Singularity deployment

A. Forti

GridPP43

28 August 2019

Use Cases

- In the system most of these use cases are satisfied only if the pilot runs the container
  - In particular payload isolation

Id	Use case
1	Installation of different OS from SL/RHEL /CentOS
2	OS upgrades don't need coordination with experiments anymore
3	Minimal installation on the nodes if sites prefer
4	Allows experiment to run tests with specific software or setups
5	May offer another approach to software distribution to sites that don't support CVMFS
6	Reduces the impact of ATLAS software on large shared file systems on HPC resources
7	Payload isolation
8	User containers
9	GPUs
10	Benchmarking suite

Requirements

- At CentOS7 sites singularity is a **requirement**.
  - Migration mostly completed
- Pilot2 deployed
  - Deployment mostly completes
- Singularity: **2.6.1 or 3.2.1+**
  - **Default configuration works** for ATLAS
  - Overlay/underlay enabled
    - User NS

PanDA Pilot 2

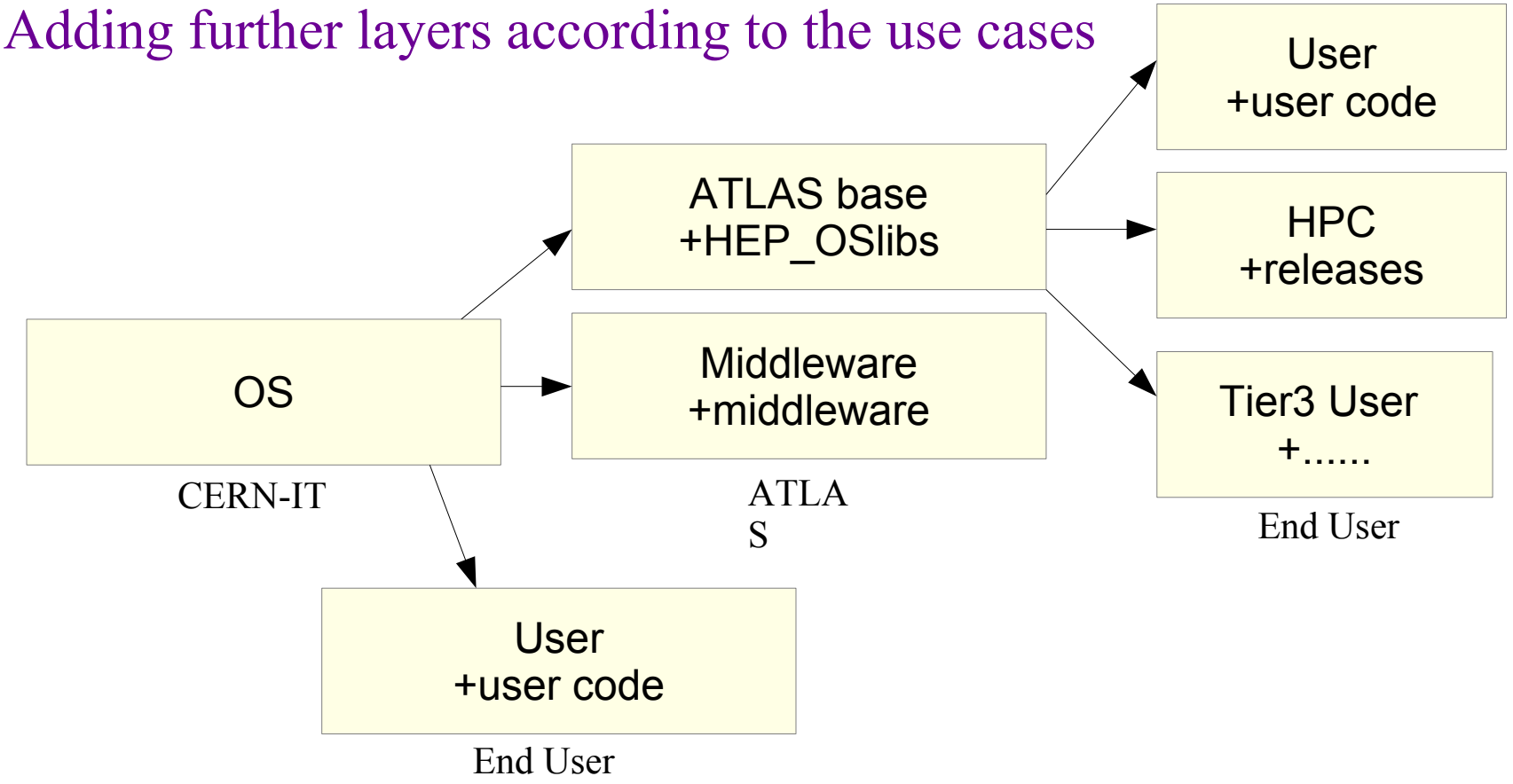
CentOS 7

Singularity container

2.6

Images

- All ATLAS images are docker images
- They are built as a hierarchy of docker layers
- CERN-IT OS image is the root layer
  - Adding further layers according to the use cases



Singularity in CVMFS

- Deploying singularity at sites like anything is a pain
  - Variety of installations and configurations
- OSG and ATLAS added singularity to CVMFS repos
  - Singularity deployment from CVMFS doesn't require sites to install the rpm
  - Caveat is that sites MUST enable user NS because the executable in /cvmfs will not use setuid for this to work.
- Agreed also at the WLCG container meeting.
  - Experiments agreed not to have a common executable to avoid to break each other workflows.
- Doesn't help other groups like SKA though

Containers

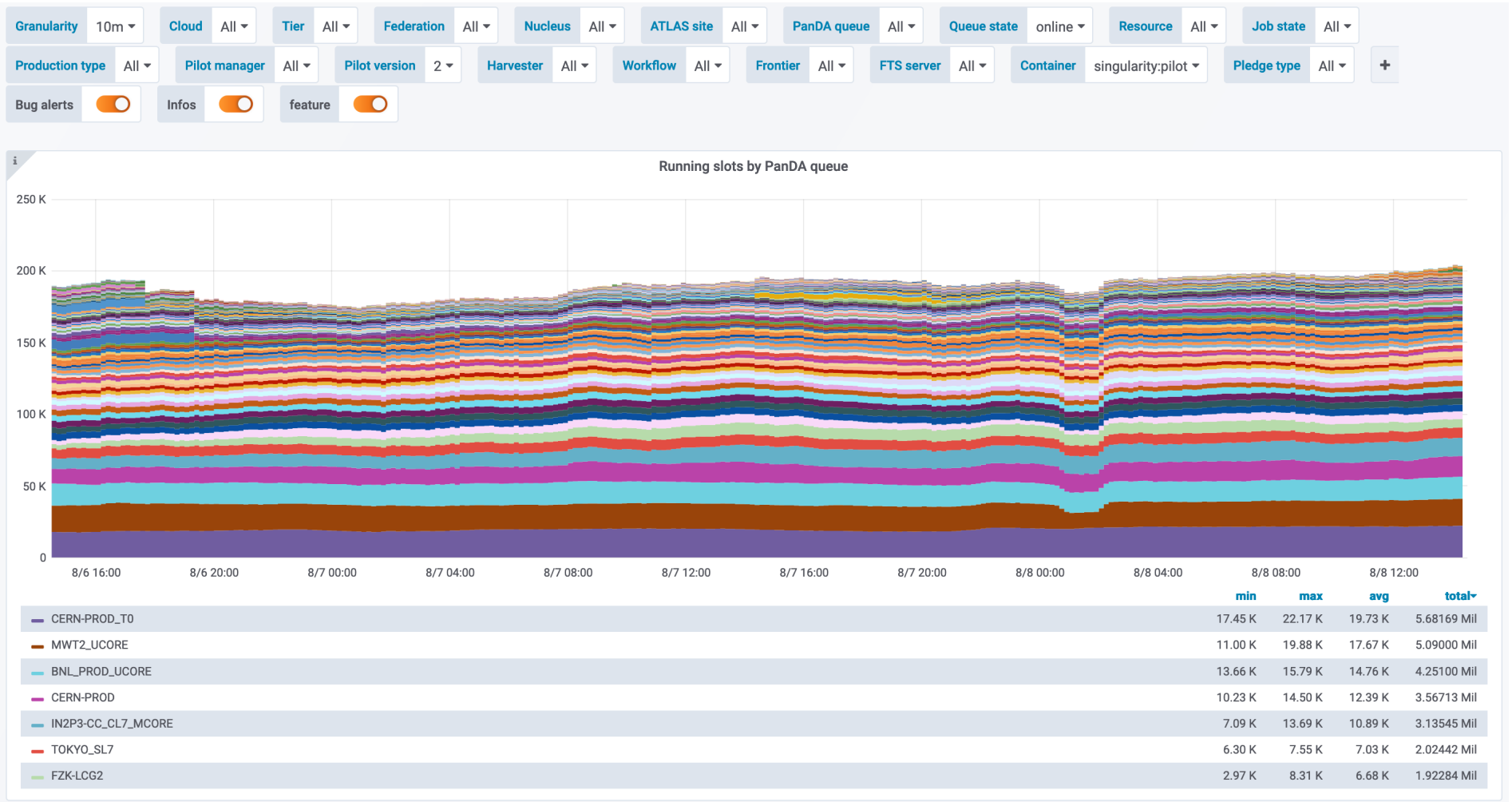
- ATLAS uses containers in 2 ways
  - Basic OS containers using the software from CVMFS
    - Production and analysis
    - Transparent for users but doesn't preserve the analysis
    - Distribution via cvmfs
  - Standalone containers software included in the containers
    - Analysis and production at HPC
    - Satisfies the preservation use case
      - requires users to build the containers or at least a layer
    - Distribution via registries or cmvfs

CVMFS containers

- Both production and users workflows
- Depend on pilot2 commissioning
  - Need container_type parameter set
  - Unpacked in /cvmfs/atlas.cern.ch/repo/containers/fs/singularity
- Doesn't need setuid, or overlay
- Limited to use software in cvmfs
- Containerization transparent to user and system administrator
  - Process tree similar to standard job

```
-runpilot2-wrapp ./runpilot2-wrapper.sh -q ANALY_MANC_TEST_SL7 -r ANALY_MANC_TEST_SL7 -s ANALY_MANC_TEST_SL7 -j user -d
  └─python pilot2/pilot.py -q ANALY_MANC_TEST_SL7 -r ANALY_MANC_TEST_SL7 -s ANALY_MANC_TEST_SL7 -i PR -j user --pilot-user=ATLAS -d
      ├─bash -c...
      │  └─startContainer. ...
      │      └─action-suid /alrb/.bashrc
      │          └─shim-init                                  /alrb/.bashrc
      │              └─.bashrc /alrb/.bashrc
      │                  └─python -u -Wignore ./runAthena-00-00-12 -a sources.20132189.derivation.tgz -r ./ --trf --useLocalIO --useCMake -
      │                      └─sh -c...
      │                          └─python /cvmfs/atlas.cern.ch/repo/sw/software/21.2/AthDerivation/21.2.33.0/InstallArea/x86_64-slc6-gcc62-
      │                              ├─MemoryMonitor --pid 2192 --filename mem.full.AODtoDAOD --json-summary mem.summary.AODtoDAOD.json --i
      │                              └─runwrapper.AODt ./runwrapper.AODtoDAOD.sh
      │                                  └─athena.py -tt/cvmfs/atlas.cern.ch/repo/sw/software/21.2/AthDerivation/21.2.33.0/InstallArea/
      │                                      └─{athena.py}
```

CVMFS ctrs deployment



	min	max	avg	total
CERN-PROD_T0	17.45 K	22.17 K	19.73 K	5.68169 Mil
MWT2_UCORE	11.00 K	19.88 K	17.67 K	5.09000 Mil
BNL_PROD_UCORE	13.66 K	15.79 K	14.76 K	4.25100 Mil
CERN-PROD	10.23 K	14.50 K	12.39 K	3.56713 Mil
IN2P3-CC_CL7_MCORE	7.09 K	13.69 K	10.89 K	3.13545 Mil
TOKYO_SL7	6.30 K	7.55 K	7.03 K	2.02442 Mil
FZK-LCG2	2.97 K	8.31 K	6.68 K	1.92284 Mil

- >200k slots now run containerized payloads
  - pilot2+containers

Standalone containers

- On the grid user workflows only
- Triggered by the user with a command line option
  - Independent from the pilot version
- Uses custom user images from a docker registry
  - Analysis images
  - Machine Learning images
  - Official docker images
- Doesn't need CVMFS to run

```
prun --containerImage docker://alpine --exec "echo 'Hello World\!'" --tmpDir /tmp --outDS user.aforti.test.20190306141519 --noBuild --site ANALY_MWT2_SL7
```

PanDA ID Attempt# of maxAttempts#	Owner Group	Request Task ID	Transformation	Status	Created	Time to start d:h:m:s	Duration d:h:m:s	Mod	Cloud Site
4267387837 Attempt 1 of 3	alessandra forti	1471 17334486	runcontainer	finished	2019-03-06 14:22:36	0:0:00:42	0:0:03:22	2019-03-06 14:31:22	US ANALY_MWT2_SL7 online no active blacklisting rules defined
	Job name: user.aforti.test.20190306141519/.4267387837 #1								
	Datasets: **Out:** user.aforti.test.20190306141519.log.235213854								

How much used?

- Not much for now
- Analysis groups imposing preservation as publishing condition
  - May increase the pressure on users
- Machine Learning people want to use GPU resources
  - Still working on their containers
    - Docker ↔ singularity creating most problems

Joao Victor Da Fonseca Pinto	221319 18951042	runcontainer	failed	2019-08-26 00:07:10	0:0:01:21	0:0:09:26	2019-08-26 00:38:05	UK ANALY_MANC_GPU_TEST online GPU.Tests	1001	0 (1)	36.54	**trans, 1:** Unspecified error, consult log file
Job name: user.jodafons.data17_13TeV.Allperiods.sgn.probes_lhmedium_EG1.bkg.VProbes_EG7.mlp.ringer_v8_et2_eta0.r53/.4462664777#3												
Datasets: **In:** user.jodafons:user.jodafons.job_config_10sorts_10inits_v2 || Rucio link **Out:** user.jodafons.data17_13TeV.Allperiods.sgn.probes_lhmedium_EG1.bkg.VProbes_EG7.mlp.ringer_v8_et2_eta0.r53_td.267226108												
Joao Victor Da Fonseca Pinto	221319 18951042	runcontainer	failed	2019-08-26 00:07:10	0:0:00:52	0:0:09:43	2019-08-26 00:38:04	UK ANALY_MANC_GPU_TEST online GPU.Tests	1001	0 (1)	48.71	**trans, 1:** Unspecified error, consult log file
Job name: user.jodafons.data17_13TeV.Allperiods.sgn.probes_lhmedium_EG1.bkg.VProbes_EG7.mlp.ringer_v8_et2_eta0.r53/.4462664776#3												
Datasets: **In:** user.jodafons:user.jodafons.job_config_10sorts_10inits_v2 || Rucio link **Out:** user.jodafons.data17_13TeV.Allperiods.sgn.probes_lhmedium_EG1.bkg.VProbes_EG7.mlp.ringer_v8_et2_eta0.r53_td.267226108												
Joao Victor Da Fonseca Pinto	221319 18951042	runcontainer	failed	2019-08-26 00:07:10	0:0:00:52	0:0:22:52	2019-08-26 00:38:03	UK ANALY_MANC_GPU_TEST online GPU.Tests	1001	0 (1)	37.63	**trans, 1:** Unspecified error, consult log file

Standalone ctrs distribution

- Currently via registries
  - Analysis images 1000 pulls a day without problems
  - Many people worried in advance
- CVMFS team setup and common repository for WLCG experiments
  - User adds the image to a list and the system downloads and unpacks in CVMFS.
    - Image gets updated the system automatically updates the copy in CVMFS
    - Still a prototype
    - Developer will be back in September more discussion then
    - https://indico.cern.ch/event/790755/
  - Other ideas circulating to merge cvmfs and registries

SKA

- SKA runs standalone containers too
- Student in Manchester running on the grid real workloads
  - Studying the magnetic field analysing images
- Singularity integrated in user scripts instead of the pilot
  - Images downloaded from registries
    - Singularity registry abandoned in favour of docker
    - 10k pulls from docker without any problem in July/August
    - Now added the image to CVMFS at RAL
- Since move to docker grid usage ramped up

SKA cont

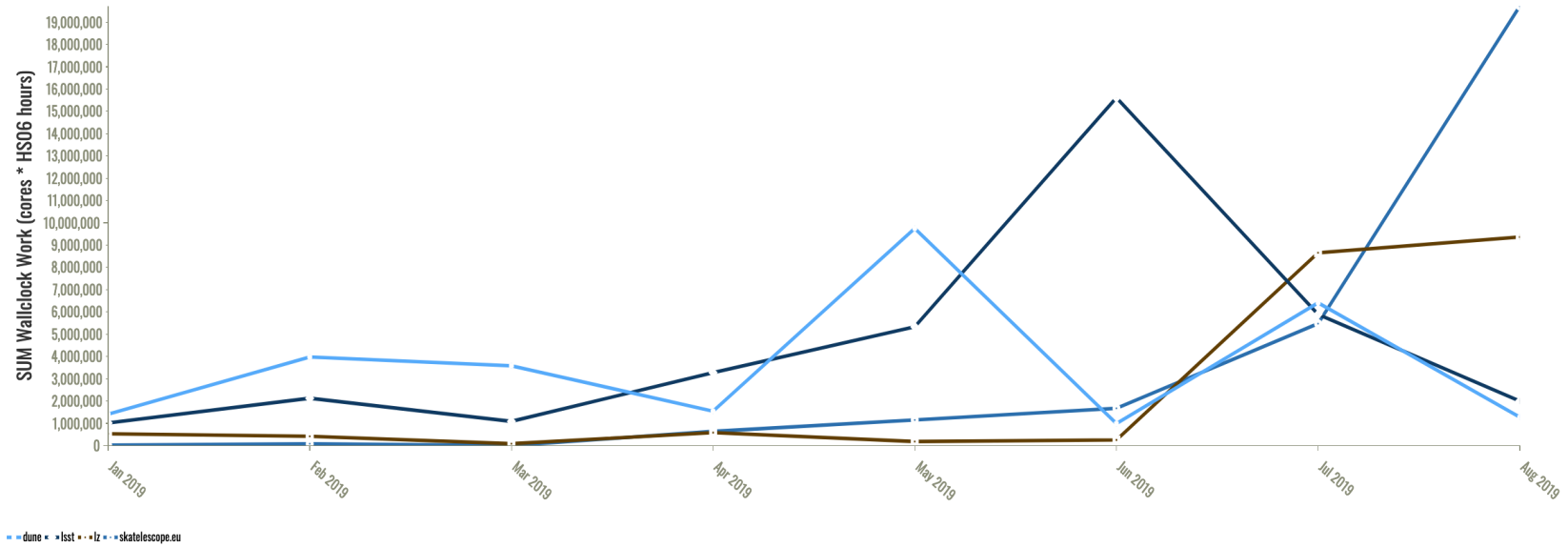
- Using Manchester until now but since last week extending to other sites
  - Now at the same level of usage as lsst, dune and lz

United Kingdom — SUM Wallclock Work (cores * HS06 hours) by VO and Quarter (Custom VOs)

VO	Jan 2019 — Mar 2019	Apr 2019 — Jun 2019	Jul 2019 — Sep 2019	Total	Percent
dune	8,975,252	12,285,721	7,689,627	28,950,600	25.36%
lsst	4,234,262	24,223,634	7,889,716	36,347,612	31.84%
lz	1,041,176	1,021,510	18,016,214	20,078,901	17.59%
skatelescope.eu	113,269	3,467,556	25,196,799	28,777,625	25.21%
Total	14,363,960	40,998,421	58,792,357	114,154,738	
Percent	12.58%	35.91%	51.50%		

1 - 4 of 4 results ‹ 1 › Number of rows per page 30

SUM Wallclock Work (cores * HS06 hours) by VO and Month



Conclusions

- Almost fully deployed on the grid in ATLAS using same model as CMS
- Standalone containers also being used not yet fully on the grid by ATLAS only a couple of users instereted in GPUs
- SKA is deploying instead and run several millions HS06 hours with them.
  - Trying to keep ATLAS and SKA with similar solutions