Phase 3

D-PIL-3.13 Total Cost of Ownership Study

Doc.:	HN-D-PIL-3.13-3001
Issue:	1.4
Date:	2018-12-14

DOCUMENT CHANGE CONTROL

Issue	Date	Chapter	Changes
1.0_DRAFT	2018-07-09	All	DRAFT
1.0	2018-09-05	All	First release
1.1	2018-09-24	4.2, 4.3, 4.4, 5.4, 6	Update after M-PIL-3.3 feedback
1.2	2018-11-09		Update for M-PIL-3.4 after additional feedback
1.3	2018-11-26	4	Update to correct errors in networking and ALICE per job costs, revised CloudFerro prices to those on their website
1.4	2018-12-14	5.4	Added paragraphs on price sensitivity to choice of VM

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Purpose

The purpose of this document is to provide an estimate of the Total Cost of Ownership (**TCO**) of two Buyers' Group use cases; PANCANCER and ALICE.

In this document we describe our approach, analyse each of the use cases and calculate the TCO for those use-cases (and the different jobs types in each).

We also consider the following public cloud services:

- Exoscale
- Amazon Web Services
- Cloudferro (www.cloudferro.pl, based in Poland, supporting one of the Copernicus Data and Information Access Systems (**DIAS**)

In this version (1.1), we provide updates following feedback and further questions form the Buyers' Group.

1.2. Scope

The Buyers' Group provided the ECAR profile as a reference model but much of that is concerned with Buyer's Group internal costs, so we have focussed what is relevant to the cloud providers.

In this version of the document we have provided an estimate of the PANCANCER and ALICE costs for each of the commercial clouds, making use of the best Virtual Machine (VM) flavours and pricing offered by each.

Together with this document, we also provide the following:

1. A spreadsheet model which summarises each use case and the resource needs. We have selected the VM flavours that we think best meet those needs but the Buyer's Group is free to select other flavours to look at the cost impact. Likewise, the Buyer's Group is free to add another cloud to its model
2. The ECAR model. We have completed some of the *qualitative* and *quantitative* fields and provided the other information that we can, with justifications.

2. REFERENCES

The following documents are referenced in this document.

Document ID	Document Title	Issue
RD-1	TCO for Cloud Services	ECAR Working Group Paper, EDUCASUE, April 24, 2015

3. TERMS, DEFINITIONS, AND ABBREVIATIONS

ECAR	EDUCAUSE Center for Analysis and Research
GPU	Graphical Processing Unit
RAM	Random Access Memory
SSD	Solid State Disk
TBC	To be Confirmed
TCO	Total Cost of Ownership
VM	Virtual Machine

4. APPROACH

4.1. ECAR

Comparing the costs of cloud solutions vs. on-premises solutions is complex and challenging to compare like-with-like.

The ECAR working group has created the Total Cost of Ownership (**TCO**) framework to provide a tool to help make such assessments in a transparent and effective way. It recognises that there are both *Tangible* and *Intangible* costs, for example "The intangible benefits of the cloud (e.g. geographic diversity of services, fault tolerance, enhanced security and compliance, and automation) can make it more attractive."

The TCO Framework addresses three main areas:

- **Foundational Risks**: These drive many of the considerations made in the TCO
  - Data Sensitivity – how securely must the data be held and protected?
  - Business Criticality – how critical is the functionality to the business of the organisation/project?
- **Quantitative Factors**: These are measurable costs that can be readily identified
  - On-Going Costs
  - One-Time Costs
  - Hidden Costs and Subsidies (on-going, one-time)
- **Qualitative Factors**: These are factors that are hard to quantify in terms of Euros but can represent significant advantages or disadvantages for a solution.

While ECAR is a good framework for comparison, what we can complete as cloud service providers is limited. Therefore, in this report we use the TCO framework to present the costs of the cloud-based solution as far as is possible but we do not have access to the information from the Buyers' Group organisations concerning costs of PANCANCER and ALICE.

4.2. Baseline Cloud Solutions

4.2.1. Overview

In the remainder of Phase 3 RHEA has been asked to offer a solution based on Exoscale alone. However, we wanted to benchmark the Exoscale prices compared to other cloud providers: Amazon Web Services because it is widely used and well known; and CloudFerro a cloud provider based in Poland supporting one of the Copernicus DIAS which provides data storage for Copernicus data (eventually 10's of PB) and access to processing of that data in their cloud. CloudFerro also have GÉANT connection. Other cloud providers such as OTC, OVH, Google etc. could have also been considered. The model provided is extensible to support these other clouds.

Based on the characteristics of each use-case we have selected a VM and purchase options for each provider that we consider would best suit the use-case, taking into account that each of the use-cases can run an instance in each core of a VM. Nevertheless, in the Excel model provided with this report, you can try different flavours to see what the cost impact could be.

In the following sections we briefly describe each cloud and the capabilities offered. We conclude the section with the Nuvla brokerage and management solution.

4.2.2. Exoscale Cloud

The Exoscale cloud offers a number of cloud VM flavours, two of which support GPU. Each supports a certain number of cores, RAM and attached SSD storage. These are summarised in the table below.

Table 1: Exoscale Compute Flavours

Type	VM Flavour Name	Cores	RAM (GB)	GPU	Storage (GB)	Default VM Disk Type
Standard Instances	Micro	1	0.5	0	10	SSD
	Tiny	1	1	0	10	SSD
	Small	2	2	0	50	SSD
	Medium	2	4	0	100	SSD
	Large 200GB	4	8	0	200	SSD
	Large 100GB	4	8	0	100	SSD
	Large 50GB	4	8	0	50	SSD
	Large 10 GB	4	8	0	10	SSD
	Extra-Large	4	16	0	200	SSD
	Huge	8	32	0	200	SSD
Larger Instances	Mega	12	64	0	50	SSD
	Titan	16	128	0	50	SSD
GPU	GPU - Small	12	56	1	100	SSD
	GPU - Huge	48	225	4	100	SSD

Note: There is a maximum of one hyper-threading core per vCPU. This means that ExoScale does not over commits the vCPUs, since each vCPU corresponds to a core.

Storage

Each VM can access the local SSD storage attached to it or Object Storage.

Hard-disk based block storage is planned for 2019, but this is likely to be limited to 10 TB per VM. Therefore, one PB of storage could require up to 100 VMs just to provide the required storage, which may not be cost-effective. Expected pricing is 0.10 CHF/GB/Month (approximately 5 times more expensive than block storage).

Networking

Exoscale offer two network connection options to the Buyers' Group:

- **GEANT**: The ports are shared amongst all the traffic coming from the GEANT Cloud VRF so the bandwidth is shared using GEANT policies if any. There is currently no plan in putting policing in place there.
- **Private Connect**: the ports are not shared and dedicated to all the private networks of an organization(s) aka a buyer's tenant(s) or subtenant(s)

During Phase 3 up to 20 Gbps was provided between GÉANT and each of the Frankfurt and Exoscale data centres (40 Gbps aggregate). The quoted prices include GÉANT cloud VRF connection which costs: 1000 €/month per 10G port.

An alternative 10Gbps private connection to Exoscale, not via GEANT, where bandwidth can be guaranteed would cost 2500 CHF for installation and then 500 CHF/month recurring fee. 1 port minimum, 2 ports preferred for redundancy. Available in Geneva and Frankfurt.

In addition, a short cable/fiber within the datacentre where the connection happens is required. It is provided by the datacentre supplier but needs to be paid for *each connection* and costs 500CHF for set-up and 120CHF monthly. Thus the total costs per 10G private connection are 3000 CHF set-up and 620 CHF monthly. Therefore, for a 50G private link it would cost 15000 CHF for set-up and then 3600CHF monthly.

Purchase Options

We provide Exoscale pricing for pay-as-you-go and various options. The relative pricing is shown in the table below.

Table 2: Exoscale Discount Scheme Used

PAYG	Reserved (1 year)	Reserved 100k	Over 1 M	Pre-emptive
100%	80%	70%	58%	50%

The baseline pricing is for PAYG. Discounts are available for reserved instances for 1 year (there are no discounts for 3- or 6-months), for reserved instance values over 100K/yr and for business over 1M/yr. Exoscale do not offer 3-year pricing. All pricing is **without** upfront payment, but reserved instances require a commitment of one-year, so the buyer has to pay whether they use it or not.

From mid-2019 Exoscale plans to launch pre-emptible instances, which would have the following criteria:

- **Maximum lifetime**: 24h is the maximum. There are no plans to increase this for pre-emptible instances.
- **Pre-emption Policy**: Can be taken down pre-emptively by Exoscale (but cost refunded). Soft shutdown initiated, then hard shutdown if the instance not stopped within 10 minutes. If the VM is shut-down by Exoscale, then the cost of the usage until that point is refunded.
- **Price**: Cost at 50% of list price. Applicable to all flavours but not available for GPU.

The pre-emptible instances make use of spare capacity in the Exoscale data centres. This can vary according to the time of day, day of the week etc. but could be very attractive for lower priority tasks that can be run whenever capacity is available for significantly lower costs than the normal pay-as-you-go or reserved instances.

Nuvla is planned to be upgraded to support this mode of purchase, including deployment and monitoring.

4.2.3. Amazon Web Services Cloud

Amazon Web Services (**AWS**) cloud offers a number of cloud VM flavours. As for Exoscale, each supports a certain number of cores, RAM and attached storage. We have extracted from the AWS catalogue the two required flavours for PANCANCER and ALICE use-cases in the table below.

The available Excel model provided with this report allows the reader to experiment and try different flavours to see what the cost impact could be. The AWS service catalogue is vast. Therefore, the following table only reproduces the flavours used in this study (the Buyers Group can simply extend this list themselves if desired).

Table 3: AWS Compute Flavours Required for Use-Cases

Type	VM Flavour Name	Cores	RAM (GB)	GPU	Storage (GB)	Default VM Disk Type
Convertible In-stances	t3.medium	2	4	0	50	Standard
	m4.2xlarge	8	32	0	50	Standard

Storage

Each VM can access the local SSD storage attached to it or Object Storage. AWS supports a range of storage solutions. However, in order to keep the comparison relevant, we only report here Exoscale comparable storage solutions. While AWS supports a managed distributed file system service, no tests were performed, such that we are **not** able to assess its ability to meet some of the use-cases' requirements.

Networking

As for Exoscale, we assume that the AWS service, in Frankfurt in this case, is connected to the GÉANT network, such that no extra connectivity costs are passed on to the buyers.

Purchase Options

We provide pricing for pay-as-you-go and reserved instances, taking 3 years as the commitment for the latter corresponding to 45% discount.

4.2.4. CloudFerro Cloud

The CloudFerro cloud offers a number of cloud VM flavours. As for Exoscale, each supports a certain number of cores and RAM. We have extracted from the CloudFerro catalogue the two required flavours for PANCANCER and ALICE use-cases in the table below.

Table 4: CloudFerro Compute Flavours Required for Use-Cases

Type	VM Flavour Name	Cores	RAM (GB)	GPU	Storage (GB)	Default VM Disk Type
Standard Instances	VM1	2	4	0	50	SSD
	VM2	8	32	0	50	SSD

Storage

Each VM can access the local SSD storage attached to it. CloudFerro does not offer object storage at this time. In order to keep the comparison relevant, we only use here local storage, which is in-line with the assumptions we made to fulfil the use-cases. However, Cloud Ferro have provided pricing for a hybrid local storage for VMs with less SSD and block storage.

Networking

The quoted price from CloudFerro service, in Poland in this case, includes a GÉANT connection, with 4 Gbps for PANCANCER and 20 Gbps for ALICE.

Purchase options

CloudFerro offers aggressive discounts for reserved instance commitments – e.g. up to ~25% 1-year, ~55% 3-year commitments.

4.2.5. Nuvla Brokerages and Management Service

The multi-cloud and hybrid-cloud management and brokerage service enables customers to monitor all cloud resources from a single point, while supporting a number of automated workflow, including deployment.

The brokerage and management fee (8% - 15% based on volume) of the Nuvla service is absorbed by the discounts compared to public pricing. The benefits of the Nuvla service include the following for multi-cloud and hybrid-cloud:

1. Frictionless cloud provider switch
2. Federated identity authentication integration
3. Service catalogue for pricing prediction
4. Dashboard and standard API (based on DMTF CIMI)
5. Rich and flexible usage metering
6. Quota monitoring

These benefits are bundled in the service fees. They flexible benefits are all optional and are expected to be used differently by the user communities and their projects.

4.3. Buyers Group Use Cases

The following subsections summarise each of the use-cases for PANCANCER and ALICE.

4.3.1. PANCANCER Use Case

During a period of 12 months PANCANCER will have a burst pattern, meaning that a minimal set of resources will be in constant use, and at periods of incoming data it will ramp up resource usage to a maximum. This means that two job types are in use: a small number of continuous VMs; and a larger number of VMs (up to 400) used in a burst mode a few times per month.

Storage Requirements:

To support 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.

Compute Requirements:

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

Network Requirements:

- 1.25 Gbps from EBI to cloud provider

4.3.1.1. Continuous Deployment

- Compute: 7 VMs, 50 CPUs, 50GB RAM
- Storage: 0.5PB, not accessed frequently (see below that we assume not to store data in the cloud)
- Network: Minimal
- Purchase option: Reserved

4.3.1.2. Burst

A burst will typically be 24-48 hours' duration, 1-2x / month (this equates to ~13% usage per annum, assuming 48 hours):

- Compute: ~250 VMs, 2000 CPUs, 8.5TB RAM
- Storage: 0.5PB, random access with high I/O requirement
- Network: 10Gbps intra-node traffic, up to 4Gbps ingress

4.3.1.3. PANCANCER Cost Analysis

We assess the best VM type(s) to use for this use-case and the storage strategies that could be employed.

Recommended VM:

Based on the characterisation of this the PANCANCER jobs, we recommend the Exoscale Huge VM with 8 cores, 32GB RAM and 50GB SSD disk. This is the smallest VM with sufficient cores. This machine type has more memory than required. The same machine type is used for continuous and bust deployments.

Storage:

Several strategies could be adopted for this use case:

1. Copy ~1PB of data to the cloud. This is time consuming, potentially very expensive if block storage has to be used and, as we understand it, each job only needs to access a small subset of the data. As we noted previously, block storage support at Exoscale will not be available until 2019 and is also likely to be limited to 10 TB per VM, requiring 100 VMs.
2. Copy files (or pre-stage) the files that are needed by each VM (e.g. using Onedata). With a 4 Gbps available bandwidth to the experiment, the largest 30GB file would take approximately one minute to copy over, the smaller 500MB files, approximately one second. If this is acceptable, storage costs would be zero (to be confirmed).
3. Use Onedata or other libraries to provide an abstraction of POSIX on S3 storage.
4. If the cost savings/performance are attractive enough, consider updating the application in the long-term to support object storage directly (if this is feasible).

We have made the following assumptions for the PANCANCER cost analysis:

- Storage costs are zero (option 2 above is acceptable) – OneData is not necessarily needed. If used, costs to provide onedata services on the cloud would be negligible. A limited number (2-5) of instances of onezone and oneprovider will be sufficient for that purpose.
- The continuous deployment has 7 VMs and runs 100% of the time.
- Since these resources are permanently deployed, we assume a reserved instance price (1-year commitment (Exoscale don't offer a 3-year price and on its own, PANCANCER does not qualify for discounts based on business over €100K for reserved instances)
- The bursts are run two times a month, each for 48 hours i.e. 96 hours per month or ~13% of the year, with 250 VMs on average in this period.
- Since the burst usage is sparse in time, pay-as-you-go pricing is considered.
- The time to upload of files to the VMs is not significant compared to the processing time
- Pricing includes €12,000 per annum for 40 Gbps for GÉANT access (clearly it can be shared between several projects or Buyers' Group members.
- **All prices exclude VAT**.

Following feedback from the Buyers Group we have added an additional use case where storage costs are included. However, there are a few caveats:

- Pricing is not yet available for Exoscale's block storage solution planned for 2019, so we provide a price based on Object Storage, which in any case is almost certainly much cheaper. The needed storage is 12 months x 0.5 PB x (1024)^2.
- As noted previously, our understanding PANCANCER cannot directly use object storage and needs an POSIX interface. We therefore assume this is provided by OneData on top-of the object storage.
- We only provide the prices for Exoscale since those for AWS and CloudFerro are not available in the time available for this update.

Based on these assumptions, the costs for PANCANCER are shown in the tables below for the non-storage and storage solutions. The costs are shown for **three years** (but for Exoscale pricing is based on a one year reserved instance pricing over three years).

Table 5: PANCANCER Summary Costs for Exoscale (No Storage)

Application Name	Job Type	Annual Utilisation (%)	Utilisation (VM-Hours)	Bulk Storage Type	Annual Bulk Storage Volume (GB months)	VM Flavour	Total Cost (€)
PANCANCER	Continuous	100%	61,362	None	NA	Huge	16,189
	Burst	13%	288,197	None	NA	Huge	94,990
	Networking						12,000
Total							**123,170**

Table 6: PANCANCER Summary Costs for AWS (No Storage)

Application Name	Job Type	Annual Utilisation (%)	Utilisation (VM-Hours)	Bulk Storage Type	Annual Bulk Storage Volume (GB months)	VM Flavour	Total Cost (€)
PANCANCER	Continuous	100%	61,362	None	NA	m5.2xlarge	17,854
	Burst	13%	288,197	None	NA	m5.2xlarge	129,721
	Networking*						12,000
Total							**147,575**

*Estimated – no figures available for GÉANT connection costs from AWS

Table 7: PANCANCER Summary Costs for CloudFerro (No Storage)

Application Name	Job Type	Annual Utilisation (%)	Utilisation (VM-Hours)	Bulk Storage Type	Annual Bulk Storage Volume (GB months)	VM Flavour	Total Cost (€)
PANCANCER	Continuous	100%	61,362	None	NA	VM2	19,851
	Burst	13%	288,197	None	NA	VM2	135,395
	Networking*						
Total							**155,247**

*CloudFerro 4 Gbps networking costs included in the figures

Table 8: PANCANCER Summary Costs for Exoscale (Object Storage)

Application Name	Job Type	Annual Utilisation (%)	Utilisation (VM-Hours)	Bulk Storage Type	Annual Bulk Storage Costs	VM Flavour	Total Cost (€)
PANCANCER	Continuous	100%		Object	53’477	Huge	69,657
	Burst	13%		Object	62’915	Huge	157,904
	Networking						12,000
Total							**239,562**

4.3.1.4. PANCANCER Summary

For each cloud provider we have selected three-year reserved instances for the continuous job type and PAYG for the burst job type. For the no-storage use case, CloudFerro is by a long way the cheapest option. Exoscale could be substantially cheaper if the pre-emptible option was chosen. No pricing is readily available for pre-emptible VMs from AWS and CloudFerro, but could also be attractive.

Adding object storage for Exoscale, increases the annual costs by €116K, almost doubling the overall cost.

4.3.2. ALICE Use Case

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

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

Each job uses a *single core*; 50,000 jobs can be run at any one time (i.e. 50,000 cores running concurrently). The Buyer's Group interest is to understand the cost to execute each job and the cost of running 50,000 jobs continuously over one year

Each type of job is described below.

4.3.2.1. 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

=> 4GB/core disk space

Network Requirements:

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

Note: Although the network bandwidth of an individual job is quite low, with many thousands of jobs running, the network bandwidth between the cloud and the institute could become significant. For example, assuming that this use-case takes 70% of the 50,000 jobs (35,000 jobs) and uses a constant and uniform bandwidth usage, the aggregate bandwidth would be 10.5Gbps.

Recommended VM:

Based on the characterisation of this the ALICE MC jobs, we recommend the Exoscale Medium VM with 2 cores, 4GB RAM and 50GB SSD disk. This is the smallest VM with sufficient RAM per core. Machines with greater number of cores cost more per core than the smaller machines, so there is no benefit to use machines with 8, 12 or 16 cores. Two jobs can be run concurrently per VM.

4.3.2.2. 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: 5 GB/core

=> 7GB/core disk space

Network Requirements:

- Input download: ~2 Gbps
- Output upload: ~0.7 Gbps
- Network bandwidth: ~1 Mbps

Note: Assuming it uses 20% of the jobs, the aggregated network bandwidth would be ~10 Gbps, which could be the limiting factor on the number of jobs running concurrently.

Recommended VM:

Like the ALICE MC jobs, we recommend the Exoscale Medium VM with 2 cores, 4GB RAM and 50GB SSD disk. Two jobs can be run concurrently per VM.

4.3.2.3. 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 (*an average of 2 is fair and assumed here*)
- RAM: at least 2 GB/core
- Swap: at least 6 GB/core in addition
- Scratch disk space: at least 10 GB/core

=> 16GB/core disk space

Network Requirements:

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

Note: The efficiency of this workload also depends on the network latency. If the efficiency is reasonable, such jobs can be run without cloud storage, which CERN will try with the current pilot platform setups. However, in this case, the aggregated network requirement would be 244 Gbps (aka 50 Mbps x 50000 jobs x 10%), which would be a challenge, even by distributing the workload to a range of cloud providers. If it could be provided, the cost would also be very high at current prices.

Recommended VM:

Like the ALICE MC jobs, we recommend the Exoscale Medium VM with 2 cores, 4GB RAM and to allow 2 jobs to be run simultaneously it will need 32GB SSD disk, so will need the flavour with 50GB. Comparable VMs have been chosen for AWS and CloudFerro.

4.3.2.4. ALICE Cost Analysis

The ALICE use-case mainly requires compute and can have large numbers of jobs running at any one time. Although the network bandwidth of the MC and RDR jobs is quite low, the network bandwidth will be a limiting factor on the actual number of jobs being run simultaneously. The Analysis Train use case could limit this to a few 100s of jobs – e.g. with a dedicated 40 Gbps link, with a requirement of constant streaming at 50Mbps per job, only 800 concurrent jobs could be run.

Network usage for ALICE MC and RDR will not be uniform over the job life. However, with a large number of jobs, one can assume that the job starts will eventually spread, to end-up behaving as the average suggested for these two use-cases. If this was not the case and the use-cases where limited by bandwidth, the job duration would extend.

To estimate the costs of the ALICE jobs we have made several assumptions in our model (which the Buyers' Group can of course change):

- Jobs can be run simultaneously on each VM (one in each core) – For ExoScale and AWS we provide pricing for 2-cores/VM and 8-cores/VM.
- The ALICE analysis trains average duration is 2 hours
- Following Buyers' Group feedback, the split of the use cases is 70% Monte-Carlo, 20% Analysis Trains and 10% Raw Data Reconstruction.
- A rough calculation of bandwidth requirements for ALICE MC and RDR indicates that a 23 Gbps (10.5 Gbps + 5 Gbps) would be sufficient for 80% of 50,000 concurrent jobs. However, ALICE AT, with a requirement of 50 Mbps per job and the remaining 20% of the jobs would need ~500 Gbps which is not feasible with the currently available network bandwidth.
- Since these resources are permanently deployed, we assume a reserved instance price (3-year commitment) for each of the job types.
- Pricing includes €12,000 per annum for 40 Gbps for GÉANT access (clearly it can be shared between several projects or Buyers' Group members.
- **All prices exclude VAT**.

It should be noted that to run 50,000 jobs per year would require 25,000 2-core VMs or 6,250 eight-core VMs running continuously and is about 5x the Phase 3 TP3 processing requirement.

The costs for the ALICE use cases are summarised in the table below, providing the cost per job and the cost to run 50,000 jobs per year. Note: These now include dedicated private network costs.

Table 9: ALICE Summary Costs for Exoscale (2 core VM)

Application Name	Job Type	Average Job Duration (Hrs)	Jobs per VM	VM Flavour	Job Mix	Networking (Private Connection)	Price for 50,000 Jobs running permanently (per year)	Cost Per Job
ALICE	Montecarlo	6.00	2	Medium 50GB	70.00%	9,991	4,181,196	0.082
	Raw data Reconstruction	7.00	2	Medium 50GB	20.00%	9,515	1,201,288	0.096
	Analysis Trains	2.00	2	Medium 50GB	10.00%	237,882	833,768	0.038
Total					**100%**	**257,388**	**6,216,252**	**0.0802**

Table 10: ALICE Summary Costs for Exoscale (8 core VM)

Application Name	Job Type	Average Job Duration (Hrs)	Jobs per VM	VM Flavour	Job Mix	Networking (Private Connection)	Price for 50,000 Jobs running permanently (per year)	Cost Per Job
ALICE	Montecarlo	6.00	8	Huge 200GB	70.00%	9,991	7,779,751	0.152
	Raw data Reconstruction	7.00	8	Huge 200GB	20.00%	9,515	2,229,447	0.178
	Analysis Trains	2.00	8	Huge 200GB	10.00%	237,882	1,347,847	0.062
Total					**100%**	**257,388**	**11,357,045**	**0.148**

Table 11: ALICE Summary Costs for AWS (2 core VM)

Application Name	Job Type	Average Job Duration (Hrs)	Jobs per VM	VM Flavour	Job Mix	Networking (Internet)	Price for 50,000 Jobs running permanently (per year)	Cost Per Job
ALICE	Montecarlo	6.00	2	t3.medium	70.00%	894,863	5,420,161	0.106
	Raw data Reconstruction	7.00	2	t3.medium	20.00%	438,300	1,731,242	0.138
	Analysis Trains	2.00	2	t3.medium	10.00%	109,575	756,046	0.034
Total					**100%**	**1,442,738**	**7,907,449**	**0.105**

Table 12: ALICE Summary Costs for AWS (8 core VM)

Application Name	Job Type	Average Job Duration (Hrs)	Jobs per VM	VM Flavour	Job Mix	Networking (Internet)	Price for 50,000 Jobs running permanently (per year)	Cost Per Job
ALICE	Montecarlo	6.00	8	m5.2xlarge	70.00%	894,863	8,303,605	0.162
	Raw data Reconstruction	7.00	8	m5.2xlarge	20.00%	438,300	2,555,084	0.204
	Analysis Trains	2.00	8	m5.2xlarge	10.00%	109,575	1,167,967	0.053
Total					**100%**	**1,442,738**	**12,026,655**	**0.160**

Table 13: ALICE Summary Costs for CloudFerro (2 Core VM)

Application Name	Job Type	Average Job Duration (Hrs)	Jobs per VM	VM Flavour	Job Mix	Networking (Internet)	Price for 50,000 Jobs running permanently (per year)	Cost Per Job
ALICE	Montecarlo	6.00	2	eo1.medium	70.00 %	179,995	6,380,472	0.125
	Raw data Reconstruction	7.00	2	eo1.medium	20.00 %	216,395	1,987,960	0.159
	Analysis Trains	2.00	2	eo1.medium	10.00 %	28,051	913,834	0.042
Total					**100%**	**424,441**	**9,282,265**	**0.123**

*CloudFerro 4Gbps networking costs included in the figures

4.3.2.5. ALICE Summary

ALICE costs are high using all three clouds - driven by the volume of the large number of concurrent jobs – but CloudFerro is by far the cheapest, followed by Exoscale, with AWS again the most expensive. As well as the number of jobs, the network bandwidth will become a limiting factor if the full 50,000 jobs were to be run.

The Analysis Train jobs in particular have a very high bandwidth, limiting concurrent jobs to 800, even if using 40Gbps at 100% efficiency, which is unlikely. The MC and RDR jobs have much lower network bandwidth requirements and therefore one strategy could be to offload only these job types to the cloud and retain the AT jobs in the BG infrastructure.

Using 8 cores substantially increases the price for both AWS and Exoscale (we do not have exact prices from CloudFerro but an 8-core instance is about 8x the price of a two core instance). Normally one would expect economies of scale to apply but the higher cost is mainly because larger VMs use large blocks of resource making it harder to schedule the usage of the remaining resource.

If the performance of each core is the same, regardless of the VM, and if there is a way to optimise the RAM use to 1GB per core (even if this is unlikely for ALICE) the price could be reduced by about 30%. If the RAM needed could be reduced to 0.5GB per core, the price could be reduced by ~66%. While we understand this would require significant reengineering of software, and might not be practical or even impossible, it is interesting to understand the type of savings such reengineering could yield and may be worth considering for other use cases.

4.4. Alternative Pricing Strategies

The pricing reported in sections 4.3.1 and 4.3.2 are pay-as-you-go and reserved instances, options that are straightforward to purchase. These options are generally well understood by most stakeholders.

The Continuous Deployment element of the PANCANCER use-case and the ALICE use-cases are good candidates for reserved instance. Knowing in advance that a minimum amount of resources will be required makes this commitment acceptable. Of course this comes at a cost of not being able to change provider mid-way through the committed period. In the same vein, the commitment period

will influence the level of available savings. This option allows cloud providers to predict more accurately capacity requirements on their infrastructure, therefore, the longer the commitment, the better planning they can make. Another aspect in this case could be payment plan, where upfront payment could yield additional savings, since it helps the cloud provider finance their equipment purchases.

For AWS, standard instances are approximately 10% cheaper than convertible instances we have assumed but are considerably *less* flexible should the users wish to use them for other purposes.

Further, "pre-emptible" or "headroom" offers should soon be available. As mentioned earlier (see 4.2.2), Exoscale expects in mid-2019 to sell pre-emptible instances at a 50% of the public pay-as-you-go price.

While AWS has been offering Spot Pricing for a while, the advantage of such a *pre-emptible* offer is that the pricing is more predictable and easier to integrate in a provisioning strategy.

In the case where spare capacity is made available at low prices, a trade-off is required, especially for workloads where intermediate state cannot be saved or where a job cannot be restarted midway. In this case, the user must trade off the savings made by purchasing spare capacity, versus the probability that resources might be reclaimed before the job has completed, resulting in the loss of the job. Looking at the specific example of Exoscale, other considerations are also important such as the maximum time for which a pre-emptible VM can run for, or even if a VM is terminated before the deadline, if the application workload can handle this case, even if refunds apply.

4.5. Over Commitment/Over Subscription

Cloud providers tend to over-book resources, just as airlines do. This is because many user's applications are IO intensive so CPUs are not fully utilised, but this can cause problems with CPU intensive workloads.

Typical over subscription is show in the table below for the three cloud providers. 2:1 means two vCPU are mapped to one physical hyper-threaded core or 1:1 per hyper threaded core. It can be avoided by using Dedicated servers (in data centre/on premise) or bare metal (which allows the VMs to be matched VMs exactly to the jobs and can help optimise on pricing/resource usage).

Table 14: Cloud Provider Over Subscription

Cloud	Over-Commitment (vCPU per Physical Core)	Comment
Exoscale	2:1	Worst case
AWS	2:1	From website for EC2
CloudFerro	3:1	Average, design is 4:1

5. RESULTS

5.1. Foundation Risks

Data Sensitivity

Considering the level of sensitivity of the data, the strategy common to both sets of use-cases is that no data is stored permanently in the cloud. Data is indeed transferred to the cloud, but is treated as transient. This does not eliminate issues in terms of data sensitivity, but represents a simpler problem to address, compared to having to protect permanent or semi-permanent storage of sensitive data.

The different use-cases will bring different level of required protection, from data integrity (often seen in High Energy Physics) to more advanced encryption for medical data.

Using transient data, directly streamed or transferred from the users' infrastructures to VMs running in the cloud and back, mature defence strategies exists, with well understood risk levels.

Business Criticality

The criticality of the workloads deployed in the public cloud for the organisations' business is not fully understood by the authors. However, from the moment where the capacity of the customer organisation is less than its critical processing needs, the public cloud extension considered here would correspondingly increase in criticality.

This is likely to change over time, as demand for compute, storage and networking increases, while private IT resources come to their natural life end, thus a natural decision point to either renew hardware or increase cloud usage.

5.2. Quantitative Factors

The authors have limited view over the organisations operating the use-cases. Therefore, our ability to comment on the overall quantitative factors involved in this study are limited.

On-Going Costs

The most significant costs involved in this TCO study are provided in the form of public cloud services delivering compute, storage and networking resources. These on-going costs are mainly driven by the volume of resources used, and therefore purchased, from the provider(s).

The first order on-going costs are linear. There are a few options identified to leverage economy of scale, by for example committing consumption. The user organisation controls its usage, therefore the costs, which is a significant advantage of this model.

One-Time Costs

Any change requires adaptation, which cause up-front costs. Consuming public cloud resources instead of in-house resources requires user organisations to adapt habits, processes and technologies. While Helix Nebula Science Cloud has provided significant scope for such adaptation, experience shows that each new use-case will require some level of adaptation.

As discussed in this study, up-front investment could also yield significant costs benefits, if for example, workloads running as jobs could be optimised to use less memory, or leverage cheaper storage strategies, such as object store. The return on investment of such optimisation is left for each use-case to evaluate. The authors hope that this study contributes to identifying rational and opportunities to improve workloads.

Hidden Costs and Subsidies (on-going, one-time)

The HNSciCloud project has shown that operations costs are higher than initially expected. It is important to understand to what extent these operations costs (e.g. quota management, training events, applications adaptation) are caused by project constraints, compared to constraints that would also exists in a commercial and operation context.

One important aspect that must be considered and monitored is the sensitivity of a given use-case on network bandwidth and the potential impact on the local user infrastructure and its connectivity to the cloud provider(s). Campus settings and the ability to secure adequate bandwidth is a challenge for many academic organisations. Therefore, this consideration must not be underestimated when using remote cloud resources.

5.3. Qualitative Factors

There are factors that are hard to quantify in terms of Euros but can represent significant advantages or disadvantages in using public cloud resources.

The advantages of public cloud computing are well documented. More specific to the potential user communities that the Helix Nebula Science Cloud represents are the following:

- The experience the project gained provides reference data in terms of risks and feasibility of integrating public cloud resources at the scale used during the project.
- Operations costs, complexity and impact could be assessed, including possible optimisation, or at least identification of areas where optimisation could yield benefit (e.g. reduction of memory requirements, usage of object store, usage of GPUs)
- Leveraging the experience of cloud providers and cloud experts, especially for smaller laboratories or teams with less IT or software expertise.

More qualitative factors must be available and applicable for each user organisation, but these are unknown by the authors.

5.4. Total Cost of Ownership

The quantifiable contribution to the Total Cost of Ownership of the user organisations carrying the use-cases put forward by the RHEA consortium is summarised hereafter:

Table 15: TCO Cost Summary for Each Use Case

Use-Case	Exoscale	AWS	CloudFerro
PANCANCER (no storage)	123,170	147,575	155,247
PANCANCER (with storage)	239,562	-	-
ALICE (all three use-cases) – 50000 jobs*	6,216,252	7,907,449	9,282,265
ALICE Monte Carlo – 1 job**	0.082	0.106	0.125
ALICE Raw Data Reconstruction – 1 job**	0.096	0.138	0.159
ALICE Analysis Trains – 1 job**	0.038	0.034	0.042

*Only the 2-core per VM prices are shown since the 8-core prices are double those of the 2-core VMs

**To avoid very small numbers in the ECAR spreadsheet we have provided the Exoscale prices for ALICE per job-year i.e. the equivalent of one job being run continuously for a year.

In the TCO, we have chosen certain VMs because we think they provide the best value for money while meeting the minimum requirements of the use cases. Other choices can be made. However, a general trend is that multicore VMs cost more per core than a single core VM. For example, we show that for ALICE, using an 8-core VM would cost twice as much as using a 2-core VM.

So purely from a price point of view, usage of smaller VMs could be attractive for some applications. However, this does not take into account other factors such as performance of the application if spread across lots of smaller VMs, the degree of over commitment (Exoscale say that in their cloud very large VMs are much less likely to be swapped-out than smaller VMs) and thus overall cost which is measured in terms of wall clock time per VM.

For very large jobs like ALICE, Exoscale (and other cloud providers) would provide a custom tailored VM to match the needs of that job and possibly dedicated hardware so there is no over-commitment. That would provide better performance and also save both resources and money.

The reader is invited to explore the model we have made and alter some of these assumptions using the companion Excel spreadsheet for the ALICE and PANCANCER use cases, which includes dedicated sheets for each cloud provider, use-cases and summary. The raw data for the cloud pricing that we have used is also included.

6. CONCLUSIONS

We have only been able to partially complete the ECAR framework in a very limited way because many of the factors are within the Buyers' Group domain. Also, some of the factors are relative to other solutions, we can only provide our assessment and justification, while the Buyers' Group can see the "big picture" of cloud compared to their internal and other solutions.

The PANCANCER and ALICE use-cases each have their own specific requirements for cloud deployment. Both use-cases have more than one job type requiring different VM flavours and/or pricing options. Also, use-cases can be supported without having to storage large volumes of data in the cloud, minimising storage costs. Therefore, we have not included the cost data management solutions, since our assumptions for use-cases are that the data can be "streamed" to the VMs.

We have created a simple Excel model to allow the Buyers' Group to compare the costs using different clouds for each of the use cases, including instructions on how to add another cloud to the model. Similarly, the additional use-cases can be easily added and the model extended. For the purposes of the TCO study we have chosen VM flavours and pricing options that we think provides the best price-performance but the Buyers' Group can also try-out different combinations and different cloud choices.

Price prediction is an important factor in choosing which cloud(s) to use for a give use-case. It is planned to incorporate such a model in Nuvla in the future to guide users on the selection of the best cloud to use. Comparison of like with like is not simple. Use cases are complex and there are many factors to consider in the TCO (e.g. performance and scaling, job efficiency, real world data throughputs, CPU contention and so-on). This requires close working between the science community and cloud experts to identify the best value for money solution. Benchmarking will be an important aspect if this process. The RHEA consortium is ready and available to support this.

The PANCANCER & ALICE use-cases show that important savings to be made by shopping around and choosing a cloud that best matches a specific use-case. The ALICE use case is sufficiently large that cloud providers could be willing to make much deeper discounts than we have used in the TCO study.

Both uses-cases use an element of reserved instances. These can provide significant savings for a period of one year or more. Depending on the use case, there is a balance between locking-in for three years to get the maximum saving or just one year and then shop-around annually and then deploying elsewhere if a cheaper option is found. We have chosen 3 years in this analysis to give maximum savings.

An important class of use case not explored is that where usage is peaky and the science community needs access to resources for hours, days, weeks or a few months. In these circumstances, the PAYG model would be much more attractive – simply paying for what is used. Batch scheduling systems like SLURM can be used to good effect to "burst out" to one or more commercial clouds when necessary.

The brokerage and management of the Nuvla service are included within the discounts compared to public pricing. The benefits of the Nuvla service include the following for multi-cloud and hybrid-cloud:

- Frictionless cloud provider switch
- Federated identity authentication integration
- Service catalogue for pricing prediction
- Dashboard and standard API (based on DMTF CIMI)
- Rich and flexible usage metering
- Quota monitoring
- Application deployment automation and management

- Initial voucher support
- First-line support

Exoscale will bring-in the concept of pre-emptive scheduling and CloudFerro have also told us they have an interest in offering it to use spare capacity in their system. We understand Amazon have a similar system for academic users but we have not explored this further. The pre-emptive approach could benefit use-cases like ALICE with low-priority jobs that can benefit from low-pricing by exploiting the spare capacity in a cloud. If no single cloud has sufficient spare capacity, jobs could be deployed on several clouds.

Pre-emptive (headroom or opportunistic) scheduling is much simpler to understand and use than spot pricing and allows costs to be determined in advance and can ideal for sporadic use-cases or to handle peak loads.

While our benchmarking has shown Exoscale to be very competitive compared to Amazon, they are not always the cheapest for all use cases, meaning the ability to easily access and deploy to multiple clouds should always be an option when striving to get the best value for money for the Buyers' Group and, ultimately, the European tax payer. Nuvla makes it commercially and contractually much easier to achieve this kind of flexibility.