CNAF experience with ARM

Andrea Chierici

On behalf of INFN-T1

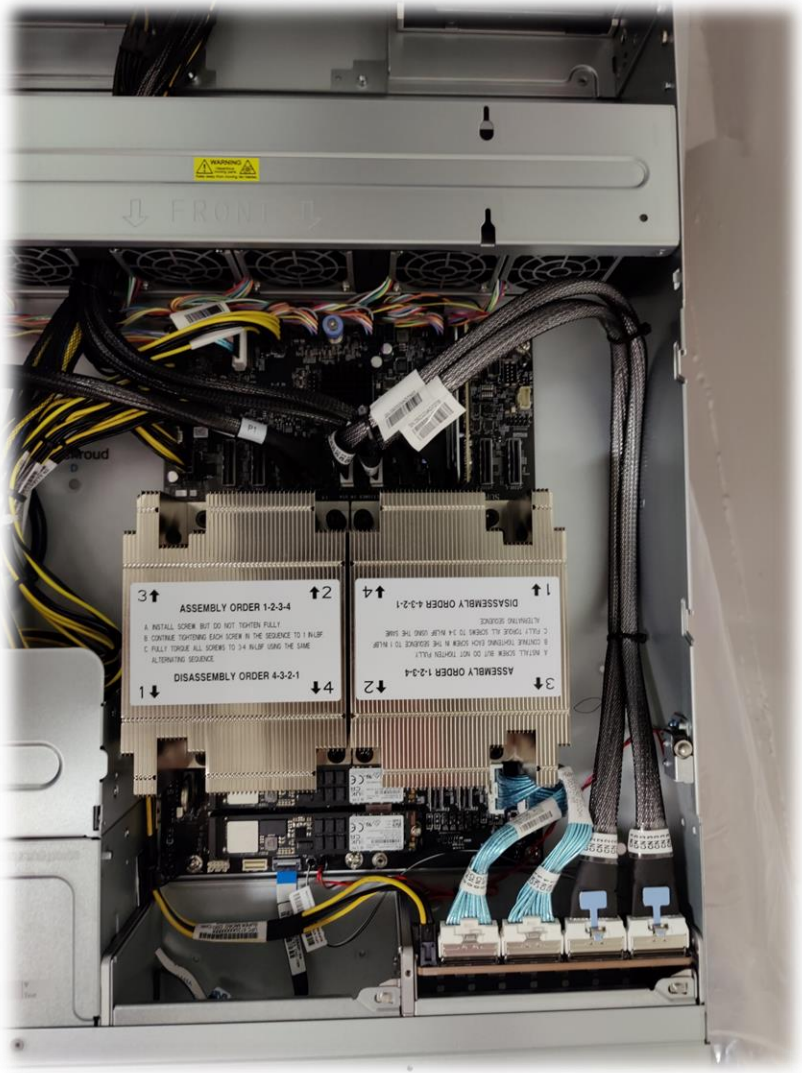
Farming group

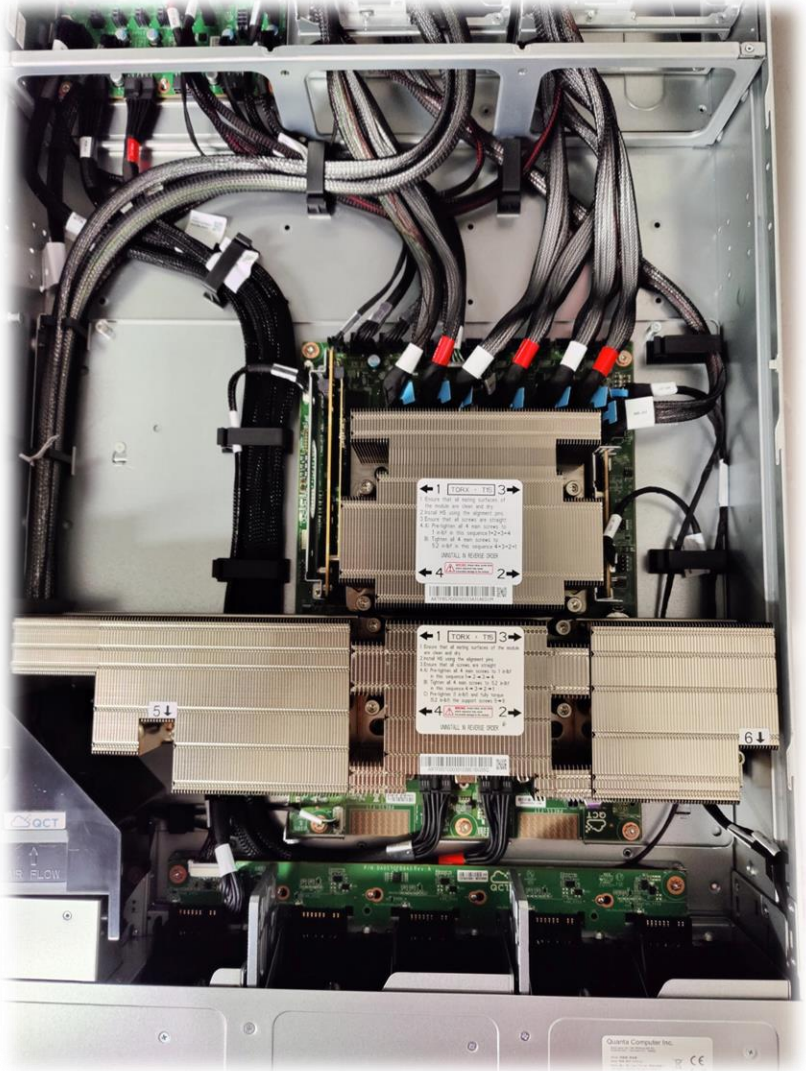
Lorenzo Rinaldi (atlas)

Resources available

- **4x** 2U Dual socket ARM – Ampere Altra Max
  - 2x Ampere AltraMax M128-30 128Core 2.8Ghz 250W
  - 1TB RAM
  - 2x NVMe U.3 3.84TB discs
  - 1x Dual port SFP28 ethernet
- 1x GraceHopper superchip (Quanta server)
  - CPU Memory: 480GB
  - GPU Memory: 96GB
  - 2x NVMe disks (1.92TB + 960GB)
  - 1x 100GBit ethernet port
- 1x Grace superchip (supermicro server)
  - CPU Memory: 480GB
  - 2x NVMe disks (7.68TB + 480GB)
  - 1x SFP28 ethernet port

Some photos


FRONT
ASSEMBLY ORDER 1-2-3-4
DISASSEMBLY ORDER 4-3-2-1



SAMSUNG
E1.S
PM9A3

Main differences – hardware side

- Procurement not so easy
  - Vendors bidded with incredible variation in price
  - Hardware support: no on-site, only on-center
    - This is slowly changing
- Form factor
  - Only 2U, 1 motherboard
  - No possibility for customization (grace/hopper net adapter)
- BMC and firmwares in general don’t seem very stable
  - BMC is slow
  - Server takes ages to boot
  - Network installation not always possible

Main differences – software side

- Started with Alma 8, since puppet build was not available on alma9
- GPFS initially not available for aarch64
  - Now it's available but only with a very recent version, not yet in production at INFN-T1
- Grid middleware not yet available
- Had to make experiments aware of the different resources and agreed on specific workflows

Some figures

- Ampere altra max
  - Each node provides 3754 hepscore, 14,66 hepscore/core
  - BMC not reachable, impossible to get power consumption ☹
- Grace
  - The node provides 4459 hepscore, 30,97 hepscore/core
  - The node consumes 1kW average, 4,459 hepscore/watt

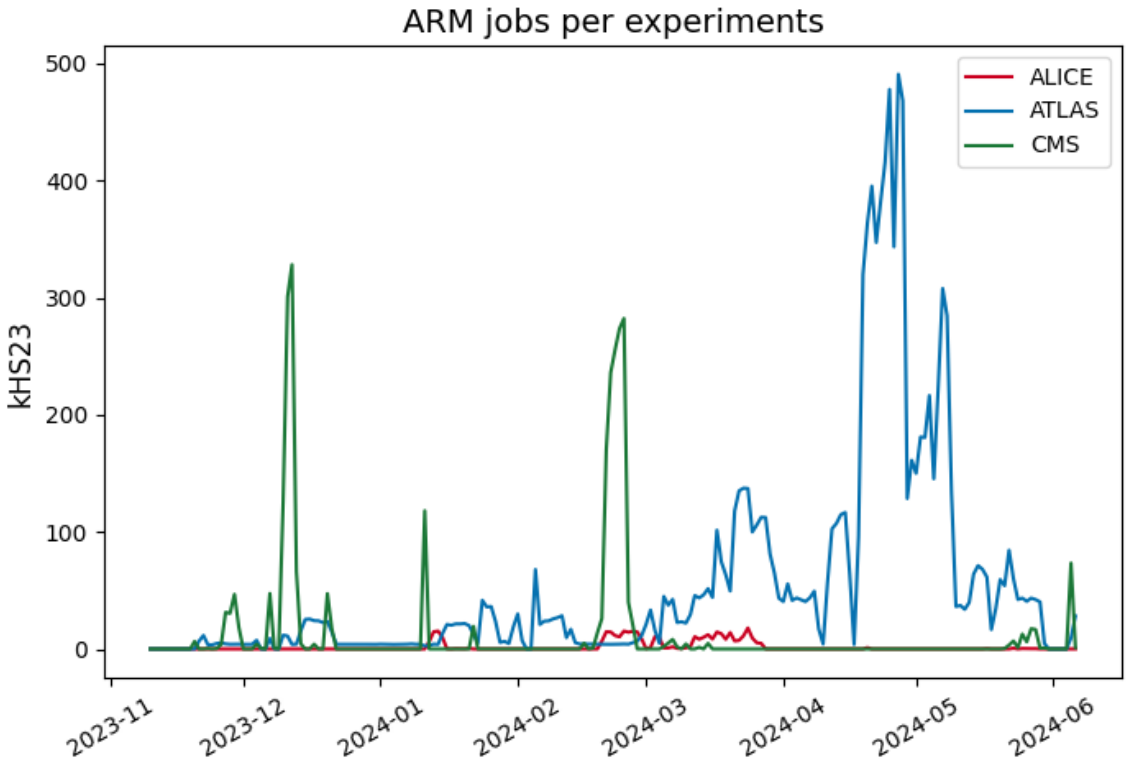
As a comparison

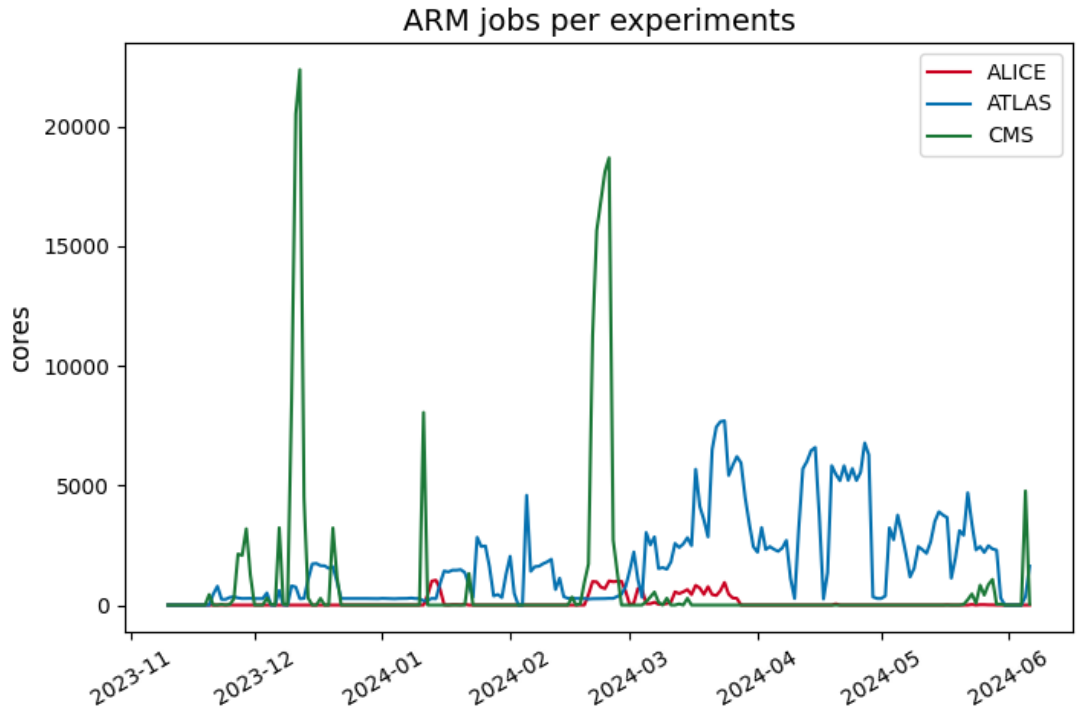
- Leonardo GP
  - The node provides 2880 hepscore, 25,71 hepscore/core (ht off)
  - The node consumes 799W (as per tender doc), 3,6 hepscore/watt

Technicalities

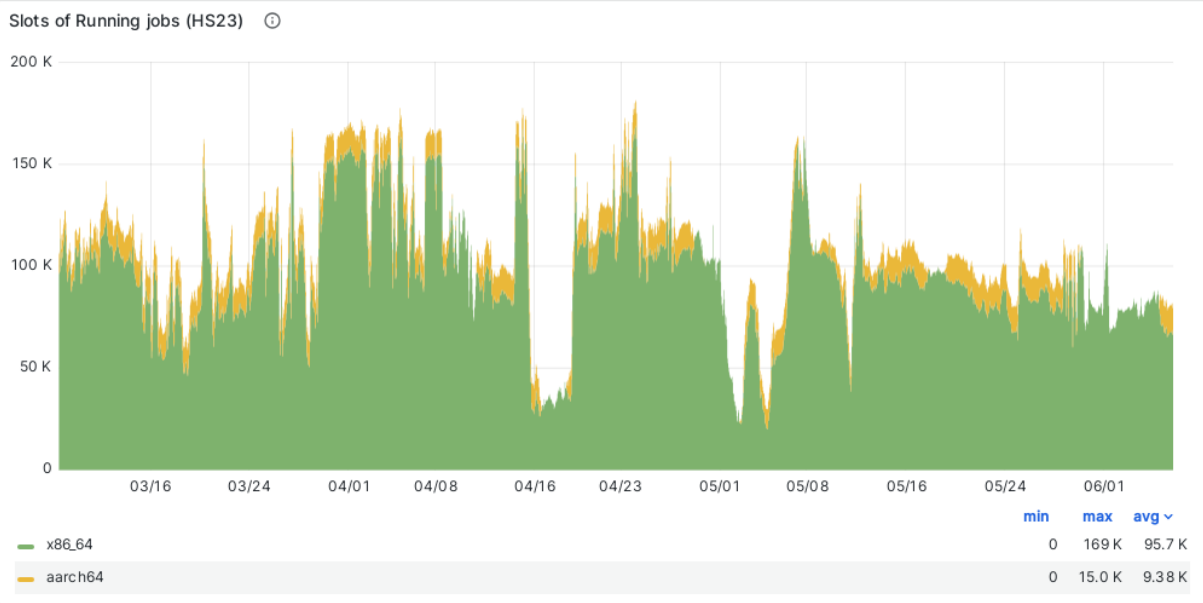
- HTCondor: classad «wantarm»
- Not for every exp
  - Atlas wants attribute «queue=arm», internally becomes wantarm classad
- Main users
  - Atlas and CMS
  - Alice tested, but then did not came back
  - LHCb showed interest, but not tested yet
- Only LHC experiments showed up
- Provided an aarch64 UI to allow for compilation and initial testing/submission
  - Available for all

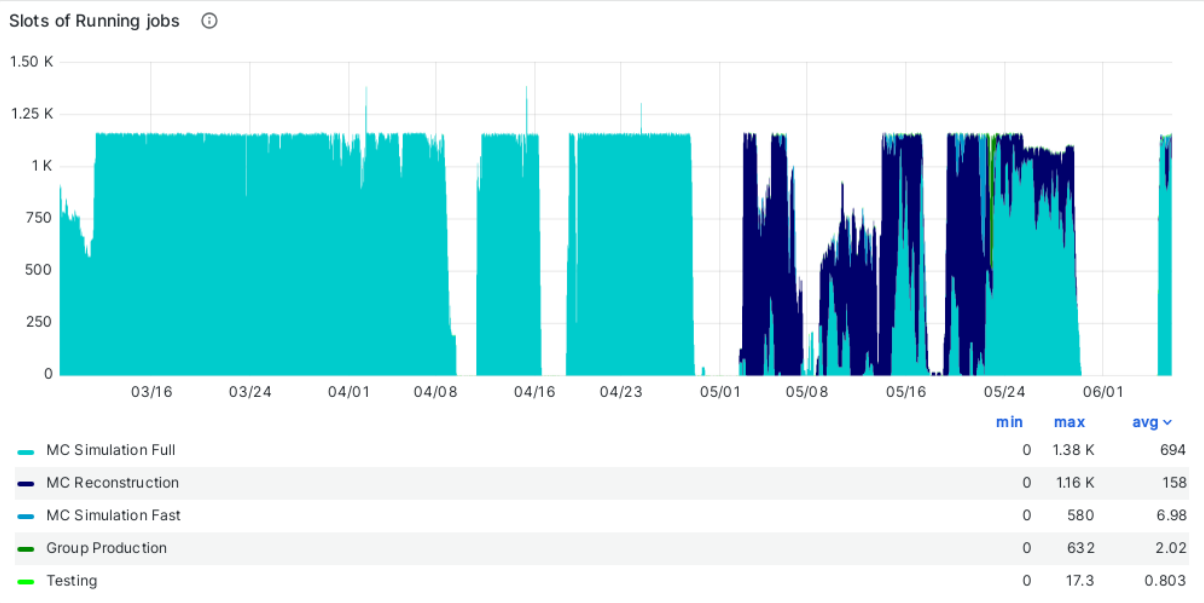
Resource utilization





Atlas point of view





- Peak usage: 19.9 kHS23
- ~15% INFN-T1 pledge
- Running since nov 2023, on el9 container, interfaced to PanDA production system

- MC simulation (FULL and FAST)
- MC reconstruction
- Few group-prord/user/testing

How fast atlas jobs run on aarch64?

MC SIM FULL	HS23sec/evnt
aarch64	1657
x86_64	2144

MC SIM FAST	HS23sec/evnt
aarch64	389
x86_64	532

MC RECO	HS23sec/evnt
aarch64	410
x86_64	365

- Source: ATLAS grafana, renormalized with measured HS23*core
- Measurements on systems in production are consistent with HS23 measurement (courtesy by Domenico Giordano)
- Differences could come from nodes setup and usage (HT, job concurrency, ram, max freq, ...)

HS23 and ATLAS_SIM ratios

SITE ↓	CPU_MODEL	HS23/HS23_at_GLASGOW_ARM	ATLAS_SIM/ATLAS_SIM_at_GLASGOW_ARM
UKI-SCOTGRID-GLASGOW_CEPH	AMD EPYC 7452 32-Core Processor	0.777	0.843
UKI-SCOTGRID-GLASGOW_ARM	Neoverse-N1	1.00	1.00
INFN-CNAF_ARM	Neoverse-V2	1.90	1.83
INFN-CNAF	AMD EPYC 7282 16-Core Processor	0.781	0.793
CERN-ARM	Neoverse-N1	1.09	1.07
CERN	AMD EPYC 7302 16-Core Processor	0.958	0.966

Conclusions

- ARM machines are interesting, but
  - service level from vendor is currently not adequate
  - some sw layers are missing
  - in a multi vo environment (with #vo >> LHC) ARM machines decrease flexibility in allocation
- Recent announcements from AMD and Intel may decrease interest in these chips (# of cores)
- Will keep tracking the market to stay up-to-date
  - 2 risc-v boards are expected to be delivered within summer