

Lawrence Berkeley National Laboratory

LBL Publications

Title

Porting WarpX to GPU-accelerated platforms

Permalink

<https://escholarship.org/uc/item/4f28s6b2>

Authors

Myers, A
Almgren, A
Amorim, LD
et al.

Publication Date

2021-12-01

DOI

10.1016/j.parco.2021.102833

Peer reviewed

Porting WarpX to GPU-accelerated platforms

A. Myers^a, A. Almgren^a, L. D. Amorim^a, J. Bell^a, L. Fedeli^e, L. Ge^{b,a}, K. Gott^a, D. P. Grote^c, M. Hogan^b, A. Huebl^a, R. Jambunathan^a, R. Lehe^a, C. Ng^b, M. Rowan^a, O. Shapoval^a, M. Thévenet^d, J.-L. Vay^a, H. Vincenti^e, E. Yang^a, N. Zaïm^e, W. Zhang^a, Y. Zhao^a and E. Zoni^a

^aLawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^bSLAC National Accelerator Laboratory Menlo Park, CA 94025, USA

^cLawrence Livermore National Laboratory, Livermore, CA 94550, USA

^dDeutsches Elektronen Synchrotron (DESY), Hamburg, Hamburg 22607, Germany

^eLIDYL, CEA-Université Paris-Saclay, CEA Saclay, 91 191 Gif-sur-Yvette, France

ARTICLE INFO

Keywords:

Exascale Computing
Particle-in-Cell Methods
Accelerator Modelling

ABSTRACT

WarpX is a general purpose electromagnetic particle-in-cell code that was originally designed to run on many-core CPU architectures. We describe the strategy followed to allow WarpX to use the GPU-accelerated nodes on OLCF's Summit supercomputer, a strategy we believe will extend to the upcoming machines Frontier and Aurora. We summarize the challenges encountered, lessons learned, and give current performance results on a series of relevant benchmark problems.

1. Introduction

WarpX [20] is a fully electromagnetic Particle-in-Cell (PIC) code that is being developed as part of the US Department of Energy's Exascale Computing Project [8]. Originally designed for particle accelerator modelling, and in particular for the study of laser- and beam-driven wakefield accelerators, it has also been used to study several other topics in the field of laser-plasma interaction, such as probing the onset of Quantum Electrodynamics (QED) in extreme fields and laser-ion acceleration.

WarpX implements the well-known electromagnetic PIC method for solving the motion of relativistic, charged particles in the presence electromagnetic fields. In addition, it also includes support for many advanced features, such as: perfectly matched layers (PMLs) [16], a pseudo-spectral (PSATD) Maxwell solver [23], and multi-physics options such a ionization, QED pair creation [14, 10], a QED vacuum polarization solver [5], and binary collisions [15]. WarpX currently supports 2D, 3D and azimuthally decomposed geometries [12] and has the ability to operate in a Lorentz-boosted reference frame [18]. It also includes support for mesh refinement and dynamic load balancing through the AMReX library [26].

WarpX was originally designed with many-core architectures in mind. While the high-level operations such as time-stepping and MPI parallelization were implemented in C++ using AMReX data structures, the core PIC operations, such as current deposition, field gathering, and various par-

ticle pushers and field solvers, were handled by the PICSAR library of Fortran kernels [22]. These routines were highly optimized for the Intel's Knight's Landing architecture found on supercomputers such as NERSC's Cori and ALCF's Theta platforms and featured hand-vectorized versions of the core PIC operations. Related works presenting early adoptions of accelerator hardware in PIC codes are presented in Refs. [2, 3, 17, 22].

As the ECP focus shifted towards GPU-based machines such as Summit, Frontier, and Aurora, the question naturally arose about what to do with the Fortran kernels in PICSAR. CUDA Fortran provided a way forward for platforms with NVIDIA hardware such as Summit [21], but it was not clear what support for Fortran would look like on AMD or Intel hardware. Likewise, OpenACC provided an easy-to-use model for offloading Fortran routines to NVIDIA GPUs, but again it was not clear how that would work with non-NVIDIA GPUs and compiler support was limited to a single vendor. OpenMP offered better prospects for portability; however, early implementations suffered performance problems when compared to OpenACC on NVIDIA hardware.

Ultimately, the choice was made to port the PIC kernels in PICSAR from Fortran to C++, and to offload kernels using either CUDA, HIP, or DPC++, depending on whether NVIDIA, AMD, Intel hardware is targeted. This removed any need for mixed language programming, which adds substantial complication to the codebase and also defeats important compiler optimizations such as inlining. Additionally, C++ usually gets better and, importantly, earlier support from vendors, owing to its prominence in industry relative to Fortran. Finally, CUDA, HIP, and DPC++ offer a relatively consistent programming model across all three target platforms. Any implementation differences between the three could be entirely hidden in a performance portability layer; in our case, in the ParallelFor routines in AMReX (see Section 2.5). In this manner, NVIDIA, AMD, and Intel GPUs could all be supported, with little to no change to the

ORCID(s): 0000-0001-8427-8330 (A. Myers); 0000-0003-2103-312X (A. Almgren); 0000-0002-1445-0032 (L.D. Amorim); 0000-0002-7215-4178 (L. Fedeli); 0000-0003-3244-5525 (K. Gott); 0000-0002-4057-8582 (D.P. Grote); 0000-0003-1943-7141 (A. Huebl); 0000-0001-9432-2091 (R. Jambunathan); 0000-0002-3656-9659 (R. Lehe); 0000-0003-2406-1273 (M. Rowan); 0000-0003-4003-4507 (O. Shapoval); 0000-0001-7216-2277 (M. Thévenet); 0000-0002-0040-799X (J.-L. Vay); 0000-0002-9839-2692 (H. Vincenti); 0000-0002-9319-4216 (E. Yang); 0000-0002-9582-5894 (N. Zaïm); 0000-0001-8092-1974 (W. Zhang); 0000-0003-4362-3630 (Y. Zhao); 0000-0001-5662-4646 (E. Zoni)

WarpX code required.

Today, WarpX is a C++14 application with an optional, standardized Python interface (PICMI) that can be used to drive simulations interactively[4], and it runs on NVIDIA, AMD, and Intel GPUs. The core of the GPU support is the `ParallelFor` kernel launching method from AMReX. In what follows, we give a brief overview of the features in AMReX that WarpX uses to enable parallelization and GPU support. Then, we summarize several key lessons learned from the experience of scaling up the code on Summit. In particular, we discuss the importance of memory management, the importance of optimizing for memory footprint as well as run time, the importance of optimizing parallel communication routines, and the importance of properly utilizing the memory hierarchy. Finally, we will present weak and strong scaling results from a uniform plasma problem setup and performance numbers obtained on a plasma acceleration stage benchmark problem.

2. Parallelization

WarpX leverages the AMReX framework for parallelization and GPU offloading. In what follows we summarize the main features of the hierarchical parallelism model that WarpX uses to run on multi-node CPU and GPU platforms.

2.1. Domain Decomposition

WarpX makes use of the AMReX-provided tools for describing block-structured adaptive mesh refinement (AMR) hierarchies. From the simplest to the most complex, these are the `IntVect`, which describes a point in an integer index space; the `Box`, which describes a region in the same index space and consists of low- and high- end `IntVects` plus a third `IntVect` that describes the staggering (i.e. is the box cell-, node- face, or edge-centered); the `BoxArray`, which describes a collection of Boxes at a given level of refinement; and the `DistributionMapping`, which describes how each of those Boxes is mapped to MPI ranks. `Vector<BoxArray>` and `Vector<DistributionMapping>` then describe the mesh hierarchy across multiple levels of refinement. When used to describe a block in an AMR hierarchy like this, we use the term ‘grid’ interchangeably with `Box`.

`DistributionMappings` can be user-generated, or AMReX can generate them for a given `BoxArray` using a number of algorithms: round robin, knapsack, and space filling curve. By default, the boxes in WarpX are assigned to MPI ranks according to the space filling curve algorithm, which attempts to put nearby boxes on the same rank. When dynamic load balancing is employed, the user can select to use either space filling curve, which attempts to maintain spatial locality, or knapsack, which provides the most flexibility to achieve a balanced work distribution.

2.2. Mesh and particle data structures

The basic mesh data structure in AMReX is the `FArrayBox`, which is a multi-dimensional array of floating point values defined on a given `Box`. `FArrayBoxes` can be single- or multi-component and used to represent scalar or vector physical

quantities. In the multi-component case, the components are stored in a Struct-of-Array style. A distributed collection of `FArrayBoxes` defined on a given `BoxArray` and `DistributionMapping` is called a `MultiFab`. The `MultiFab` data structure is what stores the core mesh data fields in WarpX - the electric and magnetic fields, the current density, and so forth.

The core particle data structure in AMReX is the `Particle`, which consists of a collection of real and integer components. In WarpX, the majority of these components are stored in Struct-of-Array style. The exceptions are the particle positions and a 64-bit integer identification number, which are stored together in a separate `struct`.

A `ParticleContainer` is a distributed collection of particles associated with a given AMR hierarchy. In WarpX, each particle species (driver beam, plasma electrons, ions, etc...) is stored in a separate `ParticleContainer`. Particles are assigned to AMR levels and grids based on their physical positions. The particles in the `ParticleContainer` can then be looped over grid-by-grid, and PIC operations such as field gathering, particle pushing, and current deposition can be performed.

2.3. Hierarchical Parallelism

The above data structures naturally lend themselves to an “MPI+X” hierarchical approach to parallelism, where “X” is one of OpenMP, CUDA, HIP, or DPC++ for on-node accelerated compute. Boxes are assigned to MPI tasks, and we typically use a form of over decomposition so that each MPI task is responsible for processing multiple boxes. This allows for more flexible grid structures and also aids load balancing via swapping boxes across ranks. AMReX provides `MultiFab` and `ParticleContainer` iterator objects that can instruct each rank to loop over their local grids, processing each one in turn. When processing individual grids, an accelerated compute backend such as OpenMP or CUDA can be selected to perform the actual computations. See the section on `ParallelFor` for more information.

2.4. Parallel Communication routines

Communicating mesh and particle data between MPI ranks is handled by the AMReX framework. In particular, WarpX makes use of the following parallel communication routines:

- **FillBoundary:** This method is used to fill guard cells for the mesh data (e.g. the electric and magnetic field components, and current / charge densities). It fills the data in the guard cells with the (possibly more recent) data from the corresponding valid cells. Here, “valid” refers to cells that are uniquely owned by the grid in question, as opposed to ghosted copies that may exist on other grids / MPI processes.
- **SumBoundary:** This operation is analogous to `FillBoundary`, except that instead of copying from valid to guard, it takes the values in the ghost cells and adds them to corresponding valid cells. This is useful when doing current and charge deposition operations on particles that are near the edge of grid boundaries. These

particles add some of their weights to guard cells, and these contributions are summed to the proper valid cell by the `SumBoundary` method.

- **SyncNodal:** This function is used when staggered or node-centered grids are employed to represent physical quantities, for example, when using the Yee grid to represent E, B, and J. In this case, some points in the discretization are represented on multiple grids and potentially multiple MPI ranks. Note that, unlike in the cell-centered case, no one grid can be said to uniquely “own” these points. To prevent spurious numerical effects, it is necessary to synchronize these shared nodal points so that they have exactly the same values to machine precision. The `SyncNodal` method accomplishes this. Several different options for deciding which value to use are implemented, e.g. simple or weighted averaging. By default, WarpX simply chooses an arbitrary value as the ‘winner’ and overrides the others.
- **ParallelCopy:** This is the most general form of parallel communication for mesh data in AMReX. It performs copy on intersection from one `MultiFab` to another, even when those `MultiFabs` have different `BoxArrays` and `DistributionMappings`. This is needed when, for example, copying data between different levels of refinement, performing regridding or load balancing operations, and when copying data between the PML grids and the rest of the domain.
- **Particle Redistribution:** This refers to putting particles back on the proper level and grid after they have been pushed. AMReX includes two versions of this operation, one in which the particles are assumed to only move between neighboring ranks, and another in which they are allowed to move between any two ranks in the MPI communicator. The former, local version is the one used most often during normal time stepping, while the latter version is used when performing load balancing.

These parallel communication routines have been optimized for hybrid CPU/GPU platforms, in particular, Summit. All run fully on the GPUs, meaning that they do not trigger any unnecessary host/device copies of mesh or particle data. Communication has been refactored to reduce the number of GPU kernels launched, in particular when performing packing and unpacking operations on MPI send and receive buffers (see Section 3.3). Finally, WarpX can take advantage of gpu-aware MPI implementations that can operate directly on device data pointers, if one is available. This operation can be enabled using a runtime option. When turned off, right before the MPI sends are performed, the data to be sent is copied into pinned memory buffers on the host and the MPI exchanges are made between host memory buffers instead.

2.5. ParallelFor

The core of the GPU support in WarpX consists of a series of `ParallelFor` functions provided by the AMReX framework. These are similar to those provided by the performance portability layers Kokkos [6] and RAJA [1], but have been tailored towards the needs of structured grid applications. The idea of these functions is that they separate the details of how the loop is performed from the loop body, which describes the operation done on each elements. The body is supplied using C++ lambda functions, which capture (device) variables from the surrounding scope. When `amrex::ParallelFor` is compiled with CUDA as the backend, they translate into a GPU kernel launch. When they are compiled without GPU support, they translate into a normal `for` loop on the host. Using this approach, a single codebase can be maintained that can run on CPUs and on multiple GPU platforms. CPU loops can be further tiled for individual threads to aid vectorization on CPUs, which is outside of the scope for this GPU-focused paper and thus omitted for brevity.

Listings 1 and 2 show two examples of `ParallelFor`. The first specifies the loop bounds using an AMReX `Box` object, which results in a 3-dimensional loop over the cells in the box. The second shows a one-dimensional `ParallelFor`, which loops over the particles in a grid.

```
1 amrex::ParallelFor(bx,
2   [=] AMREX_GPU_DEVICE (int i, int j, int k)
3   noexcept
4   {
5     dstarr(i,j,k,0) = srcarr(i,j,k,0);
6   });
```

Listing 1: A `ParallelFor` routine operating on a single box of mesh data. In this example case, the threading will be performed over the cells of a 3-dimensional box. AMReX arrays use Fortran index order.

```
1 amrex::ParallelFor(np,
2   [=] AMREX_GPU_DEVICE (int i)
3   {
4     amplitude[i] = 0.0_rt;
5   });
```

Listing 2: A one-dimensional `ParallelFor` used to thread over all the particles in a grid.

2.6. Reductions

Parallel reductions are useful in many places in WarpX, including both “on-node” reductions over OpenMP threads (or other, GPU-based compute backends) and “off-node” reductions over MPI ranks. For example, these type of reductions are found in diagnostic functions for particles and beams that act as an *in situ* data reduction technique, which can be performed in high frequency compared to costly full data output.

Performing these operations efficiently in parallel in a way that is portable between CPU and various GPU platforms is non-trivial. To aid in this task, AMReX provides generic functions for performing reduction operations in a

performance-portable way. These functions can be used at the level of contiguous arrays by passing in data pointers, or they can work on higher-level AMReX data containers to, e.g., perform reductions over all the cells in a `MultiFab`, or all the particles in a `ParticleContainer`.

A feature of the AMReX implementation of parallel reductions is that it provides an API for performing multiple reductions in one pass on any combination of data types and reduction operators. When running on GPUs, all these operations would be done in a single kernel launch. These reductions operations have been tested and implemented for CUDA, HIP, and DPC++, as well as on CPU platforms using OpenMP.

3. Lessons from Summit

In the following section, we summarize some key lessons learned from our experience of porting WarpX to Summit. These fall under three main areas: issues relating to memory management and overall footprint, issues relating to parallel communication, and finally, the importance of cache utilization on GPU platforms.

3.1. Memory Optimization

With the trend towards GPU computing, the importance of optimizing codes for memory consumption has increased. Consider the example of Summit. Summit has 4608 nodes, each of which has 608 GB of host memory (512 DDR4 + 96 HBM2), for a system total of 2.8 Petabytes. This is considerably more than Cori's KNL nodes, which have a total aggregate memory of 1.1 Petabytes. However, if we consider only device memory, each Summit node has 6 NVIDIA V100 GPUs with 16 GB of memory each, for a total of only 440 TB, substantially less than Cori phase II. This means that, provided that one wants to run in a mode in which your problem entirely fits on the GPUs (which is desirable considering the performance penalties associated with frequent host / device data transfers), one actually cannot run as big of a problem on Summit than one could fit on Cori. This makes reducing the memory footprint of a simulation code quite impactful in terms of enabling production-level science calculations.

Reducing the memory footprint can have performance implications as well. Originally in WarpX, every particle stored persistent values for the electric and magnetic fields interpolated to the particle's position. In addition to the storage overhead, these values need to be communicated every time particles change MPI domains, and shuffled around in memory every time particles are sorted (see Section 3.4.2). Additionally, if the performance of a GPU kernel is memory-bound, meaning that its performance is limited by the rate at which data can be transferred from main memory to the streaming multiprocessors on the GPUs, then increasing the arithmetic intensity of those kernels by streaming less data and recomputing values on-the-fly can improve their overall performance.

Recently, WarpX removed the persistent electric and magnetic fields at the particle positions in favor of re-gathering

these values inside GPU kernels as they are needed. For this, the field gathering and particle pushing kernels were fused together in the PIC loop, resulting in less data that needed to be streamed to the processors in a given timestep. In addition to reducing the memory footprint by a factor of ≈ 1.6 , this also led to a $\approx 25\%$ percent speedup in the overall run-time on several key benchmarks. When the field values at the particle positions are needed more than once in a step, as, for example, when modelling additional effects such as ionization or using certain diagnostics, the gather operation is simply performed multiple times.

Finally, we are currently exploring other means of reducing the overall memory footprint of WarpX, including exploiting single / mixed precision and employing compression.

3.2. Memory Arenas

Dynamic memory allocation is many times more expensive on GPU than CPU architectures. This fact, combined with common programming patterns involving temporary variables, can lead to drastic performance penalties on GPU systems. For example, consider the code in Listing 3. This snippet demonstrates how to loop over mesh data using the AMReX data structures. The `MFIter` object instructs each MPI rank to loop over the grids it owns. For each grid, we resize a temporary scratch space called `tmp`, then launch a `ParallelFor` kernel to do some calculations using it. The `Elixir` is not essential to the point, but it keeps the scratch space alive in memory until the kernel is finished working with it - this is needed due to the asynchronous nature of GPU kernel launches. If every call to `resize` the buffer ended up triggering `cudaMalloc` and `cudaFree` calls, this could easily end up becoming the dominant cost of this routine. Another place this comes up is in out-of-place sorting and partitioning operations, which require a temporary buffer in which to store the result.

One way to mitigate this is to refactor application codes to keep temporary buffers alive in memory instead of letting them go out of scope. However, this is error-prone and labor-intensive. Instead, AMReX provides a number of memory arena classes, which allocate memory in large chunks and dole out pieces of it as the application runs. Thus, even though WarpX makes frequent use of temporary variables, during most time steps that are no calls to `cudaMalloc` or `cudaFree`.

These Arenas have a number of different options for managing memory fragmentation; currently, the default in AMReX is to use a "first fit" strategy. AMReX provides memory arenas that use host, device, pinned, and managed memory. WarpX uses these Arenas for all of its mesh and particle data structures. By default, when running on NVIDIA GPUs, WarpX places all of its core data in managed memory.

```

1 FArrayBox tmp;
2 for (MFIter mfi(mf); mfi.isValid(); ++mfi)
3 {
4     const Box& bx = mf.tilebox();
5     tmp.resize(bx);
6     Elixir eli = tmp.elixir();

```

```

7  auto const& tmp_arr = tmp.array();
8
9  amrex::ParallelFor(bx,
10 [=] AMREX_GPU_DEVICE (int i, int j, int k)
11     noexcept
12     {
13         compute_tmp(i, j, k, tmp_arr);
14     }

```

Listing 3: Example of ParallelFor. This code can be compiled to run on CPU with OpenMP or GPU with CUDA, HIP, or DPC++.

3.3. Communication Optimization

Once the initial port of WarpX to NVIDIA GPUs was complete, the initial experience was that compute kernels such as current deposition and field gathering were much faster on V100 hardware than on KNL. However, the same was not true for the AMReX parallel communication routines. The primary reason for this was that the parallel communication routines involved many small, "copy on intersection" routines between neighboring boxes, especially when packing and unpacking MPI send and receive buffers. These operations involved little to no computation but launched many small kernels that packed and unpacked data buffers. Thus, the dominant cost in these routines was the latency associated with the kernel launches, which could be fused into a fewer number. After optimization, each MPI rank makes only 1 kernel launch to pack and unpack its MPI buffers, which led to greatly improved performance on Summit.

3.4. Cache utilization

As with CPU-based many-core architectures, rearranging computations so that they properly exploit the memory hierarchy can lead to significant performance increases on V100 GPUs. In this section, we discuss a case-study in this effect - specifically, on how periodic sorting of particle data, so that it is processed in a cache-friendly way, can greatly improve the performance of PIC operations like field gathering and current deposition on V100. First, however, we will describe the current deposition algorithm we use, and how it differs between CPU and GPU runs, in more detail.

3.4.1. Current Deposition

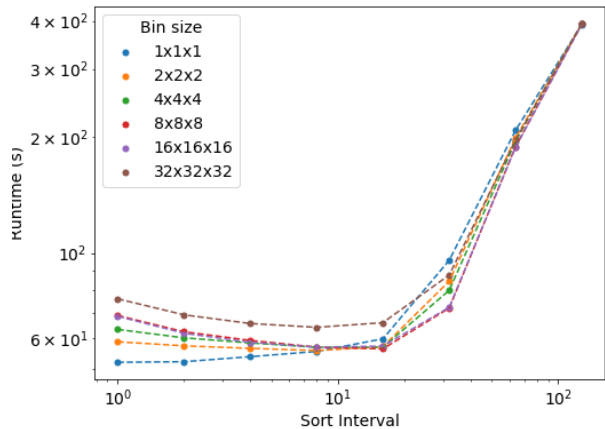
In PIC codes, most operations are straightforward to parallelize, since particles can be threaded over and processed independently without needing to worry about potential race conditions. Charge and current deposition operations, however, require special consideration, since when threading over particles there is the potential for collisions as multiple threads may attempt to update the same cell simultaneously.

In WarpX, our approach to concurrent scatter operations in particle deposition kernels varies depending on whether we are running with OpenMP or CUDA/HIP/DPC++ as the parallel backend. With OpenMP, the particles on a grid are sorted onto smaller sub-regions called tiles. OpenMP threads are mapping to tiles, which begin processing them simultaneously. Each OpenMP thread deposits particles onto its

own, private deposition buffer with enough cells to capture the support of all the particles on the tile. There is no need for atomics at this stage, since each thread has its own buffer. After deposition onto the buffer is complete, the buffer values are atomically added to the values for the full grid using atomic writes. Thus atomics are only needed on a per-cell basis, not a per-particle basis. When running on GPUs, however, we perform atomics write directly to global memory for each particle. This, along with periodic sorting, is sufficient to get good performance on NVIDIA V100 GPUs.

3.4.2. Particle Sorting

Figure 1: The effect of sorting interval (i.e., sorting every N time steps) and sort bin size on the overall performance on a uniform plasma benchmark. The x-axis shows the sort interval, while the y-axis shows the overall time to take 100 steps, including the cost of the sorting. A sort interval > 100 means that the particles are never re-sorted during the run.



Periodic sorting of the particles on each grid by their spatial locations so that particles that are close to each in memory also interact with cells that are close to each in memory exploits the memory hierarchy on the GPUs more effectively than processing them in an unordered fashion. This is particularly true in the case that particles are moving with high velocities, such that they frequently change cells. In that case, even if particles are sorted at a particular time, they will rapidly become disordered, leading to significant performance degradation in the particle-mesh operations.

Note that we differentiate between binning, which computes a permutation array that assigns particle indices to cells with user-defined bin size, and sorting, which uses this permutation array to actually reorder the particle data in memory. Cache utilization requires full sorting, but for many operations simply knowing the cell-sorted indices is sufficient. AMReX provides a GPU-capable implementation of the counting sort operation that can be used to perform both of these operations. Internally, it is built using a GPU implementation of parallel prefix sum, which is based on Ref. [13] and works on NVIDIA, AMD, and Intel GPUs.

In addition to the presented cache-utilization optimiza-

Figure 2: Roofline analysis of the 3rd-order Esirkepov current deposition [7] kernel in WarpX on a single V100 GPU, with and without particle sorting. In the memory streaming limit, three different lines are shown, corresponding to the bandwidths of the L1 and L2 caches as well as that for the main high-bandwidth memory (HBM) on the GPU. Likewise, in the compute-bound regime, two different values are used for the peak floating point performance: both with and without taking advantage of fused multiply-add instructions. The arithmetic intensity (A.I.) is measured three times for each kernel, using the memory traffic for each level of the memory hierarchy. For the sorted version, the fact the A.I. is significantly lower for the L1 and L2 data points shows that we are getting substantial reuse in both levels of cache. Conversely, the fact that the data points are all on top of each for the unsorted run indicates that without sorting, the degree of reuse is poor.

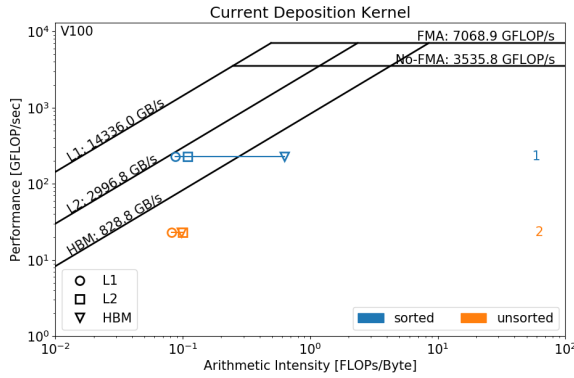
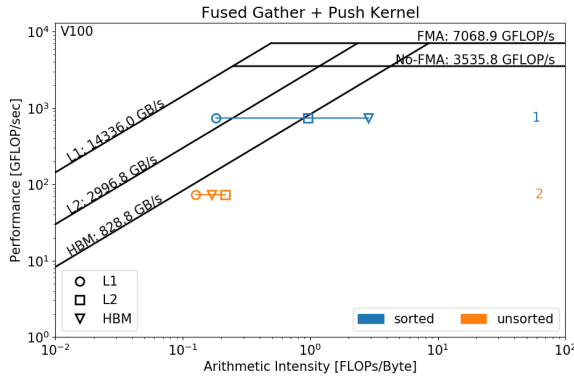


Figure 3: Same as Figure 2, but for the fused gather and push kernel in WarpX. Again, there is substantial cache reuse when sorting is employed, although for this kernel performance still appears to be limited by HBM bandwidth, even with sorting.



tion, sorting and/or binning particles is needed for the modeling of particle-particle interactions. The PIC method by default only models particle-mesh interaction and mesh updates. WarpX implements binary collisions, which depend on a prior binning of neighboring particles, to address various applications in accelerator and beam physics.

Figure 1 shows the results of a parameter study in which the bin size and sorting interval were varied. For example, a bin size of $2 \times 2 \times 2$ and sorting interval of 4 means that par-

ticles were sorted into $2 \times 2 \times 2$ supercells every 4 timesteps. On this problem, the optimal sorting is to sort by cell (i.e. a bin size of $1 \times 1 \times 1$ every time step, and the difference between sorting optimally and not sorting at all is a factor of ≈ 7.5 , with most of the improvement coming from the current deposition and fused gather and push kernels. However, this very frequent sorting interval for this problem is a special because, because the particles in this problem change cell more often than in most WarpX applications. Currently, the default in WarpX, used throughout Section 4, is to sort the particles by their PIC cell every 4 time steps.

Note that, although the `Redistribute()` function in AMReX does not maintain this cell-sorted order for particles that left one grid and been migrated to another, this only applies to particles that have changed grids - typically only a small subset of the total that are near the “surface”. The bulk of the particles on a grid will maintain their sorted order in between `Redistribute()` calls.

Figures 2 and 3 show the results of a roofline analysis [24] on the current deposition and fused gather and push kernels in WarpX, which are the two most computationally expensive operations. Our analysis followed the methodology of [25]. For this test, we used a uniform plasma setup with 8 particles per cell and gave the particles a large thermal velocity, so that they frequently change cells. To eliminate effects associated with unified memory paging, we ran the problem for a total of 100 steps and only profiled the last one.

The roofline analysis reveals three things. First, as already demonstrated, sorting the particles gives significantly better performance on V100 GPUs than not sorting them. Second, the fact that the arithmetic intensity measured using the memory bandwidth for the L1 and L2 caches is significantly lower than for HBM indicates that, in the sorted run, we are getting significant reuse in both of these levels of cache. Third, the arithmetic intensity for the current deposition for the sorted run is right up against the streaming limit for the L2 cache. This indicates that the performance of this kernel is now limited by the L2 cache bandwidth. Gather and push, on the other hand, is likely still limited by HBM bandwidth. Taken together, these results suggest that these kernels should get significantly better performance on the A100, which has a larger L2 cache and higher HBM bandwidth than the V100.

4. Performance Results

In this section, we give current performance results on Summit for two key benchmark problems. We concentrate on two areas - the scaling of the code on a uniform plasma test case and the performance on a plasma accelerator benchmark problem.

4.1. Uniform Plasma Scaling

4.1.1. Weak scaling study

In order to test the scaling of WarpX in an idealized setting, as well as to gauge the speedup associated with using accelerated nodes, we have performed a weak scaling

study using a uniform plasma setup on OLCF's Summit supercomputer. The base case for this scaling study used a $256 \times 256 \times 384$ domain with a box size of 128^3 and ran on 1 Summit node; thus, on the GPU-accelerated runs, each GPU was responsible for processing two 128^3 -sized boxes. Particles were initially distributed uniformly with 8 particles per cell. We used the standard Yee FDTD solver for these runs, with Esirkepov current deposition and third order shape functions. For the weak-scaling study, the number of Summit nodes were doubled with the number of cells (and particles therein) in the x-, y-, or z- directions, while holding everything else constant, maintaining a constant workload per node. We continued this process up to 2048 nodes - about half of the Summit machine. Overhead associated with time spent in problem initialization, memory allocation, etc., was minimized by running for a total of 100 steps.

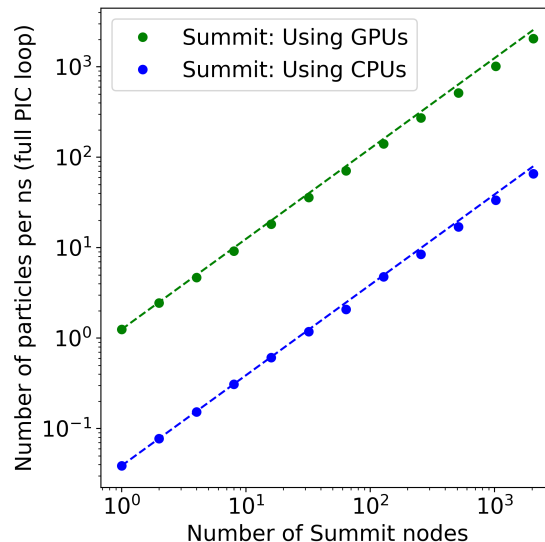
The results are shown in Figure 4. We performed the above scaling study twice, once using all six GPUs per Summit node, and again using only the POWER9 CPUs. For both runs, we used 6 MPI tasks per node. For the GPU-accelerated runs, we used one GPU per MPI task, and for the CPU-only case, we used 7 OpenMP threads per task, so that all 42 cores on the node were active. Using these results, we can characterize both the weak scaling behavior of the CPU and GPU versions of the WarpX, as well as see the overall speedup obtained on Summit from using the accelerators. In both cases, the code scales well up to 2048 nodes. The weak scaling efficiency, defined as the total time taken for 100 time steps on 1 node divided by the total taken on 2048 nodes, is 81% for the GPU case and 90% for the CPU case. The difference in scaling efficiency between the CPU and GPU can be attributed to the fact that, because the local work is significantly faster when using the V100s, communication operations like `FillBoundary`, which are inherently harder to scale, become relatively more expensive. Additionally, the speedup from the accelerators at all scales tested was a factor of 30. This speedup refers to the total run time, including time associated with host / device memory traffic and communication, not to isolated compute kernels.

4.1.2. Strong Scaling study

We have also conducted a series of strong scaling tests, using a very similar uniform plasma problem setup as before. The only difference is that the box size has been set to 64^3 , to allow for more GPUs / MPI tasks to be used as the problem is strong scaled. There is some overhead associated with doing this, since with smaller boxes, the surface to volume ratio of ghost cells is higher. Other than the box size, the parameters are all the same as before.

We use a series of problem sizes, each scaled up a factor of 2 in terms of the number of cells and particles in the domain. For each one, we conduct a series of five runs, increasing the number of MPI tasks by a factor of 2 each time. Thus, in the fifth run, the run time should have decreased by a factor of 16, assuming perfect strong scaling. By the time we have multiplied the number of MPI ranks by 16, this problem has reached the point where the compute work and the

Figure 4: Results of a weak scaling study on a uniform plasma setup on Summit. The x-axis shows the number of Summit nodes, while the y-axis is the number of particles advances per nanosecond. Both the CPU and GPU versions of the code scale well, and the overall speedup associated with using the accelerators is ~ 30 .



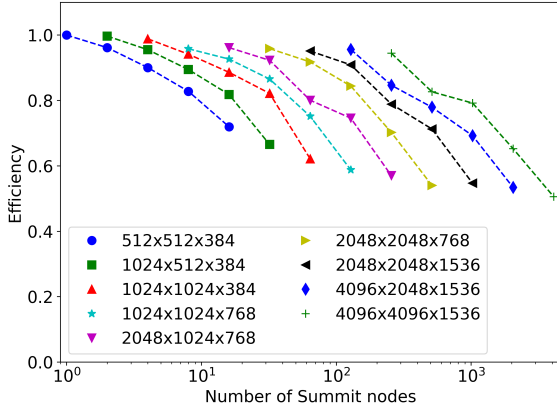
communication work take approximately the same amount of time, so we would not expect the problem to scale further than that.

The smallest scaling study in this series goes from 1 to 16 nodes, while the largest goes from 256 to 4096, nearly the entire machine. The scaling efficiency, defined as the time a run should take assuming perfect strong scaling within a problem size *and* perfect weak scaling from the base problem size divided by the actual run time, is plotted in Figure 5. The efficiencies after strong scaling by a factor of 16 for each problem size vary from approximately 70% for the smallest case to approximately 50% for the largest.

4.2. Plasma Acceleration Stage

The above tests were highly idealized in several ways. First, the workload was perfectly uniform at initial time, and approximately uniform at later times, subject only to random fluctuations in the particle density from cell to cell. Second, the number of particles per cell, 8, is significantly higher than used in some WarpX physics applications. Laser-wake-field acceleration runs, for example, tend to use about 2 particles per cell on average, which can change the performance profile of the code. Evaluating WarpX on this important science scenario, the following setup was used, designed to mimic the essential features of modelling a single plasma-accelerator stage from WarpX's challenge problem. This is also the benchmark problem used to determine a Figure-of-Merit (FOM) for the ECP Key-Performance Parameters (KPP) assessment. As a KPP-1 project, WarpX needs to show at least a factor of 50 increase in its FOM over the baseline on the eventual Exascale hardware. In this setup,

Figure 5: Strong scaling studies for a variety of problem sizes. Each tick type refers to a different problem size. The x -axis shows the number of Summit nodes, and the y -axis shows scaling efficiency, defined as the time a run should take assuming perfect strong scaling within a problem size *and* perfect weak scaling from the base problem size, divided by the actual run time.



an accelerated particle beam is tracked using the moving window feature in WarpX, in which the simulation domain itself shifts along with the beam at speed c . Additionally, the entire simulation is modeled in a Lorentz-boosted reference frame [18], using a gamma boost of 30. New plasma is continuously injected at the right-hand side of the domain, while particles that leave the domain at the left-hand side are removed from the simulation. The plasma consists of two particles per cell (one electron and one proton), while the accelerated beam is comprised of electrons. Mitigating the numerical Cherenkov instability in the modeling of a relativistically flowing plasma, the Godfrey filter [9] is applied to the electromagnetic fields prior to gathering them to particle positions. For the algorithmic options, we have used the Vay particle pusher [19], the Cole-Karkkainen-Cowan FDTD solver [11], and energy-conserving field gathering. We have again used Esirkepov current deposition with 3rd-order interpolation. To minimize the computer time needed to conduct these simulations, we initialize the problem to have the simulation domain entirely filled with plasma, which would normally not be the case when modelling an accelerator stage.

To gauge the impact of using accelerated nodes on this more realistic problem setup, we have measured the FOM on Summit, defined as

$$\text{FOM} = \text{num_cells} * (\alpha + \beta * \text{ppc}) / \text{avg_time_per_it} \quad (1)$$

where num_cells is the total number of grid points in the simulation, α is 0.1 as heuristic grid update cost, β is 0.9 for particle update costs, ppc is the average number of particles per cell, and avg_time_per_it is the average time per iteration after 1000 steps. We performed this measurement on 4263 Summit nodes, and extrapolated this number to the full machine assuming perfect weak scaling. Our baseline

FOM was measured on NERSC's Cori using the original Warp code. The baseline FOM value, measured in March 2019 on 6625 Cori nodes and extrapolated to the 9668 on the full machine, was $2.2e10$. The corresponding value on Summit, measured in July 2020, was $2.5e12$, over a factor of 100 improvement from the baseline. Additionally, the best CPU-only FOM obtained using the WarpX code was $1.0e11$, also measured in March 2019. So there is a substantial (25x) improvement in our FOM measured with WarpX from using the GPUs on Summit, as compared to Cori. These values are all summarized in Table 1, along with several other data points showing the evolution of WarpX's FOM over time.

Table 1

Progress in the FOM measurement over time. Code: either the original Warp code (baseline) or WarpX. Date: the date when the measurement was taken. Machine: which computer was used to make the measurement. Nodes: how many nodes the measurement was performed on; there are 9668 KNL nodes on Cori and 4608 nodes on Summit. FOM: the figure of merit, extrapolated from the number of nodes the measurement was taken on to the full machine.

Code	Date	Machine	Nodes	FOM
Warp	3/2019	Cori (KNL)	6625	$2.2e10$
WarpX	3/2019	Cori (KNL)	6625	$1.0e11$
WarpX	6/2019	Summit	32	$8.6e11$
WarpX	6/2019	Summit	1000	$7.8e11$
WarpX	9/2019	Summit	2560	$6.8e11$
WarpX	1/2020	Summit	2560	$1.0e12$
WarpX	2/2020	Summit	4263	$1.2e12$
WarpX	6/2020	Summit	4263	$1.4e12$
WarpX	7/2020	Summit	4263	$2.5e12$

5. Conclusion

We have summarized the approach taken to porting WarpX, which was originally designed for many-core CPU architectures, to take advantage of GPU-accelerated nodes. This approach is largely based on the `amrex::ParallelFor` set of performance portability functions. We have summarized several key lessons learned from the port, including the importance of managing memory allocation and the code's overall memory footprint, the importance of minimizing the effect of kernel launch latency in MPI communication routines, and the importance of utilizing the cache hierarchy on V100 GPUs. The GPU port of WarpX scales up to nearly all of Summit and currently sees good improvements in its KPP-1 figure of merit on Summit relative to its baseline.

6. Acknowledgements

This research was supported by the Exascale Computing Project (17-SC-20-SC), a joint project of the U.S. Department of Energy's Office of Science and National Nuclear Security Administration, responsible for delivering a capable exascale ecosystem, including software, applications, and

hardware technology, to support the nation's exascale computing imperative. This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231, SLAC National Accelerator Laboratory under contract AC02-76SF00515, and Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

WarpX is developed as open source project and available under <https://github.com/ECP-WarpX/WarpX>. Presented code versions correspond to the monthly releases of the code between 3/2019 and 10/2020. The data that support the findings of this study are available under DOI:10.5281/zenodo.4277941.

References

- [1] Beckingsale, D.A., Burmark, J., Hornung, R., Jones, H., Killian, W., Kunen, A.J., Pearce, O., Robinson, P., Ryuji, B.S., Scogland, T.R., 2019. Raja: Portable performance for large-scale scientific applications, in: 2019 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC), pp. 71–81. doi:10.1109/P3HPC49587.2019.00012.
- [2] Burau, H., Widera, R., Hönig, W., Juckeland, G., Debus, A., Kluge, T., Schramm, U., Cowan, T.E., Sauerbrey, R., Bussmann, M., 2010. Piongpu: A fully relativistic particle-in-cell code for a gpu cluster. IEEE Transactions on Plasma Science 38, 2831–2839. doi:10.1109/TPS.2010.2064310.
- [3] Bussmann, M., Burau, H., Cowan, T.E., Debus, A., Huebl, A., Juckeland, G., Kluge, T., Nagel, W.E., Pausch, R., Schmitt, F., Schramm, U., Schuchart, J., Widera, R., 2013. Radiative signatures of the relativistic kelvin-helmholtz instability, in: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, ACM, New York, NY, USA. pp. 5:1–5:12. URL: <https://github.com/ComputationalRadiationPhysics/piongpu>, doi:10.1145/2503210.2504564.
- [4] CAMPA Collaboration, . Particle-in-cell modeling interface (PICMI). URL: <https://github.com/picmi-standard/picmi>.
- [5] Carneiro, P., Grismayer, T., Fonseca, R., Silva, L., 2017. Quantum electrodynamics vacuum polarization solver. arXiv:1607.04224.
- [6] Carter Edwards, H., Trott, C.R., Sunderland, D., 2014. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. Journal of Parallel and Distributed Computing 74, 3202 – 3216. doi:10.1016/j.jpdc.2014.07.003. domain-Specific Languages and High-Level Frameworks for High-Performance Computing.
- [7] Esirkepov, T., 2001. Exact charge conservation scheme for particle-in-cell simulation with an arbitrary form-factor. Computer Physics Communications 135, 144 – 153. doi:10.1016/S0010-4655(00)00228-9.
- [8] Exascale Computing Project, . Homepage. URL: <https://www.exascaleproject.org>.
- [9] Godfrey, B., Vay, J.L., 2014. Suppressing the numerical cherenkov instability in ftdt pic codes. Journal of Computational Physics 267. doi:10.1016/j.jcp.2014.02.022.
- [10] Gonoskov, A., Bastrakov, S., Efimenko, E., Ilderton, A., Marklund, M., Meyerov, I., Muraviev, A., Sergeev, A., Surmin, I., Wallin, E., 2015. Extended particle-in-cell schemes for physics in ultrastrong laser fields: Review and developments. Phys. Rev. E 92, 023305. doi:10.1103/PhysRevE.92.023305.
- [11] Karkkainen, M., Gjonaj, E., Lau, T., Weiland, T., 2006. Low-Dispersionwake Field Calculation Tools, in: Proc. Of International Computational Accelerator Physics Conference, Chamonix, France. pp. 35–40.
- [12] Lifschitz, A., Davoine, X., Lefebvre, E., Faure, J., Rechatin, C., Malka, V., 2009. Particle-in-cell modelling of laser-plasma interaction using fourier decomposition. Journal of Computational Physics 228, 1803 – 1814. doi:10.1016/j.jcp.2008.11.017.
- [13] Merrill, D., Garland, M., 2016. Single-pass Parallel Prefix Scan with Decoupled Lookback. Technical Report NVR2016-001. NVIDIA Research. URL: https://research.nvidia.com/sites/default/files/pubs/2016-03_Single-pass-Parallel-Prefix-Scan-2016-002.pdf.
- [14] Nikishov, A., 1970. Pair production by a constant external field. Sov. Phys. JETP 30, 660.
- [15] Pérez, F., Gremillet, L., Decoster, A., Drouin, M., Lefebvre, E., 2012. Improved modeling of relativistic collisions and collisional ionization in particle-in-cell codes. Physics of Plasmas 19, 083104. doi:10.1063/1.4742167.
- [16] Shapoval, O., Vay, J.L., Vincenti, H., 2019. Two-step perfectly matched layer for arbitrary-order pseudo-spectral analytical time-domain methods. Computer Physics Communications 235, 102 – 110. doi:10.1016/j.cpc.2018.09.015.
- [17] Surmin, I., Bastrakov, S., Matveev, Z., Efimenko, E., Gonoskov, A., Meyerov, I., 2016. Co-design of a particle-in-cell plasma simulation code for intel xeon phi: A first look at knights landing, in: Algorithms and Architectures for Parallel Processing, Springer International Publishing, Cham. pp. 319–329. doi:10.1007/978-3-319-49956-7_25.
- [18] Vay, J.L., 2007. Noninvariance Of Space- And Time-Scale Ranges Under A Lorentz Transformation And The Implications For The Study Of Relativistic Interactions. Physical Review Letters 98, 130405/1–4. doi:10.1103/PhysRevLett.98.130405.
- [19] Vay, J.L., 2008. Simulation of beams or plasmas crossing at relativistic velocity. Physics of Plasmas 15, 056701. doi:10.1063/1.2837054.
- [20] Vay, J.L., Almgren, A., Amorim, L.D., Bell, J., Ge, L., Gott, K., Grote, D.P., Hogan, M., Huebl, A., Jambunathan, R., Lehe, R., Myers, A., Ng, C., Park, J., Rowan, M., Shapoval, O., Thévenet, M., Zhang, W., Zhao, Y., Zoni, E., 2020. Toward the modeling of chains of plasma accelerator stages with WarpX. Journal of Physics: Conference Series 1596, 012059. URL: <https://github.com/ECP-WarpX/WarpX>, doi:10.1088/1742-6596/1596/1/012059.
- [21] Vazhkudai, S.S., de Supinski, B.R., Bland, A.S., Geist, A., Sexton, J., Kahle, J., Zimmer, C.J., Atchley, S., Oral, S., Maxwell, D.E., Larrea, V.G.V., Bertsch, A., Goldstone, R., Joubert, W., Chambreau, C., Appelhans, D., Blackmore, R., Casses, B., Chochia, G., Davison, G., Ezell, M.A., Gooding, T., Gonsiorowski, E., Grinberg, L., Hanson, B., Hartner, B., Karlin, I., Leininger, M.L., Leverman, D., Marroquin, C., Moody, A., Ohmacht, M., Pankajakshan, R., Pizzano, F., Rogers, J.H., Rosenburg, B., Schmidt, D., Shankar, M., Wang, F., Watson, P., Walkup, B., Weems, L.D., Yin, J., 2018. The design, deployment, and evaluation of the coral pre-exascale systems, in: SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 661–672. doi:10.1109/SC.2018.00055.
- [22] Vincenti, H., Lobet, M., Lehe, R., Sasanka, R., Vay, J.L., 2017. An efficient and portable simd algorithm for charge/current deposition in particle-in-cell codes. Computer Physics Communications 210, 145 – 154. doi:10.1016/j.cpc.2016.08.023.
- [23] Vincenti, H., Vay, J.L., 2016. Detailed analysis of the effects of stencil spatial variations with arbitrary high-order finite-difference Maxwell solver. Computer Physics Communications 200, 147–167. doi:10.1016/j.cpc.2015.11.009.
- [24] Williams, S., Waterman, A., Patterson, D., 2009. Roofline: An insightful visual performance model for multicore architectures. Commun. ACM 52, 65–76. doi:10.1145/1498765.1498785.
- [25] Yang, C., Kurth, T., Williams, S., 2020. Hierarchical roofline analysis for gpus: Accelerating performance optimization for the nersc-9 perlmuter system. Concurrency and Computation: Practice and Experience 32, e5547. doi:10.1002/cpe.5547. e5547 cpe.5547.
- [26] Zhang, W., Almgren, A., Beckner, V., Bell, J., Blaschke, J., Chan, C., Day, M., Friesen, B., Gott, K., Graves, D., Katz, M.P., Myers, A., Nguyen, T., Nonaka, A., Rosso, M., Williams, S., Zingale, M., 2019.

AMReX: a framework for block-structured adaptive mesh refinement.
Journal of Open Source Software 4, 1370. URL: <https://ccse.lbl.gov/AMReX>, doi:10.21105/joss.01370.