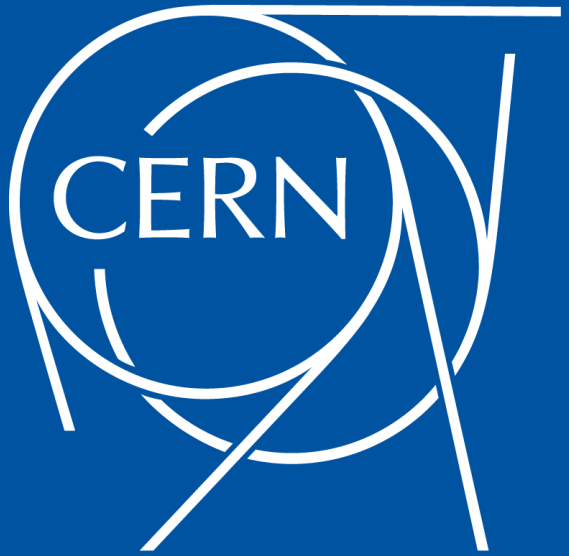
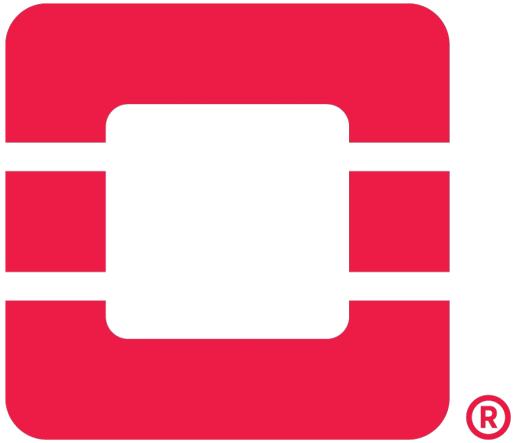
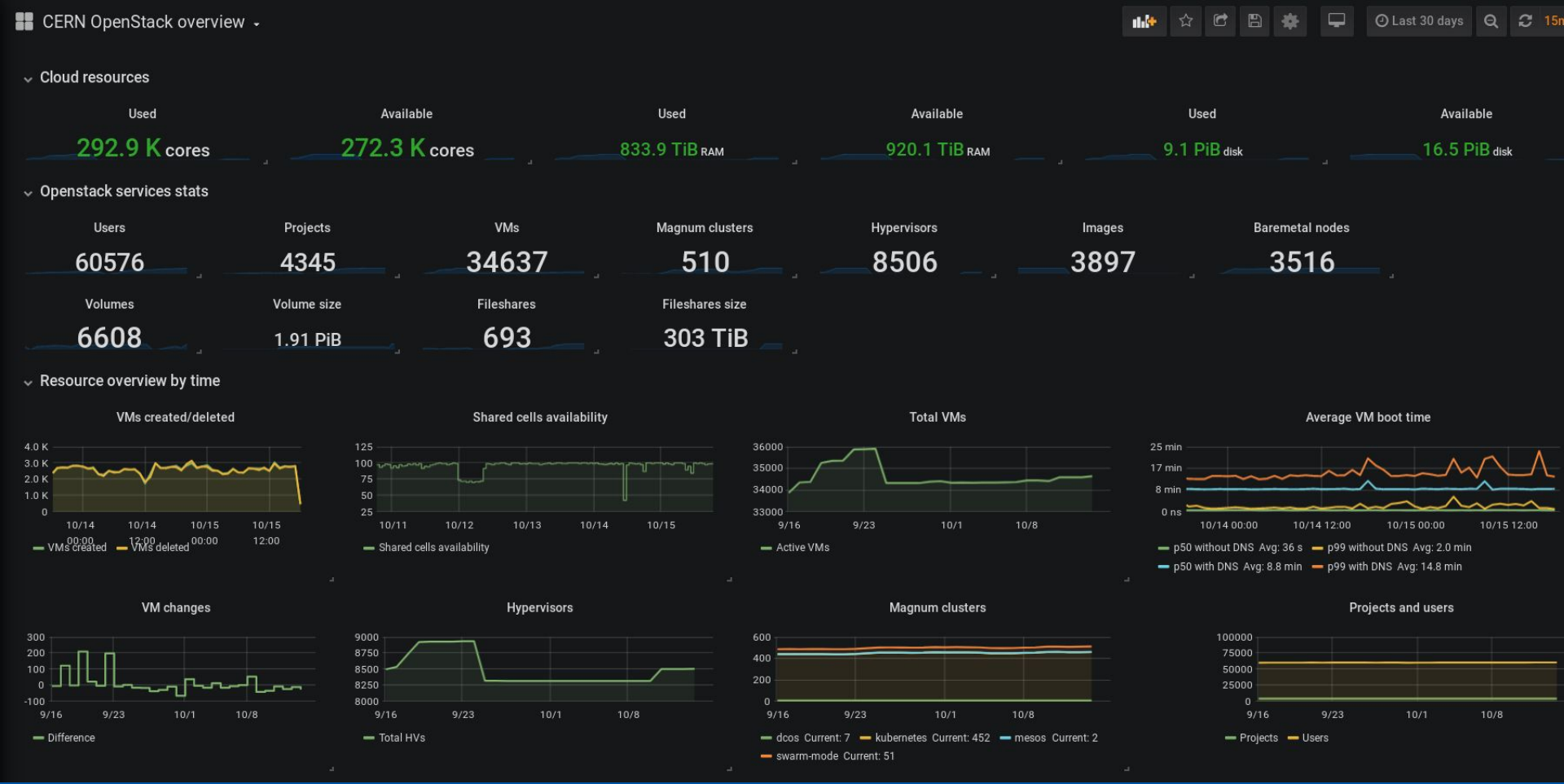

CERN



CERN Cloud Infrastructure update

Daniel Abad on behalf of the Cloud Infrastructure Team
HEPiX Autumn 2019 Workshop


CERN OpenStack overview
Last 30 days
Cloud resources
Used
292.9 K cores
Available
272.3 K cores
Used
833.9 TiB RAM
Available
920.1 TiB RAM
Used
9.1 PiB disk
Available
16.5 PiB disk
Openstack services stats
Users
60576
Projects
4345
VMs
34637
Magnum clusters
510
Hypervisors
8506
Images
3897
Baremetal nodes
3516
Volumes
6608
Volume size
1.91 PiB
Fileshares
693
Fileshares size
303 TiB
Resource overview by time
VMs created/deleted
4.0 K
3.0 K
2.0 K
1.0 K
0
10/14 00:00
10/14 12:00
10/15 00:00
10/15 12:00
VMs created
VMs deleted
Shared cells availability
125
100
75
50
25
10/11
10/12
10/13
10/14
10/15
Shared cells availability
Total VMs
36000
35000
34000
33000
9/16
9/23
10/1
10/8
Active VMs
Average VM boot time
25 min
17 min
8 min
0 ns
10/14 00:00
10/14 12:00
10/15 00:00
10/15 12:00
p50 without DNS Avg: 36 s
p99 without DNS Avg: 2.0 min
p50 with DNS Avg: 8.8 min
p99 with DNS Avg: 14.8 min
VM changes
300
200
100
0
-100
9/16
9/23
10/1
10/8
Difference
Hypervisors
9000
8750
8500
8250
8000
9/16
9/23
10/1
10/8
Total HVs
Magnum clusters
600
400
200
0
9/16
9/23
10/1
10/8
dcos Current: 7
kubernetes Current: 452
mesos Current: 2
swarm-mode Current: 51
Projects and users
100000
75000
50000
25000
0
9/16
9/23
10/1
10/8
Projects
Users

Development areas going forward

- Spot Market / Pre-emptible instances
- Software Defined Networking
  - Introducing LBaaS this month
- Magnum rolling upgrades
- Enrolling all 15K servers in Ironic
  - Containers on Bare Metal
- Evaluation of OpenStack Watcher for resource optimization
- Upgrading to Stein release

https://techblog.web.cern.ch/techblog/

OpenStack Days CERN, 27MAY2019

https://indico.cern.ch/event/776411/

L1TF (Reboot campaign again)

- Disabling SMT OFF in the BIOS makes the difference respect to acting at kernel level. We can still:
  - Overcommit resources and provide the same performance as before
  - Have servers patched for L1TF
  - D. Giordano on " CPU steal and L1TF mitigations"

- Plan to fully patch L1TF
  - Continue the reconfiguration of the remaining nodes that still have SMT ON
    - Set SMT OFF in the BIOS
    - A reboot campaign of the affected HVs was announced, VMs' users contacted and finally performed in May

Container Orchestration - OpenStack Magnum

An OpenStack API service that allows creation of container clusters

- Use your OpenStack Keystone credentials
- You choose your cluster type
- Multi-tenancy
- Quickly create new clusters with advanced features

- Deprecating Mesos/DCOS 

- Today, the CERN Cloud hosts >500 Kubernetes clusters

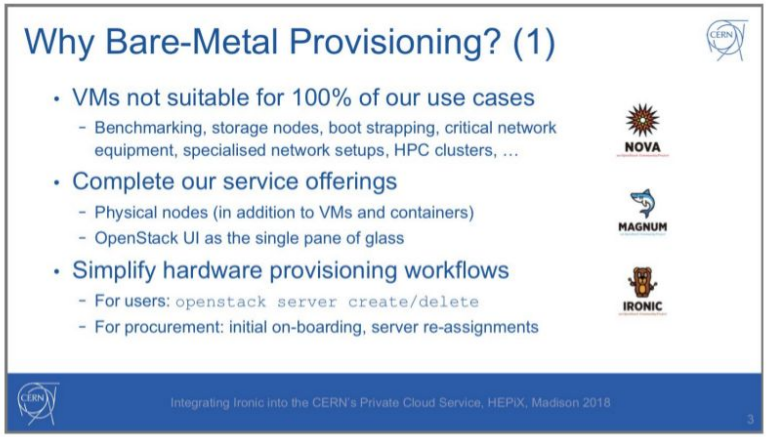
Bare metal Provisioning - OpenStack Ironic

Integrated OpenStack program which aims to provision bare metal machines instead of virtual machines

IRONIC

an OpenStack Community Project

- In production at CERN since 2018
- Consolidate accounting & bookkeeping
- Completes our service offering
- All new hardware is enrolled using Ironic
- +3700 nodes managed by Ironic
- Testing enrolment of existing hardware




OpenStack components › Ironic
Last 12 hours
2h
Total Nodes
3681
Active Nodes
3594
API Nodes
3
API Info
2
API Warn
0
API Errors
0
Service Info
508
Service Warn
10038
Service Errors
1228
Maintenance Modes
current percentage
false 3.64 K 99%
true 40 1%
Power States
current percentage
15 0%
power off 81 2%
power on 3.59 K 97%
Unusual Provisioning States
current percentage
clean failed 9 54%
clean wait 2 14%
cleaning 1 6%
deploy failed 1 6%
API Load distribution
current percentage
cci-ironic-s06.cern.ch 234 30%
cci-ironic-s07.cern.ch 286 37%
cci-ironic-s08.cern.ch 260 33%
API calls
1.3 K
1.0 K
750
500
250
0
07:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00 18:00
Log Messages
40 30 20 10 0
200 150 100 50 0
08:00 10:00 12:00 14:00 16:00 18:00
Info Warning Error
Rally Tests
Time bm_test bm_qa bm_001
6:00 PM 0 0 -
5:00 PM 0 0 1
4:00 PM 0 0 1
3:00 PM 0 0 0
2:00 PM 0 0 0
1:00 PM 0 0 0
Nodes per state
4.0 K
3.0 K
2.0 K
1.0 K
Nodes per conductor (excluding null)
1.5 K
1.0 K
500

Getting physical resources in 2011

Public Procurement Purchase Model

Step	Time (Days)	Elapsed (Days)
User expresses requirement		0
Market Survey prepared	15	15
Market Survey for possible vendors	30	45
Specifications prepared	15	60
Vendor responses	30	90
Test systems evaluated	30	120
Offers adjudicated	10	130
Finance committee	30	160
Hardware delivered	90	250
Burn in and acceptance	30 days typical 380 worst case	280
Total		**280+ Days**

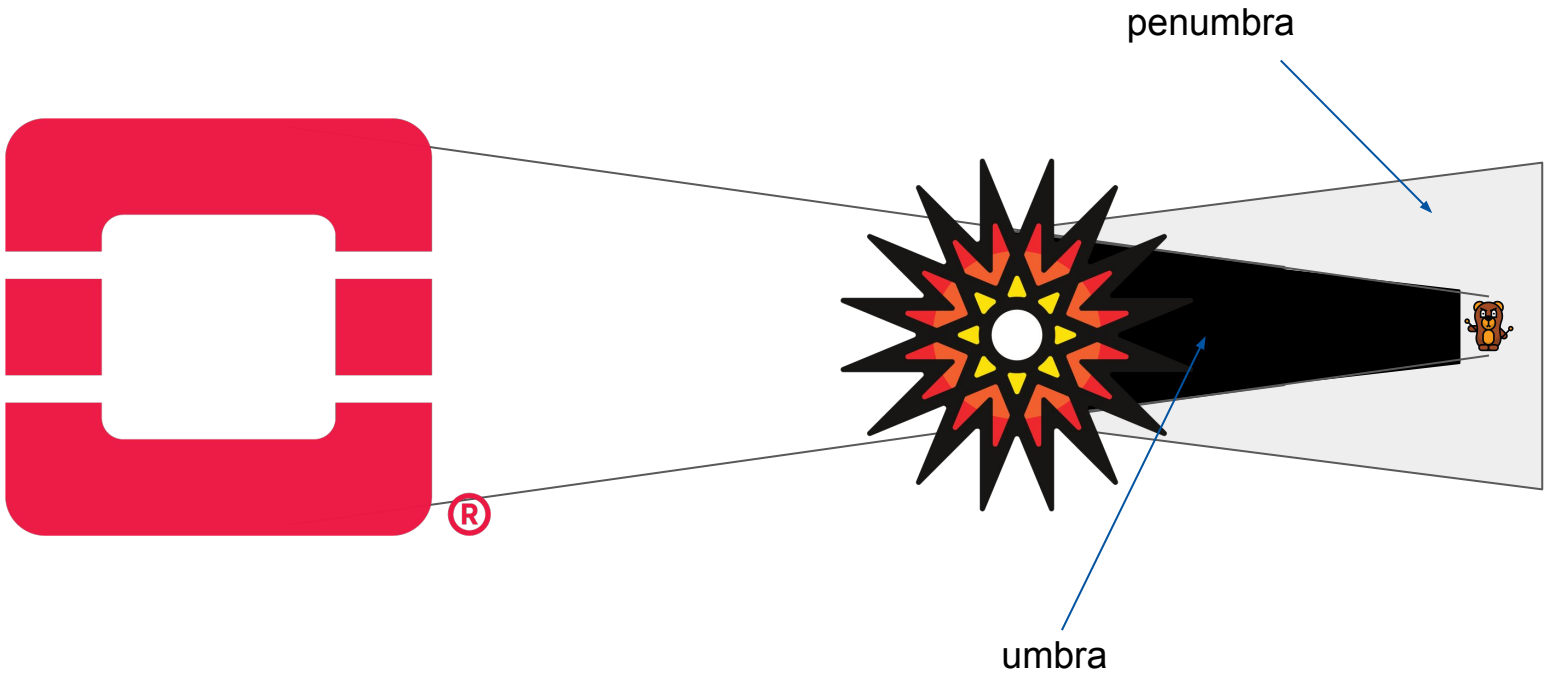
Getting physical resources in 2019

```
dabadaba@aiadm $ openstack server create \
 --image “CC7 - x86_64 [2019-09-30]” \
 --flavor p1.cd6839123.S513-H-IP71 \
 my_physical_instance
```

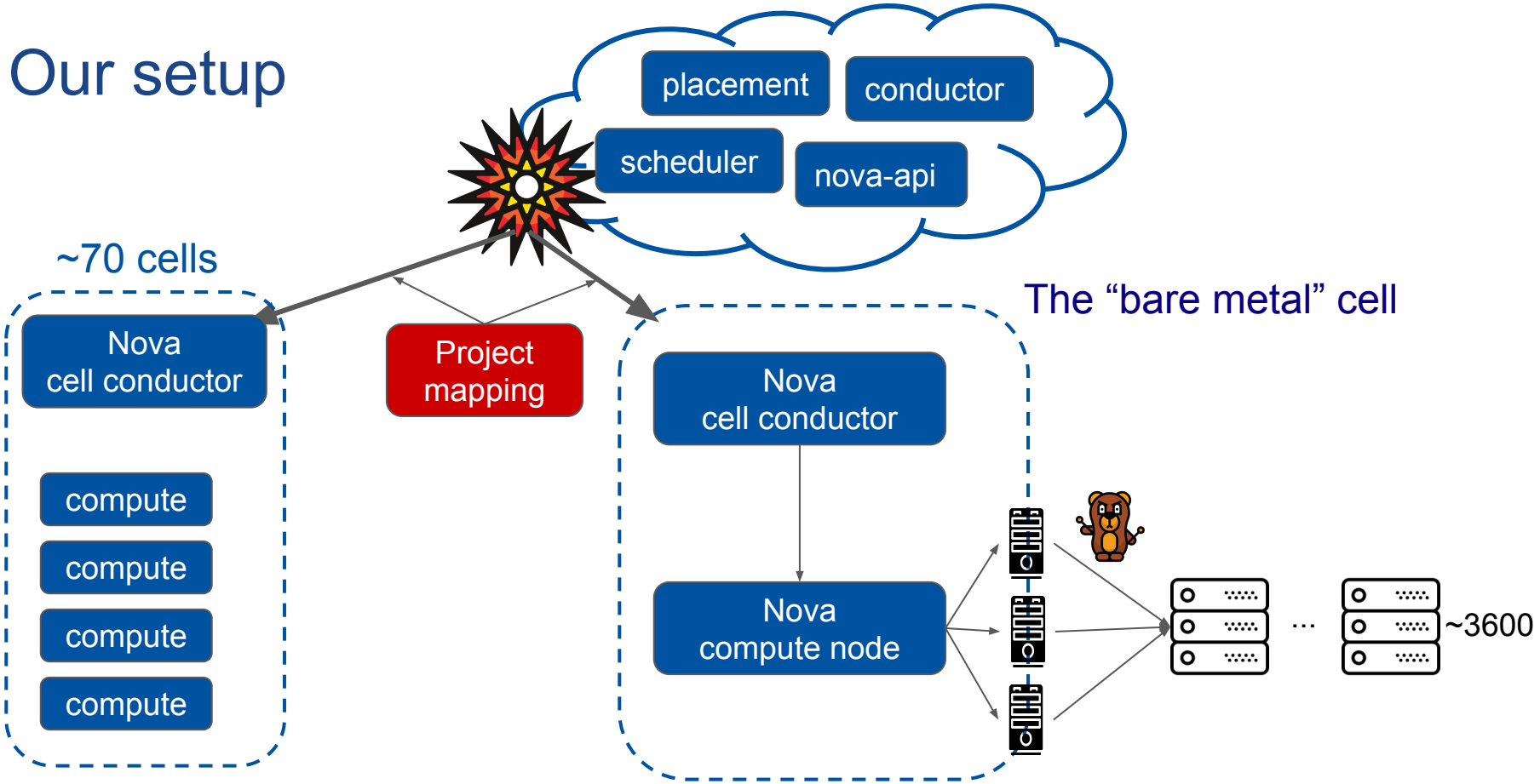
15 minutes later

```
dabadaba@aiadm $ ssh root@my_physical_instance
root@my_physical_instance #
```

Supernova eclipse



Our setup


placement
conductor
scheduler
nova-api
~70 cells
Nova
cell conductor
Project
mapping
The “bare metal” cell
Nova
cell conductor
compute
compute
compute
compute
Nova
compute node
~3600

Scalability issues

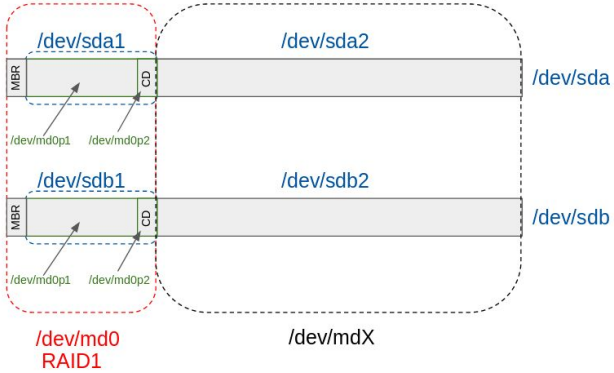
- Started with ~1500 nodes
  - (And only one level 2 compute node)
  - Overcome and now working for >3600 nodes

- Mainly synchronization between nova and Ironic
  - Upgrade to placement on nova
  - Placement designed for virtual resources (VMs, GPUs, etc.)
  - Tweaked some params for the bare metal cell:

- https://techblog.web.cern.ch/techblog/post/nova-ironic-at-scale/

Software RAID

- Sep 2018 - Proposal: Downstream Ironic Software RAID support at CERN
- Dec 2018 - Specification:
  - https://specs.openstack.org/openstack/ironic-specs/specs/not-implemented/software-raid.html
  - https://review.opendev.org/#/c/624413/
- Jan-Jun 2019 - Implementation
  - https://techblog.web.cern.ch/techblog/post/ironic_software_raid/




```
ironic node-set-target-raid-config
my-baremetal-node-1379
'{"logical_disks": [{ "raid_level":
"1", "size_gb": "MAX",
"is_root_volume": true }]}'
```

Operational issues

- Adoption of existing hardware
  - ~8000 hypervisors
  - Testing the strategy, high risk
- Mix of Physical resources:
  - Nodes outside of OpenStack
  - Hypervisors
  - End-user nodes (Storage, Windows, etc.)
- Different procedures for interventions
  - Coordination
  - Human communication

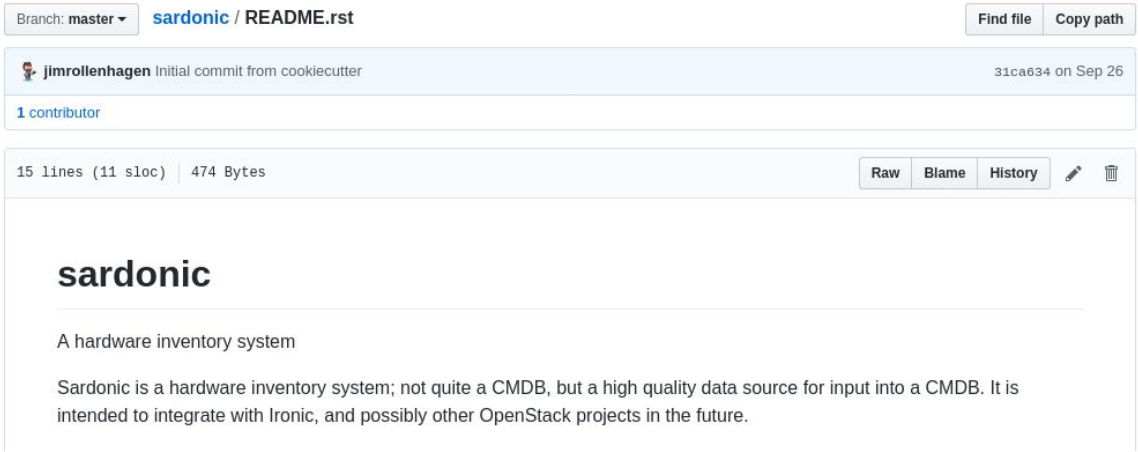
Inventory of Hardware resources: a CMDB?

CMDB: main use cases

- Single source of truth
- Inventory of servers and components
  - Accounting
  - Including temporary “leases”
- Upgrade campaigns
  - Hardware: SSDs, CPUs, etc.
  - Software: kernels, etc.
- Warranty information
  - inventory of servers and components

Sardonic - an OpenStack Proposal

Proposal: [Ideas and Requirements for a "Hardware Inventory" Functionality in OpenStack Ironic]



https://github.com/openstack/sardonic

CMDB: solutions evaluated

Solution	Open Source	License	Platforms	Business model	Language	Last release
I-doit	Yes, no repo	not specified	Linux/Windows	Free + Pro (differences)	PHP5 + JS	1.12.2 (2019-04-01)
CMDbuild	Yes	AGPL v3.0	Linux/Windows	Free + Pro (differences)	Java + Ajax	3.0.0 (2019-04-12)
SnipeIT	Yes	AGPL v3.0	All	Free + Pro (differences	PHP + JS	4.6.15/5.0.0-beta-2 (2019-03-20)
SysAid	No	/	Windows (All)	paid	???	18.1 (2019-11-04)
Spiceworks	No	/	Windows only	freeware	Ruby (2006)	7.5.00107 (2018-03-21)
GLPI	Yes	GPL v2	All (linux mainly)	libre	PHP + JS	9.4.2 (2019-04-11)
NetBox	Yes	Apache 2.0	Linux	libre	Python 3.5	2.5.10 (2019-04-08)
xCAT	Yes	EPL 1.0	Linux	libre	Perl	2.14.6 (2019-03-29)
openDCIM	Yes	GPL v3	Linux	libre	PHP + JS	19.01 (2019-03-04)
Ralph	Yes	Apache 2.0	Linux	libre	Python 3.6 + JS	20190410.2 (2019-04-10)

CMDB: solutions evaluated

Solution	Supported DBs	Schema flexibility	API	GUI
i-doit	MySQL/MariaDB	???	JSON-RPC	Web
CMDBuild	PostgreSQL only	Yes	REST, SOAP	Web (Framework)
SnipeIT	MySQL/MariaDB	At least partial	JSON REST	Web
GLPI	PostgreSQL/MySQL(InnoDB)	???	REST	Web
NetBox	PostgreSQL	???	REST	Web
xCAT	SQLite, MySQL/MariaDB, PostgreSQL	???	REST	Web
openDCIM	MySQL	???	REST	Web
Ralph	MySQL	???	REST (Django)	Web

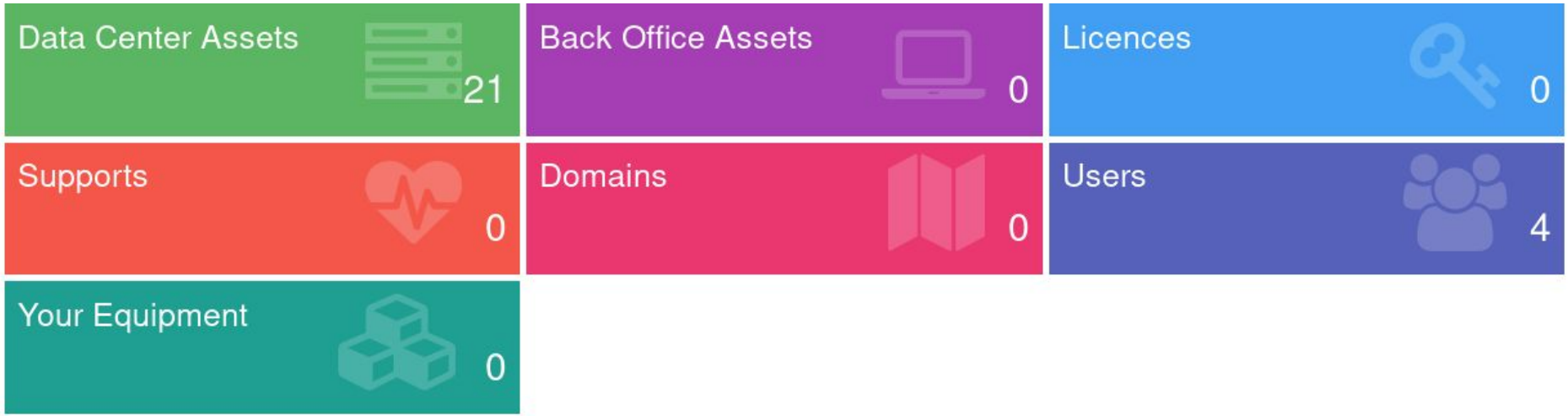
CMDB: solutions evaluated

Solution	Supported Linux distributions	Installation from repo/ Automatic install
i-doit	Debian 8/9, Ubuntu 16.06/18.04, RedHat (RHEL) 7, SUSE (SLES) 12 SP2, SP3, SP4	Yes, with bash script idoit-install
CMDBuild	Any	No
SnipeIT	Any that supports a LAMP stack or equivalent	Partial, web installation
GLPI	Any that supports a LAMP stack or equivalent	Yes, web or cli
NetBox	Ubuntu 18.04, CentOS 7.5	No
xCAT	Ubuntu Server LTS 18.0.4, RHEL 7.6 , SLES 12.3	Yes, with the bash script go-xcat or configuring repositories
OpenDCIM	Any that supports a LAMP stack or equivalent	Partial, web installation
Ralph	Ubuntu 14.04 / 18.04	Install from repo; manual configuration of web stack

CMDB: solutions tested

- NetBox
  - Complex installation
  - Mainly IP address management (IPAM) driven
- Ralph
  - Written in Python/Django
  - Better support for custom fields
  - Easy to configure
- xCAT
  - Does not have a GUI

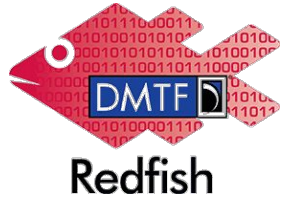
Ralph: Testing with real data



Ralph: Testing with real data

Hostname	Status	Barcode	Model	SN	Invoice date	Invoice number	Location	Service env	Configuration path	Security scan	SCM status
(None)	new	(None)	[Data Center] Manufacturer1 [redacted]6	a legit serial number v4	(None)		1	service1 - environment1	(None)	No scan	-
[redacted]51.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]59-4	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]37.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]59-3	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]73.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]59-2	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]77.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]59-1	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]ern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]83-3	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]rn.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]83-2	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]ern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]78-3	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]ern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]78-2	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]ern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]78-1	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]ern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]2	[redacted]69-3	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]642-b.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]642-B	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]642-a.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]642-A	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]641-d.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]641-D	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]641-c.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]641-C	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]641-b.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]641-B	(None)	(None)	1	service1 - environment1	(None)	No scan	-
[redacted]641-a.cern.ch	new	(None)	[Data Center] Manufacturer1 [redacted]6	[redacted]641-A	(None)	(None)	1	service1 - environment1	(None)	No scan	-

Redfish integration



- TL;DR IPMI's "successor"

- Driven by DMTF Scalable Platforms Management Forum
  - **Promoters**: Broadcom, Dell, Emerson, Hewlett-Packard, Intel, Lenovo, Microsoft, Supermicro, VMWare
  - **Supporters**: AMI, Oracle, Fujitsu, Huawei, Mellanox, Seagate
- An open industry standard specification that provides simple, modern and secure management of scalable platform hardware
- RESTful interface over HTTPs in JSON format
- A secure, multi-node capable replacement for IPMI-over-LAN

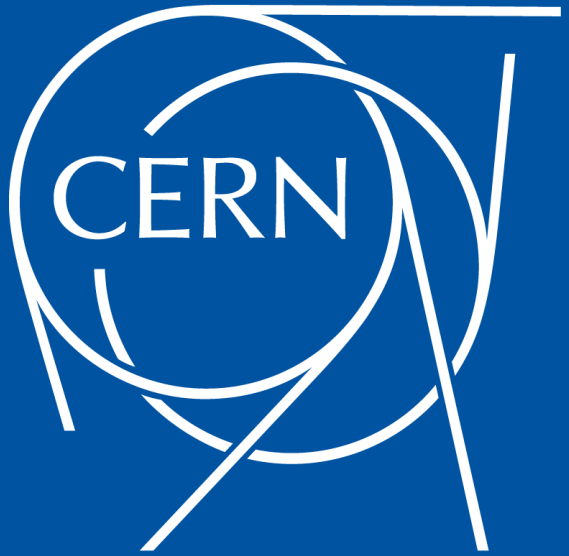
The next 5 years?

- LHC to Run 4
  - Computing must not limit the physics
- Open Infrastructure
  - OpenStack is a key part but lots of others too
  - Ceph, Tungsten Fabric, Grafana, Puppet, CI/CD, Kubernetes, …
- Open Source collaboration is the way forward
  - Natural culture fit for sharing and giving back
  - Work with the communities which are open and vibrant

We are not special anymore

Summary

- During the last 10 years, resource management and deployment model changed completely
  - From Virtualization and Server consolidation to Cloud Infrastructure
  - From Bare metal to VMs, to managed Bare metal to Containers
- Continue to adapt the infrastructure to new technologies and requirements
  - Control plane managed by kubernetes
  - New regions
  - SDN
  - Serverless
- Community and industry collaboration has been productive and inspirational for the CERN team


CERN