A puppetized Tier3 with fully containerized jobs based on HTCondor, CephFS, XRootD and CVMFS

ErUM Kickoff Workshop

Oliver Freyermuth, Peter Wienemann

Physikalisches Institut
it-support@physik.uni-bonn.de

22nd February, 2019

Physikalisches Institut
UNIVERSITÄT BONN

‘Mobility of Compute’ in 4 dimensions

Original project as associated partner in ErUM:

Add 4th dimension (time) to container mobility

- Allow to suspend / freeze containers, continue on another host
- Missing key feature for usage of opportunistic resources

- Did not get approval ($\Rightarrow$ no FTE. . .)
- In the meantime: `podman` is a feature of RHEL 7.6, supports freezing via CRIU

This talk. . .

- is not about container freezing
- presents our activities close to ErUM

Computing Resources at our Tier 3

Bonn Analysis Facility 2

- In production since Q1 / 2018
- Grid-RSE via XRootD (xroot and WebDAV protocols)
- 1120 (real) cores, ≈ 4.5 GB RAM per core on average
- CephFS offering ≈ 700 TB disk space
- workload management with HTCondor
- all jobs in Singularity containers
- OS: CentOS 7.6

A puppetized Tier 3

Configuration Management

- All hosts provisioned via Foreman, configuration managed via Puppet
  - Any service hosts (XRootD, Gateways, Condor CM)
  - Disk servers
  - Worker nodes
  - Desktops
- No manual configuration on the host
- Any machine / VM can be reinstalled with a click and reboot

- Puppetization can be reused for 'Tier 3 in a box'
- Upgrades to new OS releases strongly simplified

Containerized jobs in HTCondor

Singularity Containers

- For now, use Singularity (HTCondor & WLCG requirements)
- Plan to test `runc`, `podman`, ...
  $\Rightarrow$ Longer shelf life, larger community, upstream security focus
- Locally built containers for SL 6, CentOS 7, Ubuntu 18.04 LTS (+ `HEPOSlibs`)
- Based on Docker images, rebuilt daily
- Containers & special software distributed via local CVMFS (deduplication!), `lmod` for software selection

Containers in HTCondor

- **All** jobs run in containers, no 'login' nodes or SL 6 machines!
- Development: Interactive jobs (HTCondor: SSH with X11)
- No freezing of containers yet...

CephFS

CephFS

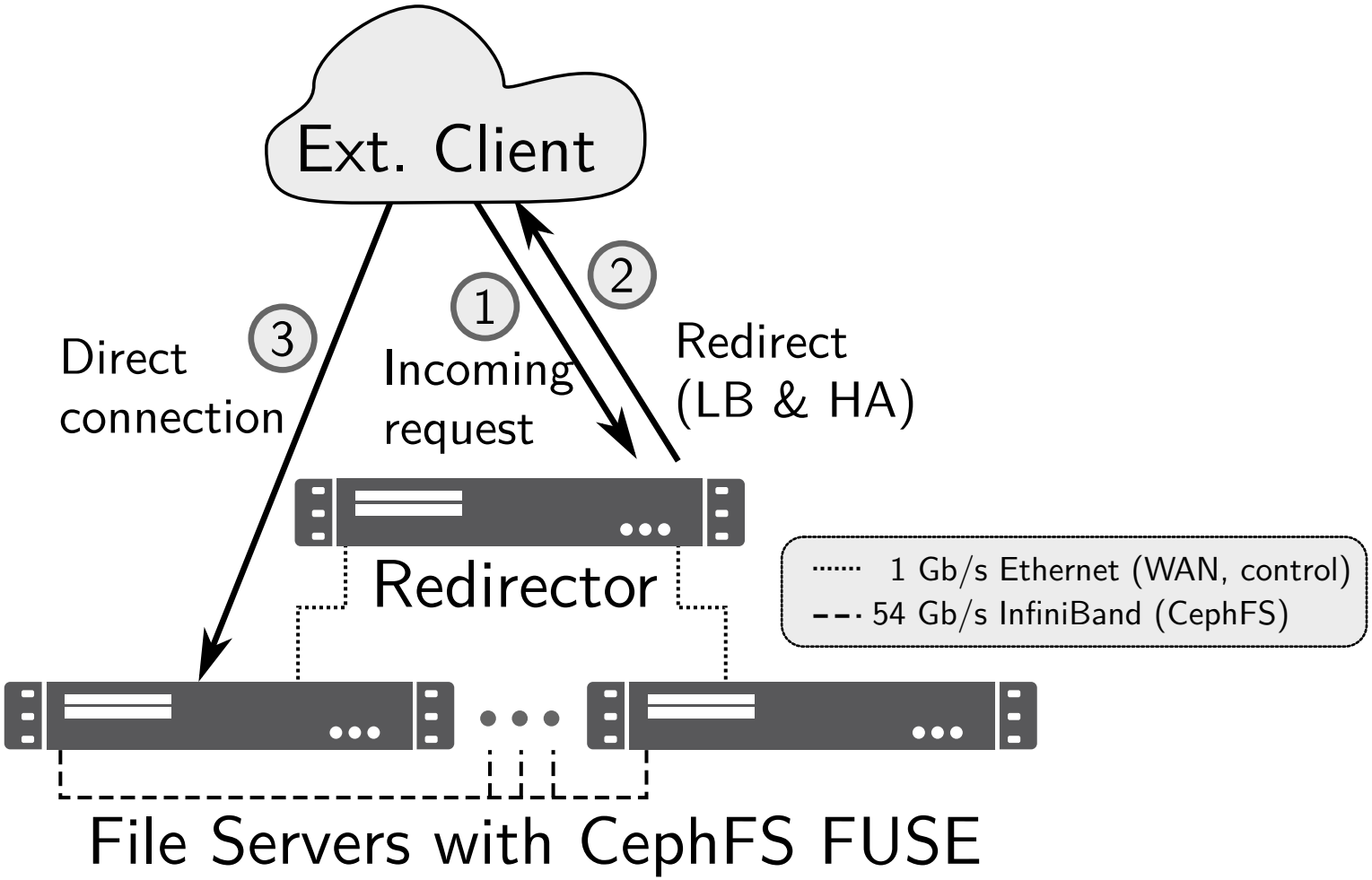
- CephFS in Erasure Coded setup ($k = 4$, $m = 2$) on 7 OSDs
- Snappy compression for all data
- Follow upstream releases closely
- CephFS FUSE client used in cluster
- On desktops: mounted via NFS (via NFS Ganesha)
- Experience:
  - Very resilient
  - Extremely flexible:
    $\Rightarrow$ Added disk server, purged complete disk server etc.
  - Performs very well, we achieve about 5 $^{\text{GB}}/_{\text{s}}$
    (limited by IPoIB performance)

XRootD

XRootD

- Protocol will replace SRM / GridFTP
- XRootD running on disk servers ($7 \cdot 1\,{}^{\text{Gbit}}/_{\text{s}}$)
- Directly operates on CephFS (makes use of `xattrs` etc.)
- Redirector automatically load-balances between disk servers, high availability
- 'Simple' configuration (as compared to dCache, GridFTP)
- Third-Party-Copy support partially there, improvements upcoming in next release

XRootD


Ext. Client
1
2
3
Incoming request
Redirect (LB & HA)
Direct connection
Redirector
1 Gb/s Ethernet (WAN, control)
54 Gb/s InfiniBand (CephFS)
File Servers with CephFS FUSE

Summary

- Tier 3 in Bonn is fully puppetized
- Complete containerization of all jobs
- First RSE in the DE-cloud with xrootd-only in production
- Production use of CephFS as storage for HTC and RSE (also use Ceph-RBD for virtualization, backup cluster planned)
- Original proposal (container freezing) has gained high momentum in the world-wide community
- Plan to publish an in-depth overview of our setup and experiences

Thank you

for your attention!