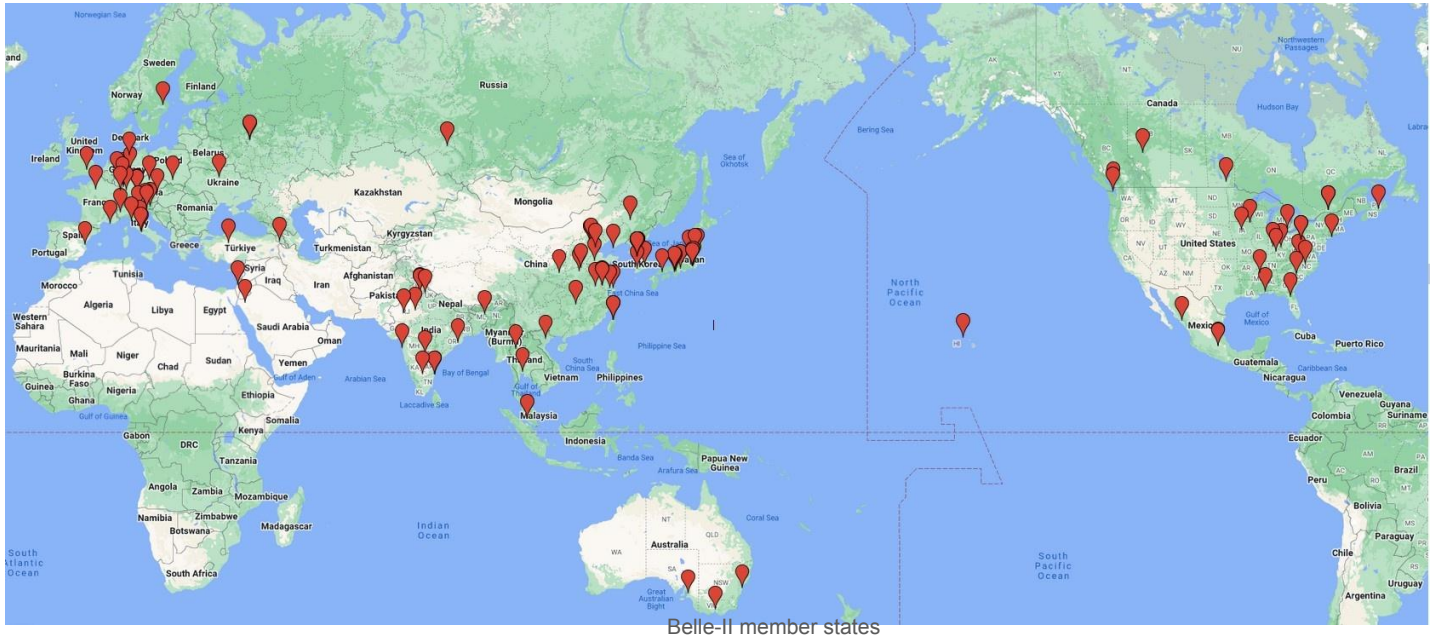
The Canadian Belle II Tier1 RAW Data Centre (RDC)

at the University of Victoria

Marcus Ebert
on behalf of the UVic HEP-RC group
University of Victoria

The Belle II experiment



Belle-II member states

- electron-positron collider experiment at

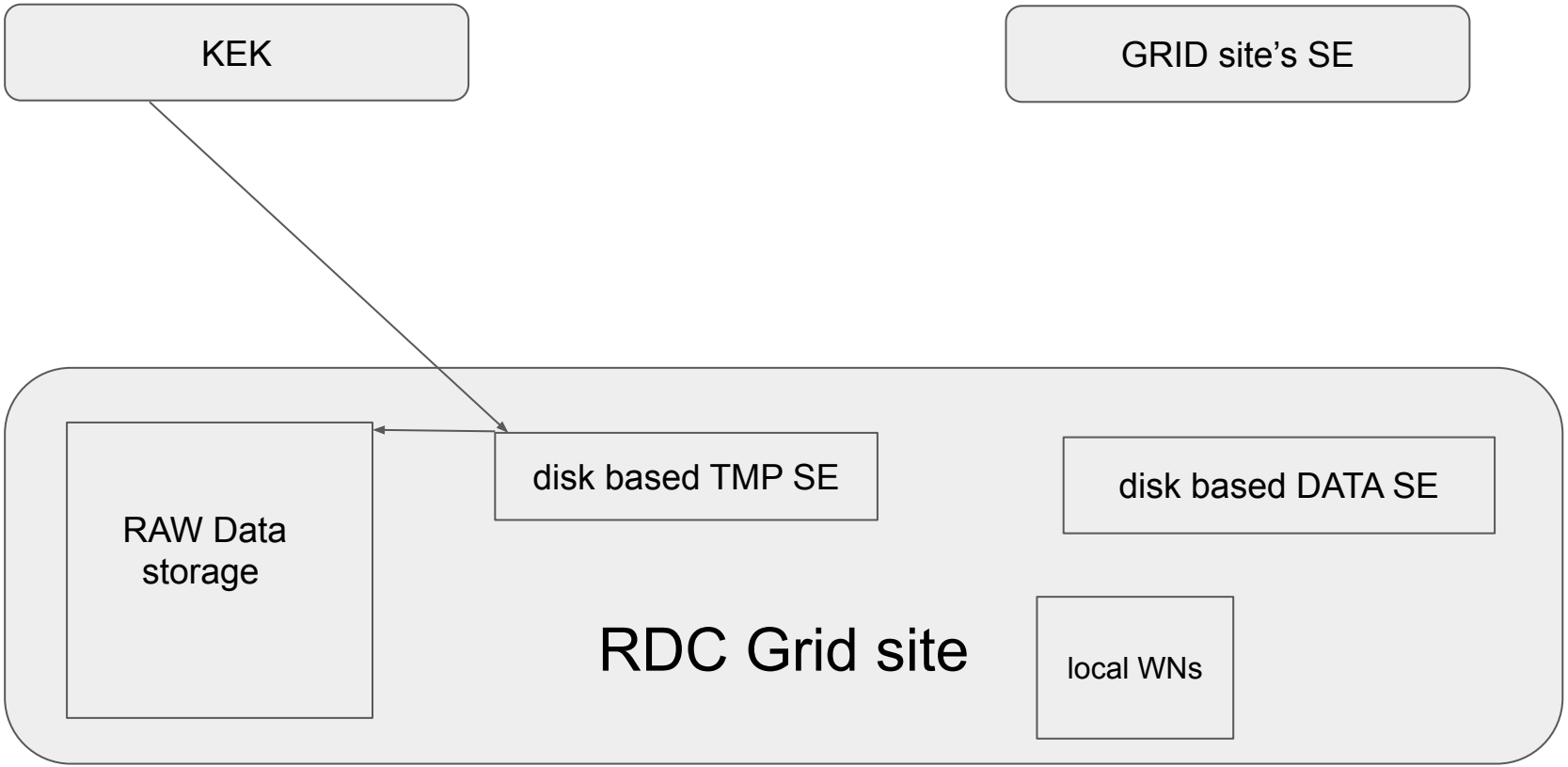
KEK Inter-University Research Institute Corporation
High Energy Accelerator Research Organization

- Data taking since 2021 (physics data)
- expected to take data for ~10 years

- copy of the RAW data is stored and processed at international centres
  (USA, Italy, France, Germany, Canada)

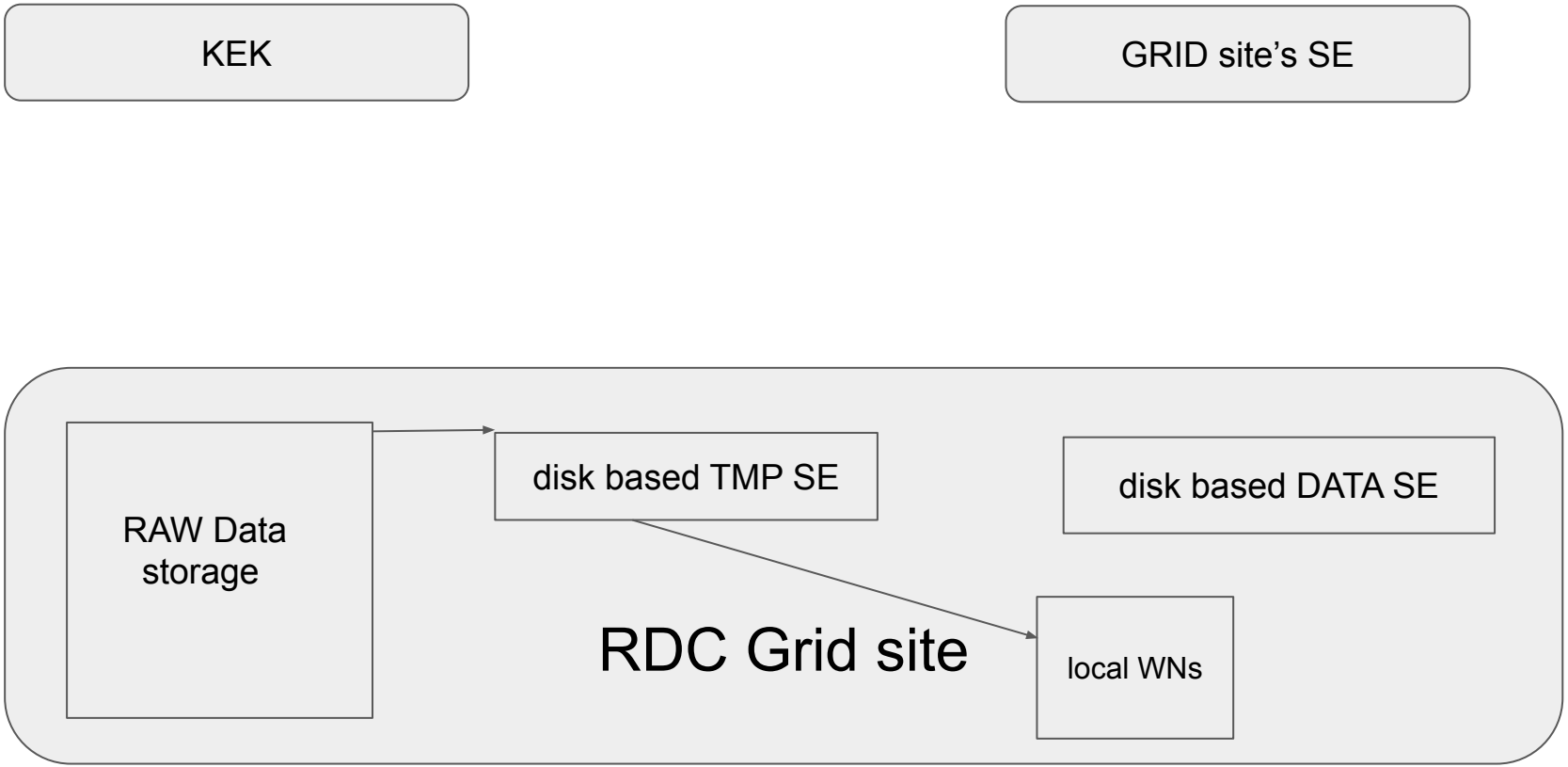
- Canada stores 15% of the collected raw data

RDC data flow from KEK to the RDC site

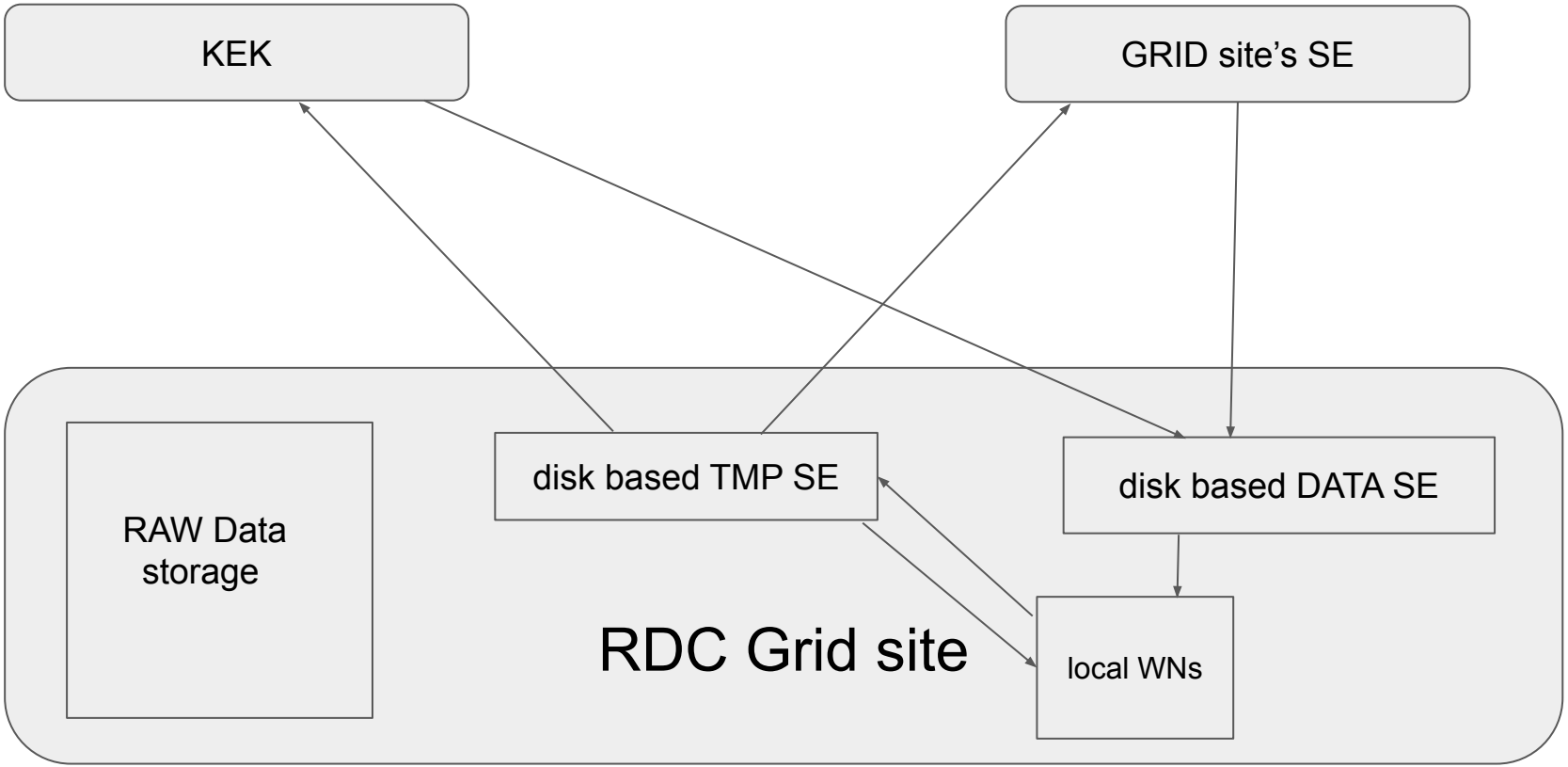


Marcus Ebert (mebert@uvic.ca)

RDC data flow for processing within the RDC site

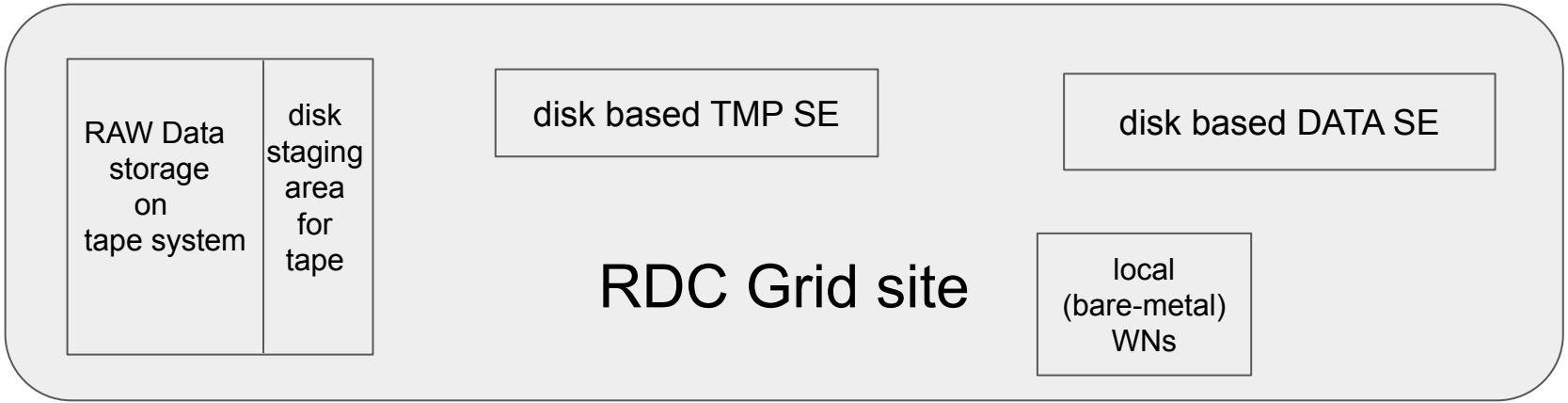


data flow for processed data



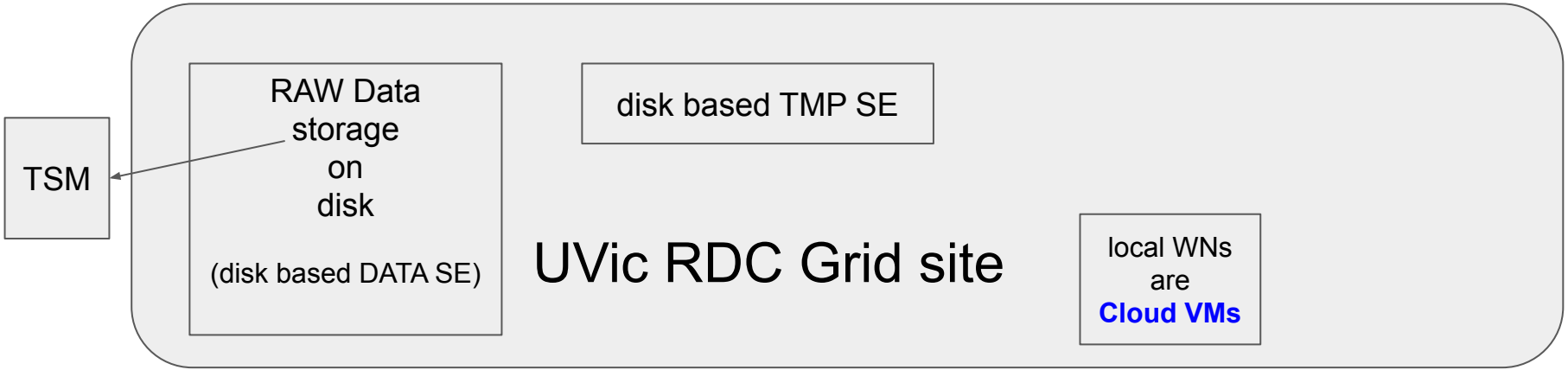
Marcus Ebert (mebert@uvic.ca)

Usual RDC site infrastructure



- TMP and DATA SE usually same infrastructure
  - often managed by dCache, EOS, or StoRM
    - have their own complex infrastructure
- RAW data stored on tape
  - its own Grid SE, exposing own /DATA area
  - stage-in (and copy to TMP SE) for processing/access initiated via grid

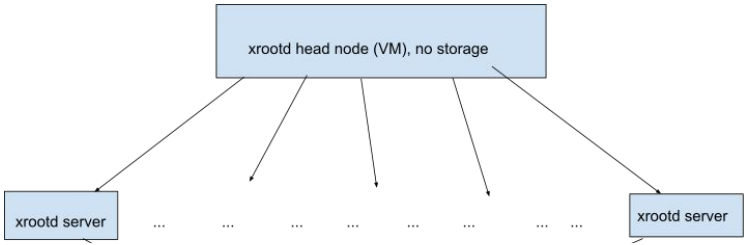
UVic RDC site infrastructure



- RAW data stored on disk in its own /DATA area
- infrastructure also shared with TMP area
  - managed by **plain xrootd** (very simple infrastructure)
  - ZFS raidz3 for data disks (**ZFS ensures on-disk data consistency**)
- backup to TSM (managed by University) in background

Canadian RDC - data access

- using plain xrootd for data access
  - can handle *davs://* as well as *root://* protocol
  - very simple to setup and configure
    - single configuration file for redirector (head node) and data servers
  - files stored under specific directory using their full logical path
    - files can be accessed/copied with any tool
    - no database dependency to associate files with their logical path and filename



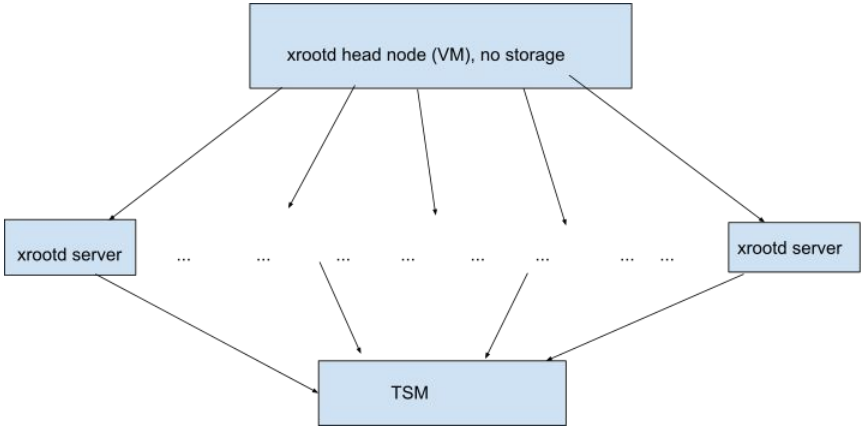

```
mebert@heplw65:~$ gfal-ls davs://xrd5.belle.uvic.ca:1094//DATA/
belle
storage-summary.json
testdirprod01
testdir

[root@xrd5 ~]# ls  /storage/RDC/DATA/
belle  storage-summary.json  testdir  testdirprod01
```

davs://rdc-redirector.belle.uvic.ca:1094//DATA/belle/Raw/e0030/physics/r01860/sub00/physics.0030.01860.HLT08.m01.f00000.root
--> */storage/RDC/*DATA/belle/Raw/e0030/physics/r01860/sub00/physics.0030.01860.HLT08.m01.f00000.root

Canadian RDC - backup

- each DATA area on servers is backed up to tape daily
  - TSM managed by University services
  - local & off-site backup

- using full path as on disk == full logical path of the experiment
  - easy to restore files
  - easy to restore to another server in case of machine failure
  - xrootd can restore automatically on-demand



Canadian RDC - performance

- Network interface (server): 25Gbps
- 28 internal disks in raidz3 (per server)

- read/write from/to disk: ~10Gbps
  - limited by disk access
  - fast enough for external transfers

- gives possibility to extend with another, independent disk array later on
  - servers can handle external disk enclosures
  - efficient way to increase capacity without purchasing new servers

Canadian RDC - infrastructure services

- in addition to data servers, also infrastructure needed
  - xrootd redirector, cobbler server, perfsonar, NFS server for shared files, ...
- perfsonar on dedicated machine
- other infrastructure services run as VM
  - have 2 Hypervisor machines
    - one runs VMs, other as redundancy in case of hardware failure
- monitoring via own scripts/web pages, as well as via Ganglia and Nagios
  - consistency between files on disk and on tape
  - monitoring tape backup process
  - monitoring of resource usage and availability of services
  - ...

Canadian RDC - computing

- Belle II uses DIRAC as distributed computing system

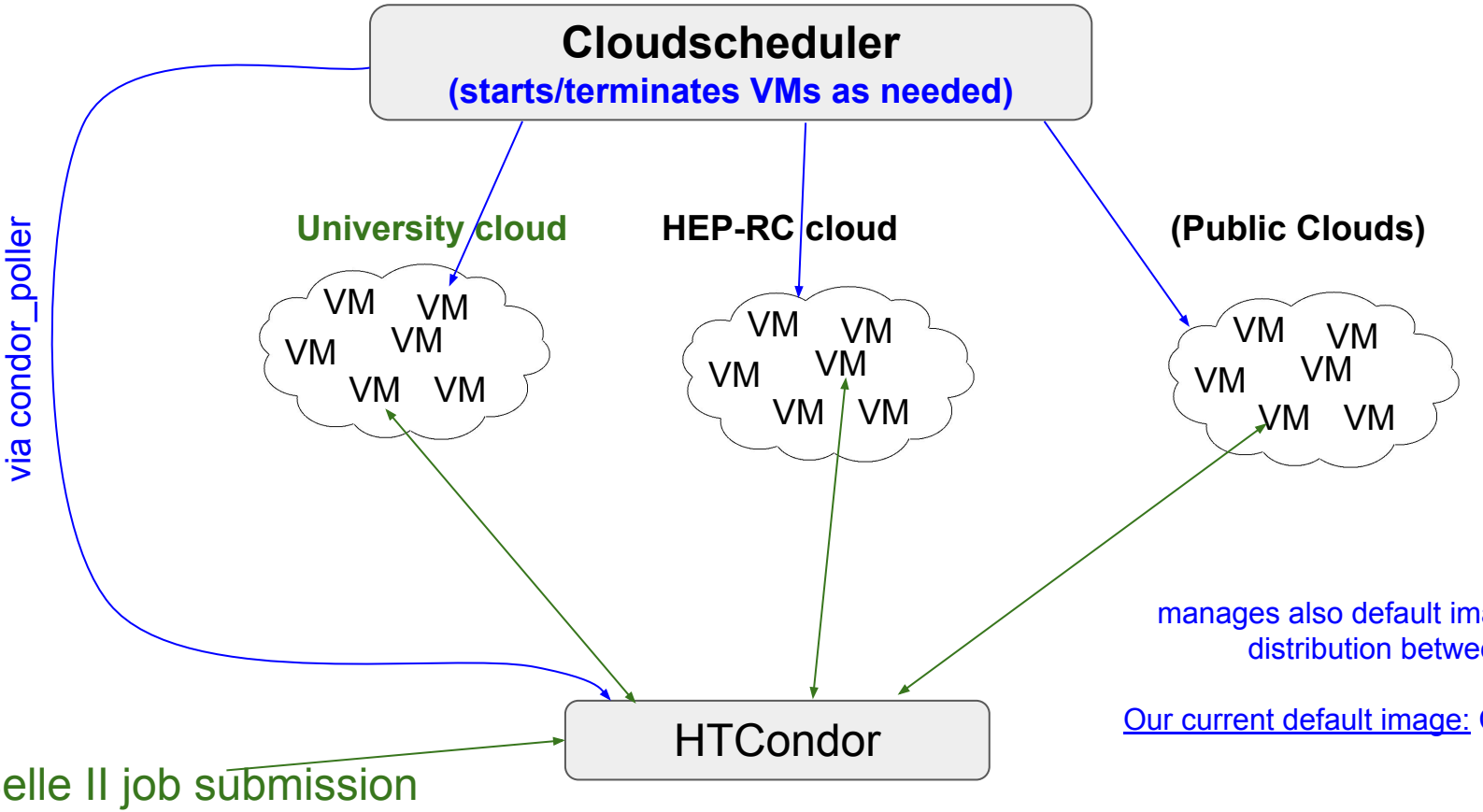
- we run local DIRAC site director and HTCondor batch system
  - site director submits pilots to the local HTCondor batch system

- HTCondor worker nodes are cloud VMs
  - cloud infrastructure managed by the university's Research Computing group
  - VMs started on demand by Cloudscheduler (https://csv2.heprc.uvic.ca/public/)

Canadian RDC - Cloudscheduler

- using clouds removes dependency on physical machines
  - in case of outages, other clouds could be used temporarily
- using VMs makes updating the system easy
  - change a single image and use it for any newly booted VM
- using VMs can give jobs exactly the resources they need
  - booting VMs with different flavors depending on what a job needs

Canadian RDC - Cloudscheduler



manages also default image and ssh key distribution between clouds

Our current default image: CERNVM4 image

Issues during deployment

- issues with xrootd's http handling with multiple servers behind redirector
  - copying via FTS using davs:// :
    - ask if path exists on SE, if not create the path
      - xrootd redirector asks all connected servers if anyone has the path
        - only server 1 may respond "yes"
    - if file doesn't exist yet, copy file to SE
      - xrootd redirects copy request to a server that has enough space
        - **can different server than the one which responded to have the directory already**

When xrootd is active party in TPC davs:// based transfer, and the path didn't exist on the server the client was redirected to, then the transfer would fail.

expected behaviour: if needed create directory on the server the client was redirected before transfer

Issue fixed in xrootd 5.7. and above!

Issues during deployment

- issues with directory listing via davs:// (*gfal-ls davs://...../DATA/*)
  - listing request sent to redirector
  - redirector sent request to a server that has the directory
  - server responds with content
    - responds to client, not to redirector
      ===> client gets content only from a single server

<u>workaround available:</u>

- add listing related http options to the servers
- redirect listing request to a proxy server
- proxy server should (via root://) query all servers and then return the request via davs://

Doesn't work as expected currently, but in contact with xrootd developers.
(fortunately not critical for normal operation as raw data SE)

Summary

Canadian RAW data centre for Belle II keeps all data on disk with backup to TSM based tape system.
(can restore files to any of the servers in case needed; manually or automatically on demand)

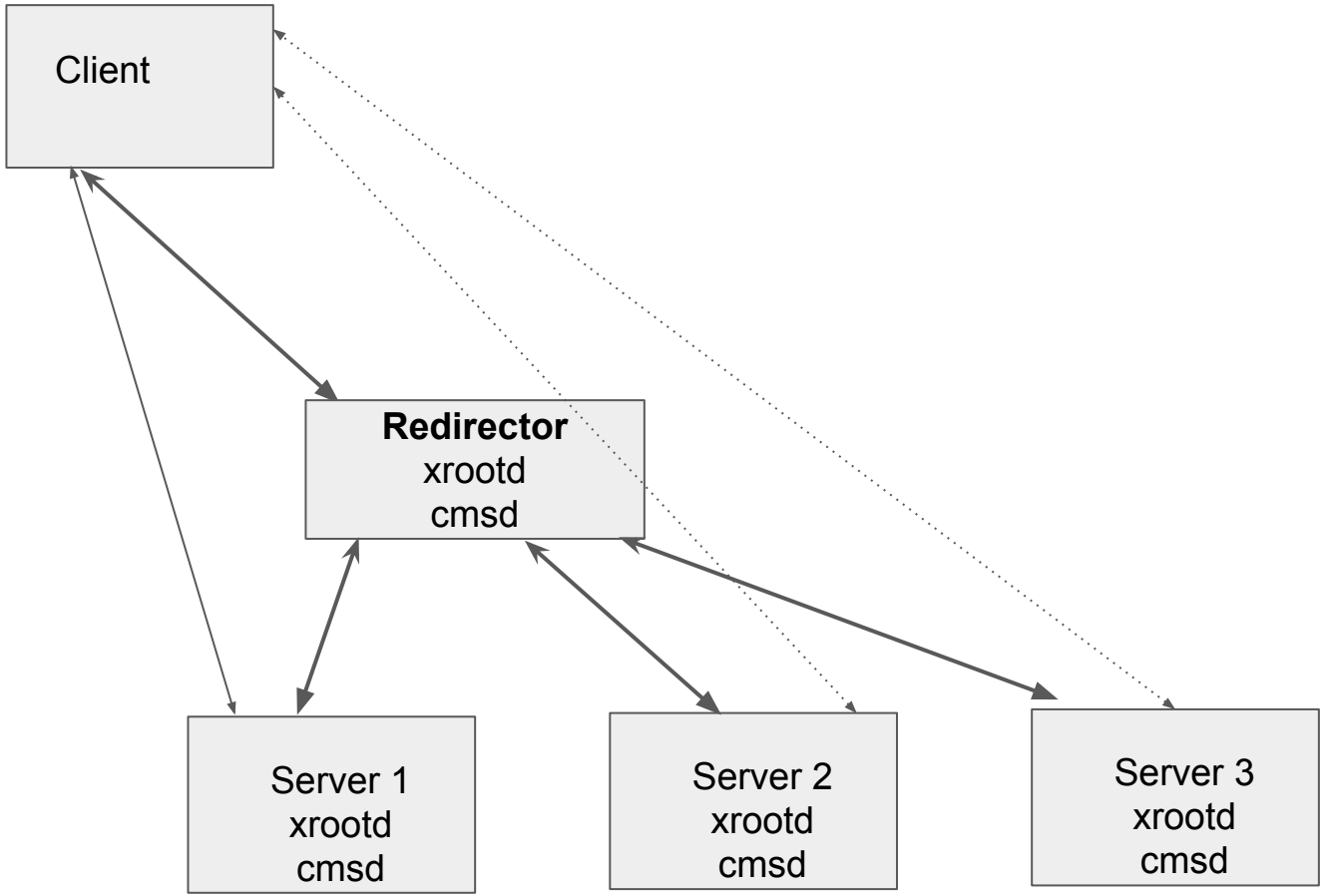
Data servers make use of ZFS raidz3 and hardware raid mirror for OS disks.
(28x20TB data disks, 25Gbps NIC)

Plain xrootd is used for access via davs:// and root://
(no complicated SE software needed)

In production since early 2024.
(based on 6 data servers and 1 redirector)

Good response to bug reports by xrootd developers!

Locally Federated XRootD



Canadian RDC - disk servers

- data servers with 28 x 20TB data disks and 2 NVMe disks for OS
  - currently 6 servers, can add more when needed
    - all servers independent
  - possibility to connect external enclosures to increase capacity
- network: 25Gbps NIC
- hardware raid1 for OS disk
  - using AlmaLinux 9
- ZFS raidz3 for data disks
  - any 3 disks can fail without data loss
  - ~450TB per server usable space
  - ZFS ensures on-disk data consistency and can restore corrupted blocks automatically