1 S2I2 Priorities from an LHCb perspective

US-LHCb is much smaller than US-ALTAS or US-CMS, so we identify only three highest priority Focus Areas and two other Focus Areas where we are most likely to engage. In large measure the highest priority areas are those with exsiting expertise and interest in US-LHCb. In addition, we identify a number of areas where we believe the core/backbone team can most usefully build expertise and contribute to efforts crossing different research focus areas and crossing experiments.

1.1 Highest Priority Research Focus Areas

As will be true more generally, these areas will overlap, at least in part.

- **Data Analysis Tools and Systems** Our interests here include
  - using many-core and multi-core architectures effectively to reduce time-to-insight
  - real-time data analysis (this will become critical in Run 3 as we will lack the offline data storage capacity to analyze all the interesting data we can process in the real-time system).
  - promoting Python as a first-class data analysis language

  The focus of all three of these items is doing more physics with limited computing resources. Although we tend to focus on the costs of developing software infrastructure and of computing hardware, the real cost-driver is analyst time. The first and third items should allow physicsts who are not software experts to do better analyses faster. The second item will allow the experiment to do physics analyses that would not be possible otherwise.
- **Trigger and Reconstruction Software** Already in Run 3, the trigger and reconstruction software stacks will be fully integrated. We expect that, in general, we will persist only reconstructed objects, not raw data. Areas of special interest include:
  - speed up charged track and vertex reconstruction for Hlt1 sufficiently to process 30 MHz of beam crossings and reduce the rate of events for additional consideration in Hlt2 to 1 MHz while retaining essentially all potentially interesting events.
  - fully reconstruct 1 MHz of events; select $\mathcal{O}(100)$ kHz for partial or full persistence
  - introduce real time analysis for final states where the statistics are too high to persist individual events; for example, we should be able to fully reconstruct 3×10^9 $D^+ \to K^{=}\pi^+\pi^+$ per fb^{-1}; and we expect to collect as much as 10 fb^{-1} of data per year.

  The focus of these items is extending the physics reach of the experiment through better use of the limited real-time CPU and storage resources. As we cache data on local disk between Hlt1 and Hlt2, both types of resources impose powerful constraints.
- **Machine Learning** We have identified three areas where ML can impact the LHCb Run 3 physics program substantially.
  - replace the most computationally expensive parts of the event pattern recognition;
  - increase the performance of the event-classification algorithms;
  - reduce the number of bytes persisted per event without degrading physics performance.

  The focus of these items is extending the physics reach of the experiment through better use of the limited resources, especially in the real-time system. The first two items are equally well considered part of the Trigger and Reconstruction effort. The third item is more general as data compression will be important both in the real time system (the limited local disk space available for caching between Hlt1 and Hlt2 is already a problem) and offline. In both these cases, we will need to understand the multi-dimensional trade-offs between reducing disk *use, increasing times for encoding and decoding data, and degrading physics when using lossy algorithms.*

The NSF funds the DIANA-HEP project through the SI2 program, and our efforts generally complement those of the ROOT team, mostly based at CERN. The DOE funds the ROOT effort at a relatively low level. So it seems like Data Analysis Tools and Systems is an area where the S2I2 can take a clear leadership role. In general, efforts in this area will provide value to more than one LHC experiment and to the wider HEP community.

Similarly, Triggering and Reconstruction seems like another area where an S2I2 can play a leadership role, at least complementary to DOE-funded efforts in the U.S. and similar efforts of international collaborators. The DOE currently funds a PIF award in this area to a CMS team. Identifying specific projects of value to more than one LHC experiment and the wider HEP community will require a bit more thought here than in some other potential Focus Areas.

Machine Learning is an area where the NSF Office of Advanced Cyberinfrastructure has indicated a strong interest. They have already recommended an SSE award to an LHCb team for ML in the context of triggering for Run 3. While some ML algorithms will be restricted to use by a single experiment, others (such as data compression) should be generally useful across LHC experiments and in the larger HEP world.

1.2 Other High Priority Focus Areas

No one from US-LHCb is currently working in the two potential focus areas discussed here, but the impact on our physics program is likely to be significant. There should certainly be opportunities to work with our non-U.S. colleagues as there are already cross-experiment collaborations in these areas. In addition, having followed the CWP process closely, they appear to be areas of substantial U.S. university interest and expertise.

- **Data Access, Organization and Management**
- **Workflow and Resource Management** Areas of special interest to LHCb might include
  - use of "opportunistic resources" at national labs
  - use of resources on commercial clouds

  US-LHCb gets no computing resources from the NSF (and no resources at all from OHEP in the DOE). We did a small experiment with Google to demonstrate that we could run our simulations on their cloud, and we are using resources at the Ohio Supercomputer Center. We are interested in learning how to use as wide a variety of resources as possible, and convincing someone to give us access.

As for the "highest priority focus areas", identifying how an S2I2 would provide value to more than one LHC experiment and to the wider HEP community might be a challenge. US-ATLAS and US-CMS seem to have built in-house/competing software in these areas. To the extent that S2I2 efforts in these areas could lead to convergence and use of common software, the impact on use of human resources, and therefore on sustainability, would be substantial.

1.3 Core/Backbone Priorities

Members of the core, or backbone, team should have expertise in a number of areas in software engineering, computer science, and data science that transcend particular "research and development focus areas". Individuals may work within one of these focus areas, or across several, but they may also povide complementary expertise and services that enhance the role of the Institute as an intellectual hub for the larger software R&D community. In no special order, areas of expertise should include:

- perforamce optimization, including use of vector processors in the next generation of Intel cores; use of math libraries; use of performance guided optimization of compilers; etc.
- software and data preservation, especially in the context of reproducible research
- effective use of emerging architectures, including many-core, multi-core, ARM processors, and HPC systems at national labs
- packaging and release management tools
- knowledge of scientific software tools beyond HEP
- deep and broad understanding of machine learning
- training and education

Individuals may have expertise in multiple areas, and multiple individuals may have expertise in any one of these areas. Members of the core team will have continuing appointments and they will provide some of the glue to ensure a coherent program. They will also be available as "consultants" on exploratory projects and projects internal to individual experiments.