PPTAP: R&D Roadmap

Neil Chue Hong (Edinburgh)
and Tim Scanlon (UCL)

Overview

- Overview
- Discussion Points

Overview

- This session will focus on a few of the general cross-cutting points for the roadmap
  - In no way an exhaustive list of all points that will be made in the roadmap
    - Many elements of roadmap already have wide consensus
    - Choice of topics focussed on points where a range of views exist and more discussion could be useful
  - To foster discussion, have written straw-man statements and included some discussion points

- Broadly it will cover the following areas:
  - Funding
  - Balance of incremental vs innovative/blue-skies R&D
  - Heterogeneous Computing (including GPUs)
  - Climate impact
  - Grid/Cloud computing
  - Cross-experiment development
  - Software ecosystems
  - Hardware engagement
  - Digital twins
  - International engagement

Funding Landscape

- Perhaps unsurprisingly there was a general consensus that:
  - More and stable funding should be invested in S&C R&D
    - This is essential for the field to fulfil its physics goals
      - It can no longer be an afterthought when planning experiments
    - It also has a significant beneficial impact upon the wider economy
      - This point will be very strongly made, with projections
  - It is essential that this also includes funding for:
    - Operational and other associated costs to be included with hardware funding
    - Innovative R&D (to be discussed on next slide)
    - RTPs (to be discussed in next session)
    - More/better training (to be discussed in last session today)
    - Retention of critical mass of specialist expertise (to be discussed in last session today)

- Discussion points to keep in mind across all roadmap topics:
  - Anything to add or emphasise?
  - Any preferences on how this should be funded?
  - What is the split between short, medium and long-term funding?
  - Examples and use-cases we can include

Blue-Skies/Innovative R&D

- Statement:
  - The majority of the R&D funding invested in S&C is spent on ensuring we can fulfil our immediate goals, however, it is essential that the field also invests in innovative and blue-skies R&D, to ensure that we can effectively identify, prepare for, steer and harness future game-changing developments (e.g. quantum computing, new data formats, ability to run on future computer architectures, intelligent networks). Such funding could be best supplied via responsive competitive bids or HEP-RSE ‘blue-skies’ fellowship.

- Discussion points:
  - How do you ensure that such funding does not cannibalise core funding and imperil our ability to successfully exploit our experimental data?
  - Would responsive short-terms funding be the most effective way to fund such research?
  - What counts as innovative R&D?
    - Survey responses suggest that significant progress could be made by applying existing R&D and making it available on production infrastructure

Heterogeneous Computing

❖ Statement:

➢ With the increasing computational demands and the greater prevalence of heterogeneous computational hardware, such solutions have already been adopted in several areas in HEP, and there is potential for their adoption across a wider range of areas. R&D should be undertaken to understand the potential, limitations and applicability of such systems. Such a R&D programme will encompass several key areas: reproducibility of the output, code portability/abstraction, performance, scheduling and sustainability (energy usage, cost, lifetime, avoiding vendor lock-in).

❖ Discussion points:

➢ It is hard to predict what the hardware landscape will look like in a decade: how do we prepare for that?

➢ What parts of the code base should be ported to heterogeneous architectures first?

➢ Are there areas which should not be considered for a heterogeneous computing approach?

➢ What are the main limiting factors we should quote (e.g. data transfer issues)?

➢ Are the performance overheads of portability libraries (e.g. Alpaka) low enough that the benefits outweigh the disadvantages? How do we compare trade-offs between performance, access to compute resources and energy efficiency?

➢ Is there a role for FPGAs in any areas apart from real-time applications?

➢ What about TPUs, IPUs or APUs?

Climate Impact

- Statement:
  - Many planned S&C improvements, for instance the use of lower-power hardware, simplified code, improved physics choices and more efficient algorithms, will also reduce the climate impact of the field. However, given the climate emergency, it is essential that the field takes into account the impact upon the environment in its S&C R&D programme, especially given potential mandatory constraints upon carbon production that could limit our fields in the future.

- Discussion points:
  - Will we reach a point where the field is forced to choose between the physics output and the environmental impact?
  - Are there other environmentally friendly developments, beyond lower-power CPUs, better code, best practices in data centres, the use of green energy and hardware re-cycling, that we can highlight?
  - How should energy efficiency be considered in overall costings?
  - What are the relevant issues around power consumption for GPU/CPU etc. that HEP R&D funding could address?

Grid/Cloud Computing

- Statement:
  - Wide spread distributed computing systems are essential to enable both large and small scale experiments to effectively both store and process their data, as more experiments from a range of areas have increasingly large data storage and processing needs, along with the need for a heterogeneous hardware environment. R&D, in addition to central support, will be needed to enable the smaller experiments to effortless integrate into these ecosystems.
  - Cloud computing should be used as an approach to provide scalable access to computing but should be kept “in-house” through specialised HEP facilities (c.f. FNAL HEPCloud) to retain expertise and avoid data egress issues. The functionality to enable cloud-bursting to commercial providers should also be provided.
  - For many existing HTC workflows, users now require HPC-level computing resources, however the way these two types of system are accessed are different. The software to access the two systems should be merged, so users can effortlessly run on either.

- Discussion points:
  - What R&D and support is required by the smaller experiments to enable them to effectively harness the power of the grid computing infrastructure?
  - What R&D is needed to allow a heterogeneous hardware landscape to be effectively harnessed by the experiments (framework alterations and scheduling)?
  - Should HTC and HPC look more alike (to users)?

Cross-experiment synergies

- Statement:
  - Cross-experimental R&D projects can be very beneficial in terms of sharing expertise and producing enhanced solutions for the same cost. The field should fully engage and invest in cross-experimental research projects (e.g. acts), to maximise the return on investment, enhance the solutions produced and ensure that the benefits are adopted as widely as possible.
  - Certain areas of HEP have also established valuable expertise in various areas (for instance in applied machine learning), whilst other areas would strongly benefit from the infusion of such expertise. Forums and collaborative projects should be setup to ensure that expertise can easily be transferred across the field to maximise research output.
- Discussion points:
  - Is pooling effort on R&D on common areas in software, algorithms and hardware, always the optimal way to produce better solutions which are widely applicable for the same money?
  - How should R&D in such projects be funded and coordinated (standalone international projects or via HSF)?
  - How do we ensure knowledge transfer both between and within experiment and theory?
  - Is the primary driver for more cross-cutting software development for the analysis stage being driven by user's frustrations at experiment specific analysis codes?

Software Ecosystems

- Statement:
  - Concentrating on the use of particular software ecosystems helps to focus development effort and reduce maintenance costs. Shared software ecosystems can provide the benefits of scale from increased numbers of users such as lower maintenance costs.
  - Splitting functionality into discrete, interoperable parts helps portability, reuse and longevity. Experiments can concentrate on the parts of the stack that are most important on them. This also makes it easier to divide required work.
  - The field should coordinate development so that new software is interoperable with existing ecosystems, and work to reduce redundancy and plan for “sunsetting” when components are superseded.
- Discussion points:
  - How can we ensure that such cross-experimental co-operation operates effectively?
    - How can smaller experiments effectively engage in this process?
  - What is the best way to identify gaps and overlaps in functionality in the current ecosystems?
  - How should the field plan for software deprecation?

Hardware Engagement

- Statement:
  - The field generally uses commercial off the shelf (COTS) computing and networking hardware to ensure cost-effectiveness and ability to deliver resources at scale. The hardware companies are keen to engage, but the field has little sway over them as it represents a tiny proportion of their income. However, the field should continue to engage with industry to understand the future developments that could enhance our S&C, to understand how to most effectively use the products and to provide feedback on product development that if adopted would be beneficial for the field.
- Discussion points:
  - How much engagement with industry should we have:
    - Which part of the field (e.g. near detector vs Grid)?
    - How far out in time (e.g. near time or longer-term developments)?
    - Which members of staff (e.g. RSEs, academics, PhD, PDRAs)
  - Do we gain knowledge from engaging with industry that helps improve the efficiency and cost-effectiveness of our procurement decisions?

Digital Twins

- Statement:
  - Digital Twins - a virtual representation that serves as the real-time digital counterpart of a physical object or system - are increasingly being used in other fields (from engineering to healthcare), driven by advances in simulation, sensors, coupled systems and machine learning.
  - Many of the approaches used in digital twins have been used in HEP to better understand the way that experiments will operate. However there may be advances from other fields which could benefit HEP.

- Discussion points:
  - What are the synergies between what we do in PP (or PPAN) and what is being done in the field of “Digital Twins”?
    - Which areas might benefit most (e.g. accelerators)?
  - How should the field engage with others doing R&D in this area?

International Engagement

- Statement:
  - The community should strengthen/expand global defined S&C R&D roadmaps produced by various bodies (HSF, WLCG, DOMA, experiments/projects) and they should be invited to officially feed into European wide S&C R&D roadmap, which should be used to support proposals at the international, European and national levels.
  - The UK should continue to invest in areas where it has international leadership, collaborating with other initiatives via various common forums such as the HSF.

- Discussion points:
  - Is there a need for an 'official' European Roadmap in this area, similar to the detector and accelerator R&D roadmaps?
  - Should the various S&C bodies producing roadmaps officially feed into that process or have a parallel process to produce a R&D roadmap?
  - Along this line:
    - Should the UK remain focussed on its strengths regardless of what others are working on and cooperate via common forums?
    - How should the UK interact with other international initiatives?
  - Swift-HEP, IRIS-UK and GridPP/WLCG already coordinate cross-experiment S&C R&D in the UK, are these sufficient or is anything more needed?
    - Should there be a body to coordination across all the S&C areas?

Further discussion points if time….

Lock-in

- Statement:
  - The field has a challenge because of the long duration of experiments which means that lock-in is more costly, particularly if key skills are not maintained. There are various APIs (both open and proprietary) and software engineering approaches that help create abstractions to reduce the chance of lock-in.
  - R&D should be undertaken into the impact of writing code that is more portable to help alleviate the typical length of the tendering process which can be vulnerable to short-term trends and issues (e.g. use of GPUs for cryptocurrency, single->many core rather than increase in clock frequency, chip shortages)

- Discussion points:
  - How much R&D should we be doing on open APIs and abstraction to ensure that they meet the long-term needs of HEP, and ensure that there is not lock-in to any particular provider?
  - Will the field need to invest R&D effort to mitigate supply chain issues and reduce dependency on particular hardware solutions?

Security

- Statement:
  - There is a significant research and reputational risk from a cybersecurity breach. In recent years, research facilities and experiments have been increasingly targeted. The field’s computing and software infrastructure is complex to secure, as it is at the forefront of distributed data storage processing & analysis, and the experiments are completely dependent on computing to conduct their research. Even if not being specifically targeted the collateral damage from a cybersecurity breach is now significant.
- Discussion points:
  - How much R&D is required to be done in the field to recover quickly - the specific weaknesses of the field because of the way that experiments organise their S&C
  - How much is just keeping up with the best practice? And how much do we need to engage with other experts in this area?
    - E.g. the AAAI paper