FENIX

RESEARCH INFRASTRUCTURE

D4.2

Infrastructure at ETHZ/CSCS

Work package:	WP4 Procurement and deployment	
Author(s):	Colin McMurtrie	ETHZ/CSCS
Reviewer #1	Anne Carstensen	JUELICH
Reviewer #2	Dirk Pleiter	JUELICH
Dissemination Level	public	
Nature	Other	

Date	Author	Comments	Version	Status
22.10.2018	Colin McMurtrie	First draft version.	v0.1	Draft
"	Colin McMurtrie	Minor fixes.	v0.2	Draft
24.10.2018	Dirk Pleiter	Review	v0.21	Draft
25.10.2018	Anne Carstensen	Review/Editorial updates	v0.22	Draft
27.10.2018	Dirk Pleiter	Split in public part and confidential annex	v0.3	Draft
30.10.2018	Anne Carstensen	Approved as is by work package leader, Jacques-Charles Lafoucriere	v0.4	Final

The ICEI project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800858.

© 2018 ICEI Consortium Partners. All rights reserved.

Executive Summary

This document outlines the status of the ICEI infrastructure available at ETHZ/CSCS. We include information on the timeline for the availability of the resources to HBP users (i.e. when the resources were made available to HBP users) and when the first projects were identified and started using the resources.[1] We do not include details on resource usage since these will be reported separately (namely in D4.9).

Contents

Acronyms

AAI	Authentication and Authorization Infrastructure
ACD	Active Data Repositories
ACL	Access Control List
API	Application Programming Interface
ARD	Archival Data Repositories
BSC	Barcelona Supercomputing Center
CapEx	Capital Expenditure
CDP	Co-design Project

[1] For an overview over allocated projects see Confidential Annex.

CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CINECA	Consorzio Interuniversitario
CLI	Command Line Interface
CSCS	Centro Svizzero di Calcolo Scientifico
DL	Data Location Service
DM	Data Mover Service
DT	Data Transfer Service
FPA	Framework Partnership Agreement
FURMS	Fenix User and Resource Management Services
GoP	Group of Procurers
GUI	Graphical User Interface
HBP	Human Brain Project
HPAC	High Performance Analytics and Computing
HPC	High Performance Computing
HPDA	High Performance Data Analytics
HPST	High-Performance Storage Tier
IaaS	Infrastructure as a Service
IAC	Interactive Computing Services
ICCP	Interactive Computing Cloud Platform
ICEI	Interactive Computing E-Infrastructure for the Human Brain Project
ICN	Interactive Computing Node
IdP	Identity Provider
IPR	Intellectual Property Rights
JP	Joint Platform
JSC	Jülich Supercomputing Centre
LCST	Large-Capacity Storage Tier
MS	Monitoring Services
NDA	Non-Disclosure Agreement
NETE	External Interconnect
NETI	Internal Interconnect
NMC	Neuromorphic Computing

NVM	Non-Volatile Memory
NVRAM	Non-Volatile Random Access Memory
OIDC	OpenID Connect
OpEx	Operational Expenditure
PaaS	Platform as a Service
PCP	Pre-Commercial Procurement
PI	Principal Investigator
PID	Persistent Identifier
PIE	Public Information Event
PRACE	Partnership for Advanced Computing in Europe
Q&A	Questions and Answers
QoS	Quality of Service
R&D	Research & Development
R&I	Research & Innovation
RBAC	Role-Based Access Control
RFI	Request For Information
SCC	Scalable Computing Services
SGA	Special Grant Agreement
SIB	Science Infrastructure Board
SLA	Service Level Agreement
SP	Subproject
TCO	Total Cost of Ownership
TGCC	Très Grand Centre de Calcul
UI	User Interface
US	User Support Services
VM	Virtual Machine Services

1. Introduction

Since the inception of the ICEI project, ETHZ/CSCS was identified as a fast-track partner for the installation and availability of (SCC, IAC, VM, ACD, ARD) resources and services to HBP users. The reason for this was that ETHZ/CSCS could leverage existing procurement contracts to procure the needed additional hardware on a short timeline and, moreover,

had sufficient immediately usable capacity to enable the co-financed contributions to be realised very early in the project timeline.

2. Details of CSCS' ICEI Infrastructure Elements

ETHZ/CSCS has provided or installed infrastructure elements for the ICEI project that run the gamut of those foreseen in the project proposal. Table 1 summarises the resources and the subsequent subsections provide more detail on each class of resource.

Table 1: Breakdown of resources made available at ETHZ/CSCS.

				Quarterly Allocation (2018)		
Component	**Service Type**	**ICEI Total Allocation (100%)**	**HBP Total Allocation (25%)**	**Q2**	**Q3**	**Q4**
Piz Daint Multicore	SCC	250 nodes	63 nodes	116'344 node-hrs	116'344 node-hrs	116'344 node-hrs
Piz Daint Hybrid	SCC + IAC	400 nodes	100 nodes	186'150 node-hrs	186'150 node-hrs	186'150 node-hrs
OpenStack IaaS	VM	35 servers	8.75 servers	8.75 servers	8.75 servers	8.75 servers
POSIX, Object and Tape	ARD	4 PB	1 PB	1 PB	1 PB	1 PB
Low-Latency Storage Tier	NVM	80 TB	20 TB	20 TB	20 TB	20 TB

2.1 Scalable Computing Services (SCC)

ETHZ/CSCS chose to fully integrate the compute resources into its flagship system, *Piz Daint*, in order to allow users to benefit from the large-scale scalability of the system. *Piz Daint* is one of the most powerful HPC systems available globally to public-good science and hence is an ideal candidate to allow users to gain access to state-of-the-art compute capability. For more general details on the *Piz Daint* system see https://www.cscs.ch/computers/piz-daint/.

The scalable computing resources are made available via a batch scheduling system (aka Workload Manager; WLM) for which, in due course, a RESTful API will be made available to provide infrastructure services to domain-specific portals. In addition to these batch-accessible scalable computing resources, ETHZ/CSCS has also planned to allow

interactive use of some of the nodes in the system, thereby enabling various interactive workloads such as those coming from the Brain Simulation Platform (BSP), the Neurorobotics Platform (NRP) and several CDPs (see D3.6 – "Scientific Use Case Requirements Documentation" for more details).

2.1.1 Multicore Nodes

As shown in Table 1, a total of 63 nodes from the multicore partition of *Piz Daint* have been assigned to the HBP as part of the ICEI allocation. Taking into consideration the uptime of the system and scheduling considerations of the WLM, these nodes equate to a total multicore compute allocation of 116'344 node-hours (node-hrs) per quarter (i.e. 465'376 node-hrs annually).

The technical specifications of the *Piz Daint* multicore nodes are summarised in Table 2.

Table 2: Piz Daint multicore node specifications.

	Processor Type	Processor Frequency	Cores/Sockets	Memory	Interconnect Configuration
Multicore Node	Intel® Xeon® E5-2695 v4	2.10 GHz	18C/dual-socket	64 or 128 GB	Cray Aries

2.1.2 Hybrid Nodes

As shown in Table 1, a total of 100 nodes from the hybrid partition of *Piz Daint* have been assigned to the HBP as part of the ICEI allocation. Taking into consideration the uptime of the system and scheduling considerations of the WLM, these nodes equate to a total multicore compute allocation of 186'150 node-hrs per quarter (i.e. 744'600 node-hrs annually).

The technical specifications of the *Piz Daint* hybrid nodes are summarised in Table 3.

Table 3: Piz Daint hybrid node specifications.

	Processor Type	Processor Frequency	Cores/ Sockets	Memory	GPU Type	GPU Memory	Interconnect Configuration
Hybrid Node	Intel® Xeon® E5-2690 v3	2.60GHz	12C/ single socket	64GB	NVIDIA® Tesla® P100	16GB CoWoS HBM2	Cray Aries

Figure 1 shows the layout of a hybrid Cray XC50 compute blade (per blade there are 4 nodes of the type summarised in Table 3).

Figure 1: Image showing a hybrid Cray XC50 compute blade from the Piz Daint system. Note that there are 4 compute nodes per blade.

2.1.3 High Bandwidth Low-Latency File System

The *Piz Daint* system features two integrated scratch file systems based on Cray's Sonexion 1600 and 3000 storage technologies and uses the Lustre file system. These file systems can be considered as ACD and each provides different capacities and performance to users (note that the Sonexion 3000 file system is the default and that the Sonexion 1600 is only provided as a fail-over and for specific needs). Table 4 summarises the specifications of both scratch file systems.

Table 4: Piz Daint scratch file system specifications.

	Usable Capacity	Performance	Lustre Version
Cray Sonexion 1600[2]	2.7 PB	138 GB/s	v2.5[3]
Cray Sonexion 3000	8.8 PB	112 GB/s	v2.7[4]

2.1.4 High Bandwidth Low-Latency Storage Tier

The *Piz Daint* system also contains two generations of Cray's DataWarp technology, which enables the provisioning of SSDs into the systems as peer nodes on the interconnect fabric. Cray provides special blades that have the SSDs integrated into them and the nodes are visible within the system. Cray provides a software layer to enable the DataWarp nodes to be used in various workloads. Unfortunately, this software environment has been targeted at use cases that differ from those found in the ICEI project and hence ETHZ/CSCS is working with Cray to augment the software environment to enable more use cases, particularly those relating to interactive reuse of data

[2] Used as a failover file system and for specific needs.

[3] No longer upgradable due to the age of the Sonexion 1600 infrastructure.

[4] Has improved metadata performance due to Lustre v2.7 and the backend metadata infrastructure.

persistently, to be hosted within the DataWarp storage tier. Work is actively on-going in this regard.

As shown in Table 1, 20 TB of the high bandwidth, Low-Latency DataWarp Storage Tier are available to HBP users within the ICEI allocation.

2.2 Infrastructure as a Service (Cloud) Resources

ETHZ/CSCS made available Infrastructure as a Service (IaaS) Cloud resource elements in a pre-existing dedicated environment created for the purpose of hosting domain-specific portals and services. The system, known as *Pollux*, was already in use by SP5 (Neuroinformatics Platform) for hosting curated data and for hosting Collaboratory services, including the JupyterHub service, as part of the SGA1 work, the resources having been offered in-kind due to a crisis situation within the HBP at that time.

With the advent of the ICEI project the opportunity arose to formalise the hosting arrangement for the SP5 curated data and Collaboratory services as well as to provide the same type of resources to other Platforms within the HBP (including the NRP).

The *Pollux* system offers a Red Hat OpenStack Platform (RHOSP) environment (originally v11, now v12 and soon to be v13) for platform service hosting. As such the system was designed with growth in mind and was installed with additional capacity. Hence ETHZ/CSCS was easily able to offer the system resources as an in-kind (i.e. co-financed) contribution from Day 1 of the ICEI project. As shown in Table 1, the unit of resource allocation was chosen as "servers" since there is a wide variation in the VM requirements coming from various platforms and thus we had difficulties to arrive at a universally acceptable "Standard VM" unit. As an indication however, each server in the *Pollux* system is capable of hosting approximately 2000 lightweight VMs so the allocated 8.75 servers represent a sizable IaaS resource for the HBP.

In addition to the VM hosting services, the *Pollux* IaaS environment features Swift Object Storage (OS). This Swift (OS) service is built upon ETHZ/CSCS' site-wide IBM SpectrumScale (aka GPFS) storage infrastructure, rather than the internal CEPH storage service of OpenStack. In this way ETHZ/CSCS is able to benefit from the economies of scale that come with its >15PB site-wide storage infrastructure. The Swift OS service is fully integrated into the *Pollux* IaaS environment and uses the same KeyStone AAI which in turn is integrated with the site-wide AAI using Red Hat Single Sign On (RH-SSO, aka KeyCloak). KeyCloak can also function as an identity broker to external IdPs and integrates well with SAML and OAuth 2.0 authorisation layers and hence, in this way, the *Pollux* system and the Swift OS service, in particular, has been prepared for these capabilities which will be provided by the ICEI project in the future.

At present 100 TB of Swift OS capacity is available to the HBP users. The backend SpectrumScale/GPFS storage infrastructure is built upon a Storage Area Network (SAN) and additional capacity can be added in multi-hundred TB increments, if needed. The *Pollux* Swift OS service can be regarded as an instance of ACD.

2.3 Archival Data Repositories (ARD)

As mentioned above, ETHZ/CSCS has a site-wide IBM SpectrumScale/GPFS parallel file system built upon a Storage Area Network (SAN). This infrastructure also includes an IBM Spectrum Protect (formerly Tivoli Storage Manager) environment, including a 30 PB tape library, for storage tiering. As such, this storage infrastructure may be regarded as an instance of ARD.

Various SpectrumScale/GPFS file systems are available at ETHZ/CSCS and these are mounted on the various compute systems available at the Centre. Specifically, the *Piz Daint* system mounts a subset of the file systems on its external login nodes and a reduced set on its compute nodes. Hence to enable users to move data between the scratch file systems and the SpectrumScale/GPFS file systems, an integrated data mover cluster is provided. Users can chain jobs together so that data can be staged into the scratch file system for use as input data and output data can be staged out after computation and analysis. It is also possible to keep persistent data sets (such as common input data) in special locations on the GPFS infrastructure so as to avoid the purge policy of the scratch file systems and to prevent the continual need to copy such data.

Work is on-going at ETHZ/CSCS to provide a RESTful API to the data mover cluster to thereby enable users to create and use platform-level data workflows. At present the capabilities in this regard are somewhat limited.

As shown in Table 1, ETHZ/CSCS provides a total of 1 PB of storage to the HBP ICEI allocation within the site wide SpectrumScale/GPFS+TSM storage infrastructure. These resources were provided as co-financed contributions to the project.

3. Timelines for Resource Availability

Due to forward planning (including the associated risk that the ICEI project would not be funded) ETHZ/CSCS was able to provide all co-financed resources to the project from 1 April 2018, as originally set forth in the project implementation plan. Furthermore, thanks to the available compute capacity on the *Piz Daint* system, the scalable compute allocations to both HBP and PRACE could be honoured with limited impact on the large number of other users on the system. This was largely possible due to some reserve capacity on the system and the ability of the WLM to schedule jobs in an efficient manner.

Once the ICEI project was approved, ETHZ/CSCS moved quickly to procure the necessary additional scalable and interactive computing resources shown in Table 1. Here again, due to delivery timelines and the late signing of the ICEI SGA, ETHZ/CSCS had to assume some financial risk to ensure that the infrastructure elements could be installed in a timely manner and in accordance with the agreed project timeline. As such, these scalable

and interactive computing resources were installed and made available in the mid-September 2018 timeframe.

4. Concluding Remarks

ETHZ/CSCS was able to assume some financial risk to ensure that the agreed HBP ICEI resource allocations could be honoured in accordance with the original project implementation plan. At the time of writing all infrastructure elements are installed and operational. Some additional work is needed to provide some additional functionality on the software environments that support these infrastructure elements, in particular the high bandwidth, Low-Latency Storage Tier and RESTful APIs for various infrastructure services. However, work is actively on-going in these latter areas and, in the meantime, users are able to continue working and are making good use of the assigned resources.