EPJ Nuclear Sci. Technol. **11**, 5 (2025)

https://doi.org/10.1051/epjn/2024034

EPJ Nuclear Sciences & Technologies

Available online at:
https://www.epj-n.org

Status and advances of Monte Carlo codes for particle transport simulation, edited by Andrea Zoia, Cheikh Diop and Cyrille de Saint Jean

REGULAR ARTICLE

OPEN ACCESS

Status of GPU capabilities within the Shift Monte Carlo radiation transport code*

Elliott Biondo*, Gregory Davidson, Thomas Evans, Steven Hamilton, Seth Johnson, Tara Pandya, Katherine Royston and José Salcedo-Pérez

Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge TN, 37830, USA

Received: 12 June 2024 / Received in final form: 18 October 2024 / Accepted: 2 December 2024

Abstract. Shift is a general-purpose Monte Carlo (MC) radiation transport code for fission, fusion, and national security applications. Shift has been adapted to efficiently run on GPUs in order to leverage leadership-class supercomputers. This work presents Shift's current GPU capabilities. These include core radiation transport capabilities for eigenvalue and fixed-source simulations, and support for non-uniform domain decomposition, Doppler broadening, free-gas elastic scattering, general-purpose geometry, hybrid MC/deterministic transport, and depletion. Transport results demonstrate a 2–5× GPU-to-CPU speedup on a per-node basis for an eigenvalue problem on the Frontier supercomputer and a 28× speedup for a fixed-source problem on the Summit supercomputer.

1 Introduction

Shift is a general-purpose Monte Carlo (MC) radiation transport code developed at Oak Ridge National Laboratory (ORNL) for fission, fusion, and national security applications [1]. Shift is distributed as part of the SCALE package [2], which is available for non-commercial use within the United States for free with a Government Use Agreement. Shift was adapted to support GPU execution [3] starting in 2017 as a component of the ExaSMR project [4] within the Exascale Computing Project (ECP). The aim of this project was to perform coupled neutronics and thermal hydraulics analysis on 3D, full-core small modular reactor (SMR) models based on the NuScale design [5,6] (referenced throughout this work) with fresh and depleted fuel, leveraging leadership-class GPU-based supercomputers, including Summit [7] and Frontier [8]. GPUs continue to dominate the high performance computing (HPC) landscape due to their comparatively low energy consumption and applicability to artificial intelligence and machine learning. As a result, demand for GPU execution of radiation transport codes has grown for both fission and non-fission applications, motivated by a desire to take full advantage of available HPC resources. Likewise, Shift's GPU capabilities have been expanded to include almost all features expected from a general-purpose MC code.

This work describes the current status of features available for GPU execution within Shift. It is noted that a full description of Shift's CPU features is found in previous work [1,9]. Section 2 describes the core capabilities of Shift on the GPU. Section 3 describes non-uniform domain decomposition (DD). Section 4 describes the Doppler broadening of cross sections. Section 5 describes free-gas elastic scattering treatment. Section 6 describes general-purpose geometry support. Section 7 describes hybrid MC/deterministic transport. Section 8 describes depletion capabilities. Finally, Section 9 provides concluding remarks and planned future work.

2 Core capabilities

The Shift MC code is written in C++ with abstracted HIP/CUDA device programming models in order to support GPU execution on both NVIDIA and AMD hardware. On the GPU, Shift supports full continuous-energy physics for neutrons and photons, both eigenvalue and fixed-source execution modes, and cell and superimposed Cartesian mesh tallies. All code for pre- and

* e-mail: veb@ornl.gov

post-processing operations such as input parsing, geometry processing, and output writing is shared by the CPU and GPU implementations of Shift. A single executable is used to run Shift on the CPU or GPU, as specified by a single parameter within a Shift input file. Shift achieves inter-node parallelism using MPI, and execution on multiple GPUs is done by assigning one GPU graphics complex die (GCD) per MPI rank.

On the CPU, Shift uses a standard history-based MC algorithm which involves simulating a full particle history with a single thread. In contrast, on the GPU, Shift uses an event-based MC transport algorithm in which operations are reordered to exploit single instruction, multiple threads (SIMT) parallelism [3]. With this approach, particle histories are stored in a large vector and sorted based on the next event (birth, a surface crossing, a collision, etc.) required by each history. Histories with the same next event are processed together within a single GPU kernel launch which results in significantly higher tracking throughput than the history-based algorithm, as it allows for smaller, simpler kernels. The simplicity of these kernels reduces thread divergence, and the small size of the kernels increases GPU occupancy, allowing the GPU to more effectively hide the latency associated with context switching.

Performance results for the fresh fuel NuScale SMR problem are shown in Figure 1. This plot shows that a considerably higher tracking rate is obtained when using all of the GPUs on a single node of Summit or Frontier compared to using all CPUs available on each node. The speedup–i.e., the ratio of single node tracking rates using GPUs vs. CPUs–is 7× and 4× for inactive and active cycles on Summit and 5× and 2× for inactive and active cycles on Frontier. Previous results show similar trends for the depleted fuel version of this problem [3].

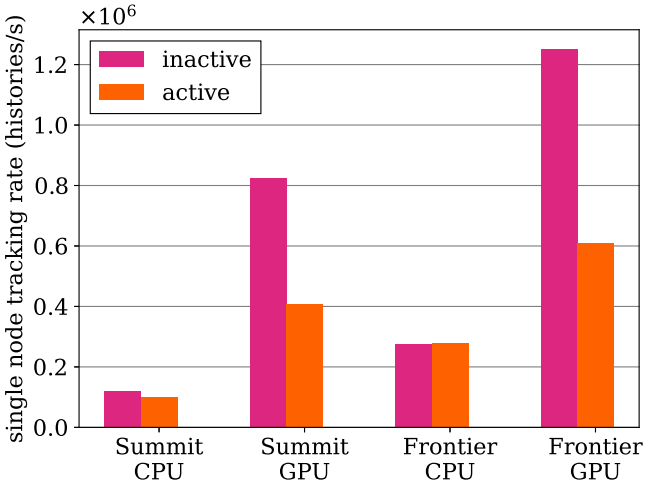


Fig. 1. Comparison between single node tracking rates using CPU- and GPU-only execution on Summit and Frontier for the fresh fuel version of the NuScale SMR problem. GPU results on Summit use a workload of 8.0×10^6 histories/cycle/GCD, and GPU results on Frontier use a workload of 2.8×10^7 histories/cycle/GCD. Total flux and fission rate were tallied on a $119 \times 119 \times 30$ Cartesian mesh during active cycles.

3 Domain decomposition

Certain classes of problems such as depletion problems, problems with large mesh tallies, and shielding problems with large weight window maps, have memory requirements that exceed the memory available on a single processor. In these cases, DD can be used to distribute the problem across multiple processors so that each processor stores a spatially contiguous subset of the problem, referred to as a *block*. Processors can only simulate particles that reside on their block. When particles cross block boundaries, they must be communicated to the adjacent block. Load balancing issues may arise if there is a large discrepancy in the amount of work available on each block, likely resulting from a large gradient in the particle flux.

To address these challenges, Shift uses the Multiple Set, Overlapping Domain (MSOD) method [1] on both the CPU and GPU with non-uniform processor allocation [10,11]. Using this approach, blocks correspond to a Cartesian grid but are expanded outwards to overlap with their neighbors. This minimizes particle communication between blocks by limiting the frequency of particles near block boundaries crossing between blocks multiple times. The ability to assign processors to blocks in a non-uniform manner is allows more processors to be assigned to blocks with high workloads, thus mitigating load imbalance. The number of processors assigned to each block is determined by an initial seed calculation to determine the tracking rate on each block. Figure 2 shows that this non-uniform DD strategy results in higher parallel efficiency for both the fresh fuel and depleted fuel versions of the NuScale SMR problem on Summit.

4 Doppler broadening

Accounting for the temperature dependence of neutron cross sections due to target motion is essential for criticality safety and multiphysics coupling applications. MC codes have traditionally handled this temperature dependence by generating separate Doppler-broadened CE data libraries at a limited number of temperatures ($\sim 10^1$) as a pre-processing step and then interpolating between these libraries or using the nearest library at runtime. A CE data library for one temperature requires ∼1 GB of data. For high-fidelity analysis requiring a refined temperature grid, these memory requirements become burdensome. This issue is exacerbated with the typically limited memory available during GPU execution.

To avoid these memory constraints, Shift uses the windowed multipole (WMP) method [12] for Doppler broadening with both CPU and GPU execution. This method relies on the rigorous pole representation of the cross section [13], which approximates the R-matrix [14]

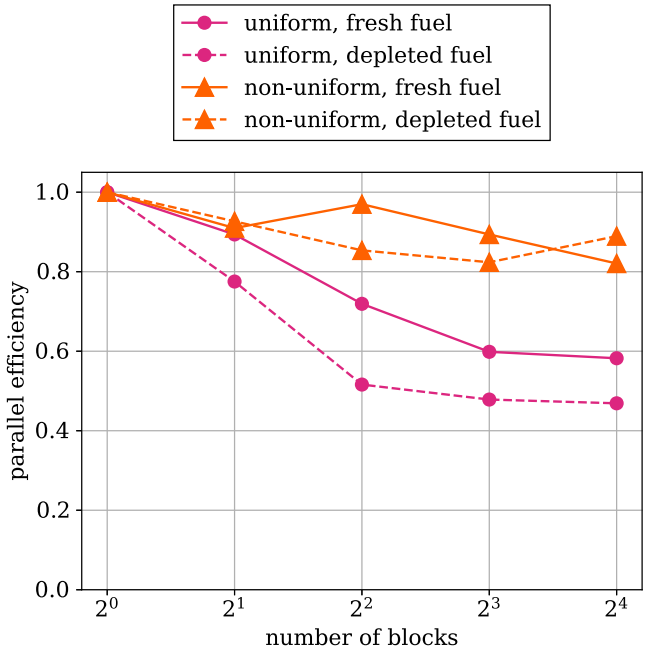


Fig. 2. Parallel efficiency as a function of the number of blocks for uniform and non-uniform DD for fresh fuel and depleted fuel versions of the NuScale SMR problem on Summit. Parallel efficiency is the ratio of the tracking rate with DD to the tracking rate with domain replication for a fixed problem size.

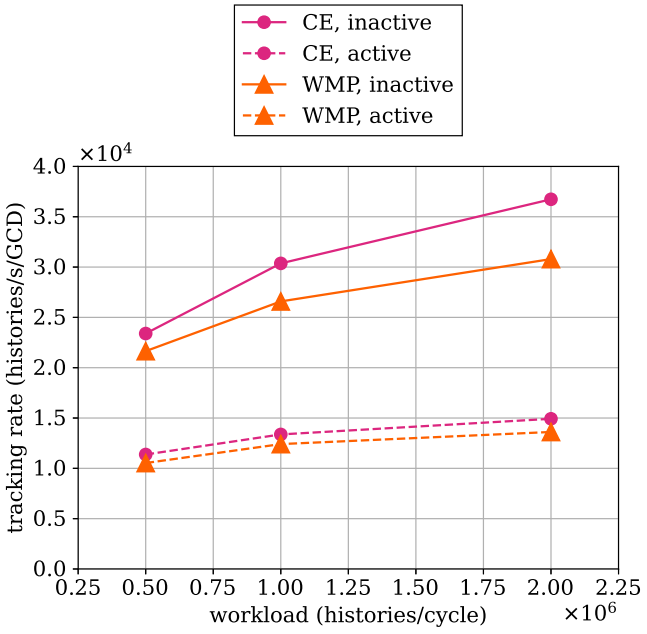


Fig. 3. GPU tracking rate comparison between simulations using the WMP method vs. CE data for the depleted core NuScale SMR problem on a single node of Summit for active and inactive cycles. Tracking rates are per GPU GCD. Total flux and five reaction rates were tallied on a $60 \times 60 \times 10$ Cartesian mesh during active cycles.

form of the cross section as a rational function. A partial fraction expansion of this rational function using the residue method provides an expression for the cross section in terms of poles and residues. Practical considerations provide an expression for the cross section in terms of a subset of these poles lying within a "window" of $\sqrt{E}$-space, with the contributions from the remainder of these poles accounted for using a polynomial fit [12].

A WMP library consists of poles, residues, and polynomials. These parameters are read from disk and used to generate Doppler-broadened cross sections at arbitrary temperatures on the fly, i.e., during the MC simulation. The most computationally intensive aspect of these calculations is evaluating the Faddeeva function, a scaled complex complimentary error function which must be evaluated numerically. Shift implements a Humlicek 8^{th}-order polynomial rational approximation of the Faddeeva function, which has been shown to significantly reduce the computational overhead of the WMP method [15]. The WMP library used by Shift (adapted from Yu [16]) contains data for 359 nuclides with 5 reactions per nuclide (total, elastic scatter, absorption, (n, γ), and fission) and is only 56 MB on disk. Figure 3 shows a GPU tracking rate comparison between simulations using the WMP method vs. using CE data at a single temperature for the depleted core NuScale SMR problem on Summit. The modest performance penalty introduced by using the WMP method is well justified when considering the memory savings are provided.

5 Free-gas elastic scattering

Epithermal scattering events can be treated as elastic collisions with free gas targets (i.e., unbound nuclei). The exit velocity of a neutron emerging from a scattering event is dependent on the target velocity (i.e., the velocity of the nucleus). Target velocities follow a Maxwellian distribution, but target velocities that cause the relative energy of a collision to coincide with a cross section resonance are significantly more likely to cause scattering events.

On the CPU, Shift captures this effect using the Doppler-Broadening Rejection Correction (DBRC) algorithm [17]. However, when DBRC is used for a collision near a resonance, the rejection sampling step has an extremely low sampling efficiency which can introduce thread divergence on the GPU. For the fresh fuel NuScale SMR problem, DBRC increased the average collision processing time on the GPU by a factor of 30, resulting in a factor of five reduction in the overall GPU particle tracking rate.

To avoid this burden, Shift uses the Relative Speed Tabulation (RST) method [18] on the GPU, which uses pre-computed tabular data. For each epithermal energy point in the CE data, a cumulative distribution function (CDF) is tabulated describing the likelihood of a collision occurring as a function of relative velocity (v_r). This is repeated for each temperature within a temperature grid. At runtime, the v_r of a collision is sampled from this tabulated data, which is then used to sample the target velocity v_T, and finally the velocity and direction of the

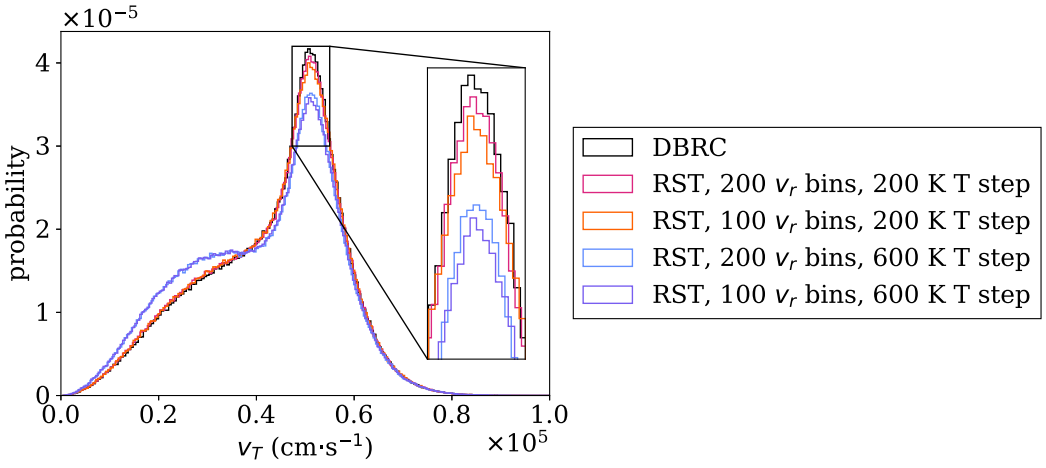


Fig. 4. Distribution of 10^6 target velocities (v_T) sampled using the DBRC and RST methods for an incident neutron energy of 36.25 eV, a ^{238}U target nucleus, and a temperature of 900 K. The RST method is performed with either 100 or 200 relative velocity (v_r) bins and temperature step sizes of 200 K or 600 K. For the 200 K step size, the nearest temperature points occur at 800 K and 1000 K. For the 600 K step size, the nearest temperature points occur at 600 K and 1200 K.

scattered neutron. For collisions at temperatures between grid points, stochastic interpolation is used to select the CDF of the temperature above or below the collision temperature. Shift tabulates v_r CDFs with 200 bins and uses a temperature grid with a maximum step size of 200 K. This configuration provides sufficient accuracy, as shown in Figure 4, with a small memory overhead (∼100 MB total for all applicable nuclides).

6 General-purpose geometry

GPU execution within Shift was originally limited to models specified in Shift's native reactor-specific geometry format, Reactor Tool Kit (RTK). RTK geometries are limited to nested rectilinear arrays of cuboidal pin cells, with each pin consisting of concentric cylinders. Shift can now perform GPU particle tracking on arbitrary constructive solid geometry (CSG), leveraging the Oak Ridge Advanced Nested Geometry Engine (ORANGE) [19]. ORANGE is a fully featured surface-based CSG implementation that supports multi-universe tracking with CSG, rectilinear array, and hexagonal array universes [20].

Within ORANGE CSG universes, tracking operations are accelerated using a bounding interval hierarchy (BIH) [21]. This acceleration structure is constructed by recursively partitioning cell bounding boxes, which is done as a pre-processing step. Particle initialization is accelerated by traversing the BIH, which reduces time complexity from $O(N)$ to $O(\log(N))$, where N is the number of cells in the geometry. Surface crossing operations are also accelerated, producing the same reduction in time complexity with respect to the number of neighbor cells of the surface being crossed.

GPU tracking rate results show that ORANGE is competitive with RTK. For the fresh fuel NuScale SMR problem, ORANGE obtained 83% of the RTK tracking rate on Frontier during inactive cycles (i.e., when geometry tracking makes up the largest fraction of the runtime). The Empire microreactor benchmark [22,23] was selected to demonstrate the use of ORANGE because the hexagonal assemblies cannot be modeled with RTK. Figure 5 shows the neutron flux distribution on the Empire midplane, obtained using 100 nodes of Frontier with 8×10^8 histories/cycle with 120 inactive and 240 active cycles.

7 Hybrid MC/deterministic transport

Shift features implementations of the hybrid MC/deterministic transport methods Consistent Adjoint-Driven Importance Sampling (CADIS) [24] and Forward-Weighted (FW)-CADIS [25] for variance reduction. Both implementations use the parallel Denovo discrete ordinates (S_N) solver [26] to rapidly produce forward and adjoint flux distributions to generate the important functions. The CADIS and FW-CADIS methods differ only in the calculation of the adjoint source used to estimate the final importance function; in FW-CADIS, the adjoint source is weighted by the forward flux, so it requires two S_N calculations (forward and adjoint). (WWs) are calculated as the quotient of the desired response and adjoint flux, where the response is the inner product of the forward source and adjoint flux. Once WWs are known, source particles are sampled from a biased source distribution, and the statistical weights of the particles are controlled using the WW method [27].

In the current GPU implementation of Shift, the Denovo S_N step(s), the calculation of weight windows, and the source biasing are all performed on the CPU, with all of the MC transport–including WW processing–performed on the GPU [28]. All methods are MPI-parallel and support the full MSOD topology. During particle flight, the Shift CPU solver has multiple options for when to check the WWs, including when crossing importance grid boundaries, geometry boundaries, per particle

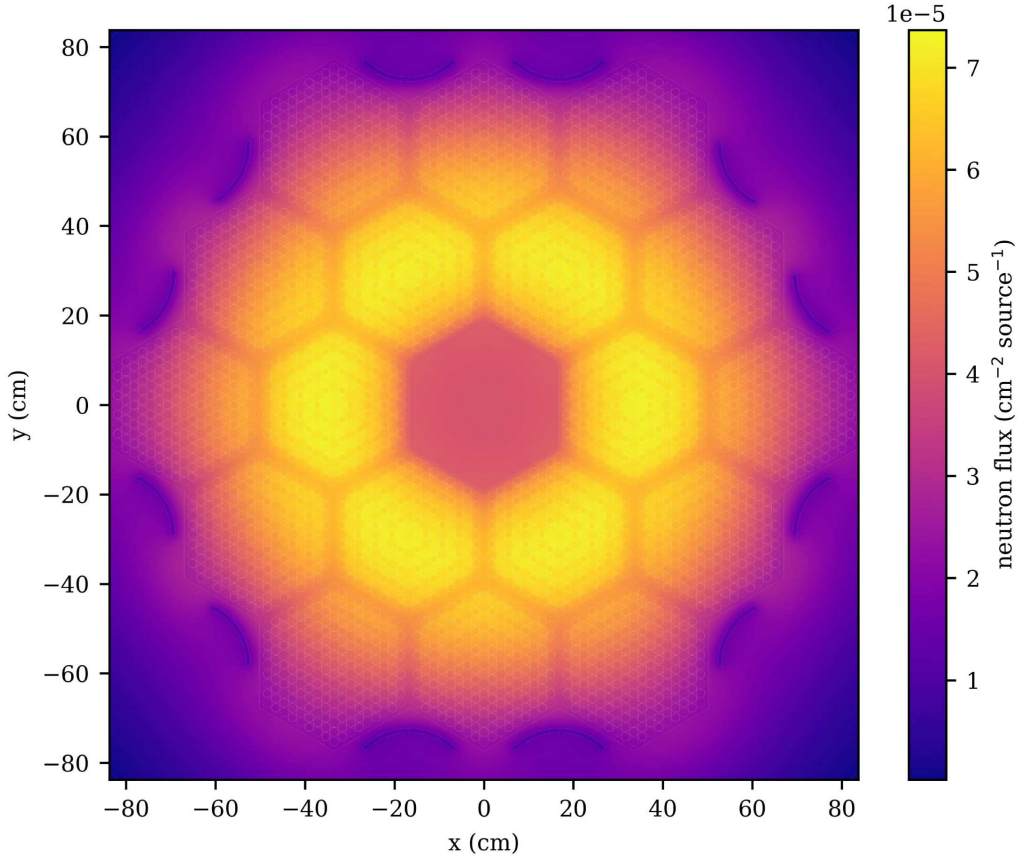


Fig. 5. Neutron flux distribution on the midplane of the Empire problem using 100 nodes of Frontier with 8×10^8 histories/cycle, with 120 inactive and 240 active cycles. Flux was tallied on a $3350 \times 3350 \times 1$ superimposed Cartesian mesh. Statistical errors in the flux on this slice are all less than 0.7% within the core region.

mean-free-path, pre-collision, and post-collision [29]. The GPU implementation supports only the most commonly used options: importance grid tracking and post-collision.

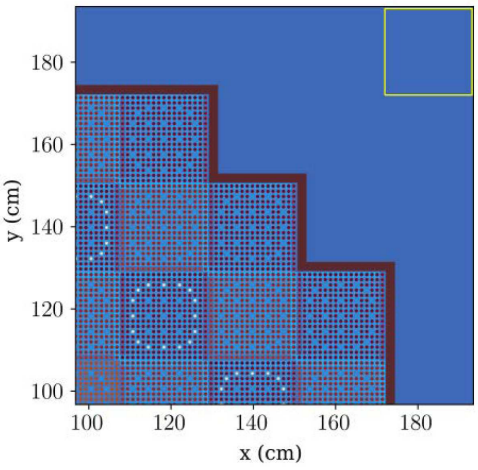
This GPU implementation was demonstrated with a fixed-source excore calculation using a quarter-core model of the fresh fuel NuScale SMR [28]. An xy slice through the geometry is shown in Figure 6a. The FW-CADIS WWs were built to optimize the solution for flux tallies located in the $+x, +y$ corner and along the axial extent of the fuel as shown by the yellow box in Figure 6a. The resulting adjoint function (importance) is shown in Figure 6b. The simulations were run using full domain replication parallelism on Summit using both the CPU and GPU solvers.

Table 1 shows performance results for importance grid tracking plus post-collision WW processing. These results show that on Summit a single GPU GCDGCD provides equivalent performance as 197 CPU cores. On a per node basis, GPUs provide a 28× speedup over CPUs.

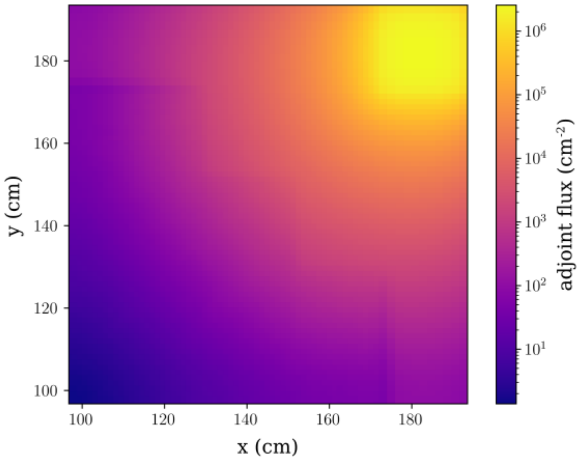
8 Depletion

Depletion simulations with Shift are enabled through a parallel domain-decomposed coupling with the Oak Ridge Isotope Generation (ORIGEN) depletion solver [2]. The initial CPU-only coupling used fine-group flux tallies to collapse one-group cross sections for use by the depletion solver [30]. Although this approach is computationally efficient, it can be prohibitively memory intensive due to the large number of energy groups (up to 43,000) required to accurately capture detailed resonance structure when integrating reaction rates. A less memory-intensive strategy is to directly tally reaction rates for every nuclide/reaction combination needed for depletion, but this dramatically increases the computational effort required by the transport solver, as many additional cross sections must be evaluated while transporting particle histories.

To balance computational efficiency and memory requirements, a hybrid approach has been implemented in Shift on both the CPU and GPU [31]. With this approach, selected nuclide/reaction combinations use reaction rate tallies, and the remaining nuclide/reaction combinations use a flux tally. Because ORIGEN has not been ported to run on GPUs, the depletion solve is performed on CPUs even when the Shift GPU solver is used. Initial performance testing led to an interesting observation: the time required to collapse one-group cross sections from flux tallies (which is performed on the CPU)



(a)



(b)

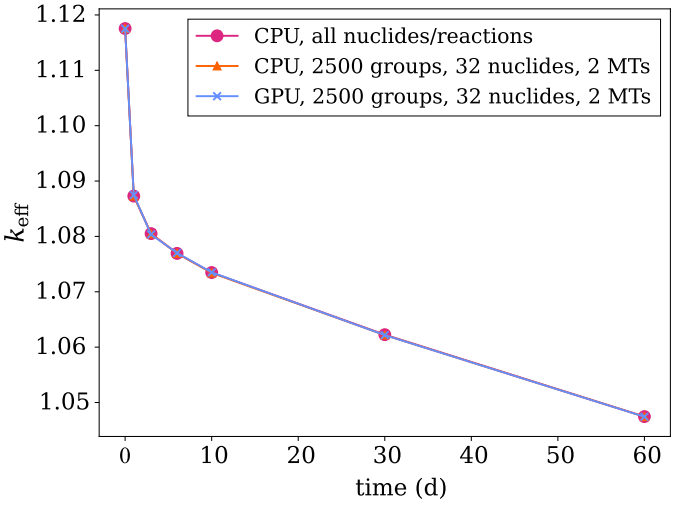
Fig. 6. Fresh fuel NuScale SMR excore problem *xy* slice at $z = 100\,\text{cm}$: (a) geometry (tally region indicated by yellow box) and (b) total adjoint flux.

frequently comprised a significant portion of the overall runtime. The reason for this is that Shift executes with a single MPI rank per GPU GCD, so the amount of computational power on the CPU is limited. Because most GPU computing systems have more CPUs than GPU GCDs on each node, this approach leaves many of the CPUs idle. As a remedy, Shift was updated to use CPU-based multithreading with OpenMP to parallelize the cross section collapse across multiple CPU cores. This approach has reduced the time required for the cross section collapse sufficiently such that it no longer significantly limits the performance of GPU depletion calculations.

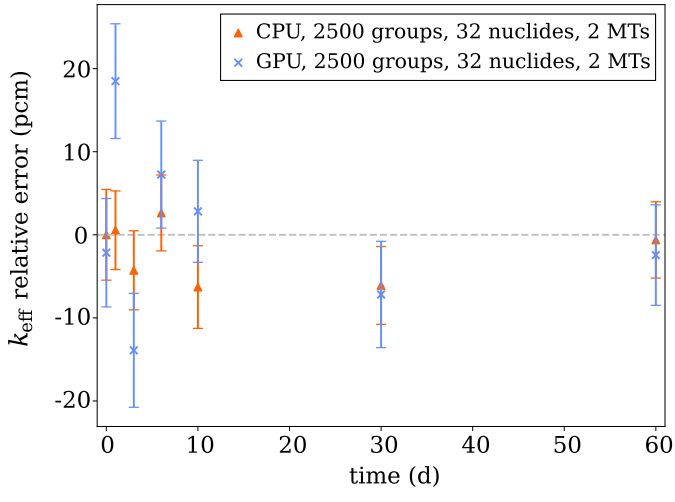
Figure 7 illustrates the accuracy of a hybrid depletion tally with two reaction rates directly tallied for 32 selected nuclides, and a flux tally with 2500 energy groups for all other nuclides/reactions, for a CE adaptation of the 3D C5G7 benchmark problem [32]. In Figure 7a, the k_{eff} for a reference case run on the CPU with all nuclides/reactions

Table 1. Tracking rates using importance grid tracking and post-collision WW tracking for the fresh fuel NuScale SMR excore problem on Summit. All simulations were run with 2.4×10^8 total histories and 1 MPI rank per GCD.

			Tracking rate (histories/s)		
Arch.	Ranks	Nodes	Total	per GCD	per node
CPU	1176	28	3.7×10^5	3.2×10^2	1.3×10^4
GPU	24	4	1.5×10^6	6.2×10^4	3.7×10^5



(a)



(b)

Fig. 7. Comparison of (a) k_{eff} calculated with all nuclides/reactions tallied and hybrid depletion tallies for both the Shift CPU and GPU solvers, and (b) k_{eff} error relative to the all nuclides/reactions solution, for a CE adaptation of the 3D C5G7 model.

tallied is plotted along with the hybrid tally results with the CPU and GPU solvers. The error in the hybrid CPU

and GPU solutions relative to the reference case is shown in Figure 7b along with the 1σ statistical uncertainty and shows that the hybrid tally k_{eff} is within 20 pcm of the reference solution at all time steps.

9 Conclusion and future work

Shift has a mature set of capabilities to enable high-fidelity analysis on leadership-class supercomputers for a variety of applications. Support for Doppler broadening, free-gas elastic scattering, and depletion will facilitate multiphysics simulations of next-generation fission reactors. Support for fixed-source execution, general-purpose CSG, and hybrid MC/deterministic transport will enable shielding calculations for fusion and national security applications. The per-node speedups demonstrated in this work—2–5× for an eigenvalue problem on Frontier and 28× for a fixed-source problem on Summit—coupled with the improved memory utilization provided by DD, will allow for higher fidelity simulations across all application spaces. Future work will include GPU support for coupled neutron-photon transport, photo-fission and photo-neutron reactions, and thick target Bremsstrahlung treatment for accelerator applications [33]. Hexagonal and cylindrical mesh tallies, as well as surface tallies, will be added for detailed fission analysis. Finally, support for MCNP [34] and/or computer-aided design (CAD) geometry may be added to facilitate fusion analysis.

Acknowledgments

The authors thank Matthew Jessee and William Wieselquist for their internal technical review.

Funding

This research was supported by the Exascale Computing Project (ECP), project number 17-SC-20-SC. The ECP is a collaborative effort of two DOE organizations, the Office of Science and the National Nuclear Security Administration, that are responsible for the planning and preparation of a capable exascale ecosystem–including software, applications, hardware, advanced system engineering, and early testbed platforms–to support the nation's exascale computing imperative.

Conflicts of interest

The authors declare that they have no competing interests to report.

Data availability statement

Data presented within this work may be available upon request.

Author contribution statement

All authors made significant methodology, software, and analysis contributions directly related to this work.

References

1. T.M. Pandya, S.R. Johnson, T.M. Evans, G.G. Davidson, S.P. Hamilton, A.T. Godfrey, Implementation, capabilities, and benchmarking of Shift, a massively parallel Monte Carlo radiation transport code, J. Comput. Phys. **308**, 239 (2016)
2. W.A. Wieselquist, R.A. Lefebvre, *SCALE 6.3.2 User Manual* (Oak Ridge National Laboratory, Oak Ridge, TN, USA, 2024) ORNL/TM-2024/3386
3. S.P. Hamilton, T.M. Evans, Continuous-energy Monte Carlo neutron transport on GPUs in the Shift code, Ann. Nucl. Energy **128**, 236 (2019)
4. E. Merzari, S. Hamilton, T. Evans, M. Min, P. Fischer, S. Kerkemeier et al., Exascale multiphysics nuclear reactor simulations for advanced designs, in *International Conference for High Performance Computing, Networking, Storage and Analysis* (Denver, CO, USA, 2023), p. 1
5. NuScale Power, LLC, About Us; 2023, nuscalepower.com/en/about
6. K. Smith, NuScale Small Modular Reactor (SMR) Progression Problems for the ExaSMR Project, Exascale Computing Project; 2017, WBS 1.2.1.08 ECP-SE-08-43
7. Oak Ridge Leadership Computing Facility, Summit: Oak Ridge National Laboratory's 200 petaflop supercomputer; 2023, olcf.ornl.gov/olcf-resources/compute-systems/summit
8. Oak Ridge Leadership Computing Facility, Frontier; 2023, olcf.ornl.gov/olcf-resources/compute-systems/frontier
9. S.R. Johnson, T.M. Evans, G.G. Davidson, S.P. Hamilton, T.M. Pandya, K.E. Royston et al. *Omnibus User Manual* (Oak Ridge National Laboratory, Oak Ridge, TN, USA, 2020), ORNL/TM-2018/1073
10. J.A. Ellis, T.M. Evans, S.P. Hamilton, C.T. Kelley, T.M. Pandya, Optimization of processor allocation for domain decomposed Monte Carlo calculations, Parallel Comput. **87**, 77 (2019)
11. S.P. Hamilton, T.M. Evans, K.E. Royston, E.D. Biondo, Domain decomposition in the GPU-accelerated Shift Monte Carlo code, Ann. Nucl. Energy **166**, 108687 (2022)
12. C. Josey, P. Ducru, B. Forget, K. Smith, Windowed multipole for cross section Doppler broadening, J. Comput. Phys. **307**, 715 (2016)
13. R.N. Hwang, A rigorous pole representation of multilevel cross sections and its practical applications, Nucl. Sci. Eng. **96**, 192 (1987)
14. E.P. Wigner, L. Eisenbud, Higher angular momenta and long range interaction in resonance reactions, Phys. Rev. **72**, 29 (1947)
15. B. Forget, J. Yu, G. Ridley, Performance improvements of the windowed multipole formalism using a rational fraction approximation of the Faddeeva function, in *International Conference on Physics of Reactors* (Pittsburgh, PA, USA, 2022), p. 1963
16. J. Yu, wmp-endfbvii.1, 2022, Commit: 7887b3144cc579c4a449c2d82ba781ae93617da3, github.com/jiankai-yu/wmp-endfbvii.1
17. B. Becker, R. Dagan, G. Lohnert, Proof and implementation of the stochastic formula for ideal gas, energy dependent scattering kernel, Ann. Nucl. Energy **36**, 470 (2009)
18. N. Choi, H.G. Joo, Relative speed tabulation method for efficient treatment of resonance scattering in GPU-based Monte Carlo neutron transport calculation, Nucl. Sci. Eng. **195**, 954 (2021)
19. S.R. Johnson, R. Lefebvre, K. Bekar, ORANGE: Oak Ridge Advanced Nested Geometry Engine (Oak Ridge National Laboratory, 2023), ORNL/TM-2023/3190

20. E. Biondo, T. Evans, S. Johnson, S. Hamilton, Comparison of nested geometry treatments within GPU-based Monte Carlo neutron transport simulations of fission reactors, International Journal of High Performance Computing Applications, 2024, Submitted March, 2024
21. C. Wächter, A. Keller, Instant ray tracing: The bounding interval hierarchy, in *Symposium on Rendering*, edited by T. Akenine-Moeller, W. Heidrich, The Eurographics Association, (2006), pp. 139–49
22. C. Lee, Y.S. Jung, Z. Zhong, J. Ortensi, V. Laboure, Y. Wang et al. *Assessment of the Griffin Reactor Multiphysics Application Using the Empire Micro Reactor Design Concept* (Argonne National Laboratory and Idaho National Laboratory, 2020), ANL/NSE-20/23 and INL/LTD-20-59263
23. C. Matthews, V. Laboure, M. DeHart, J. Hansel, D. Andrs, Y. Wang et al., Coupled multiphysics simulations of heat pipe microreactors using DireWolf, Nucl. Technol. **207**, 1142 (2021)
24. A. Haghighat, J.C. Wagner, Monte Carlo variance reduction with deterministic importance functions, Prog. Nucl. Energy **42**, 25 (2003)
25. J. Wagner, D. Peplow, S. Mosher, FW-CADIS method for global and regional variance reduction of Monte Carlo radiation transport calculations, Nucl. Sci. Eng. **176**, 37 (2014)
26. T.M. Evans, A.S. Stafford, R.N. Slaybaugh, K.T. Clarno, Denovo: A new three-dimensional parallel discrete ordinates code in SCALE, Nucl. Technol. **171**, 171 (2010)
27. T.E. Booth, *A Sample Problem for Variance Reduction in MCNP* (Los Alamos National Laboratory, 1985), LA-10363-MS
28. K. Royston, T. Evans, S.P. Hamilton, G. Davidson, Weight window variance reduction on GPUs in the Shift Monte Carlo Code, in *International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering* (Niagara Falls, ON, Canada, 2023), p. 1
29. E.S. Gonzalez, G.G. Davidson, Choosing transport events for initiating splitting and rouletting, J. Nucl. Eng. **2**, 97 (2021)
30. G.G. Davidson et al., Nuclide depletion capabilities in the Shift Monte Carlo code, Ann. Nucl. Energy **114**, 259 (2018)
31. J.L. Salcedo-Pérez, B. Forget, K. Smith, P. Romano, Hybrid tallies to improve performance in depletion Monte Carlo simulations, in *International Conference on Mathematics & Computational Methods Applied to Nuclear Science and Engineering* (Portland, OR, USA, 2019), p. 927
32. E.E. Lewis, M.A. Smith, N. Tsoulfanidis, G. Palmiotti, T.A. Taiwo, R.N. Blomquist, Benchmark specification for Deterministic 2-D/3-D MOX fuel assembly transport calculations without spatial homogenization (C5G7 MOX) (Nuclear Energy Agency and Nuclear Science Committee, 2001), NEA/NSC/DOC(2001)4
33. J. Brown, C. Celik, T. Evans, B. Jeon, R. Lefebvre, J. McDonnell et al., *Photonuclear Physics in SCALE* (Oak Ridge National Laboratory, Oak Ridge, TN, USA, 2023), ORNL/TM-2023/3201
34. D.B. Pelowitz, *MCNP6 User's Manual Version 1.0* (Los Alamos National Laboratory, 2013), LA-CFP-13-00634 Rev 0

Cite this article as: Elliott Biondo, Gregory Davidson, Thomas Evans, Steven Hamilton, Seth Johnson, Tara Pandya, Katherine Royston, José Salcedo-Pérez. Status of GPU capabilities within the Shift Monte Carlo radiation transport code, EPJ Nuclear Sci. Technol. **11**, 5 (2025) https://doi.org/10.1051/epjn/2024034.