CERN openlab

CERN openlab VI

Project Agreement

“Exploration of Google technologies for HEP”

between

The European Organization for Nuclear Research (CERN)

and

Google Switzerland GmbH (Google)

CERN K-Number	
Agreement Start Date	**01/06/2019**
Duration	**1 year**

THE EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (“CERN”), an Intergovernmental Organization having its seat at Geneva, Switzerland, duly represented by Fabiola Gianotti, Director-General,

and

Google Switzerland GmbH (Google), Brandschenkestrasse 110, 8002 Zurich, Switzerland , duly represented by [LEGAL REPRESENTATIVE]

Hereinafter each a “Party” and collectively the “Parties”,

CONSIDERING THAT:

The Parties have signed the CERN openlab VI Framework Agreement on November 1st, 2018 (“Framework Agreement”) which establishes the framework for collaboration between the Parties in CERN openlab phase VI (“openlab VI”) from 1 January 2018 until 31 December 2020 and which sets out the principles for all collaborations under CERN openlab VI;

Google Switzerland GmbH (Google) is an industrial Member of openlab VI in accordance with the Framework Agreement;

Article 3 of the Framework Agreement establishes that all collaborations in CERN openlab VI shall be established in specific Projects on a bilateral or multilateral basis and in specific agreements (each a “Project Agreement”);

The Parties wish to collaborate in the “exploration of applications of Google products and technologies to High Energy Physics ICT problems related to the collection, storage and analysis of the data coming from the Experiments” under CERN openlab VI (hereinafter “Exploration of Google technologies for HEP”);

AGREE AS FOLLOWS:

Article 1
Purpose and scope

1. This Project Agreement establishes the collaboration of the Parties in Exploration of Google technologies for HEP, hereinafter the “Project”). The detailed tasks and responsibilities of the Parties for this Project are described in the Statement of Work, contained in Annex 1 of this Project Agreement.

2. The Parties' collaboration under this Project Agreement is subject to the terms and conditions set forth in the Framework Agreement, unless otherwise specified in this Project Agreement.

Article 2
Contributions

The Parties shall make the following contributions to the Project:

1. Google Switzerland GmbH (Google) contributions:

- Deployment of technical experts to the Project;
- Secondment of or grants for recruiting dedicated researchers (at the master and doctoral level) at CERN for the execution of the projects;
- Credits towards the use of cloud technologies

Google Switzerland GmbH (Google) shall make the above contributions and issue the credit memos, as applicable, to CERN within ninety working days of receipt of the corresponding invoice from CERN.

2. CERN contributions:

- Deployment of technical experts to the Project;
- Supervision of personnel funded by Google Switzerland GmbH (Google);
- Deployment of interns for the CERN Summer Student 2019-2020 program to work on activities related to the Project;
- Access to CERN's technical environment (hardware, software) as required for the Project;
- Support to public relations and communications activities in accordance with Annex 1 of the Framework Agreement and based on Google Switzerland GmbH (Google)'s openlab VI membership status.

Article 3
Contact points

The contact persons for the Project shall be:

For Google Switzerland GmbH (Google):

XXXXXXX

For CERN:
Federico Carminati, CERN openlab
Maria Girone, CERN openlab

For the individual use cases

Use case 1.1 Tim Bell, Ricardo Rocha (IT-CM)
Use case 1.2 Felice Pantaleo, Andrea Bocci (CMS)
Use case 2.1 Sofia Vallecorsa (CERN openlab)
Use case 2.2 Jennifer Ngadiuba, Maurizio Pierini (CMS)
Use case 2.3 Luca Canali (IT-DB), Viktor Khristenko (CERN openlab)
Use case 2.4 Sofia Vallecorsa, Taghi Aliyev (CERN openlab), xxxxx (UNOSAT)
Use Case 2.5 LHCb
Use case 3.1 Federico Carminati, Sofia Vallecorsa (CERN openlab)
Use case 3.2 Federico Carminati, Sofia Vallecorsa (CERN openlab)

CERN
1 Esplanade des Particules
CH 1211 Geneva 23, Switzerland

Article 4
Entry into force, duration and termination

1. This Project Agreement shall enter into force on the date of signature by the last Party to sign with effect of June 1st, 2019. It shall remain in force for as long as necessary to give effect to the Parties' respective rights and obligations under this Project Agreement.

2. Except in case of force majeure, each Party may only terminate this Project Agreement in the event that the other Party fails to honour any of its obligations thereunder.

Signed on …………

The European Organization for Nuclear Research (CERN)

………………………………
(Signature)

………………………………
(Full Name)

………………………………
(Email address)

Signed on …………

Google Switzerland GmbH (Google)

………………………………
(Signature)

………………………………
(Full Name)

………………………………
(Email address)

Annex 1:

Project Statement of Work

OVERVIEW

Start date	June 1st, 2019
Duration	12 months with option for renewal upon mutual agreement
Members	CERN, ATLAS, Google
Total effort	
Total cash contributions	
Total in-kind contributions	

SUMMARY OF RESOURCE CONTRIBUTIONS

	FTEs	**Facilities**[1]	**Hardware**	**Software**	**Other (Summer students, events)**	**Cash (FTE+travel, other, etc.)**
CERN						
Google Switzerland GmbH (Google)						
Total[2]						

PROJECT DESCRIPTION

The following Programme of Work is organized in three main areas with specific use cases of interest. Section 1 describes explorations in data center and computing infrastructures technologies (storage, workflow managements, containers, hybrid infrastructures integration, etc.). Section 2 focuses on application of machine learning/deep learning tools and algorithms. Finally Section 3 deals with quantum computing explorations.

1. Data Center technologies

[1] Office space, rack space, access to infrastructure or facilities, etc.

[2] Provide estimated values for facilities, hardware and software contributions

Kubernetes performance and scalability (use case 1.1)

Large computing environments require high levels of automation for both resource and workload management, as well as careful resource planning to meet the needs of daily operations and periodic spikes on resource requests while minimizing cost. Balancing the two can often be challenging.

Looking forward to the upcoming challenges of the High Luminosity LHC and its significant increase in the amount of data to be processed, this use case will explore technologies with potential to ease this task while improving resource usage. In addition to the more traditional scale out options using well established network and compute provisioning methods, it will focus on the Kubernetes project and ecosystem, where Google and more recently CERN have been investing significantly along a much larger community.

The main focus will be on extending the ongoing efforts to scale out the CERN data center to the Google Cloud Platform (GCP) (already done at scale for one time demos like Kubecon Barcelona 2019), exploring Kubernetes federation and service mesh technologies (Istio). The Batch use case can be taken as the primary use case to offer a larger CPU capacity to CERN users. Another interesting area of collaboration are accelerators, and how we can offer GPUs (currently available in a very limited amount at CERN) and TPUs transparently to our users building on top of the same Kubernetes layer. The final result should include a detailed report on the technologies used as well as cost models and efficiency in the areas of compute, networking and storage.

The current CERN capacity is in the order of 300.000 hardware hyper-threads (cores) and 100s of Petabytes of storage. For a successful evaluation of this model's viability we expect to have a significant fraction of these resources available in GCP at least for a short period of time.

Additional areas of interest include research on serverless models to simplify the LHC computing workflows. The existing workflows launched by data arriving from the detectors are a good match for pipelines triggered by the storage systems, with each step executed as an independent function. Technologies like Knative (the backend of Google's Cloud Run service offering) should allow us to again build on the same Kubernetes layer, allowing experiment users to focus more on their computing tasks and less on resource management. It should also allow Google to get a large use case being deployed in this new cloud service offering. The end result is expected to be a prototype of one of the experiment's data reconstruction relying on a single pipeline backed by GCS/S3 and Knative.

Milestone	Duration	H/W	S/W	Approx. effort (FTE)	Criteria	Outcome
M1	6 months	x	x	1	Working prototype	CERN Batch system running on Kubernetes, with resources both at CERN and in the Google Cloud Platform (GCP)

M2	6 month			0.5	-	Report on cost models and service efficiency for resources running on GCP
M3	6 months		x	1	-	Experiment data reconstruction deployed using a serverless pipeline

Composable data centers for efficient HEP computing workflows (use case 1.2)

This use-case deals, in a novel way, with event processing, pushing back the frontiers of technologies that can find application in many other areas of science and technology. At High Luminosity LHC and FCC, the higher proton-proton interaction rate, pileup and event processing rate present an unprecedented challenge to the real-time and offline event reconstruction, requiring a processing power orders of magnitude larger than today. This exceeds by far the expected increase in processing power for conventional CPUs (at a fixed cost), demanding an alternative approach. This use-case will study the feasibility of allowing HEP applications to run at heterogeneous data centers, with the goal of demonstrating that they can achieve higher throughput and better energy efficiency by running each step of a computing job on the architecture that best matches its characteristics. General purpose nodes will accelerate applications leveraging the specialised nodes available within the data center. It will also investigate source-to-source code translators to improve performance portability and will quantify the benefits of this novel approach with respect to existing ones, in terms of overall cost savings, energy efficiency, throughput, flexibility and scalability.
To fully exploit the physics reach of the High Luminosity Large Hadron Collider, the LHC experiments are planning substantial upgrades of their detector technologies and increases of their data acquisition rates. Studies are ongoing to develop the Future Circular Collider (FCC) trigger and data acquisition infrastructure, which will have even higher requirements.

This use-case deals with event processing, pushing back the frontiers of technologies that can find application in many other areas of science and technology that could benefit the European scientific infrastructure. At HL-LHC and FCC many hundreds of (pileup) events are superposed on the one of interest. The event of interest has to be disentangled from all the others. The higher proton-proton interaction rate, pileup and event processing rate present an unprecedented challenge to the real-time and offline event reconstruction, requiring a processing power orders of magnitude larger than today. This exceeds by far the expected increase in processing power for conventional CPUs (at a fixed cost), demanding an alternative approach.

We will study the feasibility of allowing applications to run on heterogeneous data centers, with the goal of demonstrating that they can achieve higher throughput and better energy efficiency by running each step of a computing job on the architecture that best matches its characteristics. General purpose nodes will accelerate applications leveraging the specialised nodes available within the data center. A source-to-source code translator will also be investigated, in order to avoid code duplication when programming different accelerators and conventional CPUs. The project will quantify the benefits of this novel approach with respect to existing ones, in terms of over hall cost savings, energy efficiency, throughput, flexibility and scalability.

Substantial improvements to the current experiments at the LHC are underway, and new experiments are being proposed or discussed at future energy-frontier accelerators to answer fundamental questions in particle physics. At HL-LHC and FCC highly granular calorimeters and complex silicon vertex trackers must operate in an unprecedentedly challenging experimental environment; moreover, the real-time event selection will pose even greater challenges. Heterogeneous computing architectures, in which general purpose and specialized processors work cooperatively, hold tremendous potential for solving these issues, accelerating applications beyond what one can expect from general purpose processors, while overcoming many of the barriers that limit the application of less conventional architectures.

To achieve processing speeds orders of magnitude beyond what is available through general purpose processors, industry leaders and the HPC community are developing new strategies and exploring innovative architectures that can work around the limitations of conventional systems, leveraging specialized processors or “accelerators” that deliver enhanced performance in areas where general purpose processors fare poorly. Examples include:

- vector processors, such as the GPUs, that increase computational performance by efficiently computing identical calculations on large streams of data;
- Tensor Processing Units for Machine Learning applications, that can deliver orders of magnitude improvements over conventional processors for very specialised workflows.

The main limitation of these accelerators is that, although they provide excellent acceleration for some well-defined workflows and largely parallel operations, they have very poor performance for scalar code and control flow structures, and are often unable to run general purpose software.

Heterogeneous computing is the strategy of deploying multiple types of processing elements within a single workflow, and allowing each to perform the tasks to which it is best suited. This approach can expand the scope of conventional microprocessor architectures, taking advantage of their flexibility to run serial algorithms and control flow structures, while leveraging specialized processors to accelerate the most complex operations hundreds of times faster than what general purpose processors can achieve. Since many applications include both code that could benefit from acceleration and code that is better suited for conventional processing, no one type of processor is best for all computations: heterogeneous processing allows exploiting the best processor type for each operation within a given application, provided that the underlying reconstruction framework is able to support them.

The workflows that will run on these heterogeneous platforms (simulation, reconstruction, analysis) have very different resource demands, requiring a flexible computing farm architecture that supports all the aforementioned kind of accelerator nodes, alongside traditional ones; the composition of the farm shall be dynamically adjusted to match the requirements in term of resources used by the different workflows. A heterogenous scheduler will be able to offload data and algorithms from the traditional nodes to the various accelerator nodes, aiming to use the most efficient resources for each task, but being able to fall back to any available one. The integration of heterogeneous computing in High Energy Physics software frameworks depends on improvements to the framework and scheduling themselves, coupled with a tailoring of the reconstruction algorithms to the different architectures.

This use-case aims to develop a solution with improved physics performance at an overall lower cost. This will allow the upgraded LHC experiments, future experiments, , and experiments at the cosmic frontier (SKA, etc.) to achieve unprecedented performance, especially in the real-time reconstruction step, which is critical to fully exploit their potential. While the diverse experiments will face different problems, they will share the overall challenge of reconstructing and analysing a large data throughput - which can be efficiently addressed with a heterogeneous computing approach.

On-demand application acceleration through remote heterogeneous computing could potentially lead to leaps in performance in cases in which:

- the results are required with small latency;
- the time needed to transfer data is small with respect to the time of the computation;
- analysing data online can improve the understanding of the observed process and can reduce the amount of data stored.

MILESTONES AND EXPECTED OUTCOME

Milestone	Duration	H/W	S/W	Approx. effort (FTE)	Criteria	Outcome
M1	1 month			1	Identification of a benchmark for the evaluation of the performance	
M2	9 months			1	Development of a scheduling system	
M3	2 months			1	Test of different topologies	

2. Machine Learning/Deep Learning

We have various applications in this area, and namely

1. Generative models for detector simulation
2. Data Acquisition and Filtering
3. Event Classification
4. Event reconstruction

Generative models for detector simulation (use case 2.1)

The CERN openlab group has now a quite extensive experience in applying generative methods for detector simulation. While results are very promising, we still could not conduct extensive hyperparameter scans to determine the optimal network configuration, the main bottleneck being the performance of the training process. The optimisation of complicated networks has high computational costs and the availability of HPC facilities, on-demand access to large Cloud resources and to dedicated hardware (such as TPUs) would enable us to define the best network configuration, maximise performance and insure quick turn-around of new trainings, when needed. In fact, given the number of deep learning algorithms that are being explored for different applications (from data quality monitoring to online selection, to simulation, reconstruction and analysis) large training workloads are likely to become more and more frequent within the High Energy Physics experiments.

In order to scale out the training process of our 3D convolutional GAN model we have tested several approaches to distributed training in different environments (HPC centers and Cloud) using both CPU and GPU accelerators. We intend to leverage those results to improve performance and effectiveness of hyperparameter optimisation in order to determine the appropriate network parameters for different types of detectors. So far our initial prototype has reached a very good agreement to standard Monte Carlo approach as far as the simulation of a specific detector geometry is concerned. The aim of this work would be to test to which extent (and at what computational cost) it is possible to tune a network architecture in order to simulate a larger range of detectors (i.e. calorimeters).

Different approaches can be explored to find the best network configuration: from the easiest, but most time-consuming sequential model based strategy, defining an initial model and a specific set of alternative configurations to test, to more refined reinforcement learning approaches, in which hyper-parameter modifications represent actions to be performed by the agent and the network accuracy is the reward function. We propose to implement an evolutionary approach in order to perform the training and hyper-parameter optimisation of our network in one step, i.e. optimising the network architecture and weights at the same time. While successful examples exist in different fields [18,21,22], this strategy is still new to our field: in particular, optimisation of a large number of parameters (network weights and hyper-parameters summing up to ~millions) represents a major challenge to this approach in terms of efficiency and computing resources, however a genetic algorithm is naturally scalable on distributed systems [21] and it might have the advantage, with respect to the typical SGD-based training, to converge faster to an absolute minimum. In order to manage the 3DGAN use-case complexity, in terms of image size, network parameters and the adversarial approach, this activity will be implemented in three major steps: initially we will reduce the problem complexity from 3 to 2 dimensions, by slicing the detector volume along the direction of particle propagation and implement the training step as genetic optimisation problem for the a classifier/regression network, similar to our GAN discriminator. During a second phase the architecture hyper-parameters will also be encoded in chromosomes and optimised. At this stage we will test indirect weight encoding in order to reduce the number of parameters and stabilize the training. The complete GAN scenario will be implemented once efficient optimisation of single networks is proven. This step requires integration of the adversarial training with the evolutionary approach which is non trivial. Here we focus on phase 1 and 2 as described above: the corresponding milestones are listed in the table below.

Apart from the direct benefit of designing a prototype capable of simulating several different detectors, this project intends to understand the limits of network generalisation and to optimise the application of a GA-based approach to a real life use case, as an alternative to standard hyper-parameter scans strategies.

Milestone	**Duration**	**H/W**	**S/W**	**Approx. effort (FTE)**	**Criteria**	**Outcome**
M1	6 months		x	1	Working prototype	Training of a 2 dimensional classifier/energy regressor network implemented as a genetic optimisation task.
M2	3 month		x	1	Working prototype	Inclusion of the architecture hyper-parameter optimisation
M3	3 months		x	1	-	Performance optimisation in terms of: 1. Computing performance and test of different accelerators (GPUs, TPUs) 2. Network accuracy
M4	3 months		x	1	-	Extension to the 3 dimensional case.

Deep Learning Inference on TPU for LHC real-time collision processing (use case 2.2)

At the LHC, 40M collisions are produced every second. Of these, only 1000 can be saved for further analysis (offline), mainly due to CPU and storage limitations. In order to guarantee the possibility of carrying on a broad physics program, one cannot just select 1000 events randomly. Instead, a set of algorithms (trigger filters) run in a real-time processing system (online) to select the most interesting events (according to physics knowledge and expectations).

The CMS trigger system is structured in two levels:

1. The Level 1 (L1) trigger operates filters 100K events/sec, based on a coarse reconstruction implemented into the detector readout electronics (ASICS or FPGAs), directly attached to the detector. The selection is mainly based on local signatures (e.g., the presence of a high-energy electron) and uses to a limited extent the global event information (e.g., total energy).
2. The High Level Trigger (HLT), consisting in a CPU farm where a coarser and faster version of the offline reconstruction runs. At this stage, the final 1000 events are selected, based on a much more accurate (than at L1) description of the event.

The main structural aspect that drives the algorithm complexity is the system latency, with in average 10 μsec and 100 msec available to take a decision at L1 and HLT, respectively.

In the last years, CMS trigger experts have started looking at Deep Learning as a shortcut to run complex algorithms online while keeping latency under control. During the LHC Run II several algorithms (mainly Boosted Decision Trees) have been deployed in the HLT system. Recently, a dedicated effort to translate TensorFlow models into FPGA firmware (HLS4ML) has allowed the development of classification and regression models for the CMS L1 system, based on deep fully connected networks.

With new Deep-Learning dedicated computing architectures emerging, the CMS collaboration is interested in exploring alternative architectures for fast inference. In this respect, TPUs offer multiple possibilities:

A. They could be used as accelerator device to run inference as a service for the HLT CPU, a development direction that is currently under investigation both for FPGAs and GPUs, both considering in-situ and cloud solutions.
B. They could be adapted to be Deep Learning ASICs, to be integrated in the data flow of the L1 trigger architecture.

Starting with a set of reference use cases, developed in the context of the HLS4ML development, we propose to carry on a comparative study between TPUs and FPGAs for L1 and HLT use cases, in order to establish the technological feasibility of a TPU-for-CMS trigger solution for future upgrades. The work will move in three directions:

1. Investigate the integrability of TPUs as Deep Learning ASICs in the L1 system.

2. Develop CPU+TPU deep learning inference system integrating the C++ based CMS reconstruction software to TensorFlow on TPU.
3. Build cloud-based solution, in which the on-site CPU farm (running CMS reconstruction software) would communicate with TPU cloud resources for inference purpose.

Deep Learning data pipelines for high energy physics and event classification (use case 2.3)

High Energy Physics (HEP) experiments and their detectors at LHC produce very large quantities of data, nominally of the order of 1 PB/sec. Very sophisticated procedures have been developed and put in production in order to identify, select and filter collisions of interest, prior to storing them for further analysis. Currently, most of the algorithms that run at the “trigger level” are handcrafted traditional performant routines that physicists used over the years.

Deep Learning provides a promising alternative approach to the problem of collision event classification and filtering. Accuracy achieved by employing various architectures of neural networks proves them an ideal candidate for replacement of traditional algorithms.

This work addresses key technological challenges in the preparation of a data pipeline for DL research and aims at providing a demonstrator of tools and methods that can improve the productivity of the data scientists/physicists. Key area of investigation are: performance at scale, ease of deployment and integration of the component of the data pipeline, use of industry-standard APIs.

In particular, we plan to deploy the data pipeline and DL workload using the Google Cloud Platform. For the purpose of data preparation and feature engineering, we plan to use Apache Spark and for the setup of the DL training pipeline we plan to investigate Kubeflow. For defining and training DL model we will utilize TensorFlow and profit of hardware acceleration with TPUs and/or GPUs. The workflows will be defined using Python and Jupyter notebooks on Google Colab.

Deep Learning for Earth Observation (use case 2.4)

The United Nations Institute for Training and Research (UNITAR) hosts the UN Operational Satellite Applications Centre (UNOSAT), which analyses satellite imagery to support disaster response and humanitarian operations. Because of the high level of precision required, manual analysis of a refugee settlement in a satellite image can take many hours (sometimes days).

As a consequence and because of time and manpower constraints, UNOSAT can currently fulfill only 5-10% of UN requests.

UNITAR is researching the use of Artificial Intelligence and Deep Learning to improve efficiency and reduce this amount of time, allowing an effective deployment of field operations in critical humanitarian situations.

A unique partnership between CERN openlab, Intel, and UNITAR has been created, a few years ago, in order to use Deep Learning methods to improve the analysis of optical satellite imagery for humanitarian purposes and in particular automatize, shelter counting in refugee camps, mapping of flooded areas and infrastructure damage detection using Deep Learning models.

An initial study performed last year by a student from the CERN openlab Summer Student project, leveraged transfer learning on a Convolutional Neural Network (FaceBookAi Detectron framework) for counting in refugee camps images. The initial results achieved a 82% average precision and a x200 speedup with respect to UNOSAT standard approach (visual recognition by a trained expert). Results are now being integrated by the UN Global Pulse office in their studies to enrich and refine Deep Learning based prediction tools.

We propose to extend this approach to the problem of detecting flooded areas and determine infrastructural damages from satellite images, designing a DNN capable of reaching a higher level of accuracy (above 90-95%) as required by UNOSAT operations.

The image segmentation models are being applied to satellite imagery with mixed results (see results of a recent Kaggle challenge on the subject [19]): in general the diversity of environments including urban areas, dense vegetations or mountainous terrain and the relative small size of available training set make complicate the task. In this context, the U-Net architecture, initially developed for medical imaging tasks, has produced very interesting results [20].

Our work will initially leverage on the results obtained last year for the refugee camps image analysis, testing the performance of region-based CNNs on the detection of flooded areas. The optimisation and adaptation of the tool will take into account the specific features of water and mud in satellite imagery. Results will be compared to the performance of classical approaches and to other existing methods (U-Net based for example).

Multiple natural and human-made disasters can affect various areas of the world simultaneously. We believe the development of an automated change detection system could be an instrumental tool which could enhance our capacity to provide immediate *findings and reports to UN organizations and NGOs working in the field addressing their needs and supporting them in evidence-based decision making at the highest level. This system could be conducive in responding rapidly and timely to any specific natural or human-made disaster by providing factual and credible information leading towards achieving our common goals.*

Milestone	Duration	H/W	S/W	Approx. effort (FTE)	Criteria	Outcome
M1	3 months		x	1	Working prototype	First R-CNN based prototype on RGB images.
M2	3 month		x	1	Working prototype	Extension to multi-spectral data
M3	3 months		x	1	-	Performance optimisation
M4	3 months		x	1	-	Comparison to U-Net based approaches

RICH reconstruction using Google TPUs (use case 2.5)

The Ring-imaging Cherenkov (RICH) detectors determine the velocity of particles coming from proton-proton collisions at the LHCb detector at CERN. When particles pass by a C4F10 radiator gas, they emit cones of photons whose angle is linked to the particle speed and particle type. These photons are reflected in two mirrors prior to being detected in 64-channel Multi-Anode Photomultiplier Tubes (MaPMTs), translating into an array of pixels.

The RICH reconstruction is currently a costly process in terms of computing power. For each particle, all possible MaPMT pixels according to variations in speed and type are associated. One of the mirrors is a section of a spherical mirror, and hence calculating a single reflection means solving a quartic equation. Additionally, since different sets of pixels would be associated with each particle depending on their type, a "likelihood" is defined for each association of particles and types. The process of likelihood minimization, intractable in practice due to the number of possible associations (possible types ˆ number of particles), is tackled with a local search.

Other reconstruction mechanisms exist in literature. In this project, we propose exploring non-conventional global techniques for RICH reconstruction, applied to the LHCb use case.

Convolutional Neural Networks (CNNs) are a well-known category of Neural Networks within the Machine Learning field. In particular, they have shown to be effective in classification problems with large training datasets. In addition to CNNs other deep learning methods are studied, such as Recurrent Neural Networks, and combinations of the two. We propose to formulate the RICH reconstruction problem as a classification problem of n particles, where every particle can be classified as one of five possibilities: muon, kaon, pion, proton, electron or deuteron. We intend to employ Monte Carlo datasets as a source of training and test data, and validate our results with existing validation tools from the LHCb framework.

Such a project will greatly benefit from the use of Google's TPU accelerator hardware in order to perform the training and tuning of the neural network models that are studied to find an optimal reconstruction approach. Various Neural Network models are under study, with a large and complex architecture that would be optimally suited for the TPU's time-to-accuracy optimisation. As a result, Google's TPUs could contribute significantly to the success of the research project.

MILESTONES AND EXPECTED OUTCOME

Milestone	Duration	H/W	S/W	Approx. effort (FTE)	Criteria	Outcome

KEY RISKS

Summary description.

	Risk	Probability	Impact	Mitigation Strategy
Organizational risks				
Critical path technology risks				

3. Quantum Computing

This use case will include Quantum Computing applications to High Energy Physics. Its aim is to recast relevant High Energy Physics computational problems into quantum algorithms and explore new applications such as direct simulation of quantum systems. It is expected that the work will initially be done on quantum simulators and, when available on quantum hardware.

Quantum generative models(use case 3.1)

Generative models are being explored by the High Energy Physics community, and by the CERN openlab group, in particular, as possible fast solutions to replace Monte Carlo simulation. Models such as Boltzman Machines, Generative Adversarial Networks (GAN) and Variational Auto-Encoders (VAE) learn the underlying probability distribution from a training set and sample it during the generation step. VAEs rely on latent (non-visible) variables to encode properties of the training dataset and re-use them during the generation (de-coding step). GANs implement the adversarial training process as a zero-sum game: convergence occurs when the corresponding Nash equilibrium is reached.

Implementing generative models on quantum processors could bring potential advantages in terms of representational power and computing time (for training and inference). Recent studies on Quantum VAE (QVAE) show promising results, although no quantum supremacy in this field has been proven so far. A quantum-classical hybrid approach implements the QVAE as a classical auto-encoding network and a Quantum Boltzmann Machine-based generative process on the D-Wave annealer [12]: results on the MNIST dataset are comparable to well established classical methods (i.e. convolutional neural networks). Another approach focuses instead on the auto-enconding process: Quantum Auto-Enconders can be implemented as a sequence of gates [13] in order to compress the input qubits information into a smaller number of output qubits.

Quantum GANs, QuGAN, have been proposed in [14] and the theoretical implications of a quantum approach to the adversarial training are detailed in [15]. In QuGAN, the optimisation problem is reformulated using the quantum formalism and the cost function is minimized by evaluating the gradients on a quantum processor. The method is proved using a simple numerical example (a two labels system) and a total of 5 qubits. Convergence is reached after 100,000 training steps.

We propose to explore the possibility to extend the QuGAN system to a realistic example: the simulation of the physics quantities describing an interesting final state (i.e. the four-momentum components of the particles produced by a specific Higgs boson decay). This “ultra-fast” simulation would skip entirely the detector output reconstruction process and provide in one step the information needed by preliminary analysis design studies. Similar approaches are being studied by the High Energy Physics community using classical techniques. Given the relatively small number of outputs (in the range of 10)) this kind of application seems well suited to the dimensions of near-term quantum processors (tens of qubits). Moreover, since the input (training) data and the output are quantistic by nature, this application could benefit by the use of quantum states to represent the input.

Initial studies will be devoted to the implementation of the GAN model and the adversarial training mechanism using Cirq. The ability to reproduce the results presented in [14] (on a simplified numerical use case) will allow us to benchmark the quality of our implementation. The next step would be to understand the maximum problem size that could be solved by such a system and to design an efficient strategy to represent input and output quantities related to our fast simulation use case in terms of quantum states (i.e. quantum states amplitudes). It should be noted that this step, in itself, represents an interesting challenge, since most QML applications so far, focused on classification problems where the output variables are simple 0-1 flags in nature.

Depending on the results of this first use case we could extend this project to a second, more realistic, problem: the simulation of the detector output, interpreted as a pixelized image. Generation [12] and classification [13] quantum networks have already been applied to image datasets such as MNIST. In our case a major challenge would be the high dimensionality of the input image (typically thousands of pixels) that should be reduced to a manageable size, by classical approaches such as downsampling or Principal Components Analysis techniques.

Milestone	Duration	H/W	S/W	Approx. effort (FTE)	Criteria	Outcome
M1	3 months		x	1	Reproduce results in [14]	Implementation in Cirq of the method proposed in [14]
M2	2 month		x	1	-	Representation of the simplified fast simulation input and output variables as quantum states
M3	6 months		x	1	Working prototype	Implementation of a quGAN example to generate simplified fast simulation output

M4	1 month		-	1	-	Result publication on a peer-reviewed journal or conference proceedings

Optimisation of GRID job placement (use case 3.2)

The ALICE experiment at CERN has access to more than 70 computing centers, located in 40 countries on 5 continents, offering access to upwards of 150,000 CPU cores and 120 PB of storage for the processing needs of the experiment. The resources, both CPU and storage, are highly heterogeneous and the workflows are highly IO intensive (60GB/s reading and 6GB/s writing). The Our mission is to globally optimize the data access for both read and write operations.

For data processing jobs, the decision on where to read from or to write to is based on a distance metric between the client and all storage nodes. This metric is a weighted function of storage reliability, network topology (RTT) and remaining capacity (for writing). While having the great advantage of being fully automatic and generally correct, it doesn't take into account infrastructure elements like network links capacity and utilization nor the IO capabilities of the storage elements or the interconnect capacity between the storage and the computing nodes.

To find the correct balance between how much data to stream from the site local data storage and when it would be worth to rely on external locations is a hard problem to which we don't have a solution yet. The access to remote copies affects circa the 5-10% of the ALICE analysis jobs which means a total amount of 50PB/year in terms of data volume. With the help of quantum computing algorithms applied to the data replication protocols, this amount of data could decrease in a factor 2. This has a clear effect on the network usage whose occupancy could decrease in a 50%.

Ideally the final system should be able to:

- Interface with the monitoring system to receive the relevant information;
- Learn from the past decisions retrieving information from the ALICE MonALISA monitoring system;
- Provide hints to job and file brokers in quasi-real time;
- Sustain one order of magnitude higher access rates, from the current O(10kHz).

The potential benefit is in better utilization of the CPU cores. At the moment, the analysis uses 48% of the allocated CPU, spending thus half of the time in IO operations. With a global optimization the turnaround time for the physics analysis would be shorter and the physicists would waste less time waiting for the results to be available.

The work will make use of the extensive monitoring database that the ALICE experiment has collected over the last 10 years of Grid operation, based on the MonALISA technology [5,6], and “live” tests can be conducted on the ALICE WLCG infrastructure where jobs are continuously run.

Some of the very first applications proposed for QC have been in the domain of optimisation. In case of extremely non-linear systems, classical optimisation methods such as gradient descent have shown to be sub-optimal, leading to poor local minima or failing to converge at all, especially in quickly changing systems. Minimisation via classical sampling is also problematic in case of very high dimensionality of the parameter space, and even methods such as Gibbs sampling are known to have slow-mixing properties. The ability of QC of exploring several states concurrently has led to one of the first “quantum dominance” algorithms [7] and since then search and optimisation algorithms have been the focus of a large part of the studies in QC [8, 9] (see also the excellent survey [10] and references therein).

The challenge of this project will be to design a system that could be run (at least in principle) dynamically, to adapt to the current status of the Grid. We expect most of the work to be done on QC simulator, with tests on real QC hardware when the algorithms have been implemented and validated.

CERN openlab will provide 1.2 FTEs for this work, out of which 1 financed by the project.

The ALICE team at CERN will provide 0.3 FTEs and expertise and guidance in the interpretation of the monitoring data. It will also test the results of the optimisation on the ALICE Grid.

The Polytechnic University of Bucharest will contribute their expertise in the analysis and optimisation of the data extracted from the MonALISA monitoring system, providing a total of 0.2 FTEs.

The Laboratory of the Techniques of Informatics and Microelectronics for integrated systems Architecture will provide its expertise in optimisation and statistical analysis techniques, providing a total of 0.2 FTEs.

Google will collaborate to the project by providing support and guidance for the implementation of the quantum algorithms. Google will also provide support in the usage of the quantum simulation software (Cirq simulator) and, when appropriate during the project access to the quantum hardware to test the algorithms developed.

Several initiatives have started at CERN along this line, for example the idea of optimising WLCG data storage within geographical regions, by prototyping a distributed storage service that concentrates data (in the so-called data-lakes) and then streams it to cache locations as efficiently as possible [11].

MILESTONES AND EXPECTED OUTCOMES

Expected results	
R1	Demonstration of the algorithm on the quantum simulator;
R2	Test of the algorithm on quantum hardware;
R3	Result publication on a peer-reviewed journal or conference proceedings;

KEY RISKS

	Risk	**Probability**	**Impact**	**Mitigation Strategy**
Organizational risks				
Critical path technology risks				

References

1. Foster I, Kesselman C (1998). *The Grid: Blueprint for a New Computing Infrastructure (The Elsevier Series in Grid Computing).* Morgan Kaufmann; 1st edition.
2. https://home.cern/topics/large-hadron-collider. Last read on September 7, 2018.
3. http://wlcg.web.cern.ch. Last read on September 7, 2018.
4. http://alice-collaboration.web.cern.ch. Last read on September 7, 2018.
5. http://monalisa.caltech.edu. Last read on September 7, 2018.
6. http://alimonitor.cern.ch. Last read on September 7, 2018.
7. Grover, L. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.***79**, 325–328 (1997).
8. Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm applied to a bounded occurrence constraint problem. Preprint at arXiv:1412.6062 (2014).
9. Moll N et al. 2018. Quantum optimization using variational algorithms on near-term quantum devices. Quantum Science and Technology, 3(3).
10. Zahedinejad E and Zaribafiyan A, 2017. Combinatorial Optimization on Gate Model Quantum Computers: A Survey. https://arxiv.org/abs/1708.05294

11. Bird, I. and Campana, S., The WLCG strategy document for HL-LHC
12. A. Koshaman et al., https://arxiv.org/pdf/1802.05779.pdf
13. Lamata L. et al. https://arxiv.org/abs/1709.07409
14. Dallaire-Demers P. and Killoran N.https://arxiv.org/pdf/1804.08641.pdf
15. Lloyd S. and Weedbrook C., PHYSICAL REVIEW LETTERS 121, 040502 (2018)
16. Farhi E. and Neven H., https://arxiv.org/pdf/1802.06002.pdf
18. F. Petroski Such. et. al. arXiv:1712.06567
19. The DSTL satellite imagery feature detection, https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection
20. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, arxiv:1505.04597
21. T. Desell. IEEE eScience. 2017.
22 .K. O. Stanley, D. D' Ambrosio, J. Gauci. Artificial Life Journal, MIT Press, 2009.