SWIFT-HEP
V1.2
14-August-2020

SWIFT-HEP

SoftWare InFrastructure and Technology for High Energy Physics

Proposal to STFC

To be reviewed by the Peer Project Review Panel

TABLE OF CONTENTS

0. Executive summary

To exploit the full physics potential offered by the large data sets collected by the next-generation experiments, new software and algorithmic techniques are required to handle their production rate, volume, and complexity. The existing code, which has largely evolved from an era of entirely different hardware and software technology, is completely inadequate for the task. Over the years, serious and competent engineering has been invested to eliminate obvious inefficiencies through profiling and straightforward re-factoring where possible, but now a step-change is required. This need has led to an international roadmap for software development and the establishment of the HEP Software Foundation (HSF) to address problems that are common across HEP experiments.

This proposal identifies a set of activities we propose to undertake in the UK that are aligned with this global vision of HEP software. These concentrate on activities where the UK has leading expertise and experience and/or high profile responsibilities and can therefore make an important and valuable impact. It is centred around work that will serve the HEP experimental community by developing software tools that occupy the presently underfunded space between the Grid infrastructure and the frameworks that are specific to the experiments.

We address here a critical lack of funding for software, which is needed to fully exploit the physics capabilities of future experiments, while keeping computing costs affordable. This project will enable the full exploitation of the existing investments in the experiment construction and computing Infrastructure provision. Solutions will focus on data science tools and inference, areas of strong societal impact and will lead to the development of skills transferable to the overall economy.

The computing requirements of the highest priority projects within the HEP programme are posing significant challenges which must be addressed head-on. The four LHC experiments (ATLAS, CMS, LHCb and ALICE) recorded almost 1 exabyte of data to make groundbreaking discoveries, from the Higgs Boson to pentaquarks. Major upgrades of the LHC experiments are due to take place in the 2020s which will greatly augment their datasets, while in parallel we will see the build-up to operation of the next generation of neutrino experiments (DUNE and Hyper-Kamiokande) with LHC-scale volumes of highly granular data.

Software development is recognised by the international community as an essential aspect to achieve the physics goals. The recent update of the European Strategy for particle physics specifically notes that professional software engineers are needed for the future of the field and an international community is emerging around this research support area.

The goal of this proposal is two-fold. (i) The funding sought will provide a direct contribution to the international effort and allow us to shape the future direction of computing and software development in the field of particle physics. (ii) A new cross-experiment collaborative environment will be established in the UK to create a framework to address common problems in software development which will serve us for the next decade and beyond.

SWIFT-HEP identifies 5 work packages that leverage the expertise available in the UK to maximise the impact of the UK-wide collaboration within the international efforts. This work is complementary to the existing activities within the experiments to provide a core of research software engineers and physicist programmers with the skills needed for modern software development. Central to this platform is the training provided to the HEP community and the career development of younger colleagues who will acquire skills that are currently in high-demand both within the HEP community and industry. The use of industry-grade tools is essential to ensure long-term sustainability of the software developed, hence cooperation with industry is part of this proposal.

Conceptual design reports to outline the software and computing development for the high luminosity LHC were recently submitted and reviewed by the LHCC, with similar documents being prepared for DUNE. Technical design reports will be submitted in 2023 and will be a major deliverable for the first phase of SWIFT-HEP (this proposal) to define the UK contribution to software development in particle physics.

A second phase of SWIFT-HEP is expected to be proposed in the next 2 years to build upon the first phase, to deploy the technology defined in the technical design reports, and to drive the developments needed for experiments in the next decade. The resources requested in the current proposal will allow the UK to help shape the evolution of software towards the next generation of experiments, but will not be sufficient to resolve the challenges of the exascale era. The overriding objective of the larger project foreseen for the second phase will be to remove the software and computing bottlenecks for those experiments on time for their commencement of data-taking.

1. Science case and supporting information

1.1 STFC science and strategy

SWIFT-HEP fundamentally enables the science goals of the LHC Experiments (and others) as peer reviewed and approved elsewhere. We particularly enable the science goals of the upgrade projects approved for each experiment aimed at Run-3 for LHCb and Run-4 for CMS and ATLAS. Supporting the development of the software infrastructure to be commensurate with the data to be produced is no different to upgrading LHC to run at higher luminosity or upgrading the detector to handle the higher luminosity. Specific science goals of this project are therefore:

(i) To measure the detailed properties of the Higgs boson and its couplings to all particles;
(ii) To probe the high energy frontier to search for new heavy particles beyond the standard model;
(iii) To characterise and understand the origin of CP violation in the B meson and charmed meson sector;
(iv) To measure rare decay processes and search for evidence of physics beyond the standard model arising in loop dominated processes;
(v) To measure the mass hierarchy and observe and measure the CP violating phase in the neutrino sector;
(vi) To observe solar neutrino and supernovae neutrino events and contribute to the multi messenger physics programme.

SWIFT-HEP will enable STFC to answer the strategic goals of “World Class Research and World Class Innovation” and “What are the basic constituents of matter and how do they interact?”. This is a top strategic priority for STFC and it involves the LHC, neutrino, dark matter and flavour sector experiments. The project is also aligned with STFC's priority to maximise exploitation of UK investment in CERN (as a member state) and elsewhere.

The SWIFT-HEP technical goals are aligned with the strategic themes of “data intensive science”, “building international influence” and “developing advanced technologies”, providing the software tools needed to process and analyse the HEP vast datasets. In this we make efficient use of the UK eInfrastructure. The development of modern software tools delivers world class innovation, while training researchers with skills that are in high-demand in industry and academia.

Computing contributes significantly to the cost of operating HEP experiments, with the UK currently providing about 13.5% of the total ATLAS computing resources, 6% for CMS, 20% for LHCb and a significant fraction expected for DUNE. Resources are provided as a cross-experiment initiative via GridPP with a cost of about £5M/year. The overall resource requirement is expected to increase in the coming years as the experiment upgrades are designed to increase the overall data throughput by a factor of 10. Only a modest increase in resources is expected due to the diminishing growth rate at equal cost of computer processing and storage.

The LHC experiments have recently updated their resource projections for the 2020s, modeling the needs for processing resources (in HEP SpecInt 06, HS06) and disk storage (in Exabytes, EB). Fig. 1 shows the result of this exercise for ATLAS with similar results available also from CMS and LHCb. The "Sustained budget model" lines show the expectation of the computing resources that can be delivered by maintaining the current level of investment. The "Baseline" scenario shows the resources that would be needed if no further software development is done, clearly emphasizing that LHC computing would not be affordable towards the end of the decade. A programme of R&D is therefore needed over the next few years to enable new tools to fit the cost of computing within the expected envelope. Failure to do so will result in seriously compromising the physics output from the experiments and jeopardise UK investments of more than £100M in the upgrade of the LHC experiments.

At same time the evolution of computing hardware and software technologies calls for HEP software R&D. In particular the hardware has evolved towards parallel architectures where it is expected that the bulk of the arithmetic processing is performed using single instruction multiple process (SIMP) and single instruction multiple thread (SIMT). Optimising software to make use of this hardware is becoming the key to increase computing efficiency. Scientific computing is moving towards the use of accelerator hardware, such as GPUs and FPGAs, where thousands of threads operate simultaneously to perform fast scientific computations. This is an opportunity, as real-time applications such as low latency triggers can be developed using commodity hardware and software tools. At the same time GPUs are a promising technology with which to run simulation and data processing at lower cost and better power efficiency. To unlock these potential gains an aggressive software R&D programme is needed to modernise the current HEP codes and optimise their performance on the heterogeneous hardware architectures that are expected to be developed in the next decade.

Such developments need to be sustainable, to last for the lifetime of the current experiments, to be open source, to allow for collaboration between different teams, and to follow industry standards, to exploit state-of-the-art training and enable collaboration with industry to access and inform the latest technologies.

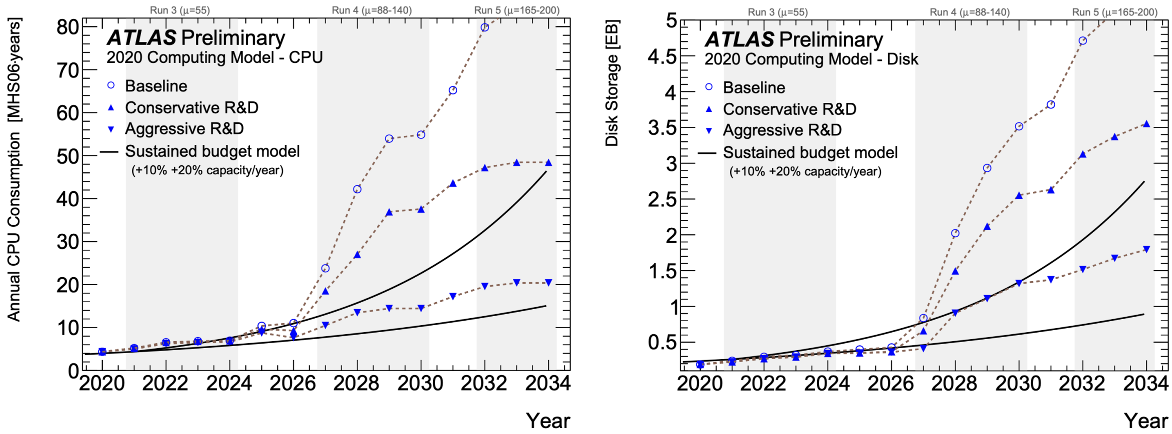

Fig 1. Projected CPU and disk resources needed by the ATLAS experiment in the next decade. Three scenarios based on different R&D efforts available are shown. The solid lines indicate the expected amount of resources available if the current level of expenditure in hardware is maintained. It can be seen that aggressive software R&D is needed if costs are to be kept within the optimistic 20% capacity increase per year.

1.2 Awareness and context

This proposal covers this crucial 3 year period between April 2021 and March 2024 for the UK to participate and lead the development of the means to execute the next generation experiments.

The general purpose detectors (GPDs) at the LHC are working on major upgrades for the LHC phase-2 which will commence operations in 2027, while LHCb and ALICE are going to exploit major upgrades in the next LHC run in 2021-24, with LHCb planning for a further upgrade in 2030. Both DUNE and Hyper-Kamiokande experiments will start to take data in the late 2020s. R&D towards this new era for particle physics is advancing with Software and Computing Conceptual Design Reports (CDRs) delivered to a dedicated LHCC committee in 2020 for ATLAS, CMS, the Worldwide LHC Computing Grid (WLCG), and the HSF. This R&D will culminate with software and computing Technical Design Reports (TDRs) for ATLAS, CMS, and WLCG and a readiness document for DUNE in late 2023. At the same time, the experience from the earlier LHCb upgrade will inform the technology choices both for its second upgrade and those made by other experiments.

The case for increased investment in software was made in 2018 in a "white paper" published by the HSF with many UK signatories who are also signatories of this proposal. This resulted in a multi-million dollar investment in the USA, initially by the NSF and then by the DOE. CERN management is very supportive of these initiatives with software workshops now held together with the WLCG workshop to also facilitate communication.The following comment was made by the LHCC review committee after discussing the CDRs:

"One area of concern shared by the experiments and WLCG is finding means to ensure that the highly skilled personnel essential for R&D in computing and storage have meaningful career paths within the LHC community to provide for sustainability and the need for continual evolution over the lifetime of HL-LHC. "

The first steps in the UK were the short-term funding (£50k) of the efficient computing for high-energy physics (ECHEP) project that enabled the discussion and organisation of this initiative to take place. Further to that, funding was secured as part of the UKRI Excalibur programme (£235k) that was used to jump-start several initiatives.

It is vital for the UK to get involved in this software and computing initiative to ensure that the skills and infrastructure needed to exploit experiments in the future are available. The investment should be compared to those made in the construction of the experiments. We initially propose a £1.2M investment for three years to drive the developments towards technical design reports for the high luminosity LHC. A second, larger phase of SWIFT-HEP is foreseen, building upon the experience acquired during the first phase. As part of the first phase project management and a collaboration organisation will be defined for this second phase.

Fig. 2 shows a simplified timeline of the infrastructure evolution together with common software initiatives in the UK and internationally, showing how a UK ramp-up in this initiative is taking place. The funds sought for the initial three-year phase will allow us to participate in the R&D projects specified in the conceptual design reports, leading to technical design reports and a leadership in well-specified developments in the second phase of SWIFT-HEP.

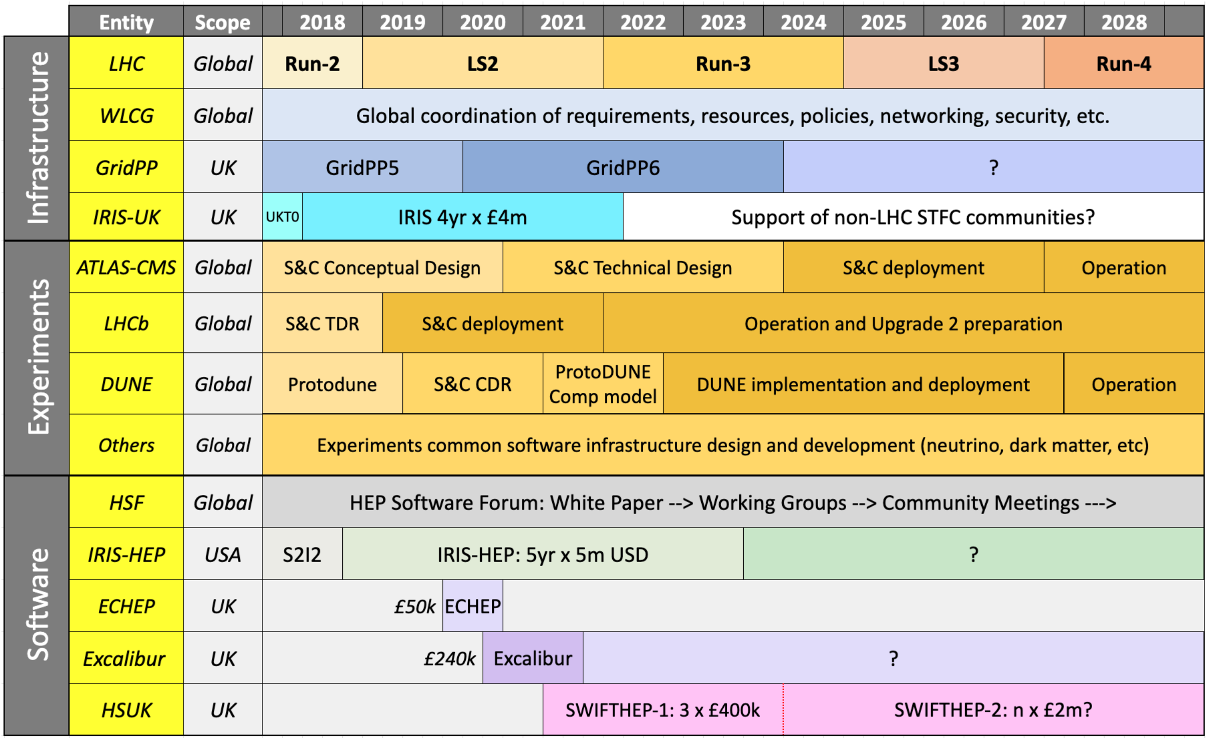

Fig. 2: Long term timeline and international context

1.3 Project description and objectives

Project description

The project aims to establish a software collaboration within HEP in the UK to ensure that future investments in HEP experiments will have access to the technology needed to exploit the data to maximise the physics return on investment. This will also create a core of knowledge and training in skills that are in high-demand in UK industry. SWIFT-HEP aims to align with the recently published European Strategy for Particle Physics[1], which calls for an increase in collaboration and effort in software and computing for HEP:

[1] https://home.cern/sites/home.web.cern.ch/files/2020-06/2020%20Update%20European%20Strategy.pdf

“The community must vigorously pursue common, coordinated R&D efforts in collaboration with other fields of science and industry, to develop software and computing infrastructures that exploit recent advances in information technology and data science. “

The first phase of the project (referred to as SWIFTHEP-1 in Fig. 2), aims to establish this effort in the UK. This is an R&D phase, which will lead to the technical design reports for the experiments’ software and computing infrastructure, which will be followed by a deployment and implementation phase.

Objectives

The main objectives are to

(i) consolidate a UK community working in the HEP software area, this is already ongoing thanks to the previous initiatives discussed in Sec. 1.2.

(ii) contribute to the delivery of the experiments software and computing technical design reports and define the UK contribution to the deployment and operation of the software tools needed by the HEP community in the era of HL-LHC and high-intensity experiments later in the decade

(iii) establish a long term community centred around software engineers that could tackle the needs of HEP experiments in future endeavours.

1.4 Complementary and competing research

The SWIFT-HEP proposal is centred around work that will serve the HEP experimental community by developing software tools that are between the Grid infrastructure and the frameworks that are specific to the experiments. This is described in Fig. 3, with the green layer 3 addressing cross-experiment issues. The full list of signatories is available in Appendix and shows the UK wide involvement in this initiative. The work proposal will be part of a growing international community and has as main reference points the Data Organisation Management and Access (**DOMA**) project within WLCG for the data management work and the **HSF** for the event generator, simulation, reconstruction and analysis work, as highlighted in the support letter attached.

A larger initiative started in the USA in 2018 with the IRIS-HEP project funded by the National Science Foundation (NSF) as a response to the HSF community paper. This is the main initiative worldwide and is funded at a level of 5 M$/year. A similar initiative, Center for Computational Excellence (CCE) was funded by the Department of Energy (DOE) at national labs in the USA to develop software to run on accelerator hardware. Other initiatives are starting elsewhere in a collaborative environment for the benefit of the wider HEP community. Our aim is to collaborate with the US initiatives (see letter of support) and we bid for a package of work that is complementary to what is currently ongoing elsewhere, by building upon the expertise present in the UK.

Such collaboration will be established early in order to avoid “competing” on topics that are of common interest. The current international organisation allows for R&D ideas to be openly presented and discussed, so that different ideas are developed in a healthy competition, with a common final product in mind driven by modern open source development. At the same time tools from other communities are considered and evaluated to achieve further efficiency. The areas addressed in SWIFT-HEP exploit expertise unique to the UK: FPGA development for data acquisition and readout has long been a UK mainstay (Imperial, RAL), and SWIFT-HEP will capitalise on this expertise to deliver FPGA-accelerator developments for offline computing. Similarly, the UK is known for expertise in generators (Birmingham, Durham, Glasgow, UCL, Warwick) and simulation (Manchester, Warwick). These are areas where SWIFT-HEP will naturally lead in the international landscape.

Participation in international workshops and meetings is essential to deliver a world-leading programme. Expertise is currently available in the UK as part of GridPP and of the experiments. SWIFT-HEP proposes to attach additional effort to this expertise to establish ourselves as a leading member of this international organisation.

Layer	Domain	Experiment 1	Experiment 2	Experiment 3	
6	Physicists	Analysis code	...	...	
5	Experiment Physicists programmer and software engineers	Analysis framework. Simulation, Reconstruction, Calibration Code	...	...	Moving Software down the stack
4	Experiment Software Engineers	Software Frameworks	...	...	
3	Common Software HSF / SWIFT-HEP	Common software components (Data management, Generators, Geant4, Accelerator integration)			
2	GridPP / WLCG	Middleware infrastructure for Distributed Computing			
1	GridPP / WLCG	Physical Hardware			

Fig 3. Complementarity of the SWIFT-HEP work with other packages of work in the HEP community.

1.5 Track record

The SWIFT-HEP project is a cross-community proposal that will develop common software tools for the benefit of all of the experiments. While this may be seen as a new project, it is one that builds upon existing experience developed within the experiments and GridPP. SWIFT-HEP will provide additional capability in the UK to exploit particle physics experiments. The establishment of a truly cross-experiment software project will enable us to strengthen the UK track record in software engineering and management, and provide career development opportunities for research software engineers and physicist programmers.

The selection of work packages emerged from dedicated discussions with the community within the context of the ECHEP project. They are based on our unique capabilities and around areas of proven world leadership at the institutes involved. For each work package the track record is summarised below.

WP0 management: **Costanzo** (PI) held the role of ATLAS Software and Computing coordinator in 2016-18 and was the editor of the ATLAS strategy document for computing at the high luminosity LHC. Previously he was one of the lead developers of the ATLAS software having contributed to the Geant4 simulation, the event data model definition and the validation of the software. **Fitzpatrick** (deputy PI) served as Project Leader of the LHCb high-level trigger from 2017-2019 and led its upgrade R&D efforts. He currently holds a UKRI Future Leaders Fellowship on real-time analysis for trigger and reconstruction, and is part of the team performing R&D for the software trigger of LHCb's second upgrade.

WP1 Intelligent data management: Developed by groups at RAL-SCD, Imperial College and Brunel, this work package leverages the expertise in data and workflow management at the institutes and will draw from their teams of research software engineers involved in grid development. RAL hosts the second largest Tier-1 for WLCG with **Dewhurst** as the Tier-1 manager directly involved in the project.

WP2 Event generators: Developed by teams at UCL and Warwick. The UCL team has a long history in involvement in event generators (**Butterworth**) and the work will be carried on by **Gutschow** (ATLAS physics modelling group convener) together with the UCL Research Software Development group, a leading programming group with more than 20 experienced RSEs. The work will be carried out with input from the Durham team (**Krauss**, **Schoenherr**) who are leading authors of Sherpa and with **Buckley** at Glasgow, lead developer of LHAPDF and Rivet. The work at Warwick on the Pythia8 fragmentation model will be overseen by **Kreps** (lead maintainer of Evtgen and current LHCb simulation convener). This will be carried out in collaboration with **Ilten** (Birmingham) and the group at Monash University.

WP3 Simulation: Developed by teams at Warwick and Manchester. The lead developer at Warwick is **Morgan**, currently Software Management Coordinator for the Geant4 simulation software and a member of the Geant4 Steering Board. He leads the Software Development Working Group for SuperNEMO and works in the ATLAS Simulation project on performance profiling and improvements. As Co-Convenor of the HEP Software Foundation's Software Tools and Packaging Working Group, he studies deployment of HEP software to end users across heterogeneous systems. At Manchester **Davis** pioneered fast calorimeter simulation and ultra-fast simulation for LHCb and is currently LHCb Simulation convener. **Gersabeck** has developed hybrid fast simulation at LHCb

WP4 Trigger and reconstruction. The UK has traditionally a large involvement in the development and operation of trigger systems for HEP experiments. The focus is shifting towards the use of software & firmware triggers with GPU and FPGA developments taking place in the UK for the major LHC experiments. This package is developed by the PPD at RAL building on the expertise in accelerator software development of **Emelyanov**, **Martin-Haugh, Olaiya, Schuh** and **Shepherd-Themistocleous,** and at Warwick by **Martin,** an expert software engineer who is in charge of the deployment and operation of the ATLAS high-level-trigger until the end of 2021. Sussex is involved via the expertise of **Cerri** as coordinator of the ATLAS hardware tracking trigger and in collaboration with **Tapper** (Imperial) as coordinator of the CMS level-1 trigger, and **Jones** (Cambridge) as a lead developer of the LHCb software reconstruction and trigger system.

WP5 Analysis systems. This work package is developed at Bristol and driven by **Kreczko** and **Krikler**, both experienced software developers and initiators of the FAST-HEP toolkit for high-level physics analysis. **Kreczko** is a software expert for CMS, LZ and DUNE, currently Deputy Simulations Coordinator for LZ and area leader for Analysis in the ECHEP project. **Krikler** is a fellow of the Software Sustainability Institute and current convenor of the PyHEP working group under the umbrella of the HEP Software Foundation. This work package will collaborate with **Rodrigues** at Liverpool, who is a member of the HSF coordination team.

1.6 Impact

Efficient HEP software: The main impact of this work is to provide software tools and infrastructure to be used by HEP experiments to improve the efficiency of their software systems. A software "upgrade" will enable us to process the data from the next generation experiments within an affordable computing budget envelope (refer to Fig. 1). The update of Fig. 1 for the HL-LHC TDRs in 2023 will incorporate developments that are planned by SWIFT-HEP.

Training: The training and career development of research software engineers (RSEs) and data scientists is crucial for the sustainability of the field. We will provide training opportunities for members of SWIFT-HEP, which will also be open to the community as a whole. The enhanced training provisions will uplift skills across the field, improving research outcomes, and given that a significant proportion of PhD/RAs move to industry, will also help address the significant skills-shortage of software engineers and data scientists in industry. The training plans are detailed in Section 4 and will be delivered in collaboration with the Centres for Doctoral Training in Data Intensive Science.

Technical engagement: Our proposed work will be of interest to the wider community. The handling of large datasets is common for all data intensive sciences which will eventually have to seriously address datasets equal to those produced by the HL-LHC upgrades. The real-time processing of data and the use of FPGA and GPU architectures will benefit other applications where throughput and latency are critical such as fast diagnostics for plasma disruptions in tokamaks, and the growing requirement for simulation and modeling concurrent with experiments. The work on simulation and the development of Geant4 is of interest to other communities where the interaction of particles with matter is studied, such as medical, nuclear or space applications. We will invite experts from other research fields to our workshops to enable exchange of ideas and foster cross-disciplinary collaborations.

Industrial and commercial engagement: The collaboration with industrial partners is vital to the development of the infrastructure and the software needed for data intensive science. At the same time our industrial collaborators can develop commercial products used by big science as well as other customers. Collaborations with industry are being established as part of ECHEP (see letters of support). As a single point of contact for particle physics applications, we plan to invite industrial collaborators to our workshops to present our use cases and discuss possible technological solutions. Interactions have been established with (i) StackHPC, a Bristol based software company, already working with IRIS (UK) on Scientific OpenStack, has agreed to partner in data management. (ii) Codeplay, an Edinburgh based software company with GPU interface products (using SYCL). (iii) Intel with interest in software development using their oneAPI portability interface. (iv) Under current plans, NVIDIA is willing to be involved in the project in order to support the development of software, where both performance and portability are critical metrics. (v) London based, Maxeler Technologies, currently working on joint projects with CMS and DUNE experiments in accelerated computing, (vi) Xilinx as the leading provider of FPGA technologies that are currently used by ATLAS and CMS in their trigger systems.

Outreach and communication: Communication of results will be achieved by reporting at specialised conferences, such as the Computing in High Energy Physics (CHEP) conference series, the yearly Workshop on Intelligent Trackers (WIT), Advanced Computing and Analysis Techniques in Physics research (ACAT) and publications in the Springer journal "Computing and Software for Big Science". We also plan to contribute to conferences with a wider participation such as the IEEE series on Scientific Computing as well as events organised by the Alan Turing Institute and the Software Sustainability Institute.

Data from particle physics experiments is used as part of the Physics masterclass, which is organised throughout Europe every year in the Spring. Several of the groups in this proposal are involved in this masterclass and can disseminate insights into the importance of software in Science to 16-18 year old school students.

2 Project organisation

The project organisation is structured to facilitate efficient collaboration across HEP experiments in the UK, in order to effectively develop the common software tools needed. In this proposal, we identify high-priority cases that are outlined in the work package sections below for which we ask for funding. A diagram showing the project organisation is presented in Fig. 4.

2.1 Project Management

The organisation is centred around six work packages, of which one is devoted to project management. The overall management of the project is carried out in WP0 by the project leader and deputy, with the help of an agile project manager as outlined in Section 3. This is appropriate given the size of the project with a maximum of 4 FTE of effort allocated at any given time and allows to streamline its management, which is at a level of 0.4 FTE total.

The remaining work packages cover areas that have been identified by HSF, ECHEP and the community as high-impact and under-resourced. These are: Data management (WP1), Generators (WP2), Simulation (WP3), Trigger/Reco (WP4) and Analysis (WP5). These are described in detail in Section 3. Each Work Package will have a responsible in the form of a Work Package coordinator. Together with the project management, they will ensure that the work is reported on regularly, adheres to the code of conduct[2], and is delivered on time. The management will liaise with the institute PIs and work package coordinators to monitor the progress of deliverables, and should a change in scope or direction be desirable for a given work package, will work to redefine the outcome and consequences for future expansion. An **advisory board** was established during the preparation phase of this proposal with the dual purpose of providing oversight and ensuring community engagement in the project. The advisory board is composed by the UK spokespersons for the major LHC experiments: ATLAS (**Farrington**), CMS (**Shepherd-Themistocleous**), LHCb (**Gershon**), a DUNE representative (**Barker**), the PI of GridPP (**Britton**) and the IRIS-UK Scientific Director (**Clarke**).

The work proposed is part of an international initiative, and such interaction with **external collaborators** is needed as indicated in Fig. 4. The main points of reference are the WLCG and HSF organisations which are charged with the international organisation into which the work proposed fits (see letter of support). Interactions with national initiatives in the USA and elsewhere will also be crucial to the success of this work. We will profit from collaboration with **industry partners** (see letters of support) to ensure the solutions delivered follow the latest standards and leverage as much as possible existing technologies.

[2] https://www.contributor-covenant.org/

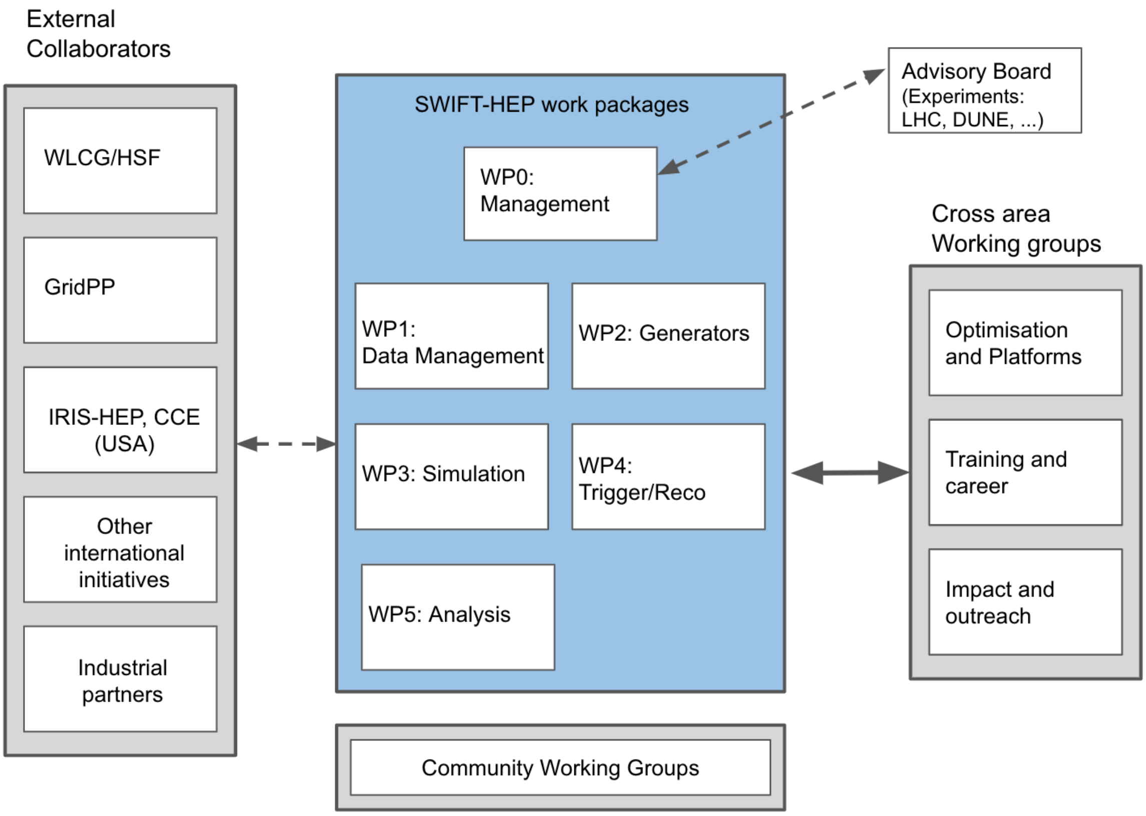

Figure 4: Organisation chart of SWIFT-HEP

2.2 Software management tools

All code developed will be open source licensed and managed through a suitable public collaborative platform such as GitHub or GitLab. These industry standard platforms will also be used to write and publish documentation for the API, usage, and technical notes, following the recommendations proposed by the HSF on best practices[3] for software projects. In particular all the code developed will adhere to the Open Source Initiative and will be licenced according to the HSF specifications.

Each project will have a Continuous Integration system set up to build and test each new development using appropriate tools for the project's programming language and use. These builds and tests will either run on the compute hosts provided by the platform, or for those projects requiring dedicated hardware (e.g. GPU/FPGA), the Software Development Platform described in Section 4.

To assist longer running or detailed benchmarking and validation testing, binary packages, including Docker/Singularity images and publication using the CERN VM file system (CVMFS) will be built and published on a regular basis. Images/CVMFS are the favoured implementation here to provide reproducible environments that can be quickly and easily deployed for further development, testing or integration on a wide range of platforms.

[3] https://hepsoftwarefoundation.org/technical_notes.html

2.3 Community engagement

Community engagement will be maintained by continuing to convene **working groups** aligned with the five WPs. These working groups will build on those established by ECHEP and provide a focus for the community to discuss common topics across experiments. It is expected that the people working on deliverables in each of the WPs will contribute to (and in some cases lead) these working groups. The coordinators of these working groups were appointed early in 2020 as part of ECHEP and are distributed both geographically and across experiments. Table 1 lists the people currently in these roles, with their university and, where possible, experiment affiliations. It is expected that a rotation mechanism is established for these roles, depending on candidate availability. A **coordination team** with the working group coordinators (area leaders in ECHEP) is already established and is expected to continue meeting monthly to steer the work of SWIFT-HEP.

Working group	Coordinator	Coordinator
Data management	D Colling (Imperial)	A Dewhurst (RAL)
Event Generator	M Schoenerr (Durham)	A Buckley (Glasgow)
Simulation	A Davis (Manchester)	B Morgan (Warwick)
Reconstruction & Trigger	S Martin-Haugh (RAL)	C Jones (Cambridge)
Analysis	L Kreczko (Bristol)	E Rodrigues (Liverpool)

The commonalities across the work packages are exploited by the creation of **cross-area working groups** that provide a meeting point to discuss common topics such as optimisation and portability. Training, career development and outreach have a very important role in this project. One of the main goals of the project is to create new skills in software development and artificial intelligence that would benefit the STFC science and society as a whole in the future decades.

This project will also provide a platform for the wider UK HEP community to discuss software development and common interests and as an interface to the software industry and the software community in the UK and worldwide, as such generating impact for the STFC investment.

2.4 Evolution towards SWIFT-HEP phase-2

The work package structure presented in this proposal is designed to efficiently deliver outputs on the proposal timescale, but it is also intended to scale up to a larger future project. Increased FTE can be assigned to the WPs outlined in Section 3 to expand the scope of the project to the ideal envelope of £2M/year. This will see the project evolve from the proof-of-concept studies proposed here to a fully-fledged deployment ready for the start of data-taking at the HL-LHC and next generation neutrino experiments. The current project management structure entails about 10% of the overall project dedicated to management.

A larger project will require an increase of management with work package managers to be funded at the 10-20% level to ensure an expanded list of milestones is met. The present structure of working group coordinators will be developed into a management structure for the second phase of SWIFT-HEP. Such a larger project will require a full time project manager and a monitoring of the progress of the deliverables and their update on a quarterly basis.

In the current proposal software management tools that are standard for HEP will be used. These tools are all open source tools industry grade tools that allow for software version control, continuous integration, testing as discussed in section 2.2. The tools used are currently maintained centrally by CERN, it is expected that with a larger project a participation effort will have to be allocated to maintain and customise this software infrastructure.

3. Detailed project plan

For each WP a table with the staff for which funding is sought is provided. The tables indicate the role of the staff, the institute where they are funded and the number of person months (**PM**) requested in FTE terms. The start and end months are also indicated in the table and reflect both the effort profile needed within the project and the availability of staff. Staff are expected to be funded at the 50% level by SWIFT-HEP, with the remaining 50% funded by an experiment grant to conduct work specific to such experimental collaboration. This ensures transfer of knowledge and expertise from the experiments to SWIFT-HEP and vice versa. Further details on the funding request and its profile is provided in Annex 1 (see separate attachment) , as per the PPRP guidelines.

Where staff names are known they are indicated in the tables. Where names are not indicated, universities will provide staff from pooled resources (in the case of RSE) or as new hires or by extending existing staff. Sources for the matched funding for the non SWIFT-HEP part of posts have been identified by universities and RAL.

Each WP provides a list of deliverables, highlighted in bold, with a timeline to achieve them. A summary of these deliverables is given in Section 5, along with a Gantt chart. The corresponding risk register is provided in Annex 2 (see separate attachment).

3.0 WP0. Management

Institute	Role		Start	End	PM	Name
Sheffield	Project Leader	WP0	1	36	6	Costanzo
Manchester	Dep Proj Leader	WP0	1	36	3	Fitzpatrick (*)
STFC Lab	Project manager	WP0	1	36	4	

(*) Provided at no cost to the project

Central to the delivery of the SWIFT-HEP project is the management team which will ensure deliverables are met and the project management plan is followed. This work package contains a project leader (**Costanzo**) and deputy (**Fitzpatrick**) and a project manager that will be funded at the 10% level to provide project management of SWIFT-HEP. The project manager will be recruited among the existing team within the STFC laboratories and will ensure the STFC project management guidelines are followed. We will use an agile methodology to ensure deliverables are deployed as soon as they are met, rather than cascading down until a final product is delivered.

The management team will prepare monthly reports for the advisory board to review the work carried out by the team and to ensure the project proceeds according to plan. An initial risk register is available in Annex 2 to establish the main risk factors at the start of the project. The project plan and risk registers will be monitored and updated typically every 3 months for review by the oversight committee to take into account factors such as:

- Institute level factors. Several institutes contribute to the proposal in order to leverage expertise that is available across the UK. Staff recruitment and retention issues will be proactively monitored and the deliverables will be rebalanced to minimise the overall impact on the project
- Change of circumstances factors. The evolution towards technical design reports will necessarily lead to design changes that will impact the expected development cycle requiring upscoping or descoping some of the deliverables
- Change in external factors. SWIFT-HEP is part of an international collaboration. External factors, such as a change in timescale and scope of the technical design reports, will have to be reflected in the project organisation

This WP will also deliver the UK contribution to the technical design reports at the end of the project funding, and is responsible for the preparation of the second phase of SWIFT-HEP

D0.1 Define the UK contribution to the HL-LHC software in the HSF and ATLAS/CMS technical design reports (Month 31)

Strong UK participation in the TDRs for the HL-LHC software is the main goal of the first phase of SWIFT-HEP. The work defined in the TDRs, expected in Q4 2023, will be executed in the years 2024-27 towards deployment of a software upgrade.

D0.2 Define a project management structure for SWIFT-HEP phase-2 (Month 24)

SWIFT-HEP is a new way to deliver software for the HEP community. The management organisation defined in this proposal is designed to scale up to a bigger project to deliver software over the next decade. Management practice will have to evolve to reflect this. The international landscape will also change as more funding agencies start software development projects, as advised in the European Strategy and by LHCC. Defining the SWIFT-HEP project management structure will also influence the international evolution of this research support area.

3.1 WP1. Intelligent data and workflow management

Institute	Role		Start	End	PM	Name
RAL-SCD	RSE	WP1.1	1	30	15	Chadwick
Brunel	RSE	WP1.1	19	24	6	Lopes
Imperial	RSE	WP1.2	1	30	15	

The purpose of this work package is to optimise the use of the diverse Digital Infrastructure available to HEP experiments in the UK. HEP experiments record and produce more than 100 PByte of data every year, and is expected to increase to more than 1 EByte/year in the next five years. These data are processed, reduced and served to the users using hardware infrastructure provided by WLCG and funded and operated via GridPP in the UK (Fig. 3). The use and deployment of data management software tools allow efficient access to the data. This efficiency results in (i) improved geographical distribution of data at Grid sites with reduced caching, hence requiring less storage to be provisioned, (ii) different quality of service storage to be used, reducing overall costs and (iii) a faster analysis turn-around time resulting in an increase in productivity of the experiments' analysts. This last aspect is further explored in WP5 at the end of this Section.

Currently, different compute and storage resources are treated uniformly[4] in a way that can be characterised as being the "lowest common denominator". Efficiency gains can be achieved by using the best available hardware for each task in a workflow. We will enable this by developing a UK data-lake (WP1.1) which will federate and make use of the diverse range of storage available in the UK. The data-lake will be based on the Rucio[[5]] data management tool. We will also develop a workflow management system (WP1.2) based on DIRAC[[6]], capable of utilising the features and functionality of this data-lake to efficiently run the complex payloads of the experiments. This work will also greatly benefit the smaller but rapidly growing community of non-LHC particle physics experiments and the wider STFC communities using these resources through IRIS.

The communities from different experiments or projects using the resources of the digital infrastructure are often referred to as a Virtual Organisation (VO) and these are used interchangeably in this application.

WP1.1 Intelligent Data Management

This task will focus on the management of experiment data between UK Sites. We will use Rucio to build a UK "data-lake" which aims to reduce the overall effort required by the experiments and infrastructure to manage data and make more efficient use of the available storage. Ultimately this will make it easier for users to process large datasets.

Currently HEP experiments manage the placement of their data across the sites that support them around the world. This means not only that each experiment must test and validate each site individually but they also need to be aware of quirks and differences between sites. In this context the aim of the data-lake is to move the lower level data management functions down a layer from being managed by each individual experiment to the infrastructure level.

D1.1 Setup a UK data-lake prototype with four sites and compare metrics with the current system (Month 1-12)

The first task is to set up a UK data-lake. This will build on the DOMA prototype, with the intention that there will be one data lake per region/country. The UK data-lake will use the Rucio instance that is already available at RAL (and being developed with IRIS funds). In this first stage the data-lake would closely follow the generic WLCG plan[[7]]. This can be broken into three roughly equal steps.

- Configure the core sites that have permanent data storage. To ensure adequate test coverage we will use at least four sites, preferably using different storage technologies.
- Configure additional sites to access data. As well as the core sites, we will configure at least one site with state-less storage (i.e. a cache), one site with no local storage

[4] There is some distinction between tape and disk, but very little distinction after that.

[5] Barisits, M., Beermann, T., Berghaus, F. *et al.* Rucio: Scientific Data Management. *Comput Softw Big Sci* **3,** 11 (2019). https://doi.org/10.1007/s41781-019-0026-3

[6] A Tsaregorodtsev et al, "DIRAC: a community grid solution", J. Phys.: Conf. Ser. 119 062048, 2008

[7] https://docs.google.com/document/d/1-0OliZcIcC3Z4hRxm9NgR7kVaOBhdg_XImiAFUxBpgw/edit

at all and one external site such as a Cloud Compute or an HPC to access the data-lake.

- Create a series of metrics which will measure: Overall time for [data analysis] tasks to complete, storage used, bandwidth used, CPU efficiency of tasks and operational burden (sites and experiment). This will allow comparison with the current infrastructure as well as being able to measure the improvements that come from subsequent tasks.

D1.2 Implement additional Quality of Service information in the data management (Month 13 - 18)

UK institutes run a variety of different storage technologies, each with their own strengths and weaknesses. In the past the only difference in Quality of Service was between Disk and Tape. We want to add in two major new features:

- High performance (i.e. SSD) storage. Some tasks (e.g. analysis) are becoming increasingly I/O bound and it makes sense to temporarily place the data on high performance storage. Not only does Rucio need to understand the concept of a high performance storage endpoint it also needs to be able to shunt the data into and out of it quickly enough to support the data intensive workflows.
- Reliability of storage (i.e. possibility of losing data). We aim to mitigate the risk of data loss by adding the concept of site reliability to Rucio. For example, as data ages the impact of file loss reduces so after 1 month moving data between new hardware and some that is 5 years old (with a higher chance of failure) could make sense.

D1.3 Produce recommendations to sites on how to optimise data access and deploy stateless storage (Months 19-24)

The UK benefits from excellent low latency network connectivity between institutions. This provides significant opportunities when it comes to data access. In many cases computing resources at one site will have excellent access to storage at others. For other sites a state-less storage service may significantly improve CPU efficiency with a minimal increase in operational effort.

D1.4 Creation of a [Virtual] Analysis Facility (Months 19-30)

As the amount of data grows, the more likely it is that something will go wrong when trying to process it all. To process datasets in their entirety users can spend a long time chasing up a small number of problematic jobs. The idea of an analysis facility is to have dedicated sites optimised to run user analysis jobs. Within the data-lake model, an analysis facility could actually be made up of multiple sites. This task will be heavily influenced by the earlier work being done in both the data management and workflow management parts of the work packages. This task will provide support to the work described in WP5 to ensure the analysis services developed have a solid infrastructure to rely on. The main goal of an analysis facility is to increase the reliability of the resources used, for example by only using sites with more reliable storage or pre-caching the data for faster access when jobs start. It will also provide tools that make things easier for users, such as easier access (via a web browser) to their job output data or relevant logs. The increased reliability will directly translate in faster analysis turn-around time and more efficient use of existing grid resources.

This work will be further exploited in WP5.1 and will foster a collaboration with the ServiceX team within IRIS-HEP in the USA.

WP1.2 Workflow Management

The workflows of HEP experiments can often be complex with different elements being suited to different forms of data storage and access - this is true for both simulation and analysis. Each of the LHC experiments uses their own workflow management systems which they have developed in house and in collaboration with others. The LHCb experiment uses DIRAC as their workflow management system and GridPP has run a DIRAC service for all non-LHC experiments since 2013. This service currently caters for 10 non-LHC HEP experiments and a handful of non-HEP STFC communities (through IRIS) and so it is the DIRAC workflow system that we will develop to use the functionality of the data-lake.

D1.5 Build a generic pilot logging system reporting to a central service (Months 1-9)

So called "pilot jobs" are used pervasively in the WLCG**.** On the GridPP DIRAC service, Pilot logging is currently limited to a subset of supported compute elements, there is no support for clouds,etc. We will build a generic pilot logging system that allows all pilots to report to a central service in a multi-vo way. This will allow VO administrators to view logs for their pilots and debug resource matching, ultimately improving job efficiency. We will add suitable instrumentation and logging to this module, with a view to improve the data staging efficiency.

D1.6 Extend and test the DIRAC workflow to allow usage from middle-sized VOs (Months 10-21)

DIRAC provides a basic Workflow management system. We will extend this to match the requirements of medium-sized VOs which could greatly increase the efficiency for large production campaigns by automating routine tasks; this can include data transfers & varied-architecture execution steps.

D1.7 Implement dynamic load management in DIRAC and allow runtime job adjustment (Months 22-27)

We plan to upgrade the DIRAC configuration based on an outdated database system (BDII). We will add support for hierarchical alternative database sources (GOCDB, JSON) to allow configuration of newer hardware types (such as GPU)

Currently, once a job is submitted to DIRAC it is immutable. We will add an option and tooling to partially remove this restriction. This would allow jobs to be retargeted after submission time, paving the way for more dynamic load management.

D1.8 Implement high-level commands to the DIRAC data handling tools (Months 28 - 30)

The built-in DIRAC data tools were never really designed to be user facing: They were built for development purposes, with a view of having a separate frontend. Now DIRAC has become multi-user, a number of user-interface improvements are needed. This will open up the possibility of adding advanced data management commands in the higher-level tool that are not currently possible with the basic one.

3.2 WP2. Event generators

Institute	Role		Start(*)	End	PM	Name
UCL	PDRA	WP2.1	7	30	12	Gutschow
UCL	RSE	WP2.1	10	18	3	
Warwick	PDRA	WP2.2	5	17	6	(**)
Warwick	RSE	WP2.2	5	17	4	

(*) The start month is driven by the availability of Gutschow at UCL and of Morgan as the Warwick PI

(**) Provided at no cost to the project

The UK is world-leading in Monte Carlo event generator development, with key authors of the programs Sherpa, Herwig and Pythia8 all based here. The UK hosts also the main developers for widely used support tools, such as LHAPDF, EvtGen, HepMC, and Rivet. As well as making major contributions to and leadership of the ATLAS physics modelling group, the UK is ideally positioned to drive future developments to address the computing problem faced by the LHC experiments.

MC event generators act as a bridge between developments in state-of-the-art field-theory calculations and experimentally useful particle-level simulations. These programs are subject to mixed incentives: their main developers are theorists and as such are primarily motivated to produce “theory” outputs such as phenomenology papers; meanwhile, their main users in terms of computing resources are the experimental collaborations for whom computational efficiency is increasingly as important as theoretical precision. This experimental need has been brought into sharp focus by studies of the CPU budget for the HL-LHC programme: MC generation alone is expected to oversaturate the available processing resources by a factor of 2 in the absence of major efficiency improvements. In this WP, several deliverables are defined that address this challenge:

WP2.1 Hard-scattering process performance study and technical optimisation

D2.1 Detailed report on CPU-performance profiling of high-precision Monte Carlo event generation to identify bottlenecks for cost-driving setups used by experiments. (Month 7 to 10)

Building on similar ongoing efforts across the experiments to pinpoint limiting performance issues in the current event generation model, this work will provide a complementary assessment of the setups used by ATLAS to generate very large Monte Carlo samples with Sherpa, such as W+jets and top-quark pair+jets. Preliminary ECHEP studies indicate that more than 40% of the CPU cost of ATLAS high-precision event generation with Sherpa is going into the evaluation of parton densities, as well as transcendental functions such as logarithms or exponentials. The report will identify the part of the code where optimisation is expected to return the highest improvements.

Software expertise can also be usefully applied to perform a detailed profiling of the efficiency of automatically generated code, e.g. by the MadGraph5 and Sherpa matrix-element builder modules.

D2.2 Release of optimised LHAPDF version, aiming to reduce cost of PDF evaluation using knowledge gained from D2.1 (Month 11 to 18)

Drawing on the preliminary ECHEP studies and the profiling results from D2.1, the widely used LHAPDF[8] library will be optimised with the aim of a major new package release. The development of a caching system and improved event-generation call strategies may be able to save tens of percent of event generation CPU. The performance impact will be examined for a number of representative setups used by the experiments.

D2.3 Generator code release (e.g. Sherpa, Herwig) incorporating technical optimisations and refinements for usage by experiments (Month 19 to 32)

This work is expected to be primarily informed by the profiling studies from D2.1 with the aim of major software contributions to general purpose event generators used by the experiments. Capitalising on software proficiency, the evaluation strategies are to be refined by providing higher-precision interpolation routines, e.g. using Chebyshev polynomials, with the aim of reducing CPU costs for expensive high-precision calculations, such as V+jets and top-quark pair+jets production.

In order to exploit the improvements from D2.2 to its fullest, the PDF access in multi-leg generators such as Sherpa will be adapted to suit the newly developed LHAPDF call strategies, in particular for setups using a large set of PDF variations to ensure efficient scaling up to approximately one thousand variations.

In addition, professional software expertise could help pave the way towards prototyping the rewrite of core modules (e.g. matrix-element generators) with modern compute architectures in mind.

WP2.2 Development of efficient hadronization models

The hadronization component of event generators is based on non-perturbative physics models, which by necessity are less rooted in *a priori* theory than evaluation of the hard-scattering process. Hadronization modelling uses stochastic methods for the reconfiguration of colour-strings or preconfinement clusters, and their eventual resolution into colour-singlet hadrons. These mechanisms are computationally expensive to the extent of being simulation bottlenecks, particularly for LHCb but increasingly at other experiments pursuing expanded B physics programmes, and are well-suited to development by non-theorists. The particles produced in the hadronisation process can be passed to "after-burner" software packages such as EvtGen, which is particularly crucial to model the properties and decays of heavy-flavoured (beauty and charm) hadrons. The EvtGen codebase still retains the essentials of its original development 20 years ago, and must be modernised to benefit from the opportunities of parallelisation on modern architectures.

The first two deliverables are to be performed on the Pythia 8 generator, both because it is the main MC generator in use by LHCb and hence will have the largest immediate impact, and also because the codebase is relatively approachable by a non-expert. Moreover, the project will benefit from the Monash-Warwick alliance, which will provide matching funding of

[8] https://lhapdf.hepforge.org/

PDRA effort as well as access to the lead Pythia 8 developers in Monash University. These developments will serve as prototypes for equivalent future work in Sherpa and Herwig. The third will prevent EvtGen becoming a limiting factor in the improved generation chain.

D2.4 Develop a biased hadronization mechanism for Pythia 8 (Month 1 - 6)

New biasing “user hook” mechanisms will be developed in Pythia 8, to enable efficient generation of specific primary hadron species, using probabilistic weights to recover physical distributions, to replace the current inefficient rejection-sampling approach. This is primarily targeted at expensive LHCb heavy-flavour samples, but will have side-benefits for b-physics studies in ATLAS and CMS (in which the UK has a leading role).

D2.5 Develop a vectorised colour-reconnection module for Pythia 8 (Month 7 - 12)

The most sophisticated Pythia 8 colour-reconnection model will be ported to opportunistically use accelerator hardware, via Sycl or OneAPI, to enable use of its improved physics modelling by LHCb. A prototype version will be written, validated, and benchmarked.

D2.6 EvtGen modernisation (Month 1-12)

The codebase will be modernised to comply with the latest C++ standards and to facilitate multithreading. Legacy dependencies on obsolete or unmaintained packages will be replaced with suitable alternatives. The updated software will be released as a major version change, and a new user manual will be published.

3.3 WP3. Simulation

Institute	Role		Start	End	PM	Name
Warwick	Senior RSE	WP3.1	5	36	16	Morgan
Manchester	PDRA	WP3.2	1	12	6	Davis

Simulation covers the transport of primary particles through a detector geometry, modeling their interactions with its materials and components. The Geant4 Toolkit [9]is the main particle transport code used in High Energy Physics (HEP), and is widely used in other fields including medical, space, and bio-physics. Within HEP, and especially the LHC, there is heavy use of so-called ``fast" methods, offloading computationally expensive parts of the transport/physics (e.g. showering in EM calorimeters) from Geant4 to less CPU intensive parametric methods . Division into ``full" and ``fast" simulation is artificial - detectors require the use of Geant4-based ``full” simulation to develop the ``fast" component(s) and ensure that they reproduce detector observables to the required accuracy.

Simulation accounts for a large amount of CPU usage throughout HEP. For ATLAS in 2019, 35% of experiment wide CPU usage was simulation, with 85% of that being simulation of the EM calorimeter. In LHCb, roughly 50% of simulation time is dominated by optical photon transport in the RICH detectors, and 30% in calorimeter response. For the LZ experiment, optical photon transport accounts for roughly 95% of all CPU usage in MC, necessitating faster optical photon propagation options. With the much larger datasets to be collected in the future across HEP, assessing CPU usage and developing better use of resources is a top priority.

[9] https://doi.org/10.1016/S0168-9002(03)01368-8, https://doi.org/10.1016/j.nima.2016.06.125

As part of the upcoming HL-LHC Computing Review, the HEP Software Foundation prepared a document reviewing common software tools and the challenges faced in preparing for the increased demands on available processing resources.

Detector Simulation is a key area here (alongside the upstream Physics Generators), and while focused on the LHC, the challenges outlined will be faced by any experiment using simulation and needing to increase their throughput, sample sizes, or make optimum use of the increased availability of HPC/Accelerator platforms. Ongoing developments and R&D on new techniques for improving simulation performance break down into three layers:

- Ongoing development and releases of the Geant4 toolkit, including optimization for use across the community, and R&D on refactoring to ease use of SIMD/Accelerators
- Offloading parts of overall simulation application to ``fast" techniques, such as parameterisation of EM showers, trading some accuracy for improved throughput
- R&D on particle transport/physics process algorithms suitable for offload to other architectures such as GPUs and FPGAs

WP3.1 Core Geant4 development

The Geant4 collaboration has started a R&D programme to modernise the Geant4 codebase for efficient use of modern architectures and to prototype data models and algorithms suitable for SIMD/Accelerators. This R&D programme is still in the initial phase and is expected to last a few years. UK participation in this programme started thanks to the ExCALIBUR programme with involvement in development of CUDA based prototypes for electromagnetic physics processes. The work we propose here enables us to participate in this initial R&D phase, with the intention of increasing our contribution once the R&D is better defined based on the prototypes developed.

D3.1 Develop a prototype Geant4 application for calorimetry with CUDA-based EM Physics (Month 5 - 24)

A CUDA-based prototype for electron/positron/photon transport in a single volume with a limited description of electromagnetic physics will be available from the ExCALIBUR-HEP at the start of this project. Building on that work and its collaborations with the Geant4 R&D Task Force, HSF, and US Celeritas projects, this task will integrate the prototype in a Geant4 application modelling a simplified tracker/calorimeter HEP detector with scoring of energy depositions in the calorimeter component(s).

Extension to more complex geometries with multiple materials will allow the throughput and memory impacts of their interplay with physics processes (e.g. multiple scattering) and computation (data table lookups) to be studied. This setup will also allow improvements to the CUDA-based physics algorithms to be developed, including investigations of Ray Tracing methods, as geometry/material effects can be included.

Modelling a combined tracker/calorimeter detector will be used to prototype hybrid Host/Device computation: all particles in tracker on the Host, electrons/positrons/photons in calorimeters offloaded to the Device. This maps to electromagnetic calorimeters as the highest CPU user in typical HEP simulations, and the suitability of its physics models for computation on Accelerators. Nevertheless, Host-Device offload/onload is a critical item for effective Accelerator utilization, so workflows for optimal use will be a key area of investigation. The hybrid computation permits both Host-only and Host-Device versions of the application to be built, allowing later physics validation and overall throughput comparison in D3.2.

D3.2 Physics Validation and Optimisation of the prototype CUDA application (Month 25-36)

The prototype Geant4 application developed in D3.1 will be run in both Host-only and Host-Device modes to generate metrics comparing physics (via calorimeter observables) and event throughput performance. In both cases, the Host-only output data provides a suitable baseline as it uses standard Geant4 geometry and physics models.

These metrics will be used to study and optimise the Host-Device application for physics accuracy and event throughput. The outcome of this work will be a report with a preliminary evaluation of the accuracy and performance of the Host-Device application versus its Host-only version, together with identified issues and directions for future development.

WP3.2 Geant4+Opticks Example Application

The transport of large numbers of optical photons within the Geant4 framework is one of the more CPU intensive processes currently dealt with by HEP experiments. For example, photon transport in the RICH detector accounts for 50% of LHCb simulation time[10]. With only a single particle type and a relatively simple set of physics processes, optical photon transport algorithms are well suited to parallelization on Accelerators. Such work has been started with the Opticks package[11], which enables offload of optical photon generation, transport, and scoring from a CPU Geant4 application to an NVIDIA GPU with NVIDIA's OptiX Ray Tracing Engine handling the transport and physics. By keeping optical photons on the GPU and only transferring detector hits back to the Host, Opticks has demonstrated speedups over Geant4 that scale linearly with the number of CUDA cores for the case of the JUNO detector[12].

Opticks can be of potential benefit to any experiment where optical photons dominate the simulation time, and LHCb, LZ and DUNE are investigating its use. Its use of Ray Tracing methods is also of interest for application to charged particle tracking, with work started under the Geant4 R&D Task Force with contributions from ExCALIBUR-HEP. However, Opticks remains a young project and requires resources for both ongoing maintenance and further development and application.

A key goal of Opticks, and adopted under the Geant4 R&D Task Force, is to become an easy to use and standard way to perform optical photon transport in Geant4-based simulations. The work we propose here will contribute to this effort, with the intent of building experience and allowing future contributions to longer term Opticks developments in geometry and physics.

D3.3 Develop an example Geant4 application with optical photon transport offloaded to Opticks (Months 1-12)

A project has been started under the Geant4 R&D Task Force to create a Geant4 example application demonstrating offload of optical photons to a GPU using Opticks, with non-optical processes on the Host. In collaboration with the developers at FNAL and the Opticks project, this task will contribute to the development, testing, validation, documentation, and release of this application as part of the Geant4 Toolkit.

This work is expected to have strong synergy with D3.1 via the common need to carefully manage Host-Device data transfers and compute. Experience gained with the use of the

[10] https://inspirehep.net/literature/1760902
[11] https://doi.org/10.1051/epjconf/201921402027
[12] https://iopscience.iop.org/article/10.1088/1742-6596/898/4/042001

underlying Ray Tracing algorithms and the NVIDIA Engine will also inform investigations on applying these techniques to charged particle transport.

3.4 WP4. Reconstruction and Trigger

Institute	Role		Start	End	PM	Name
RAL	PDRA	WP4.1	1	24	12	
RAL	PDRA	WP4.2	9	24	8	
Sussex	PDRA	WP4.1	1	24	12	Trovato (?)
Warwick	RSE	WP4.1/4.2	10	36	14	Martin

Core components of real-time and offline detector data processing are beginning to benefit from access to different architectures as part of a heterogeneous system. Triggering systems operate in dedicated farms in a more controlled environment, thus are the first place where this move away from single-architecture computing systems is taking place. FPGA and GPU technologies are starting to be exploited in triggering systems to “accelerate” or even replace conventional CPU computing. These heterogeneous computing systems have the potential to address the more general aims of SWIFT-HEP, namely that attempting to scale current offline CPU infrastructures to meet HL-LHC and other future HEP computing needs.

The UK has considerable expertise built up over many years and in many projects on triggering systems which make use of FPGAs and GPUs (as well as CPUs). This will be exploited to develop the use of accelerators based on these technologies and the code portability which is required to make effective use of heterogeneous computing systems.

A recent example of effective use of GPUs is the LHCb implementation of its entire first trigger stage, on GPUs alone. Through inclusive beauty and charm identification this stage reduces the event rate from 30MHz to 1MHz. This implementation required the development of a new, cross-architecture application framework that supports cross-event parallelism, necessary to fully exploit the parallel potential of modern GPUs, and also a custom memory manager to accommodate the memory restrictions inherent in the current generation of GPUs. It was shown that the GPUs were able to implement the full physics functionality present in the CPU implementation, while demonstrating a significant cost saving. LHCb’s developments have so far resulted in a native CUDA implementation of a trigger framework. At the time this had its advantages, but in order to leverage a truly heterogeneous hardware environment, heterogeneous software will be required.

In this work package, we will develop and benchmark heterogeneous computing solutions for trigger and reconstruction algorithms in order to inform and advance the adoption of this approach for future computing systems. In WP4.1 specific knowledge of FPGA-based accelerator best practices are developed, in synergy with WP4.2 where this knowledge is incorporated into more a general heterogeneous computing study that permits the same algorithms to run on CPUs, GPUs, FPGA, and hybrid/mixed-architecture environments.

The initial benefits of these developments will be in computing farms for trigger systems because these structures are directly under experiment control. The portability of algorithms between online and offline systems will be key to successfully exploiting accelerators in either type of system. In a second stage these portable and scalable solutions will enable the optimal use of more generic data centres where GPU and FPGA resources are available.

D4.1 A report on benchmarking tools and practice to be used to evaluate the performance of the algorithms developed in WP4.1 and WP4.2 (Months 10-15)

The prototypes described in this WP as a whole will be evaluated against a set of benchmarks prepared by **Martin**.

WP4.1 Deployment and benchmarking of tracking and vertexing algorithms on FPGA accelerators

The recent development of FPGAs as accelerators provides a significant opportunity for the UK LHC community to exploit its long-standing expertise in using FPGAs for hardware triggers. This has traditionally required specialised firmware development, but hardware vendors are now providing increasingly automated High Level Synthesis (HLS) solutions. To ultimately be able to use FPGA accelerators as seamless devices in heterogeneous computing systems, HLS approaches must work effectively. The UK has considerable firmware engineer expertise and a growing level of expertise in the use of HLS. The proper exploitation of FPGAs provides the UK with an opportunity to capitalize on its expertise to develop the use of FPGA accelerators and adopt them in offline computing, as has happened in e.g. Microsoft data centres. A number of UK institutes have already fostered industrial contacts with FPGA vendors (e.g. Xilinx, Intel) and in this WP these contacts will be of benefit.

In order to perform R&D towards implementing algorithms such that they can be deployed in heterogeneous systems in WP4.1 we will implement and benchmark examples of realistic and representative algorithms on FPGA accelerator cards, including those algorithms already developed for the HL-LHC upgrades. The goal of this work is to identify the techniques required to enable algorithms to be deployed cost effectively both online and offline in computing environments that employ both accelerators and CPUs.

CMS has developed and deployed a tracking reconstruction chain in a realistic system of FPGAs as well as vertex finding in neural nets[13]. ATLAS has demonstrated that FPGAs using HLS solutions can replace the need for dedicated processing units in ad-hoc silicon devices (ASICs). This combined experience will be built upon to develop best practices in the deployment of algorithms of common interest to experiments that have tracking abilities. We will also work collaboratively with a PDRA from Imperial College drawing on their experience from the CMS trigger upgrade work. The challenge here is to be able to take advantage of the abilities of FPGAs to process data at high throughputs without having to explicitly write algorithms in hardware-specific firmware. The combination of expertise that exists in the institutes engaged in this WP is what is required to bridge the gap between highly tailored engineering and generic high-level software. This will enable techniques to be developed that will allow the exploitation of FPGA accelerators.

The first stage of the study will take existing algorithms and deploy them in FPGA accelerators making use of HLS, comparing with the performance in CPUs and with the VHDL versions of the same algorithms, documenting best practice for higher level language programming and documentation of experience with the vendor-specific development environments and tools. The close connections with industry, formed specifically with Xilinx and Intel will be exploited to enable these developments. We will use dedicated accelerator cards and cloud resources to gain experience in both approaches. Two FPGA boards[14] will be purchased at the beginning of the project. While it is too early to obtain a precise quote,

[13] https://iopscience.iop.org/article/10.1088/1748-0221/12/12/P12019

[14] https://www.digikey.co.uk/products/en/computer-equipment/server-acceleration-cards/986

this is expected to cost in the region of £10k. We expect further hardware resources to be available on loan, and add additional costs as a working allowance in case such resources need to be purchased or in case additional hardware is needed to utilise resources on loan/purchased.

In order to benchmark against CPU, GPU and FPGA technologies, the use of generic interface systems (OneAPI, SYCL) will be explored to enable algorithm execution in specific accelerators as well as more ambitiously in commercial cloud systems, resulting in a suite of heterogeneous algorithm implementations to be benchmarked in terms of overall cost/benefit.

The resulting deliverables are:

D4.2 Deployment and performance comparison of tracking and vertexing algorithms in FPGA accelerators for real-time and offline reconstruction (Months 1-15)

Good candidates are a classical type of track reconstruction algorithm, consisting primarily of a Hough Transform followed by a Kalman Filter, and a Machine Learning algorithm where vertices are constructed using reconstructed tracks. The work will be carried out jointly by the **Sussex PDRA** and **RAL PDRA1**, in order to exploit the expertise with ATLAS and CMS.

D4.3 Benchmarking of the prototypes developed on CPU and accelerators (Months 16 - 23)

The algorithms developed will be benchmarked to assess the cost and power performance of the heterogeneous implementation relative to the existing CPU-based solutions.

D4.4 Report on the use and performance comparison of generic interfaces (e.g. oneAPI) to interface the algorithms to the hardware. (Months 23-24)

WP4.2 Prototype of tracking and vertexing algorithms on heterogeneous compute platforms

While WP4.1 increases the level of expertise in the field on an entirely new architecture, in WP4.2, we build on this by incorporating and benchmarking both newer and more established architectures in a truly heterogeneous manner: The aim of WP4.2 is to deliver cross-architecture algorithm development and benchmarking using more generic algorithms that are suited to test the latest advances in heterogeneous computing.

In order to develop a cross-community benchmark a more general tracking implementation will be studied. This implementation will be based on FasTrack, a tracking algorithm that exploits machine learning techniques, and will be deployed on the universally accessible TrackML dataset that was released as part of a Kaggle challenge[15] in 2018. FasTrack was developed by **Emelyanov** at RAL and placed second in this challenge.

This algorithm will be ported to run on FPGA accelerators, building on the expertise developed in WP4.1, and GPUs. OneAPI will be tested for suitability for this task. At present OneAPI is under active development and requires significant R&D in order to determine if it is a suitable tool for future use: The OneAPI implementation will be benchmarked against the native implementation to determine the portability overhead and to study specific optimisations. It will be integrated with the ATLAS and CMS multi-threaded frameworks and we will compare throughput to native CPU-only and CPU + GPU solutions. The resulting software will be open-source and documented as part of the deliverables. This constitutes a

[15] https://www.kaggle.com/c/trackml-particle-identification

genuinely detector-independent methodology that can be used to test algorithms independently from the specific experiment design technology.

The resulting deliverables are:

D4.5 Open source tracking algorithm (FasTrack) implemented through OneAPI and benchmarked to the native implementation. (Months 9-24)

Carried out by the second PDRA at RAL, will require the porting of the existing code and delivery of the corresponding documentation.

D4.6 Performance comparison of FasTrack on CPU, GPU and FPGA technologies. (Months 25-36)

The performance will be compared in several cases and integrated with the ATLAS and CMS environments. **Martin** will develop this integration, benchmark against a set of use cases and write a final report.

3.5 WP5. Analysis systems

Institute	Role		Start	End	PM	Name
Bristol	PDRA	WP5	13	36	12	

The ultimate goal of particle physics experiments is to analyse the data and produce the final results that feature in publications. The production of high-impact results, such as the Higgs boson discovery or the measurement of neutrino oscillations rely on a data processing pipeline and several layers of software shown in Fig. 3.

The analysis work package concentrates on common tools that are needed to serve the data to the analysers and allow a fast-turn around analysis work. Furthermore higher job efficiency translates into a more efficient use of grid compute resources with fewer cycles lost due to misconfigurations and failed attempts.

Analysis tools concern a wide user base spread across multiple institutes, with many users working on an experiment for a period of 2-3 years to e.g. complete a PhD. As such analysis tools need to be robust, easy to use and provide user support and training. This WP concentrates on tools that allow serving analysis data in a fast and reliable way building upon the work proposed in WP1, and tools to streamline the analysis process by simplifying the complexity of the programming.

WP5.1 Caching of intermediate analysis results

The purpose of this work package is to minimize the use of computing resources through repetition. The typical analysis workflows are inherently iterative and analysts will often resubmit workflows with small modifications. By analysing the compute operations in an analysis and storing and managing intermediate analysis in a federated storage (e.g. a UK "data-lake" -- D1.1), operations can be skipped resulting in less computing resource usage and faster response times.

D5.1 Interoperability with UK data lake (Month 12-18)

Develop a local caching mechanism suitable for individual workflow steps in HEP. Both FAST-HEP and IRIS-HEP tools can use Parsl to distribute workflows, which also includes an access point for caching (App caching). Due to the complexity of HEP analysis this will likely require a custom memoization solution to compute the equivalence of individual workflow steps across analyses.

In order to facilitate caching at a meaningful scale, intermediate results have to be stored in a location accessible to all compute resources. Furthermore, access needs to be moderated per VO, group or even user level which is beyond this WP. The UK data lake project, however, is an ideal candidate for this endeavour. Close collaboration with the project will allow us to implement the optimal solution.

D5.2 Integration of caching mechanism with workflow management (Month 18-24)
With implicit sub-job level caching enabled (D5.1), individual workflow steps only need to be scheduled if no cache entry exists. To improve scheduling efficiency, and thus user turn-around time, an exchange mechanism between workflow tools (e.g. Parsl) and dynamic load management (here DIRAC, see D1.7) will be created. This will ensure that only non-cached parts of an analysis are submitted for processing, prioritising locations where cached results of the previous steps are present. Such an automated setup will not require any additional inputs from the users and will therefore reduce barriers to make full use of caching.

WP5.2. Portability of analysis code

D5.3 Per-site optimization of workloads (Month 25-30)
Typical grid infrastructure is very heterogeneous. Certain workloads (e.g. machine learning, iterative processing) are better suited for specific platforms (GPUs, NVMe storage, etc). We will implement a mechanism to choose the implementation detail based on the computing infrastructure present. This will require a dedicated set of information that needs to be present on the execute nodes (e.g in the form of HTCondor machine features) as well as the necessary mechanism in the analysis steps to select implementation detail at such a stage. Collaboration with the virtual analysis facility (D1.4) will be required to find the optimal exchange mechanism.

D5.4 Dedicated workload scheduling for the Workflow management (Month 31-36)
This deliverable builds upon the work of D1.7, D5.2 and D5.3. It will implement the introspection of workloads by the scheduling system in order to extract the ranking (i.e. preferred) of required computing infrastructure for each individual workload step. Dynamic load management (here DIRAC, see D1.7) will then be able to match available resources based on that information. This will not only provide users with the flexibility of specifying preferred and fallback resources, but also present a unique opportunity to automatically monitor (changing) computing resources requirements.

4. Cross activity working groups

4.1 Optimisation and platforms

Working group on accelerators and profiling

A common theme in several work packages in section 3 (WP4, WP3.1 and WP2.1) revolves around the deployment of software on multiple platforms and on its optimisation to improve the efficiency and portability of the code. Efficiency can not easily be quantified in a heterogenous system; the actual cost of a computing system depends on many aspects, market factors when procurement is made, financial model of the data centres used, and other parameters such as cost of networks required, error tolerance, maintenance and upgrade costs etc. Under all circumstances for a given workflow the aim is to minimise the compute time needed for a unit of work, allow a wide spectrum of architectures to be used, reduce total memory footprint, reduce the total I/O needed, etc. All these improvements have an impact on the overall computing costs and a handle on all these parameters will ensure that the overall computing cost is kept within a reasonable range to warrant the physics exploitation of experiments in the next decade (see Fig. 1).

Optimisation of software and its deployment on multiple architectures is a common topic of all work packages in SWIFT-HEP, and of the experiments software teams. Exchange of ideas and good practices both between work packages in SWIFT-HEP and the community in general is very important as profiling and compilation tools evolve and become more sophisticated. Hence a cross-activity working group will be established with monthly meetings to discuss the software engineering challenges connected with optimisation and portability.

Software development environment

The work proposed is based on the development of software and its optimisation on a variety of platforms and using different software tools. A central software development facility is needed to ensure people participating in this project have access to common hardware and tools such as SYCL and OneAPI instances. Access to a generic development environment can be achieved through the existing general use PPD cluster at RAL and the STFC cloud. This will allow all collaborators to access hardware and tools for compilation, testing and profiling on both CPUs and GPUs. Support and instruction will be provided by the PPD Computing team (specifically including C.Brew and A.Djaoui).

We expect to loan additional FPGA boards for testing from industrial partners, we have an indication that this is possible, and in fact a Xilinx system is currently hosted at the University of Sussex. The lack of FPGA boards to be used by the project is a potential risk for WP4 (R8 in the risk register, Annex 2). Similarly additional capacity may be needed to host the loaned boards (R9 in the risk register, Annex 2).

Large production jobs that will run for large scale validation of the software developed by SWIFT-HEP will use CPUs and GPUs already provisioned by GridPP.

4.2 Training and careers

Training and career development will be a key component of this programme. This will ensure that the direct participants in the programme have their skill-base enhanced, and more broadly that skills in this area are uplifted across the entire field, from PhD students to academics, via open access to and wide promotion of the training offerings.

The skills training will be offered in collaboration with STFC's Data Intensive Science (DIS) Centres for Doctoral Training (CDT) at UCL, who have been successfully delivering training in this area for 3 years. The CDTs were set up to enhance the training provided in software/computing to PhD students, with many elements also offered to non-CDT PhD students and RAs, this programme will build upon and expand these offerings to the entire field. The programme will be co-delivered by the CDTs and will include:

- Advanced courses: These will include software engineering, firmware engineering, recent language developments (C++/Python), programming for heterogeneous architectures, parallel/optimised scientific software, machine learning and tools for software development/testing/profiling.

- Industry led workshops: These will focus on exploiting the latest hardware developments (FPGAs, GPUs, CPUs, intelligent networks), delivered by Intel, NVidia, ARM and Mellanox.
- Software carpentry courses: Focussed on key software skills and the attendees will return as tutors in future iterations, building networks of experts and propagating knowledge throughout the field.
- STFC's DIS Summer School: STFC's annual summer school which hones key DIS skills, with significant industry involvement in the training.
- Industry placements: The opportunity to undertake a three month industry placement will be offered to the RAs/RSEs, using the CDT schemes. The placement will be funded by the industry partners and will be in addition to the time committed to this programme. Placements will enhance the PDRAs knowledge of the tools, techniques, work-flows and language used in industry, whilst building links with industry, promoting knowledge transfer and creating impact.
- Industry group projects: A group project undertaken for 1-2 days a week by a team of 3-4 PhD students for 3 months, which focuses on working with experts in industry to solve an issue they have. This develops similar skills to the placement and the PDRAs will be offered the opportunity to supervise one of these projects.
- Project management, leadership and supervision: Training in effectively managing large software projects and leadership training/mentoring, to provide the RAs/RSEs with the skills necessary to become future leaders in this area. The RAs/RSEs will also co-supervisor CDT PhD students who will contribute to the projects outlined in the programme, which will more closely align the two programmes.

All these elements have been successfully undertaken in the STFC's CDTs, with great success. The RAs and RSEs recruited by SWIFT-HEP, along with the wider community, will participate in the training programme and engage with the training offerings, and will see a significant improvement in their career prospects either within research or beyond. They will develop a much sought after skillbase which is essential not only to research, but also to industry, where there is a significant shortage of both data scientists and software engineers. With the ever increasing role that online and cloud services play in the economy, which has been accelerated by the Covid-19 situation, these skills will become ever more essential and in demand in the coming years. A large proportion of the PhD students and RAs in this field currently move to industry, and this training programme will help ensure the community is even better placed to fulfil some of this demand. This has been clearly demonstrated via a 60-70% success rate of RAs obtaining industry jobs after a placement or industry group project in the UCL DIS CDT.

4.3 Impact and outreach

There are several areas where impact from SWIFT-HEP can be generated. These are centred around (i) training (ii) technical engagement (iii) industrial and commercial engagement and (iv) outreach and communication. The details of these four pillars are discussed in section 1.5. During the execution of the project particular attention will be paid to impact via the four items above. An impact and outreach champion will be nominated from the communities and activities will be monitored throughout the project. An internal report on impact and outreach will be provided to the advisory board every quarter, with a discussion held at the workshops to ensure the whole community is involved in this particular item.

4.4 Workshops and community participation

This proposal builds on carefully selected work packages that are useful to the wide HEP community in the UK, in collaboration with the experiments and GridPP developers, as evidenced in Fig. 3. A community participation model was established in 2020, thanks to the funding of ECHEP and further the establishment of an ExCALIBUR design and development working group. Due to the lockdown in 2020, most of this activity took place remotely after a successful workshop in Edinburgh in February.

This collaboration work will revolve around:

- Working groups have been created (Section 2) and meet remotely every 4 weeks
- A general SWIFT-HEP (currently ECHEP) meeting to be held remotely approximately every month
- Management meeting to be held every fortnight to monitor the project progress
- Two workshops every year for two days to enable people to discuss common topics and ideas and to connect
- Hackathons on dedicated topics to be arranged on a need-to basis
- Travel for hands-on meetings and to allow short-term visits at different sites to collaborate on specific topics

5. Deliverables and finance summary

5.1 Justification of resources

The overall structure of the SWIFT-HEP proposal was organised to fit within the overall budget envelope of £1.2M agreed by Science Board at the April 2020 meeting. The costs reported are the costs to STFC at 80% FEC for university grants, with variation of overhead costs at different institutes. Details are available in the JeS forms attached to this proposal, with a summary provided in Annex 1.

This allows for funding about 4 FTE of effort across the three years of the project. This is considered a reasonable level of investment as it creates enough critical mass to establish ourselves as a key international player and will allow to build a UK leadership within the international organisation. Reducing funding below this envelope would not allow enough effort to establish ourselves as a "project" at the national level and would not result in a functioning collaboration. Further effort could be used, building upon the experience gained in the project first phase, and is the desired evolution which is discussed in Sec. 2.4.

It is a strategy decision to request funding for staff at the 50% FTE level at different universities and laboratories across the UK. The posts will be shared with related roles within the experiments, enhancing collaboration both across experiments and geographically. Groups have strategies in place to fill these positions and, in most cases, staff already allocated to the project. Staff are a combination of PDRA and RSE, with some of the RSE being trained in the particle physics field, and some being part of the RSE pool at UCL and Warwick. In the case of RSE from a pool, funding is requested at the 33% FTE level. Finally approximately 0.4 FTE/year are dedicated to management as detailed in Sec. 3.0.

Travel and training funds will be managed within the RAL SLA for groups across the UK to access them using an invoice mechanism similar to experiment travel budgets. The request is made under the assumption that COVID travel restrictions will cease during this project. If, however, travel restrictions remain in place the expenditure will inevitably be lower. We request £20,000 of travel per year. This is to cover travel costs of staff employed on this project to major international events (HSF workshops, CHEP conference) for 5 trips/year at £2,000 each, some travel within Europe/CERN for hackathons and collaborative work for about 10 trips/year at £500 each and £5,000 for UK travel for meetings and workshops.

We request £30,000 for training across the three years to cover the costs of training courses for staff and the organisation of training events for the UK community. The cost per year is estimated based on the costs of the data intensive CDT, which is used as a blueprint for training needs as detailed in Sec 4.3.

Most of the computing hardware needed to deliver this project is available as part of GridPP and the STFC cloud, with additional development resources available at individual institutes and RAL-PPD. Some of the work for WP1 and WP4 entails specialist equipment which may not be available as part of existing resources at the institutes involved, or may not be loaned from industry partners. A small equipment cost is included as part of the working allowance to mitigate risks connected with hardware availability.

The total cost to STFC of this proposal is £1,138,987 for staff, £60,000 for travel and £30,000 for training and £10,000 to procure two FPGA cards needed for WP4 at the start of the project. The cost is distributed almost uniformly over the three years of the project. The final year of the project is kept lower than the first two years to allow for possible delays with the recruitment process at institutes.

5.2 Risks and working allowance

As a mostly R&D programme, a £40,000 equipment **working allowance** is included to mitigate the risk of hardware (R8, R9) not being available during the course of the project either at the institutes involved or on loan for the development planned (Sec. 4.1). A £30,000 **working allowance** is included to hire 3 months of effort from a research software engineer pool at one of the institutes to mitigate delays in the deliverables (R4). This is calculated using the average cost of £10,000 per month for a RSE and would allow to avoid descoping of a central part of the programme.

As an R&D project, it is expected that some of the deliverables will need rescoping during the three years to match the work done by our international collaborators and the findings from such R&D (R5). This is mitigated by engaging with our partners.

The main risk exposure remains with delays with the timeline for the technical design reports (R3), which is largely beyond our control as it is driven by the international schedule and by the timeline for the experiments start up. This can be mitigated by ensuring that our contribution is defined before the end of this project, but will necessarily result in some schedule slippage, as with the experiment construction and R&D projects.

The loss of key personnel is a large risk (R1), due to the fact that software engineers skills are in high demand in industry, hence managing staff turnover is a key concern in the software development area. To mitigate the risk, we should offer a long term career perspective to HEP software specialists and a thriving working environment. While matching industry salaries is not possible, the intellectual interest of a scientific environment coupled with career stability is key to staff retention. Achieving this is one of the major goals of SWIFT-HEP.

5.3 List of deliverables

ID	Year	Q	Milestone
D0.1	23	4	UK contribution to HL-LHC technical design reports
D0.2	23	2	Define phase-2 of SWIFT-HEP
D1.1	22	1	UK data lake prototype
D1.2	22	3	Implement additional QoS information in the data lake
D1.3	23	1	Recommendation to sites on data access optimisation
D1.4	23	3	Virtual analysis facility prototype
D1.5	21	4	Generic pilot logging system
D1.6	22	4	Extend DIRAC workflow for middle-sized VOs
D1.7	23	2	Implement dynamic load management in DIRAC
D1.8	23	4	Implement high level commands in DIRAC
D2.1	22	1	Report on CPU performance profiling of MC event generators
D2.2	22	3	Release of optimised LHAPDF version
D2.3	23	4	Release of MC event generator code w/ technical optimisations
D2.4	21	3	Develop a biased hadronization model for Pythia8
D2.5	22	1	Develop a color-reconnection model for Pythia8
D2.6	22	1	EvtGen Modernisation
D3.1	23	1	Develop an EM prototype for Geant4 in CUDA
D3.2	24	1	Validation and optimisation of the EM CUDA prototype
D3.3	22	1	Example Geant4 application using Optiks
D4.1	22	2	Report of benchmarking tools
D4.2	22	2	Performance comparison of tracking and vertexing algs on FPGAa
D4.3	23	1	Benchmarking of CPU and accelerator algorithms
D4.4	23	2	Report on comparison of generic interfaces
D4.5	23	1	FasTrack algorithm implemented in OneAPI
D4.6	24	1	FasTrack comparison between CPU, GPU and FPGA
D5.1	22	3	Demonstrate interoperability with UK data lake
D5.2	23	1	Integration of caching mechanism with workflow management
D5.3	23	3	Per-site optimisation of workloads
D5.4	24	1	Deliver workload scheduling for the workflow management

Table: List of deliverables organised by WP

5.4 Gantt Chart

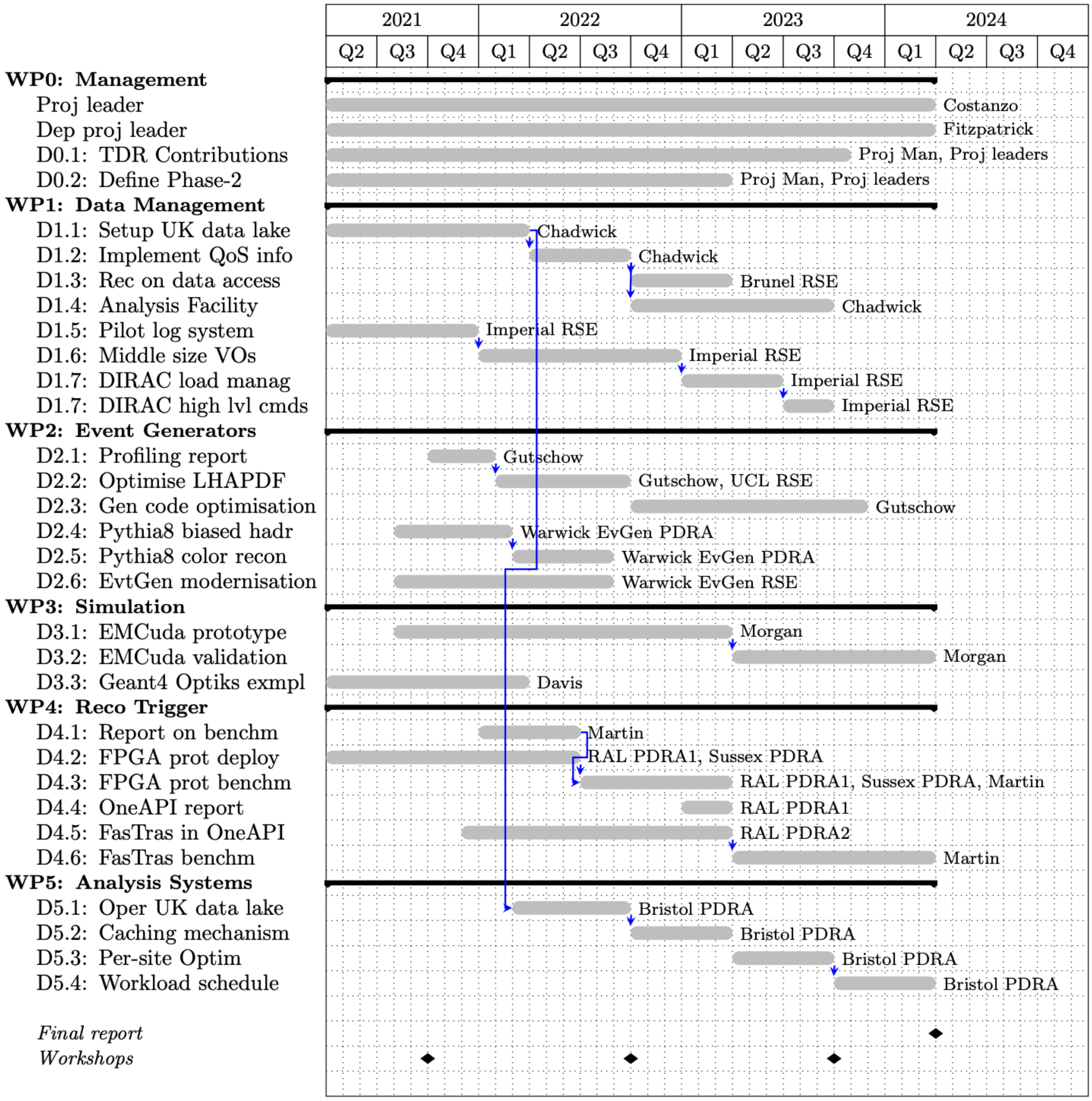

Figure: Gantt chart for the project

Appendix: List of SoI signatories

Birmingham: Philip Ilten, Mark Slater, Chris Hawkes, Jacob Kempster, Cristina Lazzeroni

Bristol: Henning Flaecher, Benjamin Krikler, Jonas Rademacker, Sudarshan Paramesvaran, Daniel O'Hanlon, Luke Kreczko, Joel Goldstein, Jim Brooke

Brunel University London:

Akram Khan, Paul Kyberd, Joanne Cole, Raul Lopes, Ivan Reid

Cambridge: Christopher Jones, John Chapman, Valerie Gibson, Leigh Whitehead, Oleg Brandt, Sarah Williams

Durham: Marek Schoenherr

Edinburgh: Pete Clarke, Sinead Farrington, Christos Leonidopoulos, Victoria Martin, Ben Wynne, Rob Currie, Miquel Nebot, Mark Williams

Glasgow: David Britton, Andy Buckley, Tony Doyle, Lars Eklund, Phillip Litchfield, Mark Owen, Aidan Robson, Manuel Schiller

Lancaster: Roger Jones, Jaroslaw Nowak, Helen O'Keeffe, Andy Blake

Liverpool: Monica d'Onofrio, Carl Gwilliam, Eduardo Rodrigues, Tara Shears, Karol Hennessy, Kurt Rinnert, Helen Hayward

Imperial College London: Alex Tapper, Henrique Araujo, Paul Dauncey, Mitesh Patel, Yoshi Uchida, Antonin Vacheret, Nicholas Wardle, Morgan Wascko

King's College London: Francesca Di Ludovico, Teppei Katori, Jeanne Wilson

UCL: Jon Butterworth, Louie Corpe, Christian Gutschow, Nikos Konstantinidis, Sébastien Rettie, Tim Scanlon, Peter Sherwood, Jim Dobson

Royal Holloway: Simon George, Pedro Teixeira-Dias

Queen Mary: Adrian Bevan, Jonathan Hays

Manchester: Conor Fitzpatrick, Marco Gersabeck, Darren Price, Chris Parkes, Adam Davis

Oxford: Daniela Bortoletto, James Frost, Claire Gwenlan, B. Todd Huffman, Ian Shipsey, Guy Wilkinson, Alfons Weber (also STFC/RAL)

Sheffield:

Davide Costanzo, Mark Hodgkinson, Dan Tovey, Christos Anastopoulos, Trevor Vickey, Kristin Lohwasser, Lee Thompson

STFC Rutherford Appleton Laboratory Tim Adye, Chris Brew, Linda Cornwall, Alastair Dewhurst, Katy Ellis, Dmitry Emeliyanov, Sam Harper, Dave Kelsey, Julie Kirk, Jacob Linacre, Stewart Martin-Haugh, Raja Nandakumar, Dave Newbold, Emmanuel Olaiya, David Petyt, Thomas Schuh, Claire Shepherd-Themistocleous, Fergus Wilson, John Baines, James Walder, Alfons Weber (also Oxford)

Sussex: Iacopo Vivarelli, Simon Peeters, Antonella De Santo, Fabrizio Salvatore, Alessandro Cerri

Warwick: Gary Barker, Tim Gershon, Michal Kreps, Thomas Latham, John Marshall, Ben Morgan, Mika Vesterinen

CERN: Graeme A Stewart, Pere Mato