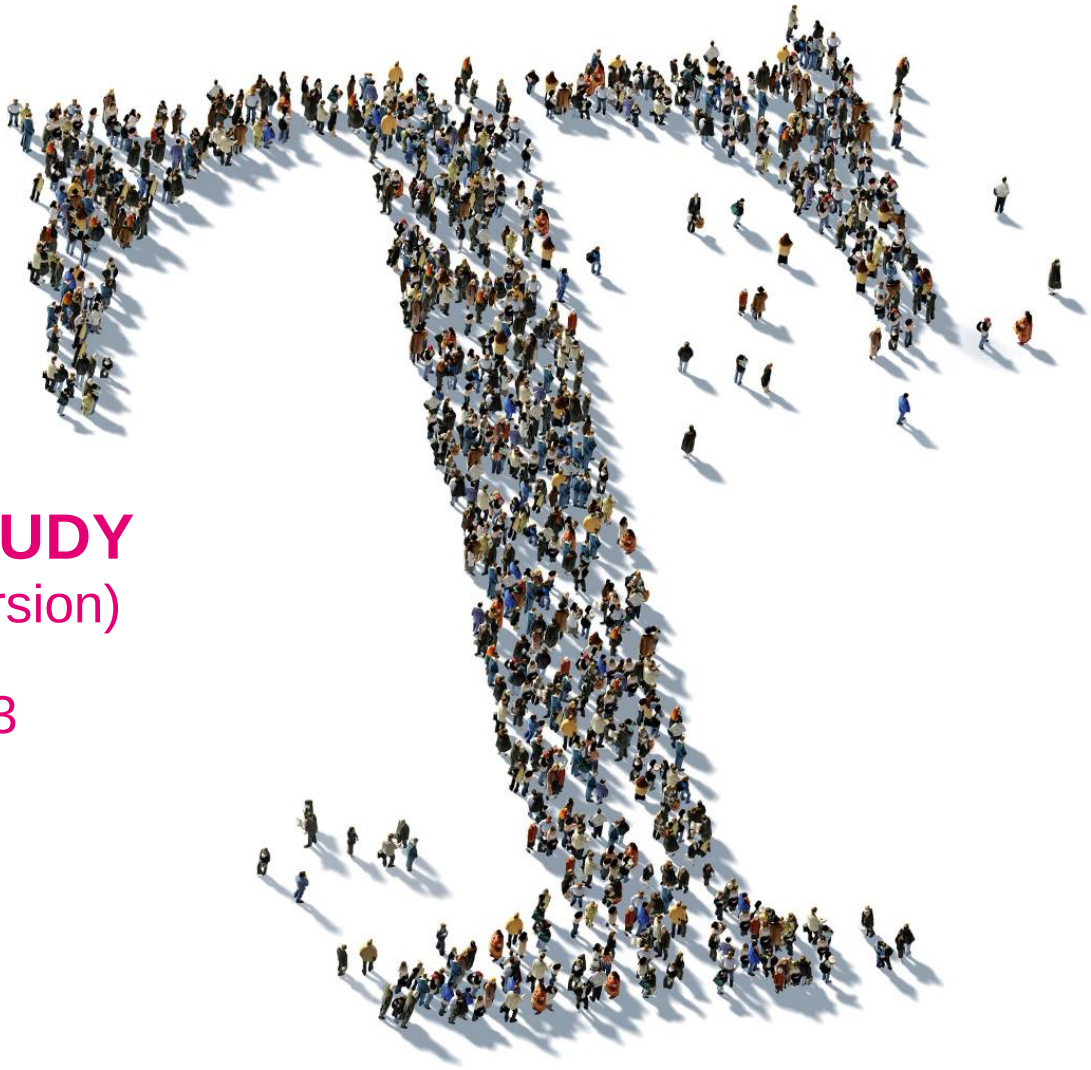

TCO STUDY
(public version)

D-PIL-3.13

HNSciCloud
for
CERN
CH-1211
Genève 23

Issue: 1.0
Date: 28/01/2019

T · · Systems ·

VERSION HISTORY

Version	Date	Editor	Changes / Remarks
1.0	28/01/2019	Jurry de la Mar	Issue for public dissemination

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Purpose

The document contains a summary of a detailed TCO study of selected scientific use cases of the Helix Nebula project and is released for public dissemination. The buyers group have selected relevant use cases and provided the high-level input requirements and workflow information to perform a TCO study. Based on this, T-Systems have evaluated alternative approaches how the use cases can be supported most efficiently by commercial cloud services and derived the TCO per use case.

The document includes a description of the general approach taken and how this may be used for other use cases or TCO studies in future.

1.2 Scope

Summary TCO study of selected scientific use cases for the HNSciCloud PCP project.

1.3 General approach

The Buyers Group has proposed two use cases providing extensive information: Pan-Cancer (supported by EMBL) and ALICE Monte Carlo, reconstruction and analysis trains (supported by CERN, CNRS, INFN, STFC and SURFsara) as the basis for the TCO study to be performed during the final Pilot Phase of the PCP procurement.

The TCO should use the same elements as those found in the ECAR framework for public sector higher education organisations [RD-01]. It was agreed with the BG, that the study can be provided as a report and does not have to include the ECAR Excel-Templates. The report – with separate sections for each of the selected use-cases – follows the ECAR structure including following elements:

- Tangible and Intangible value of using the proposed service solutions
- Quantitative factors including One-time and Ongoing costs
- Qualitative factors including Agility, Elasticity and Scalability, Regulatory and Policy Requirements, Security and Service Levels
- Recommendations for how the TCO can be reduced
- Suggestions for how to make the TCO more accurate/relevant

Regarding networking alternative approaches how to connect users to a commercial cloud may be considered and compared.

The TCO study is based on a use period of 12 months. It is understood that longer term procurement contracts will be financially advantageous and alternative periods as an improvement to the TCO are included in this version.

Since the essence of the study is to compare the TCO to alternative deployments e.g., based on on-premise e-Infrastructures of the BG, the report includes a high-level comparison with on-premise alternatives. The on-premise costs were derived from public available documents and related science publications.

1.4 Reference Documents

Table 1: Reference Documents

Number	Description	Source
RD-01	TCO for Cloud Services – a framework	ECAR working group paper, April 24, 2015 http://www.educause.edu/library/resources/tco-cloud-services-framework
RD-02	Open Telekom Cloud Pricing Models	T-Systems

1.5 Glossary

Table 2: Abbreviations used

Abbreviation	Description
ALICE	A Large Ion Collider Experiment
BG	Buyers Group
CEPH	A free-software storage platform, that implements object storage on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage.
CERN	Conseil Européen pour la Recherche Nucléaire
ECAR	Educause Center for Analysis and Research, a nonprofit association that helps higher education elevate the impact of IT
EMBL	European Molecular Biology Laboratories
EOSC	European Open Science Cloud
GDPR	EU General Data Protection Regulation
HEP	High-Energy Physics
HS06	HS06 is the HEP-wide benchmark for measuring CPU performance, based on SPEC2006
LHC	Large-Hadron Collider, the world largest and most powerful particle accelerator
LLSM	Live-cell Lattice light-sheet microscopy
OBS	Object-based storage, a computer data storage architecture that manages data as objects and is provided as a service option on OTC
Standard OBS	A version of OBS service on OTC, that is optimized for millisecond access and has commercially attractive conditions for dynamic short-term use and frequent access
Warm OBS	A version of OBS service on OTC, that is optimized for millisecond access and has commercially attractive conditions for longer term use and less-frequent access
OTC	Open Telekom Cloud
PAINT	Points accumulation for imaging in nanoscale topography, an easy-to-implement approach to localization-based super-resolution microscopy
PALM	Photoactivated Localization Microscopy, a fluorescence microscopy imaging method that allows obtaining images with a resolution beyond the diffraction limit
PCP	Pre-Commercial Procurement, a competitive R&D procurement programme of the European Union
POSIX	Portable Operating System Interface (POSIX)[1] is a family of standards specified by the IEEE Computer Society for maintaining compatibility between operating systems
PPR	Price-Performance-Ratio
R12M, RU36M	Price Models of OTC, explained in the text
ROM	Rough-Order-of-Magnitude
SATA	Serial Advanced Technology Attachment, a computer block storage technology provided as a service on OTC

Abbreviation	Description
SSD	Solid State Drive, a computer block storage technology provided as a service on OTC
SLA	Service Level Agreement
SPEC	Standard Performance Evaluation Corporation, www.spec.org
STORM	Stochastic Optical Reconstruction Microscopy, , a fluorescence microscopy imaging method that allows obtaining images with a resolution beyond the diffraction limit
TCO	Total-Cost-of-Ownership
VM	Virtual Machine
VRF	Virtual Routing and Forwarding, a technology that allows multiple instances of a routing table to co-exist within the same router at the same time. Used to connect BG members to OTC through GÉANT, without being impacted by other users
WLCG	World-wide LHC Computing Grid

2 USE CASE DESCRIPTIONS

2.1 Pan-Cancer (EMBL)

The Pan-Cancer initiative primary goal is to compare 12 tumor types profiled in the context of The Cancer Genome Atlas Research Network. Cancer can take hundreds of different forms, depending on external factors such as localization and cell type. Pan-Cancer will look for shared molecular aberrations and will try to identify their impact on the evolution of the disease. Eventually, this will help extend therapies that are already known to be effective against a certain type of cancer to similar cancer types at the genomic level.

Storage requirements:

To deploy a dataset size of 1 PB with more than 4,000 files in the 5-30GB range and more than 4,000 files in the 100-500 MB range. Outputs ca. 92,000 files in kb range. POSIX access.

Compute requirements:

Up to 400 VMs (burst), with at least 250 VMs running concurrently, where each VM will be a 8-16 vCPUs, 16G RAM and 30 GB scratch disk.

Network requirements:

1.25 Gbit/s from EMBL-EBI to provider. Load during these 12 months will have a burst pattern, meaning that a minimal set of resources will be in constant use, and at periods of incoming data we will ramp up resource usage to a maximum. Ballpark figures below:

Continuous deployment (will be running for the full 12 months):

- Compute: 7 VMs, min. 50 vCPUs, 50GB RAM, 8 vCPU per VM
- Storage: 0.5PB, not accessed frequently
- Network: Minimal

During burst (24-48 hours duration, 1-2x / month):

- Compute: ~250 VMs, 1000 vCPUs, 8,5TB RAM
  Alternative to be based on 250 VMs with 16 cores and 1 GB RAM per core.
- Storage: 0.5PB, random access with high I/O requirement
- Network: 10GBit/s intra-node traffic (within the vLAN), up to 4GBit/s ingress

2.2 ALICE (HEP)

ALICE is a heavy-ion detector on the Large Hadron Collider (LHC) ring. It is designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms.

There are 3 types of jobs (workloads) within the ALICE use-case:

- Monte Carlo (detector simulation)
- Raw Data reconstruction
- Analysis Trains

2.2.1 ALICE Assumptions

Each job uses a single core. Each type of job is described below.

In overall size, it shall be assumed, that 50,000 concurrent jobs will be running throughout the year. The best metric shall be the cost per job per type, considering that the cloud resources would be filled with a given type throughout the year, should the pricing allow it. That is, there should be no shortage of workloads.

Multiple jobs can be run on a single VM. It implies that a VM does not have to be sized specifically for an ALICE job. Instead, the VM specs must be compatible with running N jobs concurrently, for some value of N that is deemed desirable by the cloud provider or the (CERN-IT) group instantiating the VMs.

At CERN alone ALICE run ca. 50k+ concurrent jobs at all times. If some of the resources were located in the cloud, they would get a proportional fraction of that workload.

2.2.2 ALICE Monte Carlo

The Monte Carlo jobs often have no fixed deadline for execution and can be considered low priority suitable for ‘backfilling’ low-cost unused capacity. There is a very large quantity of Monte Carlo jobs that can be executed throughout the year.

Storage requirements: None

Compute requirements:

- Typical job duration: 6 hours
- RAM: at least 2 GB/core
- Swap: at least 2 GB/core in addition
- Scratch disk space: at least 2 GB per core

Network requirements:

- Input download: ~200 MB
- Output upload: ~350 MB
- Network bandwidth: ~0.3 Mbps

2.2.3 ALICE raw data reconstruction

There often is a steady supply of ALICE raw data reconstruction jobs throughout the year, either to process recently recorded data or to reprocess older data.

Storage requirements: None

Compute requirements:

- Typical job duration: 7 hours
- RAM: at least 2 GB/core
- Swap: at least 2 GB/core in addition
- Scratch disk space: at least 10 GB per core

Network requirements:

- Input download: ~2 GB

- Output upload: ~0.7 GB
- Network bandwidth: ~1 Mbps

2.2.4 ALICE analysis trains

These jobs neither download a lot at the start, nor upload a lot at the end, but do read a lot of input data that is streamed while the jobs are running. Hence, they have by far the highest network requirements. There usually is a steady supply of such jobs throughout the year.

Storage requirements: None

Compute requirements:

- Typical job duration: between 15 min and 5 hours, with an average of 2 hours
- RAM: at least 2 GB/core
- Swap: at least 6 GB/core in addition
- Scratch disk space: at least 10 GB per core

Network requirements:

- Input download: (small)
- Output upload: (small)
- Network bandwidth: ~50 Mbps for streaming remote input at good efficiency (+80%)

3 EVALUATION

3.1 Pan-Cancer (EMBL)

3.1.1 Tangible and intangible values

The storage and compute requirements for the Pan-Cancer use case represent an attractive use case for commercial public cloud services. The overall dimension of resources is well-defined, and most of the compute resources are required with a burst pattern and for a limited time only. With a traditional deployment based on an on-premise infrastructure, such provision of resources will be only economically viable if the organisation has other applications that can make use of the resources when idle.

Tangible values

Storage can be based on both block and object storage. To achieve the best balance in price-performance and optimize TCO, the use case will benefit from the use of a Global Namespace solution e.g., Onedata. With Onedata, less-frequent accessed data can be shifted to lower cost object storage and high-I/O data can based on a high-performing object/file-system with block storage. Onedata will maintain a full transparent POSIX access across the user data collection and the various cloud storage pools. It has been demonstrated during the PCP project that CEPH in combination with Onedata provides the best I/O-performance with block storage. An overview of the data management scenario is depicted in the Figure 1 below.

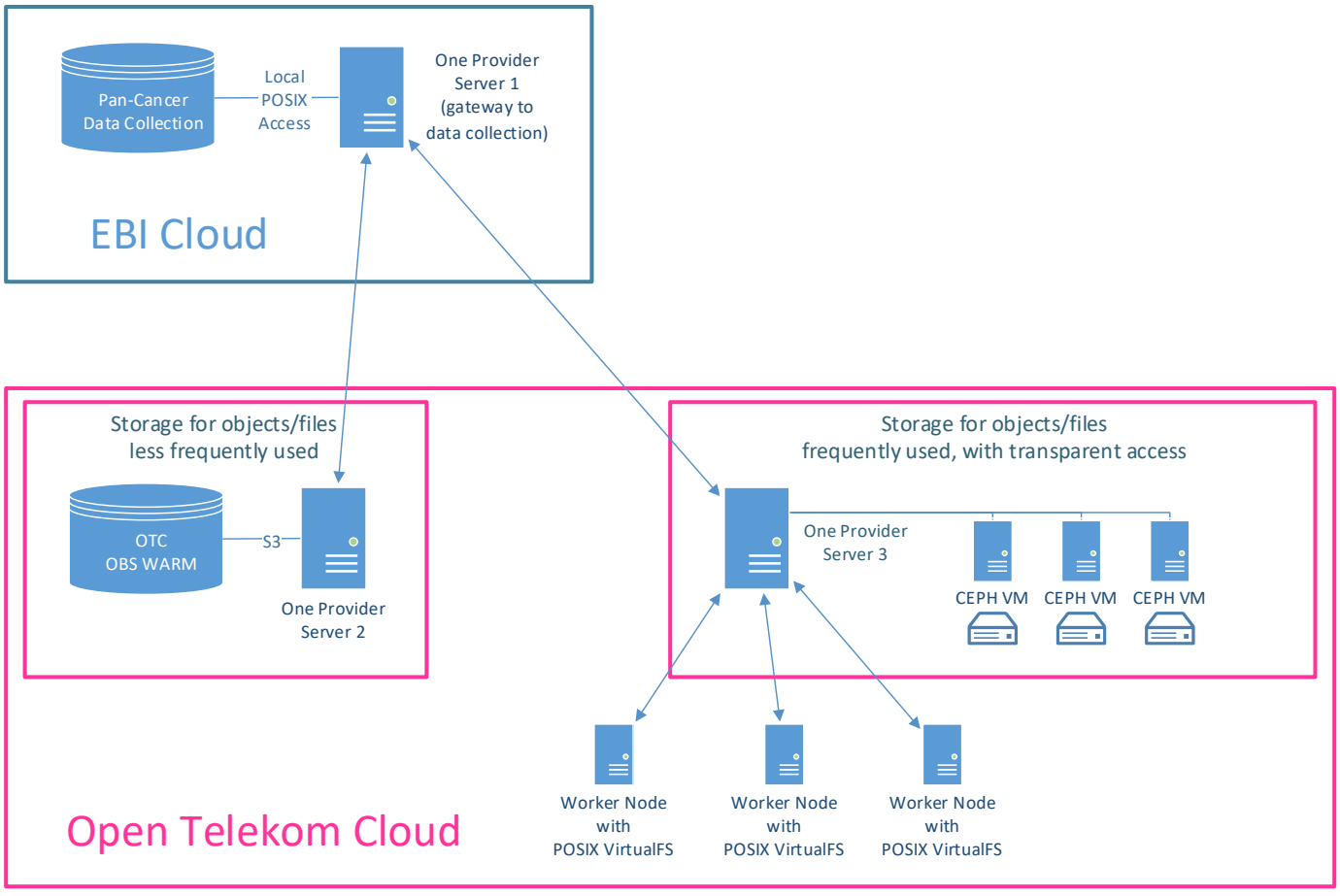

Figure 1: Pan-Cancer Data Management scenario

A Onedata deployment increases the resources required for data management, but this is offset against the lower TCO for object storage. Since the Onezone service for IAM will be available as integral OTC service, only the resources for the Oneprovider service are reflected in the TCO.

The compute resources are proposed to be based on the OTC S2 flavour-type with Linux OS, that can provide the required vCPU/RAM ratios and have the lowest costs, both per hour as well as per month in the reserved model. The lowest TCO can be obtained through a mix of pay-per-use and reserved pricing models, that has been applied.

Regarding the network requirement, a 10 Gbps network capacity is assumed, to provide sufficient reserve capacity to start a burst pattern quickly, thereby minimizing the risk that initial resources might be used longer than needed.

Intangible values

For organisations that need to support a wide range of communities with diverse workloads e.g. EMBL, an important intangible value will be how they can effectively manage the resources required without “over-spending” budget. Avoiding over-spending could become more difficult if the overall aggregated workload is quite dynamic as depicted in Figure 2.

The intangible value of the hybrid cloud solution would be, that the organisation can plan its on-premise resource requirement in a much more predictable way. The risk of having idle resources running on-premise could be significantly reduced e.g., by moving the assumed red dotted line down to the straight red line in Figure 2. The red line could be determined by comparing the TCO of the more static on-premise resources and the TCO of the dynamic off-premise resources supported by a public cloud. The most-economic overall TCO would determine the optimal position of the red line and the future level of procurement for each resource type.

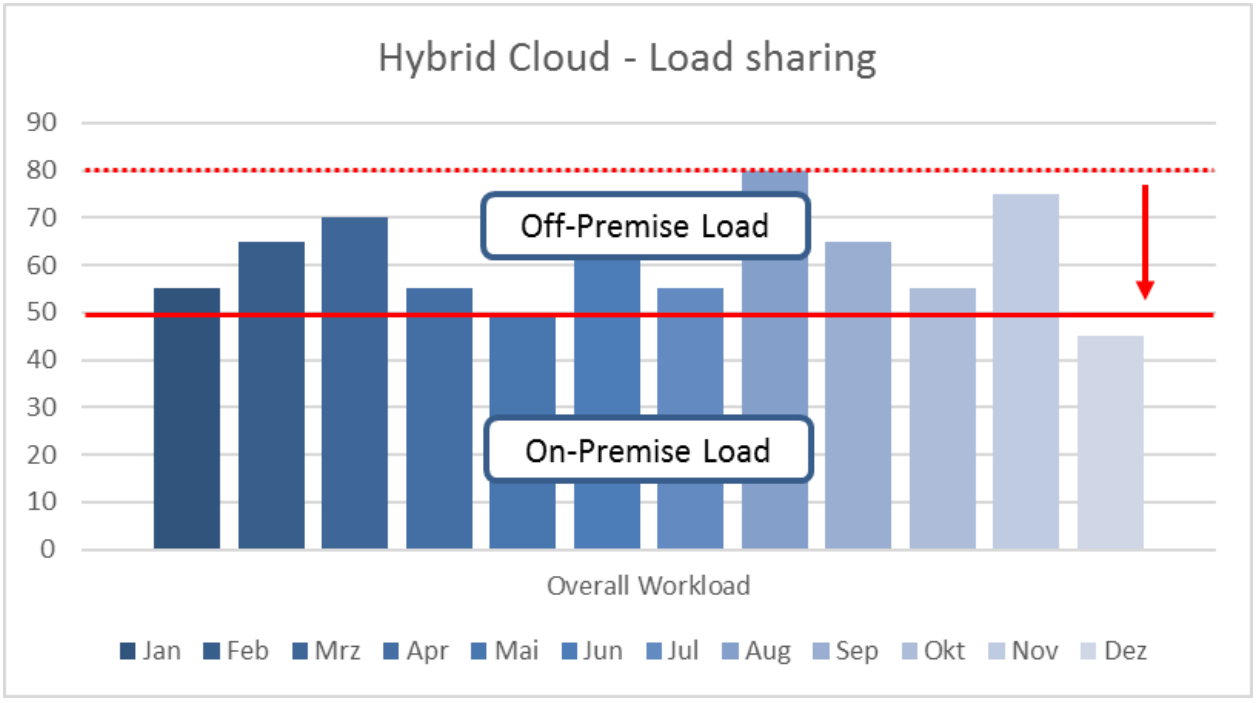

Figure 2: Hybrid Cloud Load sharing

Microbiology and Bioinformatics are also characterized by continuous and recent breakthroughs in analytics, detector and imaging technologies. E.g. Live-cell Lattice light-sheet microscopy (LLSM) together with PALM, STORM and PAINT technologies seem to be such a breakthrough, that have recently created significant attention amongst science teams. An example is the recent publication in Science (see Figure 3). Usually such breakthrough technologies take significant time to become available in a wider community, since the instrument costs may be very high and the analytics complex and resource consuming.

A hybrid cloud model could make the new innovative analytics available more easily and quicker by setting them up in the public (dynamic) part of the hybrid cloud, whereby the on-premise part would be connected to the (expensive) experiment setup.

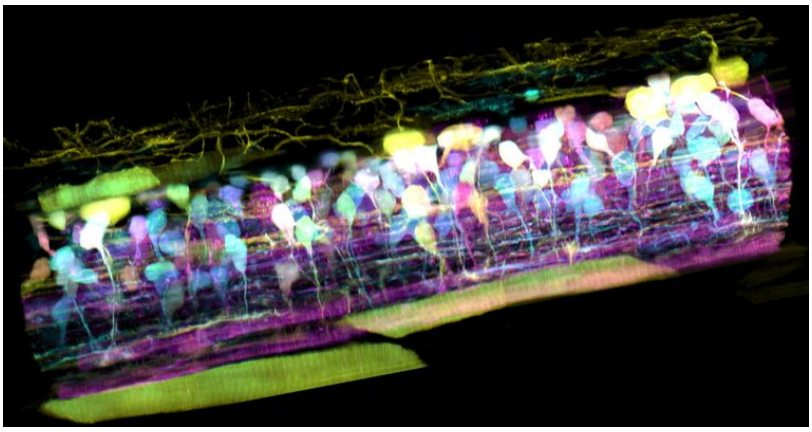

Figure 3: Intangible value example: Neurons lighting up in the bone mark of a zebrafish embryo - obtained with LLSM (published in Science 2018)

Another characteristic of data analytics in Microbiology and Bioinformatics is the complex data privacy compliance and regulations that must be considered. What is becoming more relevant for more and better science, is that data can be made more quickly available when the appropriate data privacy has been clarified. In a hybrid cloud, especially when supported by Global Namespace solutions it would be possible to almost instantaneously implement or change the required compliant classification of data and datasets. As soon as data classification can be changed e.g., from initially confidential to restricted, or from restricted to public and vice versa, the access to the data can be changed in the system accordingly, without hassle and without moving the physical location(s). It would enable science organisations to make much more data available faster to restricted users group or the public for further analysis.

3.1.2 Quantitative factors

The quantitative factors were analysed for 12- and 36-month contract terms and alternative compute flavors. The longer-term commitment only provided a slight TCO improvement, since only few compute resources are required to support baseline, most are required as pay-per-use. All services can be provided for a TCO in the order of 400.000 Euro per year.

3.1.3 Qualitative factors

The qualitative analysis of using the public cloud alternative was more straightforward. It has been demonstrated during the PCP project, that agility, elasticity and scalability can be achieved at a very high level and can fulfil the requirements.

The BG has been made available the Data Processing agreements and the BG confirmed that these documents and the test incidents, that were initiated confirm the required regulatory, policy and security levels. Life sciences data privacy is complex and may involve confidential, restricted and public data to be managed. OTC and T-Systems are certified to manage such data with the appropriate configurations and processes.

Service levels during the HNSciCloud project have been managed based on the OTC Enterprise Agreement. In the agreement silver, gold or platinum service levels can be agreed. The baseline for the TCO was the gold level.

Table 3: Pan-Cancer qualitative factors

Factor	Importance (H, M, L)	Effectiveness	Notes	Additional Information
Agility	H	+1	The solution provides easy and quick deployment and changes to the processing and storage functionality	
Contract Review and Negotiation	M	+1	Can be implemented with minimum review	
Elasticity and Scalability	H	+1	Solution facilitates easy and quick expansion of available processing and/or storage capacity	
Regulatory and Policy Requirements	H	+1	Solution adequately enables compliance with external regulations	User data can be stored encrypted in the public cloud. Through Onedata, data privacy can be maintained between public, restricted and confidential data. Confidential data would only be hosted on-premise.
Security	H	+1	Solution provides effective mechanisms by which constantly escalating security threats are prevented and security events or breaches are constantly monitored	Cloud service is certified e.g. for relevant ISO standards, EU GDPR and German national regulations.

Factor	Importance (H, M, L)	Effectiveness	Notes	Additional Information
Service Levels	M	+1	The solution would be able to match or improve the current SLA based on on-premise services. Especially with regards to performance, the cloud offers more granularity and diversity to tune performance to requirements when compared to on-premise resources.	Service Availability of 99,95% is achievable when using 2 availability zones and load balancing. Support is provided 24/7.

3.1.4 Recommendations and suggestions

The comparison with on-premise costs indicated that these are currently lower than those of a commercial cloud service. Compute costs are on a par, but on-premise storage costs seem to be lower by a fair amount. However, on-premise costs may not have been transparent and known for all service elements and the TCO for the commercial services were not based on a best-and-final-offer, that could be offered as part of a public procurement.

Therefore, an assessment under which circumstances the commercial cloud services might be even or more advantageous than on-premise costs for the Pan-Cancer or similar use cases will need to be performed on a case-by-case basis following the approach described and considering a procurement competition might result in further benefits for the commercial cloud services.

3.2 ALICE (HEP)

3.2.1 Tangible and intangible values

The ALICE use case is significant in size with an overall and continuous requirement of 50,000 continuous concurrent jobs to be supported by 50,000 cores (vCPUs). A key factor to determine the price per job is to analyse how many jobs could be completed over a fixed time e.g., a year.

Tangible values

As has been thoroughly studied by High-Energy Physics scientist the wall-clock time for a job depends most on the CPU performance. For typical HEP job performance, the best correlation is obtained with subsets of the SPECfp- and SPECint-RATE benchmarks, the SPEED benchmarks are less relevant.

Since it was not possible in the context of the TCO study to perform a representative benchmarking and review, T-Systems have used the published results of the CPU2017 and CPU2006 benchmarks for the relevant OTC CPU-types, that have been considered to determine the best price/performance on job level. Since the CPU2017 benchmarks were not available for older CPU-types, we have used the CPU2006 RATE benchmarks to compare the older and newer CPU performances. And when comparing flavour CPU-performance, the allocation of vCPU to physical cores, hyperthreading and over-commitment strategies should be considered. The relative performance per vCPU (core) are listed in Table 4.

Table 4: OTC CPU relative performance ratios

OTC flavour	CPU	Cores/ Chips/ Threads	CPU2006 RATE relative performance per core
S1/C1/C2/M1[1]	Xeon E5-2658A v3	24/2/2	1.00
S1/C1/C2/M1[1]	Xeon E5-2658A v4	28/2/2	1.03
S2[1]	Xeon 6161	44/2/2	1.08
C3[2]	Xeon 6151	36/2/2	1.37
M2[2]	Xeon E5-2690 v4	28/2/2	1.17
M3[2]	Xeon 6151	36/2/2	1.37
HP I[2]	Xeon E5-2690 v3	24/2/2	1.11
HP II[2]	Xeon E5-2667 v4	16/2/2	1.45

In general, cloud pricing is more attractive for smaller machines and flavours that include OC. It enables cloud providers to distribute workloads of thousands of users more efficiently over an infrastructure. Following, the best compute option for the ALICE use case was to use small S2-flavours for processing. The storage requirement is limited and does not have a significant impact on the TCO.

The network requirement is mainly determined by the Analysis Train jobs with a bandwidth of ca. 50 Mbps per job. Assuming on average the 50,000 jobs would be a mix of 40% monte carlo, 40% raw data and 20% analysis, the overall network requirement would be in the order of magnitude

[1] Flavour including over-commitment

[2] Flavour without over-commitment, each vCPU allocated to a physical core

of 700 Gbps. Therefore, the study determined a potential implementation and included cost for a full-diverse 700 Gbps network capacity.

Intangible values

Intangible values can be identified when studying the WLCG Strategy towards the future High-Luminosity Experiments at CERN (HL-LHC). Two of the key challenges identified for HL-LHC in 2026-2027 are better management of operations costs and optimisation of hardware costs. Although the promise of the potentials identified still must be demonstrated, intangible values of the proposed solution in this respect can already be identified.

The proposed solution follows the data streaming to remote processors approach, providing CERN with an early opportunity to learn and optimise implementation of the future compute and data management model. At the same time, it will enable CERN to reduce compute power at the Geneva premise, freeing up space and power capacity for the envisaged storage consolidation. These are first intangible values identified.

3.2.2 Quantitative factors

The TCO study identified that a significant commercial benefit can be achieved with longer-term commitments, due to the very stable baseline requirement over time. To simplify a comparison of ALICE job TCO costs on OTC with the job costs at CERN using on-premise resources, Table 5 highlights green colour the Monte Carlo and Reconstructions jobs that can be run more economical with the commercial cloud service. The Analysis jobs (in red) currently can be run more economical using on-premise resources due to high network costs.

Table 5: ALICE job costs compared between CERN cloud and OTC public cloud

Job Type	TCO Benefit OTC Public Cloud
Monte Carlo	
Reconstruction	
Analysis	

3.2.3 Qualitative factors

The qualitative analysis of using the public cloud alternative was more straightforward. It has been demonstrated during the project, that agility, elasticity and scalability can be achieved at a very high level and can fulfil the dynamic job management requirements.

The BG has been made available the Data Processing agreements and the BG confirmed that these documents and the test incidents, that were initiated confirm the required regulatory, policy and security levels.

Service levels during the HNSciCloud project have been managed based on the OTC Enterprise Agreement. In the agreement silver, gold or platinum service levels can be agreed. The baseline for the TCO was the gold level.

Table 6: ALICE qualitative factors

Factor	Importance (H, M, L)	Effectiveness	Notes	Additional Information
Agility	L	+1	The solution provides easy and quick deployment and changes to the processing and storage functionality	
Contract Review and Negotiation	M	+1	Can be implemented with minimum review	
Elasticity and Scalability	L	+1	Solution facilitates easy and quick expansion of available processing and/or storage capacity	
Regulatory and Policy Requirements	L	+1	Solution adequately enables compliance with external regulations	User data can be stored encrypted
Security	H	+1	Solution provides effective mechanisms by which constantly escalating security threats are prevented and security events or breaches are constantly monitored	Cloud service is certified e.g. for relevant ISO standards, EU GDPR and German national regulations.
Service Levels	H	+1	The solution would be able to match or improve the current SLA based on on-premise services. Especially with regards to performance, the cloud offers more granularity and diversity to tune performance to requirements when compared to on-premise resources.	Service Availability of 99,95% is achievable when using 2 availability zones and load balancing. Support is provided 24/7.

3.2.4 Recommendations and suggestions

The impact of flavour, VM size and Storage I/O on TCO can be significant. Users are well advised to perform benchmarking of real-world scenarios to determine which cloud service implementation will perform best and leads to the lowest TCO.

A further alternative that may be considered is to extend the commercial cloud services with a local on-premise component to reduce the impact of variable and network costs. T-Systems has recently introduced a version of its Open Telekom Cloud – Hybrid OTC, that includes an On-Premise or local element installed on customer premise or a data centre with good network connectivity in the direct vicinity of the customer. The local element has identical services, functions, interfaces and APIs as the public OTC. The local element is connected to the public OTC to enable various use case scenarios e.g. burst, fail-over and backup. However, the local element can continue to work completely independent of the connectivity to the public OTC – as stand-alone facility.

4 MISCELLANEOUS

All price and cost information provided in the context of the TCO study were for planning purposes and non-binding. All prices and costs were provided excl. VAT and in Euro.