

Finanziato dall'Unione europea NextGenerationEU







## Centro Nazionale di Ricerca in HPC,

Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Heterogeneous computing at INFN-T1

D.Lattanzio\*, A.Chierici, D.Michelotto, A.Pascolini, G.Sergi

Conference on Computing in High Energy and Nuclear Physics. October 19-25, 2024 - Kraków

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing

Missione 4 • Istruzione e Ricerca









Outline



- Introduction
- ARM resources
- RISC-V resources
- Sierra Forest









## Introduction



- At INFN we started a technology tracking program
- At INFN-T1, in farming division, we mainly focused on computing
  - Investigate new processors technologies
  - Power consumption
  - CPU architectures
  - Understand middleware and general software readiness











## **ARM Resources**

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









## ARM resources

- 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















## 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 alma8, 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 still an issue (full availability only with el7)
- Had to make agreements with experiments to run only specific workflows









## Some numbers



- Ampere altra max
  - Each node provides 3754 hepscore, 14,66 hepscore/core
  - BMC not reliable, couldn't 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







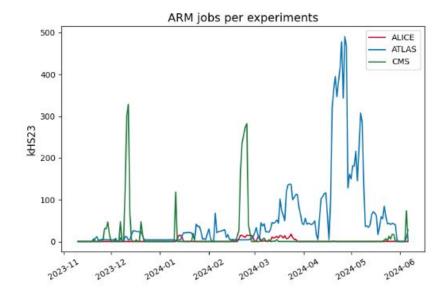


## ARM experiment usage at CNAF



#### Current setting (still work in progress)

- Cvmfs available
- Access to external network
- Gpfs client -> now available on ARM but only with a very recent version, not yet in production at INFN-T1
- Condor/GRID -> in production









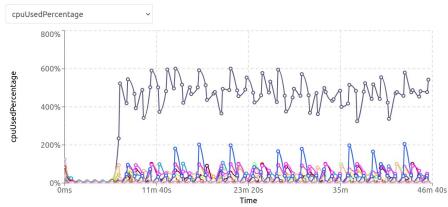


## Experiment point of view: ALICE



Preliminary tests showed good balancing in using resources: still some instability in ARM builds didn't allow to run a validation so far.

Data reconstruction



- GRID setting tuned for 8 cores per job
- CPU efficiency consistent with what observed in the GRID node
- Physics validation on the output not yet done

Recent builds on Almalinux9 presented some issues (mainly fixed)  $\rightarrow$  tests going to be resumed

GRID submission @CNAF (GRID submission + aptainer container)  $\rightarrow$  validated







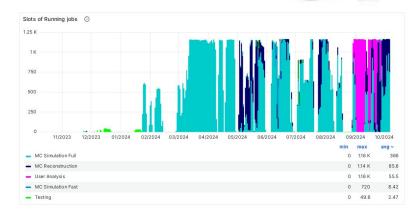


## **Experiment point of view: ATLAS**

#### ATLAS has been running jobs on ARM@CNAF for one year

- initially, only test jobs
  - ATLAS Software already Physics-validated on ARM
  - Technical validation performed at CNAF (HTCondorCE, pilot/PanDA, containers)
- workflows:
  - Full and Fast MC Simulation
  - MC reconstruction
  - Group production
  - User Analysis





#### Very good performance observed

 Steady use of available resources. On peak: ~12% of ATLAS-dedicated resources at INFN-T1









The ARM nodes at T1\_IT\_CNAF have been integrated as a sub site of the regular Tier1 and thus accessed via GRID

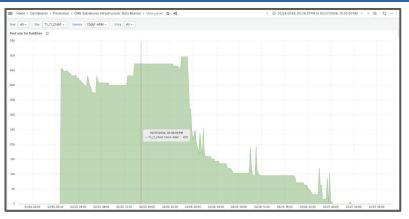
- Minimize the effort on CMS and simplify the site admins life
- At the moment data are accessed via Xrootd protocol. The plan is to provide direct access also via GPFS.

#### Technical validation fully done.

- Being the first allocation at a Tier site, the CNAF nodes where essential to finally validate that CMS Computing is multi-architectures enabled.

#### **Physics validation in progress:**

- most of the subsystems report green light (especially when looking at MC).
- Some discrepancies spotted on DATA require further analysis for a better understanding
- This step has been carried on both at CNAF and at Glasgow temporary allocation



RelVal		TICKETS RELVALS DASHBOARD Logged in as Da	niele Spig
	PrepID	Workflows (jobs in RegMgr2)	
	CM50V, 14.3.9.3m3_Des_2023.CM4760H9ursHMET2020-00001	<ol> <li>pdmvserv, RVDMSSW_14_0.0, pre3RunJetMET2023D_CNAFARM, ReiVaL2023D_240215, 092819_7275 spen in: Stats2 staus: normal-archived adapter: FEVTDEBUGHL, complexe: 109.72%, event: 1266.761, pper VALID JAVKETOLMSER: 14.0, app. 4144(statshare). RTS/mexerv.146/MARAIVA.2020.2019/PTEFEUHAT</li> </ol>	
		<ul> <li>datatie: AOD, completed. 106.33%, events: 1,227.653, type: VALID / verset try/catalogy, 14, 0, p.exi-148x, detailund, Promet, horset, y1, 2042484, 36942 (2010) / ACD</li> </ul>	
		<ul> <li>datatier: MINIADD, completed: 106.33%, events: 1,227,653, type: VALID / aetwic=0,04655w; 14,0,0,pen3-1450; detailund; Promer, fromer, 1, 054246M, 56/14(2):0235-1/100846D</li> </ul>	
		<ul> <li>desister: NANOAOD, completed: 106.33%, events: 1,227,653, type: VALID /aetwt10.1M55m; 14.0.0, pee-148K, datakura; Promot Instein, v1, DMARAM, Serval; 32230-VIRMADADD</li> </ul>	
		<ul> <li>detative: DQMIO, completed: 0.00%, events: 0, type: VALID /art/#ET0.145506_116_0_0_peg3_168_cdta1a.in.8_home_frozen_it_CMARAM_RVId_20230_v1600480</li> </ul>	
	CM20W_14_0_0_ard_0_bet_2023_CM44AMM Budbisplooed.et20220-00001	<ol> <li>pdmwserv, RVCMSSW, 14,0,0, pre3RunDisplaced.Jet2023DCNAFARM_BetVol.2023D_240215,092843_9419 open in: Stats2 status normal-archived</li></ol>	
		<ul> <li>dutative: AOD, completed: 120.25%, events: 769,596, type: VALID Displaced.et/CMSSin: 14.0.0, prept 4486, dataBack, Promet, Pages, v1, CMA4368, Parka, 2020 v1/300</li> </ul>	
		<ul> <li>detative: MINIAOD, completed: 119.45%, wwwtx: 764.476, type: VALID Displaned.art(SMStat; 14.0, t), pres 1466, dataBack, Promet, Pasen, v1, CMARMM, Revise), 2020 v1 AMMAOD</li> </ul>	
		<ul> <li>dutative: NANOAOD, completed: 120.25%, events: 769,596, type: VALID ministered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, invited. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, invited. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, invited. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, invited. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, provider. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, provider. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, provider. 2020 v / international providered.act/chargesci.14.0.0, pres 1496, distance, proved. Statem - 1, convirtant, provider. 2020 v / international providered.act/chargesci.14.0, pres 1496, distance, proved. Statem - 1, convirtant, provider. 2020 v / international providered.act/chargesci.1200, pres 1496, distance, proved. 2020 v / international providered.act/chargesci.1200, pres 1496, distance, provider. 2020 v / international providered.act/chargesci.1200, pres 1496, distance, providered.act/chargesci.1200, pres 1496, distance, providered.act/chargesci.1200, distance, prov</li></ul>	
		<ul> <li>dutative: DQMID, completed: 0.00%, events: 0, type: VALID Displaced.et/CMIDIN: 14:0.0; pre3-1406; displaced: promote (type); 1: OX473Mb; Biv(a); 20230-1; 0OM0;</li> </ul>	
	CM20W, 14.3, 0. am304x, 2023. CNAI AMM Rund Comma 20220 00001	<ol> <li>pdmvserv, RVCMSSW_14_0_0, pre3RunECamma2023DCNAFABM, ReiVal_2023D_240215,097825,2415 open in: Stats2 status: normal-archived</li></ol>	
		dutative: AOD, completed: 135.99%, wwets: 1,809,262, type: VALID     TOammed/CM02W, 14.0., p.ee3-149X, dyniferd: Proved, freex, vi. CM47498, Mir/el, 20202-47/ADD	
		<ul> <li>dutatie: MINIAOD, completed: 135.98%, events: 1.809,252, type: VALID (Commod/CMSSW 14.0 0 pere3-140X databad) fromet froein v1 CM3/488 Mit/al 2020-1/MINIAO</li> </ul>	
		<ul> <li>deate:: NANOAOD, completed: 135.98%, revents: 1,809,262, type: VALID (50ammeD/CM55W, 14.0, 0, pre3140K, danibard; Promet, Freen, vf. CMFARR, Berlail. 2020D+r/ NANOAOD</li> </ul>	
		<ul> <li>datatie: DQMIO, completed: 0.00%, events: 0, type: VALID (Commed/CMSSW, 14, 0, 0, eve3.149X, daniberd, Promet, Ergen, v1, CMFAMB, MYGL 2020 v1, 50M0</li> </ul>	
	CMSSW_14_0_0_pm3_Data_2022_CM4740M BinTac2022D 00001	<ol> <li>pdmvserv_RVCMSSW_14_0_0_pre3RunTau2023D_CNAFARM_RetVal_2023D_240215_072626_2546 open rx Stat32 status: normal-archived</li> <li>mature: FEVTDEBUGHLT_completed 83.12%, events 053,398, ppc VALID</li> <li>mature 10.444 bit 40.4454 databat 44.150 mm in 10.4444 bit 40.2120214 (StreBackAtz)</li> </ol>	
		<ul> <li>Identifier: AOD, completed: 82.25%, events: 671.245, type: VALID</li> <li>Tau/CMSSW 16.0 p.mei-140X dataBard Power (none v) CM-ESBA Barkal 2020 v(JAD)</li> </ul>	
		<ul> <li>Ideate: MINAOD, complete: 82.25%, events: 671,245, type: VALID ThurCheStor, I.e. do. press 1460, doublest Parent, From v1. CN47588, Molecular Data VLMMAD</li> </ul>	
		<ul> <li>Ideate: NANOAOD, completed: 82.25%, events: 671,245, type: VALID The/CMCOV III 0.0 cm-2-140X database Parent freew vi CM4/348, ReVial 2020-4/34840Aco</li> </ul>	
		<ul> <li>destate: DOMID consistent 0.00% exercise 0. sees: VALID</li> </ul>	

#### ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing

#### Missione 4 • Istruzione e Ricerca









### **RISC-V** Resources

ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing









## **RISC-V Resources**

- 2x Milk-V Pioneer
  - 64 Core RISC-V CPU up to 2GHz
  - 128GB DDR4 3200
  - 1TB NVMe disk
  - 2x10Gbps network cards
  - OS pre-installed: Fedora 38
- An interesting architecture for the future
  - Not competitive today (cores at the level of Rpi cores), but evolving fast!
  - Open ISA managed by the RISC-V Foundation (riscv.org)











## Sophgo SG2042



- 64 RISC-V Cores
  - T-Head XuanTier C920s (high performance)
- 2Ghz frequency
- 64MB system cache
- 32x PCIe Gen4 lanes
- good enough to test larger projects like porting CMS full software stack (CMSSW)













- INFN is deploying compiled codes for tests in: /cvmfs/datacloud.infn.it/repo/riscv64-pioneer/
- Please see the CMS <u>CHEP talk on RISC</u> benchmarking
- The MILK-V performance is comparable to a good desktop pc, but the number of cores is very interesting and can increase more.











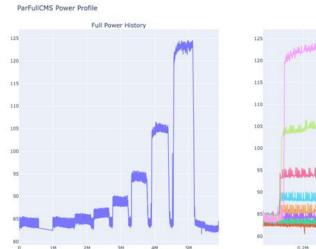
- db12 CMS benchmark
  - milk-v: 378.3, 5.8 per core
  - e5-2640v3: 248, 3.8 per core
- Milk-v today performs better than a 2016 CPU...
- If we take out the hyperthreading, milk-v is 2-3 times slower than a modern xeon
- 64 real cores, not so common on modern CPUs
- Trade-off between core number and power (take into account power consumption too)













0.6M

0.45

0.8M



1/2/4/8/16/32/64 threads test of CMS Fast Simulation → power consumption



Centro Nazionale di Ricerca in HPC Big Data and Quantum Computing

# Supercomputing shaping the future