Unleash the Power of Ceph Across the Data Center

TUT18972: FC/iSCSI for Ceph

Ettore Simone
Alchemy LAB
ettore.simone@alchemy.solutions

SUSE
We adapt. You succeed.

Agenda

- Introduction
- The Bridge
- The Architecture
- Benchmarks
- A Light Hands-On
- Some Optimizations
- Software and Hardware
- Q&A

Introduction

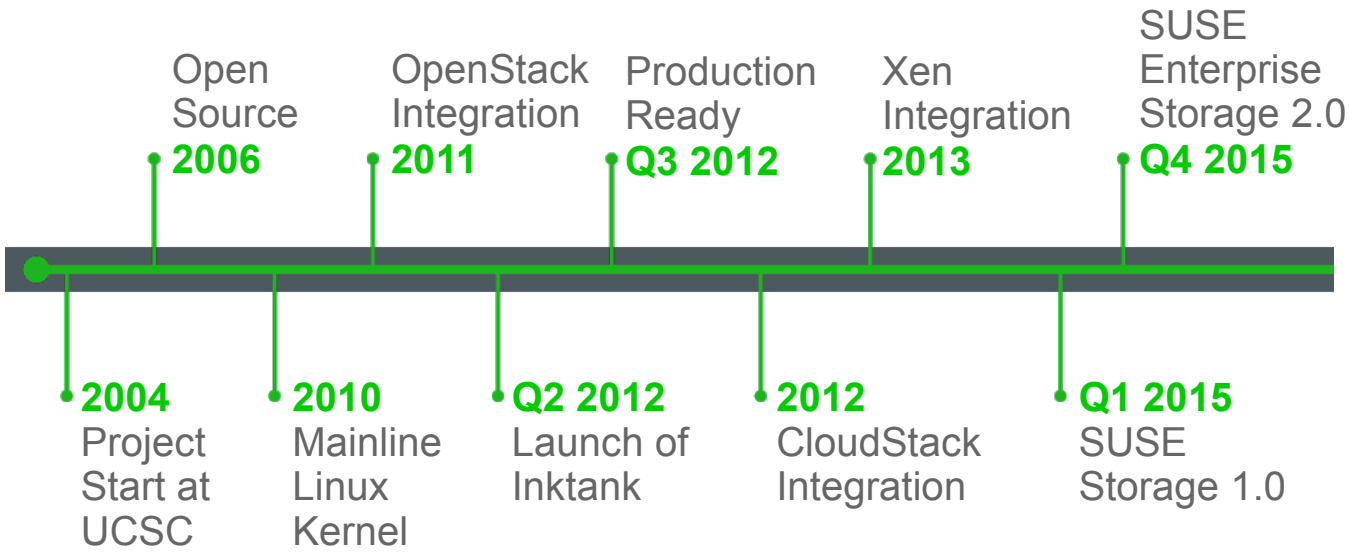
About Ceph

"Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability." (http://ceph.com/)

FUT19336 - SUSE Enterprise Storage Overview and Roadmap

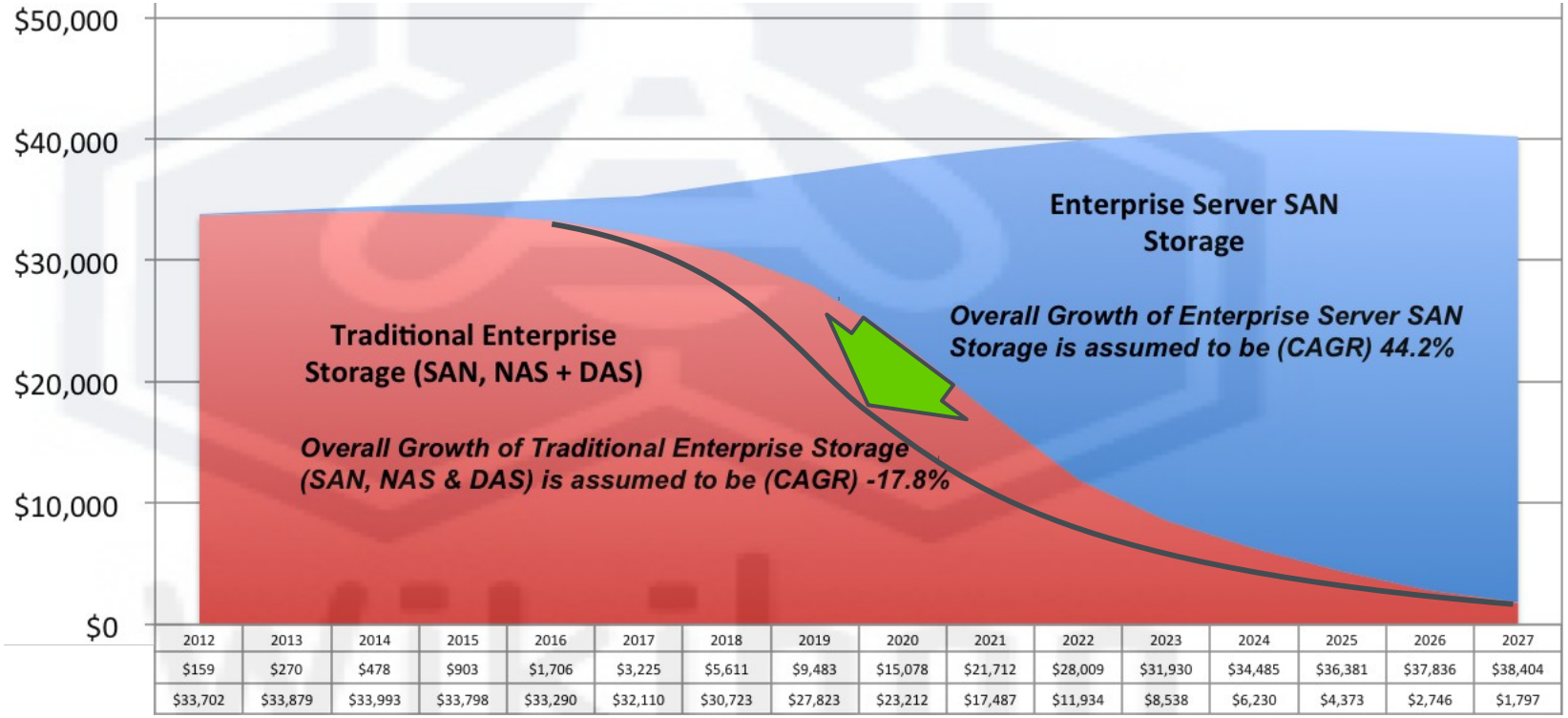
TUT20074 - SUSE Enterprise Storage Design and Performance

Ceph timeline



Some facts

Common data centers storage solutions are built mainly on top of Fibre Channel (yes, and NAS too).



2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027
$159	$270	$478	$903	$1,706	$3,225	$5,611	$9,483	$15,078	$21,712	$28,009	$31,930	$34,485	$36,381	$37,836	$38,404
$33,702	$33,879	$33,993	$33,798	$33,290	$32,110	$30,723	$27,823	$23,212	$17,487	$11,934	$8,538	$6,230	$4,373	$2,746	$1,797

Source: Wikibon Server SAN Research Project 2014

Is the storage mindset changing?

New/Cloud

- Micro-services Composed Applications
- NoSQL and Distributed Database (lazy commit, replication)
- Object and Distributed Storage

SCALE-OUT

Classic

- Traditional Application → Relational DB → Traditional Storage
- Transactional Process → Commit on DB → Commit on Disk

SCALE-UP

Is the storage mindset changing? No!

New/Cloud

- Micro-services Composed Applications
- NoSQL and Distributed Database (lazy commit, replication)
- Object and Distributed Storage

Natural playground of Ceph

Classic

- Traditional Application → Relational DB → Traditional Storage
- Transactional Process → Commit on DB → Commit on Disk

Where we want to introduce Ceph!

Is the new kid on the block so noisy?

Ceph is cool but I cannot rearchitect my storage!

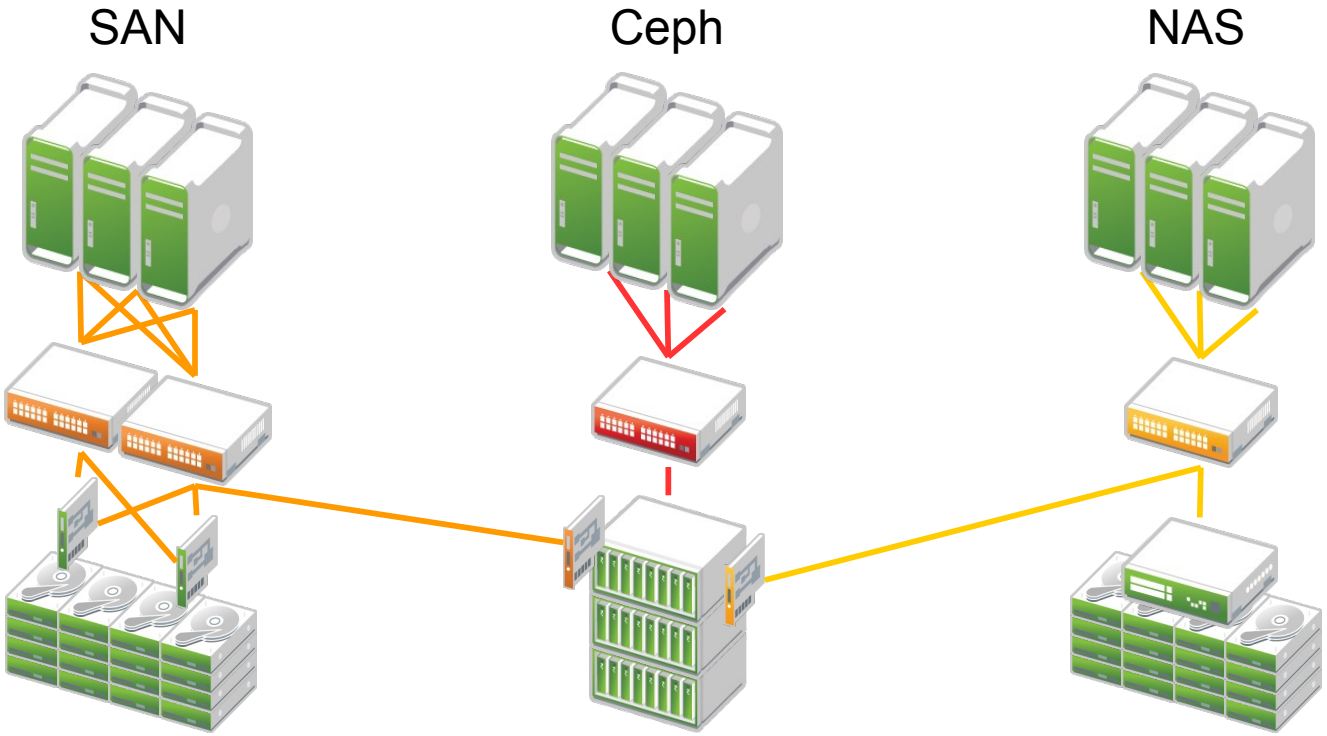
And what about my shiny big disk arrays?

I have already N protocols, why another one?

<Add your own fear here>

Our goal

How to achieve a non disruptive introduction of Ceph into a traditional storage infrastructure?



How to let happily coexist Ceph in your datacenter with the existing neighborhood

(traditional workloads, legacy servers, FC switches etc...)

The Bridge

FC/iSCSI gateway

iSCSI

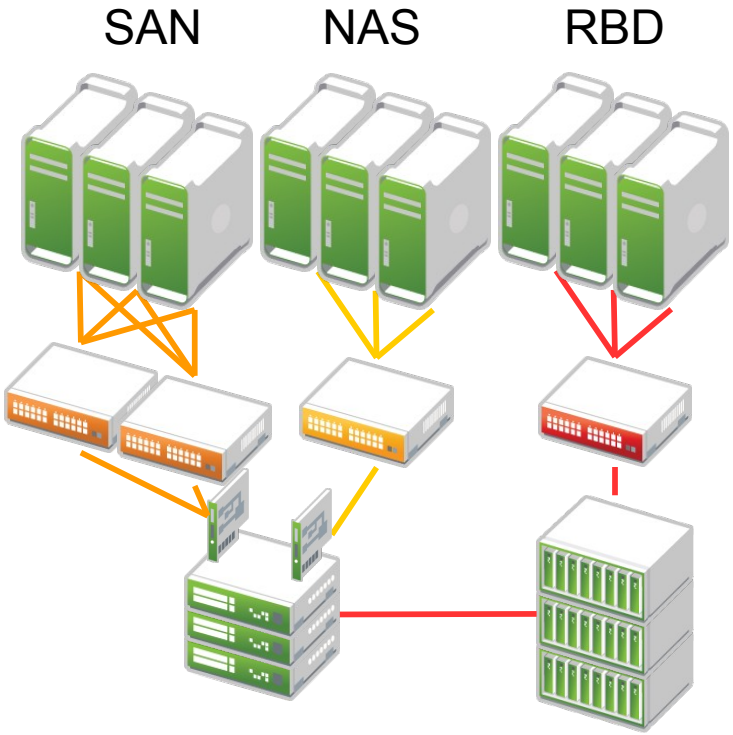
- Out-of-the-box feature of SES 2.0
- TUT16512 - Ceph RBD Devices and iSCSI

Fiber Channel

- That's the point we will focus today

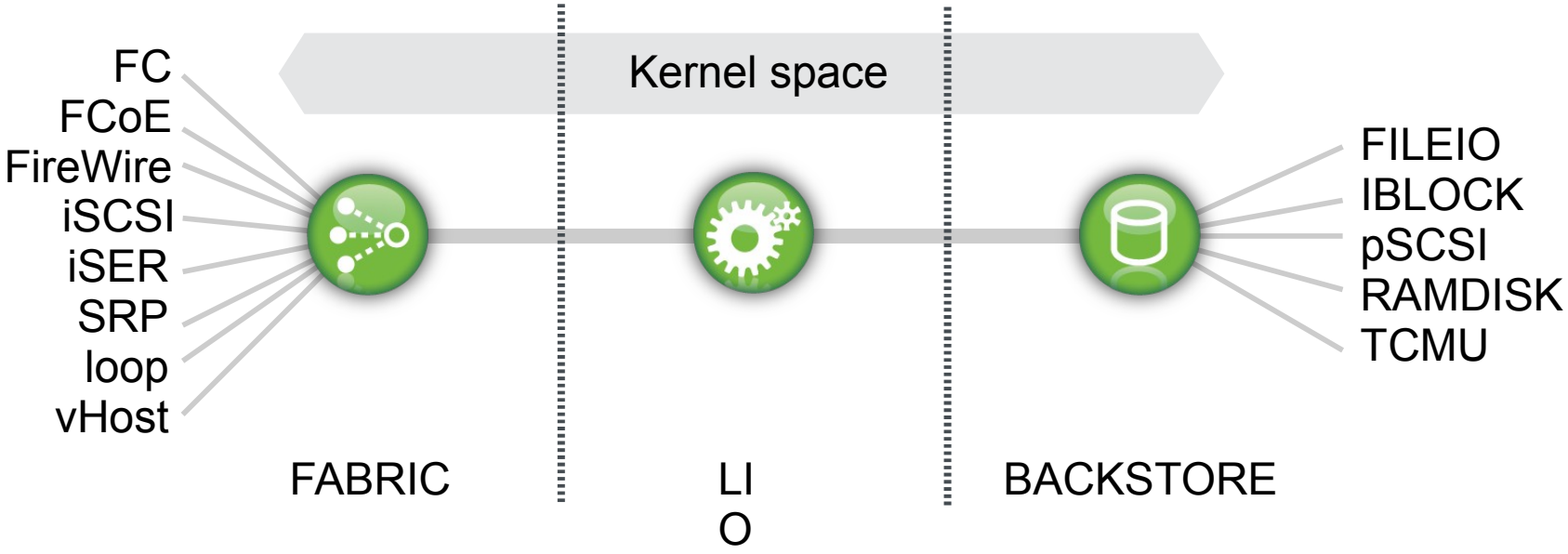
Back to our goal

How to achieve a non disruptive introduction of Ceph into a traditional storage infrastructure?



Linux-IO Target (LIO™)

Is the standard open-source SCSI target in Linux.



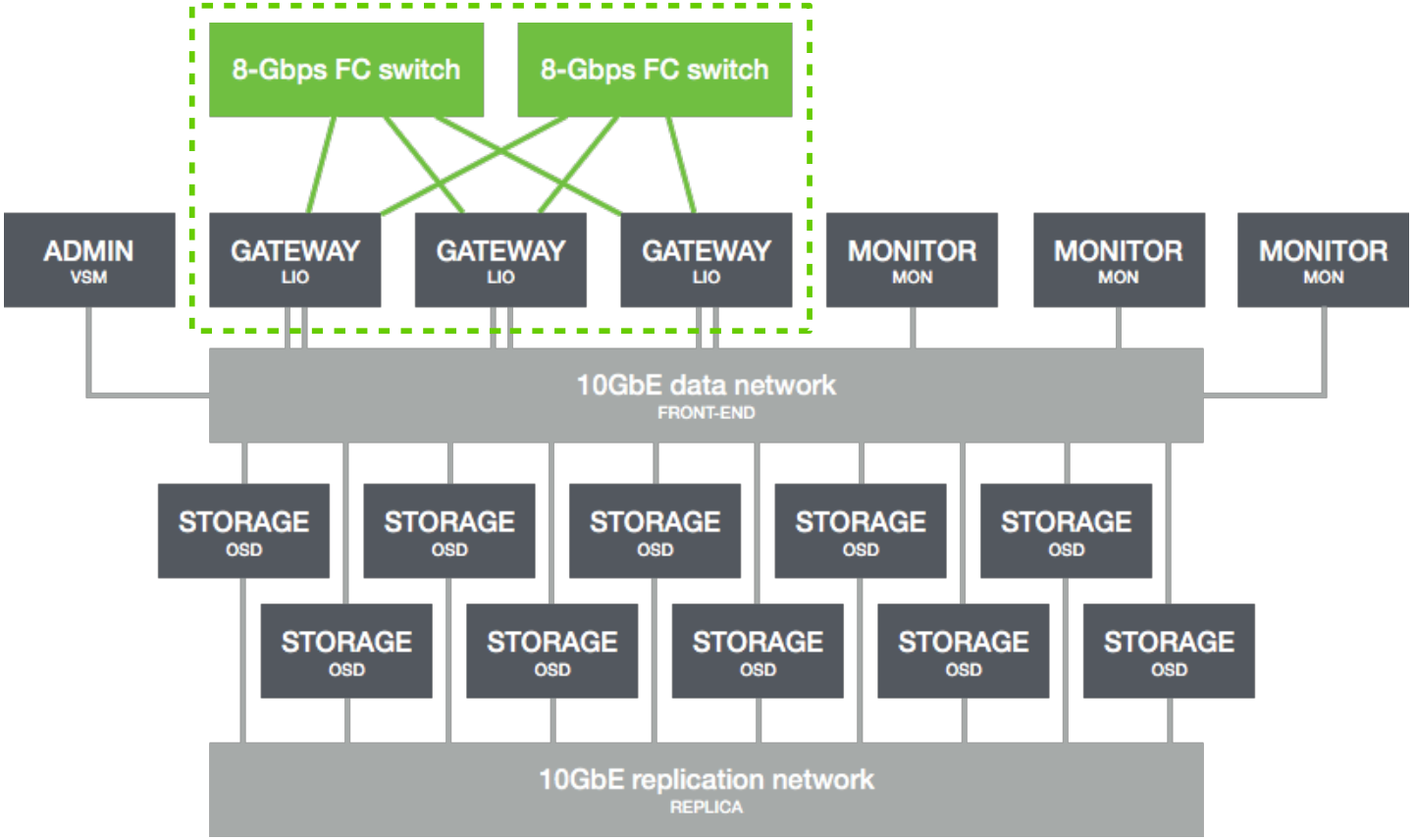
Linux-IO command line

TODO:

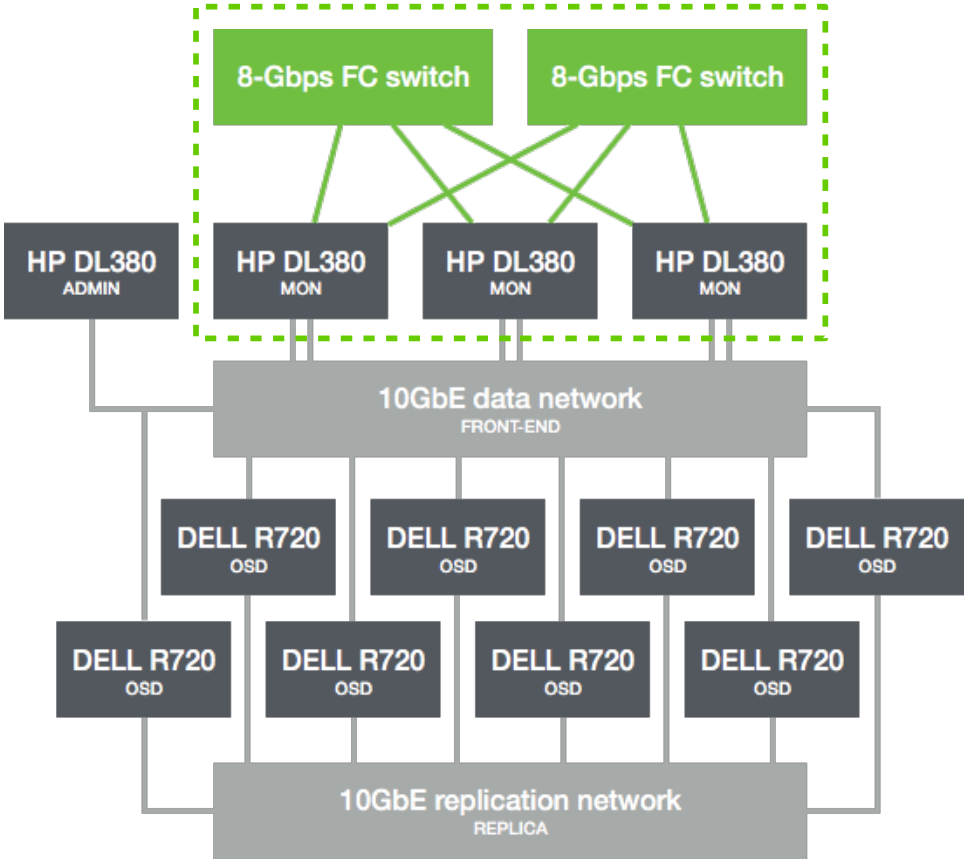
- better define user-space interaction
- targetcli examples

The Architecture

Solution Architecture



The LAB Architecture



Benefits

Smooth Transition

Unlimited scalability

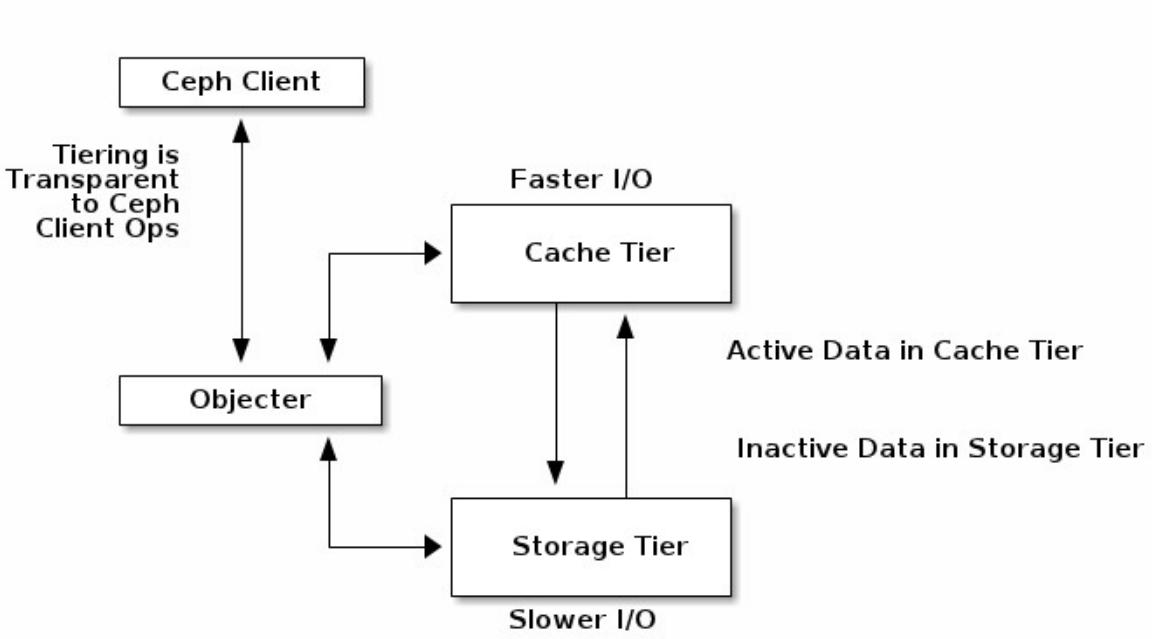
Cheaper

But what about performance?

Benchmarks

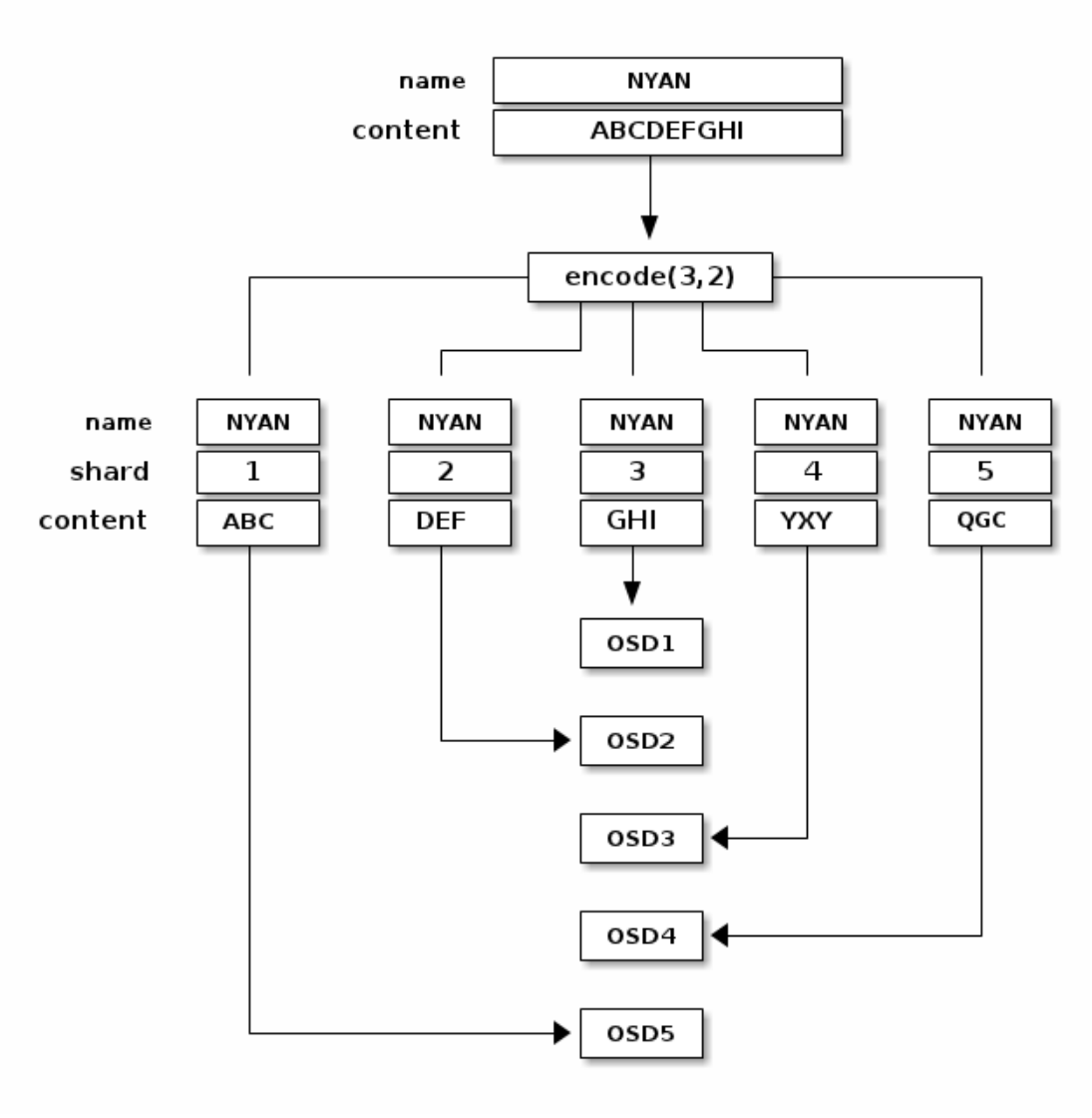
Erasure Code and Cache Tiering

- TBD



Erasure Code and Cache Tiering

- TBD



Benchmarks and Tests

Standard Replication with and without Cache Tiering

CONFIGURATION	SSD	HDD	IOPS	THROUGHPUT
R3SATA	none	3 SATA		
RJ3SATA	journal	3 SATA		
RC3SATA	cache	3 SATA		
RJC3SATA	journal+cache	3 SATA		
R10SAS	none	10 SAS		
RJ10SAS	journal	10 SAS		
RC10SAS	cache	10 SAS		
RJC10SAS	journal+cache	10 SAS		

Benchmarks and Tests

Erasure Code and Cache Tiering

CONFIGURATION	SSD	HDD	IOPS	THROUGHPUT
~~E3SATA~~	~~none~~	~~3 SATA~~	~~N/A~~	~~N/A~~
~~EJ3SATA~~	~~journal~~	~~3 SATA~~	~~N/A~~	~~N/A~~
EC3SATA	cache	3 SATA		
EJC3SATA	journal+cache	3 SATA		
~~E10SAS~~	~~none~~	~~10 SAS~~	~~N/A~~	~~N/A~~
~~EJ10SAS~~	~~journal~~	~~10 SAS~~	~~N/A~~	~~N/A~~
EC10SAS	cache	10 SAS		
EJC10SAS	journal+cache	10 SAS		

Workloads to Test

- 2 nodes cluster with VMware ESX
  - GNU/Linux VM
  - Microsoft Windows 2008/2012 VM
  - TBD: number of VMs, iozone test for Linux, for MS?
- 3 nodes cluster with Oracle RAC
  - Data Warehouse DB
  - Other DB type?
- Other: TBD

Benchmark Results

- TBD
- First results will be available on October
-

A Light Hands-On

A Vagrant LAB for Ceph and iSCSI

- 3 all-in-one nodes (MON+OSD+iSCSI Target)
- 1 admin Calamari and iSCSI Initiator with multipath
- 3 disks per OSD node
- 2 replicas
- Placement Groups: 3*3*100/2 = 450 → 512

Ceph Initial Configuration

Login into ceph-admin and create initial ceph.conf

```
# ceph-deploy install ceph-{admin,1,2,3}
# ceph-deploy new ceph-{1,2,3}
# cat <<-EOD >>ceph.conf
  osd_pool_default_size = 2
  osd_pool_default_min_size = 1
  osd_pool_default_pg_num = 512
  osd_pool_default_pgp_num = 512
  EOD
```

Ceph Deploy

Login into ceph-admin and create the Ceph cluster

```
# ceph-deploy mon create-initial
# ceph-deploy osd create ceph-{1,2,3}:sdb
# ceph-deploy osd create ceph-{1,2,3}:sdc
# ceph-deploy osd create ceph-{1,2,3}:sdd
# ceph-deploy admin ceph-{admin,1,2,3}
```

Linux IO Target Initial Configuration

TBD

```
# Not yet defined
```

LRBD “auth”

```
"auth": [
  {
    "authentication": "none",
    "target": "iqn.2015-09.ceph:sn"
  }
]
```

LRBD “targets”

```
"targets": [
  {
    "hosts": [
      {
        "host": "ceph-1", "portal": "portal1"
      },
      {
        "host": "ceph-2", "portal": "portal2"
      },
      {
        "host": "ceph-3", "portal": "portal3"
      }
    ],
    "target": "iqn.2015-09.ceph:sn"
  }
]
```

LRBD “portals”

```
"portals": [
  {
    "name": "portal1",
    "addresses": [ "10.20.0.101" ]
  },
  {
    "name": "portal2",
    "addresses": [ "10.20.0.102" ]
  },
  {
    "name": "portal3",
    "addresses": [ "10.20.0.103" ]
  }
]
```

LRBD “pools”

```
"pools": [
  {
    "pool": "rbd",
    "gateways": [
      {
        "target": "iqn.2015-09.ceph:sn",
        "tpg": [
          {
            "image": "data",
            "initiator": "iqn.1996-04.suse:cl"
          }
        ]
      }
    ]
  }
]
```

Multipath

```
# cat /etc/multipath.conf
  devices {
     device {
        vendor "(LIO-ORG|SUSE)"
        product "RBD"
        path_grouping_policy "multibus"
        path_checker "tur"
        features "0"
        hardware_handler "1 alua"
        prio "alua"
        failback "immediate"
        rr_weight "uniform"
        no_path_retry 12
        rr_min_io 100
     }
  }
```

Screenshot Here

Some Optimizations

Design optimizations

- SSD on monitor nodes for LevelDB
- SSD Journal decrease IO latency and 3x IOPs

Software optimizations

- deadline elevator on physical disk
- noop elevator on RAID0

Software and Hardware

Software Projects Used

- SUSE Linux Enterprise Storage 2.0 Beta 3
- LRBD
  - Some modification needed to manage FC
- Intel VSM
  - Some plugin written to manage FC and iSCSI

Software Details

- Write detailed usage of previous listed software
-

Hardware Details

- Write Hardware details
- OSD Nodes
- Monitor Nodes
- Gateway Nodes

Q&A

lab@alchemy.solutions

http://alchemy.solutions/lab

Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
www.suse.com

Join us on:
www.opensuse.org