# Operational Intelligence @ Glasgow





#### **Outline**

- Glasgow Tier2 site
- Cluster Monitoring & Logging
- Visualization & Querying
- Alerts & Automation

Latest Addition: auto-update



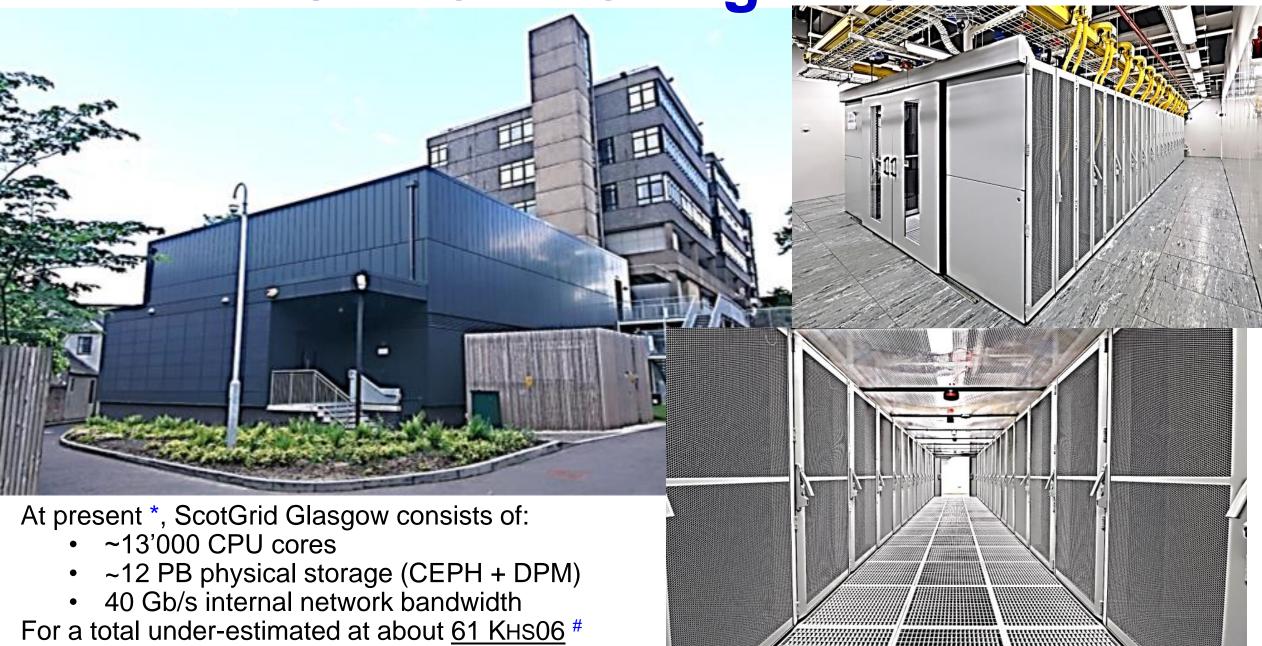




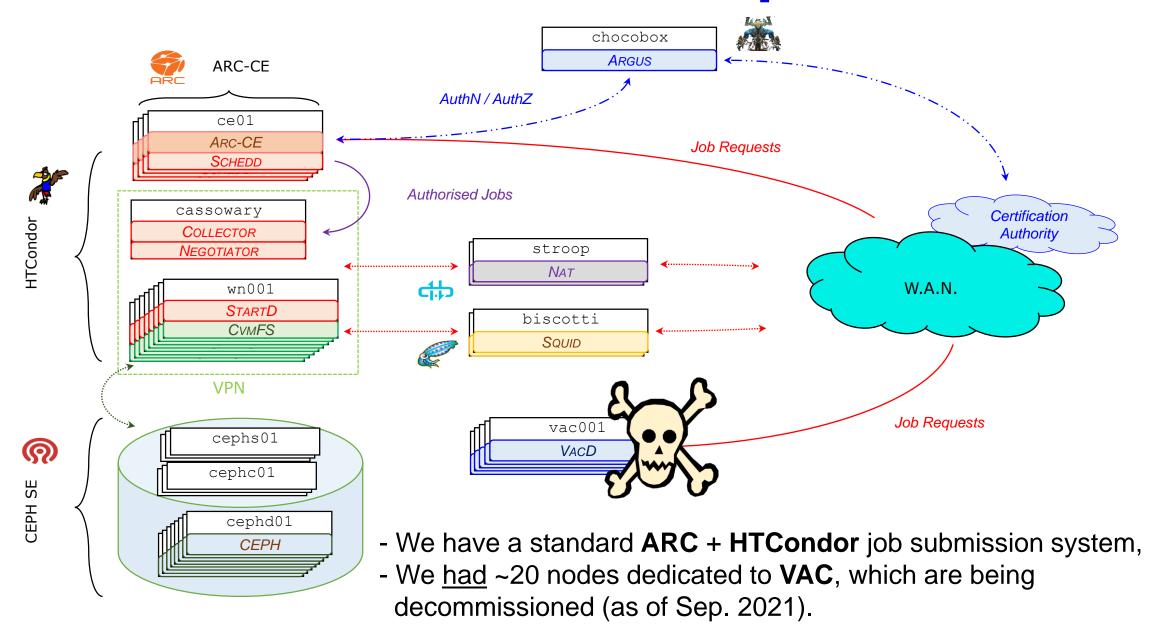
#### **ScotGrid Glasgow**:

Emanuele Simili, Gordon Stewart, Samuel Skipsey, David Britton

ScotGrid Glasgow Site



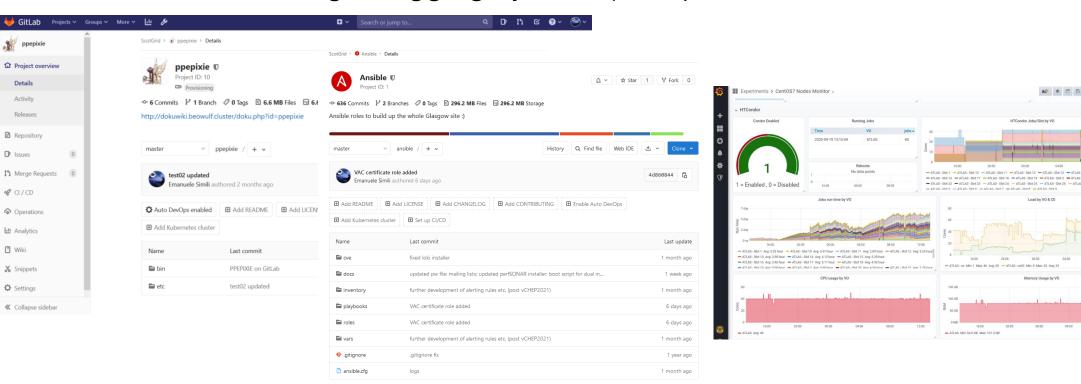
#### **Cluster Map**



#### **Automation Tools**

As a site of significant size, we use many sys-admin empowering tools:

- A network provisioning system (PPE PiXiE)
- A remote management tool (Ansible)
- A central monitoring & logging system (**PLG**) ... which I will talk about



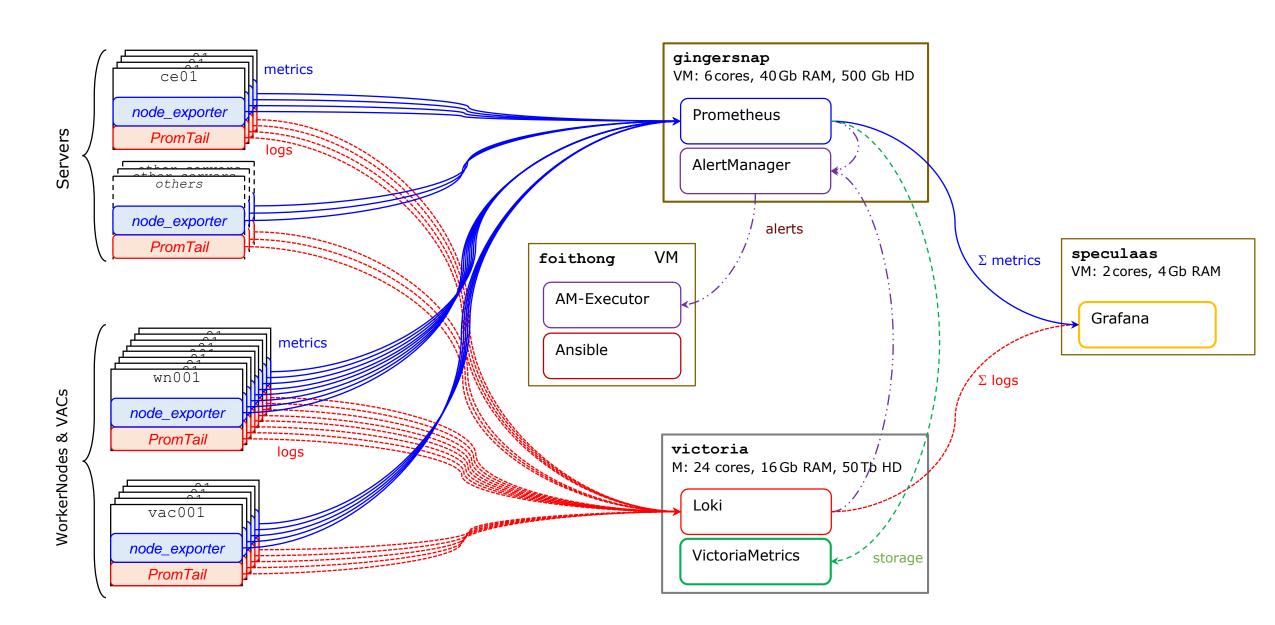
We have invested quite some time in preparing such environments because, in the long run, it has saved us more time than we spent making it.

# **Monitoring & Logging (1)**

Our system is built on **Prometheus**, **Loki** and **Grafana** (**PLG**)

- Metrics are exported by node\_exporter (installed on nodes) and collected by a central Prometheus instance
- Logs are exported by PromTail (nodes) and collected by Loki (central)
- Data visualization and querying is done by Grafana, which pulls data from both Prometheus and Loki
- Alert are sent out by AlertManager, installed on the Prometheus server, and received by AM-Executor, installed on the configuration management server (which also runs Ansible)
- Long term storage of metrics is done by VictoriaMetrics, installed on the same server as Loki to take full advantage of the large storage

# Monitoring & Logging (2)



### **Exported Metrics**

**node\_exporter** produces by default a large set of metrics (e.g., CPU load, network activity, disk I/O, system info) and it can be further customised with ad-hoc metrics:

Metric	Purpose	Tool
Temperature HD	check for overheat	S.M.A.R.T. tools
Temperature CPU	check for overheat	IPMI tools
HTCondor jobs stats	check the status and ownership of running jobs	condor_status
VAC jobs stats	check the number and status of virtual machines	check-vacd
Running processes	runtime and resource usage	ps + grep
Number of Reboots	static counter to keep track of unexpected restarts	cron job
Static info	check for unexpected changes in hardware (e.g., memory bank fail)	Iscpu, vmstat , Isblk

Metrics produced with any script must be written in a properly formatted text file and placed in a folder where **node\_exporter** can scrape them:

node\_exporter\_dir: "/var/lib/node\_exporter/textfile\_collector"

### **Exported Logs**

**PromTail** tails the content of local logs and sends it to a central **Loki**. We have chosen this tool (instead than the widely spread **ElasticStearch**)

because:

it well integrates with Grafana

it is very easy to set-up ...

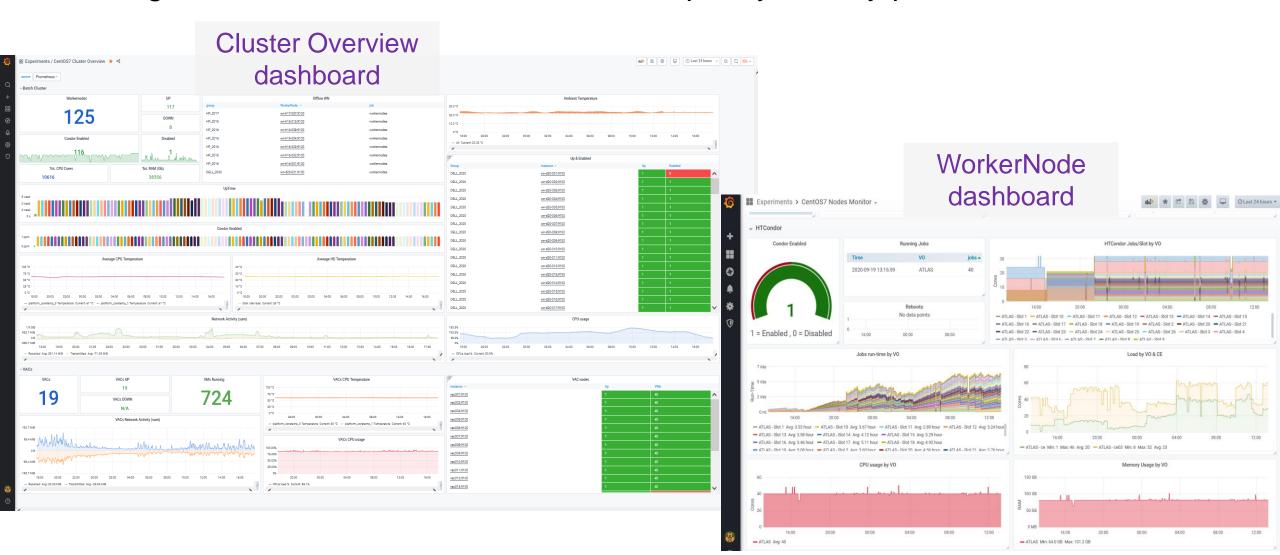
Logs to scrapes are specified in the configuration (we scrape default system logs\* and specific service logs)

<sup>\*</sup> we do not export the journal because it clogs the system

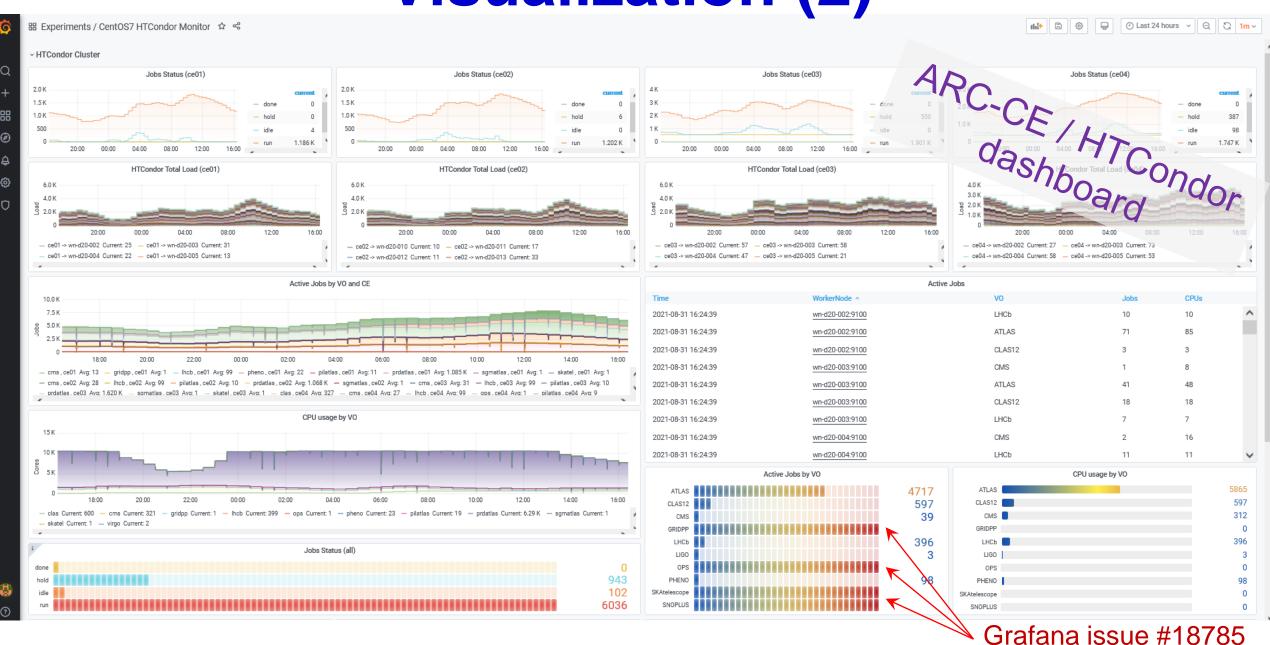
Service	Exported Logs		
all	/var/log/messages		
	/var/log/secure		
an	/var/log/cron		
	/var/log/yum.log		
HTCondor	/var/log/condor/		
ARC-CE	/var/log/arc/		
ARC-CE	/var/log/arc/bdii/		
	/var/log/argus/pap/		
ARGUS	/var/log/argus/pdp/		
	/var/log/argus/pepd/		
Squid	/var/log/squid/		
Apache	/var/log/httpd/		
Ansible	/var/log/ansible/		
Alisible	/var/log/alerts/		

# Visualization (1)

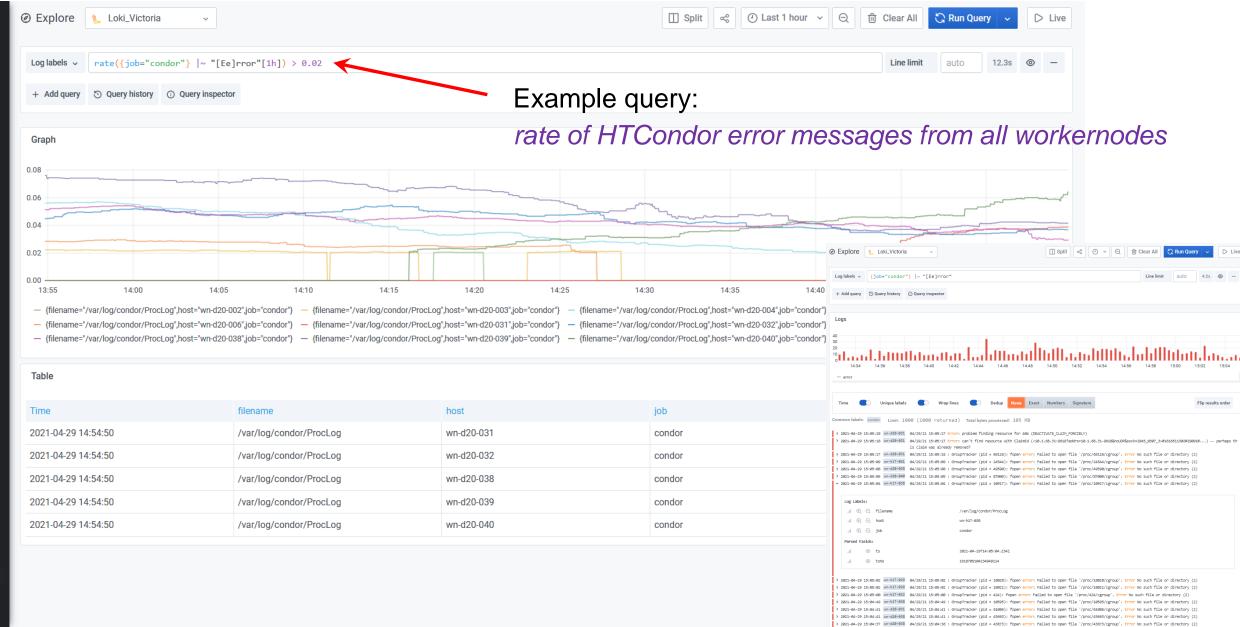
With a few custom dashboards and over 100+ graphs, it is easy to check the correct functioning of all machines from remote and to quickly identify potential issues.



Visualization (2)

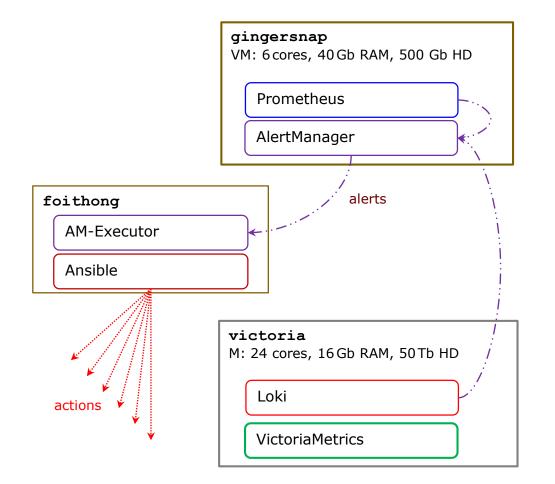


# Logs DB



# Alerts & Automation (1)

Alerts are generated according to the <u>outcome of conditional tests</u> involving metrics or queries against certain criteria. Both **Prometheus** and **Loki**\* can be configured to trigger alerts when a custom set of conditions is met.

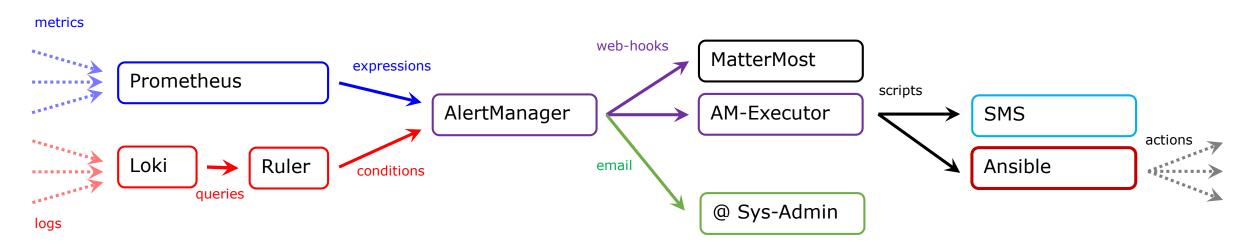


Alerts are handled and sent out by the **AlertManager** ...

... and are received by the **AM-Executor**, which performs some pre-encoded actions.

<sup>\*</sup> Loki's alert system is new in v2.0 and uses a specific plug-in (ruler), which in turns calls AlertManager ...

# **Alerts & Automation (2)**



In summary, these are the stages of an alert:

- 0) Prometheus or Loki verify a test condition and triggers AlertManager
- 1) An alert messages is generated by AlertManager and sent as a web-hook
- 2) **AM-Executor** is an HTTP server that listens for alerts, parses their content, and triggers custom commands matching the alert type and source
- 3) By running **AM-Executor** on the same management server that runs **Ansible**, we can harness the full power of Ansible remote management, with a variety of play-books prepared beforehand

#### **AM-Executor**

AM-Executor is configured with a custom script that defines the action to

be taken based on the alert:

- Actions implemented in a <u>case-switch</u>
- Multiple alerts handled in a for loop
- The script can call an Ansible playbook or trigger another script (SMS)
- Privileges for Ansible remote management are set by a password-less SSH key \*
- \* This is not the best practice, but so far it has been an ok solution
- \* New addition: jobs draining, (wait), custom update and reboot

```
# case-switch to assign an action
case $AM ALERTNAME in
        echo "> ansible-playbook -1 $AM NODE do cvmfsclean.yml -b" >> $AM LOGFILE
        echo "> ansible-playbook -l $AM NODE do shutdown.yml -b" >> $AM LOGFILE
        echo "> ansible-playbook -l $AM NODE do shutdown.yml -b" >> $AM LOGFILE
        echo "> ansible-playbook -l $AM NODE do clean-enable.yml -b" >> $AM LOGFILE
        echo "> ansible-playbook -1 $AM NODE do update.yml -b" >> $AM LOGFILE
        echo "> ansible-playbook do_shutdown.yml -b" >> $AM LOGFILE
        echo "> . /etc/am-executor/send_sms_alert.sh" >> $AM_LOGFILE
    *) echo "Alert not recognised: $AM_ALERTNAME !" >> $AM_LOGFILE
esac
```

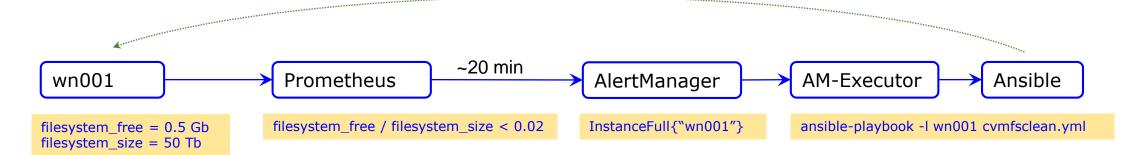
#### **Alert Conditions**

So far, only a limited number of alert-cases have been implemented:

Alert condition		Alert Name	Action	Mail Chat	SMS
$avg(node\_sensors\_ambient\{all\}) > 25$		TooHot	/	✓	✓
$avg(node\_sensors\_ambient\{all\}) \ge 30$		WayTooHot	shutdown{all}	✓	
$node_hwmon_temp\{node\} > 80$		CpuHot	shutdown{node}	<b>✓</b>	
$\lor$ node_sensors_hddtemp{ $node$ } > 35		HdHot			
$\frac{\text{node\_filesystem\_free}\{node\}}{\text{node\_filesystem\_size}\{node\}} < 0.02$		InstanceFull	clean{node}	<b>✓</b>	
avg(node_coi	$ndor_cpu\{node\})[1h] \le 1$ $condor_cpu)[1h] > 10$	NodeLazy	reboot{node} + enable{node}	<b>✓</b>	

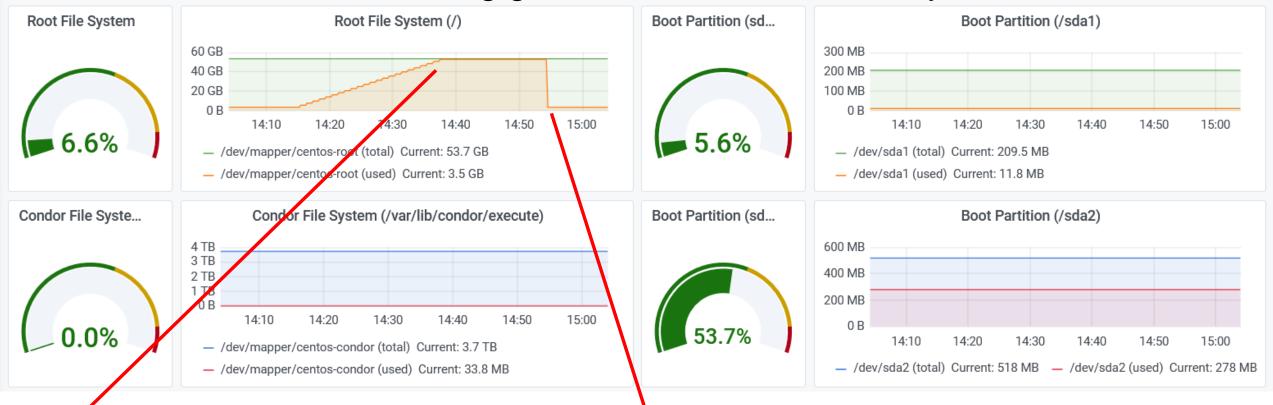
+ NodeDrain ...

#### Example (InstanceFull alert):



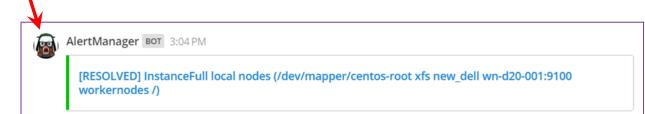
### **Example Alert & Solution**

See the InstanceFull Alert being generated and automatically resolved:

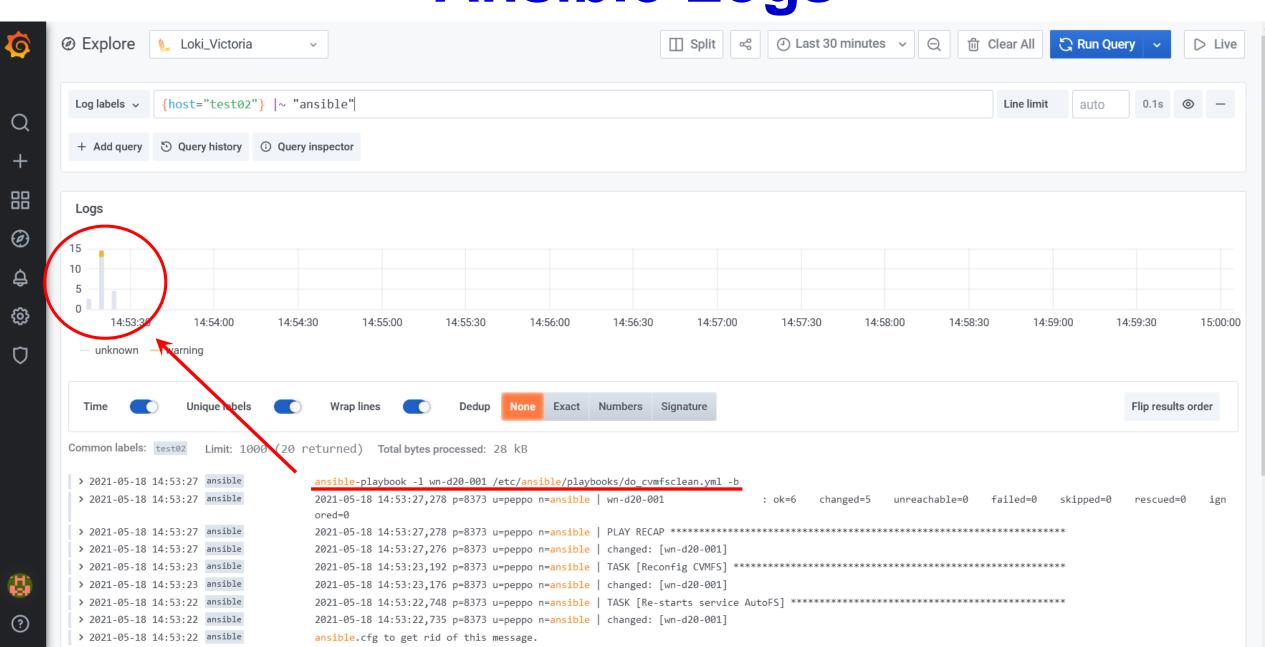


MatterMost messages (ClusterComms channel):





### **Ansible Logs**



> 2024 OF 40 44.F2.22 profblo

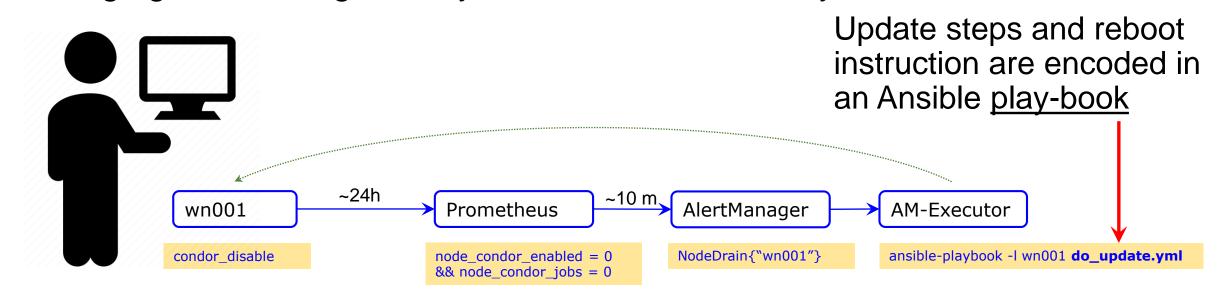
# Patching & Updating

Some maintenance tasks, such as HTCondor updates, Kernel updates or other CVE patches, do require the node to be rebooted.

In order not to loose running job, this is normally done in two steps:

- first the node is drained (HTCondor stops accepting new jobs)
- when all running jobs are finished (~24h), the node can be updated, rebooted and then put back online.

Leveraging the existing Alert system, this can be easily automated:



#### **Conclusions & Outlook**

The system is being evaluated since Jan. 2021 and it seems well-behaved \*:

- ✓ The alerting system can identify issues and send Alerts to site admins via dashboard/emails/SMS/chat.
- ✓ In a few, well known cases, it can trigger simple recovery actions that run automatically (clean HD, reboot, ...)
- ✓ The system can be exploited to cover maintenance tasks, such as installing patches and updates that require the node to be drained and rebooted
- ✓ It was presented at vCHEP and shared with CERN Operational Intelligence #
- Can be extended to identify complex alerts based on combinations of metrics and multi-step recovery actions (e.g., re-provisioning)
- ? Next idea: use Prometheus metrics for Anomaly Detection ...
- ? To be added: Networking metrics (sFlow) ...

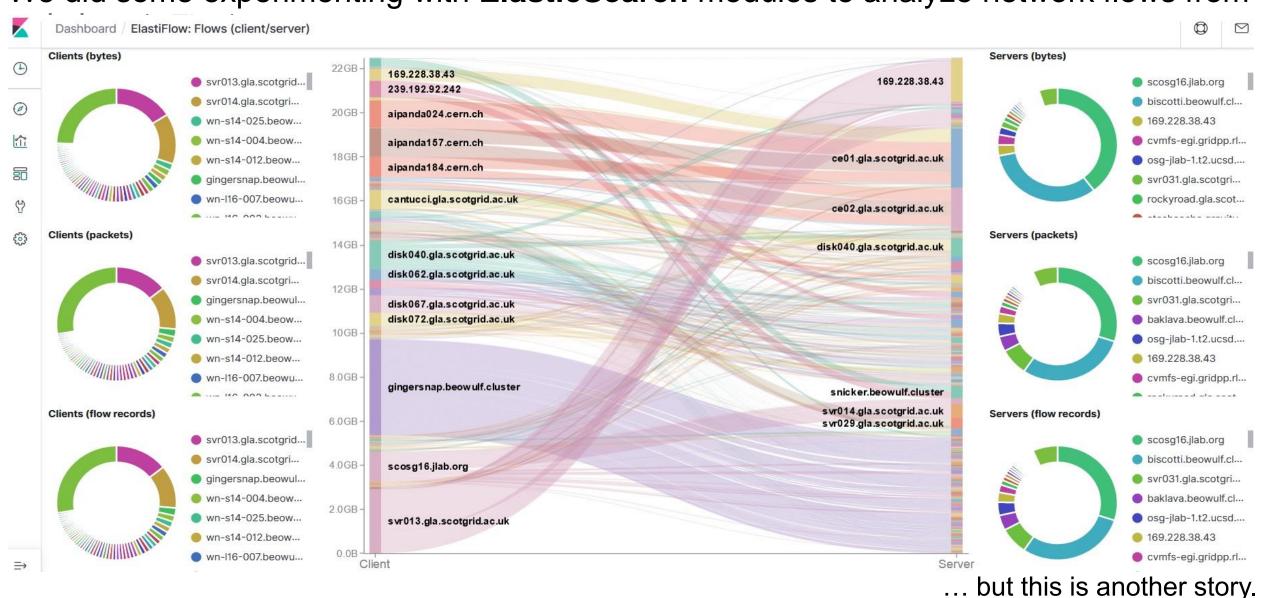
<sup>\*</sup> It is well known that giving full control to the machines without proper evaluation will eventually lead to

#### Thanks.



#### ElasticFlow

We did some experimenting with ElasticSearch modules to analyze network flows from



### **Example Procedures**

#### Create a new workernodes (assuming the node is cabled and plugged in):

- Fill in MAC, IP and HostName (in Ansible), then run the play-book to update the network vi roles/dhcp/files/hosts.conf roles/dnsmasq/files/10.0.0.0\_8 ansible-playbook update\_networking.yml
- Create a PXE system profile, then boot the machine to install OS vi systems/nodes/wn00x.yml systemctl restart pxecli
- Finally run ansible to install the software and configure the workernode ansible-playbook -I wn00x workernodes.yml

#### **Add a new VO** to the standard job submission system:

Fill in the VO coordinates in a proper Ansible role, then run it cp -r roles/vo-template roles/vo-xxx vi roles/vo-xxx/files/xxx.grid-mapfile ansible-playbook add\_vo.yml -extra-vars "vo=xxx"

#### Clean CVMFS cache and re-enable node 00x:

Just run the ansible play-book ansible-playbook -I wn00x cleanCVMFS.yml

#### Occasional Kernel or HTCondor update (needing a reboot):

> Drain jobs (24h), custom play-book, reboot, re-enable ...

# **Example Logging**

As an example, these are the bits in the **PromTail** configuration (promtail.yml) that export general and HTCondor logs:

```
# PromTail Configuration defines the rules for scraping local logs
 2
    -server:
                                                                                 ## Service Logs
       http listen port: 9080

    job name: condor

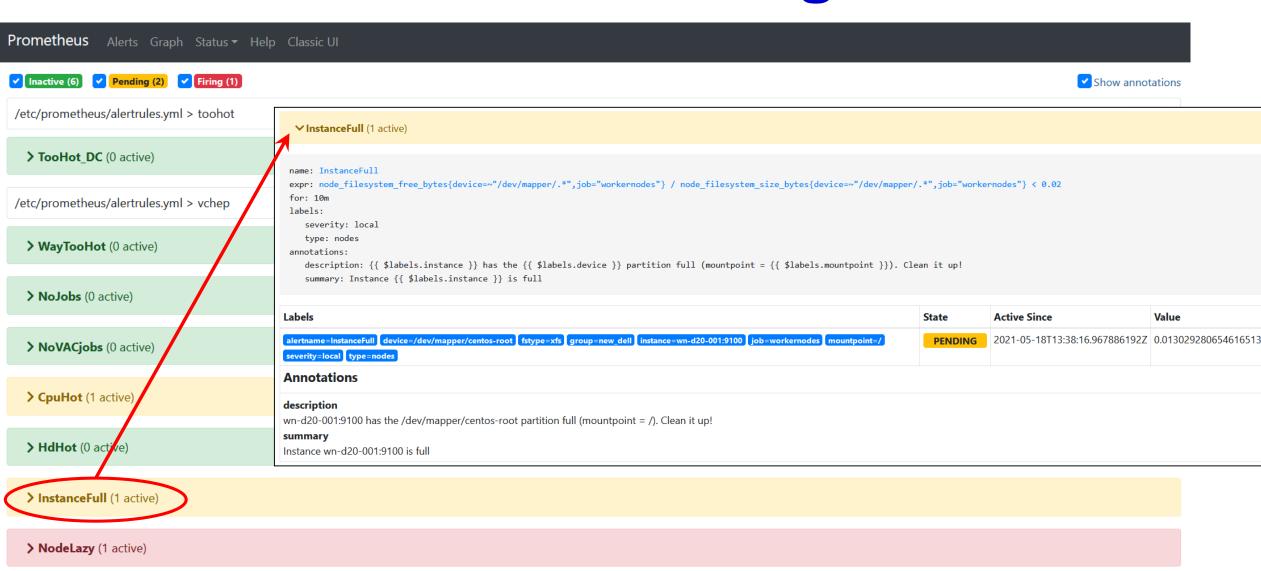
        grpc listen port: 0
                                                                         29
                                                                                   static configs:
                                                                         30
                                                                                      targets:
      # Positions
                                                                         31
                                                                                          - "{{ inventory hostname }}:9080"
    -positions:
        filename: /tmp/positions.yaml
                                                                         32
                                                                                       labels:
10
                                                                         33
                                                                                          iob: condor
11
      # Loki Server URL
                                                                         34
                                                                                          host: "{{ inventory hostname }}"
    -clients:
                                                                         35
                                                                                          path : /var/log/condor/*og
        - url: http://{{ central log server }}:3100/loki/api/vl/push
13
14
15
    -scrape configs:
16
        ## Common Logs
                                  # log messages (/var/log/messages)

    job name: messages

          static configs:
19
            - targets:
                - "{{ inventory hostname }}:9080"
20
              labels:
                                                                                       HTCondor logs:
22
                job: messages
                 path /var/log/messages
24
                                                                                         /var/log/condor/
25
```

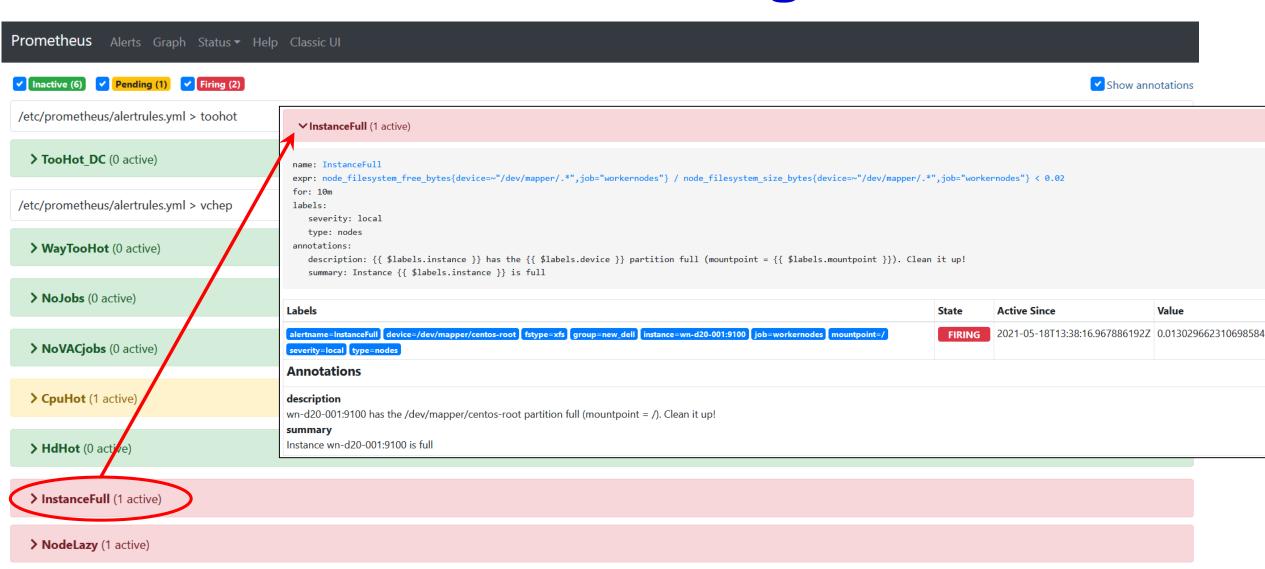
General Logs (system messages)

# **Alert Pending**



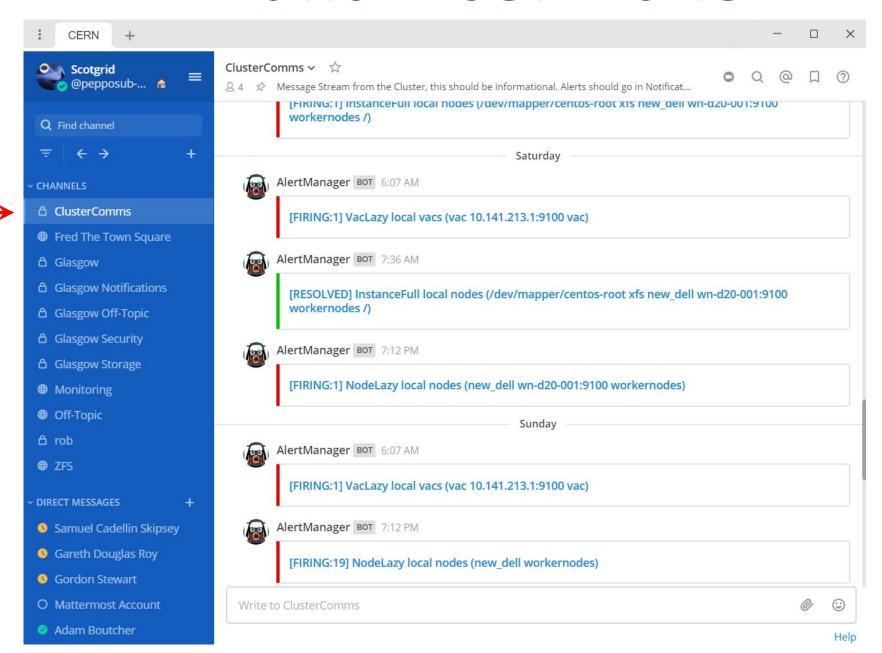
> VacLazy (0 active)

# **Alert Firing**



> VacLazy (0 active)

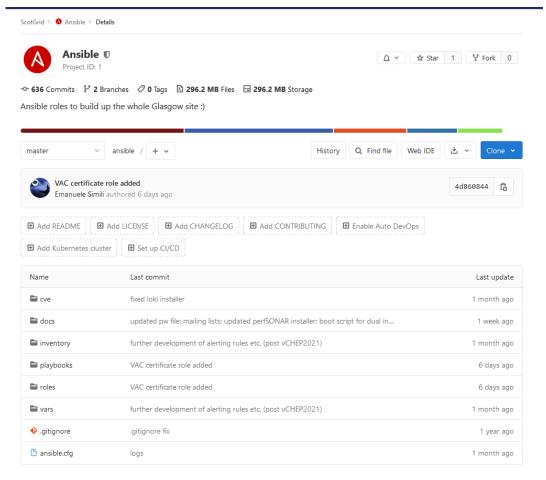
#### **MatterMost Alerts**



#### Git & Wiki

Ansible repository on our internal Git:

- ~70 Ansible roles
- ~50 Ansible play-books (incl. maintenance)



Internal wiki with detailed procedures and code snippets for cluster built: ~50 pages

