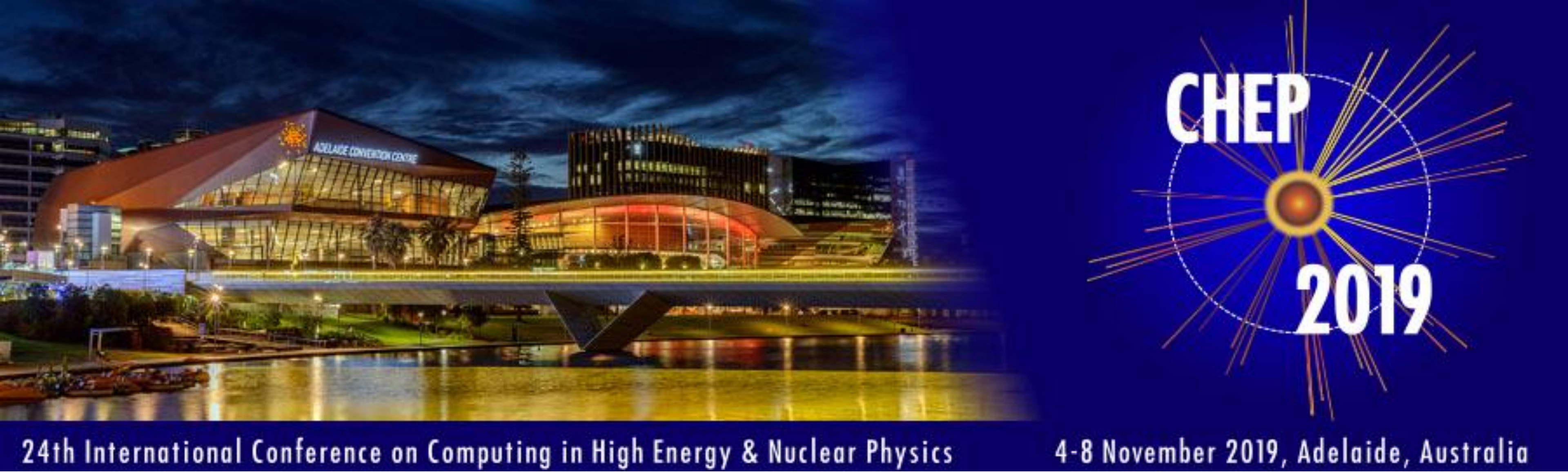


24th International Conference on Computing in High Energy & Nuclear Physics

4-8 November 2019, Adelaide, Australia

High Performance Computing for High Luminosity LHC

Maria Girone, Domenico Giordano, Viktor Khristenko, Gavin McCance, Hannah Short, CERN

Motivation

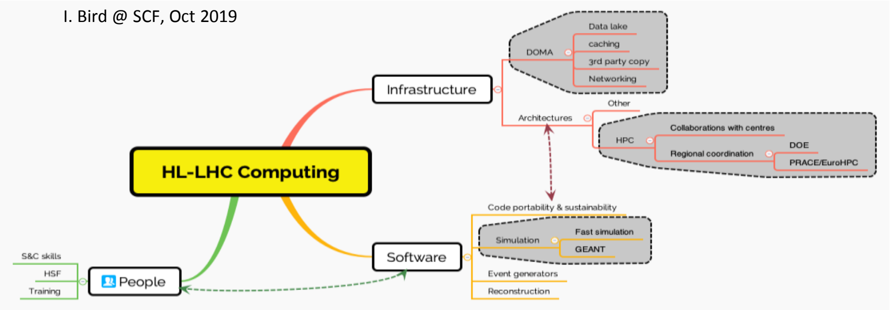
- HPC sites will grow by a factor of 20 on the timescale of the HL-LHC
  - Large investments in the US, Europe, and Asia. **Pushing to Exascale**
  - Architectures chosen are well suited to AI and ML applications

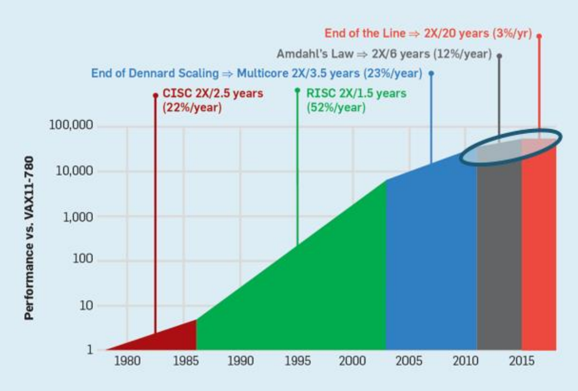
- Technology improvements on traditional CPU are slower than the average from last decade
  - Our resource estimates within flat budgets might be optimistic

- Super Computers may help to close the resource gap at HL-LHC

P.Perez (CMS), F.Stagni LHCb,
D.Benjamin F Sciacca, ATLAS,
CHEP 2019

- All experiments have independent efforts to adopt HPC sites
  - Extracting commonalities is needed


I. Bird @ SCF, Oct 2019



Top10
Super Computers

- HEP relies on x86, but most of the Top10 SC use **different architectures**
- We are in **pre-Exascale**
- **Exascale expected in 2021**

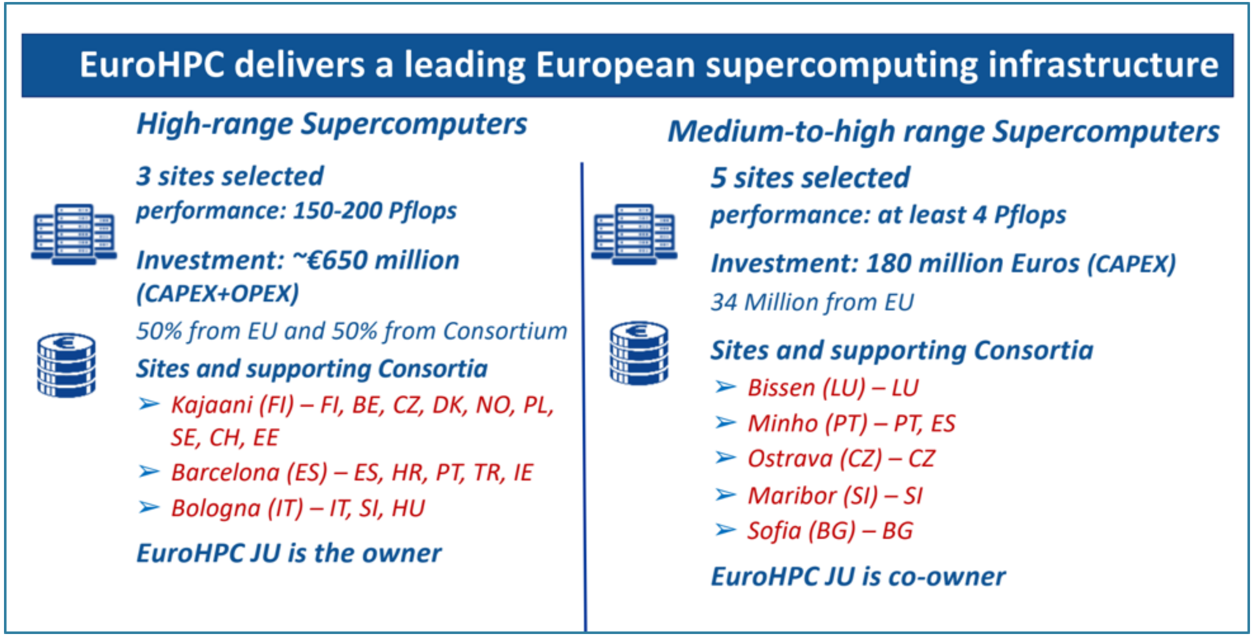
🟩	Intel Xeon+ NVIDIA GPUs
🟦	Intel Xeon
🟥	Custom RISC
⬜	Power9 + NVIDIA GPUs

Maria.Girone@cern.ch

	10. Lassen	LLNL, USA	IBM Power9 and NVIDIA V100	18.2 PFlops
	9. SuperMUC-NG	LRZ, Germany	Intel Xeon Platinum	19.5 PFlops
	8. AI Bridging Cloud Infrastructure (ABCI)	AIST, Japan	Intel Xeon Gold and NVIDIA V100	19.9 PFlops
	7. Trinity	LANL, USA	Intel Xeon and Xeon Phi	20.2 PFlops
	6. Piz Daint	CSCS, Switzerland	Intel Xeon and NVIDIA V100	21.2 PFlops
	5. Frontera	TACC, USA	Intel Xeon Platinum	23.5 PFlops
	4. Tianhe-2A (Milky Way 2A)	Guangzhou, China	Intel Xeon and Matrix 2000 (RISC)	61.4 PFlops
	3. Sunway TaihuLight	Wuxi, China	Sunway 26010 (RISC)	93.0 PFlops
	2. Sierra	LLNL, USA	IBM Power9 and NVIDIA V100	94.6 PFlops
	1. Summit	ORNL, USA	IBM Power9 and NVIDIA V100	**143.5 PFlops**

EuroHPC Roadmap to pre-Exascale

EuroHPC
Joint Undertaking



EuroHPC delivers a leading European supercomputing infrastructure

High-range Supercomputers

3 sites selected

performance: 150-200 Pflops

Investment: ~€650 million (CAPEX+OPEX)

50% from EU and 50% from Consortium

Sites and supporting Consortia

- *Kajaani (FI) – FI, BE, CZ, DK, NO, PL, SE, CH, EE*
- *Barcelona (ES) – ES, HR, PT, TR, IE*
- *Bologna (IT) – IT, SI, HU*

EuroHPC JU is the owner

Medium-to-high range Supercomputers

5 sites selected

performance: at least 4 Pflops

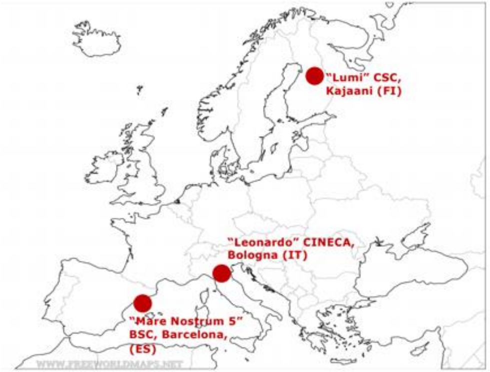
Investment: 180 million Euros (CAPEX)

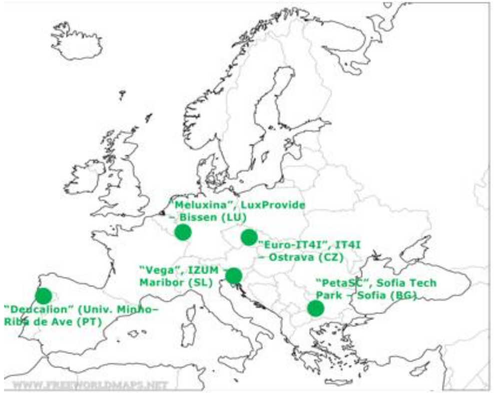
34 Million from EU

Sites and supporting Consortia

- *Bissen (LU) – LU*
- *Minho (PT) – PT, ES*
- *Ostrava (CZ) – CZ*
- *Maribor (SI) – SI*
- *Sofia (BG) – BG*

EuroHPC JU is co-owner


"Lumi" CSC, Kajaani (FI)
"Leonardo" CINECA, Bologna (IT)
"Mare Nostrum 5" BSC, Barcelona, (ES)
WWW.FREEWORLDMAPS.NET


"Meluxina", LuxProvide – Bissen (LU)
"Euro-IT4I", IT4I – Ostrava (CZ)
"Vega", IZUM – Maribor (SL)
"PetaSC", Sofia Tech Park – Sofia (BG)
"Deucalion" (Univ. Minho-Riba de Ave (PT)
WWW.FREEWORLDMAPS.NET

Courtesy of S. Girona, BDEC2 workshop, San Diego, Oct. 2019

Working together

Engaging and working together with HPC centers is essential for HEP

(DOE & NSF in USA, PRACE and EuroHPC in Europe)

- Regional and HPC community coordination
  - PRACE discussions from the workshop in October 2018 led to a proposal for MoU with SKA, CERN and GÉANT under evaluation
  - Representing WLCG input to **PRACE, EuroHPC and BDEC2** WGs

https://indico.cern.ch/event/760705/

- Direct collaborations with centres
  - Ongoing discussions with CSCS, Jülich (through the DEEP-EST project), Oak Ridge (Summit) via CERN

- Code portability and sustainability
  - Portability libraries: Alpaka, SYCL, Kokkos, etc
  - Co-organizing training/hackathons and hands-on in 2020

Challenges and R&D

Not an exhaustive list!!!!!

Main Challenges

Software and Architectures

Runtime Environment and Containers

Monitoring and Accounting/benchmarking

Authorization and Authentication

Provisioning

Data Processing and Access

Wide and Local Area Networking

R&D activities

Heterogeneous Computing

Benchmarking

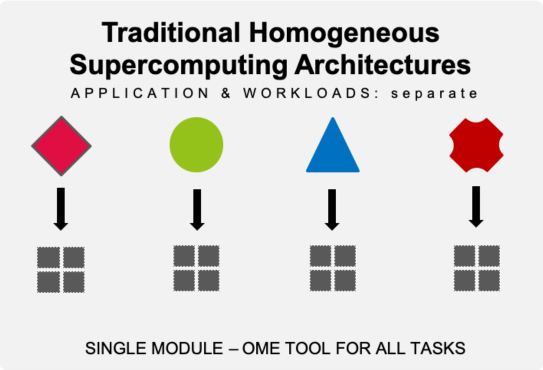
Authenticated Workflow

Dynamic Workflow

Data Access

Challenges outlined in the WLCG Common Challenges document
https://docs.google.com/document/d/1AN1d6Nu-khBsKnNH1MVvszqWdpcMfFGaYQMEVnS01Tc/

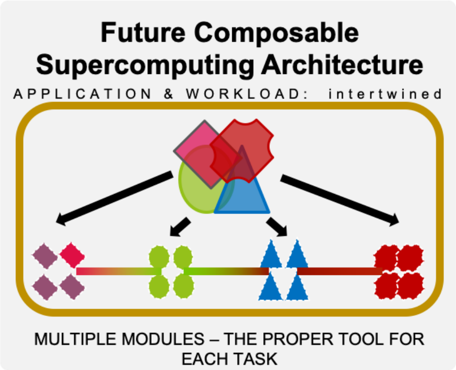
Modular Super Computer Architectures



Traditional Homogeneous Supercomputing Architectures

APPLICATION & WORKLOADS: separate

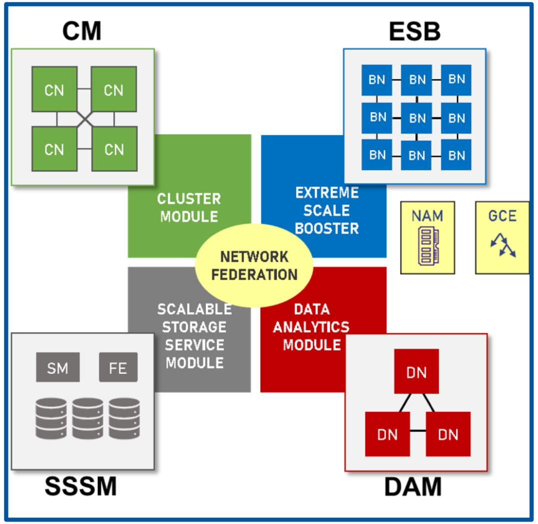
SINGLE MODULE – OME TOOL FOR ALL TASKS

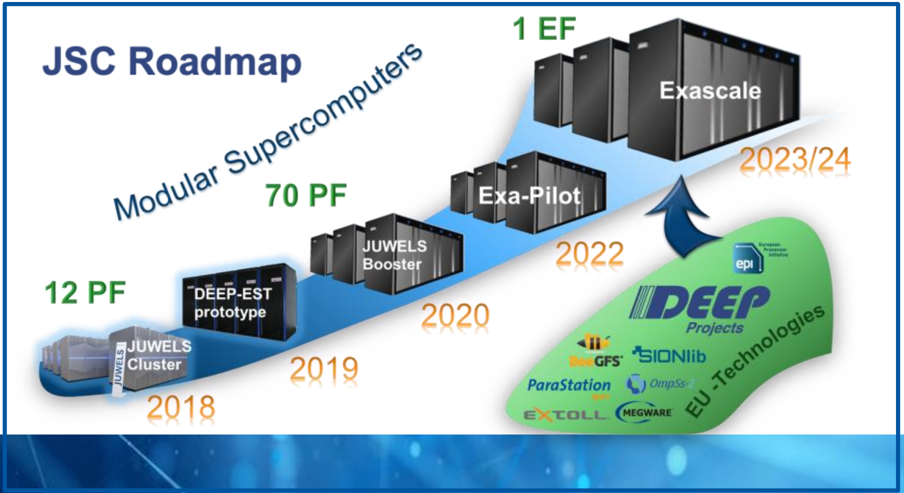


Future Composable Supercomputing Architecture

APPLICATION & WORKLOAD: intertwined

MULTIPLE MODULES – THE PROPER TOOL FOR EACH TASK





DEEP-EST: Modular Architecture Prototype at JCS

- Key for JSC and EU Strategy

Goals and Motivation

- Participate in the **co-design** of the future Modular HPC and provide HEP-specific feedback to the HPC community

- Explore CMS reconstruction workflows for HLT on modular HPC infrastructures

- R&D: Explore heterogeneous hardware for CMS HLT reconstruction workflows
  - GPUs / FPGAs

- R&D: Explore the usability of HPDA resources for CMS Data Analysis
  - HPDA - High Performance Data Analytics

Architectures and Software

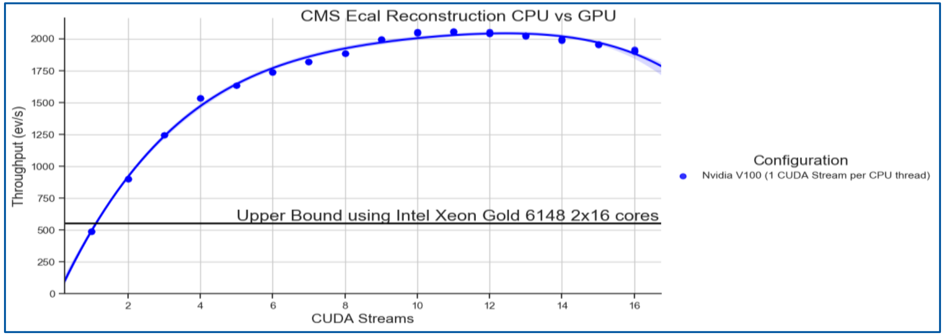
ECAL/HCAL local reconstruction for HLT on Heterogeneous Architectures in CMS

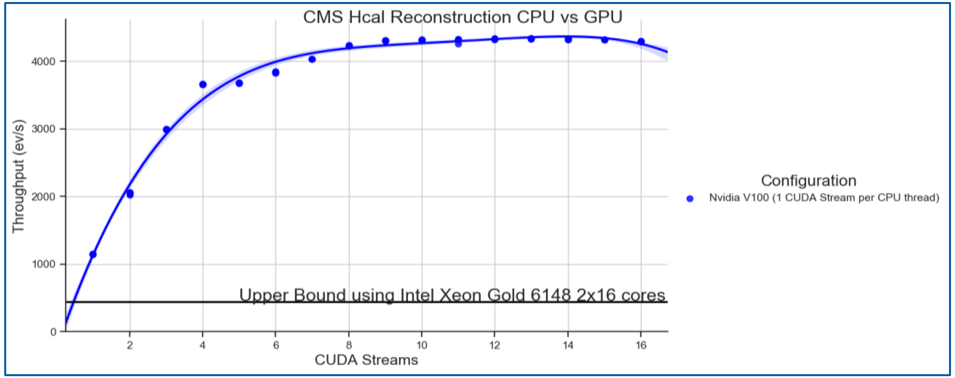
Demonstrate the ability to do a local reco on GPU with very good efficiency

- Fully CUDA-based CMS Hcal/Ecal reconstruction for HLT, integrated in **CMSSW**
- Results are reproducible (within 0.1% or better)
- Intel Xeon Gold 6140/6148 used for comparison (similar to HLT Cluster Intel Xeon Gold 6130)
- Provided tests use 1 CUDA Stream per 1 CPU thread

A. Bocci, CHEP 2019

Results on Open Data:
http://opendata.cern.ch/record/12303





Runtime Environment, Containers and Benchmarking

CVMFS: efficient technology for the global distribution of massive application software stacks

- >1B files under management, >150 production sites
- example of LHC software R&D that evolved into a HEP de-facto standard

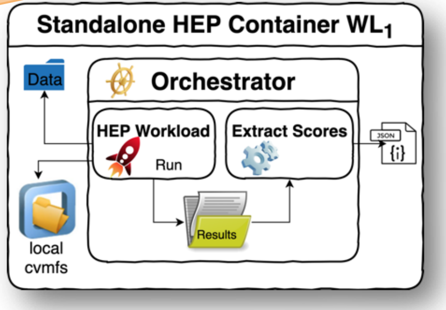
Critical for the LHC experiments

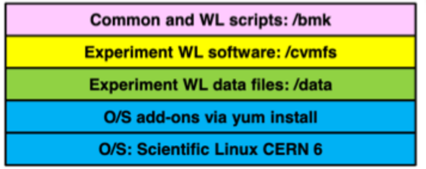
- HPC centers could **support CVMFS centrally** (and some have)
- solutions to support lightweight and dynamically deployable versions of the existing infrastructure are already avaiable

Establish a standalone containerized **benchmark suite** to measure the workflow performance

- representative applications by each experiment for both HTC and HPC
- standalone containers encapsulating all and only the dependencies to run each workflow as a benchmark

A. Valassi, CHEP 2019




Common and WL scripts: /bmk
Experiment WL software: /cvmfs
Experiment WL data files: /data
O/S add-ons via yum install
O/S: Scientific Linux CERN 6

Container images are made up of layers

Authorization and Authentication

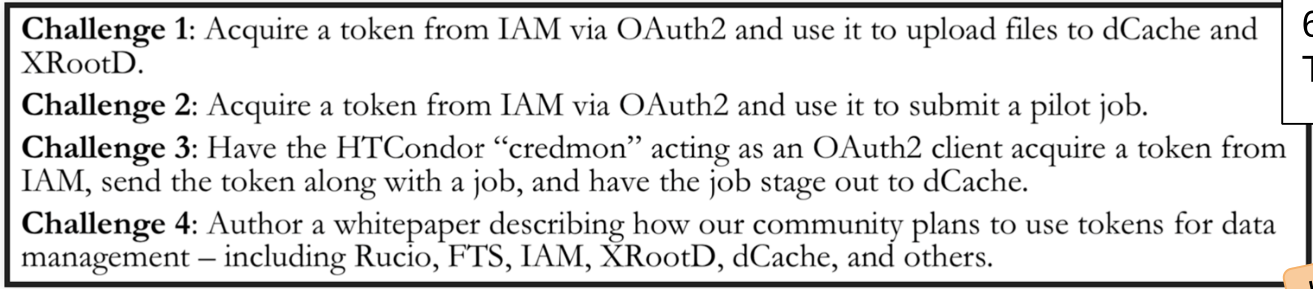
LHC collaborations have thousands of active submitters

- Essentially use a “trust the VO model”
  - VOs are relied on to log the source of work and respond to suspicious activity
- HPC sites often have stricter cybersecurity policies
  - Not clear whether our existing security model applies

Show that the workflows performed by VO users can be securely supported by HPC

- HPC sites should support standard OAuth2.0 flows
- Sites should trust WLCG Virtual Organisations as OAuth2.0 Token Issuers, and validate bearer tokens accompanying incoming jobs

OAuth Challenges proposed by Brian Bockelman



Challenge 1: Acquire a token from IAM via OAuth2 and use it to upload files to dCache and XRootD.

Challenge 2: Acquire a token from IAM via OAuth2 and use it to submit a pilot job.

Challenge 3: Have the HTCondor “credmon” acting as an OAuth2 client acquire a token from IAM, send the token along with a job, and have the job stage out to dCache.

Challenge 4: Author a whitepaper describing how our community plans to use tokens for data management – including Rucio, FTS, IAM, XRootD, dCache, and others.

5. Traceability
6. Multi-tenancy

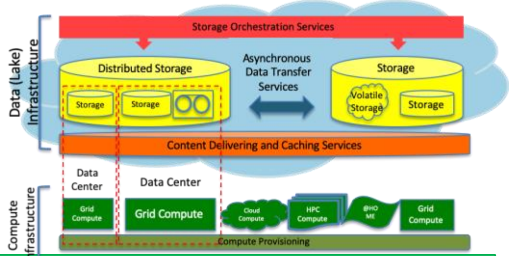
Token Issuer Model

Augmented by WLCG AuthZ WG

H. Short, CHEP 2019

Data Access

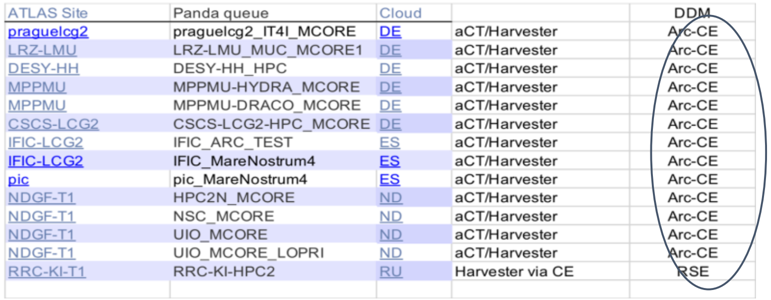
- HEP workflows are data intensive
  - HPC sites are optimized for tightly coupled calculations
  - HPC sites often have stricter firewalls and often no permanent storage
- HEP moves data
  - Will need to demonstrate filling multiple 100Gb/s network links



Requirements for data access

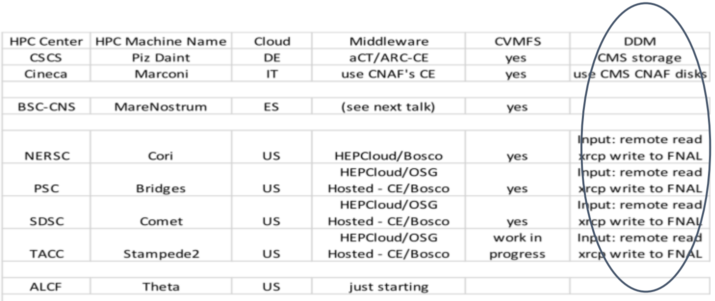
- Deliver and validate multi-petabyte datasets to local storage
- Real-time delivery to maintain CPU efficiency
  - Use edge caching
- Exercise Local Site Storage
  - Creating and storing data locally at scale

ATLAS HPC Sites/ PanDA



ATLAS Site	Panda queue	Cloud		DDM
praguelcg2	praguelcg2_IT4I_MCORE	DE	aCT/Harvester	Arc-CE
LRZ-LMU	LRZ-LMU_MUC_MCORE1	DE	aCT/Harvester	Arc-CE
DESY-HH	DESY-HH_HPC	DE	aCT/Harvester	Arc-CE
MPPMU	MPPMU-HYDRA_MCORE	DE	aCT/Harvester	Arc-CE
MPPMU	MPPMU-DRACO_MCORE	DE	aCT/Harvester	Arc-CE
CSCS-LCG2	CSCS-LCG2-HPC_MCORE	DE	aCT/Harvester	Arc-CE
IFIC-LCG2	IFIC_ARC_TEST	ES	aCT/Harvester	Arc-CE
IFIC-LCG2	IFIC_MareNostrum4	ES	aCT/Harvester	Arc-CE
pic	pic_MareNostrum4	ES	aCT/Harvester	Arc-CE
NDGF-T1	HPC2N_MCORE	ND	aCT/Harvester	Arc-CE
NDGF-T1	NSC_MCORE	ND	aCT/Harvester	Arc-CE
NDGF-T1	UIO_MCORE	ND	aCT/Harvester	Arc-CE
NDGF-T1	UIO_MCORE_LOPRI	ND	aCT/Harvester	Arc-CE
RRC-KI-T1	RRC-KI-HPC2	RU	Harvester via CE	RSE

CMS HPC sites



HPC Center	HPC Machine Name	Cloud	Middleware	CVMFS	DDM
CSCS	Piz Daint	DE	aCT/ARC-CE	yes	CMS storage
Cineca	Marconi	IT	use CNAF's CE	yes	use CMS CNAF disks
BSC-CNS	MareNostrum	ES	(see next talk)	yes	
NERSC	Cori	US	HEPCloud/Bosco	yes	Input: remote read xrcp write to FNAL
PSC	Bridges	US	HEPCloud/OSG Hosted - CE/Bosco	yes	Input: remote read xrcp write to FNAL
SDSC	Comet	US	HEPCloud/OSG Hosted - CE/Bosco	yes	Input: remote read xrcp write to FNAL
TACC	Stampede2	US	HEPCloud/OSG Hosted - CE/Bosco	work in progress	Input: remote read xrcp write to FNAL
ALCF	Theta	US	just starting		

D. Benjamin
May 10, 2019

HPC Access Policies

- HPC resources are often proposal-driven annual allocations
- Sites supporting HEP have arrangements that last for many years
  - required for planning
- Longer term allocations and arrangements will be needed for HEP to rely on HPC sites
- **Ongoing discussions with EuroHPC and PRACE to adjust the resource allocation model to more longer lived**

21 October 2019 | Brussels |

Workshop on EuroHPC Systems Access Policy

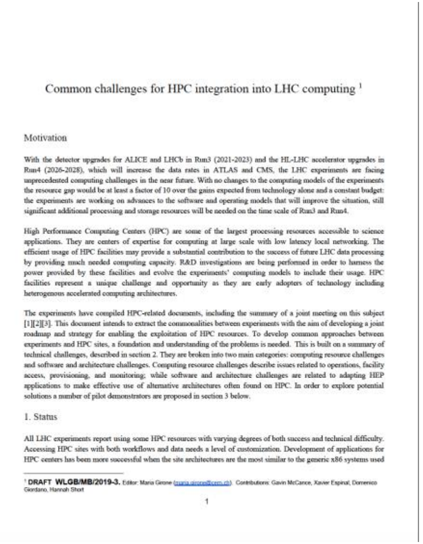
Conclusions

- HPC resources are large computing facilities with the potential to significantly increase the resources available to HEP
  - There is valuable expertise in computing at scale and application porting at the HPC sites
  - **Access to testbed systems** to port and optimize applications will be key (non x86 systems)

- Using them involves challenges
  - Different hardware architectures, software, cyber security, provisioning, data access models

- Through engagement and active R&D programs we can address the challenges to integrate these powerful resources into the WLCG computing environment
  - Common HPC challenges document from WLCG to facilitate discussions with HPC centres.

https://docs.google.com/document/d/1AN1d6Nu-khBsKnNH1MVvszqWdpcMfFGaYQMEVnS01Tc/



Common challenges for HPC integration into LHC computing [1]

Motivation

With the detector upgrades for ALICE and LHCb in Run3 (2021-2023) and the HL-LHC accelerator upgrades in Run4 (2026-2028), which will increase the data rates in ATLAS and CMS, the LHC experiments are facing unprecedented computing challenges in the near future. With no changes to the computing models of the experiments the resource gap would be at least a factor of 10 over the gains expected from technology alone and a constant budget: the experiments are working on advances to the software and operating models that will improve the situation, still significant additional processing and storage resources will be needed on the time scale of Run3 and Run4.

High Performance Computing Centers (HPC) are some of the largest processing resources accessible to science applications. They are centers of expertise for computing at large scale with low latency local networking. The efficient usage of HPC facilities may provide a substantial contribution to the success of future LHC data processing by providing much needed computing capacity. R&D investigations are being performed in order to harness the power provided by these facilities and evolve the experiments' computing models to include their usage. HPC facilities represent a unique challenge and opportunity as they are early adopters of technology including heterogenous accelerated computing architectures.

The experiments have compiled HPC-related documents, including the summary of a joint meeting on this subject [1][2][3]. This document intends to extract the commonalities between experiments with the aim of developing a joint roadmap and strategy for enabling the exploitation of HPC resources. To develop common approaches between experiments and HPC sites, a foundation and understanding of the problems is needed. This is built on a summary of technical challenges, described in section 2. They are broken into two main categories: computing resource challenges and software and architecture challenges. Computing resource challenges describe issues related to operations, facility access, provisioning, and monitoring; while software and architecture challenges are related to adapting HEP applications to make effective use of alternative architectures often found on HPC. In order to explore potential solutions a number of pilot demonstrators are proposed in section 3 below.

1. Status

All LHC experiments report using some HPC resources with varying degrees of both success and technical difficulty. Accessing HPC sites with both workflows and data needs a level of customization. Development of applications for HPC centers has been more successful when the site architectures are the most similar to the generic x86 systems used

[1] **DRAFT WLGB/MB/2019-3.** Editor: Maria Girone (maria.girone@cern.ch). Contributions: Gavin McCance, Xavier Espinal, Domenico Giordano, Hannah Short

1