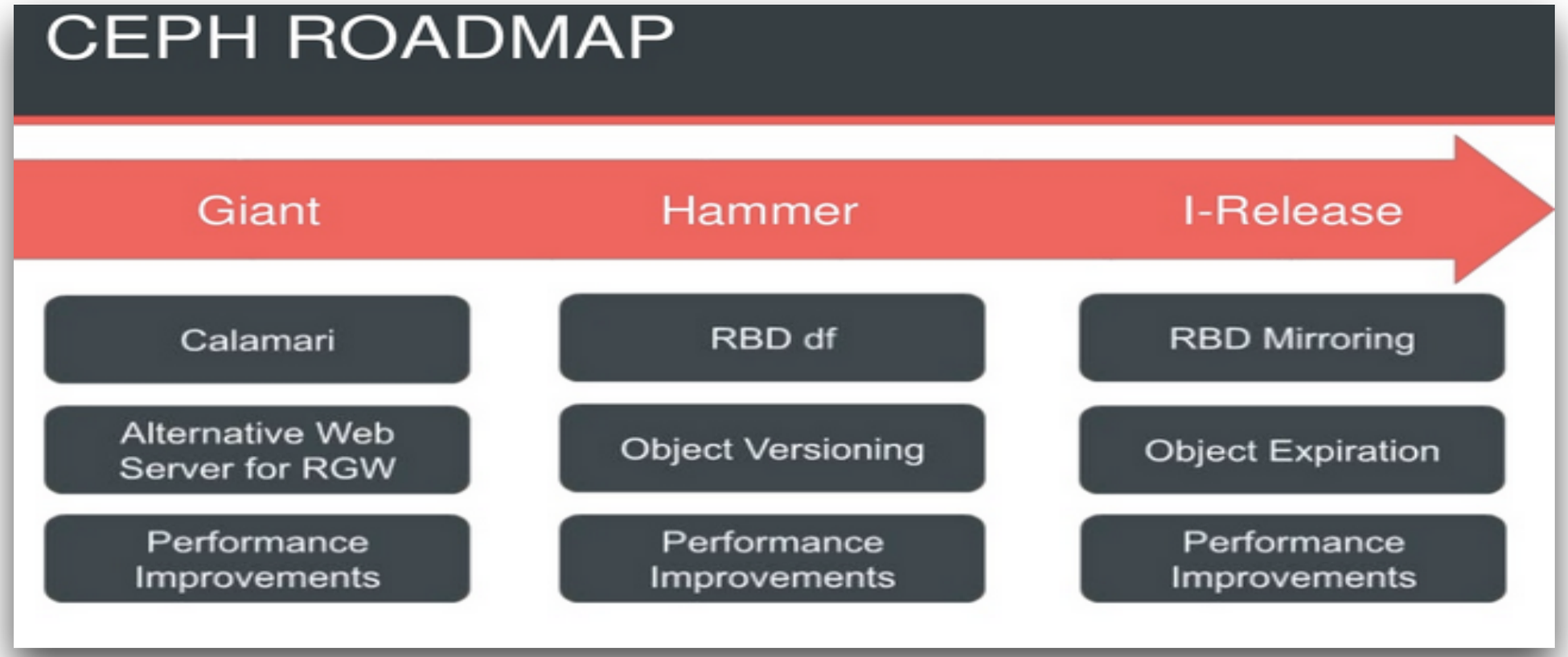
OpenStack + CEPH: advantages & disadvantages

- Advantages of using CEPH in cloud platform:
  - Open source software solution
  - Distributed storage
  - Provides higher availability
  - More scalable in comparison with default LVM backend
- Disadvantages:
  - The learning curve for Ceph is pretty steep :)

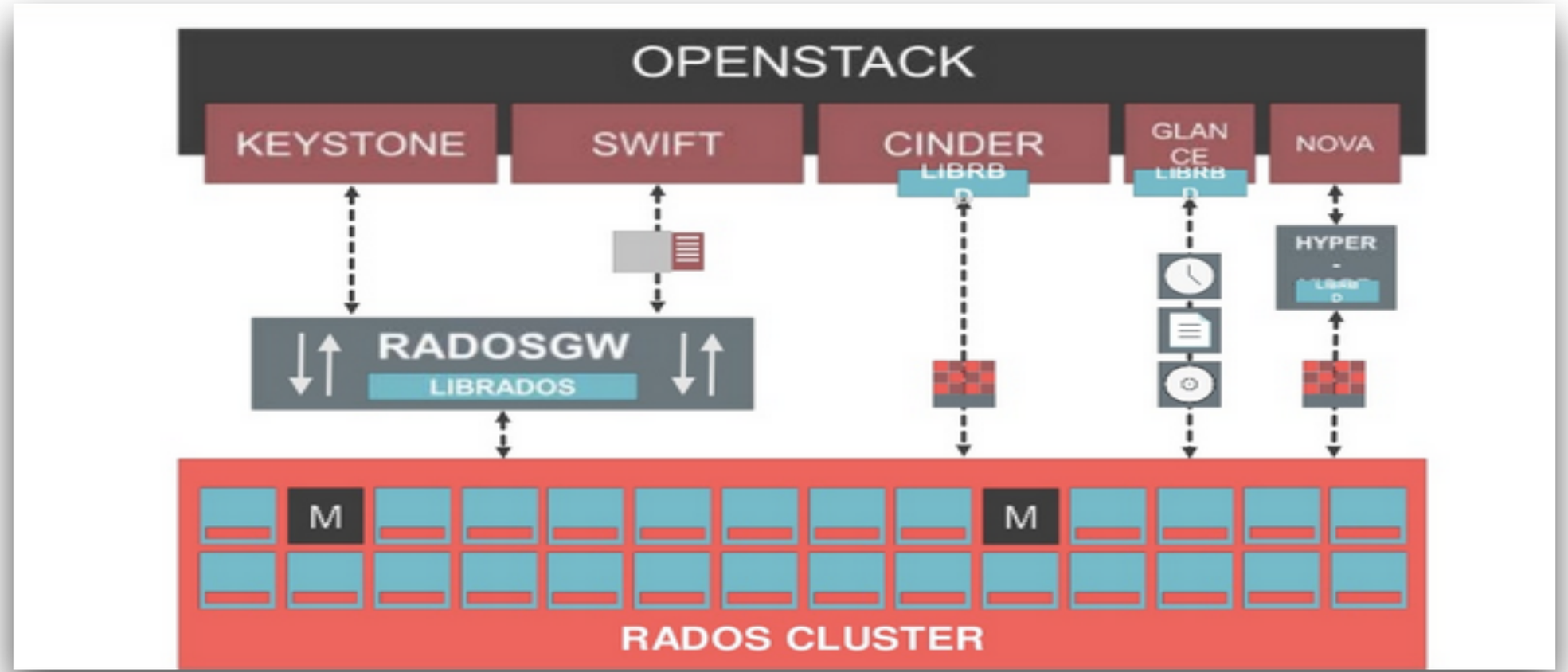
Current status of OpenStack development (in term of features for Ceph backend)

- Ceph as backend was introduced a long time ago…
- Openstack **Icehouse** (RHEL6)
    - Introduced RADOS as a backend for Swift
    - Clone non-raw images in Glance RBD backend
- New release of Openstack **Juno** (RHEL7)
    - Stable live migration for CEPH based images
    - DevStack Ceph: ease the adoption for developers
- Openstack **Kilo** (Next release)
    - Volume migration support with volume retype : move block from Ceph to other backend and the other way around
    - Use RBD snapshoting instead of qemu-img: efficient since we don't need to snapshot, get a flat file and upload it into Ceph
    - Improving Backup system

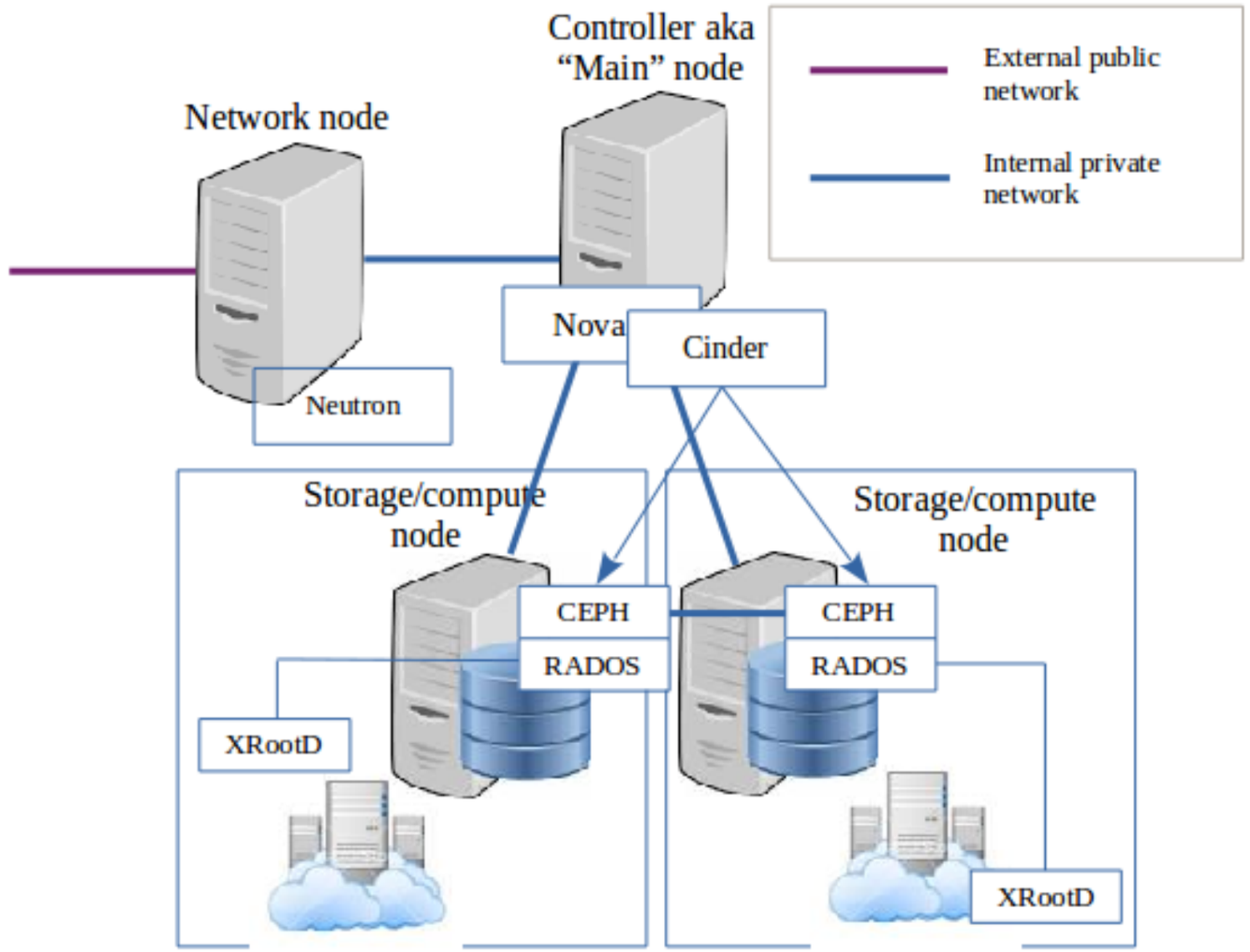
Ceph Roadmap



Interaction between OpenStack and CEPH



Our solution schema



RADOSFS and XRootD interaction

- For organisation interaction of Xrootd and object storage we are using intermediate layer: RadosFS

- RadosFS - A filesystem library based in librados that offers a simple interface for file operations on top of a Ceph Cluster.

- Written by Joaquim Rocha (IT, CERN): https://github.com/joaquimrocha/radosfs

RADOSFS and XRootD interaction

- We are using radosfs-python bindings Python wrapper for RadosFS and fuse-plugin as a basic instrumentation tool for creation specific folder architecture for XRootD

- XRootD is running on Nova hosts to avoid traffic multiplication and each VM connects to XRootD server which launched on the same physical machine

Part 2.Organization of distributed storage in our cloud

ALICE authorisation for XrootD/Ceph [1/6]

- Used software:
  - xrootd4-4.0.4 {from eos-diamond repository}
  - radosfs {from github.com/cern-eos}
  - ceph-0.87.1-0.el6.x86_64
  - xrootd-alicetokenacc
  - tokenauthz

ALICE authorisation for XrootD/Ceph [2/6]

- Creation of 2 pools for xrootd:
  - radososs.datapools /:data0
    radososs.metadatapools /:data0-mtd
- Creation of catalogs (using radosfs-fuse):
  - data0/01
  - …
  - data0/09

  *with permission **777** (!) for simple user

ALICE authorisation for XrootD/Ceph [3/6]

- Now we can copy files:

using aliensh (**alien.v2-19.192**):

{

aliensh:[alice] [3] /alice/cern.ch/user/a/azaroche/ >cp file:///etc/passwd mk16@ALICE::SPbSU::CEPH_TEST

=> Creating replica 1/1 ...

Overriding 'TransactionTimeout' with value 60. Final value: 60

Overriding 'RequestTimeout' with value 30. Final value: 30

Overriding 'ConnectTimeout' with value 10. Final value: 10

Overriding 'FirstConnectMaxCnt' with value 4. Final value: 4

[xrootd] Total 0.00 MB |====================| 100.00 % [inf MB/s]

aliensh:[alice] [4] /alice/cern.ch/user/a/azaroche/ >

}

ALICE authorisation for XrootD/Ceph [4/6]

- BUT not works from versions from **CVMFS**:

{
$ /cvmfs/alice.cern.ch/bin/alien -user azaroche -exec "add mk32 file://`hostname -f`/etc/passwd ALICE::SPbSU::CEPH_test"
.....
Last server error 3016 ('Unable to fchmod /06/28261/1f0f1e9e-8090-11e4-bbce-075cee6a1698/passwd; is a directory')
ALIEN_XRD_SUBCALL_RETURN_VALUE=255
ERROR storing file://alice11.spbu.ru/etc/passwd in ALICE::SPbSU::CEPH_TEST
Dec 10 20:15:32 **error We couldn't upload any copy of the file.**
}

- Problem is in the key option “-P” in the xrdcpamon command - RadosFS can not submit “chmod”.

ALICE authorisation for XrootD/Ceph [5/6]

- Problem can solved by the **old alien version**:

{[alicesgm001@alice11 ~]$ /cvmfs/alice.cern.ch/x86_64-2.6-gnu-4.1.2/Packages/AliEn/v2-19-218/bin/alien -user azaroche -exec "add $file file://`hostname -f`/etc/passwd ALICE::SPbSU::CEPH_test"

……..

Mar 19 19:48:31 notice We will upload the file file://alice11.spbu.ru/etc/passwd to ALICE::SPbSU::CEPH_TEST

Mar 19 19:48:36 info   File saved successfully in SE: ALICE::SPbSU::CEPH_TEST

Mar 19 17:48:36 info   Inserting the lfn

Mar 19 17:48:36 info   File(s) inserted

Mar 19 17:48:36 info   File /alice/cern.ch/user/a/azaroche/file-alice11.spbu.ru-1426783705 inserted in the catalog

Mar 19 17:48:36 notice **Authorize: EXCELLENT! All of the 1 PFNs where correctly registered.**

Mar 19 19:48:37 notice OK. The file /alice/cern.ch/user/a/azaroche/file-alice11.spbu.ru-1426783705 was added to 1 SEs as specified. Superb!

[alicesgm001@alice11 ~]$ alien -user azaroche -exec "whereis file-alice11.spbu.ru-1426783705"

Mar 19 17:49:23 info   **The file azaroche/file-alice11.spbu.ru-1426783705 is in**

SE => ALICE::SPbSU::CEPH_TEST pfn =>root://alice05.spbu.ru:1094//04/28804/ c629df9c-ce57-11e4-8fee-d3c60f5805e8

}

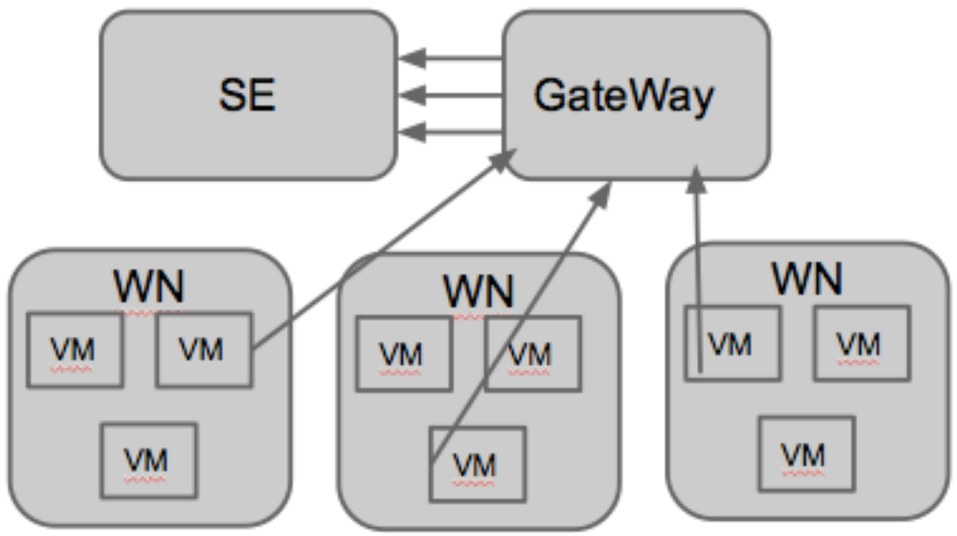
ALICE authorisation for XrootD/Ceph [6/6]

- Two existing problems:
  - user mapping
  - “chmod” in RadosFS
- First problem can be can solved through XrootD settings
- Second can be solved through patching RadosFS, or (may be) just to use CEPHFS as XrootD backend?

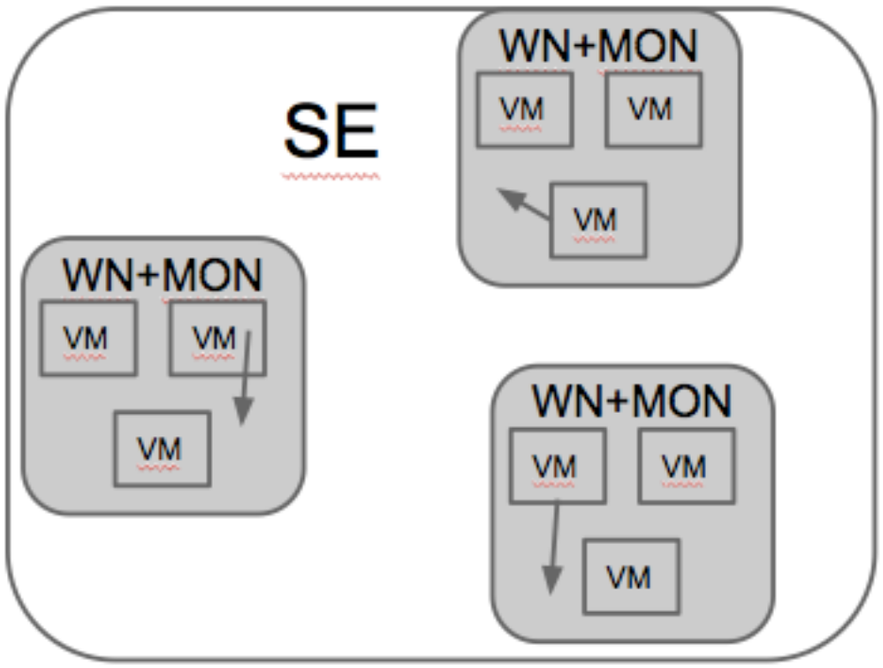
XrootD/Ceph on cloud [1/3]

- On distributed storage we can use any point access:

[normal storage]



[distributed storage]



XrootD/Ceph on cloud [2/3]

- Problem: we need access from VM to WN
  - Creation of second private network for communication with SE access point from VM to WN
  - Launching set scripts on WN to create network interface for second private network
  - Launching set of scripts on VMs for organisation special query address of interface on WN

XrootD/Ceph on cloud [3/3]

1. VM take address of WN:

[zar@host-10-0-0-44 ~]$ ifconfig eth1 | grep HWaddr| awk '{print $5}'

FA:16:3E:EE:D3:D5

[zar@host-10-0-0-44 ~]$ curl -i http://195.19.226.152:8080/?mac=FA:16:3E:EE:D3:D5

host=['alice14'][zar@hostail -1 /etc/hosts

10.30.30.224 alice22.spbu.ru alice22

2. When alilce22 - address of gateway for xrootd server:

[zar@host-10-0-0-44 ~]$ /cvmfs/alice.cern.ch/x86_64-2.6-gnu-4.1.2/Packages/AliEn/v2-19-218/bin/alien -domain spbu.ru -user azaroche -exec "add $file /etc/passwd ALICE::SPbSU::CEPH_test"

……..……..……..

Mar 20 03:26:54 info File /alice/cern.ch/user/a/azaroche/zar17344 inserted in the catalog

Mar 20 03:26:54 notice Authorize: EXCELLENT! All of the 1 PFNs where correctly registered.

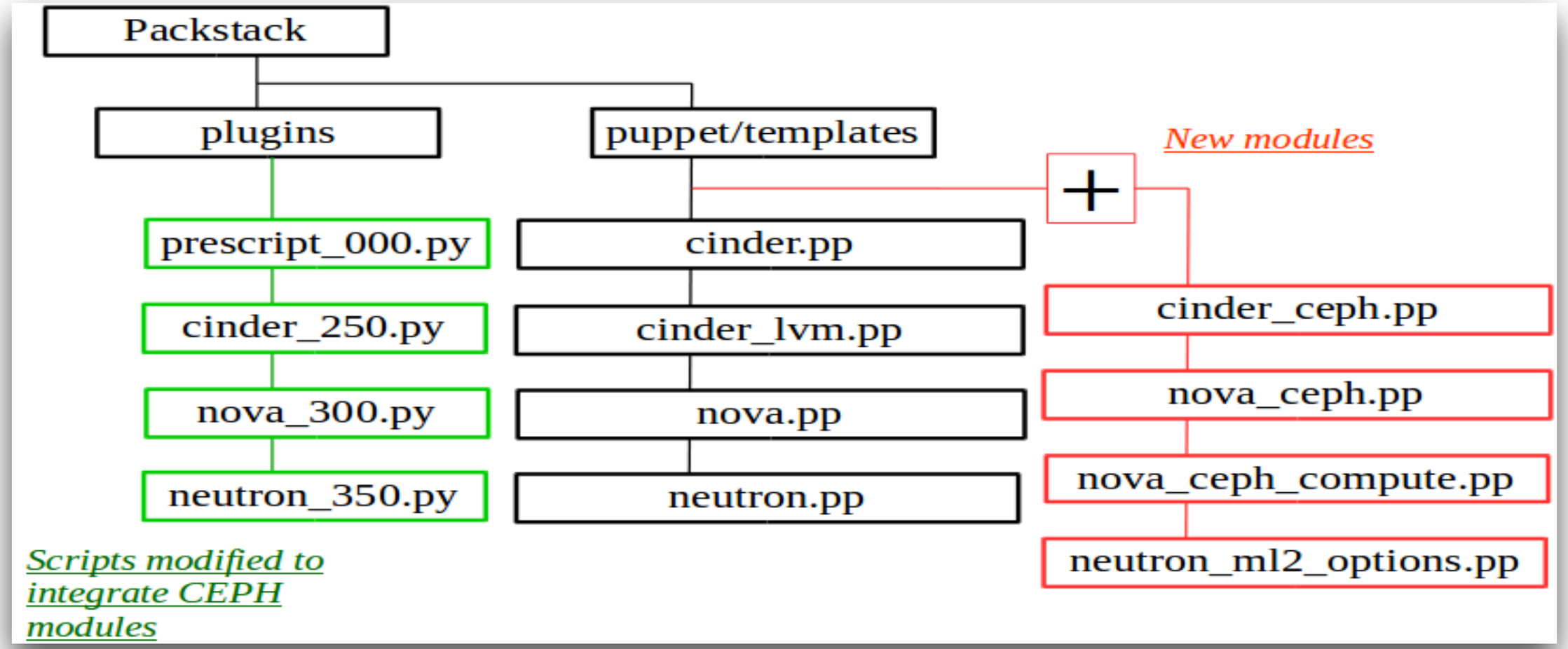
Mar 20 05:26:53 notice OK. The file /alice/cern.ch/user/a/azaroche/zar17344 was added to 1 SEs as specified. Superb!

Part 3. Additional instruments

PackStack as a choice for automatisation tool

- Easy and small tool for fast deployment of OpenStack infrastructure
- Support for RHEL6/RHEL7 & all possible versions of Openstack
- Based on Puppet solution
  - Easy & scalable management
- Modular structure
  - Allows to add features and integrate CEPH and XRootD into OpenStack installation

CEPH module for Packstack



Deployment on BITP & SPSU

- Tested on RHEL6 -> works
- Tested on RHEL7
  - Required small changes in syntax (new branch under testing)
- Usage only with Neutron
  - Tested with GRE/Vxlan networks

Issues

- Usage of elastic & CernVMOnline as a service for organisation elastic batch system and contextualization process
  - EC2 API is not working with Nova+Neutron API

```
Python 2.6.6 (r266:84292, Jan 22 2014, 05:06:49)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-3)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import boto
>>> AWS_ACCESS_KEY_ID = 'xxxx'
>>> AWS_SECRET_ACCESS_KEY = 'xxxx'
>>> conn = boto.connect_ec2_endpoint('http://194.44.37.120:8773/services/Cloud',AWS_ACCESS_KEY_ID, AWS_SECRET_ACCESS_KEY)
>>> conn.get_all_instances()
[Reservation:r-f4vv41d1, Reservation:r-bwdfs9rg]
>>> img = conn.get_image('ami-00000003')
>>> img.name
u'mcernvm'
>>> conn.run_instances(image_id='ami-00000003', instance_type='m1.tiny')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/python2.6/site-packages/boto/ec2/connection.py", line 974, in run_instances
    verb='POST')
  File "/usr/lib/python2.6/site-packages/boto/connection.py", line 1204, in get_object
    raise self.ResponseError(response.status, response.reason, body)
boto.exception.EC2ResponseError: EC2ResponseError: 400 Bad Request
<?xml version="1.0"?>
<Response><Errors><Error><Code>NetworkAmbiguous</Code><Message>Multiple possible networks found, use a Network ID to be more specific.</Message></Error></Errors><RequestID>req-88e32b03-
c24f-4f9c-a42a-47ecafc676eb</RequestID></Response>
```

Solution:new ec2-api

- https://github.com/stackforge/ec2-api
- Checked with AWS cli
- Extension of existing ec2 interface in Openstack
- Initialised on different port 8788 (could be changed)
    - Work with Neutron through special nova_metadata_port = 8789

```
| +-------------+----------------------------------+
| | Field       | Value                            |
| +-------------+----------------------------------+
| | adminurl    | http://194.44.37.129:8788/       |
| | id          | 178338a2318f43b88d78091a0a738425 |
| | internalurl | http://194.44.37.129:8788/       |
| | publicurl   | http://194.44.37.129:8788/       |
| | region      | RegionOne                        |
| | service_id  | b6250deb95ac4c0daf053eb2f744f2ec |
| | service_name | ec2                             |
| | service_type | ec2                             |
| +-------------+----------------------------------+
```

Openstack network management through EC2 API

[stack@horst-9 devstack]$ aws --endpoint-url http://194.44.37.129:8788/services/Cloud **ec2 create-vpc --cidr-block 10.5.5.0/24**
VPC10.5.5.0/24 default False available vpc-60dc43d4
[stack@horst-9 devstack]$ aws --endpoint-url http://194.44.37.129:8788/services/Cloud **ec2 create-subnet --vpc-id vpc-60dc43d4 --cidr-block 10.5.5.0/24**
SUBNET252 10.5.5.0/24 False False available subnet-c6e6ad83 vpc-60dc43d4
[stack@horst-9 devstack]$ **neutron net-list**
+--------------------------------------+--+
| id | name | subnets |
+--------------------------------------+--+
| 42fbec40-b29f-4371-b980-ae1dedfe376a | **subnet-c6e6ad83** | 44bfb5ae-eaaf-4c55-8f70-8696d6db0787 10.5.5.0/24 |
+--------------------------------------+--+

Start of instance {example}

[stack@horst-9 devstack]$ aws --endpoint-url http://194.44.37.129:8788/services/Cloud **ec2 run-instances --image ami-6098b252 --subnet-id subnet-c6e6ad83 --instance-type m1.small**
d12600c684964f59943869b7a67a7736 r-rurilshe
INSTANCES 0 ami-6098b252 i-f4337886 m1.small None 2015-03-11T12:42:45Z r-rurilshe-0 10.5.5.2
None /dev/sda1 instance-store subnet-c6e6ad83 vpc-60dc43d4
NETWORKINTERFACES None fa:16:3e:c7:3d:8d eni-2f9cca72 d12600c684964f59943869b7a67a7736 10.5.5.2 True
in-use subnet-c6e6ad83 vpc-60dc43d4
ATTACHMENT 2015-03-11T12:42:45.504973Z eni-attach-2f9cca72 True 0 attached
GROUPS sg-13eb6794 default
PRIVATEIPADDRESSES True 10.5.5.2
PLACEMENT nova
SECURITYGROUPS sg-13eb6794 default
STATE 0 pending
[stack@horst-9 devstack]$

[stack@horst-9 devstack]$ aws --endpoint-url http://194.44.37.129:8788/services/Cloud **ec2 describe-instances**
RESERVATIONS d12600c684964f59943869b7a67a7736 r-rurilshe
INSTANCES 0 **ami-6098b252** i-f4337886 m1.small None 2015-03-11T12:42:45Z r-rurilshe-0 10.5.5.2
None /dev/sda1 instance-store **subnet-c6e6ad83 vpc-60dc43d4**
NETWORKINTERFACES None fa:16:3e:c7:3d:8d eni-2f9cca72 d12600c684964f59943869b7a67a7736 10.5.5.2 True
in-use subnet-c6e6ad83 vpc-60dc43d4
ATTACHMENT 2015-03-11T12:42:45.504973Z eni-attach-2f9cca72 True 0 attached
GROUPS sg-13eb6794 default
PRIVATEIPADDRESSES True 10.5.5.2
PLACEMENT nova
SECURITYGROUPS sg-13eb6794 default
STATE 0 pending

Next steps:

- Openstack with Ceph and new ec2-api (SL7 only)
- Tiny changes in code of elastic or CernVM Online —add subnet=xxx feature
- Voila!

Part 5.Conclusions

Results

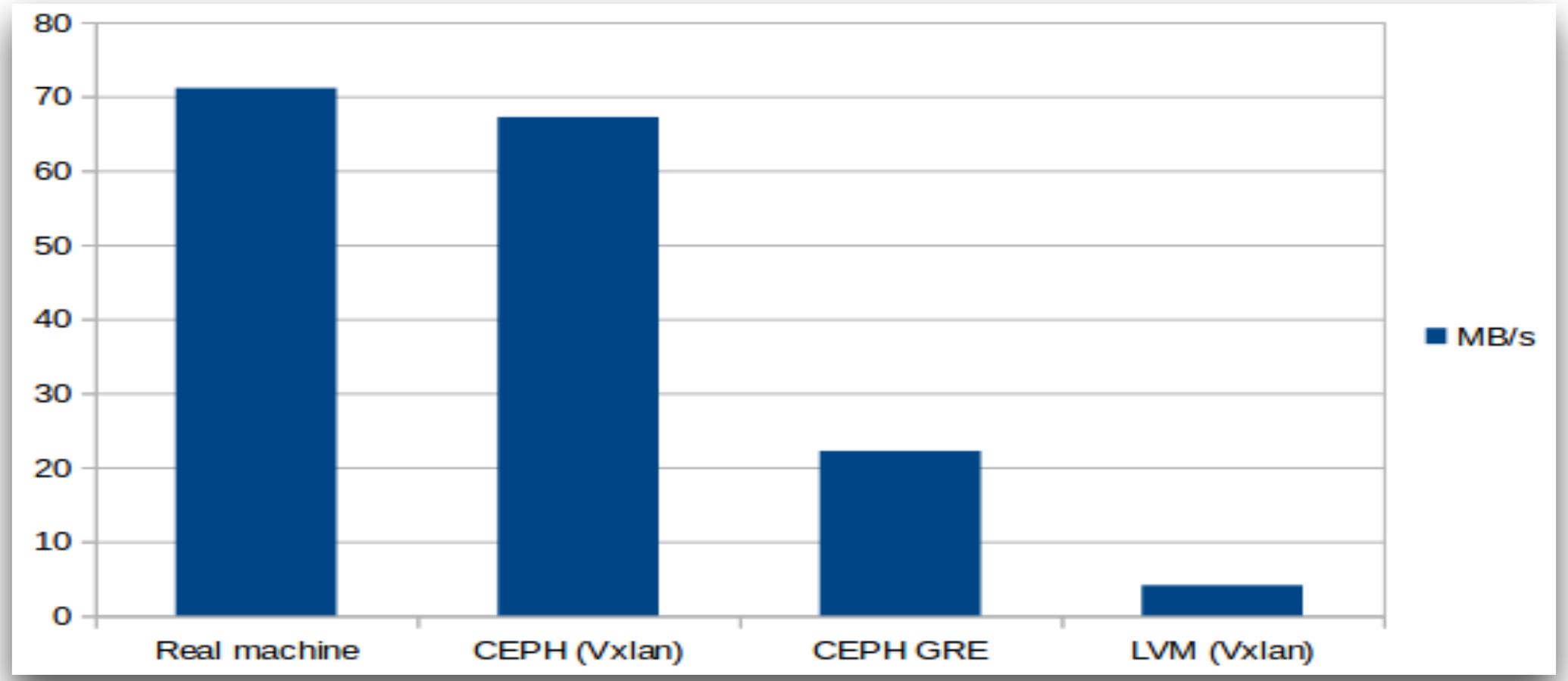
- Automatic installation of Openstack cloud with support of CEPH & Xrootd: DONE
- XRootD+Ceph testing: DONE
- aliEn [specific version] + XRootD + Ceph: DONE
- Elastiq+Openstack [with Neutron]: still not successful for our solution

Plans for future

- Solve small problems based on interaction Openstack and elastic batch system (elastiq)
  - EC2-API usage in Openstack Juno with Packstack & Ceph
- Tests on production cluster {more nodes}
- Deep benchmarking to check performance
- Tuning & optimisation of system configuration

Part 5.Backup

Benchmarking – simple write on disk



Bonnie++ test suite results