Cloud based Tier 2 Computing at the University of Melbourne

David Dossett

University of Melbourne

Martin Sevior

University of Melbourne

Marcus Ebert

University of Victoria (Canada)

Australia Tier 2 status

At its peak the physical Tier 2 resources were:

- ~1200 CPU cores and 1.2 PB DPM storage for ATLAS
- ~300 CPU cores and 20 TB DPM storage for Belle II (+ a few hundred CPU cores in OpenStack)

Funded and maintained by the ARC Centre of Excellence for Particle Physics at the Terascale (CoEPP) → wound down in 2018

Storage is now down to ~0.8 PB DPM as the old disks die off

In November 2022 our final network switch (of 3) that was powering the compute nodes failed. This put the ATLAS compute site offline

Current Situation

In the last year we have won grants for ATLAS/Belle II that include infrastructure:

- 600k AUD for hardware
- 300k AUD for manpower (about 2.5 years for a FTE employee)

Research Computing Services (RCS) at Melbourne now centrally maintain Infrastructure as a Service (IaaS) for all faculties across the university

They provide (among other things):

- The Melbourne Research Cloud (MRC) [OpenStack]
- Ceph object storage with Swift/Mediaflux/S3 API access
- Expert support for everything they maintain

Initial Cloud Resources

1024 VCPUs in a MRC OpenStack project and **800 TB** Ceph object storage available via S3 gateway

- These can be moved around the cloud to other projects or storage if needed

Still most of the budget left to make sizable purchases (~2x more) depending on the outcome of the commissioning

An additional preemptiable **1200 VCPUs** provided for free by RCS for us to opportunistically use

Grid Storage Restrictions

Installing and managing our own disks/servers in the RCS infrastructure is not feasible, and funding + expertise isn’t at the level where we could build our own

- Standard servers in front of disk storage like XRootD and dCache are not possible

RCS resources are common across faculties, so we cannot require specific configurations for them

No direct/admin access to current Ceph clusters as they are shared and in a different network than the OpenStack VMs

- Must use the provided APIs (S3/Swift/Mediaflux)
- No CephFS provided so we can’t put dCache/XrootD in front of a filesystem

Grid Storage Requirements

Unfortunately we can’t migrate our DPM storage to a solution that would fit ATLAS + Belle II preferences e.g. dCache

But if we can use the RCS cloud resources and their support we can save a lot of money and manpower

→ Put the funding directly into purchasing storage/compute

Want a solution that will work *reliably* and:

- Can support future WLCG goals e.g. HTTP/WebDAV and IAM auth tokens
- Uses our available access to Ceph storage (S3/Swift)
- Reasonable to maintain or support with our manpower funding

For now the only technology that works with the S3 API directly is **Dynafed**

Dynafed?

A storage federation tool built using an Apache, dmlite, and memcached

Uses HTTP/WebDAV and redirects requests to the correct storage endpoint URLs e.g. on our S3 gateway

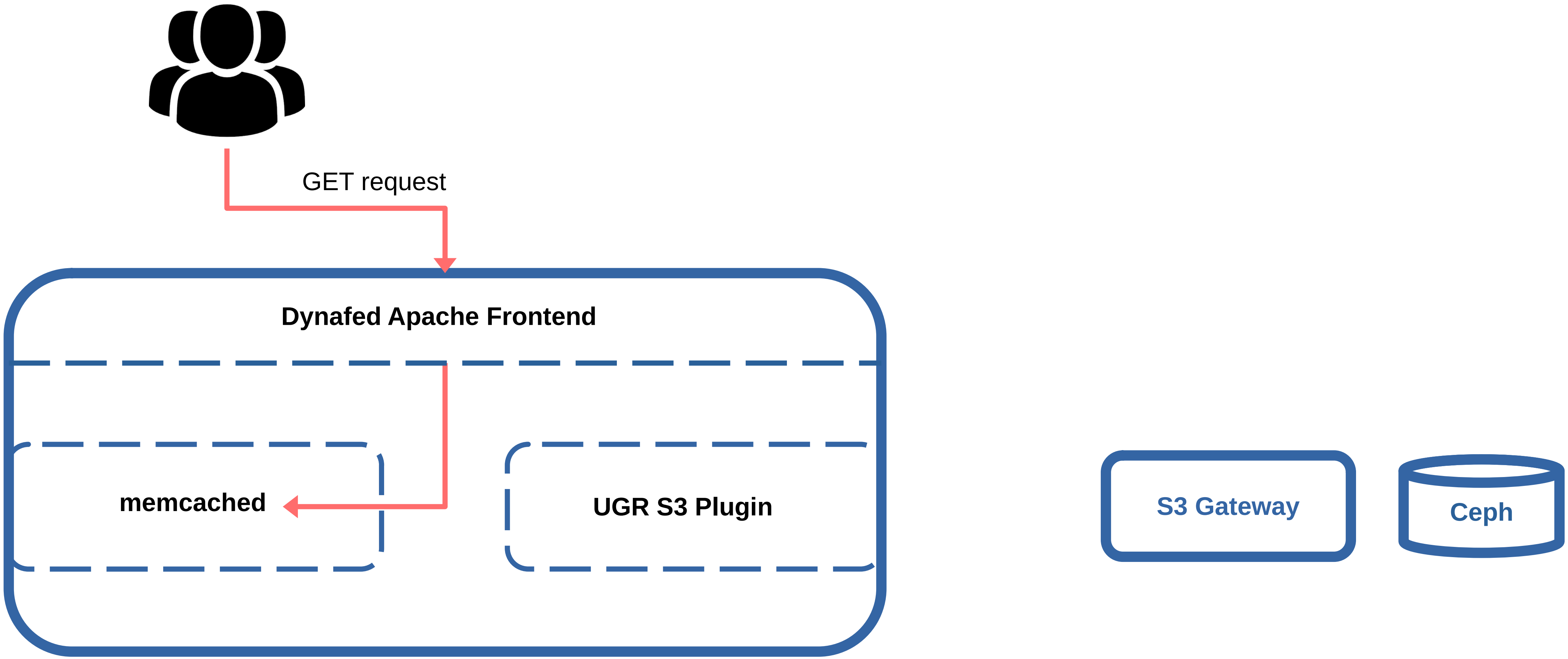
Very lightweight and simple to maintain

- No complex databases or task schedulers
- Monitoring added using Apache logs via **Elastic Stack** Logstash/Beats and U. Vic's **Grafana**

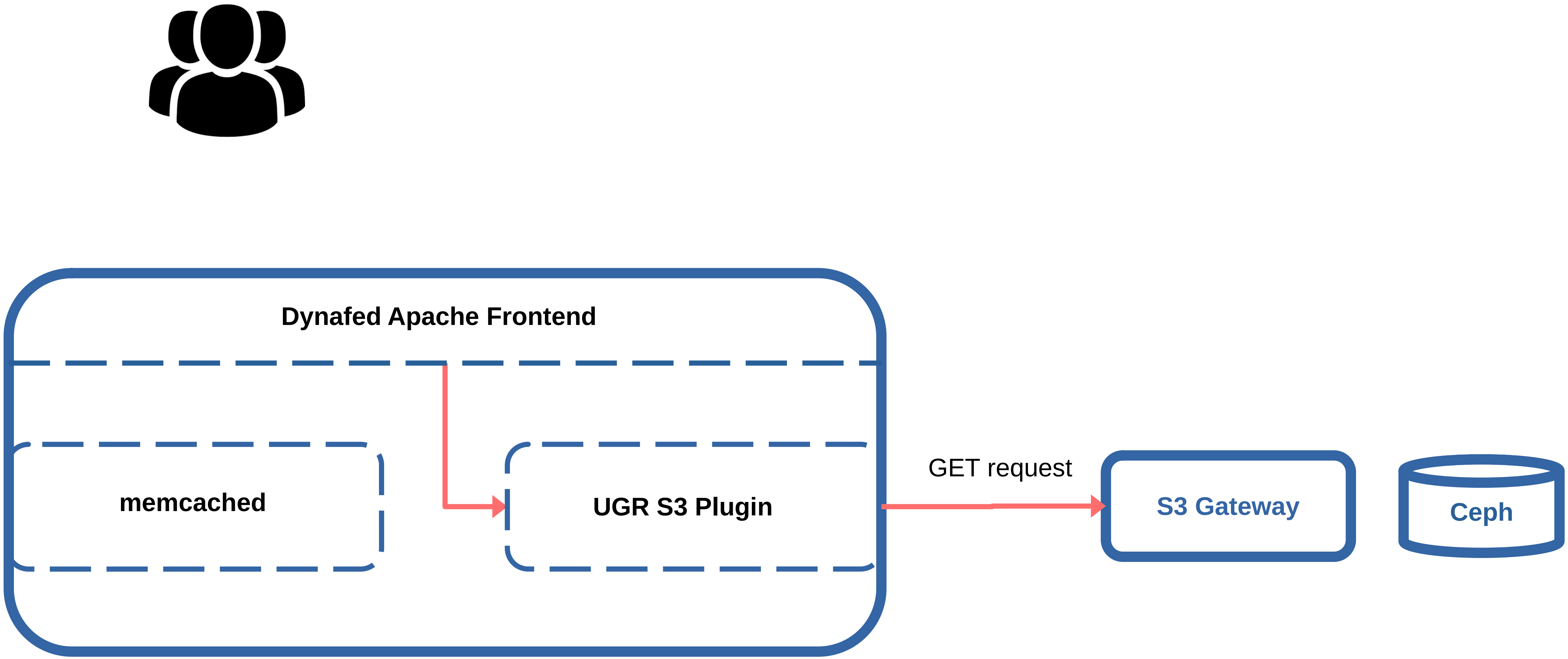
Main issues to solve for Dynafed

- DPM project goes EOL in summer 2024 (Dynafed and dmlite are part of it)
- Identity Access Management tokens instead of X.509+VOMS should work but require development

Dynafed With S3 API



Dynafed With S3 API



Dynafed Apache Frontend

memcached

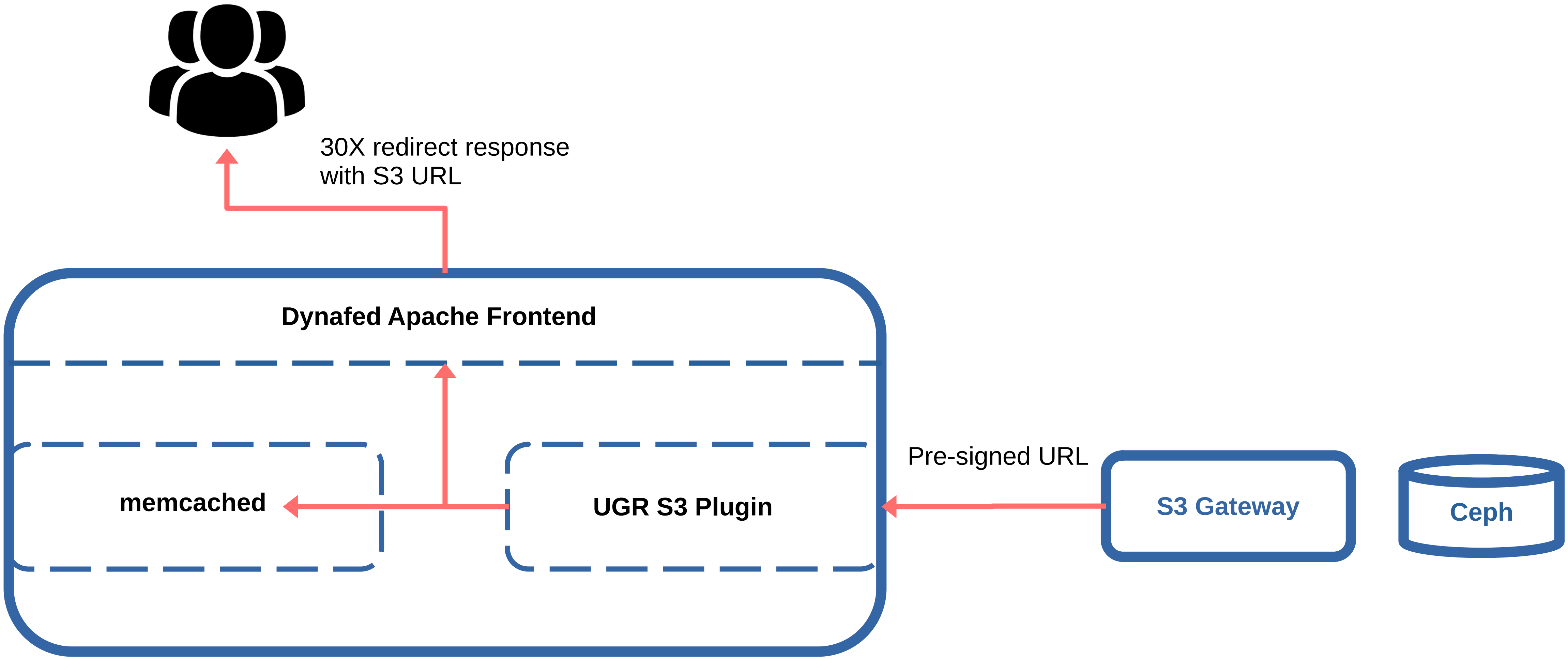
UGR S3 Plugin

GET request

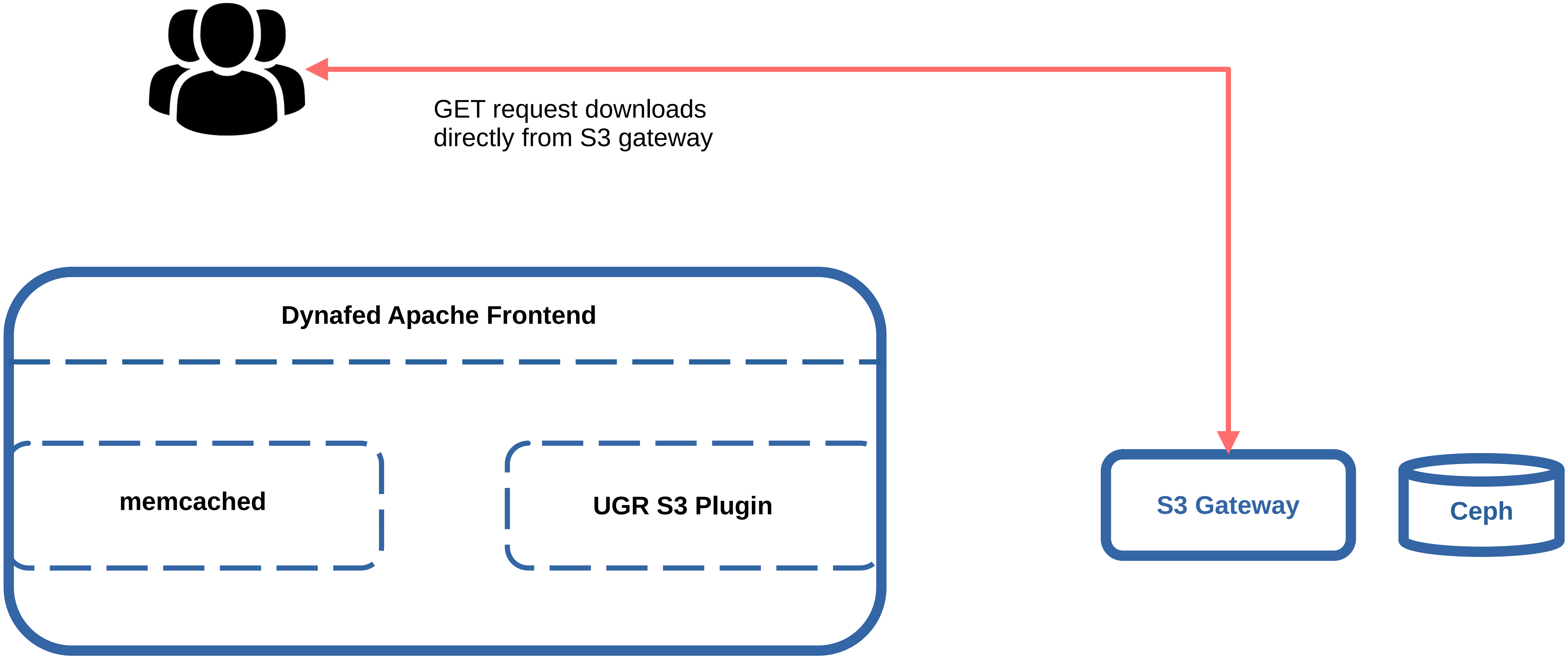
S3 Gateway

Ceph

Dynafed With S3 API



Dynafed With S3 API



Belle II Dynafed Server Commissioning

VOMS + grid certificate auth working

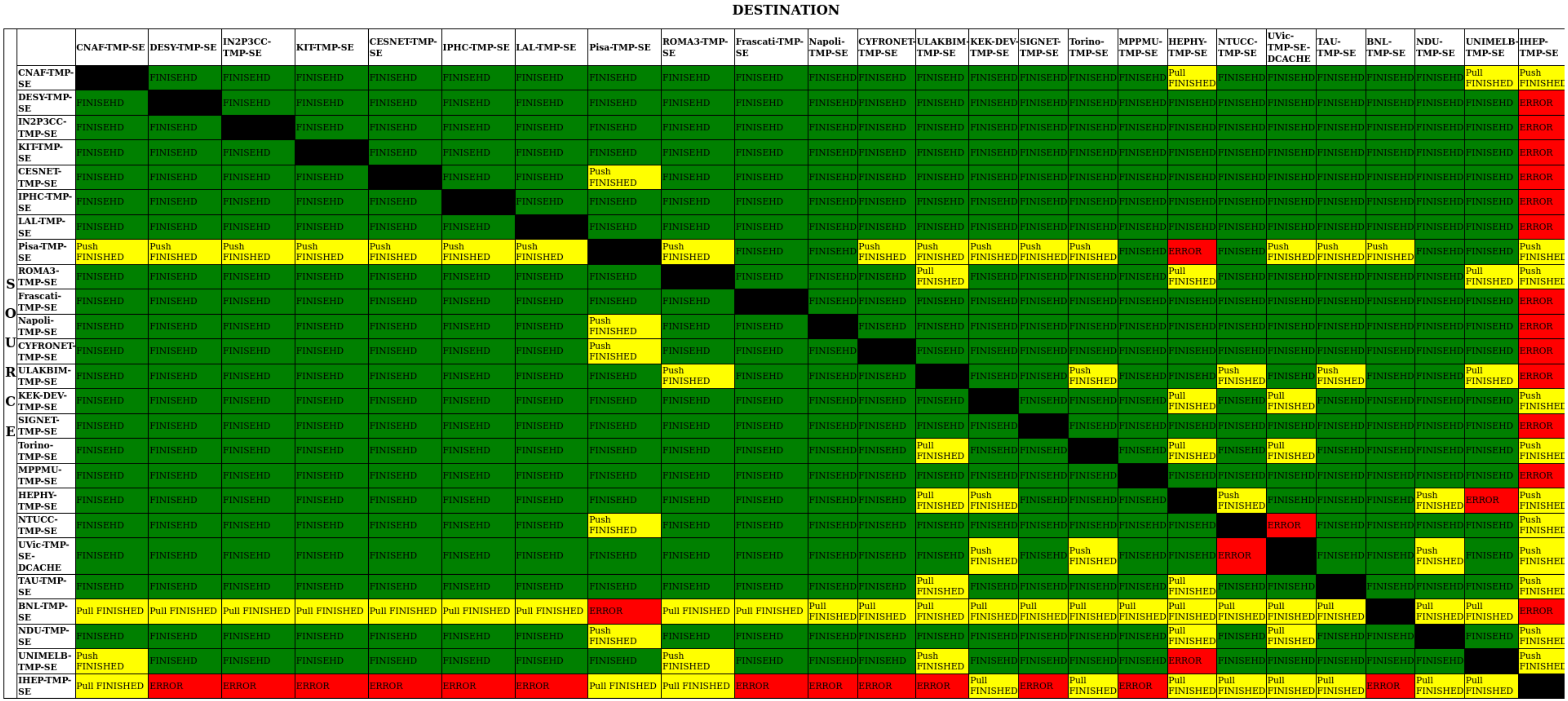
Macaroon bearer tokens used during third-part copy (TPC)

MD5 and adler32 checksum calculation works:

- Memcached is checked first
- Then the object store file metadata
- Then streamed download and calculation. Checksum gets added to object metadata and memcached for next time it is queried

Installation and setup of our hosts with ANSIBLE

Belle II Dynafed TPC Tests



DESTINATION

SOURCE

	CNAF-TMP-SE	DESY-TMP-SE	IN2P3CC-TMP-SE	KIT-TMP-SE	CESNET-TMP-SE	IPHC-TMP-SE	LAL-TMP-SE	Pisa-TMP-SE	ROMA3-TMP-SE	Frascati-TMP-SE	Napoli-TMP-SE	CYFRONET-TMP-SE	ULAKBIM-TMP-SE	KEK-DEV-TMP-SE	SIGNET-TMP-SE	Torino-TMP-SE	MPPMU-TMP-SE	HEPHY-TMP-SE	NTUCC-TMP-SE	UVic-TMP-SE-DCACHE	TAU-TMP-SE	BNL-TMP-SE	NDU-TMP-SE	UNIMELB-TMP-SE	IHEP-TMP-SE
CNAF-TMP-SE		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	Push FINISHED
DESY-TMP-SE	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
IN2P3CC-TMP-SE	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
KIT-TMP-SE	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
CESNET-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
IPHC-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
LAL-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
Pisa-TMP-SE	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED		Push FINISHED	FINISEHD	FINISEHD	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED	Push FINISHED	FINISEHD	ERROR	FINISEHD	Push FINISHED	Push FINISHED	Push FINISHED	FINISEHD	FINISEHD	Push FINISHED
ROMA3-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	Push FINISHED
Frascati-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
Napoli-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
CYFRONET-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
ULAKBIM-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	Pull FINISHED	ERROR
KEK-DEV-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED
SIGNET-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
Torino-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD		FINISEHD	Pull FINISHED	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED
MPPMU-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR
HEPHY-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	Push FINISHED	FINISEHD	FINISEHD	FINISEHD		Push FINISHED	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	ERROR	Push FINISHED
NTUCC-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		ERROR	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED
UVic-TMP-SE-DCACHE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	ERROR		FINISEHD	FINISEHD	Push FINISHED	FINISEHD	Push FINISHED
TAU-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD		FINISEHD	FINISEHD	FINISEHD	Push FINISHED
BNL-TMP-SE	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	ERROR	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED		Pull FINISHED	Pull FINISHED	ERROR
NDU-TMP-SE	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pull FINISHED	FINISEHD	Pull FINISHED	FINISEHD	FINISEHD		FINISEHD	Push FINISHED
UNIMELB-TMP-SE	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	Push FINISHED	FINISEHD	FINISEHD	FINISEHD	FINISEHD	ERROR	FINISEHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD		Push FINISHED
IHEP-TMP-SE	Pull FINISHED	ERROR	ERROR	ERROR	ERROR	ERROR	ERROR	Pull FINISHED	Pull FINISHED	ERROR	ERROR	ERROR	ERROR	Pull FINISHED	ERROR	Pull FINISHED	ERROR	Pull FINISHED	Pull FINISHED	Pull FINISHED	Pull FINISHED	ERROR	Pull FINISHED	Pull FINISHED	

Other Options and ATLAS

ATLAS would prefer not to add S3 storage if possible due to worries about disproportionately adding to the Ops load

Our plan is to commission a server to test if Dynafed S3 storage works reliably in their current ecosystem

- If we can’t convince ATLAS we may be forced to pivot to only providing compute resources for them

We can investigate getting RCS to build and maintain a Ceph cluster for Belle II and ATLAS that we control. Then use it with XRootD/dCache and CephFS

- But there are no guarantees that it is feasible for us

Grid Compute

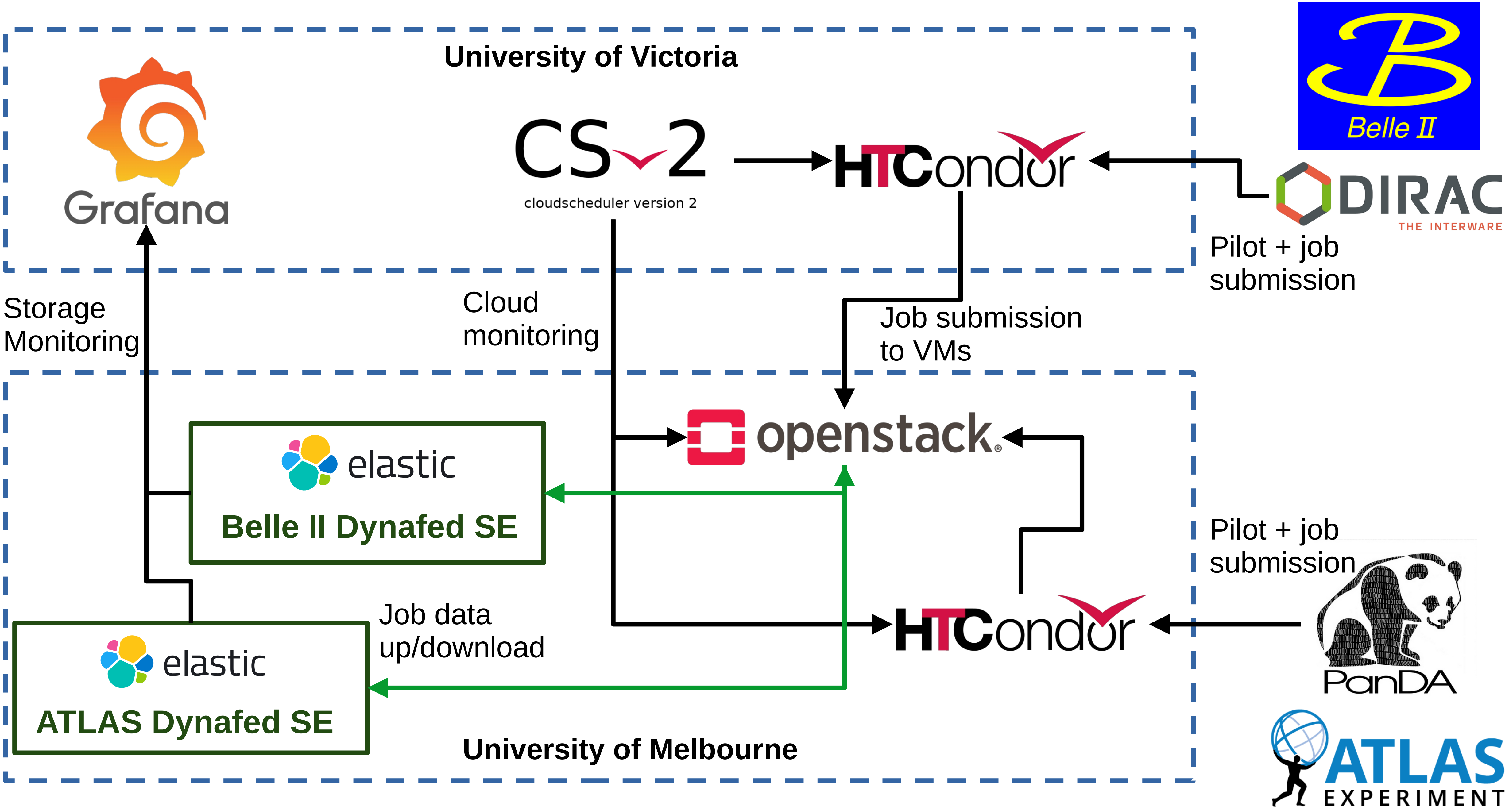
Belle II:

- Already uses our OpenStack VMs managed by U.Vic’s HTCondor and Cloudscheduler setup
- We will just add the preemptiable CPUs to the config

CS2
cloudscheduler version 2

ATLAS:

- Currently testing our own HTCondor manager host in our cloud
- Still use U.Vic’s Cloudscheduler to monitor, configure quotas, and dynamically start/stop VMs on our cloud
- Will allow us to respect Belle II and ATLAS quotas while opportunistically using any extra or preemptiable CPUs



University of Victoria

Grafana

CS2

cloudscheduler version 2

HTCondor

Belle II

DIRAC

THE INTERWARE

Pilot + job submission

Storage Monitoring

Cloud monitoring

Job submission to VMs

openstack

elastic

Belle II Dynafed SE

Pilot + job submission

Job data up/download

HTCondor

PanDA

elastic

ATLAS Dynafed SE

University of Melbourne

ATLAS EXPERIMENT

Summary

We are replacing our Tier 2 with cloud resources while attempting to re-use as much institutional resources from Melbourne and Victoria as possible

Grid job submission to the cloud is relatively simple, but the centrally managed storage creates a lot of restrictions on which grid technologies are feasible

We are commissioning our Dynafed S3 storage with Belle II now and in discussion with ATLAS storage experts

With our funding for manpower we believe we can keep Dynafed supported beyond 2024 if it stays as our main solution

We can investigate what it would take in funding and manpower to get our own Ceph so that we could have the option to move to dCache/XRootD