

# ATLAS usage of ARM

*Frédéric Derue, LPNHE Paris*

LCG France meeting, CC-IN2P3 Lyon

28<sup>st</sup> november 2023

Largely based on presentation of J. Elmsheuser (BNL) at CHEP2023 ([link](#))  
+ recent presentations at ATLAS Software & Computing week

- **Access to ARM resources**

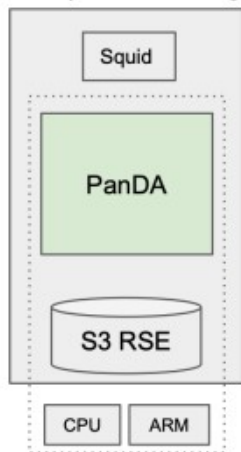
J. Elmsheuser (BNL) at CHEP2023 ([link](#))

- CERN (usually single instances)
  - lxplus8/9-arm (in Oracle Cloud)
  - ATLAS build machine provisioned by CERN IT and Openlab on prem and in Oracle Cloud (Neoverse and Cavium ThunderX2)
- AWS (Graviton2 and 3)
  - fully integrated in ATLAS PanDA/Rucio based production system
- Google cloud (Ampere Altra/Neoverse-N1)

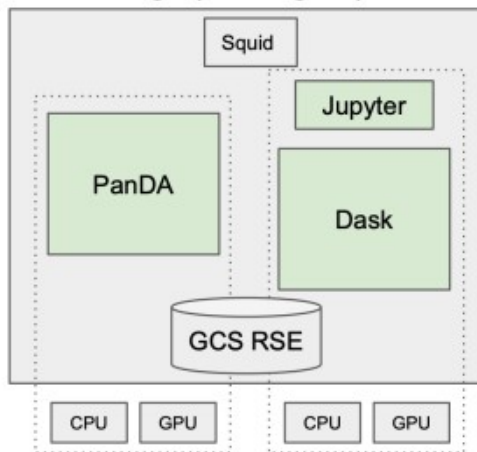
- **Current and future large scale resources**

- Fugaku HPC at Riken Center for Computational Science, Japan is #2 in TOP500 supercomputer list of November 2022 ([link](#))
- a few other HPCs plan to have ARM partitions
- grid sites are starting to be interested in ARM when the experiment software is validated (e.g Glasgow, CNAF)

Amazon (CSU Fresno grant)



Google (ATLAS grant)



- self-container, cloud-native, vendor-agnostic, auto-scaled, auto-healing
- Rucio
  - object store + signed URL + http protocol
  - integrated in Rucio and WLCG MW
- PanDA
  - Harvester submitting to K8S directly
  - CVMFS K8S plugin
  - Frontier squid either in K8S cluster or as separate load balanced VM group
  - auto-scaling : no jobs, almost no cost

- Full aarch64 grid setup available with OS container, middleware, Kubernetes etc.
- More details in talks from Fernando Barreiro ([link](#)) and Mario Lassnig ([link](#)) at CHEP 2023 conference

arm-Node = 2x160 cores

x86-Node  $\approx$  1000 cores

## External View

- Nothing changes for other VO's. They submit jobs as normal to the same queue.

**UKI-SCOTGRID-GLASGOW\_CEPH**  
ce01.gla.scotgrid.ac.uk (x86)  
ce04.gla.scotgrid.ac.uk (x86)

**UKI-SCOTGRID-GLASGOW\_ARM**  
ce\_test.gla.scotgrid.ac.uk (x86)

- Users that want to access the ARM resources at Glasgow have to submit to a specific queue

## Internal Architecture

ce01...ce04

ARC-CE

UKI-SCOTGRID-GLASGOW\_CEPH  
(x86 resources)

ce\_test

ARC-CE

UKI-SCOTGRID-GLASGOW\_ARM  
(aarch64 resources)

**CASSOWARY**  
(Condor 8)

**HUMMINGBIRD**  
(Condor 10)



- Want to move to condor 10 on all worker nodes so we will have two types of architecture in one condor pool.

- Have to ensure that jobs run on nodes that have the right architecture!

## Setup: ATLAS-side

- PanDA setup stores data locally (now)
- Points to our temporary CE and our condor\_arm queue in the ARC
- aarch architecture only
- Now we just have to run some proper jobs on them

### PanDA Queue: UKI-SCOTGRID-GLASGOW\_ARM

#### Object details

|                |                                    |
|----------------|------------------------------------|
| ID             | 930                                |
| Name           | UKI-SCOTGRID-GLASGOW_ARM           |
| PanDA Site     | UKI-SCOTGRID-GLASGOW-CEPH          |
| Description    |                                    |
| Default object | UKI-SCOTGRID-GLASGOW_VIRTUAL       |
| Corecount      | 8 <span>overwritten</span>         |
| Corepower      | 8.74 <span>inherited.rcsite</span> |
| Coreenergy     | 10.0 <span>inherited.rcsite</span> |

#### State

|                        |                                                            |
|------------------------|------------------------------------------------------------|
| Object state           | ACTIVE                                                     |
| State comment          | Object was cloned from UKI-SCOTGRID-GLASGOW-CEPH via WebUI |
| Last modification date | 2023-08-11 13:26:51.509877                                 |

#### Associated DDMEndpoints

| DDMEndpoint                           | Type        | Experiment Site      | Activities                                    |
|---------------------------------------|-------------|----------------------|-----------------------------------------------|
| UKI-SCOTGRID-GLASGOW-CEPH_DATADISK    | DATADISK    | UKI-SCOTGRID-GLASGOW | pl/0, read_lan/0, write_lan/0                 |
| UKI-SCOTGRID-GLASGOW-CEPH_SCRATCHDISK | SCRATCHDISK | UKI-SCOTGRID-GLASGOW | write_lan_analysis/0, read_lan/1, write_lan/1 |

Showing 1 to 2 of 2 entries

#### Associated Queues

| CE                         | Queue      | flavour | version | qstatus    | site                 | cputime | wclock | ETF | Ops |
|----------------------------|------------|---------|---------|------------|----------------------|---------|--------|-----|-----|
| ce-test.gla.scotgrid.ac.uk | condor_arm | ARC-CE  | None    | production | UKI-SCOTGRID-GLASGOW | 0       | 0      | ✗   |     |

Showing 1 to 1 of 1 entries

#### Associated PandaQueue architectures

| Type | Architectures | Vendor | Instructions | Model | Ops |
|------|---------------|--------|--------------|-------|-----|
| cpu  | aarch64 excl  |        |              |       | ✗   |



## ATLAS Testing

- The ARM nodes are running a job starting with 50M events and with simulation, reconstruction and derivation tasks.
- Running since Late August, and still running!

| Evgen                                                                                                                | Evgen Merge | Simul | Merge | Digi | Reco | Rec Merge | Attfast | Attf Merge | TAG | Deriv | Deriv Merge |                                                                                                                           |
|----------------------------------------------------------------------------------------------------------------------|-------------|-------|-------|------|------|-----------|---------|------------|-----|-------|-------------|---------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                      |             |       |       |      |      |           |         |            |     |       |             | Replace empty <input type="checkbox"/>                                                                                    |
| 0 + 601229/mc.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.py<br>(Fullsim)(0)PhPy8EG_A14_ttbar_hdamp258p75_SingleLep       |             |       |       |      |      |           |         |            |     |       |             | mc23_13p6TeV.601229.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.merge.EVNT.e8514_e8528_tid33116249_00<br>events: 50000000      |
| <div> <div>s4159</div> <div>s4114</div> <div>r14799</div> <div>r14811</div> <div>p5667</div> <div>p5669</div> </div> |             |       |       |      |      |           |         |            |     |       |             | partially_submitted                                                                                                       |
| T: <div>running running aborted aborted</div>                                                                        |             |       |       |      |      |           |         |            |     |       |             | Produced events: 39370000                                                                                                 |
| 1 + 601229/mc.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.py<br>(Fullsim)(0)PhPy8EG_A14_ttbar_hdamp258p75_SingleLep       |             |       |       |      |      |           |         |            |     |       |             | mc23_13p6TeV.601229.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.merge.EVNT.e8514_e8528_tid33116249_00<br>events: -1 (50000000) |
| <div> <div>s4159</div> <div>s4114</div> <div>r14799</div> <div>r14811</div> <div>p5667</div> <div>p5669</div> </div> |             |       |       |      |      |           |         |            |     |       |             | partially_submitted                                                                                                       |
| T: <div>running aborted</div>                                                                                        |             |       |       |      |      |           |         |            |     |       |             | Produced events: 36570000                                                                                                 |
| 3 + 601229/mc.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.py<br>(Fullsim)(0)PhPy8EG_A14_ttbar_hdamp258p75_SingleLep       |             |       |       |      |      |           |         |            |     |       |             | mc23_13p6TeV.601229.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.merge.EVNT.e8514_e8528_tid33116249_00<br>events: -1 (50000000) |
| <div> <div>s4159</div> <div>s4114</div> <div>r14799</div> <div>r14811</div> <div>p5667</div> <div>p5669</div> </div> |             |       |       |      |      |           |         |            |     |       |             | partially_submitted                                                                                                       |
| T: <div>running</div>                                                                                                |             |       |       |      |      |           |         |            |     |       |             | Produced events: 35740000                                                                                                 |
| 7 + 601229/mc.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.py<br>(Fullsim)(0)PhPy8EG_A14_ttbar_hdamp258p75_SingleLep       |             |       |       |      |      |           |         |            |     |       |             | mc23_13p6TeV.601229.Phy8EG_A14_ttbar_hdamp258p75_SingleLep.merge.EVNT.e8514_e8528_tid33116249_00<br>events: -1 (50000000) |
| <div> <div>s4159</div> <div>s4114</div> <div>r14799</div> <div>r14811</div> <div>p5667</div> <div>p5669</div> </div> |             |       |       |      |      |           |         |            |     |       |             | submitted                                                                                                                 |
| T: <div>running</div>                                                                                                |             |       |       |      |      |           |         |            |     |       |             | Produced events: 34520000                                                                                                 |

Estimated CO<sub>2</sub> emission in grams, based on the weighted average of regional electricity sources for ATLAS Grid computing

| 34412357 task |          |         |      |                 |               |         |                      |         |                                      |                                     |                 |                                                                   |                                                  |       |                                  |         | regional electricity sources for ATLAS Grid computing |                                          |
|---------------|----------|---------|------|-----------------|---------------|---------|----------------------|---------|--------------------------------------|-------------------------------------|-----------------|-------------------------------------------------------------------|--------------------------------------------------|-------|----------------------------------|---------|-------------------------------------------------------|------------------------------------------|
| Task ID       | Campaign | Request | Type | Processing type | Working Group | User    | Nucleus              | Status  | N input files<br>finished<br>failed  | N input events<br>used              | N output events | HS06*sec<br>Expected<br>Total<br>done<br>failed                   | Time stamps:<br>created<br>last modified         | Cores | Priority:<br>original<br>current | Attempt | Tracker                                               | gCO <sub>2</sub> total<br>done<br>failed |
| 34412357      | MC23c    | 50572   | prod | simul           | AP_MCGN       | janders | UKI-SCOTGRID-GLASGOW | running | 25,000<br>(94%)<br>23,564<br>(0%) 11 | 50,000,000<br>(94.3%)<br>47,128,000 | 47,128,000      | 54,150,000,000<br>37,482,088,818<br>36,757,221,900<br>724,866,918 | 2023-08-18<br>12:45:09<br>2023-10-03<br>11:13:09 | 8     | 220<br>341                       | 0       | JIRA                                                  | 3,128,346<br>3,065,700<br>62,646         |

|     |                                                                |                 |                 |       |       |     |     |                 |
|-----|----------------------------------------------------------------|-----------------|-----------------|-------|-------|-----|-----|-----------------|
| ARM | <a href="#">master_AnalysisBase_aarch64-centos7-gcc11-opt</a>  | 2023-04-18T0220 | 18-APR<br>10:34 | 0 (1) | 0 (0) | N/A | N/A | 18-APR<br>11:11 |
| ARM | <a href="#">master_Athena_aarch64-centos7-gcc11-opt</a>        | 2023-04-18T2300 | 19-APR<br>06:16 | 0 (1) | 4 (4) | N/A | N/A | 19-APR<br>09:21 |
| ARM | <a href="#">master_AthSimulation_aarch64-centos7-gcc11-opt</a> | 2023-04-18T2101 | 18-APR<br>21:37 | 0 (0) | 1 (1) | N/A | N/A | 18-APR<br>22:11 |
| ARM | <a href="#">master_DetCommon_aarch64-centos7-gcc11-opt</a>     | 2023-04-18T2000 | 18-APR<br>20:02 | 0 (0) | 0 (0) | N/A | N/A | 18-APR<br>20:21 |

- 4 nightly builds for Athena, AthSimulation, AnalysisBase and DetCommon ([link](#)) projects fully integrated in standard ATLAS build system and available on CVMFS
- Selected stable Athena releases like 23.0.3 and 23.0.14 installed on CVMFS - used in physics validation
- since December 2022 dedicated build machine provided by CERN IT hosted in Oracle Cloud, 40 vCPUs (Ampere Altra A1), 250 GB RAM, special manual CentOS7 installation
- used before techlab-arm64-thunderx2-01 provided by CERN OpenLab and shared with CERN SFT
- LCG\_102b\_ATLAS\_\* / LCG\_103 layers and tdaq/tdaq-common builds available and used in ARM/aarch64 builds
- can easily build stand-alone docker/podman container for e.g. AthSimulation

## • Build flags

- using Armv8 defaults (gcc 11.2 allows up to armv8.6-a, gcc docu link)
- no special arch option and no special compilation flags set apart from different linker flag in max-page-size
- able to build Athena/AthenaExternals with clang16 as well

## • Floating Point Exception (FPE)

- Athena FPE auditor code not working on ARM/aarch64 since it uses x86 specifics
- discussions about different FPE behaviours of x86\_64 vs. ARM/aarch64 compilers: StackOverflow 1, StackOverflow 2, ARM compiler

## • Potential numerical differences

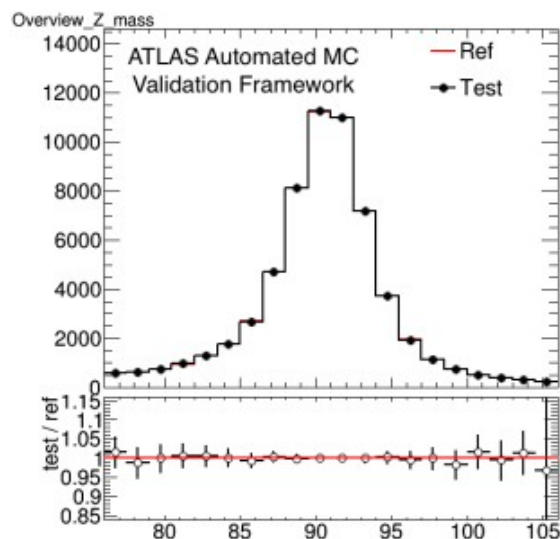
- e.g. due to different floating point libraries used - see StackOverflow link
- some fluctuations in physics objects at the level of ( $10^{-4} - 10^{-6}$ )
- NB. small numerical differences or Intel vs. AMD if IntelMathFunction used



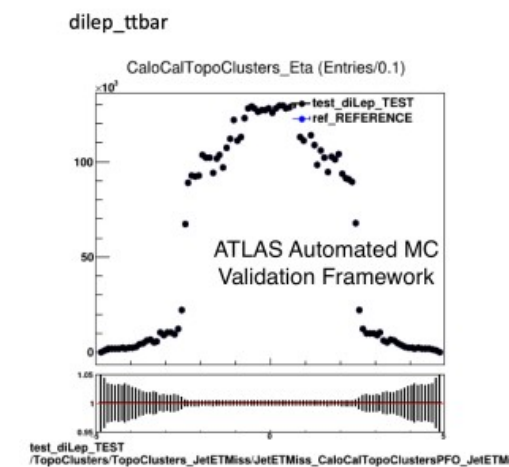
## • using Athena 23.0.14

J. Elmsheuser (BNL) at CHEP2023 ([link](#))

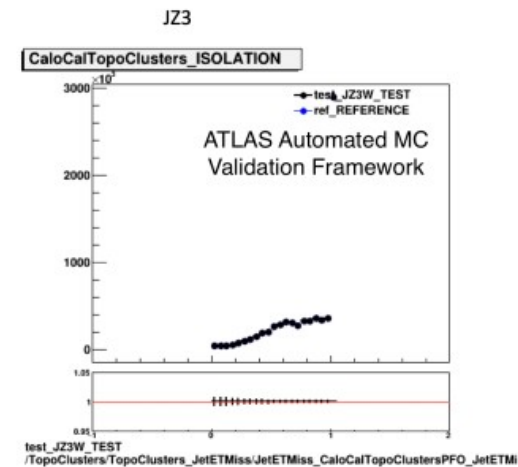
- March 2023 successfully passed reconstruction physics validation
- standard procedure of 13 MC physics processes (100k events each), no pile-up
- Digitization+Reconstruction step on AWS Graviton2 PanDA queue (130 jobs, 215 GB output, 1.5 days, 450 USD in total)
- compared with same events produced on x86\_64 grid sites - all subsequent steps (merging, histogramming) on x86\_64 - task speed comparison in the backup



Reconstructed MC  $Z \rightarrow \mu\mu$   
invariant mass



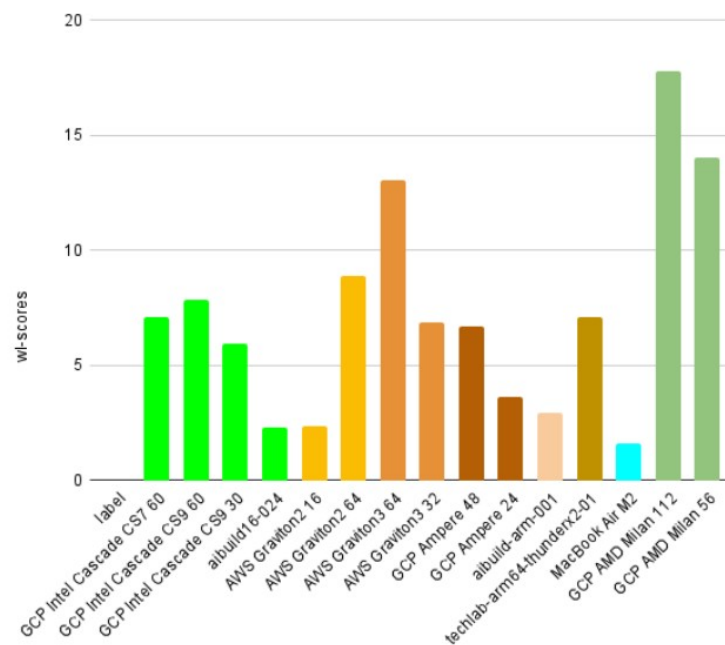
Perfect agreement



Example plots from calorimeter cluster validation

J. Elmsheuser (BNL) at CHEP2023 ([link](#))

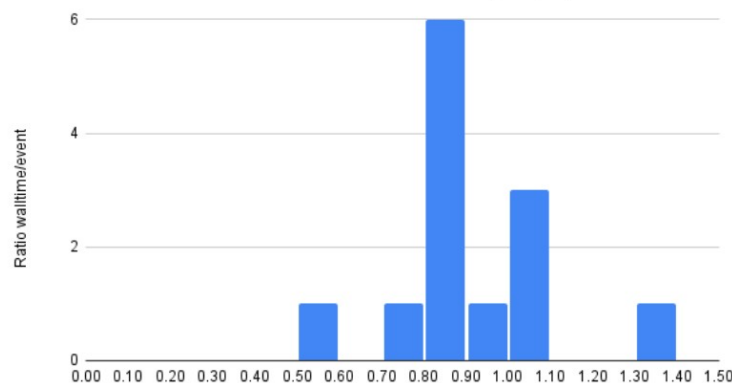
wl-scores



Larger values are better

- HepScore numbers for ATLAS reconstruction
- Node fully packed with  $n \times 4$  threads  
→ different number of available cores explains different scales !
- Orange/Brown/Blue: ARM flavours
- Green: Intel/AMD flavours
- Some hyperthreading or IO related differences
- Similar trends for Event generation and simulation

Ratio ARM vs. x86 tasks for walltime/event per physics task



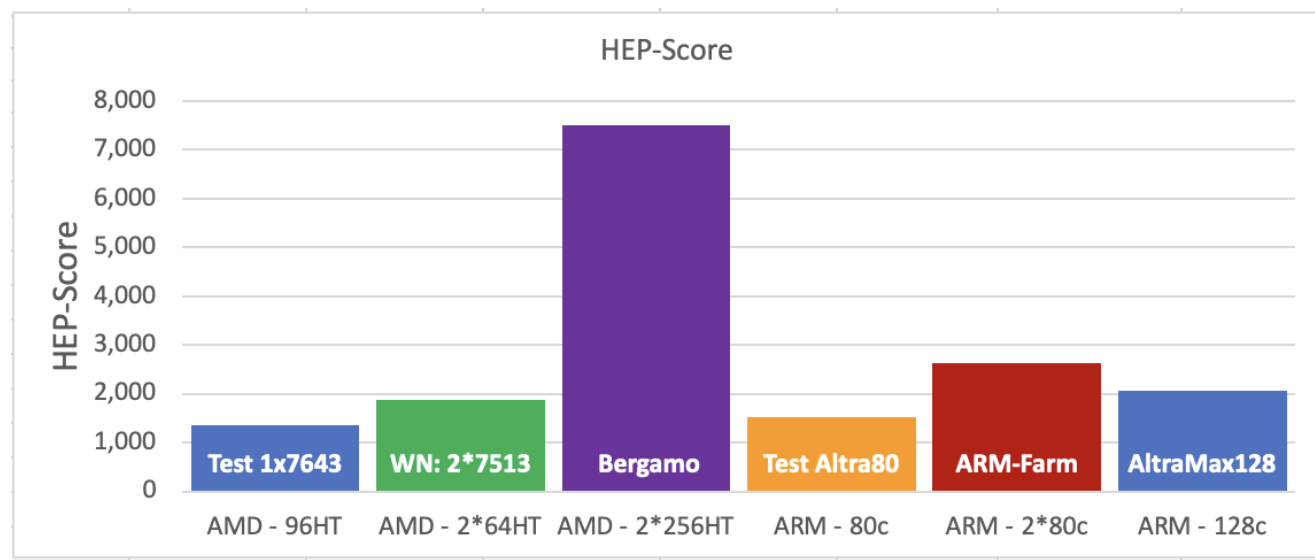
Ratio ARM vs. x86 Grid job averages:

- Walltime/event: 0.92
- CPU time/event: 0.81
- HS06sec/event: 0.79  
(with ARM HS06=10)

| Machine     | CPU             | Threads | HEPScore |
|-------------|-----------------|---------|----------|
| Test AMD    | 1x EPYC 7643 HT | 96      | 1359     |
| Std AMD WN  | 2x EPYC 7513 HT | 128     | 1887     |
| AMD Bergamo | 2x EPYC 9754 HT | 512     | 7497     |
| Test ARM    | 1x ALTRA 80     | 80      | 1513     |
| Farm ARM    | 2x ALTRA 80     | 160     | 2619 ??  |
| Ampere Max  | 1x ALTRA Max    | 128     | 2065     |

→ x4  
←

→ Not x2 ??  
←

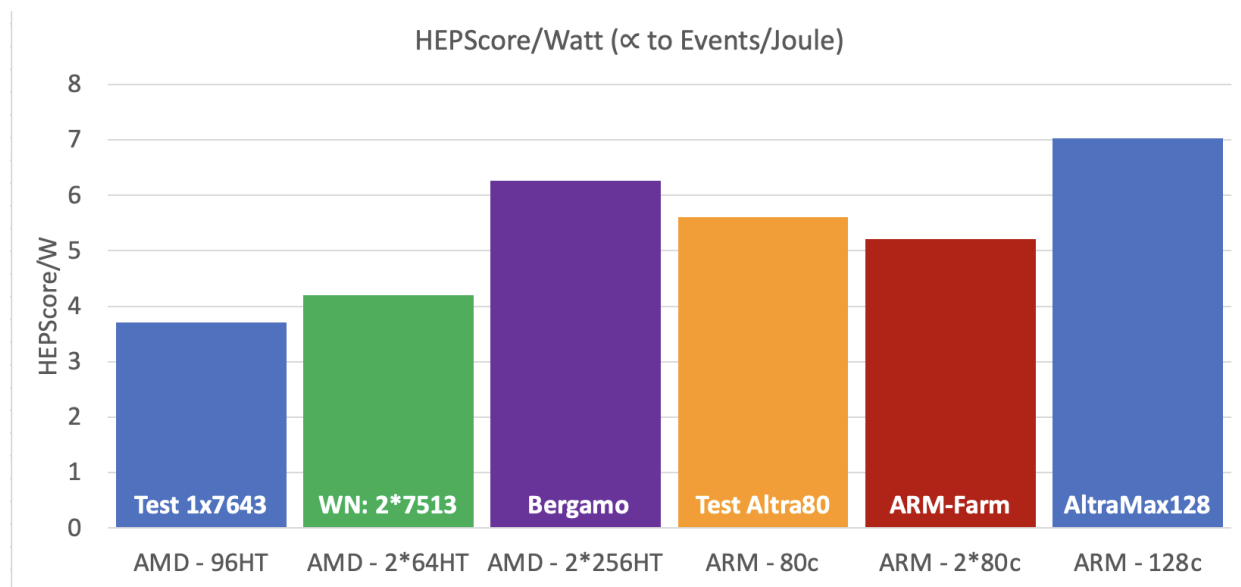


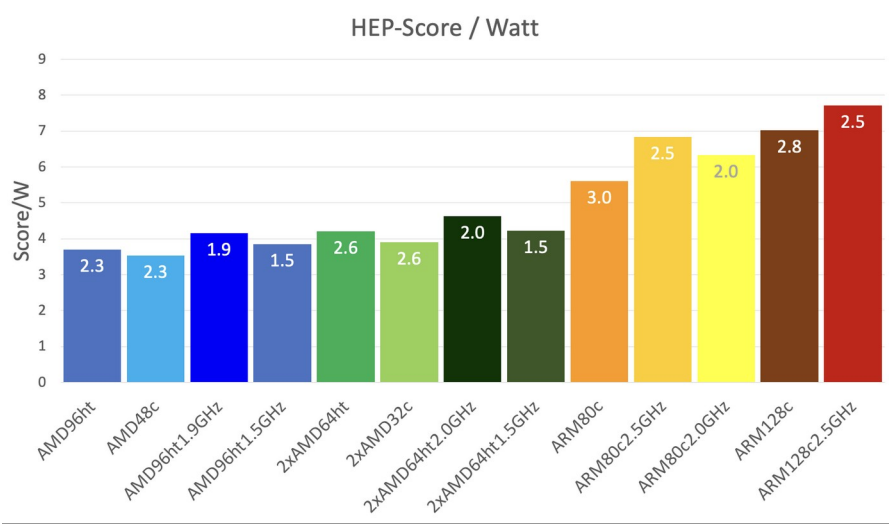
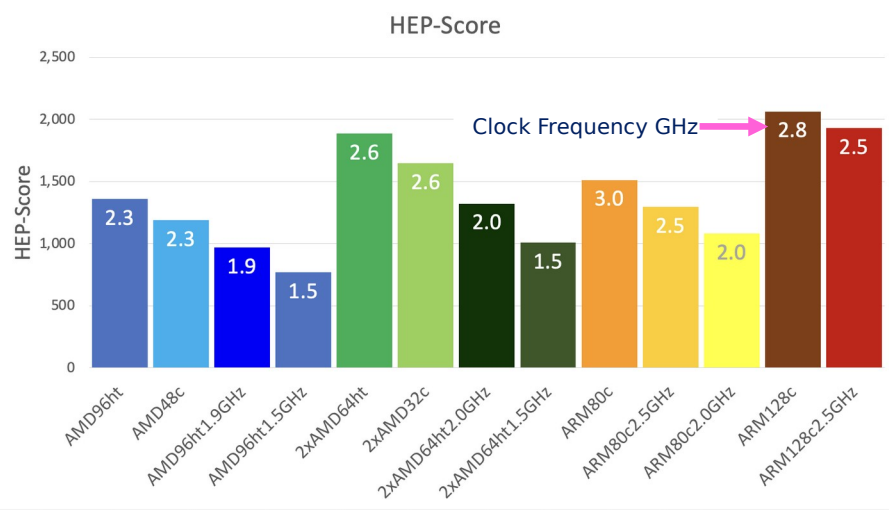
| Machine     | CPU             | Threads | HS/<br>Watt |
|-------------|-----------------|---------|-------------|
| Test AMD    | 1x EPYC 7643 HT | 96      | 3.7         |
| Std AMD WN  | 2x EPYC 7513 HT | 128     | 4.2         |
| AMD Bergamo | 2x EPYC 9754 HT | 512     | 6.3         |
| Test ARM    | 1x ALTRA 80     | 80      | 5.6         |
| Farm ARM    | 2x ALTRA 80     | 160     | 5.2 ??      |
| Ampere Max  | 1x ALTRA Max    | 128     | 7.0         |

50% improvement from 7513 WN

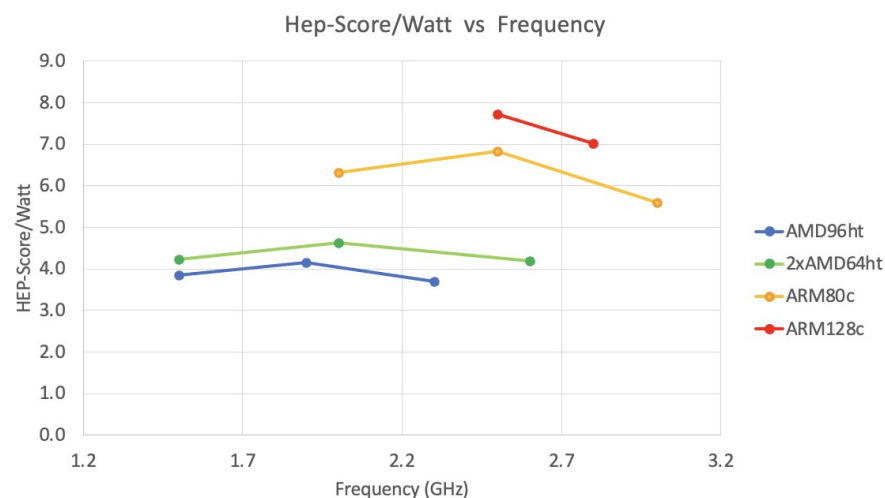
25% improvement from 80c

Note added: We use Average Watts to calculate HS/Watt. Other people also use Max Watts, so care needs to be taken comparing numbers.





- Running one step below max clock frequency seems to be the sweet-spot for all CPUs tested.
- On ARM128, frequency reduction lowers HEP-Score by 6.5% but saves another 9% of the energy.
- No data on Bergamo (bug?)



## ● Further plans:

- migrate to AlmaLinux9 (all pieces already available)
- investigate further production workflows (e.g. MC generators)
- power consumption tricky to assess on virtualized hardware

## ● Summary

- full ARM/aarch64 ATLAS nightly builds on CVMFS through regular nightly ATLAS build system
- Geant4 and reconstruction physics/technical validation successfully passed using AWS through PanDA/Rucio production system
- Athena, 23.0.3 sherpa, simulation and reco workflows integrated into new HepScore benchmark

## ● Energy efficiency

- 128-core ALTRA-Max is most energy efficient CPU tested by Glasgow
- AMD has made big improvements to energy efficiency with Bergamo.
- running CPU one step below max-frequency seems to be sweet-spot in terms of energy efficiency

Project is part of ATLAS HL-LHC R&D projects and will be part of ATLAS computing technical design report for HL-LHC