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

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

Maria.Girone@cern.ch – CHEP2019
## 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**

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

Intel Xeon+ NVIDIA GPUs

Intel Xeon

Custom RISC

Power9 + NVIDIA GPUs
EuroHPC Roadmap to pre-Exascale

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

**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**
- **High-range Supercomputers**
  - Kajaani (FI) – FI, BE, CZ, DK, NO, PL, SE, CH, EE
  - Barcelona (ES) – ES, HR, PT, TR, IE
  - Bologna (IT) – IT, SI, HU
- **Medium-to-high range Supercomputers**
  - Bissen (LU) – LU
  - Minho (PT) – PT, ES
  - Ostrava (CZ) – CZ
  - Maribor (SI) – SI
  - Sofia (BG) – BG

**EuroHPC JU is the owner**

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

Maria.Girone@cern.ch – CHEP2019
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

• 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

https://indico.cern.ch/event/760705/
Challenges and R&D

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/

Not an exhaustive list!!!!!
Modular Super Computer Architectures

Traditional Homogeneous Supercomputing Architectures
- Application & Workloads: separate
- Single Module – One Tool for All Tasks

Future Composable Supercomputing Architecture
- Application & Workload: intertwined
- Multiple Modules – The Proper Tool for Each Task

JSC Roadmap
- Modular Supercomputers
  - 12 PF (2018)
  - 70 PF (2019)
  - 1 EF (2020)
  - Exa-Pilot (2022)
  - 2023/24 Exascale

DEEP-EST: Modular Architecture Prototype at JCS
- Key for JSC and EU Strategy

Maria.Girone@cern.ch – CHEP2019
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
Viktor Khristenko and Maria Girone, DEEP-EST Project

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

Demonstrate the ability to do a local reconstruction 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

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

A. Bocci, CHEP 2019

Maria.Girone@cern.ch – CHEP2019
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 available

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

Container images are made up of layers

A. Valassi, CHEP 2019
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

5. Traceability
6. Multi-tenancy
Token Issuer Model

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.

Maria.Girone@cern.ch – CHEP2019
**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

**ATLAS HPC Sites/ PanDA**

<table>
<thead>
<tr>
<th>ATLAS Site</th>
<th>Panda queue</th>
<th>Cloud</th>
<th>DAGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prague-LMU</td>
<td>prague-lgg2_0T4L_MOORE</td>
<td>DE</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>Lanzhou</td>
<td>Lanzhou-MUC_MOORE1</td>
<td>DE</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>DESY-HH</td>
<td>DESY-HH_HPC</td>
<td>DE</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>MPML</td>
<td>MPML-HYDRA_MOORE</td>
<td>DE</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>CSGS-LCG2</td>
<td>CSGS-LCG2_HPC_MOORE</td>
<td>DE</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>IFIC-LCG2</td>
<td>IFIC_ARC_TEST</td>
<td>ES</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>IFIC-LCG2</td>
<td>IFIC_MareNostrum4</td>
<td>ES</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>pic</td>
<td>MareNostrum4</td>
<td>ES</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>NICEF-T1</td>
<td>NICEF_MOORE</td>
<td>ND</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>NICEF-T1</td>
<td>NICEF_MOORE</td>
<td>ND</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>NICEF-T1</td>
<td>NICEF_MOORE</td>
<td>ND</td>
<td>aCT/Harvester</td>
</tr>
<tr>
<td>RRC-KIT</td>
<td>RRC-KIT-HPC2</td>
<td>RU</td>
<td>Harvester via CE</td>
</tr>
</tbody>
</table>

**CMS HPC sites**

<table>
<thead>
<tr>
<th>HPC Center</th>
<th>HPC Machine Name</th>
<th>Cloud</th>
<th>Middleware</th>
<th>CVMFS</th>
<th>DDN</th>
<th>CMS storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSN</td>
<td>Piz Daint</td>
<td>DE</td>
<td>aCT/ARC-CE</td>
<td>yes</td>
<td>use CMS CNAF disk</td>
<td></td>
</tr>
<tr>
<td>Gineca</td>
<td>Marcini</td>
<td>IT</td>
<td>use CNAF's CE</td>
<td>yes</td>
<td>use CMS CNAF disk</td>
<td></td>
</tr>
<tr>
<td>BSC-CNS</td>
<td>MareNostrum</td>
<td>ES</td>
<td>see next talk</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NERSC</td>
<td>Cori</td>
<td>US</td>
<td>HEPCloud/Bosco</td>
<td>yes</td>
<td>Input: remote read, data write to FNAL</td>
<td></td>
</tr>
<tr>
<td>PSC</td>
<td>Bridges</td>
<td>US</td>
<td>HEPCloud/OSG</td>
<td>yes</td>
<td>Input: remote read, data write to FNAL</td>
<td></td>
</tr>
<tr>
<td>SDSC</td>
<td>Comet</td>
<td>US</td>
<td>HEPCloud/OSG</td>
<td>yes</td>
<td>Input: remote read, data write to FNAL</td>
<td></td>
</tr>
<tr>
<td>TACC</td>
<td>Stampede2</td>
<td>US</td>
<td>HEPCloud/OSG</td>
<td>yes</td>
<td>Input: remote read, data write to FNAL</td>
<td></td>
</tr>
<tr>
<td>ALCF</td>
<td>Theta</td>
<td>US</td>
<td>just starting</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

---

**D. Benjamin**

May 10, 2019

Maria.Girone@cern.ch – CHEP2019
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
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-khBsKnh1MVvszqWdpcMFGaYQMEVnS01Tc/

Maria.Girone@cern.ch – CHEP2019