An implementation of cloud-based grid CE and SE for ATLAS and Belle II

Jonathan Woithe ¹ Martin Sevior ² Paul Jackson ¹ David Dossett ² Marcus Ebert ³

> ¹University of Adelaide, Australia ²University of Melbourne, Australia ³University of Victoria, Canada

CHEP2024, Kraków, Poland October 2024

Outline

Motivation

Intrastructure description

Servers (storage, compute) Storage Compute

Current status

Benchmarks

Challenges

Future plans

Conclusions

Research Computing groups at our institutions now provide and maintain hardware and access through cloud platforms

 \rightarrow Need to fit in with what our universities provide

Industry standard interfaces

→ No esoteric filesystems are exposed to cloud users

Exploit economies of scale in commercial cloud resource providers for Grid computing.

Compute and storage can be easily increased as funding allows and demand grows.

Melbourne Research Cloud (MRC) VMs

Orchestration by OpenStack

Server configuration managed by Ansible, tracked in git

▶ 750 TB of S3 compatible object store from MRC

- Not a traditional filesystem
- Each "file" is an object in a database
- The object's "key" is interpreted as its filesystem path
- No explicit objects for filesystem directories

Currently use a single bucket for flexibility

- Belle II and ATLAS have separate key namespaces
 - Gives illusion of separate top-level directories
- Transports: root, davs, https
- Enabled by the xrootd-s3 work at SLAC (https://cds.cern.ch/record/2857626/files/ATL-SOFT-SLIDE-2023-125.pdf)

Storage

- Xrootd redirector VM
 - Authenticates incoming requests
 - Generates access token
 - Redirects requests to one or more proxy servers
- Xrootd proxy server VM
 - Validates access token
 - Serves requested resource
 - Currently have 1 proxy server
 - Can deploy more when bandwidth requirements increase



Storage Resource Reporting (SRR) json file

- Defines Belle II and ATLAS storage shares, space usage and capacities
- Generated hourly by python script on primary proxy server
- Boto3 library used for S3 access
- Adler32 checksums
 - Managed on primary xrootd proxy server
 - Maintained with python script using boto3 library for S3 access
 - Calculated on first request, stored as metadata attribute on S3 object



Third Party Copy

- Executed on xrootd proxy servers
- Uses bash shell script to interface with xrootd
- root:// transfers: xrdcp streams content from source, s3cmd sends content into S3 object
- davs:// transfers: handled by libXrdHttpTPC.so: http.exthandler xrdtpc libXrdHttpTPC.so

- Slightly different architectures used for Belle II and ATLAS
- Cloud resources managed by Cloud Scheduler v2 (CSv2) instance at UVic (https://csv2.heprc.uvic.ca)

Compute



Compute Belle II

HTCondor host VM at UVic

- Jobs submitted to HTCcondor host via local DIRAC site-director (Belle-II still uses GSI, no HTCondor-CE is involved)
- ► CSv2 monitors HTCondor, starts HTCondor worker VM in MRC if needed
- ▶ Worker node set up via cloud-init as configured in CSv2
- Worker node registers with HTCondor when ready
- HTCcondor runs job on appropriate VM
- CSv2 shuts down worker VMs that remain idle for too long

A VM in MRC OpenStack hosts HTCondor and HTCondor-CE instances

- Host is running AlmaLinux9
- Token authentication is supported
- Jobs submitted to HTcondor-CE on the HTCondor host
- After authorisation, jobs passed onto HTCondor by HTCondor-CE on same host
- CSv2 processes proceed as for Belle II

Belle II storage is operational (400 TB)

Belle II compute is operational (900 vCPUs)

Current status Belle II - TPC matrix

DESTINATION																												
	CNAF-TM SE	DESY-TMP- SE	IN2P3CC- TMP-SE	KIT/TMP- SE	CESNET- TMP-SE	IPHC-TMP- SE	LAL-TMP- SE	Kisa-TMP- SE	ROMA3- TMP-SE	Francati- FMP-SE	Napoli- TMP-SE	CIFRONET TMP-SE	CLAKBIM TMP-SE	KEK- DEV- TMP-SE	SIGNET- TMP-SE	Tariao- TMP-SE	MPPMU- TMP-SE	HEPHY- TMP-SE	NTUCC- TMP-SE	UNIO- BAWTMP- SE	UVIO TMP-SE	TAU- TMP-SE	BNL- TMP-SE	NDU- TMP-SE	IHEP- TMP-SE	KISTI TMP-S	Australia- TMP-SE	RL.
CNAF-TN SE	7.	FINISTHD	FINISEHD	FINISEHD	FINISEHD	FINISEHD	Pall FINISHED	FINISEHD	UNISEHD	FINISEHD	FINISEHI	PMI FINISHED	FINISEHD	FINSTHE	Pall FINISHEE	FINELHE	FINISTER	Pad FINISHE	FINISERI	Pad FINISHED	FINISEHD	Pati FINISHET	FINISEHE	FINISEHE	Pull FINISHED	FINISE	Pall	i ace
DESY-TM SE	FINISTHD		FINISEHD				Pash FINISHED	EINISEHD				FINISEHD			FINSTRE	FINISTEE		FINISEH	FINSTH	FINISTHD		FINISEHI			FINISEHD	FINISE	FINISEHID	ROF
IN2P3CC- TMP-SE	FINISTED	FINISTERD		INISEHD			TINISTRE	TINISTRID																		TINISE	STREET OF	100
KIT/TMP SE	INSING				FINISEHD																					TINISE	FUNDATION	100
CESNET- TMP-SE	FINISIND					FINISHID																				TININ.	TABLE IN	100
IPHC-TM SE	FINELIED						Pash FINISHED	INISEHD														Pall FINISHEI	FINISTER			TIMM	TABLIC	100
LAL-TMP SE	FINISTED							FINISEHD				FINISHID										Pall FINISHEE	INSIDE			Push FINISH	SAMPLED	100
Nu-TMP SE	FINISTED						Pali FINISHED		FINISEHD			PMI PENISHID	FINISTHD					PAR FINISHER	FINISTHE	Pat FINSHID	FINISEHD	PAR FENENTER	FINISERI		Pall EINISTEED	FINISE	PMI FUNISHED	ROH
ROMA1- TMP/SE	FINISTHD						Pall FINISHED	EINISEHD		FINISEHD	FINISEHI	FINISHED	FINISTHD					PAI FINISHEI	FINISTHE	P44 FINSHID	FINISEHD	PAR FINISHER	FINISEH		Fall FINISHEED	UNISE	PAIL FUNISHED	100
Francati- TMP-SE	FINISTEED										CINISEHI															rinisi	CONSERIO	100
S Napeli-Th SE	FINSUAD											FINISEHD	FINISTHD	FINISTIA								FINISEHI		FINISEH	FINISEHD	rinisi	CONSCRED	100
O TMPSE	FINEDAD												FINISTHD	PAR								PAR FINISHER	TINISEH	Pash FINISHEE	FINISEHO	rinisi	TRANSFILD	100
U TMP-SE	INSTRO	FINISTRED	FINISTERD	FINISEHD		PENISIHID	TNISTRO	FINISIEID	TINISIEIID			FINISTHD		FINISTIE				rasan	111225110	FINETID		FEMALIE	21NISTER		INMERI	TINIS	TRAVELO	100
R TMP-SE	FINISTED	FINISTERD					FINISHED	TINISTERD				FINISHED	FINSTHD		FINISTER	WINISTIN		Pat FINISHEI	FINSTH	FINSHED	UNISERE	EMSHE	FINISTING	EDMANDER	FINISHED	TINISI	FINISHED	100
C TMP SE	FINISHED	FINISTERD	FINISTERD				TINISTIND	ENISTER				TRANSFILD	FINISIHD			T INIS I DE	OF INIS LINE	FINAL	PERSIN	FINSHID		ENISHE	TINISTING	FINISHEE	INNER	TINISE	CONTRACTOR OF	
SE E	INSUE	FINISTREE					FINISHED	TINISTERD				FINISHED	FINISTIID				# 1NES 1.340	FINISHE	FINISIHI	FINSHED	UNISTHE	FINISHEE	FINISTER		FINISHED	120151	FINISHED	100
IMPSE	FINISHED	FINISTERD			FINISEHD	PENISEHD		TINISTERD	TINISIEND	FINISEHD				FINESTIC		NIST IN		PINSIH	IT INSTITU	FINETHD	TRADEHO			INSTITU		FINISH	TEMELIND	
IMPSE	FINEHID	FINISTERD			TINISIED	FINISEHD		INBIED	INISIED	FINISEHD				FINEHIT		TINISHIT	FINISTHE		FINSHI	FINSTHD	TASIES	CREOK	FINISEHE	INISIGE	FINISEHD	TINISI	FINISE HD	100
TMPSE	ERROR																			ERROR	TASIE	CIMINITI	TINISEH			TINISI.	A INTRE HID	100
RANTMP	INSIN	FINISTERD	FINISISID	INSERD	FINISEHD	PENISIDID	TINISTREE	TINISTERD	TIMISTIND	FINISEHD	FINASALHI	FEREND	FINSTID	FINISTIA		W 15125 I J 02	07 INES I 242	PENSIN	TINSIII		FINISTER	Pall FINISHEE	, TINISTING	FINISIER	UNISERE	CIN154	STREETING	100
UVie-TMI SE	FINISTHD																	PAI FINISHEI	FINISTHE	FINSIND		IND FINISHED	FINISEH			FINISE	FINISEHID	ROF
TAU-TMP SE	Pash FINISHED	FINISEHD				Pusa FinishilD	ERROR	Fush FINISHED	Fush FINISHED	FINISEHD		ERROR		Pash FINISHII	Pash Pash PINISHLE	Pash WINISHII	FINISTHE	ERROR	Pash FINISHEE	ERROR	Pana FINASHEE		FINISEHE	Fush FINISHEE	ERROR	FINISE	ERROR	ROP
BNL-TMF SE	Pall FINISHED	Pall FINISHED	FINISHED	FILL	Pdl FINISHED	Pull FINISHED	Pall FINISHED	FINISHED	FINISHED	Pdl FINISHED	Pull FINISHEE	Pall FINISHED	Pall FINSHED	Pall FINISHEE	Pall	Pall FINISHEE	Pall	Pall FINISHE	Patt	Pail	Pull FINISHED	Pall FINISHEI	,	Pull FINISHEE	Pull FINISHED	Pull FINISH	Poll FINISHED	100
NDU-TMF SE	INSIAD	FINISTREE	TINBLED	INSERD	FINISERD	FINISHID	Pall FINISHED	TINISTERD	TIMISTIND	FINISIEIID	FINASILIU	Pall FINISHED	FINISTIID	FINISTIC		WIND IN	DU INIS I DE	Pall FINISHE	FINSIN	Pail FINISHED	FINISILIU	Pall FINISHEE	TINBEE		Pull FINISHED	FINISI	Pull FINISHED	101
THEP-TM SE	FINISHED	TINISTAD	FINISISID				FINISIAD					FEREND										Pall FINISHEE	TINISTIN			EULOR	UNIMERID	100
RISTON	18						Pall					Pall					Pall	Pad									UNISERIO	
Australia IMP/SE	INBIND						TINISTIND	Push FINISHED	Fush FINISHED	TINISERID		TIMESIND		Pash FINISHEE			TNBIN	TIMESTIC				Pall FINISHEE	INBER					
LINESE.																											ERROR	

(from https://people.na.infn.it/~spardi/tpc-davs-latest.html)

Current status Belle II - Compute jobs, running and idle



Current status ATLAS

ATLAS storage is ready for production testing (350 TB)

- ATLAS compute is being finalised (200 vCPUs initially)
 - HTCondor-CE accepts local job submissions and passes remote token-based access test at https://novastore.farm.particle.cz/cgi-bin/condor.cgi
 - All CSv2 processes work
 - Jobs are run by VMs as required
 - Remote HTCondor-CE access is being debugged
 - Looking to add additional vCPUs to ATLAS pool

Benchmarks

	Within cloud	In Australia
davs:// read	108 MB/s	40 MB/s
davs:// write	123 MB/s	74 MB/s
Checksum calc	3.2 s	3.4 s
Checksum fetch	0.72 s	0.98 s
s3 read	213 MB/s	n/a
s3 write	165 MB/s	n/a
root:// read	6.6 MB/s	5.9 MB/s
root:// write	132 MB/s	70 MB/s

Read/write tests used gfal-copy, checksum tests used gfal-sum. s3 tests on xrootd proxy server. Results are the average of 5 tests, each using a 1 GB test file.

Invisible application firewalls

Slow root:// read

Read/write speed variability, particularly outside Australia

AlmaLinux 9 environment

Future plans

Resolve remaining issues with ATLAS compute infrastructure

- Bring ATLAS SE and CE into production together (the approach preferred by ATLAS)
- Monitor production transfers for Belle-II and ATLAS, add extra proxy servers as needed
- Increase storage and compute resources as funding allows.
 - Tentatively planning for an additional 1 PB in 2025, mostly directed towards ATLAS
 - Add 1000 vCPUs to ATLAS pool

- ► A grid site using cloud storage and compute is feasible
- ► The "Melbourne" site is in production for Belle-II (CE and SE)
- The "Melbourne" site is expected to also provide CE and SE resources for ATLAS soon