Full Detector Simulation on GPU Updates

Ben Morgan

THE UNIVERSITY OF WARWICK

Detector Simulation Work and R&D

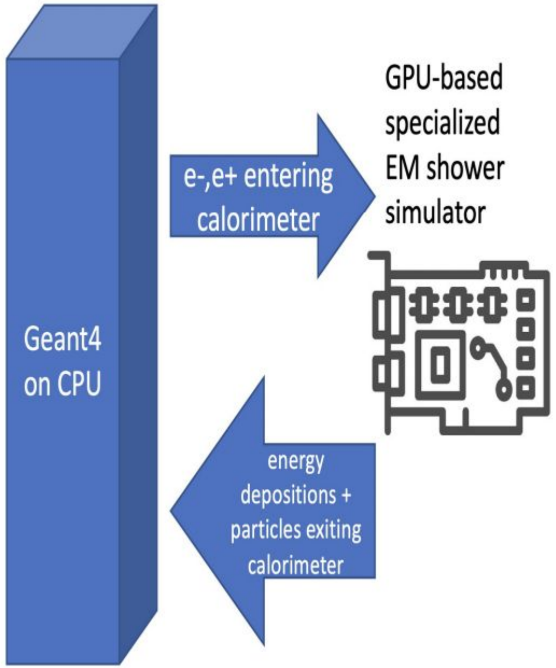
- Short overview of status today - follow links to recent presentations at **Geant4 Collaboration Workshop**, **GPU R&D Review** for in depth details
- **Geant4 Project (Worldwide)**
  - *https://geant4.web.cern.ch*
- **AdePT R&D Project (CERN-SFT)**
  - *https://github.com/apt-sim*
  - *Detailed presentation tomorrow*
- **Celeritas R&D Project (ECP: ORNL, FNAL, Argonne, LBL)**
  - *https://github.com/celeritas-project*
- **Vecgeom/ORANGE Surface Based Geometry (CERN, Celeritas/ORNL)**
  - *https://gitlab.cern.ch/VecGeom/VecGeom (See surface_model branch)*
  - *https://github.com/celeritas-project/celeritas/tree/develop/src/orange*

Community Engagement: Geant4 GPU R&D Review

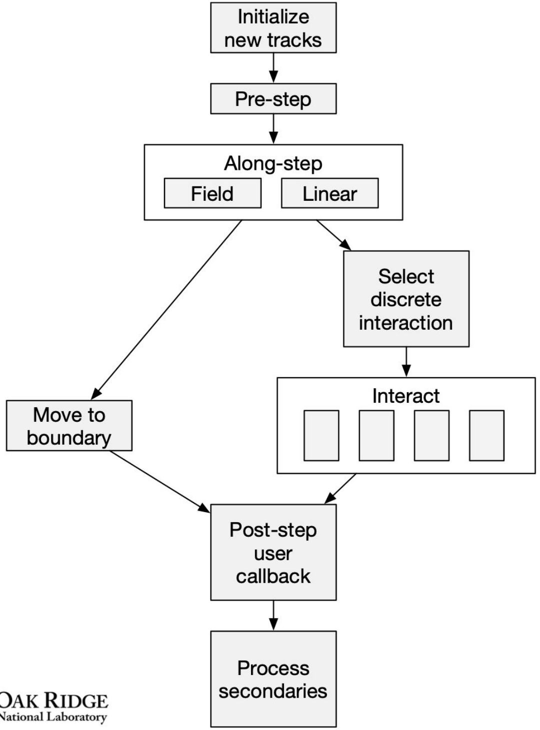
- 13th-14th Dec 2023@CERN: Presentations by AdePT/Celeritas teams to small panel of experts drawn from across LHC experiments and Geant4, together with questions/discussion
  - *Gather feedback, suggestions on progress, recommendations on next steps.*
  - *https://indico.cern.ch/event/1332507/*
  - **Results shown today are drawn from the presentations at the meeting**
- Panel Report available on request, but **very positive**!
  - *One main recommendation was **closer co-working with experiments**, see later for current status here*
- AdePT/Celeritas teams also held series of discussions on common efforts and areas on potential co-working.
  - ***Continued through bi-weekly technical/hackathon meetings this year***

Geant4 R&D: $e^-/e^+/\gamma$ Particle Transport on GPUs

- AdePT/Celeritas developing data structures, workflows to transport **EM particles($e^-/e^+/\gamma$)** using GPUs
  - *Goal: improve event/s/W throughput for LHC detector simulations*
  - ***Focus on EM showers as most expensive fraction of detector portion of "typical" production runs***
- Integrate with existing CPU Geant4 simulations by "offloading" EM particles to GPU, e.g. via "Fast Simulation" hooks in Geant4, with **main challenges**:
  - *Minimizing number/size of on/offload actions*
  - *Synchronization between CPU/GPU (event boundaries)*
  - *Handing back particles (e.g. exiting particles, hadrons from photonuclear processes) from GPU to CPU*
  - *Allow user-defined actions on GPU, such as field/scoring*



Track-parallel Stepping Workflow on GPUs




```
extend_from_primaries              ▷ Copy primaries to device, create track initializers
while Tracks are alive do
    initialize_tracks              ▷ Create new tracks in empty slots
    pre_step                       ▷ Sample mean free path, calculate step limits
    along_step                     ▷ Propagation, slowing down
    boundary                       ▷ Cross a geometry boundary
    discrete_select                ▷ Discrete model selection
    launch_models                  ▷ Launch interaction kernels for applicable models
    extend_from_secondaries        ▷ Create track initializers from secondaries
end while
```

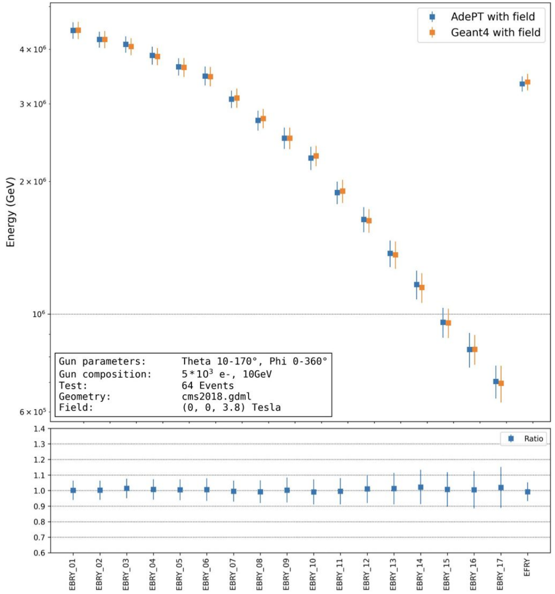
- **CPU**: parallel Events, ***sequential*** **Tracks**
- **CPU+GPU**: parallel Events, ***parallel*** **Tracks (1 per GPU thread)**
    - ***Action*** *based control flow*
    - *Kernels determine next* ***Action****, or perform an* ***Interaction***
    - *Example from Celeritas, AdePT's is similar though with larger, per-particle, kernels*

OAK RIDGE National Laboratory

Credit: Seth Johnson (ORNL)

Physics Validation

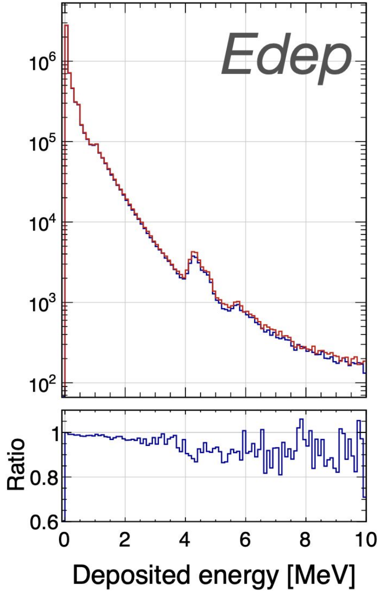
- [G4HepEM](#) in AdePT, CPU/GPU implementation of Geant4 models/data in Celeritas.
- ***Excellent agreement with Geant4, but studies ongoing across problem space***

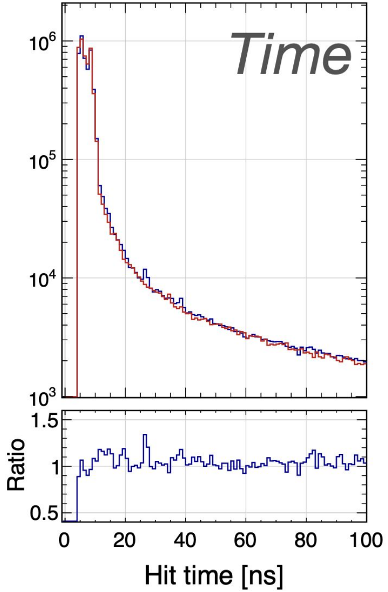


Example: CMS ECAL

← AdePT: 10GeV e⁻

Celeritas: 14TeV $t\bar{t}$ →



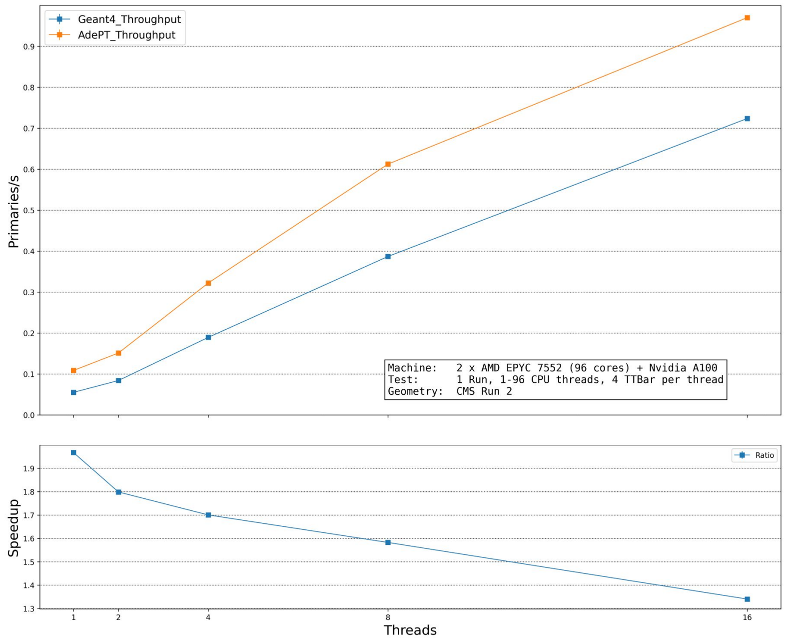


Credits: *[Jonas Hahnfeld (CERN)](#)*, *[Seth Johnson (ORNL), Amanda Lund (Argonne)](#)*

Some Benchmarking Results: AdePT w/CMS2018

- **CMS geometry, 14TeV ttbar Events**
- **2xAMD CPU** feeding **1xA100**
  - ***~90 - 30% speedup*** *as number of threads on CPU increased*

- Decreasing AdePT speedup with increasing CPU threads due to the GPU becoming saturated
  - ***Geometry is a major factor in how quickly this occurs***

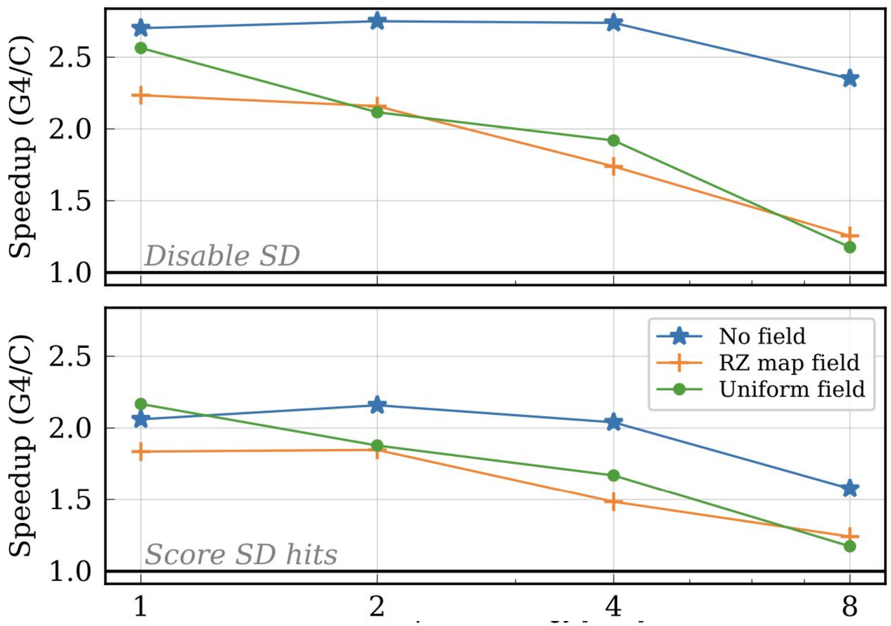


Credit: *[Jonas Hahnfeld (CERN)]*

Some Benchmarking Results: Celeritas w/CMS Run3

- Initial performance comparison in standalone Geant4+Celeritas application
  - *CMS GDML geometry/Sensitive Detectors*
- 8CPU+1GPU standalone simulation with 14TeV tt **17-87% faster**
  - *Theoretical maximum speedup (all e^-/e^+/g tracks take zero time) in full CMSSW ~230%*

Standalone app results (Run 3)

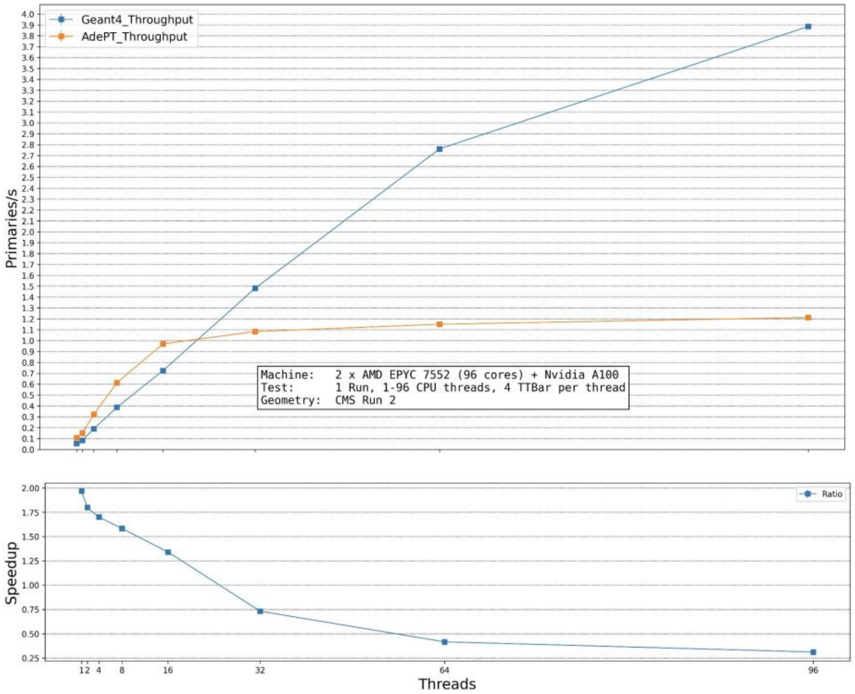


Hardware: Intel Xeon Gold 6152 CPU 22c 2.10GHz + NVIDIA Tesla V100 SXM2
Geometry: CMS detector (Run 3 configuration)
Input: 8 $t\bar{t}$ events @ 14 TeV from LHC pp collision

Credits:Seth Johnson (ORNL), Amanda Lund (Argonne)

Potential Optimization Strategies for Workflow

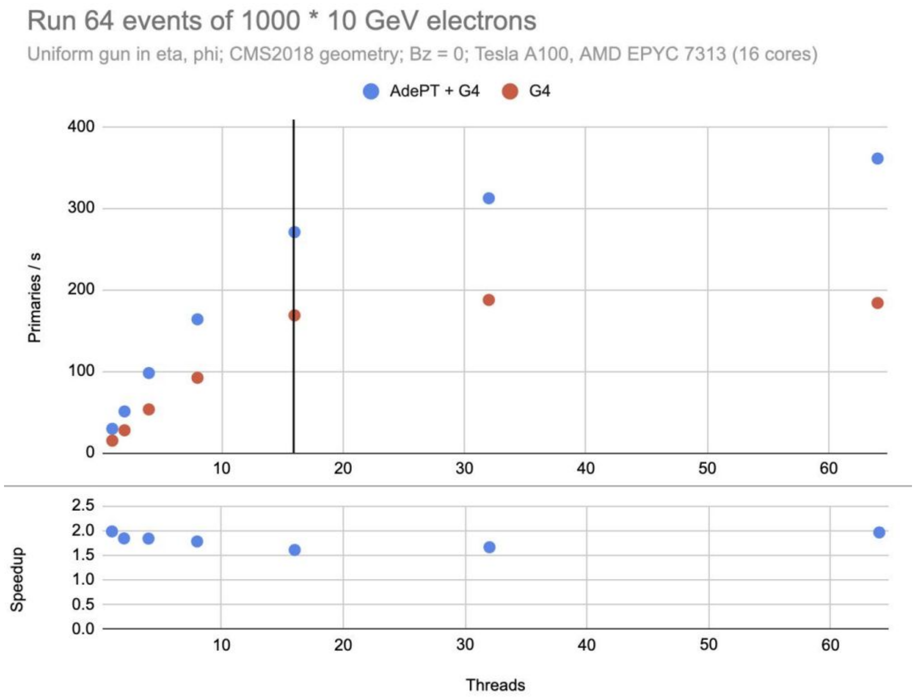
- Sorting tracks on type/energy/etc at specific points
  - *Reduce kernel grid sizes, maybe divergence*
- Use of single/mixed precision
  - *Mostly of benefit to consumer grade GPUs*
- Shared offload "service" to improve CPU/GPU concurrency
  - ***Less blocking of CPU when GPU processes (e.g. Hadrons on CPU whilst GPU processes EM)***
  - *=> Event mixing on device*



Credit: [Juan Gonzalez (CERN)]

Potential Optimization Strategies for Workflow

- Sorting tracks on type/energy/etc at specific points
  - *Reduce kernel grid sizes, maybe divergence*
- Use of single/mixed precision
  - *Mostly of benefit to consumer grade GPUs*
- **Shared offload "service" to improve CPU/GPU concurrency**
  - ***Less blocking of CPU when GPU processes (e.g. Hadrons on CPU whilst GPU processes EM)***
  - ***=> Event mixing on device***



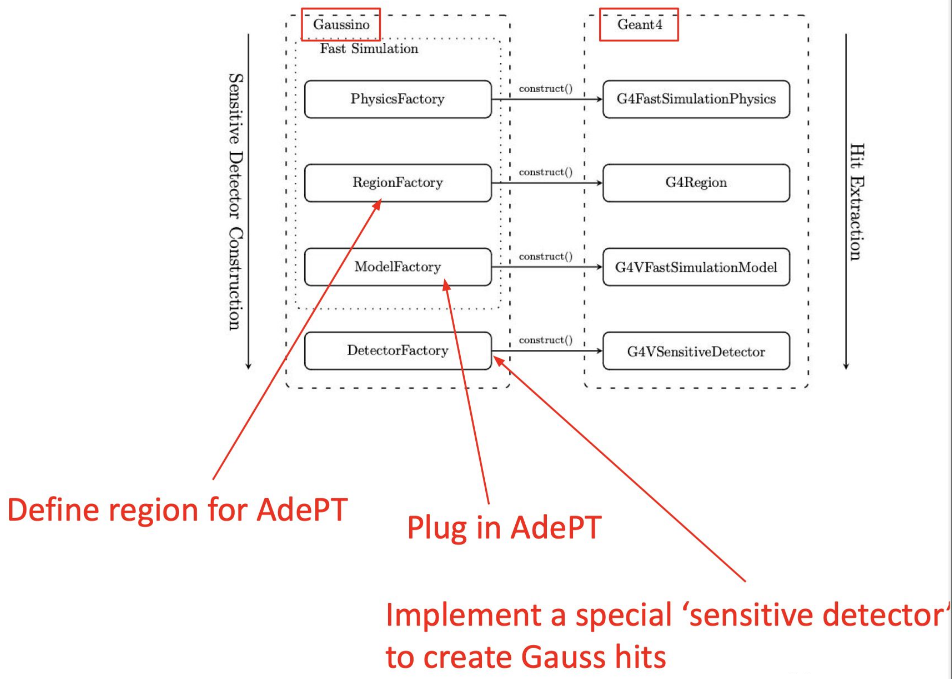
Credit: [Juan Gonzalez (CERN)]

Integration/Testing in Experiments: ATLAS

- TileCal test beam standalone Geant4 application as testbed
  - *Code: https://github.com/lopezzot/ATLTileCalTB (see presentation)*
- **AdePT**: Initial integration by Davide Costanzo:
  - *See https://indico.cern.ch/event/1215829/contributions/5306569/*
- **Celeritas**: integration both standalone and as FullSimLight plugin by Warwick
- **Combined offload interface being prototyped by both teams as outcome of December review/co-working discussions**
  - *Choose AdePT/Celeritas at runtime to compare, help consolidate/test*
  - *Enables easier integration, testing in Athena*
- **Topics being worked on as part of UK contribution to ATLAS Full Simulation WG and both R&D teams**

Integration/Testing in Experiments: LHCb

- **AdePT**: Combine standalone application example with Gauss-on-Gaussino machinery
  - *Fill AdePT pipeline with particles entering LHCb Calo region*
  - *Generate Gauss hits from AdePT (to give equivalence with plain Geant4)*
  - *Working with Juan Bernardo Benavides (LHCb Doctoral Student)*
- **Celeritas**: Warwick will assist in trialling integration of **new combined offload interface** this year.



Integration/Testing in Experiments: CMS, ALICE

- Celeritas already integrated in CMSSW, with testing continuing
  - *CMS also investigating* ***use of G4HepEM, the main physics engine of AdePT***
  - *Review meeting highlighted "tuning" of CPU physics for performance as an important consideration for comparison of CPU/CPU+GPU performance*
  - ***Apples-to-Apples physics comparison may require adding this tuning capability to the GPU codes***

- Preliminary discussions with ALICE on testing integration during December 2023 AdePT/Celeritas review/discussion week

Surface-based Geometry: VecGeom, ORANGE

- **Current CSG model of VecGeom known GPU bottleneck in AdePT and Celeritas**
  - *Divergence from different algorithmic complexity in different solids, etc*
- Effort to develop and use surface- based geometry models, navigation
  - *Reduce divergence from smaller number of surfaces, simpler algorithms*
  - ***VecGeom**: bounded surfaces (explore potential for work reduction in LHC-complexity geometries by reducing checks on "virtual" crossings)*
  - ***ORANGE**: unbounded surfaces (approach from nuclear engineering codes for reactor geometries)*
- Defer to the following presentations at the Geant4 Collaboration meeting for details:
  - *Surface-based GPU model in VecGeom, Andrei Gheata et al*
  - *ORANGE surface geometry progress, Seth Johnson et al*
- **RSEs at Warwick and Sheffield contributing to profiling/optimization under the ExaTEPP grant, common Geometry interface for VecGeom/ORANGE through SWIFT-HEP co-working.**
  - ***VecGeom/ORANGE surface models now in more stable, developed state than previously.***
  - ***Still some CSG->Surface conversion implementations to complete***
- ***Possible commonalities with detray component of ACTS should also be explored further.***

Optical Photon Simulations

- Work continuing at Manchester on use of Mitsuba3 renderer for optical photon simulations
  - *Current work focussing on transport in dense media and geometry input from Geant4*
- Celeritas also starting to develop optical photon physics/transport using same workflow as for EM particles
  - *Interesting comparison with raytracing/rendering methods (Mitsuba, NVidia OptiX)*
- **Broader scope of applications than just LHC (e.g. Dark Matter, Neutrinos, Detector Development) and also a key bottleneck for simulations with optical detectors**
  - ***How can SWIFT-HEP help to engage these communities?***

Other Topics: Power Efficiency, AI methods

- Energy is a critical factor for overall compute budgets, so don't just want to increase Events/Second, but **Events/Second/Watt**!
  - *Results from KEK on Geant4 benchmarks on AMD/Intel/Apple/Fujitsu x86/ARM*
  - *Celeritas have promising preliminary results on CPU+GPU workflows presented at recent HEPiX meeting*
  - ***More work needed in this area**, especially in developing **realistic (geometry, primary events, scoring) benchmark problems** for wider use (e.g. see later presentation on ARM Compute)*
- Interesting presentation at Geant4 Collaboration Meeting on the use of Differentiable programming for simulation:
  - *Related to wider topic of use of AI methods for simulation*
  - *Should we look to engage SWIFT-HEP with broader efforts on "Fast Sim"?*

Slight Aside: Geant4 Work in 2023/24

- Release 11.2 on 8th December 2023: Release Notes
    - *Minor release, with general improvements and fixes*
    - ***Significant UK input** on rollout of Qt6 and VTK visualization support (Cockroft, Warwick)*
- This year, UK (Warwick) will take leading role in Run/Event/Track **workflow management** topics:
    - *Implement new requirements coming from AdePT/Celeritas R&D*
    - *Modernization/Sustainability of Multithreading/Tasking infrastructure (with CERN/JLAB)*
    - *Revival/Modernization of MPI support (with CERN)*
- **Critical that UK maintain contribution to core Geant4 to support HEP in general *and* bring SWIFT-HEP R&D outcomes to production!**

Summary

- AdePT/Celeritas continue to develop and work together
    - *Positive review meeting in December 2023*
    - *Testing in LHC frameworks underway*
- **Co-working underway on common APIs/codes**
    - *Currently on Geant4<->GPU offload*
    - *Next on geometry calls*
- Optical photon on GPU work continuing, an important capability for many experiments
- **Benchmarking for Power Efficiency with realistic problems increasingly important**