New capabilities of the Monte Carlo dose engine ARCHER-RT: Clinical validation of the Varian TrueBeam machine for VMAT external beam radiotherapy

David P. Adam
Medical Physics, University of Wisconsin Madison, 1111 Highland Avenue, Madison, WI 53705, USA

Tianyu Liu
Nuclear Engineering Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

Peter F. Caracappa
Virtual Phantoms, Inc, Albany, NY 12205, USA

Bryan P. Bednarz
Medical Physics, University of Wisconsin Madison, 1111 Highland Avenue, Madison, WI 53705, USA

Xie George Xu[a)]
Nuclear Engineering Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA

(Received 24 June 2019; revised 3 February 2020; accepted for publication 10 February 2020; published xx xxxx xxxx)

Purpose: The Monte Carlo radiation transport method is considered the most accurate approach for absorbed dose calculations in external beam radiation therapy. In this study, an efficient and accurate source model of the Varian TrueBeam 6X STx Linac is developed and integrated with a fast Monte Carlo photon-electron transport absorbed dose engine, ARCHER-RT, which is capable of being executed on CPUs, NVIDIA GPUs, and AMD GPUs. This capability of fast yet accurate radiation dose calculation is essential for clinical utility of this new technology. This paper describes the software and algorithmic developments made to the ARCHER-RT absorbed dose engine.
Methods: AMD's Heterogeneous-Compute Interface for Portability (HIP) was implemented in ARCHER-RT to allow for device independent execution on NVIDIA and AMD GPUs. Architecture-specific atomic-add algorithms have been identified and both more accurate single-precision and double-precision computational absorbed dose calculation methods have been added to ARCHER-RT and validated through a test case to evaluate the accuracy and performance of the algorithms. The validity of the source model and the radiation transport physics were benchmarked against Monte Carlo simulations performed with EGSnrc. Secondary dose-check physics plans, and a clinical prostate treatment plan were calculated to demonstrate the applicability of the platform for clinical use. Absorbed dose difference maps and gamma analyses were conducted to establish the accuracy and consistency between the two Monte Carlo models. Timing studies were conducted on a CPU, an NVIDIA GPU, and an AMD GPU to evaluate the computational speed of ARCHER-RT.
Results: Percent depth doses were computed for different field sizes ranging from 1.5 cm^2 × 1.5 cm^2 to 22 cm^2 × 40cm^2 and the two codes agreed for all points outside high gradient regions within 3%. Axial profiles computed for a 10 cm^2 × 10 cm^2 field for multiple depths agreed for all points outside high gradient regions within 2%. The test case investigating the impact of native single-precision compared to double-precision showed differences in voxels as large as 71.47% and the implementation of KAS single-precision reduced the difference to less than 0.01%. The 3%/3mm gamma pass rates for an MPPG5a multileaf collimator (MLC) test case and a clinical VMAT prostate plan were 94.2% and 98.4% respectively. Timing studies demonstrated the calculation of a VMAT plan was completed in 50.3, 187.9, and 216.8 s on an NVIDIA GPU, AMD GPU, and Intel CPU, respectively.
Conclusion: ARCHER-RT is capable of patient-specific VMAT external beam photon absorbed dose calculations and its potential has been demonstrated by benchmarking against a well validated EGSnrc model of a Varian TrueBeam. Additionally, the implementation of AMD's HIP has shown the flexibility of the ARCHER-RT platform for device independent calculations. This work demonstrates the significant addition of functionality added to ARCHER-RT framework which has marked utility for both research and clinical applications and demonstrates further that Monte Carlo-based absorbed dose engines like ARCHER-RT have the potential for widespread clinical implementation.
 [https://doi.org/10.1002/mp.14143]

Key words: dose calculation, GPU, Monte Carlo, radiation therapy, software

1. INTRODUCTION

External beam radiotherapy (EBRT) has remained an essential modality in the armamentarium of oncologists over the last several decades. Over 50% of all cancer patients receive EBRT for curative or palliative purposes.[1] EBRT requires accurate absorbed dose computation to safely and effectively deliver radiation treatment regimens to patients. Monte Carlo methods are regarded as the "gold standard" for performing these absorbed dose calculations.[2] Thoroughly benchmarked general-purpose Monte Carlo codes have been used for decades to support research efforts related to EBRT including EGSnrc,[3,4] MCNP,[5–7] PENELOPE,[8,9] and GEANT4.[10,11] Although highly accurate, Monte Carlo-based absorbed dose calculation engines require a large amount of computational resources. To this end, several CPU-based Monte Carlo codes that utilize algorithmic approximations and modifications have been developed to improve run-time efficiency including DPM,[12] VMC++,[13] and MCDOSE.[14] Additionally, several groups have demonstrated significant performance gains in comparison to CPU-based MC codes for EBRT Monte Carlo simulations on graphics processing units (GPUs).[15–20]

While GPU-accelerated Monte Carlo codes exhibit desirable performance characteristics, the accuracy of these codes can be compromised by GPU architecture-specific optimization considerations, mainly, (a) the necessity of atomic-add operations to accurately tally absorbed dose and (b) the utilization of single vs double-precision floating point representation of real numbers. For Monte Carlo radiotherapy absorbed dose calculations, a single array shared by all threads is implemented to accumulate tally data because GPU thread-specific memory is usually not large enough to hold an entire local tally array. Atomic-add operations are required to avoid race conditions where two threads try to update the same memory location.

There are also trade-offs in accuracy vs performance when utilizing single-precision vs double-precision. Using single-precision generally results in better performance (at least twice as fast[21]) at the expense of accuracy, whereas using double-precision results in better accuracy at the expense of performance. To date, single-precision has been widely adopted in various other studies of GPU-accelerated Monte Carlo absorbed dose calculation due to better performance and the lack of support of double-precision numbers on several GPU architectures. Specifically, NVIDIA GPUs prior to the Pascal generation and all AMD GPUs do not offer the same hardware level atomic-add operation support for double-precision numbers and suboptimal software emulation can be prohibitively slow. One major drawback of using native implementations of single-precision atomic-add operations is that round-off errors can occur in highly absorbed dose regions, where small absorbed dose increments are added to counters with large absorbed dose values. The lower digits of the absorbed dose increments may be truncated, resulting in an underestimate of the absorbed dose. Magnoux et al. found in voxel-based absorbed dose calculations that single-precision calculations can differ from double-precision calculations by over 40%.[21] Liu et al. found that in CT scan absorbed dose calculations the lung dose can be underestimated by as much as 20% using single-precision calculation methods.[22]

Additionally, the majority of the aforementioned GPU codes all rely on the CUDA architecture, which is not desirable in terms of portability. Tian et al. utilized OpenCL to promote an architecture independent simulation platform,[23,24] but computational speeds for this OpenCL platform were slightly slower than its CUDA counterpart and as of 2018, OpenCL is deprecated on MacOS.

This work builds upon our previously published ARCHER-RT work[25] and seeks to address these pitfalls by (a) describing and demonstrating the implementation of a single-precision Kahan summation-based atomic-add algorithm to ensure dosimetric accuracy for all GPU architectures, (b) describing architecture-specific optimal atomic-add algorithms to provide older NVIDIA GPUs software emulation of double-precision atomic-add methods to ensure accurate absorbed dose calculations, (c) implementing ARCHER-RT on AMD's Heterogeneous-Compute Interface for Portability (HIP), a C++ Application Programming Interface (API) which allows for device-independent execution on NVIDIA and AMD GPUs, and (d) adding VMAT capable source modeling by utilizing patient-independent phase spaces as input and performing radiation transport simulations through patient-dependent collimators into patient geometries. We demonstrate this through the modeling and benchmarking of a flattened photon 6 MV Varian TrueBeam linear accelerator in the ARCHER-RT framework.

2. METHODS

2.A. ARCHER-RT description

The following section describes both the software design and the underlying physics models used in absorbed dose calculation engine called ARCHER-RT (Accelerated Radiation-transport Computations in Heterogeneous EnviRonments).

2.A.1. Software design for GPU and multithread CPU codes

ARCHER-RT is designed and optimized for both CPU and GPU processors.[26–28] The CPU code uses open multiprocessing (OpenMP) API for parallel computing, whereas the GPU code uses HIP) — a new C++ API developed by AMD.[29] The advantage of choosing HIP is portability and simplicity. HIP allows the same source code to be compiled into different binaries to run on AMD and NVIDIA GPUs respectively. On AMD's platform, the HIP functions are compiled into the Instruction Set Architecture (ISA) of AMD GPUs using the hcc compiler, whereas on NVIDIA's, they are wrappers of their CUDA counterparts (for instance, hipMalloc calls cudaMalloc) and are compiled into NVIDIA GPUs' ISA using the nvcc compiler driver. In ARCHER-RT,

there are only a few cases where HIP cannot directly provide a uniform interface, due to architectural differences between AMD and NVIDIA GPUs, and platform-dependent code becomes necessary. For example, a warp consists of 32 threads on NVIDIA GPUs but 64 on AMD's. Consequently, in our highly-optimized absorbed dose accumulation functions, the intrinsic function __popc (requiring a 32-bit integer parameter, each bit representing the status of a lane in a warp) should be exclusively applied to NVIDIA GPUs and __popcll (requiring a 64-bit integer parameter) to AMD GPUs.[30] HIP also provides an element of simplicity. HIP is a high-level API designed to closely resemble the syntax of CUDA runtime API, the use of which has dominated NVIDIA GPU-accelerated applications. HIP significantly reduces the amount of boilerplate code required by some alternative GPU computing APIs such as OpenCL or CUDA driver API.

ARCHER-RT is written in C++11, with a strong emphasis on both performance (achieved by low-level optimizations) and maintainability (using object-oriented design). All computationally intensive components are GPU accelerated (i.e. phase space particle coordinate transforms, particle transport in the Linac source model, and particle transport in the patient). Meanwhile, the multithreaded CPU fallbacks are available, serving two purposes: to allow fair CPU-GPU performance comparison, where both codes are sufficiently parallelized and optimized, and to allow code verification, where the developed GPU code is constantly verified with the CPU code in a variety of unit tests. The absorbed dose engine in ARCHER-RT supports both single-precision and double-precision formats implemented by a C++ template. In general, the GPU results are expected to have 10–15 identical digits with the CPU result for double-precision, and 4–7 for single-precision.

2.A.2. Architecture-specific atomic-add methods for single and double-precision

ARCHER-RT is now designed to operate in either single or double-precision modes depending upon the needs and constraints of the end-user. In each of these modes atomic-add algorithms are implemented to provide optimal performance when used for GPU-accelerated absorbed dose computation.[30]

Although double-precision is significantly more reliable than native single-precision for absorbed dose accumulation calculations, NVIDIA GPUs prior to the Pascal generation and all existing AMD GPUs do not directly support double-precision "atomic-add" operations at the hardware level. GPU implementation of "compare-and-swap" (CAS) is a commonly used algorithm for the software emulation of double-precision "atomic-add" operations but can be prohibitively slow. In the CAS algorithm, a calling thread adds the absorbed dose increment to a memory location using a do-while loop structure. Specifically, the calling thread repeatedly executes instructions in the do-while loop until no other thread in the same warp has meanwhile updated the same memory location. This takes indefinite number of steps to complete. The root cause of CAS low performance lies in GPU's nominal SIMT (single instruction, multiple threads) architecture where threads in a warp execute the same instruction at nearly the same time. If several threads in a warp attempt to update the same memory location, contention arises and each thread tend to repeat the do-while loop indefinitely many times before succeeding.

ARCHER-RT addresses the limitation of CAS by implementing the Warp-aggregated method (WAG) algorithm to eliminate intra-warp thread contention on supportive architectures. The WAG reduces intra-warp contention by utilizing GPU warp shuffle methods.[31,32] The general idea behind WAG is that threads of a warp attempting to update the same memory location are put into the same subgroup. In each subgroup, a leader thread sums all absorbed dose increments from its peer threads and singly updates the memory location without contention. It is worth pointing out that theoretically inter-warp thread contention exists as well but happens significantly less frequently,whereas GPUs implement SIMT for threads within the same warp, threads from different warps are not synchronized on the GPU grid level by design. This means that the inter-warp thread contention in the WAG algorithm theoretically exists but only occurs by chance and is significantly less likely than the intra-warp thread contention in the CAS algorithm caused by GPU's SIMT design.

ARCHER-RT supports hardware-level double-precision atomic-add operations for NVIDIA GPUs following the Pascal generation. While recent iterations of NVIDIA GPUs enable double-precision atomic-add operations at the hardware level, the problem of thread contention, despite being less severe than the software counterpart, still inevitably exists, where threads in a warp updating the same memory location are serialized by the hardware. In ARCHER, WAG is used again to have the leader threads alone update the memory locations and avoid serialization.

For architectures that do not support double-precision, single-precision must be used. Bossler et al.[33] and Liu et al.[30] have performed studies to identify appropriate single-precision atomic-add algorithms that retain accuracy at only a small performance penalty. Based on these results, the Kahan summation algorithm (KAS) has been implemented in ARCHER-RT to allow for both fast and accurate absorbed dose accumulation computations. The algorithm has been historically used in sequential code to reduce numerical rounding errors for single-precision summation but it has been adopted in ARCHER-RT for parallel GPU computations to accumulate 64-bit absorbed dose values (32-bit absorbed dose increment and 32-bit numerical error) while using the same techniques as in WAG to eliminate intra-warp thread contention.[30,34]

2.A.3. ARCHER-RT workflow

ARCHER-RT is designed using the general workflow depicted in Fig. 1 using the unified modeling language activity diagram for radiation therapy applications. The white blocks indicate code run on the host system, and the shaded blocks on the GPUs (with CPU fallbacks).

The host system initializes several key modules of ARCHER-RT, including Linac source modeling, phase space file I/O, DICOM, and radiation transport in patient. The Linac source module transports particles in X/Y jaws and multileaf collimator (MLCs). The phase space file I/O module reads phase-space-1 files from the disk and allocates/deallocates the memory for particle data storage. The DICOM module parses CT images, RT plan, RT dose, and RT structure files, sets up radiotherapy simulation parameters, and generates patient-specific phantoms. The transport in patient module simulates photon-electron transport inside the patient body. For each beam in an IMRT plan, ARCHER-RT passes one or more (customizable) batches of phase-space-1 files to the Linac source module that tracks particles through the rotated collimator comprised of the X/Y jaws and MLCs until they reach the phase-space-2 plane. ARCHER-RT then applies the geometry transformation to the particles according to the gantry angle and couch angle before saving them to phase-space-2 particle container. For a VMAT plan, the DICOM module implicitly converts the control point sequence into beam sequence, so that ARCHER-RT is capable of simulating VMAT plans in the same way as IMRT. Once phase-space-2 particles from all beams are obtained, ARCHER-RT proceeds to photon-electron coupled transport inside the patient, using the batch simulation scheme. Because this process is usually more computationally intensive, multi-GPU support is provided based on dynamic scheduling, where each GPU, if idle, retrieves one batch of particles to simulate. After all batches are finished, the voxel absorbed dose and the relative standard deviation arrays are stored.

2.A.4. Source modeling

Source modeling, in the context of Monte Carlo absorbed dose engine, refers to a method that calculates information about particles passing through linear accelerator beamline components. In the method that uses phase space information collected from treatment head simulations,[35] source particles originated from the target are transported through the detailed geometric model of the treatment head, in which the energy,

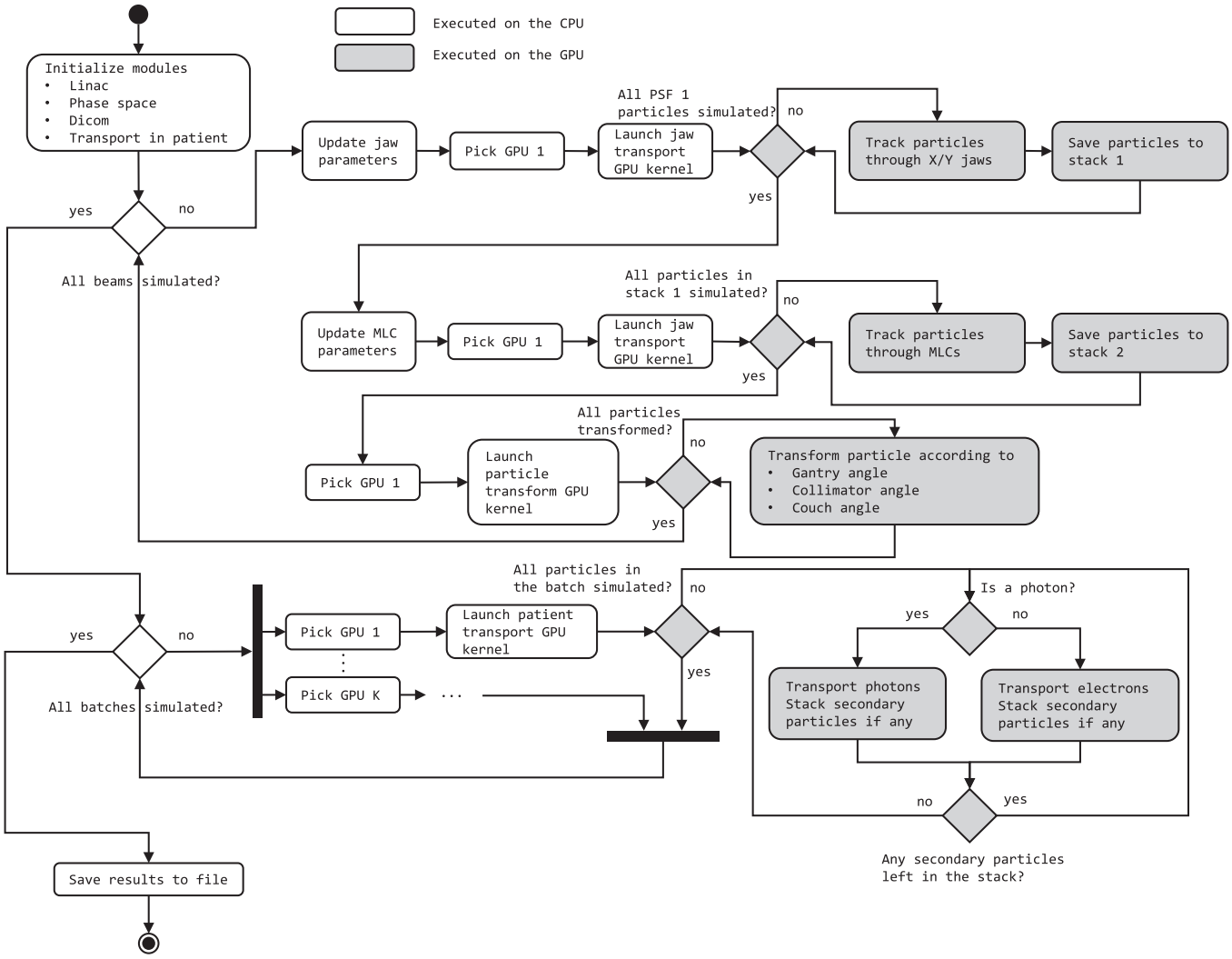

FIG. 1. Unified modeling language activity diagram of ARCHER-RT for radiation therapy applications. The white blocks illustrate the tasks of the sequential code executed on the CPU and the shaded blocks illustrate the tasks of the parallel code on the GPU.

position, direction and statistical weight information of a particle are recorded. With the correct setup of geometry parameters and energy spectrum of the linear accelerator, this method provides the most accurate source model.[35]

The source modeling is initiated at the patient-independent phase space directly below the primary collimators but just above the secondary X/Y collimator of the Varian TrueBeam. These patient-independent phase spaces are generated by GEANT4 simulations performed and validated by Constantin et al.[36] As such, the task of source modeling for ARCHER-RT is narrowed down to modeling of secondary X/Y jaws and MLC, and efficient sampling of particles through these components.

To balance the accuracy of source term representation and sampling efficiency, a First-Compton-based approximate transport method[37,38] was used for the particle transport through the secondary X/Y jaws and MLC in ARCHER-RT. ARCHER-RT only transports photons in the beam collimation routines. Only the Compton scattering effect is considered in the source modeling method because any electron generated is assumed to be absorbed locally for both photoelectric interactions and Compton scattering interactions. Pair production is ignored because of the low interaction probability for any photon less than 5 MeV. As photons traverse through the secondary jaws or the MLC, the interaction site is sampled over the slab thickness assuming an exponential attenuation of incident photons. The probability of Compton scattering is evaluated by the ratio of the Compton and total attenuation coefficients and the energy and angle of Compton scattered photons are determined according to the Klein-Nishina formula.[39] In accordance with the methods outlined by Keall et al.,[37] an interacting particle's weight is modified based on scattered photons' energy and direction, while the emerged Compton electron histories are terminated.[38] The remaining thickness after the interaction and the corresponding attenuation coefficient of the Compton scattered photons are then used to attenuate the photon as they exit the secondary jaws/MLC leaves.

In this work, we modeled the jaws and specifically the Varian HDMLC using the Siebers-Keall method.[37,38] The Varian HDMLC is a multileaf collimator consisting of two banks of 60 tungsten-alloy leaves with 32 $cm^2 \times 0.25$ cm^2 wide (projected at isocenter) leaves in the central 8 cm of the field and 28 $cm^2 \times 0.50$ cm^2 leaves on the outer 14 cm of the field. MLC-specific parameters that were modeled include intraleaf thickness, interleaf leakage, leaf tip radius and thickness, tongue-and-groove effects, and leaf-edge effects. Physical dimensions and material composition of both the jaws and HDMLC leaves were obtained from the proprietary Varian Monte Carlo data package. The MLC is modeled in a method similar to that described by Bergman et al.[40] The leaf geometry for ARCHER-RT is specified in two input files that account for the distance from the upper surface of an MLC region from the source, the leaf number, and the leaf thickness. In a separate file, the physical density of the tungsten alloy, the leaf tip radius of curvature, the leaf tip 'tip angle,' the maximum thickness of the leaf tip, and the physical leaf offset between closed leaf pairs are specified.[40]

2.A.5. Coupled electron-photon transport

The coupled electron-photon transport kernel in ARCHER-RT is based on the DPM open-source code.[12] Photon transport is explicitly modeled, that is, all particle interactions including those of secondary particles generated along the particle tracks are explicit and independently simulated until they reach the cutoff energy or leave the geometry. In the photon transport module, ARCHER-RT takes photoelectric effect, Compton scattering, and pair production into account. Rayleigh scattering is safely neglected since it has very little impacts on absorbed dose distributions considering the energy range used for radiation therapy (keV to 20 MeV).[12]

A Class II condensed history method is employed in ARCHER-RT for electron simulations. Class II condensed history method basically divides electron simulations into two categories: (a) Hard collisions which are simulated explicitly since they can lead to significant changes in the direction or kinetic energy of the electron, and (b) Soft collisions which are frequent interactions resulting in an energy loss below a predefined threshold and are modeled using the Continuously Slowing Down Approximation. Energy loss for soft collisions is calculated using restricted stopping powers and the direction change is calculated using the Multiple Scattering methods.[39] In this study, the energy threshold is set to 200 keV (the default value in DPM) since the range of MeV electrons being transported in soft tissues is about 1.0 mm — a typical voxel size of a patient phantom.

2.A.6. Patient modeling

ARCHER-RT consists of a DICOM processing module to parse CT images, RT plan, RT dose, and RT structure files, built on top of the DCMTK API.[41] Conversion of CT images into a patient-specific phantom is implemented according to a simplistic, four-material scheme originating from EGSnrc's "ctcreate" program which maps the Hounsfield Units (HU) of each image pixel to a density value and a material type (dry air, lung, soft tissue, or compact bone). Absolute absorbed dose calibration was performed for ARCHER-RT according to a simplified dose conversion factor expression of Popescu which ignores chamber backscatter.[40,42]

2.B. EGSnrc description

The dosimetric accuracy of ARCHER-RT was evaluated using a validated model of a flattened 6X Varian TrueBeam implemented using a coupled EGSnrc simulation using BEAMnrc "Source 21" for source modeling and DOSXYZnrc "Source 20" for in-phantom particle transport. The source models implemented in both BEAMnrc/DOSXYZnrc allow for time-dependent beam configurations. Similar to

ARCHER-RT, a patient-independent phase space generated by Constantin et al. is used as input for the BEAMnrc source model.[36] The component modules "SYNCJAWS" and "SYNCHDMLC" were used to model the jaws and HDMLC respectively. MLC-specific parameters are input into the "SYNCHDMLC" component module using data obtained from the proprietary Varian Monte Carlo data package. A workflow for extracting patient specific beam parameters was developed using the "pycom" utilities developed by Lixin Zhan.[43] This suite of utilities is specifically designed to automatically populate the BEAMnrc/DOSXYZnrc input files. EGSPHANT phantoms used in DOSXYZnrc simulations were generated using modified Computational Environment for Radiotherapy Research scripts.[44] The CT calibration curves used in the phantom generation scripts matched those input for ARCHER-RT. Absolute absorbed dose calibration was separately performed for the EGSnrc beam model according to a simplified dose conversion factor expression of Popescu which ignores chamber backscatter.[40,42]

2.C. Test cases

2.C.1. Single vs double-precision test case

Supplementing the work of Liu et al.,[30] a test case was conducted in line with Magnoux et al.[21] to compare the impact of native single-precision vs double-precision on the computational accuracy and the effectiveness of the Kahan-summation method as implemented in ARCHER-RT GPU code. A 1-MeV monoenergetic electron volumetric source (a cube of 2mm sides) was placed directly above a 10 cm^3 × 10 cm^3 × 10 cm^3 with a voxel size of 1 mm^3 × 1 mm^3 × 1 mm^3. Electrons were chosen as the simulated particle in order to isolate the electron transport in the photon–electron transport. The number of histories that was run included 7e8 particles. Absorbed dose was tallied to the phantom using three different algorithms, (a) native single-precision, (b) double-precision, and (c) single-precision utilizing the KAS algorithm. Absorbed dose differences were computed to assess the agreement between the three algorithms and timing studies were conducted to assess the performance of each algorithm. double-precision calculations were taken as the gold standard. The calculations were performed on an NVIDIA GTX 1080Ti GPU which is capable of computing all three algorithms at the hardware level; for example, software emulation of the algorithms was not conducted. The NVIDIA GTX 1080Ti is capable of 1134 GFLOPS (giga-floating point operations) in single-precision computations and 354.4 GFLOPS in double-precision computations.

2.C.2. PDD and Axial profiles

Percent depth dose (PDD) curves and lateral absorbed dose profiles in a cubic water phantom were calculated in both ARCHER-RT and EGSnrc codes. The phantom was a 40 cm^3 × 40 cm^3 × 40 cm^3 water phantom with a voxel size of 0.1 cm^3 × 0.1 cm^3 × 0.1 cm^3. The source-to-surface distance was set to 100.0 cm. For the PDD verification, open field sizes of 1.5 cm^2 x 1.5 cm^2, 3 cm^2 × 3 cm^2, 6 cm^2 × 6 cm^2, 10 cm^2 × 10 cm^2, 20 cm^2 × 20 cm^2, and 22 cm^2 x 40 cm^2 were simulated, and the absorbed dose distributions were scored using a voxel size of 0.2 cm^3 × 0.2 cm^3 × 0.2 cm^3. The lateral absorbed dose verification was performed with the open field size of 10 cm^2 × 10 cm^2 and the absorbed dose distributions are scored at different depths including 1.5, 5.0, 10.0, and 20.0 cm. The absorbed dose distributions are scored using a voxel size of 0.2 cm^3 × 0.2 cm^3 × 0.2 cm^3.

2.C.3. Picket fence test

To verify the accuracy of the HD120 MLC model, a picket fence MLC pattern was simulated in ARCHER-RT and compared against absorbed dose distributions calculated by EGSnrc. The phantom was 10 cm^3 × 10 cm^3 × 10cm^3 with a voxel size of 0.2 cm^3 × 0.2 cm^3 × 0.2 cm^3 with an SSD of 90 cm. Leaves were moved in the cross-plane direction from -5 to 5 cm in 1 cm intervals. Results were normalized to the same voxel index for both the EGS and ARCHER-RT dose grid in a region near d_{max} (e.g., voxel exposed to fluence).

2.C.4. MPPG5a test case

As an additional check to verify the accuracy of the MLC model, a static MLC pattern recommended by MPPG5a[45] for treatment planning system commissioning was simulated in ARCHER-RT and compared against absorbed dose distributions calculated by EGSnrc. The phantom was 10 cm^3 × 10 cm^3 × 10 cm^3 with a voxel size of 0.2 cm^3 × 0.2 cm^3 × 0.2 cm^3. To quantify the differences, an absolute absorbed dose difference and 3D gamma test were calculated.

2.C.5. VMAT

A clinical prostate VMAT treatment plan was calculated in ARCHER-RT, and EGSnrc to evaluate ARCHER-RT capability of clinical treatment plan absorbed dose calculations. The voxel size used in the simulation was 0.25 cm^3 × 0.25 cm^3 × 0.25cm^3. Sufficient histories were simulated to ensure the relative standard deviation of critical regions, that is, the uncertainty for voxels with absorbed dose greater than 20% was under 1.7% in ARCHER-RT. Absorbed dose difference and 3D gamma tests were performed to evaluate the dosimetric accuracy of ARCHER-RT.

2.D. VMAT efficiency studies

Timing studies were conducted to evaluate the relative speed of ARCHER-RT being executed under three different

hardware architectures: (a) an Intel i7-8700K CPU with six cores (12 hardware threads), (b) an NVIDIA 1080Ti GPU with 28 streaming multiprocessors and 11GB GDDR5X memory, and (c) an AMD Vega 56 GPU with 56 compute units and 8GB HBM2 memory. The host system has 16GB DDR4 memory and a solid-state drive (SSD). Timing of different modules in ARCHER-RT (i.e., time to read the phase space file, time for source modeling execution, and time for particle transport in the patient) was also investigated. Timing studies were also conducted on the VMAT plan to compare the relative speed of the computation conducted in native single-precision, KAS single-precision, and double-precision.

3. RESULTS

3.A. Single vs double-precision test case

Figure 2(a) depicts the absorbed dose distribution of the test case executed in double-precision, Fig. 2(b) depicts the difference between the test case executed in native single-precision vs double-precision, and Fig. 2(c) depicts the difference between the test case executed utilizing the KAS single-precision algorithm vs double-precision. The results of the test case in Fig. 2 are presented for a 20 mm^2 × 20 mm^2 subsection of the phantom for a slice 2mm from the source. The results presented are for voxels that are close to the source and as expected, these voxels will be accessed and written to most, thus creating a scenario to highlight numerical truncation errors. A maximum difference of 71.73% between the native single and double-precision case was calculated. In the second case where KAS was implemented, the maximum difference between KAS and double-precision was less than 0.01%. The native single-precision, KAS single-precision, and double-precision took 39, 55, and 70 s, respectively to calculate. The results both confirm the findings of Liu[30] and Magnoux[21] and proffer a solution to the inaccuracies of the native single-precision atomic-add method through the implementation of the KAS atomic-add algorithm.

3.B. PDD and Axial profiles

Figure 3 compares relative depth dose and lateral dose profiles for ARCHER-RT and EGSnrc. Absorbed doses are normalized to the maximum absorbed dose of the 10 cm^2 × 10 cm^2 PDD field on the central axis. Outside the buildup region for the PDDs, there was less than 3% difference between the two codes for all points. Similarly, outside the penumbra region there was less than 2% difference for all points in the axial profiles. The large difference between the two codes in the buildup region of the PDDs could be that ARCHER-RT transports only photons in the beam collimation routines (Note: ARCHERRT transports both photons and electrons inside the patient), whereas EGSnrc transports both photons and electrons in beam collimation routines. It is also possible that the buildup region of the PDD, a region with a large gradient, contain large local absolute differences. Either of these are plausible explanations for the underestimation of the absorbed dose in the buildup region for static beam configurations as demonstrated in the PDDs. The latter reason is certainly applicable for the differences noted in the axial profiles; large differences are only seen in the penumbra region.

3.C. Picket fence test

Figure 4 depicts the calculated cross-plane dose profile on the central axis of the in-plane direction and at a depth of 5 cm for the picket fence test demonstrating good agreement in the MLC model between ARCHER-RT and EGSnrc. Slight differences in the in-field scatter are likely attributable to the difference fluence models each code uses; EGSnrc models the MLC geometry explicitly while ARCHER-RT uses an approximate form of the MLC. Slight differences in the out-of-field scatter are likely attributable to ARCHER-RT's use of the first Compton scatter approximation for MLC photon transport.

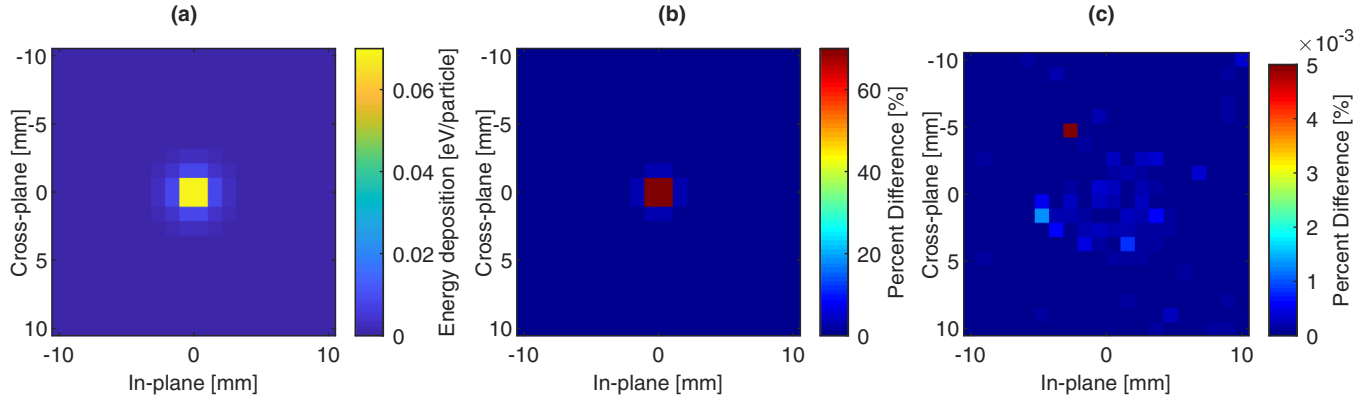

FIG. 2. Results from the test case (a) absorbed dose distribution of the case run in double-precision, (b) relative difference (data shown in percent difference) between native single-precision and double-precision and (c) the relative difference between KAS single-precision and double-precision. The figure is displaying data in a slice 2mm from the surface of the phantom. Note the scales for the relative difference between (b) and (c) are dramatically different to highlight differences.

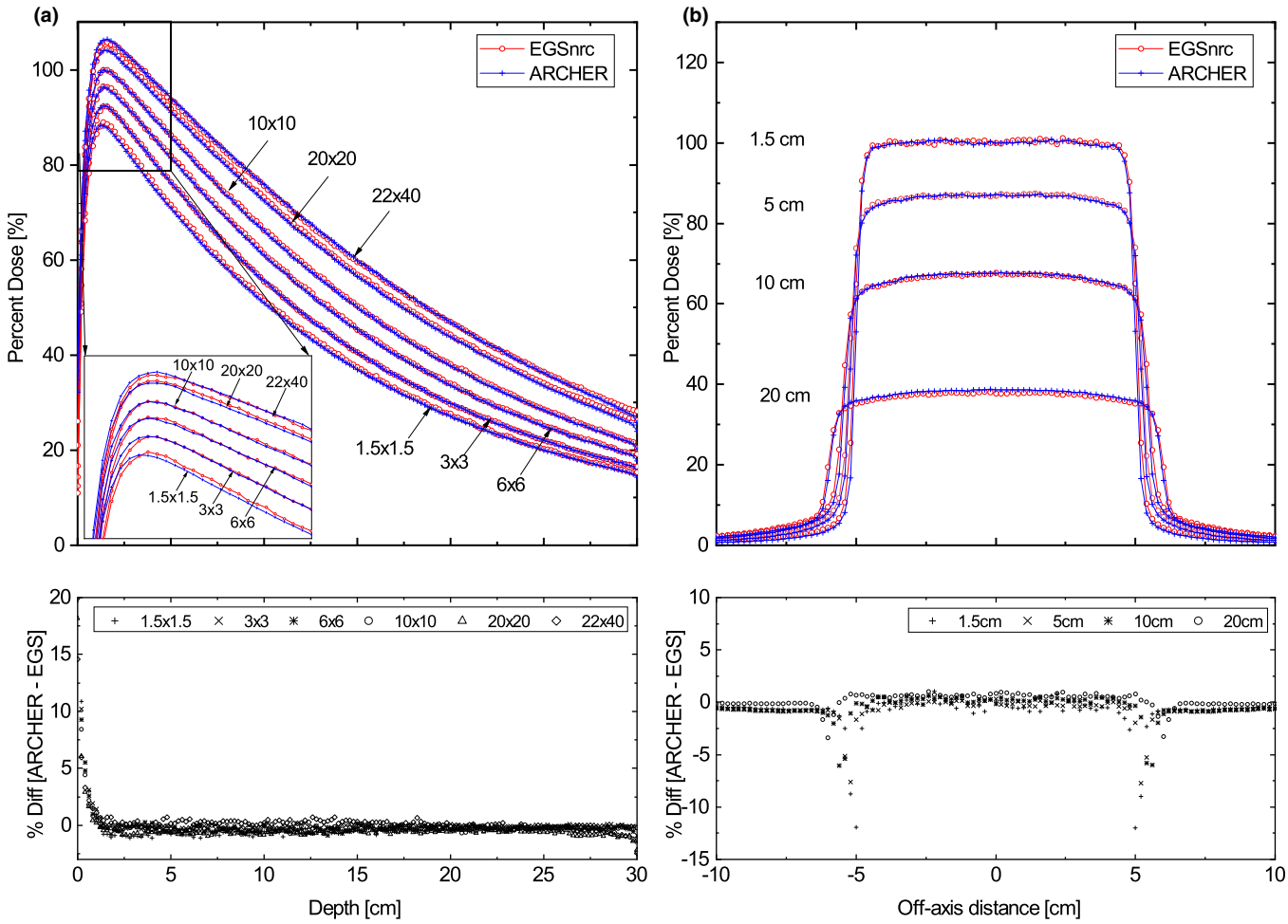

FIG. 3. Comparison of percent depth dose and lateral dose profiles for ARCHER-RT and EGSnrc. (a) percent depth dose and calculated differences and (b) axial profiles and calculated differences for a 10 cm^2 × 10 cm^2 field at depths of 1.5, 5, 10, and 20 cm.

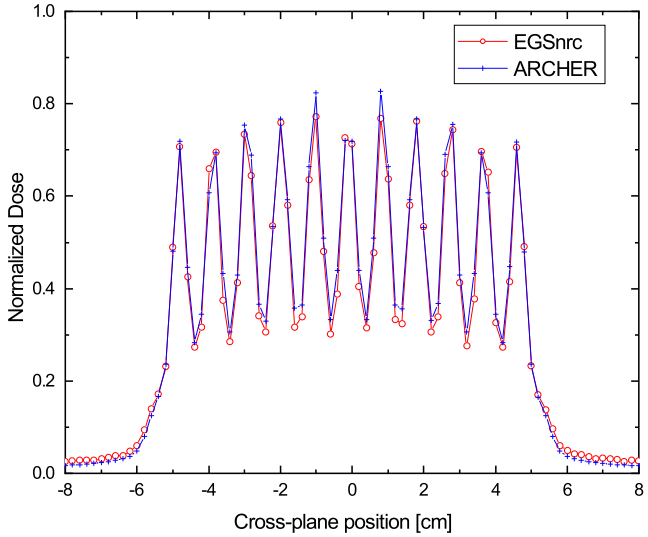

FIG. 4. Comparison of the picket-fence test for ARCHER-RT and EGSnrc.

3.D. MPPG5a test case

Figure 5 depicts the calculated absorbed dose for the MPPG5a clinical test plan for (a)–(c) ARCHER-RT, and (d)–(f) the absolute difference between the two (EGSnrc-ARCHER-RT), demonstrating good qualitative agreement between ARCHER-RT and EGSnrc. A gamma index test is performed for voxels equal to or greater than 20% of the maximum absorbed dose to quantitatively evaluate the agreement between the two codes. The passing rate is found to be 94.2% for 3%/3 mm criterion and 86.4% for 2%/2 mm, thus further confirming the agreement of these two codes and the accuracy of the HD120 MLC model. Similar to that for the PDDs, the difference between the two codes near the surface of the phantom in the electron contamination region could be that ARCHER-RT transports only photons in the beam collimation routines, whereas EGSnrc transports both photons and electrons in all media and the buildup region is a region with a large gradient. Either of these are plausible explanations for the underestimation of the absorbed dose in the surface contamination region for static beam configurations by ARCHER-RT as shown in Figs. 5(e) and 5(f).

3.E. VMAT

Figure 6 depicts the calculated absorbed dose for the clinical VMAT plan for (a)–(c) ARCHER-RT, and (d)–(f) the absolute difference between the two (EGSnrc- ARCHER-RT). From these visual inspections, it is clear that ARCHER-RT and EGSnrc agree well. A gamma index test was performed for voxels equal to or greater than 20% of the maximum absorbed dose to quantitatively evaluate the agreement between the two codes. The passing rate was 98.35% for 3%/3mm criterion, suggesting that the accuracy of ARCHER-RT is satisfactory. Furthermore, Fig. 7 depicts the dose volume histogram (DVH) for the VMAT prostate case including the PTV, and organs at risk including the bladder, rectum, femoral heads, and penile bulb. It can be seen that these two codes agree with each other very well. The penile bulb showed the most notable difference between the two codes, likely because the volume of the ROI is quite small and thus will manifest in exaggerated differences on a DVH plot.

3.F. VMAT efficiency studies

A comparison of the total wall time and execution time of different modules in ARCHER-RT for the Intel CPU, NVIDIA GPU, and AMD GPU conducted on the clinical VMAT prostate plan are presented in Table I. ARCHER-RT was compiled with the fast-math flag and executed using single-precision atomic-add methods for the timing studies presented. While these two options may theoretically limit the accuracy of the absorbed dose accumulation, the agreement to EGSnrc indicated the computational approximations were valid. Statistical uncertainties are kept under 1.7% for critical absorbed dose regions, that is, for voxels with absorbed dose greater than or equal to 20% of D_{max}. The NVIDIA 1080Ti card executed the fastest and completed the absorbed dose calculation in 50.3 s. This time was 4.3x faster than the i7-8700K CPU and 3.7x faster than the AMD Vega 56 GPU. In theory, the AMD GPU has competitive computing power with NVIDIA GPU, but it seriously underperformed in our analysis primarily due to identified deficiencies within the compiler. Specifically, the AMD hcc compiler generates a hanging kernel code for particle transport in patient. A workaround was utilized by combining several C++ classes in the kernel into one large class, but the GPU register spill resulted in a side effect causing remarkable performance degradation. Overall, the most time-consuming part was the Monte Carlo particle transport in the patient, which is expected considering the heterogeneities and dimension of the patient phantom.

There were no voxel level dosimetric differences in excess of 0.01% between native single-precision, KAS single-precision, and double-precision; however, there were differences in the timing studies. The major difference in timing between the three cases was the patient transport time; the other routines took approximately the same amount of time. The patient transport time for the native single-precision case took 25.6 s and was 3.04x faster than that for double-precision, which took 77.9 s, whereas the KAS single-precision case took

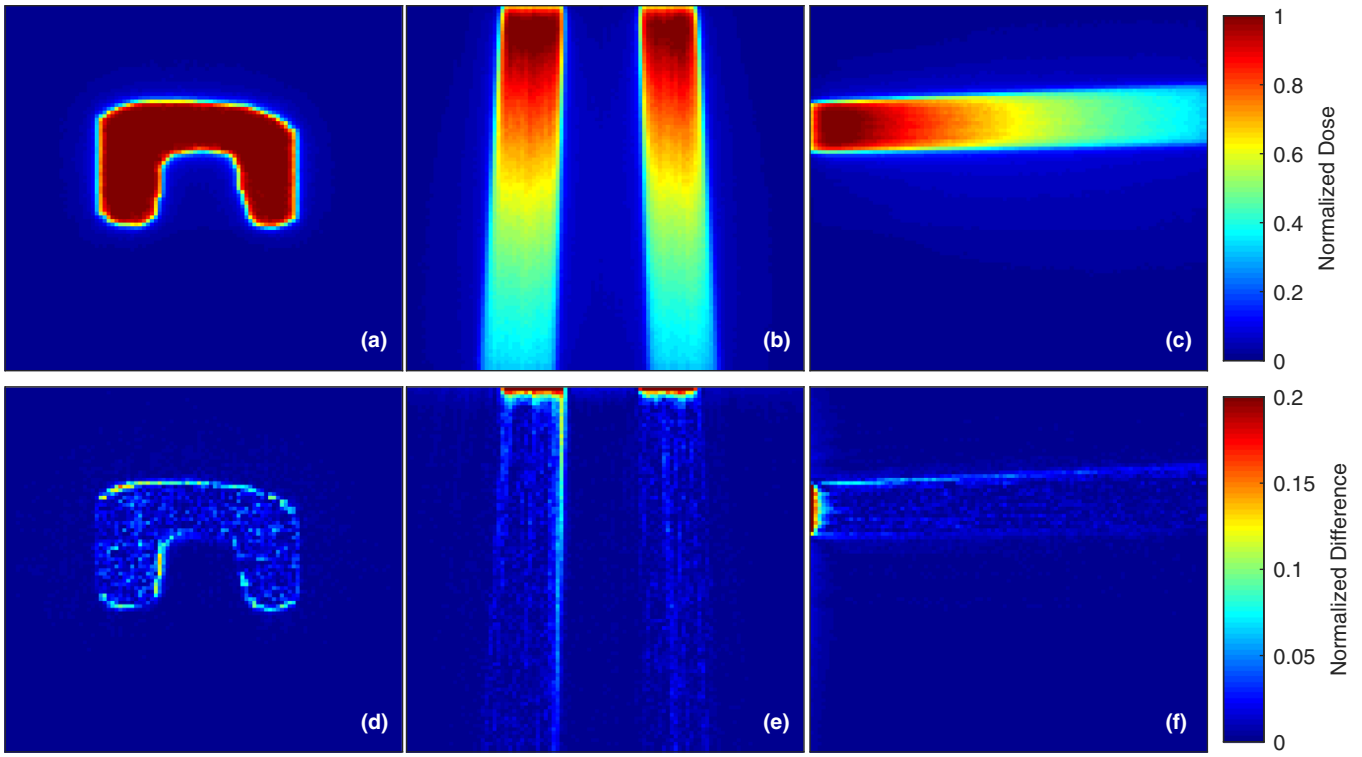

FIG. 5. Visual inspection of absorbed dose distributions of two codes for the MPPG5a test plan showing excellent agreement. (a)–(c) ARCHER-RT, and (d)–(f) absolute difference (EGSnrc-ARCHER-RT).

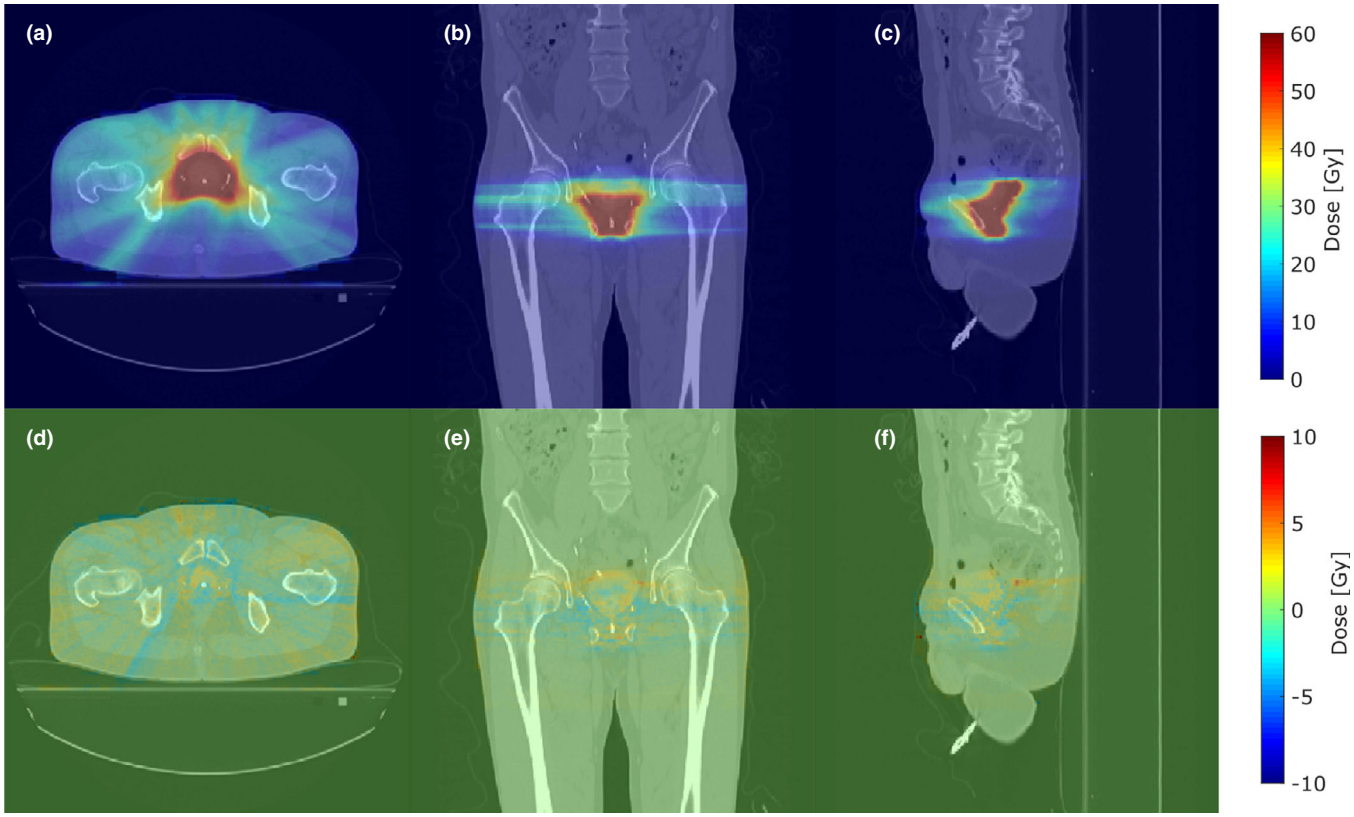

FIG. 6. Clinical VMAT prostate calculated absorbed dose distributions showing excellent agreement of two codes. (a)–(c) ARCHER-RT, and (d)–(f) absolute difference (EGSnrc-ARCHER-RT).

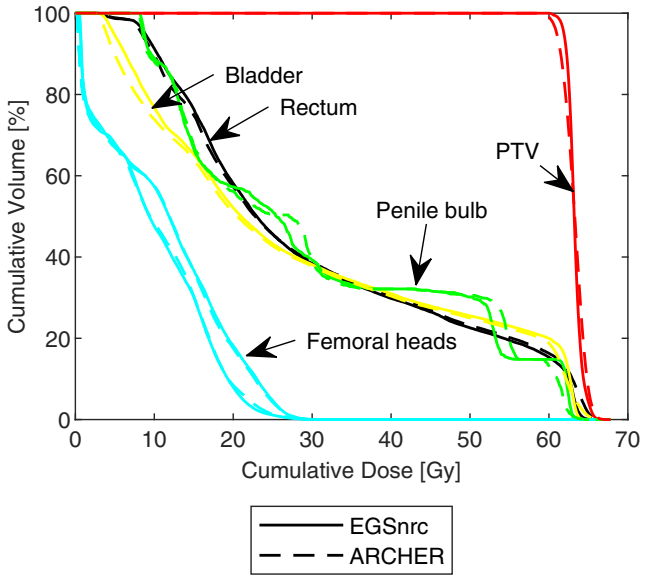

FIG. 7. Dose volume histogram comparison between ARCHER-RT and EGSnrc for the VMAT prostate case showing excellent agreement between the two MC codes.

30.1 s and thus was 2.59x faster than the double-precision case. The results indicate there is a small performance penalty in the use of the KAS single-precision algorithm but in accordance with the results of the test case, ensures there is computational accuracy.

TABLE I. Architecture timing results for the ARCHER-RT simulation of the VMAT prostate plan.

Architecture	Total wall time (s)	PSF reading time (s)	Linac transport time (s)	Patient transport time (s)
Intel i7-8700K CPU	216.8	3.1	100.8	108.2
NVIDIA 1080Ti GPU	50.3	3.4	16.8	25.6
AMD Vega 56 GPU	187.9	3.3	28.0	151.6

4. DISCUSSION

This work demonstrates that ARCHER-RT is a versatile, cross-platform Monte Carlo absorbed dose calculation engine and is compatible with multiple hardware architectures in the clinical setting. We have benchmarked ARCHER-RT by comparing calculated absorbed dose distributions to results from EGSnrc. With an NVIDIA GPU, we demonstrated that a clinical VMAT prostate case can be executed in less than 51 s and dosimetrically verified its results against well-benchmarked codes. Source modeling has been implemented in ARCHER-RT in which patient-independent phase spaces are used as input for transporting particles through secondary X/Y jaws and MLCs. The Siebers-Keall first Compton-based approximate transport method is used to balance the accuracy of source term representation and sampling efficiency. The

results reported in this work indicate that the source modeling implementation is accurate and reproducible.

Cross platform compatibility is an important feature for clinical deployment onto different computing architectures. Previous groups have implemented OpenCL to employ cross-platform compatibility; however, OpenCL is currently deprecated on MacOS. We implemented HIP for cross platform compatibility so that that ARCHER-RT can run on CPUs, and both AMD and NVIDIA GPUs. While HIP allows for cross-GPU compatibility, the AMD GPU implementation seriously underperformed because of two unresolved inadequacies currently residing in the HIP compiler. Specifically, the compiler cannot correctly handle a C++ "aggregation" class where, for instance, class A references external objects B, C, D by pointers. The compiled binary crashes upon execution. A workaround is to switch the design to a C++ "composition" class where objects B, C, D are instantiated inside of A as a data member. The downside of this workaround is that the size of class A is inflated by a large degree, causing register spills, a common culprit for GPU performance degradation. The second shortcoming of the compiler we have found is that the HIP compiler cannot correctly handle in branch code the warp vote functions, which constitute the centerpiece of our WAG and KAS algorithms for fast atomic-add tallies.[30] The only viable workaround is to switch back to the slow, default CAS algorithm for absorbed dose accumulation. These slower execution times on AMD GPUs are similar to other studies in which HIP was utilized.[46] While we have successfully implemented a cross-platform code, further performance increases will come from architecture-specific algorithmic development and HIP API maturation.

The motivation behind the test case comparing the accuracy and performance of native single-precision, double-precision, and KAS single-precision was to proffer a better solution to that offered by Magnoux than the software emulation of double-precision computational methods. The test case is indeed an idealized case enacted to demonstrate a scenario in which native single-precision is inadequate and large dosimetric differences are present. The reason there are greater voxel-level discrepancies between atomic-add methods in the test electron case is simply because there is a greater frequency of voxel-specific interactions (voxels near the source) thus leading to more opportunity for voxel specific truncation errors. This is in contrast to the VMAT case in which the spatial distribution of interacting photons is much more diffuse. While there was little dosimetric difference between the different algorithms in the VMAT case, the identification and implementation of algorithms to ensure computational accuracy are an important consideration in the software development of a Monte Carlo absorbed dose engine such as ARCHER-RT and there could be clinical scenarios in which this may be important. Additionally, the implementation of KAS is important to preserve the accuracy on all NVIDIA and AMD devices, some of which do not support double-precision calculations at the hardware level. The timing results of each algorithm showed there is some penalty in performance utilizing KAS over native single-precision, but it is a small price compared to the software emulation of double-precision as described by Magnoux. There was a smaller performance penalty in utilizing KAS over single-precision for the VMAT case in comparison to the test electron case. Theoretically the penalty should be the same for every dose accumulation event, but this discrepancy between the two cases is simply because the electron case only transported electrons, whereas the VMAT case included coupled photon–electron transport and thus the dose deposition (electron transport) was only a portion of the particle transport in the VMAT case.

ARCHER-RT was validated by comparing against a validated EGSnrc TrueBeam model. Benchmarking tests included PDD and axial profiles, an MPPG5a MLC test shape, and a clinical prostate VMAT plan. The agreement was quantitatively evaluated using absorbed dose differences and gamma tests. PDD's for both ARCHER-RT and EGSnrc was found to match within 3%. Slight differences between ARCHER-RT and EGSnrc in the MPPG5a static beam case are attributable to how the MLC geometry is specified in each code package and because ARCHER-RT does not include the source modeling of electrons. EGSnrc explicitly models the HDMLC in entirety including small features like tongue and groove, whereas ARCHER-RT's MLC model is based upon that described by Siebers and Keall and represents the complexity of the MLC's geometry by breaking down the MLC into simple geometrical regions.[37] The rationale for using the MLC representation described by Siebers and Keall was to simplify the radiation transport calculation for complex IMRT beam delivery; small differences in individual beamlets effectively "wash-out" in evaluating full IMRT deliveries because the average interaction probability is determined by evaluating the probability of an incident particle in the MLC multiple times. This lends itself to systemic MLC collimator edge differences between ARCHER-RT and EGSnrc as depicted in Figs. 5(d)–5(f). Considering these slight deficiencies, they were shown to be acceptable approximations. As demonstrated in the clinical VMAT plan results, individual beamlet differences do effectively cancel each other out and patient surface absorbed dose differences are not appreciable.

There are known limitations of utilizing the gamma test to compare Monte Carlo dose distributions. These limitations are generally influenced by the presence of statistical noise, especially in low dose gradient regions.[47,48] We have run enough particles for each scenario in which the gamma test is used such that we can be confident the gamma test is an accurate representation of the agreement between the two codes.[49]

5. CONCLUSIONS

ARCHER-RT's capabilities have been dramatically extended from the previous publication to include newer modalities, and, with these improvements, the accuracy, speed, and computational precision have been demonstrated in this work through the modeling and benchmarking of a flattened photon 6 MV Varian TrueBeam. ARCHER-RT

fulfills the clinical requirements of fast yet accurate radiation dose calculation that are essential for absorbed dose engines to be introduced into clinical workflows. There are examples today for how a Monte Carlo absorbed dose engine like ARCHER-RT can be adapted into the clinical workflow as part of the Monte Carlo-based treatment planning system. Under this auspice, this work demonstrates the significant addition of functionality to ARCHER-RT framework which has marked utility for both research and clinical applications.

ACKNOWLEDGMENTS

This project was supported by NIH/NIBIB (R42EB019265-01A1) and NIH SPORE CA196513-01. We would like to thank the anonymous reviewers and associate editor for carrying out an extremely thorough review that has helped improve the quality and impact of this manuscript.

CONFLICT OF INTEREST

X. George Xu is a cofounder of Virtual Phantoms, Inc (Albany, New York) that commercializes software technologies—VirtualDose™ for medical CT dose reporting and ARCHER™ for real-time Monte Carlo dose computing.

[a)]Author to whom correspondence should be addressed. Electronic mail: xug2@rpi.edu

REFERENCES

1. Delaney G, Jacob S, Featherstone C, Barton M. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer*. 2005;104:1129–1137.
2. Rogers DWO. Fifty years of Monte Carlo simulations for medical physics. *Phys Med Biol*. 2006;51:R287–R301.
3. Kawrakow I.The EGSnrc Code System, Monte Carlo simulation of electron and photon transport. NRCC Rep. PIRS-701, 2001.
4. Kawrakow I, Walters BRB. Efficient photon beam dose calculations using DOSXYZnrc with BEAMnrc. *Med Phys*. 2006;33:3046–3056.
5. Forster RA, et al. MCNP™ Version 5. *Nucl Instrum Methods Phys Res Sect B Beam Interact Mater Atoms*. 2004;213:82–86.
6. Lewis RD, Ryde SJS, Hancock DA, Evans CJ. An MCNP-based model of a linear accelerator x-ray beam. *Phys Med Biol*. 1999;44:1219–1230.
7. Mesbahi A, Reilly AJ, Thwaites DI. Development and commissioning of a Monte Carlo photon beam model for Varian Clinac 2100EX linear accelerator. *Appl Radiat Isot*. 2006;64:656–662.
8. Baró J, Sempau J, Fernández-Varea JM, Salvat F. PENELOPE: An algorithm for Monte Carlo simulation of the penetration and energy loss of electrons and positrons in matter. *Nucl Instrum Methods Phys Res Sect B Beam Interact Mater Atoms*. 1995;100:31–46.
9. Sempau J, Badal A, Brualla L. A PENELOPE -based system for the automated Monte Carlo simulation of clinacs and voxelized geometries-application to far-from-axis fields. *Med Phys*. 2011;38:5887–5895.
10. Agostinelli S, et al. Geant4—a simulation toolkit. *Nucl Instrum Methods Phys Res Sect A Accel Spectrom Detect Assoc Equip*. 2003;506:250–303.
11. Grevillot L, Frisson T, Maneval D, Zahra N, Badel J-N, Sarrut D. Simulation of a 6 MV Elekta Precise Linac photon beam using GATE/GEANT4. *Phys Med Biol*. 2011;56:903–918.
12. Sempau J, Wilderman SJ, Bielajew AF. DPM, a fast, accurate Monte Carlo code optimized for photon and electron radiotherapy treatment planning dose calculations. *Phys Med Biol*. 2000;45:2263–2291.
13. Kawrakow I. VMC++, Electron and Photon Monte Carlo Calculations Optimized for Radiation Treatment Planning, In *Advanced Monte Carlo for Radiation Physics, Particle Transport Simulation and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg. 2001;pp. 229–236.
14. Ma C-M, et al. A Monte Carlo dose calculation tool for radiotherapy treatment planning. *Phys Med Biol*. 2002;47:305.
15. Jia X, Gu X, Graves YJ, Folkerts M, Jiang SB. GPU-based fast Monte Carlo simulation for radiotherapy dose calculation. *Phys Med Biol*. 2011;56:7017–7031.
16. Jia X, Ziegenhein P, Jiang SB. GPU-based high-performance computing for radiation therapy. *Phys Med Biol*. 2014;59:R151–R182.
17. Jahnke L, Fleckenstein J, Wenz F, Hesser J. GMC: a GPU implementation of a Monte Carlo dose calculation based on Geant4. *Phys Med Biol*. 2012;57:1217–1229.
18. Hissoiny S, Ozell B, Bouchard H, Després P. GPUMCD: a new GPU-oriented Monte Carlo dose calculation platform. *Med Phys*. 2011;38:754–764.
19. Townson RW, Jia X, Tian Z, Graves YJ, Zavgorodni S, Jiang SB. GPU-based Monte Carlo radiotherapy dose calculation using phase-space sources. *Phys Med Biol*. 2013;58:4341–4356.
20. Tian Z, Graves YJ, Jia X, Jiang SB. Automatic commissioning of a GPU-based Monte Carlo radiation dose calculation code for photon radiotherapy. *Phys Med Biol*. 2014;59:6467–6486.
21. Magnoux V, Ozell B, Bonenfant É, Després P. A study of potential numerical pitfalls in GPU-based Monte Carlo dose calculation. *Phys Med Biol*. 2015;60:5007–5018.
22. Liu T, Xu XG, Carothers CD. Comparison of two accelerators for Monte Carlo radiation transport calculations, Nvidia Tesla M2090 GPU and Intel Xeon Phi 5110p coprocessor: a case study for X-ray CT imaging dose calculation. *Ann Nucl Energy*. 2015;82:230–239.
23. Tian Z, Shi F, Folkerts M, Qin N, Jiang SB, Jia X. A GPU OpenCL based cross-platform Monte Carlo dose calculation engine (goMC). *Phys Med Biol*. 2015;60:7419–7435.
24. Tian Z, Li Y, Hassan-Rezaeian N, Jiang SB, Jia X. Moving GPU-OpenCL-based Monte Carlo dose calculation toward clinical use: Automatic beam commissioning and source sampling for treatment plan dose calculation. *J Appl Clin Med Phys*. 2017;18:69–84.
25. Su L, Yang Y, Bednarz B, et al. ARCHERRT- A GPU-based and photon-electron coupled Monte Carlo dose computing engine for radiation therapy: software development and application to helical tomotherapy. *Med Phys*. 2014;41:1–13.
26. Liu T. Development of ARCHER — a Parallel Monte Carlo Radiation Transport Code — for X-Ray CT Dose Calculations Using GPU and Coprocessor Technologies. Rensselaer Polytechnic Institute. 2014.
27. Su L. Development and application of a GPU-based fast electron-photon coupled monte carlo code for radiation therapy by major subject: nuclear engineering. Rensselaer Polytechnic Institute. 2014.
28. Lin H. GPU-Based Monte Carlo Source Modeling and Simulation for Radiation Therapy Involving Varian TrueBeam LINAC. Rensselaer Polytechnic Institute. 2018.
29. AMD. HIP Programming Guide. 2019. Available: https://rocm-documentation.readthedocs.io/en/latest/Programming_Guides/HIP-GUIDE.html Accessed: 28-Apr-2019.
30. Liu T, Wolfe N, Lin H, Carothers CD, Xu XG. Performance study of atomic tally methods for GPU-accelerated Monte Carlo dose calculation. in 20th Topical Meeting of the Radiation Protection and Shielding Division of the American Nuclear Society. 2018.
31. Adinets A. "CUDA Pro Tip: optimized Filtering with Warp-Aggregated Atomics", 2014. [Online]. Available: https://devblogs.nvidia.com/cuda-pro-tip-optimized-filtering-warp-aggregated-atomics/ [Accessed: 16-Sep-2019].
32. Westphal E. "Voting and Shuffling to Optimize Atomic Operations", 2015. [Online]. Available: https://devblogs.nvidia.com/voting-and-shuffling-optimize-atomic-operations/ [Accessed: 16-Sep-2019].
33. Bossler KL. "Methods for Computing Monte Carlo Tallies on the Gpu", PHYSOR 2018 React. Phys. Paving W. Towar. More Effic. Syst., pp. 166–177, 2018.
34. Kahan W. Pracniques: further remarks on reducing truncation errors. *Commun ACM*. 1965;8:40.
35. Chetty IJ, Curran B, Cygler JE, et al. Report of the AAPM Task Group No. 105: issues associated with clinical implementation of Monte Carlo-

based photon and electron external beam treatment planning. *Med Phys*. 2007;34:4818–4853.
36. Constantin M, Perl J, LoSasso T, et al. Modeling the TrueBeam linac using a CAD to Geant4 geometry implementation: dose and IAEA-compliant phase space calculations. *Med Phys*. 2011;38:4018–4024.
37. Siebers JV, Keall PJ, Kim JO, Mohan R. A method for photon beam Monte Carlo multileaf collimator particle transport. *Phys Med Biol*. 2002;47:3225–3249.
38. Keall PJ, Siebers JV, Arnfield M, Kim JO, Mohan R. Monte Carlo dose calculations for dynamic IMRT treatments. *Phys Med Biol*. 2001;46:929–941.
39. Turner JE. *Atoms, Radiation, and Radiation Protection*. Hoboken, NJ: Wiley; 2007.
40. Bergman AM, Gete E, Duzenli C, Teke T. Monte Carlo modeling of HD120 multileaf collimator on Varian TrueBeam linear accelerator for verification of 6X and 6X FFF VMAT SABR treatment plans. *J Appl Clin Med Phys*. 2014;15:148–163.
41. Eichelberg M, Riesmeier J, Wilkens T, Hewett AJ, Barth A, Jensch P. Ten years of medical imaging standardization and prototypical implementation: the DICOM standard and the OFFIS DICOM toolkit (DCMTK). *Med Imaging*. 2004;5371:57.
42. Popescu IA, Shaw CP, Zavgorodni SF, Beckham WA. Absolute dose calculations for Monte Carlo simulations of radiotherapy beams. *Phys Med Biol*. 2005;50:3375–3392.
43. Zhan L. pycom –- Python DICOM Processing Toolkit. 2013.
44. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. *Med Phys*. 2003;30:979–985.
45. Jacqmin DJ, Bredfeldt JS, Frigo SP, Smilowitz JB. Implementation of the validation testing in MPPG 5.a 'Commissioning and QA of treatment planning dose calculations–megavoltage photon and electron beams'. *J Appl Clin Med Phys*. 2017;18:115–127.
46. Dong T, Haidar A, Tomov S, Dongarra J. Accelerating the SVD bi-diagonalization of a batch of small matrices using GPUs. *J Comput Sci*. 2018;26:237–245.
47. Low DA. Gamma dose distribution evaluation tool. *J Phys Conf. Ser*. 2010;250:349–359.
48. Low DA, Dempsey JF. Evaluation of the gamma dose distribution comparison method. *Med Phys*. 2003;30:2455–2464.
49. Graves YJ, Jia X, Jiang SB. Effect of statistical fluctuation in Monte Carlo based photon beam dose calculation on gamma index evaluation. *Phys Med Biol*. 2013;58:1839–1853.