Interactive, On-demand and Bespoke Services on Hybrid Supercomputing and Cloud technologies (*HPC Friendliness*)

19th International Workshop on Advanced Computing and Analysis Techniques in Physics Research
Sadaf Alam
Chief Technology Officer
Swiss National Supercomputing Centre
March 13, 2019
About Swiss National Supercomputing Centre (CSCS)
Mission, Infrastructure and Services

- CSCS develops and operates cutting-edge **high-performance computing systems** as an essential service facility for Swiss researchers [https://www.cscs.ch](https://www.cscs.ch)

- High Performance Computing, Networking and Data Infrastructure
  - Piz Daint supercomputing platform
    - 5000+ Nvidia P100 + Intel E5-2690 v3 nodes
    - 1500+ dual-socket Intel E5-2695 v4 nodes
    - Single network fabric (10s of Terabytes/s bandwidth)
    - High bandwidth multi-Petabytes of scratch
  - Storage Systems
    (10s of PetaBytes online and offline)

- Services
  - Computing services
  - Data services
  - Cloud services
Users and Customers

- User Lab
  - Allocation based on scientific merit
- MeteoSwiss
- CHIPP (Swiss Higher Energy Physics Community)
- European research infrastructure projects

Consolidation of customers and services

10s of Peta ($10^{15}$) Floating-point operations/second

100s of Peta ($10^{15}$) Bytes storage
The largest data centre in Europe is located in a tiny village in Norway. Opening in the fourth quarter of 2018, the Kolos Data Centre will cover 6.5 mn sq. ft across four storeys, and is being billed as a hyper-scalable data centre, with plans to consume up to 1000 megawatts of power by 2027.

1 | China Telecom Data Centre, China
2 | China Mobile Hohot, China
3 | The Citadel, United States
4 | Harbin Data Centre, China
5 | Kolos Data Centre, Norway
6 | Range International Data Centre, China
7 | Switch SUPERNAP, United States
8 | Dupont Fabros Technology
9 | Lakeside Technology Centre, United States
10 | Tulip Data Centre, India

https://www.gigabitmagazine.com/top10/top-10-biggest-data-centres-world
HYPERSCALE DATA CENTER MARKET TREND WILL BE FUELED BY RISE IN THE ADOPTION OF CLOUD COMPUTING AND ONLINE SERVICES

Compute, storage, network

https://newatlas.com/inside-google-data-centers/24654/

X-as-a-Service (Cloud & HPC Data Centre)

**Performance & Scaling-as-a-Service**
- Parallel computing
- Parallel file system technologies (POSIX based)
- Bulk processing, scale out with fast, integrated ecosystem
- High bandwidth networking subsystems
- Internal and external connectivity for high throughput data transfers
- …

**Automation and Interactivity-as-a-Service**
- IaaS (ownership infrastructure and services)
- On-demand
- High availability through service migration
- Roles based access control
- Storage models for role based access controls
- Isolation, security and QoS
LHC Workflow on Piz Daint → A Success Story (Bespoke Service)
The LHConCRAy project at CSCS

- Consolidation project to run LHC jobs on Piz Daint
  - Partners: CSCS, CHIPP (Swiss Institute of Particle Physics - ATLAS, CMS, LHCb)
  - Started ~2 year ago with preliminary studies on a Cray TDS
  - **Started production in April 2017 on Piz Daint:** 25 Cray nodes/1600 cores (ATLAS:CMS:LHCb - 40:40:20)
  - Operated in parallel with Phoenix
  - The goal is to run **ALL VO workloads without changes to the experiments’ workflows**

- Normal workflow:
  - Plugs transparently in to the experiments’ WMSs

- Roadmap
  - Measure performance in the production environment
  - Produce a cost study (until Dec. 2017)
  - Decision due: **migrate to the Cray or revert to invest on Phoenix**
CHALLENGES

WLCG computing on HPC systems

- Several challenges arise
  - Processor architecture and/or OS might not always be suitable complex software re-builds, environment tweaking, etc..
  - Compliance with tight access rules single-user access, username/password
  - Application provisioning a single ATLAS release is ~20GB, release cycles are very short/unpredictable
  - Workload management integration requires in general outbound IP connectivity
  - Data input and retrieval for real data processing: ~0.2MB/s/core IN, ~0.1MB/s/core OUT
Strategic and Operational issues (1) – HPC - LHC

- Challenge for computing resources to LHC experiments – over next 8 years need a factor of ~50 more resources.

- Switzerland started project LHConCRAY in 2016 (initiated at AEC-Bern) to test possibility and economy of LHC workloads on HPCs.
- December 2017: concluded tests successfully.
  - Team CSCS+CHIPP succeeded to run ALL LHC job-types on CRAY! found same job efficiency as PHOENIX, but higher economic value.
  - Meeting of "CHIPP LHC computing board" on 7.12.2017, decided to go for using HPC for providing the Swiss T2-resources at CSCS.
    1) CSCS will provide shared HPC resources for LHC computing, based on same FLAT budget by FLARE/SNF (and ETHZ+Uni contributions)
    2) We will continue to provide the pledges of Switzerland towards WLCG
    3) PHOENIX as a "separate dedicated cluster" will be phased out eventually.
    4) AEC at Bern continues providing additional ATLAS-T2 resources
All experiments running on Piz Daint, a supercomputing platform

No more operations on a dedicated cluster

Service consolidation benefits for the users and customers

High throughput computing for HEP middleware & workflows transparently (for users) running on a Petascale system
  - HPC friendliness?

Dedicated multi-year effort and collaboration between CSCS and our customers
Bridging the Gap → Creating New Abstractions

- Light-weight operating system (SLES based)
  - Possible solution: containers or other virtualization interfaces

- Diskless compute nodes
  - Possible solution: exploit burst buffer or tiered storage hierarchies

- Computing nodes connectivity (high speed Aries interconnect)
  - Possible solution: web services access with no address translations overhead
HPC Friendliness

what are other words for friendliness?

amiability, warmth, affability, cordiality, friendship, geniality, kindness, kindliness, amity, benevolence
Experiences & Expectations

- Classical HPC users
- Data science HPC users
- Extreme data workflow / experimental facilities

```bash
#!/bin/bash -l
#SBATCH --job-name=job_name
#SBATCH --time=01:00:00
#SBATCH --nodes=2
#SBATCH --ntasks-per-core=2
#SBATCH --ntasks-per-node=12
#SBATCH --cpus-per-task=2
#SBATCH --partition=normal
#SBATCH --constraint=gpu
export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK
export CRAY_CUDA_MPS=1
module load daint-gpu srun
./executable.x
```
Classification & Abstractions

- **Classic HPC use cases**
  - Performance & Scaling as a Service
  - Batch processing
  - Access to low level toolchains
    - Already friendly
- **Data Science HPC use cases**
  - Software as a Service
  - (Big) Data as a Service
  - Interactivity
  - Elasticity
- **Extreme Data Workflow use cases**
  - Workflow as a Service
  - Automation as a Service
  - Need privileged access to infrastructure services
  - Need composable platform services
    - Need cloud++ delivery model or hybrid cloud & HPC to be friendly
Swissuniversities funded project with PSI
SELVEDAS (Services for Large Volume Experiment-Data Analysis utilizing Supercomputing and Cloud technologies at CSCS)
Highlights:
Archival storage for the new SwissFEL X-ray laser and Swiss Lightsource (SLS)
A total of 10 to 20 petabytes of data is produced every year
A dedicated redundant network connection between PSI and CSCS, 10 Gbps
CSCS tape library current storage capacity is 120 petabytes, can be extended to 2,000 petabytes
By 2022, PSI will transfer around 85 petabytes of data to CSCS for archiving. Around 35 petabytes come from SwissFEL experiments, and 40 come from SLS.
Growth in Data Volume → Online & Offline Data Processing

User applies for beamtime

Prior to upgrade

After the upgrade
Hybrid Cloud and HPC

**Today**

IT infrastructure (experimental facility)

CSCS IT infrastructure (data centre)

Custom interfaces for compute & data services

Experimental facility users

**Target**

IT infrastructure (experimental facility)

CSCS IT infrastructure (data centre)

Infrastructure services
- AAI
- Containers

Exp. facility community platforms

Experimental facility users

**Future**

IT infrastructure

Other IT infrastructure (data centre)

CSCS IT infrastructure (data centre)

Infrastructure services
- AAI
- Containers

Other community platforms

Other community users

End users
Future Outlook
Going Forward with X-as-a-Service (Hybrid Cloud and HPC)

- **On-demand**
  - Reservation service for HPC resources
  - Batch with delegated, privileged access
  - Coordination of experiments with access to Petascale computing resources

- **Bespoke**
  - Offer Infrastructure-as-a-Service (IaaS), e.g. virtual machines
  - Privileged access without compromising performance and security
  - Web service access to HPC resources

- **Interactive**
  - Resource management and scheduling (batch and service-oriented)
  - Resource utilization metrics (policy)

- Making HPC ecosystems becoming friendlier …
Thank you for your attention.