



How do we manage 40K machines in the CERN Computer centre

[ZHECHKA TOTEVA \(CERN/IT\)](#)

CERN – 17/07/2024

Outline

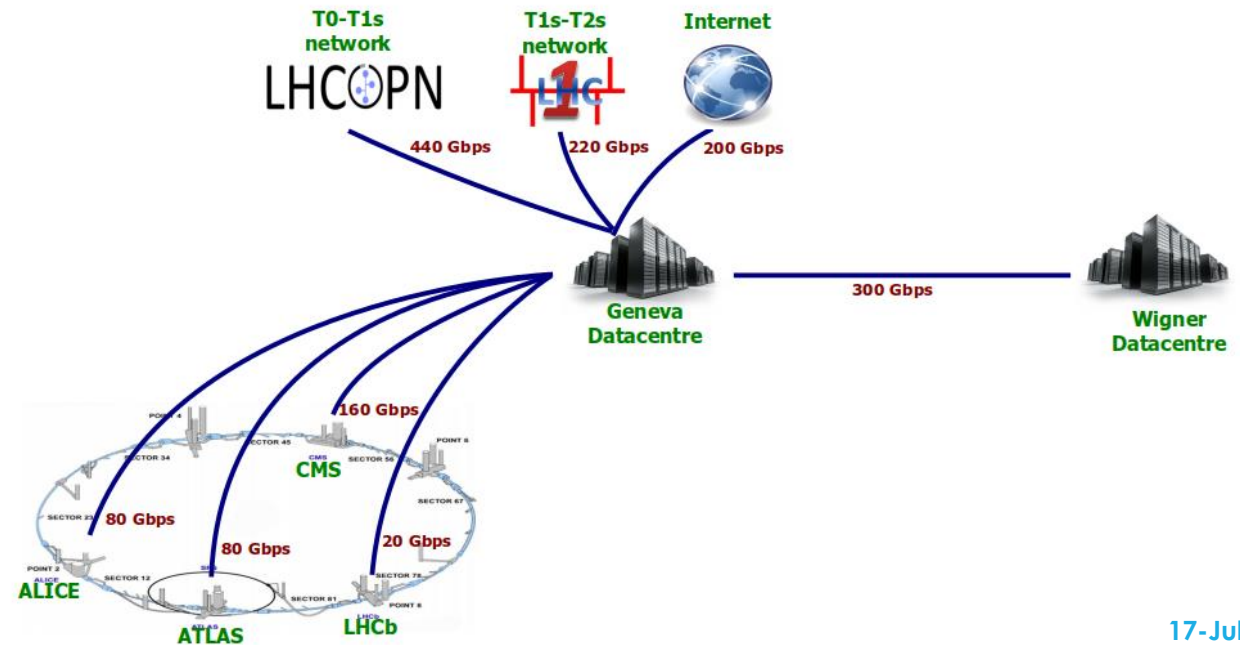
- ▶ CERN Computer Centre (CC) in numbers
- ▶ Overview of the CERN network and data storage
- ▶ Overview of electricity and cooling
- ▶ WLCG is couple of numbers
- ▶ Configuration management at CERN IT
- ▶ CERNMegabus@CERN
 - ▶ Architecture
 - ▶ Overview of major implemented use cases
 - ▶ CERN Computer Centre (CC) power cut management

CERN Computer Centre (CC) in numbers



Computing network

- ▶ 250 routers, 4100 switches, 1200 Wi-Fi points
- ▶ 35 000 km optical fibre (only ~5 000 less than the equator length)
- ▶ Wigner Data centre in Hungary
 - ▶ 1200 km distance
 - ▶ with three 100 Gb/s



Data storage

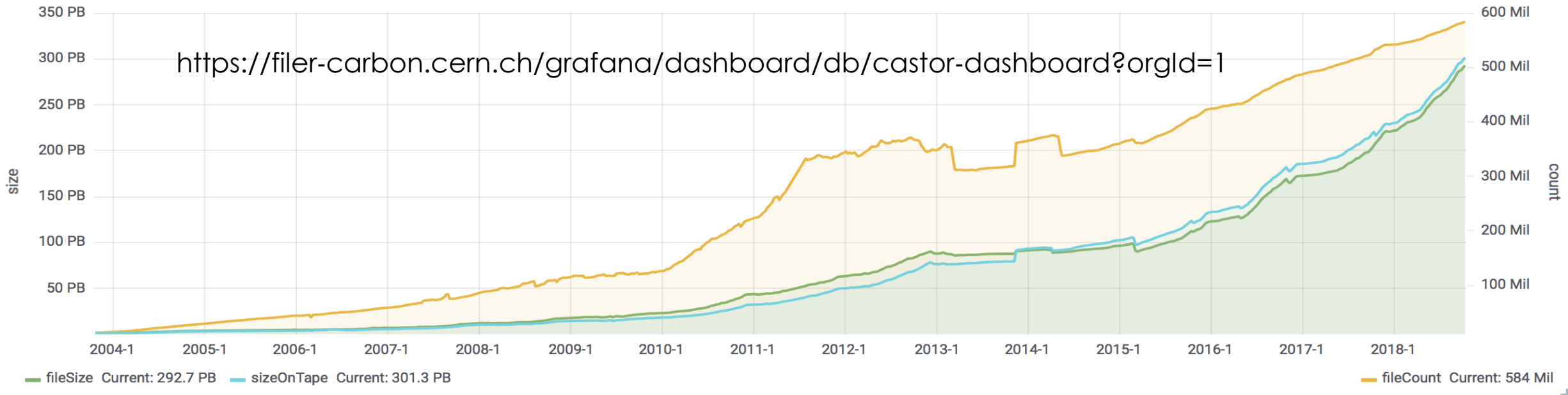
- ▶ 50 Pbytes / year from LHC
 - ▶ + 25 Pbytes / year from non-LHC experiments
- ▶ **RECORD:** August 2018: 13.8 Pbytes of data written on tape (of which 11.56 is LHC data)
 - ▶ More than 2 PB read/write daily
- ▶ Tape drives faster than disks; but slower in mounting (latency)
 - ▶ 90 K disk drives (of which 10-15% are SSD, providing less than 10% capacity)
 - ▶ SSDs are 5-10 times more expensive than spinning disks

www.cern.ch/eos

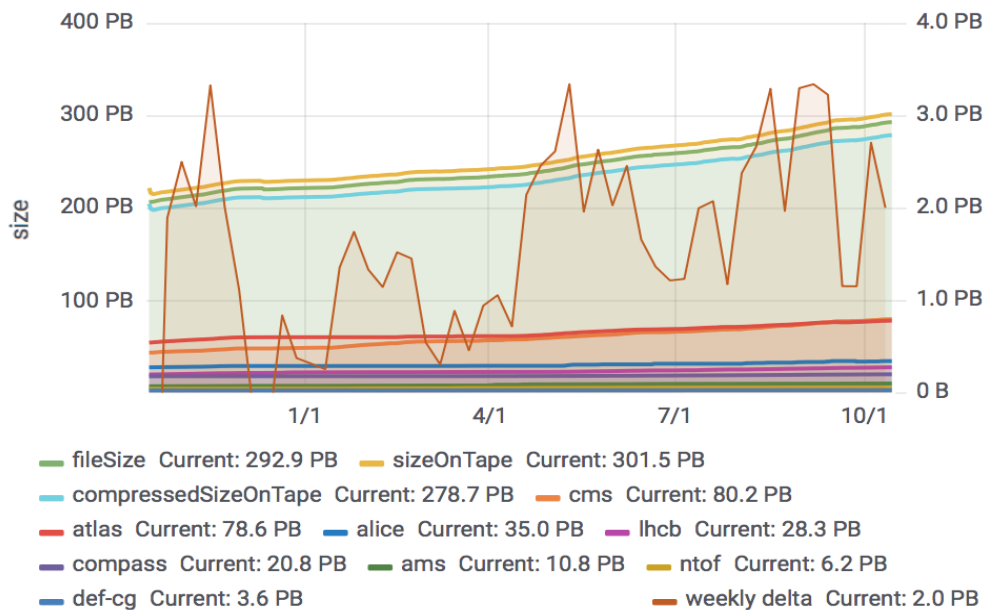
www.cern.ch/castor

Physics Data in CASTOR

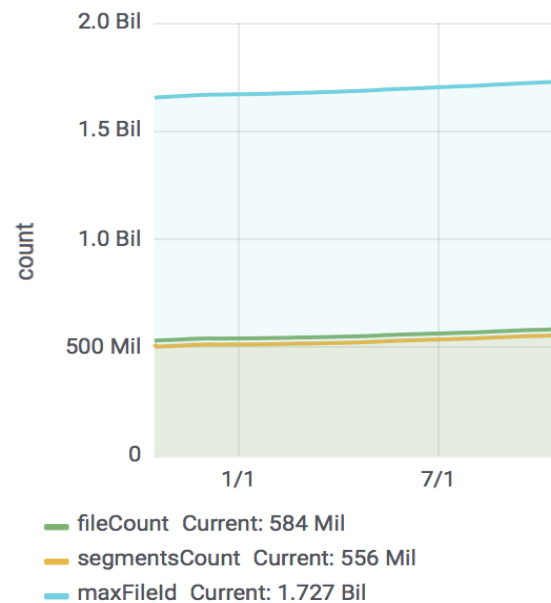
<https://filer-carbon.cern.ch/grafana/dashboard/db/castor-dashboard?orgId=1>



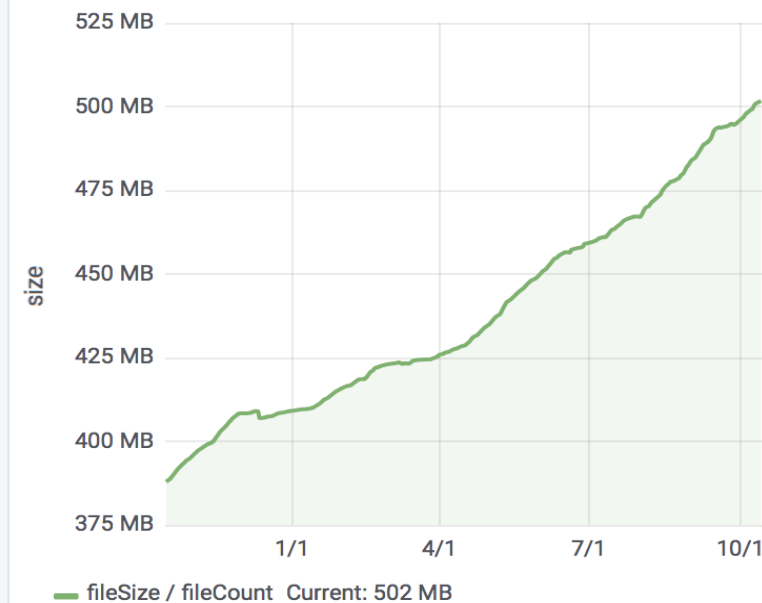
Total data



File counters

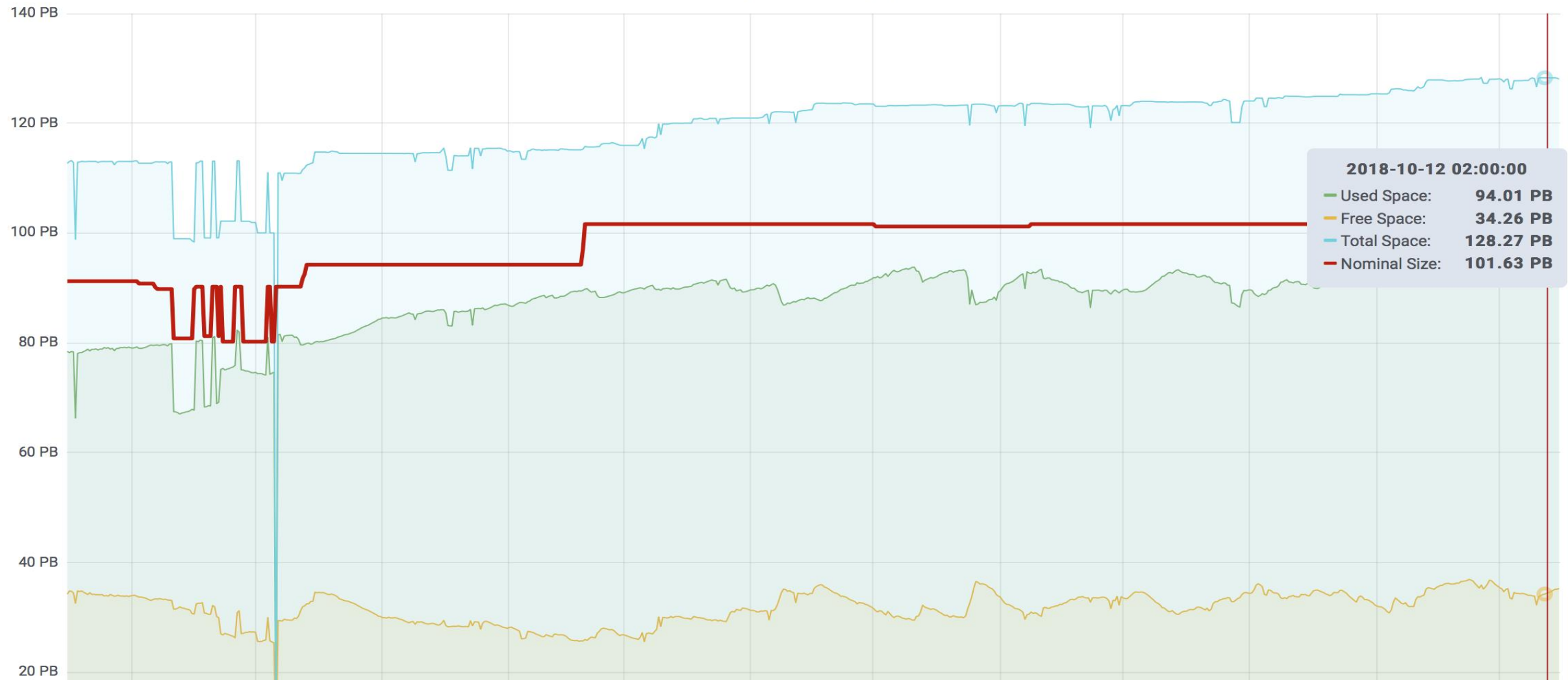


Mean file size



Instance All ▾

EOS Space Monitor (All) ▾



Teachers Program - Jul 24

17-Jul-24

Electricity and cooling

- ▶ 2.7 MW consumption (+ ~ 1 MW cooling) from maximum 3.5 MW
 - ▶ 480 KW diesel generators
- ▶ Protected by UPS
 - ▶ Enough to start the diesel generators
 - ▶ Enough to shut down non-critical machines*
- ▶ Cooling
 - ▶ Chilled air via silver ducts enters the false floor and the into the closed server aisles
 - ▶ Water-cooled racks in the vault in the basement

WLCG – Worldwide LHC computing grid

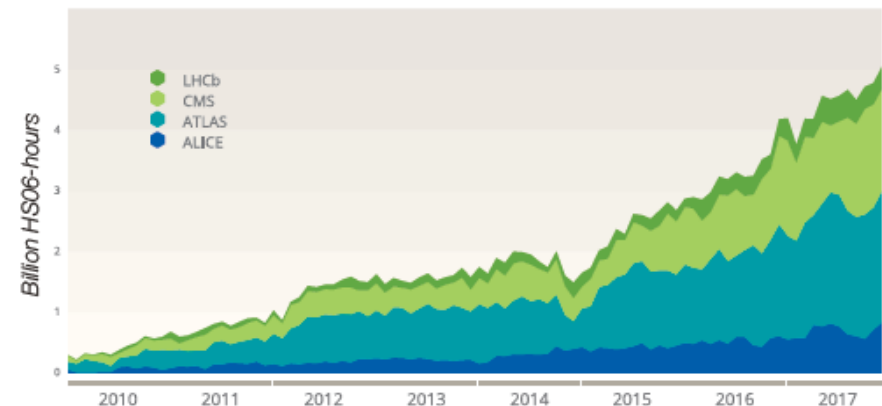
- ▶ More than 170 data centres in 42 countries with about 800,000 cpu cores
 - ▶ CERN provides about 20% of the WLCG resources
 - ▶ Allows more than 10,000 physicists to access LHC data
 - ▶ >250,000 jobs run concurrently on the Grid
 - ▶ Storage is about 400 PB disk and 400 PB of tape globally
 - ▶ In 2016, global transfer rates have regularly exceeded 35GB/s

- ▶ Key facts and numbers
(<http://information-technology.web.cern.ch>)

www.cern.ch/wlcg
www.cern.ch/wlcg-public

Evolution of the global core processor time delivered by the Worldwide LHC Computing Grid (WLCG)

As seen on the graph, the global central processing unit (CPU) time delivered by WLCG (expressed in billions of HS06 hours per month, HS06 being the HEP-wide benchmark for measuring CPU performance) shows a continual increase. In 2017, WLCG combined the computing resources of about 800 000 computer cores.



Configuration management at CERN IT



FOREMAN



HAPROXY

Certification
Manager

CERNMegabus

43 000
*Puppet managed
machines*

TEIGI Tool suite



puppet

MCOLLECTIVE

PuppetDB

AI-TOOLS

...

riakKV

python™

django

The Puppet cycle

Interactions with the server and the agent

- 01** **Store manifests into Git**
As a first step, manifests (our config) have to be generated and stored in GitLab.
- 02** **Register a machine**
A machine will then be created, in a specific hostgroup (eg. webchat/frontend/atlas). It will be registered in Foreman.
- 03** **Run Puppet**
With the machine ready, the Puppet agent can be executed interactively (or let it run by itself). This will request the catalog (final state) of the machine.
- 04** **Master asks for hostgroup**
The Puppet master handling the request will ask Foreman for the hostgroup of the machine.
- 05** **Master asks for manifests**
Once it has the hostgroup, it will obtain the manifests that we defined in GitLab.
- 06** **Catalog generation**
As a final step, the Puppet master will generate the catalog and return it to the agent, which will apply it to the machine.

Thanks Config team fro the slide

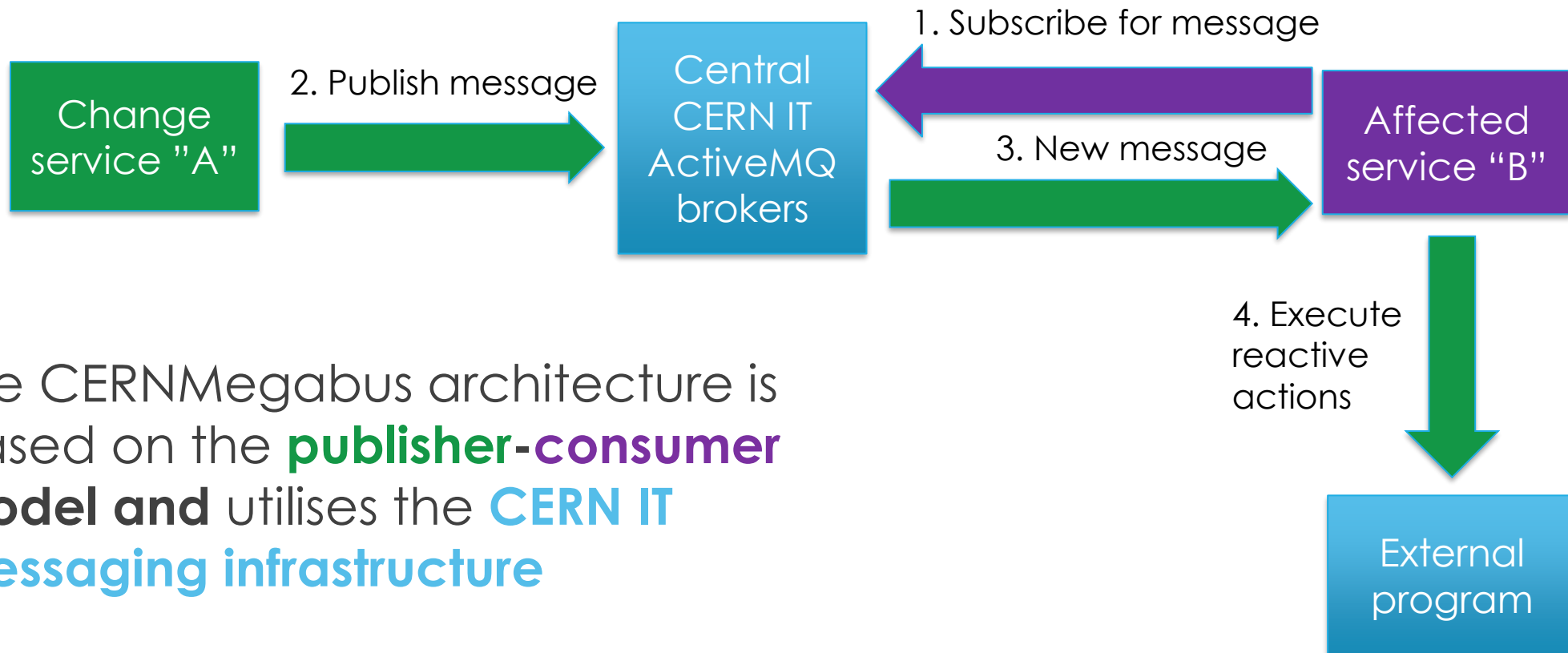
CERNMegabus at CERN IT

A service that provides for **instant communication between services**

CERNMegabus

- The CERNMegabus architecture is based on the **publisher-consumer model** and utilises the **CERN IT messaging infrastructure**
- The publisher and the consumer services comprises of building blocks
 - **configured with Puppet**
 - to use the **CERNMegabus python libraries**
- **Installed on all** Puppet managed **machines** in the **CERN CC**

CERNMegabus architecture



The CERNMegabus architecture is based on the **publisher-consumer model** and utilises the **CERN IT messaging infrastructure**

Already our clients

CASTOR



HAPROXY

BATCH

EOS

CERNMegabus

CLOUD

43 000
Puppet managed
machines

TEIGI Tool suite
(roger)

DNS Load Balancing

IT Monitoring

CERN Computer
Centre Power Cut
Management



From roger to EOS/CASTOR/Puppet HAProxy



Set roger state of
*castor-lhcb-
disknode-X* to
disabled

2. Publish message to
`/topic/roger.**group.castor`

Central
CERN IT
ActiveMQ
brokers

1. Subscribe for topic `/topic/roger.**group.castor` and
hostgroup selector in message

3. Message de-queued

*castor-lhcb-
headnode-Y*

4. Execute
`\modifydiskserver`
**Disab castor-
lhcb-disknode-X**

Set *castor-lhcb-
disknode-X* in
read-only mode

hostgroup header: ``castor-lhcb-diskservers``

```
{
  "new": {"update_time": "1538633786", "updated_by": "blueuser",
    "hostname": "castor-lhcb-disknode-X", ..., "appstate": "disabled"},
  "old": {"update_time": "1538633774", "updated_by": "somebody",
    "hostname": "castor-lhcb-disknode-X", ..., "appstate": "production"}
}
```

~1 sec

In practice - 1

CERN CC Power Cut event



1. Preserve data



Power back

2. Shutdown
(all machines which we can)



CERN CC Power cut event detection

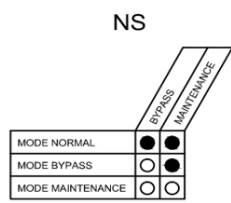
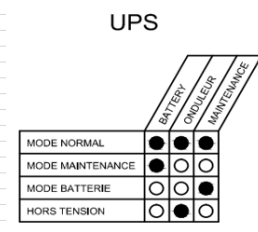
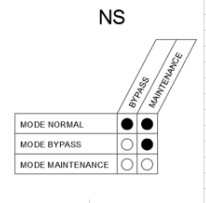
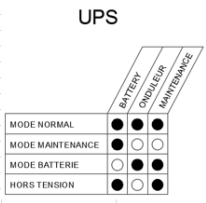
ccpcoX
programmatically
detects power
cut/power back event

UPS and PLC

UPS STATES							PLC STATUS													
UPS	Type de UPS	Descriptive	Data Type	MW	Bit		A	B	C	D	E	F	G	H	I	J	K	L	M	
GROUP 1	Module 1	EB5104*43	Galaxy 6000	Sur batteries	Single	20	0	Act	GROUP 5	Module 4	EB5406*43	Galaxy 7000	En maintenance	Single	54	0	Actif	Inactif	G4-M3-MAINTENANCE	(54.0) = 1 @ 2017-12-04 15:37:32
		EB5104*43	Galaxy 6000	Sur onduleur	Single	20	8	Ina			EB5407*43	Galaxy 7000	Sur batteries	Single	54	8	Actif	Inactif	G4-M4-BATTERY	(54.8) = 1 @ 2017-12-04 15:37:32
		EB5104*43	Galaxy 6000	En maintenance	Single	21	0	Act			EB5407*43	Galaxy 7000	Sur onduleur	Single	55	0	Inactif	Actif	G4-M4-ONDULEUR	(55.0) = 1 @ 2017-12-04 15:37:32
	Module 2	EB5105*43	Galaxy 6000	Sur batteries	Single	21	8	Act		EB5407*43	Galaxy 7000	En maintenance	Single	55	8	Actif	Inactif	G4-M4-MAINTENANCE	(55.8) = 1 @ 2017-12-04 15:37:32	
		EB5105*43	Galaxy 6000	Sur onduleur	Single	22	0	Ina		EB5408*43	Galaxy 7000	Danger Bypass	Single	56	0	Inactif	Actif	G4-NS-BYPASS	(56.0) = 1 @ 2017-12-04 15:37:32	
		EB5105*43	Galaxy 6000	En maintenance	Single	22	8	Act		EB5408*43	Galaxy 7000	En maintenance	Single	56	8	Actif	Inactif	G4-NS-MAINTENANCE	(56.8) = 1 @ 2017-12-04 15:37:32	
	Module 3	EB5106*43	Galaxy 6000	Sur batteries	Single	23	0	Act		EB5704*43	Galaxy 7000	Sur batteries	Single	60	0	Actif	Inactif	G5-M1-BATTERY	(60.0) = 1 @ 2017-12-04 15:37:32	
		EB5106*43	Galaxy 6000	Sur onduleur	Single	23	8	Ina		EB5704*43	Galaxy 7000	Sur onduleur	Single	60	8	Inactif	Actif	G5-M1-ONDULEUR	(60.8) = 1 @ 2017-12-04 15:37:32	
		EB5106*43	Galaxy 6000	En maintenance	Single	24	0	Act		EB5704*43	Galaxy 7000	En maintenance	Single	61	0	Actif	Inactif	G5-M1-MAINTENANCE	(61.0) = 1 @ 2017-12-04 15:37:32	
	Module 4	EB5107*43	Galaxy 6000	Sur batteries	Single	24	8	Act		EB5705*43	Galaxy 7000	Sur batteries	Single	61	8	Actif	Inactif	G5-M2-BATTERY	(61.8) = 1 @ 2017-12-04 15:37:32	
		EB5107*43	Galaxy 6000	Sur onduleur	Single	25	0	Ina		EB5705*43	Galaxy 7000	Sur onduleur	Single	62	0	Inactif	Actif	G5-M2-ONDULEUR	(62.0) = 1 @ 2017-12-04 15:37:32	
		EB5107*43	Galaxy 6000	En maintenance	Single	25	8	Act		EB5705*43	Galaxy 7000	En maintenance	Single	62	8	Actif	Inactif	G5-M2-MAINTENANCE	(62.8) = 1 @ 2017-12-04 15:37:32	
GROUP 2	Module 1	EB5208*43	Galaxy 6000	Danger Bypass	Single	26	0	Ina	EB5706*43	Galaxy 7000	Sur batteries	Single	63	0	Actif	Inactif	G5-M3-BATTERY	(63.0) = 1 @ 2017-12-04 15:37:32		
		EB5208*43	Galaxy 6000	En maintenance	Single	26	8	Act	EB5706*43	Galaxy 7000	Sur onduleur	Single	63	8	Inactif	Actif	G5-M3-ONDULEUR	(63.8) = 1 @ 2017-12-04 15:37:32		
		EB5204*43	Galaxy 6000	Sur batteries	Single	30	0	Act	EB5706*43	Galaxy 7000	En maintenance	Single	64	0	Actif	Inactif	G5-M3-MAINTENANCE	(64.0) = 1 @ 2017-12-04 15:37:32		
	Module 2	EB5204*43	Galaxy 6000	Sur onduleur	Single	30	8	Ina	EB5707*43	Galaxy 7000	Sur batteries	Single	64	8	Actif	Inactif	G5-M4-BATTERY	(64.8) = 1 @ 2017-12-04 15:37:32		
		EB5204*43	Galaxy 6000	En maintenance	Single	31	0	Act	EB5707*43	Galaxy 7000	Sur onduleur	Single	65	0	Inactif	Actif	G5-M4-ONDULEUR	(65.0) = 1 @ 2017-12-04 15:37:32		
		EB5205*43	Galaxy 6000	Sur batteries	Single	31	8	Act	EB5707*43	Galaxy 7000	En maintenance	Single	65	8	Actif	Inactif	G5-M4-MAINTENANCE	(65.8) = 1 @ 2017-12-04 15:37:32		
	Module 3	EB5205*43	Galaxy 6000	Sur onduleur	Single	32	0	Ina	EB5708*43	Galaxy 7000	Sur batteries	Single	66	0	Actif	Inactif	G5-M5-BATTERY	(66.0) = 1 @ 2017-12-04 15:37:32		
		EB5205*43	Galaxy 6000	En maintenance	Single	32	8	Act	EB5708*43	Galaxy 7000	Sur onduleur	Single	66	8	Inactif	Actif	G5-M5-ONDULEUR	(66.8) = 1 @ 2017-12-04 15:37:32		
		EB5206*43	Galaxy 6000	Sur batteries	Single	33	0	Act	EB5708*43	Galaxy 7000	En maintenance	Single	67	0	Actif	Inactif	G5-M5-MAINTENANCE	(67.0) = 1 @ 2017-12-04 15:37:32		
	Module 4	EB5206*43	Galaxy 6000	Sur onduleur	Single	33	8	Ina	EB5720*43	Galaxy 7000	Sur batteries	Single	68	8	Actif	Inactif	G5-NS-BYPASS	(67.8) = 1 @ 2017-12-04 15:37:32		
		EB5207*43	Galaxy 6000	Sur batteries	Single	34	0	Act	EB5720*43	Galaxy 7000	Sur onduleur	Single	69	0	Inactif	Actif	G5-M6-BATTERY	(68.8) = 1 @ 2017-12-04 15:37:32		
		EB5207*43	Galaxy 6000	Sur onduleur	Single	35	0	Ina	EB5720*43	Galaxy 7000	En maintenance	Single	69	8	Actif	Inactif	G5-NS-MAINTENANCE	(68.0) = 1 @ 2017-12-04 15:37:32		
GROUP 3	Module 1	EB5308*43	Galaxy 6000	Danger Bypass	Single	36	0	Ina	EB5108*43	Galaxy 7000	Danger Bypass	Single	67	0	Inactif	Actif	G5-M6-ONDULEUR	(69.0) = 1 @ 2017-12-04 15:37:32		
		EB5308*43	Galaxy 6000	En maintenance	Single	36	8	Act	EB5108*43	Galaxy 7000	En maintenance	Single	68	8	Actif	Inactif	G5-M6-MAINTENANCE	(69.8) = 1 @ 2017-12-04 15:37:32		
		EB5304*43	Galaxy 6000	Sur batteries	Single	40	0	Act												
	Module 2	EB5304*43	Galaxy 6000	Sur onduleur	Single	40	8	Ina												
		EB5304*43	Galaxy 6000	En maintenance	Single	41	0	Act												
		EB5305*43	Galaxy 6000	Sur batteries	Single	41	8	Act												
	Module 3	EB5305*43	Galaxy 6000	Sur onduleur	Single	42	0	Ina												
		EB5305*43	Galaxy 6000	En maintenance	Single	42	8	Act												
		EB5306*43	Galaxy 6000	Sur batteries	Single	43	0	Act												
	Module 4	EB5306*43	Galaxy 6000	Sur onduleur	Single	43	8	Ina												
		EB5306*43	Galaxy 6000	En maintenance	Single	44	0	Act												
		EB5307*43	Galaxy 6000	Sur batteries	Single	44	8	Act												
GROUP 4	Module 1	EB5307*43	Galaxy 6000	Sur onduleur	Single	45	0	Ina												
		EB5307*43	Galaxy 6000	En maintenance	Single	45	8	Act												
		EB5307*43	Galaxy 6000	Sur batteries	Single	45	0	Ina												
	Module 2	EB5709*43	Galaxy 6000	Danger Bypass	Single	46	0	Ina												
		EB5709*43	Galaxy 6000	En maintenance	Single	46	8	Act												
		EB5404*43	Galaxy 7000	Sur batteries	Single	50	0	Act												
	Module 3	EB5404*43	Galaxy 7000	Sur onduleur	Single	50	8	Ina												
		EB5404*43	Galaxy 7000	En maintenance	Single	51	0	Act												
		EB5405*43	Galaxy 7000	Sur batteries	Single	51	8	Act												
	Module 4	EB5405*43	Galaxy 7000	Sur onduleur	Single	52	0	Ina												
		EB5405*43	Galaxy 7000	En maintenance	Single	52	8	Act												
		EB5406*43	Galaxy 7000	Sur batteries	Single	53	0	Act												
Module 5	EB5406*43	Galaxy 7000	Sur onduleur	Single	53	8	Ina													
	EB5406*43	Galaxy 7000	En maintenance	Single	54	0	Act													
	EB5406*43	Galaxy 7000	Sur batteries	Single	54	8	Ina													

GALAXY 6000 LOGIC

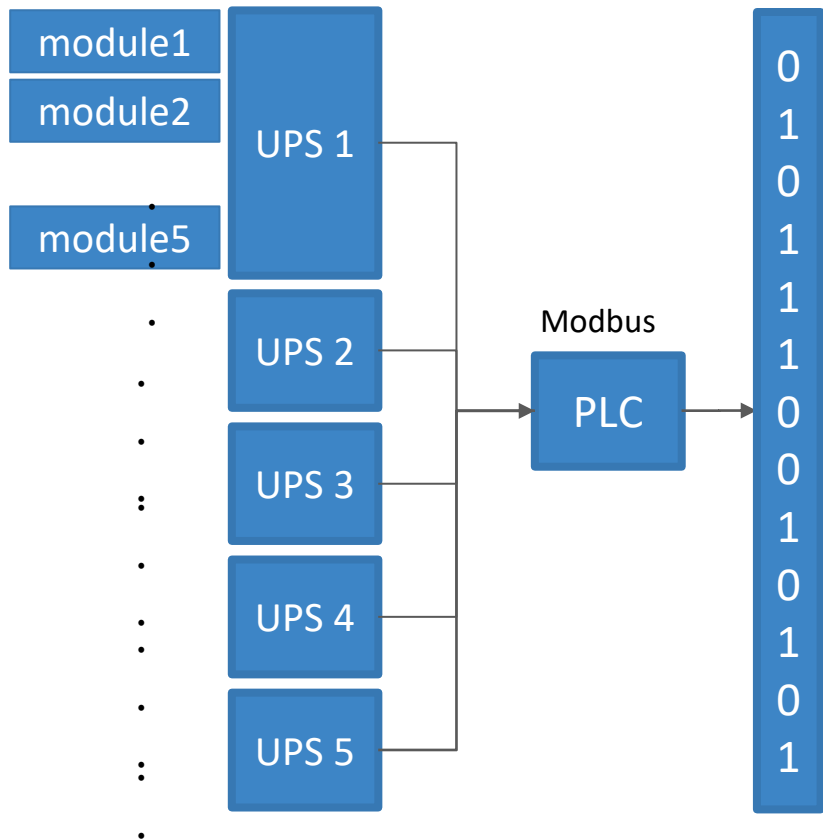
GALAXY 7000 LOGIC



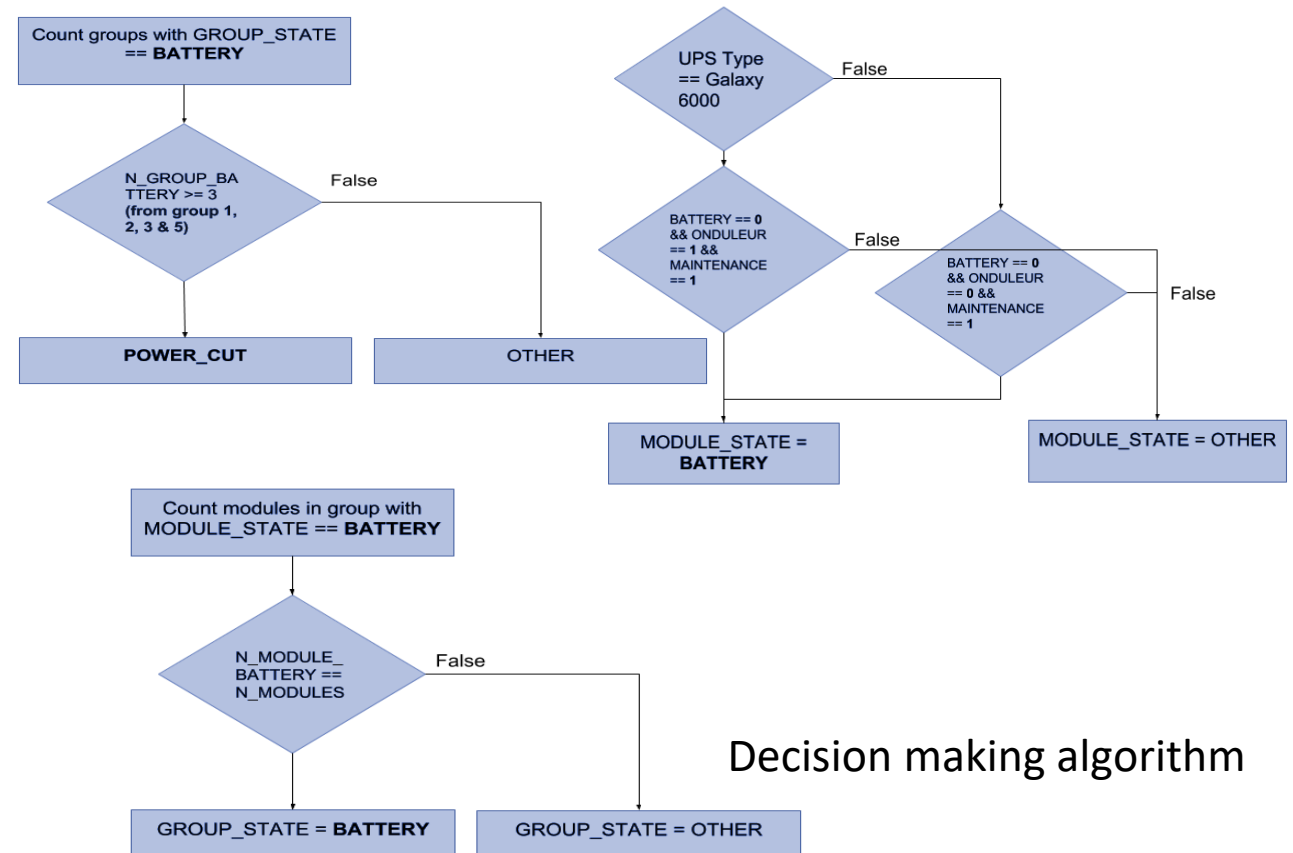
Raw data (in case there is some error with pa

Address	Value	
20		257
21		257
22		257
23		257
24		257
25		257
26		1
27		0
28		0
29		0
30		257
31		257
32		257
33		257
34		257
35		257
36		1
37		0
38		0
39		0
40		257
41		257
42		257
43		257
44		257
45		257
46		257

CERN CC Power cut event detection algorithm

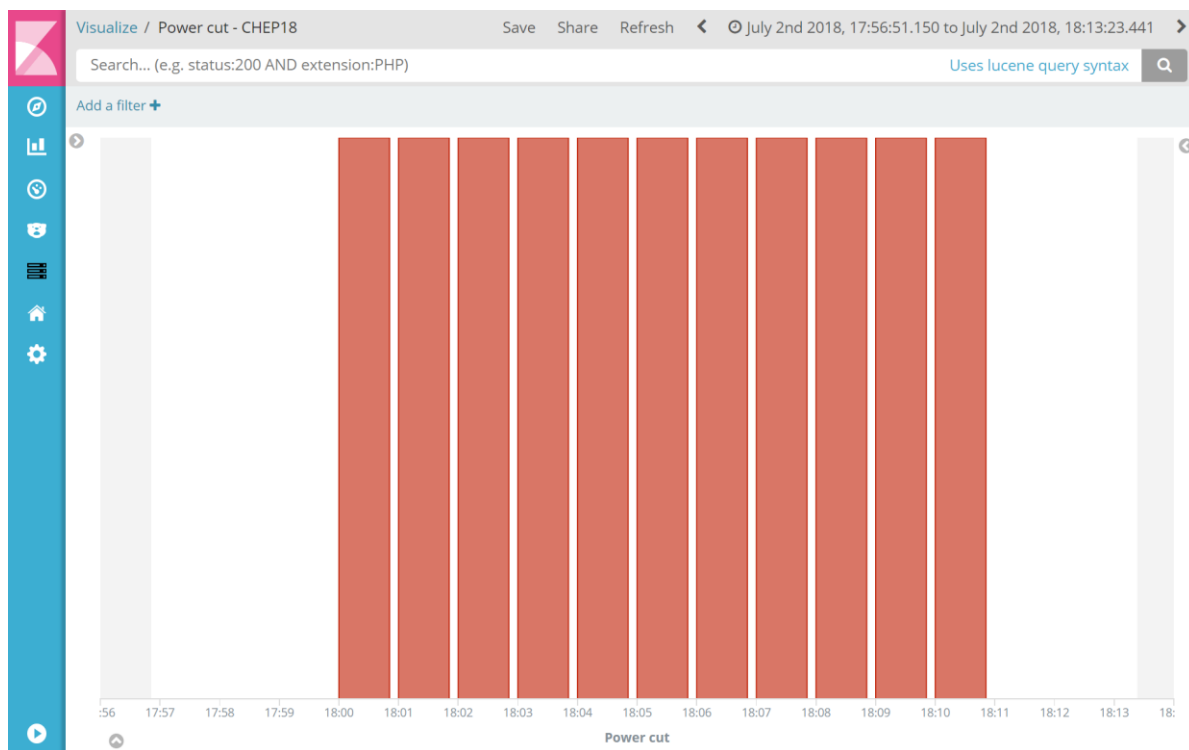


Data collection



Decision making algorithm

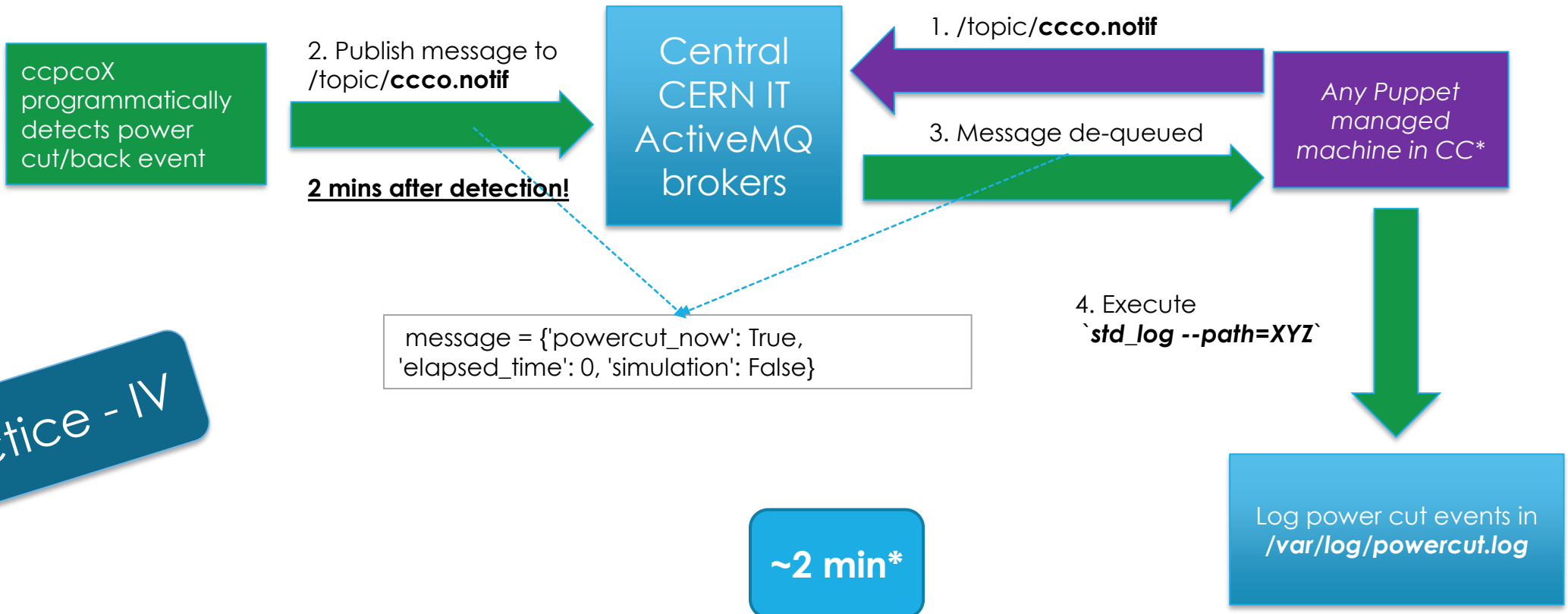
CERN CC Power cut tests



- ▶ During mid-annual power cut test on the 2nd of July, 2018
 - ▶ Detected power cut
 - ▶ Notified the subscribed machines
 - ▶ Shutdown the machines, which had been predefined to be shutdown
 - ▶ Detected the power back
 - ▶ Notified the machines, which had been predefined to wait

Presented at CHEP'18

From CERN CC UPS PLC to CERN CC shutdown



In practice - IV

Thanks

THANKS for listening and **ENJOY** your visit at CERN



Zhechka