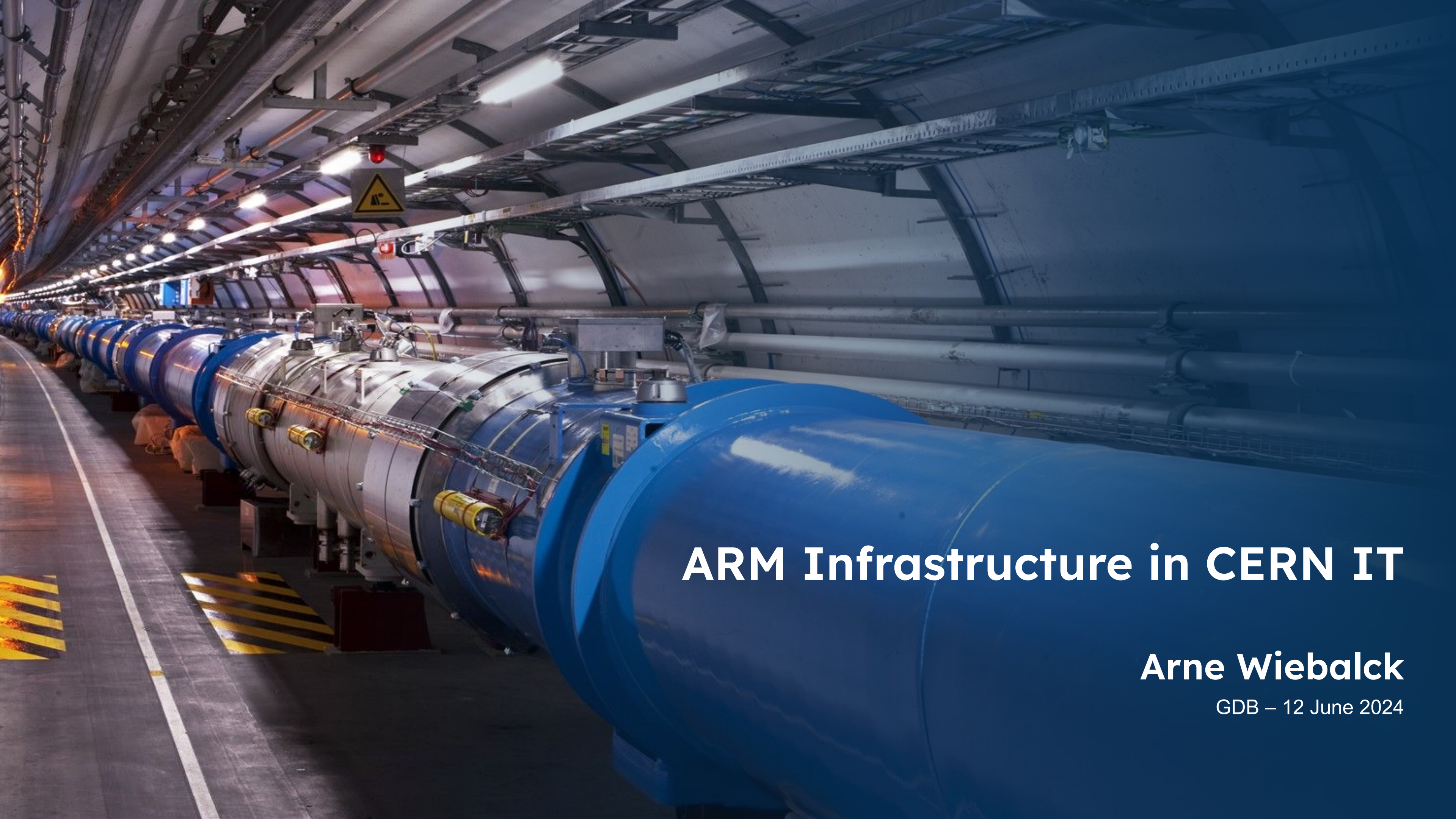


ARM Infrastructure in CERN IT

Arne Wiebalck

GDB – 12 June 2024

Overview

→ ARM infrastructure integration

→ ARM resource inventory

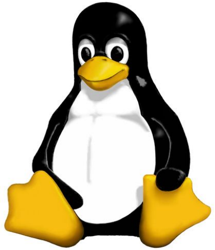
→ ARM resource provisioning

→ Power Considerations

Integration: Bootstrapping ARM Images

➔ **Linux image created before access to first ARM hardware**

- ➢ `koji` built AArch64 packages on a AARCH64 VM running on “Franken”-x86_64
- ➢ `koji imagebuild` (QEMU emulator, kickstart file, snapshot) ➩ ARM image
- ➢ Later (and easier setup): ARM VM as builder in Oracle cloud
- ➢ Now: x86 and AARCH64 images built the same way on local builder VMs



Integration: Bootstrapping Physical ARM (1)

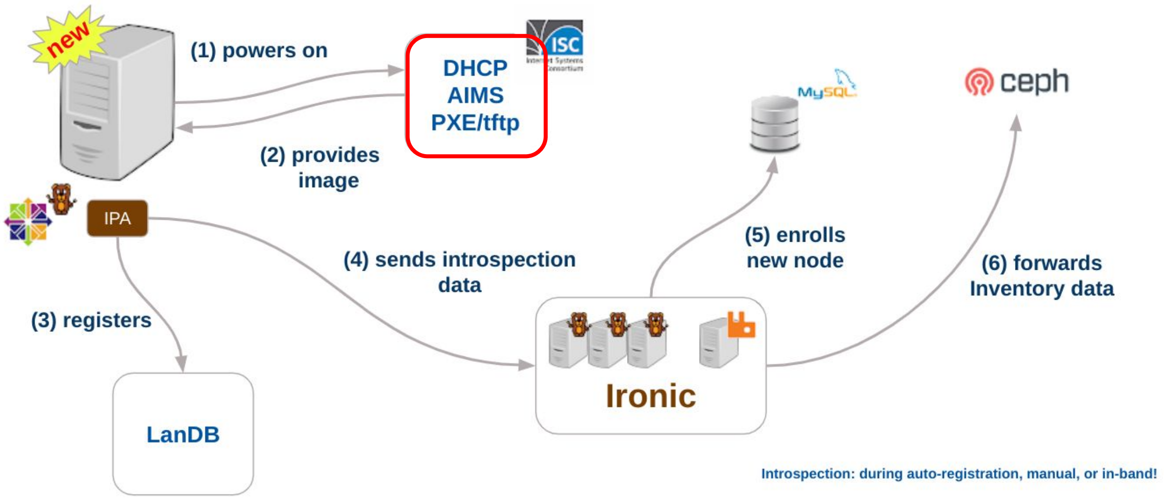
➔ Spawning physical machines

- ➢ OpenStack Ironic is CERN’s framework for bare metal fleet management
- ➢ Multi-arch node discovery

openstack

IRONIC

an OpenStack Community Project



new

(1) powers on

DHCP
AIMS
PXE/tftp

(2) provides image

IPA

(3) registers

LanDB

(4) sends introspection data

Ironic

(5) enrolls new node

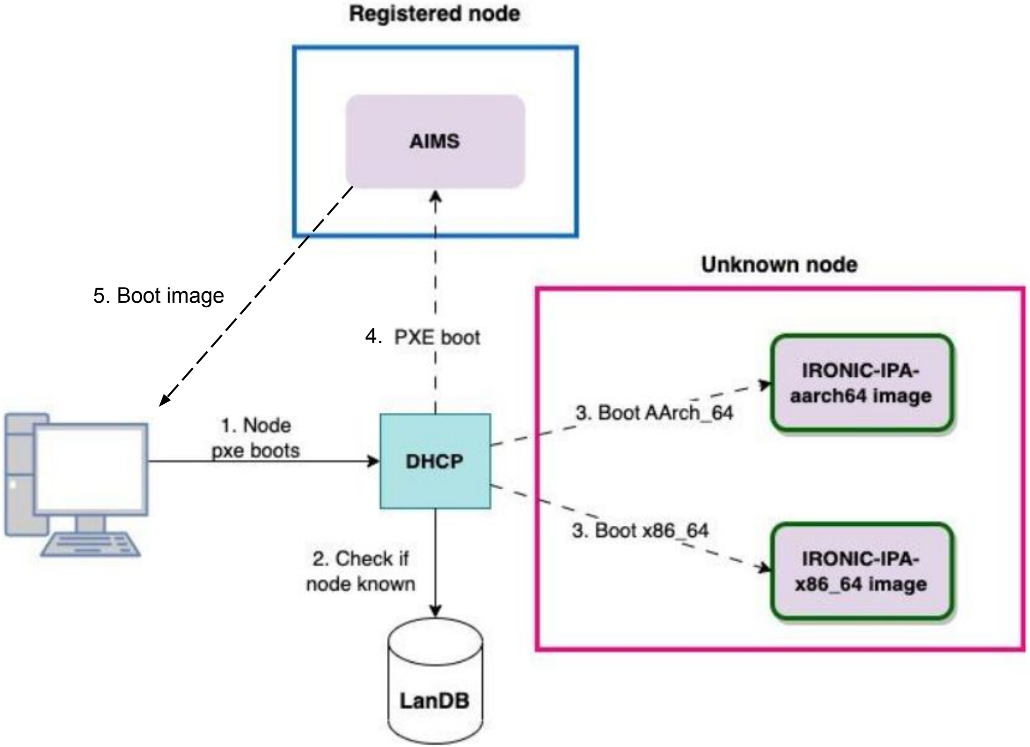
MySQL

ceph

(6) forwards Inventory data

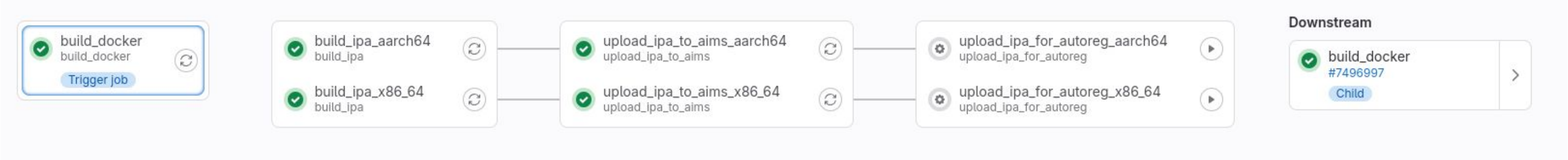
Introspection: during auto-registration, manual, or in-band!

Integration: Bootstrapping Physical ARM (2)

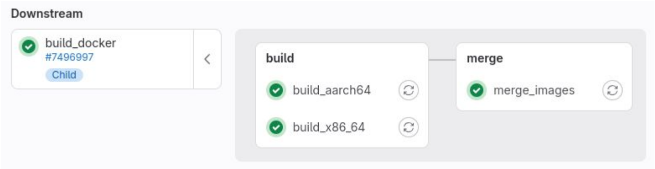


- ➔ DHCP/PXE/TFTP (AIMS) need to “decide” on architecture
- ➔ Need {x86_64,aarch64} PXE images
- ➔ Need {x86_64,aarch64} docker images for CI image building pipeline

Integration: Bootstrapping Physical ARM (3)



- ➔ **Code changes hitting the gitlab repository trigger image (re-)generation**



- ➔ **Pipeline builds multi-arch docker images**
  - ➢ Containers are then spawned to build the IPA (PXE/deploy) images

- ➔ **This requires an AARCH64 GitLab runner …**

Bootstrapping ARM VMs

➔ **Spawning virtual machines … easy!**

- ➢ On ARM based hypervisors
- ➢ Required EL8 and fixing a libvirt bug
- ➢ Image capabilities filtering to direct request to correct host
- ➢ Detailed talks at HEPiX & CHEP

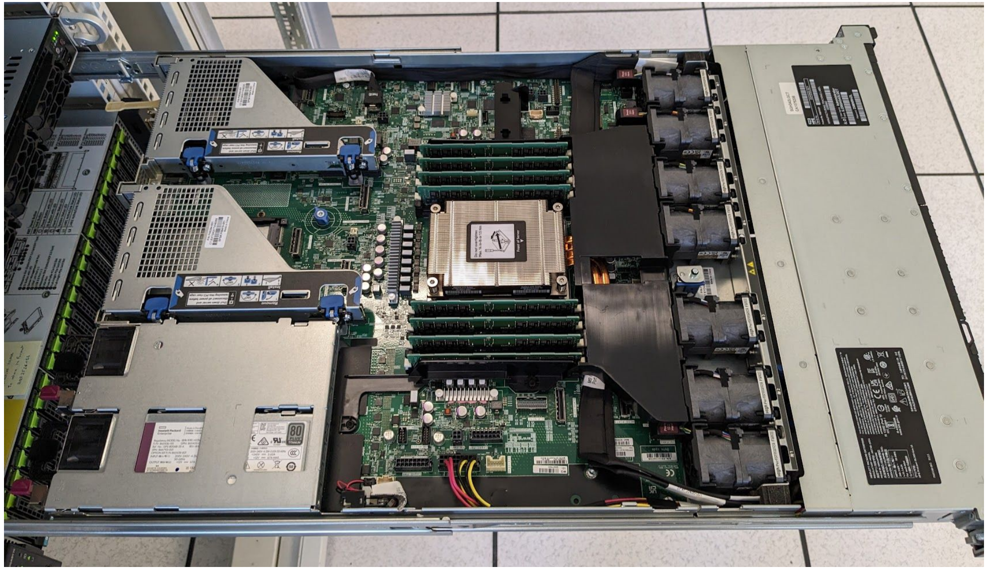
Inventory

48x Ampere Altra systems

- 80-core 3GHz Altra CPU
- 256GB of DDR4 RAM
- 1x 3.84TB NVMe
- 10GbE NIC
- On 3 different platforms:
  - GIGABYTE (5) “Mt Snow” R272-P30
  - SUPERMICRO (19) “Mt Hamilton” ARS-110M-NR
  - HPE (24) ProLiant RL300 Gen11

GIGABYTE SUPERMICRO Hewlett Packard Enterprise

AMPERE



Provisioning: Cloud (physical & virtual)

➔ **Virtual machines instantiated with ARM VM flavors and images**

- ➢ `openstack server create --flavor a1.xlarge --image “ALMA9 - aarch64”`

➔ **Physical nodes instantiated with ARM BM flavors and images**

- ➢ Instantiated just like virtual machines with a specific flavor
- ➢ `openstack server create --flavor p1.ca1076207.S513-C-IP503 ...`

➔ **GitLab runners, LxPlus nodes, …**

Provisioning: Batch

- **Used public cloud ARM VMs, now all physical nodes on-premises**
- **In batch the main issue is to ensure that jobs only run on the intended arch**
- **For “locally” submitted batch jobs, usual HTCondor semantics apply**
  - Jobs automatically have a requirement for the Arch of the submitting machine
  - Only jobs submitted by users submitting from lxplus-arm, or explicitly requesting aarch64 in their submit files will run on ARM machines
- **Grid lacks this default, so the HTCondor-CE router ensures the ARCH**
  - Submitter jdl should use `+WantARM=True` or `remote_queue=ARM`

Provisioning: HTCondor-CE routing for ARM

➔ **Make x86 the default and …**

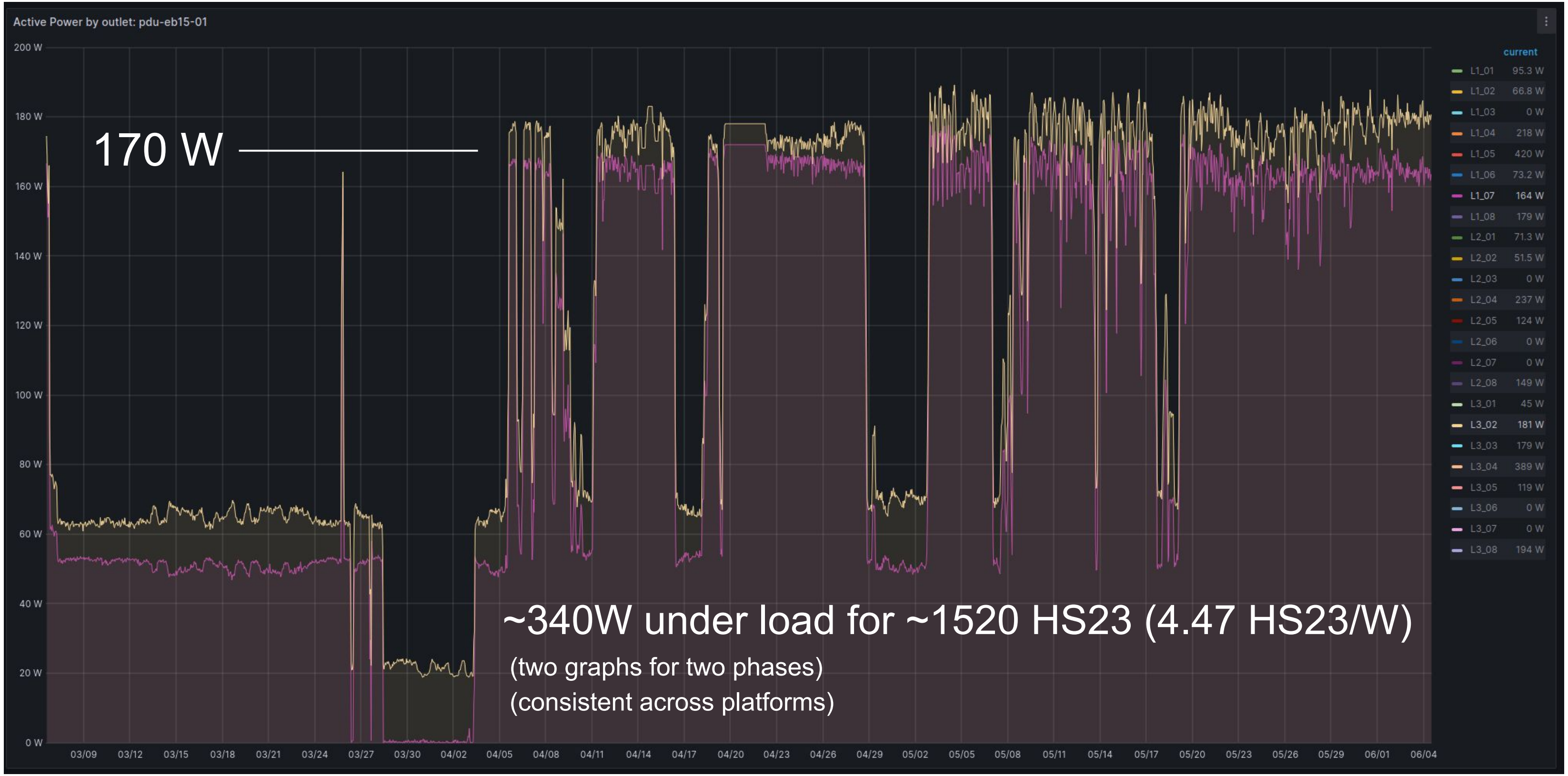
```
JOB_ROUTER_TRANSFORM_x86 @=end
    NAME x86
    REQUIREMENTS (MY.WantARM =!= True && MY.queue =!= "ARM")
    COPY Requirements PreArchRequirements
    SET Requirements (Arch =?= "X86_64" && PreArchRequirements)
@end
```

… route ARM jobs to AARCH64 resources

```
JOB_ROUTER_TRANSFORM_Arm @=end
    NAME Arm
    REQUIREMENTS (MY.WantARM =?= True) || (MY.queue =?= "ARM")
    SET WantARM True
    COPY Requirements PreArchRequirements
    SET Requirements (Arch =?= "aarch64" && PreArchRequirements)
@end
```

Power Considerations Disclaimer

Single ARM Server Power Consumption



Active Power by outlet: pdu-eb15-01

		current
🟩	L1_01	95.3 W
🟨	L1_02	66.8 W
🟦	L1_03	0 W
🟧	L1_04	218 W
🟥	L1_05	420 W
🟦	L1_06	73.2 W
🟪	L1_07	164 W
🟪	L1_08	179 W
🟩	L2_01	71.3 W
🟨	L2_02	51.5 W
🟦	L2_03	0 W
🟧	L2_04	237 W
🟥	L2_05	124 W
🟦	L2_06	0 W
🟪	L2_07	0 W
🟪	L2_08	149 W
🟩	L3_01	45 W
🟨	L3_02	181 W
🟦	L3_03	179 W
🟧	L3_04	389 W
🟥	L3_05	119 W
🟦	L3_06	0 W
🟪	L3_07	0 W
🟪	L3_08	194 W

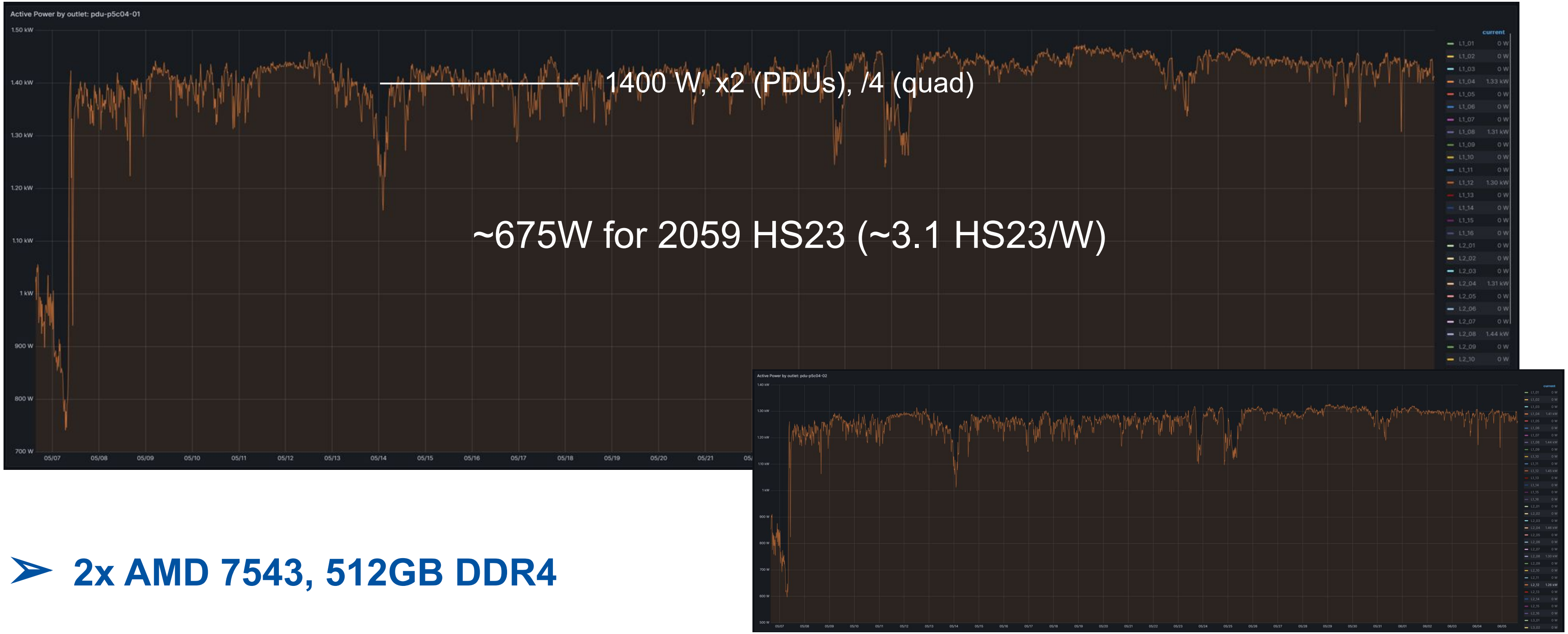
170 W

~340W under load for ~1520 HS23 (4.47 HS23/W)

(two graphs for two phases)

(consistent across platforms)

Single AMD Server Power Consumption



1400 W, x2 (PDUs), /4 (quad)

~675W for 2059 HS23 (~3.1 HS23/W)

- ➢ **2x AMD 7543, 512GB DDR4**

Power Considerations

This is not a scientific study!

	Power [W]*	HS23	HS23/Power	Price**	HS23/Power/Price
ARM	340	1520	4.47	1	4.47
AMD	675	2059	3.1	1.15	2.7
Δ (ARM ↦ AMD)	+200%	+73%	-36%	+15%	-50%

* At PDU (prod, not benchmark)
**ARM price for initial purchase

CERN

www.cern.ch