Preparing for a new Data Center:

Automated Management of a 10’000 node bare-metal fleet

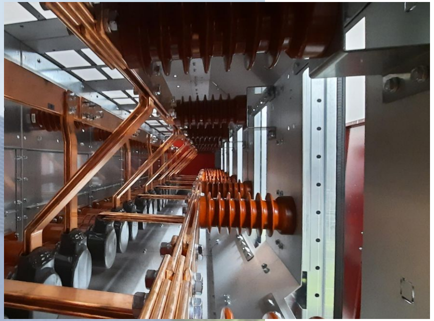
Arne Wiebalck, Luca Atzori, Nikos Papakyprianou,
Michał Piszczek, Maryna Savchenko

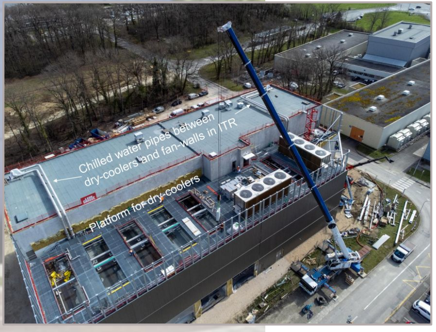
CERN IT

Norfolk, 9 May 2023

The new CERN Prévessin Data Center

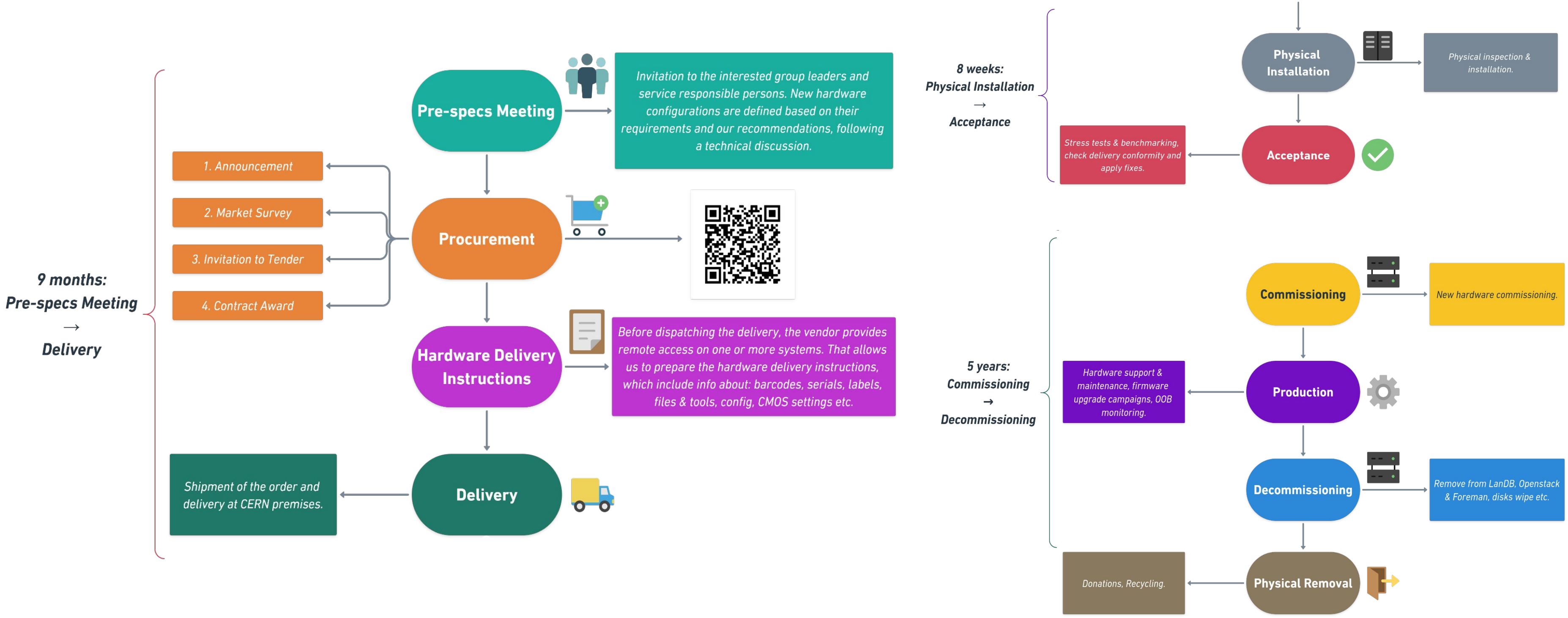
- **Three phases on three floors**
  - 4MW – 2nd floor
  - 8MW – 1st + 2nd floor
  - 12MW – ground floor + 1st + 2nd floor
- **Air-cooled racks with hot-aisle containment**
- **Two redundant power feeds: red and blue**
  - Red feed: 20% UPS coverage





Chilled water pipes between dry-coolers and fan-walls in ITR
Platform for dry-coolers

From Specification to Removal



9 months:
Pre-specs Meeting
→
Delivery

Pre-specs Meeting

Invitation to the interested group leaders and service responsible persons. New hardware configurations are defined based on their requirements and our recommendations, following a technical discussion.

Procurement

1. Announcement

2. Market Survey

3. Invitation to Tender

4. Contract Award

Hardware Delivery Instructions

Before dispatching the delivery, the vendor provides remote access on one or more systems. That allows us to prepare the hardware delivery instructions, which include info about: barcodes, serials, labels, files & tools, config, CMOS settings etc.

Delivery

Shipment of the order and delivery at CERN premises.

8 weeks:
Physical Installation
→
Acceptance

Physical Installation

Physical inspection & installation.

Acceptance

Stress tests & benchmarking, check delivery conformity and apply fixes.

5 years:
Commissioning
→
Decommissioning

Commissioning

New hardware commissioning.

Production

Hardware support & maintenance, firmware upgrade campaigns, OOB monitoring.

Decommissioning

Remove from LanDB, Openstack & Foreman, disks wipe etc.

Physical Removal

Donations, Recycling.

The Ironic Bare Metal API

➔ **Idea: Extend the cloud approach to bare metal servers!**

- ➢ Bare Metal offering to complement VMs and containers
- ➢ Provided via the same interface
- ➢ Simplification: workflows, approval, accounting, …

➔ **API service to manage/interact with physical servers**

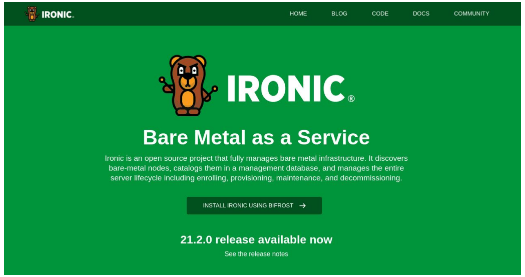
- ➢ Originally a provisioning driver in Nova

➔ **Can be used with OpenStack or stand-alone**

- ➢ ironicbaremetal.org

➔ **Leverages OSS standards and tooling**

- ➢ IPMI/Redfish, PXE, DHCP, … but allows for vendor plugins

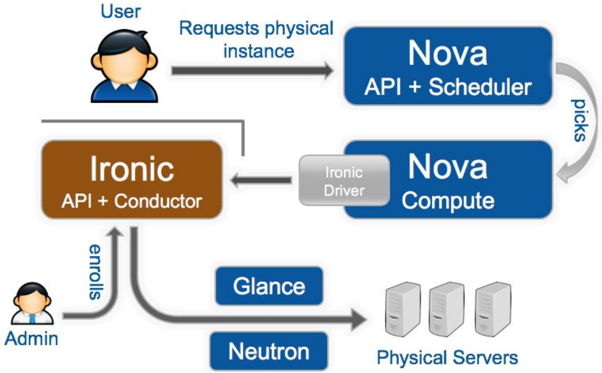


Bare Metal as a Service

Ironic is an open source project that fully manages bare metal infrastructure. It discovers bare-metal nodes, catalogs them in a management database, and manages the entire server lifecycle including enrolling, provisioning, maintenance, and decommissioning.

21.2.0 release available now

See the release notes

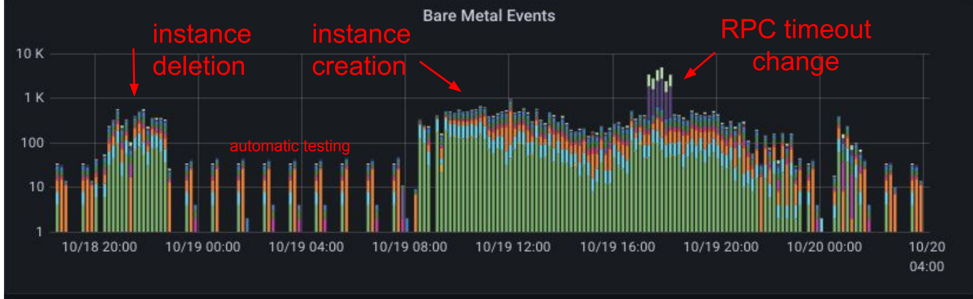


BACK TO THE FUTURE : Physical Batch

- **Conversion of virtual to physical batch**
  - with the availability of a bare metal API, we revisited the virtualisation tax
- **3’775 hypervisors recreated as physical batch instances**
  - done in multiple chunks over several months
- **Terraform as the ‘Infrastructure-as-Code’ tool to interface with OpenStack/Ironic**

HashiCorp Terraform

Bonus: 16’000 VMs less than one year ago … 10k+ IPv4 addresses free’d up.

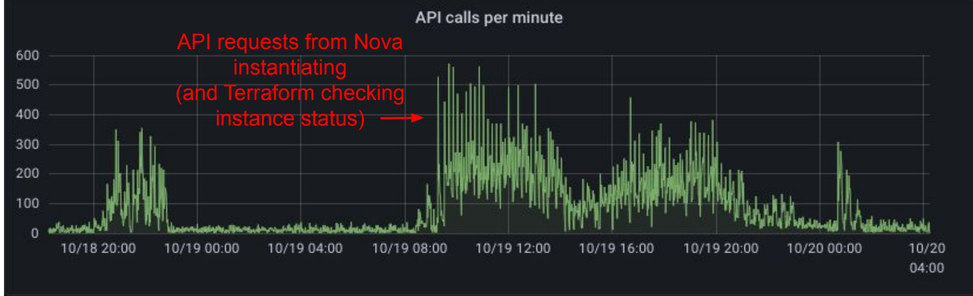

Bare Metal Events

instance deletion

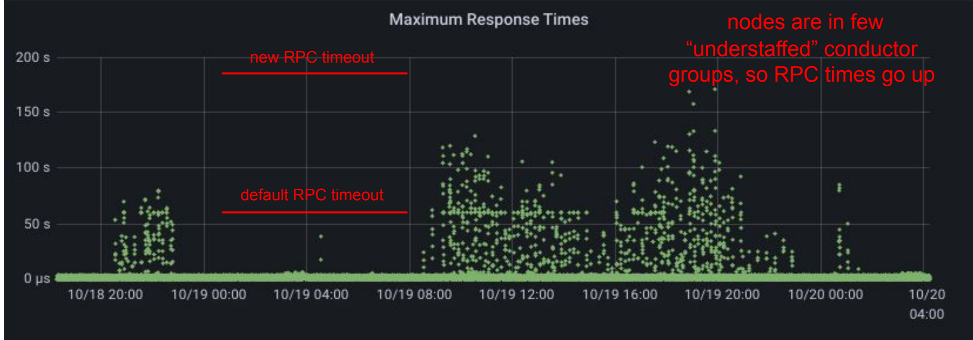
instance creation

RPC timeout change

automatic testing


API calls per minute

API requests from Nova instantiating (and Terraform checking instance status)

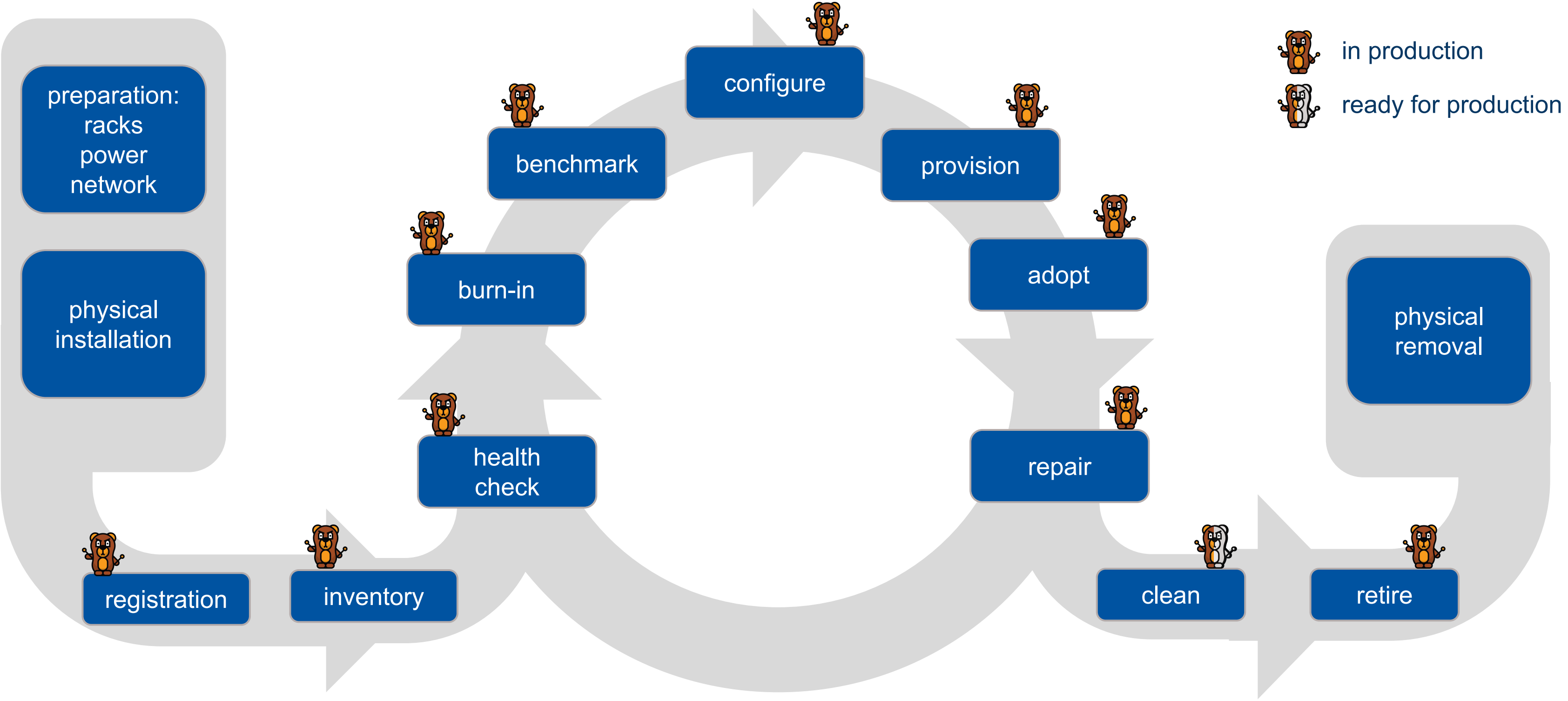

Maximum Response Times

new RPC timeout

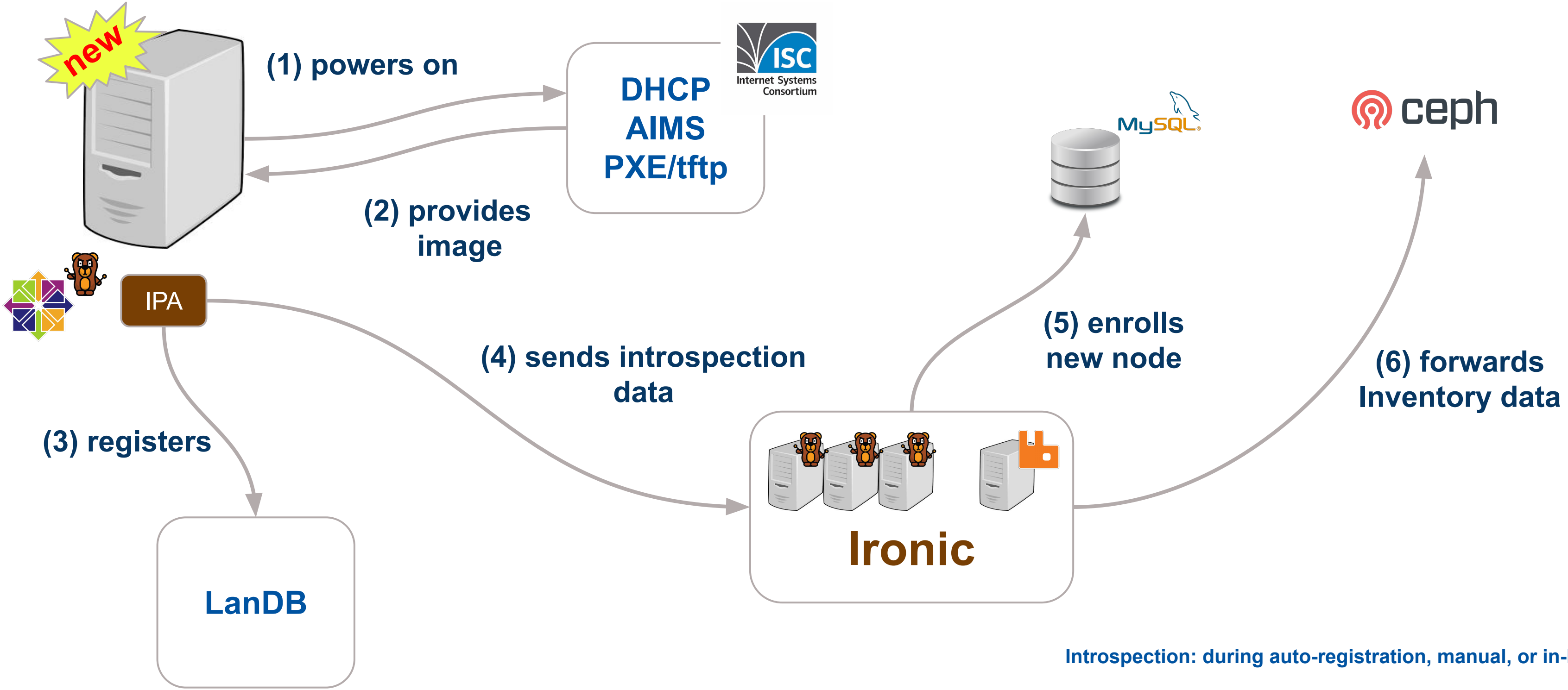
default RPC timeout

nodes are in few “understaffed” conductor groups, so RPC times go up

Server Life-cycle Management

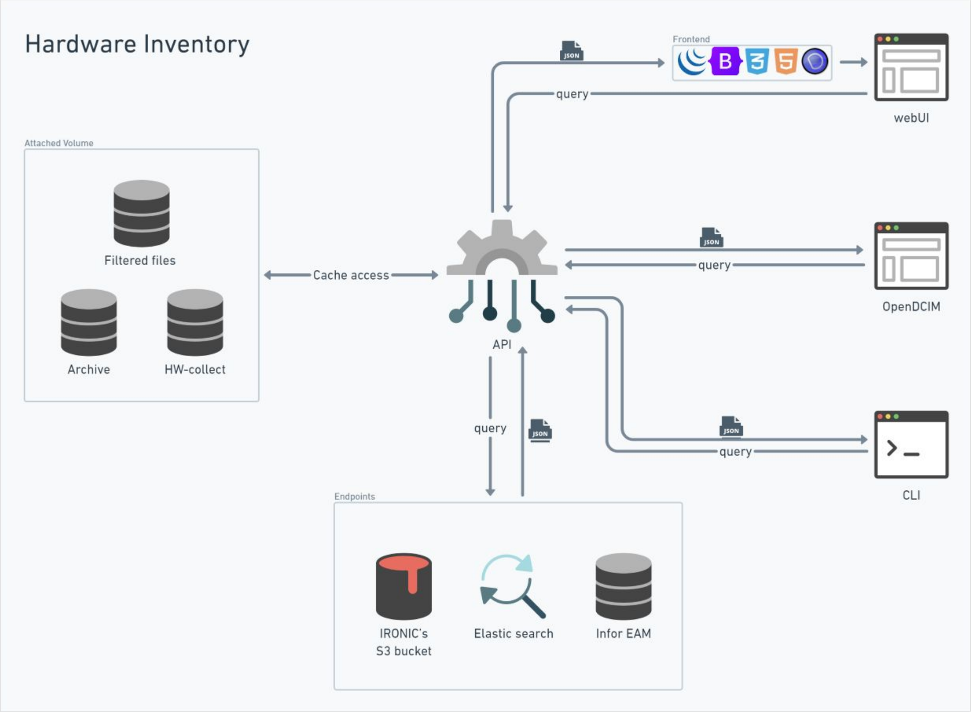


Auto-Registration with Ironic



Introspection: during auto-registration, manual, or in-band!

Hardware Inventory

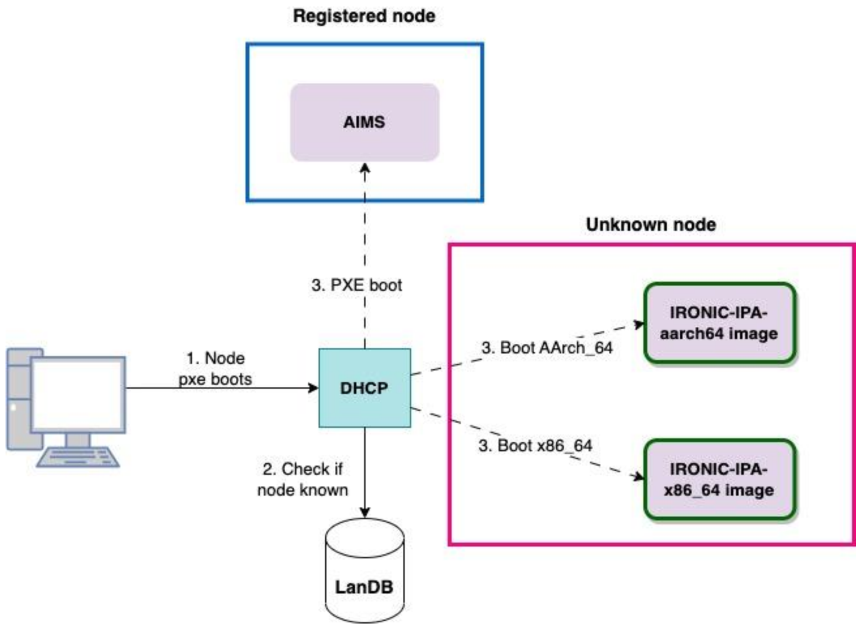


➔ **Pool data from Ironic + Infor EAM**

➔ **Combine with OpenDCIM**

➔ **Provide info via CLI or webUI**

Detour: Auto-registration (ARM)



- ➔ IPA image is architecture dependent!
- ➔ Automatically build for x86 & ARM
- ➔ DHCP decides based on PXE data

Hardware Burn-in

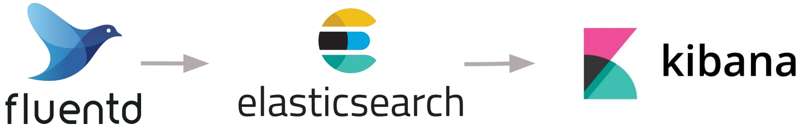
➔ Provoke early failure via stress-test

- ➢ CPU (stress-ng)
- ➢ Memory (stress-ng)
- ➢ Disk (fio)
- ➢ Network (fio)

➔ Integrated upstream

- ➢ Released with Xena
- ➢ Implemented via cleaning steps

➔ Real-time log handling



➔ Network requires pairing

- ➢ Initial version: static pairs
- ➢ Works, but clumsy

➔ Dynamic pairing: distributed arbiter

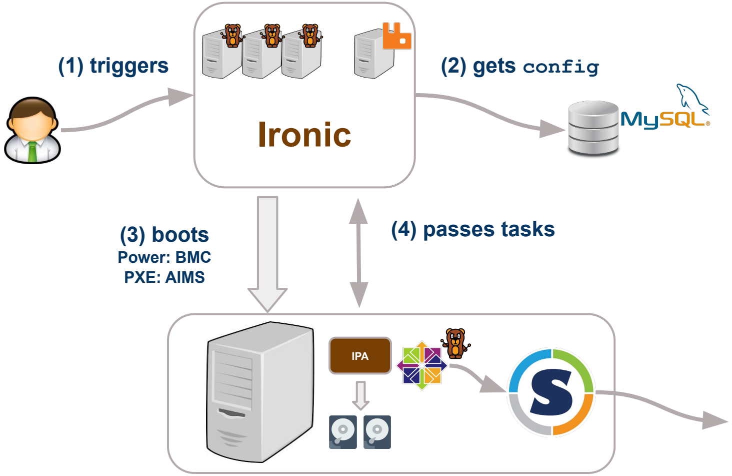
- ➢ OpenStack/tooz with ZooKeeper
- ➢ delivery and interface separation
- ➢ merged upstream, in-prod downstream

APACHE ZooKeeper™

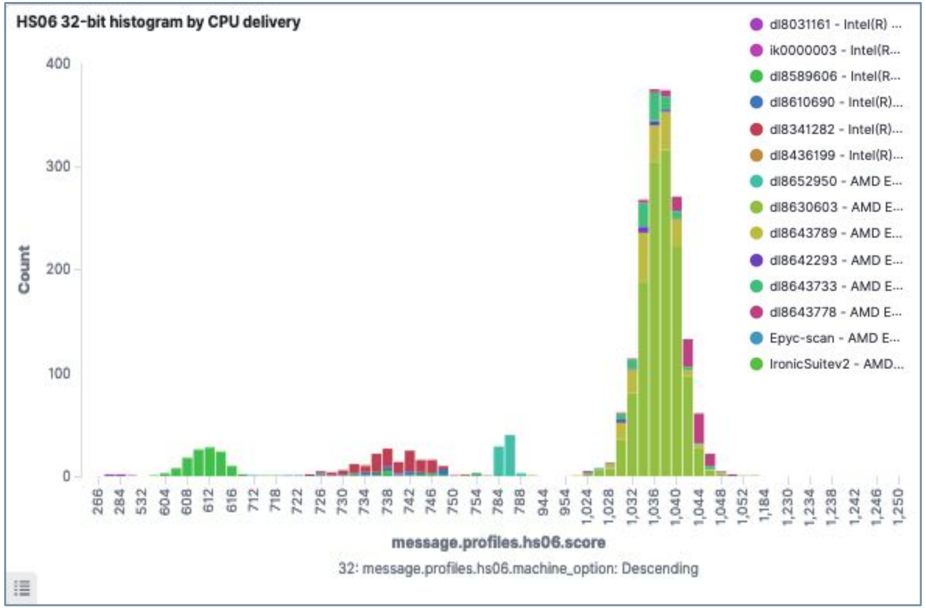
Benchmarking

➔ Based on a cleaning step

- ➢ Downloads singularity image
- ➢ Executes configured benchmarks
- ➢ Sends results into OpenSearch

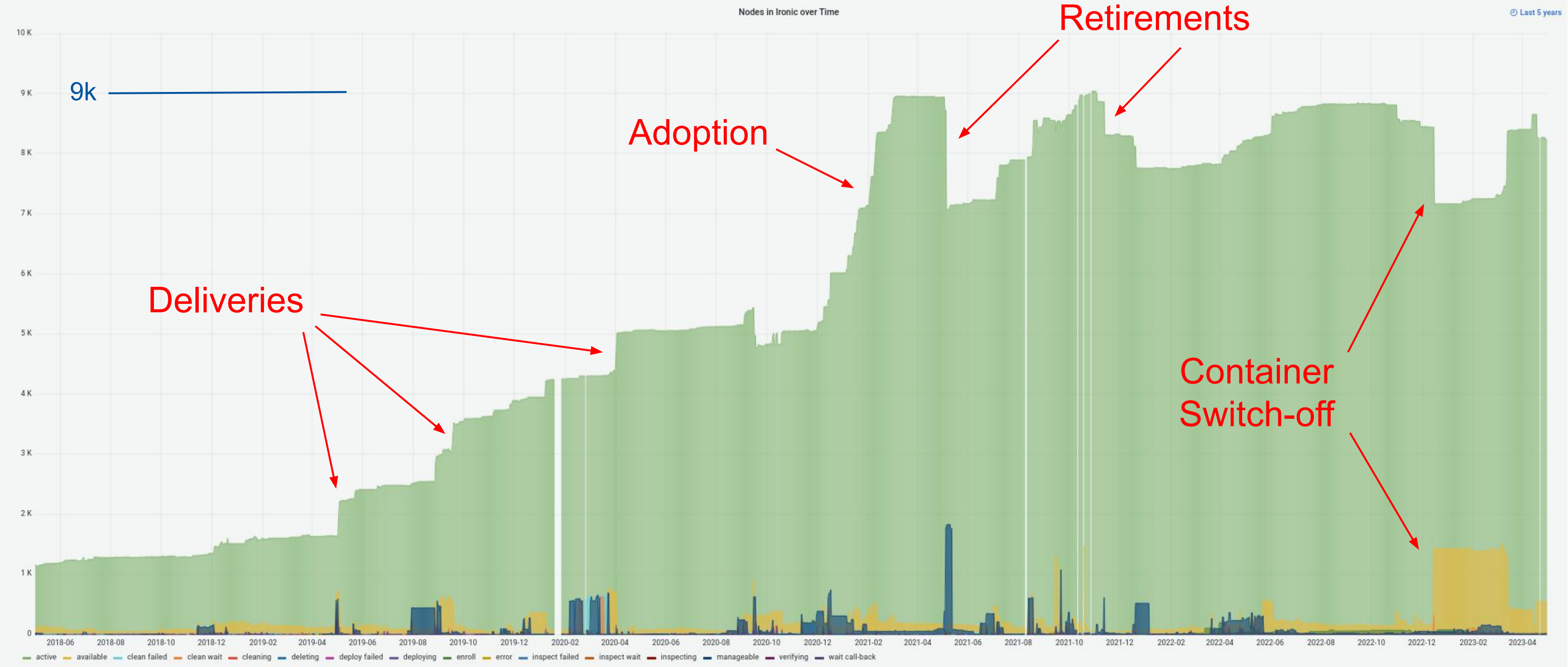


HEP Spec06 32bit by delivery



Ironic Evolution over Time

Number of Ironic-managed Nodes



9k

Retirements

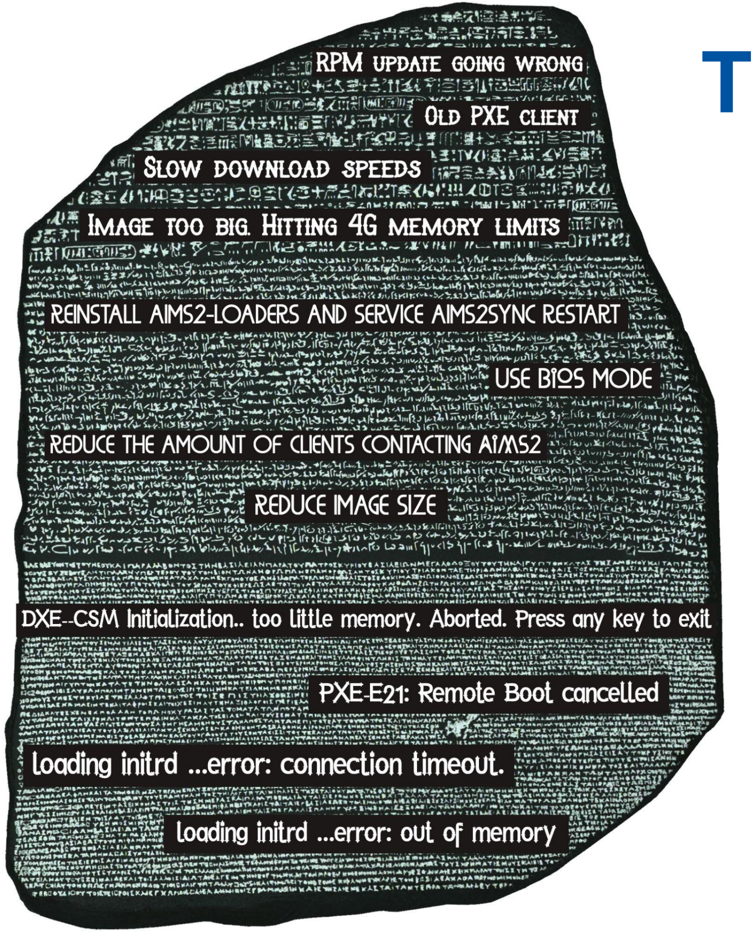
Adoption

Deliveries

Container Switch-off

2018 2020 2022

The GRUBsetta Stone



RPM UPDATE GOING WRONG

OLD PXE CLIENT

SLOW DOWNLOAD SPEEDS

IMAGE TOO BIG. HITTING 4G MEMORY LIMITS

REINSTALL AIMS2-LOADERS AND SERVICE AIMS2SYNC RESTART

USE BIOS MODE

REDUCE THE AMOUNT OF CLIENTS CONTACTING AIMS2

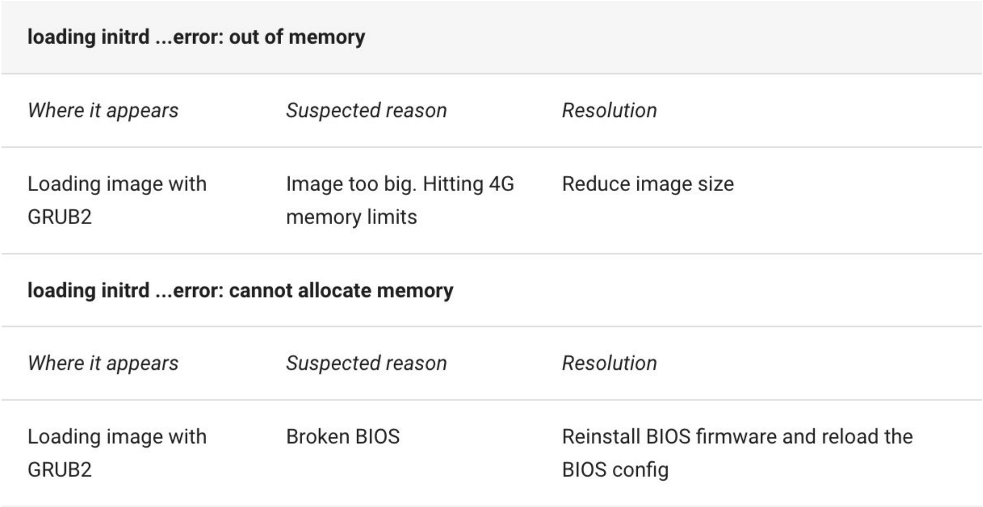
REDUCE IMAGE SIZE

DXE--CSM Initialization.. too little memory. Aborted. Press any key to exit

PXE-E21: Remote Boot cancelled

loading initrd ...error: connection timeout.

loading initrd ...error: out of memory



loading initrd ...error: out of memory

Where it appears	*Suspected reason*	*Resolution*
Loading image with GRUB2	Image too big. Hitting 4G memory limits	Reduce image size

loading initrd ...error: cannot allocate memory

Where it appears	*Suspected reason*	*Resolution*
Loading image with GRUB2	Broken BIOS	Reinstall BIOS firmware and reload the BIOS config

Thanks!

Arne.Wiebalck@cern.ch