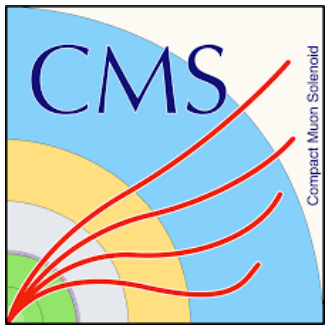
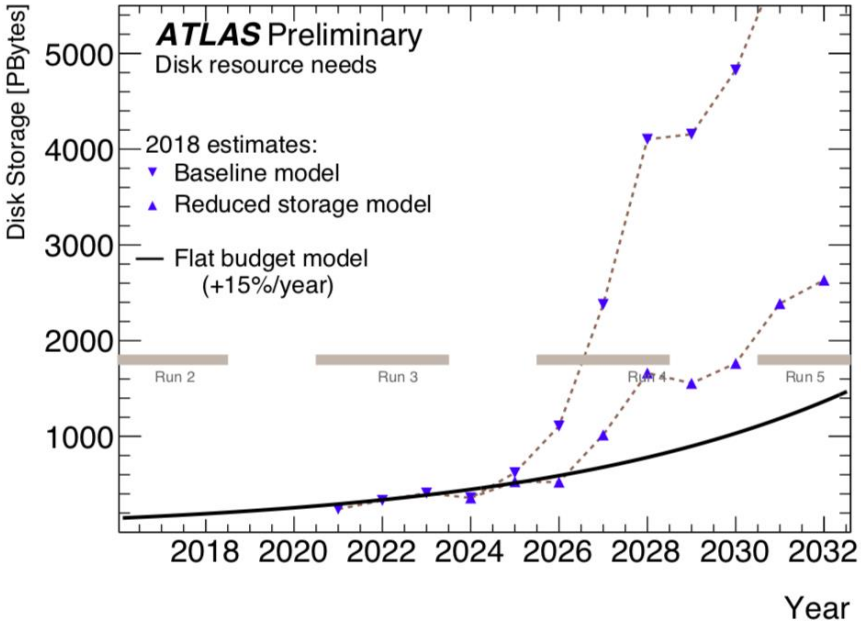


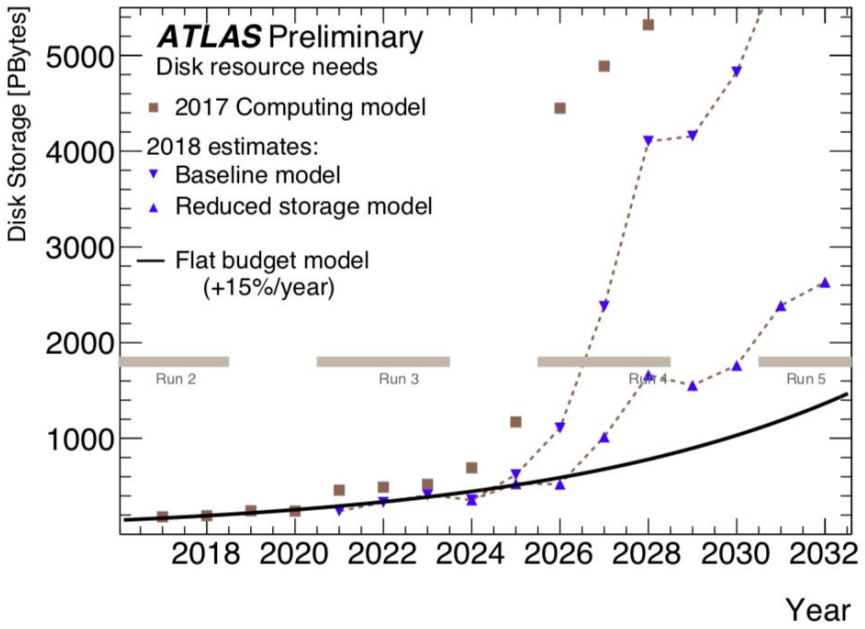
Efficient Computing for High Energy Physics

Sinéad Farrington, Conor Fitzpatrick, Dave Newbold

Why are we here?

- Significant computing challenges lie ahead
  - Price per CPU performance has been in decline and will continue to be





Why are we here?

- Particle physics in the HL-LHC era will generate unprecedented datasets, to deliver unprecedented precision and searches
- We must prepare ourselves for that era, equipping ourselves with techniques that optimise the use of available methods in order to give us *the best capacity to exploit fully the scientific opportunity* of the HL-LHC- era particle physics experiments
  - First we must equip ourselves with proofs of principle, define figures of merit, and demonstrate where gains are achievable

Goals of the ECHEP Project (6 months)

- Discuss the challenges of the HEP software stack (focused on HL-LHC but others welcome) (✓w.s.)
- Inform ourselves on alternative architectures
  - Reach out to industry (✓w.s.)
- Discuss work already done/ongoing to address this, within the UK and internationally
  - Engage with existing organisations e.g. HSF, IRIS-HEP(✓w.s.)
- Identify proof-of-principle demonstrations of new platforms (started, w.s.)
- Form a new community of experts in the UK with international engagement (✓w.s. and continuing)
  - Requires engaging with, and providing, training opportunities (to do)
- Establish working parties, define deliverables (happening)
- Detailed plan and impact strategy for subsequent, ~three year R&D project will be written

Link to other projects + overall plan

- ECHEP is funded to inform the Statement of Intent STFC submission (PI Costanzo)

- ECHEP aims to provide substantial content to the PPRP
- Part of ECHEP's proposal to the opportunity call, was that it will provide a detailed plan of work for a 3 year R&D project to follow it, together with impact plans
  - This needs to be realistic, in the international context and not solve problems in a silo or ignore already-existing solutions
  - Cannot be an open-ended wish list, it should be needs-based

Draft SoI

- Draft SoI prepared by Davide Costanzo (with input from cross-experiment and GridPP/IRIS PIs) has been circulated on the Excalibur and ECHEP lists
  - Comments to be sent in via your group leaders or experiment PIs

How ECHEP will operate

- Hold open overview meetings every month to discuss the work of that month
- Hackathons/training events/mini-meetings (always open and advertised on our mailing list) to be set up as needed

- The people responsible are the PI/Co-I's/funded PDRAs/ area co-leaders

Working group areas

1) Generators (adaptation to multi-threaded methods, address negative weights issues – phase space sampling)
2) Simulation (both fast simulation and approaches to vectorise full simulation)
3) Trigger/Reconstruction
4) Analysis methods (UpROOT, vectorisation etc.)

Area leaders now appointed – thank you for the very strong (self-)nominations received, many highly qualified people, stay tuned for the sign-up sheets as we would like to gather all effort together.

ECHEP as a project should deliver conclusions on all of these areas and the UK needs/capacities/abilities. It's of course a given that the SoI (project) may be able only to draw on a subset.

Area Leaders

- Generators: Andy Buckley (Glasgow, ATLAS), Marek Schoenherr (Durham, Sherpa)
- Simulation: Ben Morgan (Warwick, Geant collab, DUNE/ATLAS), Adam Davis (LHCb)
- Trigger/Reco: Stewart Martin-Haugh (RAL, ATLAS), Chris Jones (Cambridge, LHCb)
- Analysis: Eduardo Rodrigues (Liverpool, LHCb), Lukasz Krecko (Bristol, CMS)

How you can get involved

- Add your name to a sign-up sheet for the area you want to work in (see mailing), mattermost will be set up for each
- Follow the main ECHEP mailing list announcements about hackathons and mini-meetings
- Contribute your knowledge, current work (think about what you'd like to get involved with if not already)
  - *Talk to the relevant area leader(s)*
- Already start thinking about where this can go after the 6 month project is over – what resource would be needed to turn your ideas into something real
  - Is the work doable within the six months with resource you already have (your own time…)
  - Does it need dedicated funding
  - What training is needed that you currently can't access either nationally or internationally

A Legacy

- We should aim for this to be a productive six months
  - And a productive workshop (✓), with a second to follow in six months time (looking at extension given current CV19 circumstances)
- We should aim for ECHEP to be a project that delivers our best ideas on what needs to be done next
  - Define benchmarks
  - Proofs of principle within the six months where possible
  - Full engagement with industry and international colleagues
    - Several at the workshop – good discussions
  - Maximised scientific output in the HL-LHC era is the ultimate and overarching goal

Workshop Agenda

- The software stack
- International efforts
- Generators
- Simulation
- Trigger and Reconstruction
- Analysis codes
- Alternative Architectures tech talks
- Aiming to draw out existing UK (unique/complementary) expertise and have invited international and industry experts so that we are fully engaged

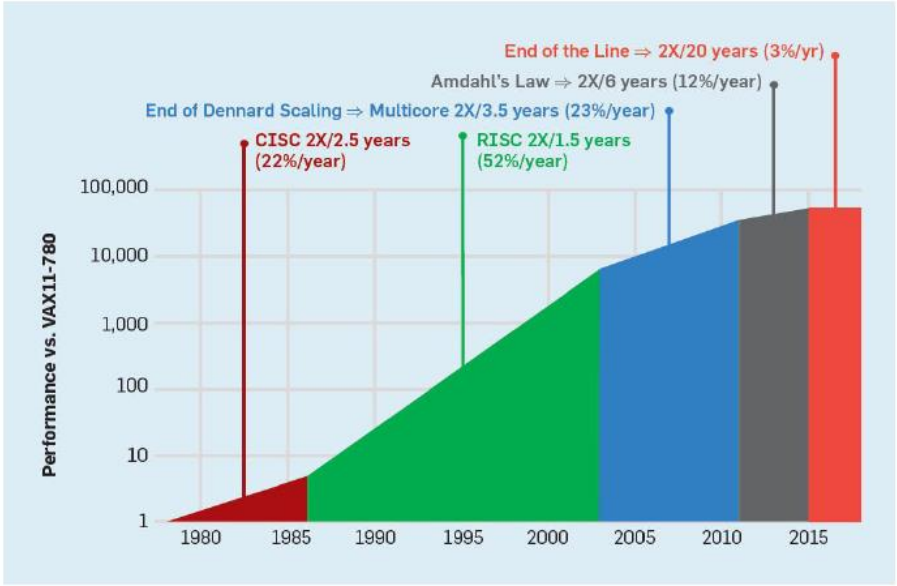
Workshop summary: HSF

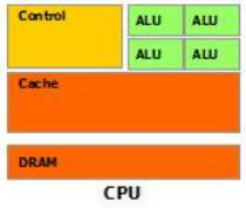
- Graeme Stewart – overview of HEP Software Foundation
  - We should join HSF groups, ECHEP aims are the same as HSF and will not duplicate, but we have the short term goal that we

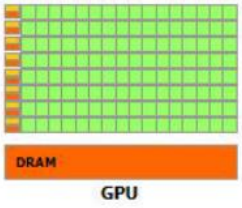
Decreasing Returns over Time

- Conclusion: diversity of new architectures will only grow
- Best known example is of GPUs

[link]





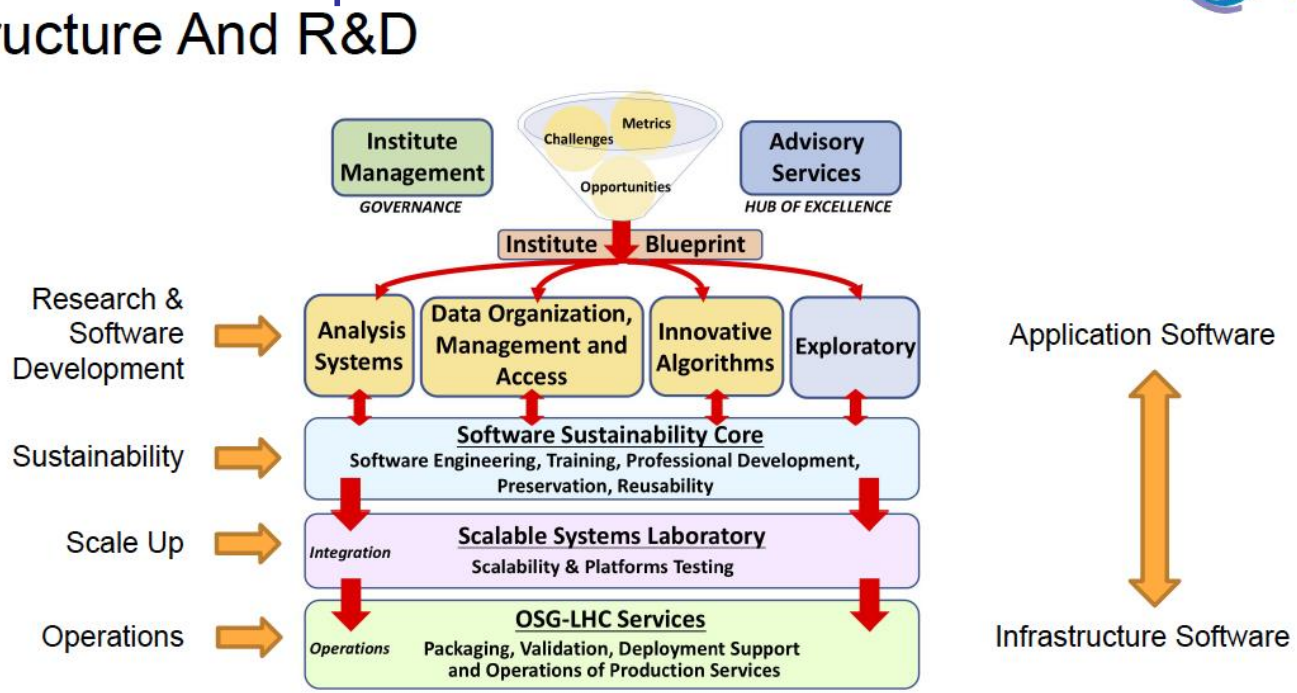


GPUs dedicate far more transistors to arithmetic

5

Workshop summary: IRIS-HEP

- IRIS-HEP (Henry Schreiner) –went through a large scale preparatory phase in the US
- Remit:1) innovative algorithms in reco/triggers 2) analysis systems 3) data organisation 4) the Open Science Grid (note: not generators or simulation)
- 5 year project from September 2018



Workshop summary: Frameworks

- Adam Barton (complemented by talk from CodePlay)

Lancaster University

Conclusions

- The after the core boom the future appears to be heterogeneous computing
- CMS have started working on applying this
- ATLAS has been investigating various systems
- ALICE has an interesting alternative design
- Intel OneAPI is, on paper, a promising framework for handling heterogeneous hardware with “minimal” code rewrites

24

Workshop summary: Generators

- Weights are a difficult challenge; suggestion for binaries dedicated to physics processes (Marek Schonherr, Christian Gutschow)

Conclusions

- **matrix elements**
  unweighting is the bottleneck
  - high complexity of the integrand (large memory requirements)
  - poor understanding of its structure (unweighting efficiency)
    (machine learning currently does not seem to be the answer)

  → nonetheless, compared to other fields we need relatively few evaluations of the integrand, but one evaluation is complex

- **parton showers & non-perturbative modeling**
  currently fine (small part of the overall computational budget)
  - may need attention when improvements lead to large weight spread
- **code structure**
  not compatible with currently favoured computing infrastructure
  - trying to make computers do physics, not exploit their strength
  - too many run-time decisions
  - any theory code is essitially a prototype

Workshop: CPU Optimisation

- CPU profiling, optimisation, library pre-load, heap profilers (Stewart Martin-Haugh)

- Rich seam of easy and hard optimisations to apply to HEP code
- Use off-the-shelf tools as much as possible
- Needs revision as algorithms and frameworks change
- Opportunity to design new experiment software: fast and correct from the start

Workshop: Simulation

- Full Simulation: Ben Morgan, noted there are 2 UK HEP physicists in Geant, a handful more in medicine/industry

Geant4 Task Force for R&D

- Promote and survey research activities:
  - *Potential **software architect** updates to Geant4*
  - *use of **emerging technologies/computing architectures** of benefit to Geant4*
- Ensure the visibility of such explorations and act as the focal point for such **activities inside and outside of the Collaboration**
- Where appropriate, conduct benchmarking comparison and provide/assist communication/support among R&D activities
- Make timely assessment reports to Steering Board with **solid proof of benefits**
- ***Based on assessments of this Task Force, once a concrete and beneficial architectural revision is identified, the Steering Board launches a new, dedicated, task force to create workflow, estimate required resources and drive that particular development for integration into the code base.***
  - ***As was done for Multithreading and prior revisions***

Workshop: Simulation

- Fast Simulation: Hasib Ahmed, Adam Davis
- Vital part of the picture, improvements in recent years show it to be accurate method (validate against FullSim)
  - Current versions now more accurate as well as being fast

Workshop: industry talk take-homes

- (subjective summaries)
- nVidia have a next generation coming soon, currently being tested
- Xilinx are able to tune their FPGA offerings to the use case (“more than an FPGA”)
- ARM and nVidia are targeting energy efficient supercomputing
- CodePlay: the future is heterogeneous and they are designing solutions to address it
- GraphCore: designing dedicated chips for ML

Workshop: Timing

- Tracking with timing (Mark Williams)

Open Questions: Avenues for R&D

How to benchmark computing performance & value-for-money?

⇒ CPU / GPU / FPGA

Benefits of **per-hit** versus **per-track** timestamps

⇒ How essential is 4D tracking for HL-LHC and FCC applications?
(both for physics performance and resource use)

Can we gain even more by considering timing globally – time-aware Kalman filter?

⇒ CMS now working to incorporate timing into particle flow, but no timing in tracker / vertex detector

Matching objects between detectors relies on knowledge of particle type (=speed)

⇒ i.e. we really need 5D reconstruction (space + time + particle ID)

Can we reduce simulation resources? Generate only in-time part of events?

CHEP Workshop: Timing and 4D Reconstruction 17-18 February 2020 Mark Williams

Workshop: Accelerators

- FPGA promising in tracking (Alex Cerri); GPU studies (Ben Wynne)

- It was discussed that we should keep records of studies that did/didn't work well, what the processors were and the quantified outcomes. *We need to define benchmark studies* see area leader talks.

Workshop: Trigger/reco contd

- LHCb (Conor Fitzpatrick): found it cheaper and more flexible to meet its physics programme requirements by reading out the detector at the bunch crossing frequency.
  - Required significant FTE in software optimisation. Both GPU-based and CPU-based triggering solutions under study and have proven capable of operating at 30MHz in a modest budget.

- Real time analysis (Caterina Doglioni): ATLAS, CMS, LHCb all use reduced data formats in some shape or form.
  - Extension for future consideration on ATLAS + CMS, already used on LHCb: Selective raw event/partial event building can be used to reduce bandwidth and offline processing loads for certain analyses. Needs access to physics objects at trigger level and quantification of 'how good is good enough?'

Workshop: Machine Learning tracking

- Kurt: Using ML for track stub reconstruction on FPGAs looks like promising R&D.
  - Machine Learning hybrid algorithm is 99.73% efficient on tracks from B mesons, compared to 99.12% state of the art algorithm.

Summary

Particle physics faces major computing challenges

- Lots of data
- Fewer relative resources

Python is a first class analysis language

- E.g. industry, astrophysics
- We seem to be at a tipping point within HEP?

Many new approaches to integrate HEP analyses with other tools

- PyHEP and scikit-hep projects
- Columnar Data Analysis

FAST-HEP has been exploring new approaches within the UK

- Resulting tools seeing use on several experiments

How can we best capitalise on these existing UK-led endeavours ?

ECHEP 2020, b.krikler@cern.ch

What next?

- Today: kick-off talks from the Area Leaders, food for thought

- Sign-up sheets for the areas
- Each area will set up a mattermost channel for discussion and organise meetings, area leaders to decide on their own frequency/structure since each is different in nature
  - Business to be conducted through mattermost and through the meetings, meetings to be announced on the main ECHEP list
  - https://mattermost.web.cern.ch/signup_user_complete/?id=yademxm4bpfc7qjfes34gi1k1y
- Want the focus to be on quantifying outcomes, proofs of principle where possible – value substance over talk

Charge for the Areas

- Review and make contact with international work already started [3 months]
- Gather quantitative data (disk/CPU/usage patterns) in the area of interest [3 months]
- Review potential avenues for efficiency / resource use improvements, and critically examine which ones may be the most useful [3 months]
- Where possible, carry out rapid modelling, trials and feasibility studies of different approaches [6 months]
- Document in a short report the work, findings and recommendations for the next stage [6 months]
- The main deliverable is an evidence-based set of recommendations on which areas and approaches to pursue with UK resources in the coming years, and a first appraisal of what benefits that may bring if successful.

Dates of next meetings

- Subject to change (will doodle nearer the time, and need to check on conflicts e.g. HSF will have some plenary sessions)

- 17th April
  15th May
  19th June