Highly-parallelized simulation of a pixelated LArTPC on a GPU

Article (Published Version)


This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/112536/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk
Highly-parallelized simulation of a pixelated LArTPC on a GPU

To cite this article: A. Abed Abud et al 2023 JINST 18 P04034

View the article online for updates and enhancements.

You may also like

- Narrowband Searches for Continuous and Long-duration Transient Gravitational Waves from Known Pulsars in the LIGO-Virgo Third Observing Run

- A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals
- B P Abbott, R Abbott, T D Abbott et al.
Highly-parallelized simulation of a pixelated LArTPC on a GPU

The DUNE Collaboration

E-mail: roberto@lbl.gov

ABSTRACT: The rapid development of general-purpose computing on graphics processing units (GPGPU) is allowing the implementation of highly-parallelized Monte Carlo simulation chains for particle physics experiments. This technique is particularly suitable for the simulation of a pixelated charge readout for time projection chambers, given the large number of channels that this technology employs. Here we present the first implementation of a full microphysical simulator of a liquid argon time projection chamber (LArTPC) equipped with light readout and pixelated charge readout, developed for the DUNE Near Detector. The software is implemented with an end-to-end set of GPU-optimized algorithms. The algorithms have been written in Python and translated into CUDA kernels using Numba, a just-in-time compiler for a subset of Python and NumPy instructions. The GPU implementation achieves a speed up of four orders of magnitude compared with the equivalent CPU version. The simulation of the current induced on $10^3$ pixels takes around 1 ms on the GPU, compared with approximately 10 s on the CPU. The results of the simulation are compared against data from a pixel-readout LArTPC prototype.

KEYWORDS: Detector modelling and simulations II (electric fields, charge transport, multiplication and induction, pulse formation, electron emission, etc); Simulation methods and programs; Noble liquid detectors (scintillation, ionization, double-phase); Time projection Chambers (TPC)

ArXiv ePrint: 2212.09807
The idea of using a liquid argon time projection chambers (LArTPC) for the detection of neutrino interactions was first proposed in 1977 [1]. The detection mechanism is the following: charged particles produced by neutrino interactions ionize the argon, leaving a trail of ionization electrons. In addition, liquid argon also produces scintillation light, which provides calorimetric information and a fast timing signal ($O(10 \text{ ns})$ [2]). A fraction of the ionized electrons recombine immediately with the positive argon ions, while the remaining ones drift towards the anode side of the detector in a homogeneous electric field applied to the argon volume, which is usually $O(100 \text{ V/cm})$. Impurities present in the LAr (e.g. $\text{O}_2$, $\text{H}_2\text{O}$, $\text{N}_2$) can attach a portion of the drifting electrons. The amount of drifting electrons declines as a function of the distance from the anode, since the electrons need to travel a longer path.

Typically, two or more arrays of sense wires are placed at the anode and assembled into planes. The drifting of negative charges in a constant electric field induces a signal on the wires. Each plane provides...
a two-dimensional image of the ionization: the position of the wire provides one dimension, and the
time of the arrival provides the second one, since the drift velocity of the electrons in the LAr is known
and is typically $O(1 \text{ m/ms})$. Using multiple wire planes can help estimate the position of the ionization
in three dimensions. However, ambiguities arise when drifting electrons are isochronous or parallel to
a wire orientation. Unambiguous 3D imaging of LArTPC charge signals is possible using a readout
system based on a pixelated array of charged-sensitive pads, which has been demonstrated in ref. [3].
Both the typical distance between adjacent wires and the typical pixel pitch are in the order of few millimeters. The position on the anode plane of the involved pads provides two spatial dimensions, and the
time of the induced signal provides the third one. This truly three-dimensional readout provides better
reconstruction efficiency and purity than the 2D combined wire readout, as demonstrated in ref. [4].

Pixel readout requires the channel count used to be increased by a factor of 10 to 100 with
respect to wire planes. Thus, with this increased channel count and granularity simulation burden,
the transport of electrons in LAr and the signal induction on the pixel pads represent an ideal use
case for highly-parallelized, concurrent simulation algorithms. The development of general-purpose
computing on graphics processing units (GPGPU), which went hand in hand with the advances in
the machine learning and deep learning fields, has driven the design of the current generation
of supercomputers, such as Perlmutter at NERSC [5], which is a heterogeneous system with both
GPU-accelerated and CPU-only nodes. Implementing highly-parallelized simulation chains allows
full advantage of these new systems to be taken and enables the simulation of future pixelated
LArTPCs, which would otherwise not be viable with current resources.

The US-based neutrino physics program relies on present and future experiments using LArTPC
technology. The flagship experiment is the Deep Underground Neutrino Experiment (DUNE),
which will consist of a high-intensity accelerator neutrino beam, measured by near and far detectors [6].
The Far Detector will consist of four 17-kiloton LArTPC modules located deep underground at the
Sanford Underground Research Facility in Lead, South Dakota, located 1285 km from the beam
source [7]. The Near Detector will be located at Fermilab, 574 m from the beam source, and will
contain a 67 t modular LArTPC called ND-LAr [8].

There are already some software toolkits for LArTPCs [9, 10] and there has been some effort
towards parallelizing the reconstruction stage [11]. However, the simulation stages have remained
mostly sequential and their speed up has been recognized as a priority by the community [12].

In this document we will describe the implementation of a set of highly-parallelized algorithms,
organized in a module called larnd-sim [13], that run on GPUs. They simulate the ionized electrons
recombination and drifting towards the anode, the generation of electronics signals on the pixelated
readout, and the processing of the signal by the front-end electronics.

2 Technical implementation

Recent rapid developments in the field of machine learning have stimulated the creation of several tool-
kits for GPU-accelerated applications. In particular, the NVIDIA® CUDA platform [14] allows to use
GPUs for general purpose computing via different programming languages. We opted for Numba [15],
which generates CUDA computing kernels using a subset of native Python and NumPy code [16].

A CUDA kernel is a function that is executed $N$ times in parallel by $N$ CUDA threads. The
threads can be organized in one-dimensional, two-dimensional, or three-dimensional *blocks*, which
in turn can be organized in one-dimensional, two-dimensional, or three-dimensional grids. Blocks in the same grid contain the same number of threads, can run independently, and can be executed in any order, while threads in the same block can co-operate through shared memory. CUDA kernels typically store the result of the computation in a pre-allocated array passed to the kernel function. The CUDA programming model requires a careful design of the algorithm: the shape and size of the array where the result is stored must be known in advance and the threads must avoid race conditions during execution,\footnote{In software, a race condition can happen when the behavior of the program depends on the relative timing of multiple threads or processes.} thus the result of the algorithm must not depend on the order of execution of the threads.

The software described in this document contains several CUDA kernel functions, separated into two logical categories: one for the charge simulation, described in section 3, and one for the light simulation, described in section 4. The functions simulate the detector response, including: (1) the recombination of the electrons with the argon ions, (2) the drifting of the electrons towards the anode, (3) the induction of electronic signals on the pixel pads and optical detectors, and (4) the electronics response of the charge and light readout systems.

The simulation of the passage of the initial particles through matter is performed using edep-sim\cite{17}, a wrapper around Geant4\cite{18}, which is independent from the larnd-sim package described here. The output consists of a set of short particle track segments, in the order of few millimeters, which describe the energy deposition trail of each particle. The length of the segments depends on the derivative of the stopping power $dE/dx$: the portion of a particle trail where the $dE/dx$ changes abruptly will be divided in finer segments than the portion where the $dE/dx$ is mostly constant. Thus, in a single segment, the energy deposition per unit length is assumed to be constant. This approximation is valid when the segment length is of the same order as the detector resolution — at much shorter lengths, fluctuations from the long tail of the $dE/dx$ distribution fall outside of the applicable region of recombination models, and at much longer lengths, correlations in the dE/dx fluctuations become significant and the $dE/dx$ width will be under-simulated. Within these broad considerations, reducing the minimum size of the segments has not shown a significant impact on the result of the charge simulation. This set $S_i$ is stored in a bi-dimensional NumPy array containing the energy deposition and the spatial distribution of the segments:

$$S_i = (\vec{r}_s, \vec{r}_e, E)_i,$$

where $\vec{r}_s$ and $\vec{r}_e$ are four-dimensional vectors containing the spatial and timing coordinates of the segment start and end points, and $E$ is the deposited energy. This array is used as input for our module. In order to minimize the memory transfer between the host and the device (in our case the GPU), we allocate the NumPy array directly on the device memory using CuPy\cite{19}, a GPU array backend that implements a subset of the NumPy interface. The output of the larnd-sim simulation is then saved in a HDF5 file\cite{20}. The entire simulation workflow is shown in figure 1.

3 Charge simulation

3.1 Electron recombination

The first step of our simulation is to calculate the number of electrons that remain after the recombination and start drifting towards the anode. We denote the initial charge ionized by the
particle as $Q_0$, while the charge remaining after the recombination is given by $Q_R = R \cdot Q_0$, where $R$ is our recombination factor.

Two different models are commonly used to describe this phenomenon: the Birks model [21], which gives spurious values when applied to high-ionization particles [22], and the modified Box model [22], which doesn’t suffer from these issues but is inadequate to describe particles at low stopping power (low $dE/dx$).

The recombination factor for the Birks model $R_{\text{Birks}}$ can be parametrized as:

$$R_{\text{Birks}} = \frac{A_b}{1 + k_b/\epsilon \cdot dE/dx}, \quad (3.1)$$

where $A_b$ and $k_b$ are free parameters that usually depend on the detector and $\epsilon$ is the product of the electric field with the liquid argon density. The ICARUS collaboration obtained $A_b = 0.800$ and $k_b = 0.0486$ (kV/MeV)(g/cm$^3$) [23], which are the values used in our simulation.

The recombination factor the modified Box model $R_{\text{Box}}$ is defined as:

$$R_{\text{Box}} = \frac{\log(\alpha + \beta/\epsilon \cdot dE/dx)}{\beta/\epsilon \cdot dE/dx}, \quad (3.2)$$

where $\alpha$ and $\beta$ are free parameters which were measured by the ArgoNeuT collaboration to be $\alpha = 0.93$ and $\beta = 0.207$ (kV/MeV)(g/cm$^3$) [22]. In both cases, the typical recombination factor for a minimum ionizing particle (MIP) is around 0.7. Our simulation assumes the Birks model by default.

The implementation of the calculation of the recombination factors $R_{\text{Birks}}$ or $R_{\text{Box}}$ on the GPU is trivial: the $i$-th thread of the $K_{\text{recomb}}(S_i, R)$ CUDA kernel takes as input the $i$-th row of the NumPy array containing the segments $S$ and applies the recombination formula, so eq. (3.1) or eq. (3.2). The result is stored in an appropriate column of the NumPy array. Currently, the simulation treats the recombination factor as a fixed number: although this is an approximation, the level of fluctuations associated with it is significantly lower than the expected intrinsic noise of the detector, which was measured with the prototype described in section 6.

Even if this operation is computationally inexpensive, it’s instructive to take a look at the performance comparison between a sequential, interpreted Python for loop, a loop compiled on the CPU using Numba, and the GPU implementation using a CUDA kernel. Figure 2 shows the processing time needed to calculate $R_{\text{Birks}}$ of eq. (3.1) as a function of the number of GeANT4
segments given as input. The CPU-compiled version is obviously faster than the sequential interpreted loop since the function is now translated into machine code. While the CUDA kernel processing time is initially the largest, it doesn’t immediately scale with the number of input segments, so it starts being the exponentially faster implementation with more than $O(10^4)$ segments, where it starts taking advantage of the massive parallelization achievable by the GPU. The NVIDIA® Tesla® V100 GPU used for this study can run more than $10^5$ parallel threads. To give a figure of merit, a typical neutrino beam spill in the ND-LAr corresponds on average to $O(10^5)$ segments.

![Processing time graph](chart.png)

**Figure 2.** Processing time for the calculation of the recombination factor $R_{\text{Birks}}$ of eq. (3.1) with the GPU implementation using a CUDA kernel (blue), the CPU implementation using Numba (orange), and a sequential Python for loop (green). For reference, a neutrino beam spill corresponds on average to $O(10^5)$ energy deposition segments in the ND-LAr. This computation was performed on a node of the NERSC Cori supercomputer, which contains two sockets of 20-core Intel Xeon Gold 6148 (Skylake) at 2.40 GHz and 8 NVIDIA® Tesla® V100 (Volta) GPUs.

### 3.2 Electron transport in liquid argon

The electrons remaining after the recombination travel towards the anode at a constant velocity $v_{\text{drift}}$ (assuming perfect uniformity of the electric field), which is typically $O(1 \text{ mm/\mu s})$. The time they take to reach the anode is given by:

$$t_{\text{drift}} = \frac{z - z_{\text{anode}}}{v_{\text{drift}}},$$

where the electron drift direction is assumed to be along the $z$ axis and where $z_{\text{anode}}$ is the $z$ coordinate of the anode.

The impurities in the liquid argon, such as $O_2$, $N_2$ and $H_2O$, can attach a portion of the drifting electrons, so electrons farther from the anode will have a higher chance to be attached. This effect is usually parametrized by a negative exponential, so the charge $Q_a$ that effectively reaches the anode, assuming uniform impurities and a perfect electronics response, is:

$$Q_a = Q_R \cdot \exp(-t_{\text{drift}}/\tau),$$

where $Q_R$ is the total charge generated, $\tau$ is the characteristic time constant, and $t_{\text{drift}}$ is the drift time to the anode.
where $t_{\text{drift}}$ is the time the charge takes to reach the anode and $\tau$ is a parameter that depends on the concentration of impurities and it is usually called electron lifetime, which is in the order of milliseconds for concentrations of O$_2$ at tens of parts per trillion.

Electrons drifting in strong electric fields do not diffuse isotropically, so it is necessary to estimate both the longitudinal and transverse components with respect to the drift direction [24]. The diffusion length is given by:

$$\sigma = \sqrt{2D t_{\text{drift}}},$$

(3.5)

where $D$ is the longitudinal or transverse diffusion coefficient, which depends on the electric field and liquid argon temperature. In our simulation we set the longitudinal and transverse diffusion coefficients to $D_L = 4 \text{ cm}^2/\text{s}$ and $D_T = 8.8 \text{ cm}^2/\text{s}$, respectively. These values were obtained by a preliminary ProtoDUNE-SP [25] analysis.

This information is calculated and stored in appropriate columns of the NumPy array in a way analogous to the one described in section 3.1, where each thread processes a single segment independently.

### 3.3 Electronic signal induction on a pixel

#### 3.3.1 Field response

The current induced by a point charge on a given pixel within the anode is calculated using the Shockley-Ramo theorem [26]:

$$I_{\text{pixel}} = q \bar{v} \cdot \nabla W,$$

(3.6)

where $W$ is the weighting field potential, the normalized contribution of a single electrode to the overall field, $q$ is the particle charge, and $\bar{v}$ is the velocity of the particle. Within larnd-sim, the induced current is pre-calculated for a point-like charge and referenced as a $I_{\text{pixel}}[t,x,y]$ look-up table (LUT). The table contains the current induced by an electron placed at discrete $(x, y)$ locations on the anode plane at a discrete time $t$, where $t = 0$ is the time when the electron is at $z = 0.5 \text{ cm}$. This value is chosen because of the observed flatness of the electric and weighting fields at this distance.

The $I_{\text{pixel}}[t,x,y]$ values are calculated as follows. In the region very close to the anode, the electric field and the weighting field are calculated numerically. The geometry of a small volume, including a central pixel with a pitch of 4.4 mm and its 8 nearest neighbor pixels, is modeled using CAD software and converted into a 3-dimensional mesh using the Gmsh package [27]. The fields are then calculated using the successive over-relaxation method as implemented in the Elmer FEM [28] software package. The electric and weighting fields share the same geometry, but differ in the boundary conditions imposed on the problem. In the electric field calculation, the pixels are grounded, with the backing plane at a small offset voltage, and the field on the cathode-facing side of the volume is set to the nominal field of 500 V/cm. In the calculation of the weighting field, all electrodes other than the central pixel are grounded, the central pixel is set to unit voltage, and the voltage on the remaining conductors is set to zero. The results of these two field calculations are shown in figure 3.

Next, the idealized drift paths (ignoring diffusion and attenuation effects) are integrated using the ICARUS and Walkowiak [29, 30] electron transport models evaluated within the calculated electric field. This model allows for the drift velocity to change as a function of the local electric
(a) Drift electric potential field in the region near the pixel plane. Shown is a slice in $y$ along on a pixel center. The pixel surfaces are set to 0 kV, while the gradient of the far field at the surface of $z = 10 \text{ mm}$ is set to 0.5 kV/cm. The color scale corresponds to the value of the drift potential.

(b) Weighting potential field in the region near a “pixel of interest”. The central ($x = y = 0$) pixel is set to a unit potential, while all other surfaces are set to 0 potential. The resulting unitless field defines the susceptibility to current induction by charges moving nearby.

Figure 3. The drift (left) and weighting (right) field potentials obtained by finite-element analysis in the region near a generic pixel. Shown are slices of the potential fields along an $x - z$ plane which crosses through the centerline of a pixel.

field, as the assumption of a perfectly uniform field is not necessarily true very close to the anode. The drift paths are calculated for a grid of four hundred $(x, y)$ positions, 0.5 cm from the anode on the $z$ axis. The resulting granularity on the $x$ and $y$ axes is 0.33 mm. Finally, we compute the time-derivative of the weighting field potential along each path, which yields a charge-normalized current series for a given initial charge position. The sampling in time is 0.1 $\mu$s. A select few drift paths are shown in figure 4(a), with their corresponding current series in figure 4(b).

Farther from the pixel at $z > 0.5 \text{ cm}$, the weighting potential is solved by treating the pixel as point-like and using a method-of-image-charges approximation to fix the boundary conditions at the anode and the cathode. The drift velocity is assumed to be uniform. The relative normalization to the near-field calculation is then fixed by enforcing continuity in the current series at the time boundary. Some re-scaling is performed on these current series to ensure they integrate to one electron charge, as error is introduced in the drift path integration, as well as numerical error present in the FEM solutions and their interpolated values. Figure 5 shows the tabulated induced current on a pixel up to 30 $\mu$s from the time of arrival of the drift electron. The agreement of the near-field FEM model and the dipole far-field approximation across this surface seen in figure 5 demonstrates that the transition surface at 0.5 cm is sufficiently far from the pixel plane.

### 3.3.2 Induced current calculation

The electrons drifting towards the anode are diffused in the longitudinal and transverse directions, as described in section 3.2. The ionization electrons corresponding to a GEANT4 segment will then form a three-dimensional charge cloud. The induced current on a given pixel from this can be calculated by taking the convolution of the pre-calculated pixel response model (see section 3.3.1) and the charge density of the track segment:

$$I(t) = \int I_{\text{pixel}}(t - z/v_d, x, y)\rho(x, y, z)dxdydz$$  \hspace{1cm} (3.7)
Figure 4. To determine the near-field pixel response, we simulate the paths and velocities of drift electron within a 3D FEM field model. Here we show the simulated electron drift paths (left), and the corresponding induced current on the pixel (right) for a subset of the starting points used. The color of the trajectory on the left corresponds to the color of the current induced curve on the right.

Figure 5. Induced current on a 4.4 mm pixel from a drifting electron at a given drift time and x, y position relative to the pixel, in units of electrons per 0.1 µs. The different calculations used for the near-field and far-field regions are described in the text in section 3.3.1 with the near-field boundary highlighted with a dotted line.
where $\rho(x, y, z)$ is the 3-dimensional charge density including diffusion and $I_{\text{pixel}}(t - z/v_d, x, y)$ is the current response model at the given position and time tick, which is stored in the LUT described in section 3.3.1. The positively charged ions also produce a current on the pixel, however the magnitude of this current is more than 7 orders of magnitude smaller than the signal from the drift electrons due to the small drift velocity of the ions, and so it is neglected in the detector simulation.

The calculation of $I(t)$ is the most computational-intensive step of the detector simulation. In order to reduce its complexity, we apply two approximations to the charge density: first, we discretize the track segment along the track length into $N$ points; and second, we approximate the effect of diffusion summing the contribution from $M$ random 3D perturbations of the sample point. With these approximations, we can write the charge density as

$$
\rho(x, y, z) \approx \frac{Q_a}{NM} \sum_{\mu=1}^{M} \sum_{\nu=1}^{N} \delta(x_\mu + \delta x_{\mu\nu} - x) \delta(y_\mu + \delta y_{\mu\nu} - y) \delta(z_\mu + \delta z_{\mu\nu} - z)
$$

(3.8)

where $Q_a$ is the total charge magnitude of the segment from eq. (3.4), $\delta(x)$ is the Dirac delta function, $N$ is the number of sampled points along the track segment, $M$ is the number of Monte Carlo iterations per sample point, $(x_\mu, y_\mu, z_\mu)$ is a point uniformly distributed along the track segment, and $(\delta x_{\mu\nu}, \delta y_{\mu\nu}, \delta z_{\mu\nu})$ are normally distributed in proportion to the diffusion in that dimension. This simplifies the calculation to a sum over the pixel response model at each sampled position

$$
I(t) \approx \frac{Q_a}{NM} \sum_{\nu=1}^{N} \sum_{\mu=1}^{M} I_{\text{pixel}} \left( t - (z_\mu + \delta z_{\mu\nu}) / v_d, x_\mu + \delta x_{\mu\nu}, y_\mu + \delta y_{\mu\nu} \right).
$$

(3.9)

In our implementation, $N$ is calculated for each track with a 10 µm sampling interval, with the option of increasing the resolution and the Monte Carlo sampling for higher precision. For reference, at this sampling interval each sample point contributes a charge of roughly 50 electrons for a MIP-like track, and the error introduced by the approximations is at the percent level and much smaller than the noise level.

The calculation of the induced current is particularly well-suited to be implemented on a highly-parallelized architecture, since it is necessary to perform the sum of eq. (3.9) for each time tick in the time window, each pixel, and each Geant4 detector segment. The length of the time window depends on the orientation of the segment with respect to the anode, since the arrival time is directly proportional to the $z$ coordinate. In order to minimize the memory usage, the length of the longest signal and the largest number of pixels per segment are calculated before execution and used to allocate the array where the induced signals are going to be stored.

In our CUDA kernel $K_{\text{current}}(s_i, p_j, t_k)$, the parallel threads are organized in a three-dimensional grid, where the dimensions are the segment index $s_i$, the pixel index $p_j$, and the time tick $t_k$.

It is possible to compare the performances of this GPU algorithm with an equivalent CPU version, compiled with explicit parallelization of the loops using the Numba prange function. This function runs the loop in parallel threads, one for each core (20 in our case), and tries to merge adjacent loops together, reducing loop overhead [31]. The GPU version provides a speed-up of around four orders of magnitude for $O(10^3)$ simulated pixels, as shown in figure 6 for different values of the time sampling. Interestingly, in this case the GPU version is significantly faster than the CPU one even for only 10 pixels, since we are parallelizing also in two other dimensions (segment index and time). The processing time for the GPU implementation does not increase with the number of simulated pixels.
and with different values of the time sampling, which shows that the calculation is effectively being performed in parallel. Compilation times and memory allocation times are not taken into account here.

![Graph](image)

**Figure 6.** Processing time for the calculation of the induced current with the GPU implementation using a CUDA kernel (blue) and with a CPU implementation compiled with Numba with explicit parallelization of the loops (orange). For reference, a neutrino beam spill will induce a signal on around $5 \times 10^4$ pixels of the ND-LAr. The calculation has been performed with different values of the time sampling (0.05 $\mu$s, 0.1 $\mu$s, 0.2 $\mu$s) on a node of the NERSC Cori supercomputer, which contains two sockets of 20-core Intel Xeon Gold 6148 (Skylake) at 2.40GHz and 8 NVIDIA® Tesla® V100 (Volta) GPUs.

The result of the simulation of a cosmic-ray muon interacting in the TPC active volume is shown in figure 7(a). Given the effect of transverse diffusion, signals are induced also on the pixels that do not lie exactly on the $xy$ projection of the muon trajectory. The high granularity of the pixel layout shows clear imaging of features such as delta rays. Figure 7(b) shows the current induced on a pixel both by the cosmic-ray muon and by the subsequent delta ray.

### 3.4 Electronics response

In the LArPix system described in ref. [3] the pixel pads are uniquely instrumented by application-specific integrated circuits (ASICs), which provide a charge-sensitive amplifier (CSA) with self-triggered digitization. First, the signal from each pad is input to the CSA. Then, as signal accumulates on the pad, the voltage at the output of the CSA grows until it exceeds the discriminator threshold, which in our simulation is set to 28 mV, equivalent to the signal induced by $7 \times 10^3$ electron charges. The discriminator, in turn, triggers the output digitization through an 8-bit analog-to-digital converter (ADC). The ADC stores the CSA output voltage and then the CSA is reset, discarding the collected charge. The CSA is then ready to collect the subsequent signal. The minimum time between two consecutive ADC counts is 11 clock cycles, which with our 10MHz clock correspond to 1.1 $\mu$s. The time between the threshold crossing and the signal digitization is tunable in the LArPix ASIC. In practice, this is 1.6 $\mu$s, which is tuned to the typical signal size. While leakage current is negligible, sub-threshold charge could collect on the channel introducing a bias in the charge measurement. A periodic reset limits the impact of this effect.
Three kinds of noise are included in the simulation: (1) a reset noise, which corresponds to a random pedestal shift after every trigger reset, (2) a discriminator noise which affects the discriminator threshold, and (3) an uncorrelated noise. These are set in the simulation 900, 650, and 500 electron charges, respectively. These values have been tuned from the data described in section 6.

In our implementation, after summing the current induced by different tracks on the same pixel $p_i$, a CUDA kernel $K_{\text{electronics}}(p_i)$ simulates the trigger logic and the noise sources described above. The threads are organized in a one-dimensional grid, where each thread processes in parallel the current induced on a single pixel $p_i$. The CUDA kernel is exponentially faster than the CPU implementation and is more than three orders of magnitude faster with $O(10^3)$ pixels, as shown in figure 8.

Figure 9 shows a comparison of the integral of the induced current on each pixel and the corresponding ADC counts for a cosmic-ray muon. When the integral signal on a certain pixel does not reach the discrimination threshold no ADC count is stored, so it is possible to have pixels with a non-null induced current but no ADC count.

The resulting ADC counts and their corresponding timestamps are stored in a bi-dimensional array, where one dimension is the pixel index and the other is the ADC count index. They are then saved in the same HDF5 format used by the LArPix readout system, facilitating comparison between data and simulation.

4 Light simulation

4.1 Incident light calculation

The scintillation light produced by the excitation of the liquid argon provides additional calorimetric and topological information and a fast timing signal. Within larnd-sim, we include GPU algorithms to model scintillation light production, timing, and the associated electronics response.
Figure 8. Processing time for the simulation of the electronics response with the GPU implementation using a CUDA kernel (blue) and with a CPU implementation compiled with Numba with explicit parallelization of the loops (orange). For reference, a neutrino beam spill will induce a signal on around $5 \times 10^4$ pixels of the ND-LAr. The simulation has been performed on a node of the NERSC Cori supercomputer, which contains two sockets of 20-core Intel Xeon Gold 6148 (Skylake) at 2.40 GHz and 8 NVIDIA® Tesla® V100 (Volta) GPUs.

Figure 9. Bi-dimensional projection of a cosmic-ray muon interacting in the detector with the integral of the current induced on each pixel (left) and the sum of the ADC counts after the digitization stage (right).
Light intensity is calculated using a self-consistent recombination model. Namely, each track segment $dE/dx$ is converted to a total number of emitted photons $N_\gamma$ using the same recombination parameterization used for the $dQ/dx$:

$$\frac{dN_\gamma}{dx} = \left( W_{ph}^{-1} - \mathcal{R} W^{\text{iso}}_{\text{ion}}^{-1} \right) \frac{dE}{dx} \quad (4.1)$$

where $W_{ph}$ is the photon production per deposited energy at zero field, and $\mathcal{R}$ is from one of the recombination models described in section 3.2.

Light propagation and material quantum efficiencies are pre-tabulated using a dedicated Geant4 simulation into a LUT containing an average visibility $V_{i} [i_x, i_y, i_z]$ and relative time distribution $f_{i} [i_x, i_y, i_z, i_t]$ for each 3D voxel $(i_x, i_y, i_z)$ and each photosensor $i$ in the active volume [32]. For each track segment and photosensor, the total number of photoelectrons (p.e.) is calculated as

$$N_{\text{pe},ij} = \epsilon_i V_{i} [i_x, i_y, i_z] \left( \frac{dN_\gamma}{dx} \right)_j dx_j \quad (4.2)$$

where $\epsilon_i$ is the overall detection efficiency of photosensor $i$ and $dx_j$ is the length of the track segment $j$. To implement this, we use a two-dimensional grid across the track segment index and detector index, such that each thread calculates the number of photoelectrons observed on a light detector from its respective track segment. In testing, we observe a speed-up factor of about $56 \times$ by using the CUDA kernel for this task.

### 4.2 Photocurrent simulation

Once the total light signal is determined for each track segment, they are summed into a photocurrent time profile on each photosensor using the time distribution stored within the LUT:

$$I_{\text{pe},i} \left[ i_t \right] = \frac{\Delta t_{\text{LUT}}}{\delta t} \sum_j f_i \left[ i_x, i_y, i_z, i_t - t_0/\Delta t_{\text{LUT}} \right] N_{\text{pe},ij} \quad (4.3)$$

where $t_0$ is the segment deposition time, $\delta t$ is the simulation time tick, and $\Delta t_{\text{LUT}}$ is the time bin size of the LUT. The implementation of this calculation first determines the total number of ticks needed to fully simulate the event, with a configurable time pre- and post-buffer, and then pre-allocates the photocurrent time profile $I_{\text{pe},i} \left[ i_t \right]$ array on the GPU. A CUDA kernel then iterates over the incident light array $N_{\text{pe},ij}$ summing the contribution at each time tick. To maintain independence between threads, we use a two-dimensional grid on the dimensions of the output time profile array. Each thread is responsible for calculating the sum over all track segments for a single photodetector and time tick. We observe a factor of 45 improvement by using the GPU kernel for this calculation at scale. However below approximately 1000 channels $\times$ 1000 ticks the parallel CPU calculation becomes faster due to the copying of static configuration data to the GPU, see figure 10(a).

After the time profile has been created, the simulation calculates the smearing effect due to the scintillation light emission. LAr scintillation light consists of two components: a prompt component with a characteristic decay time of $\sim 7$ ns; and a slow component produced with a characteristic decay time of $\sim 1600$ ns [33, 34]. These values, which are assumed to be constant in our simulation, depend on the concentration of non-argon impurities in the liquid [35, 36] and on the electric
Figure 10. A summary of the runtime of the CUDA-compiled light simulation kernels described in the text using 100 identical track segments in the center of a TPC. The three most computationally intensive steps of the light simulation are compared to an equivalent parallelized but CPU-based calculation. For the light scintillation calculation and the electronics response calculation, the run time does not increase with more track segments. For reference, the drift time window of ND-LAr will be around $3 \times 10^3$ time ticks.

A two-component exponential model for the scintillation time profile

$$f_{\text{scint}}(t) = \frac{f_p}{\tau_p} e^{-t/\tau_p} + \frac{1 - f_p}{\tau_s} e^{-t/\tau_s}$$  \hspace{1cm} (4.4)$$

is used to simulate the broadening of the light signals due to these processes, where $f_p$ is the fraction of total light arising from the prompt component ($\sim 0.3$ [2]), and $\tau_p$ and $\tau_s$ are the prompt and slow decay times, respectively.

The convolution is direct and truncated:

$$I_{\text{scint},i}[i_t] = \sum_{j_t = \tilde{i}_t, \text{min}}^{i_t} I_{\text{pe},i}[j_t] \ast f_{\text{scint}}((i_t - j_t) \delta t),$$  \hspace{1cm} (4.5)$$

where $\tilde{i}_t, \text{min} = i_t - 5 \frac{\delta t}{\tau_s}$. The convolution kernel is truncated at $5 \times$ the scintillation slow component lifetime. Within this interval, > 99% of the total light from an energy deposit will have been emitted, thus minimally impacting the accuracy of the simulation. However, by using a truncated convolution, we improve the scaling of the simulation from $O(N^2)$ to $O(N \times N_{\text{trunc}})$, where $N_{\text{trunc}}$ is
the number of truncated time ticks. This convolution is implemented in a CUDA kernel and threads are distributed in a two-dimensional grid across the light detector index $i$ and the time tick index $i_t$, such that each thread calculates the weighted sum across the input array $I_{pe,i}$. The direct convolution of the scintillation light profile along with the electronics response convolution, discussed later, are the most computationally intense stages of the light simulation. Figure 10(b) and figure 10(c) show the performance of these kernels with the number of time ticks simulated and the number of light channels. Within the event sizes tested, regardless of the number of time ticks and light channels, a significant speed-up is realized by moving this calculation to the GPU.

Fluctuations in the number of photoelectrons produced by the incident light due to counting statistics is included for each time profile bin. An inverse transform sampling Poisson random number algorithm [38] is used to generate the number of observed photoelectrons per time tick. The average number of steps required by this algorithm scales in proportion to the mean of the distribution. To improve the runtime for large signals, at greater than 30 expected photoelectrons, a normal distribution $\mathcal{N}(I_{scint,i}[i_t] \Delta t, \sqrt{I_{scint,i}[i_t] \Delta t})$ is used. The choice of 30 photoelectrons is selected to limit the error from this approximation to less than 1%.

4.3 Electronics response

The electronics response of the light detector readout is simulated assuming a perfectly linear ADC, a user-specified unit-normalized impulse response model, and purely uncorrelated noise:

$$V_i[i_t] = n_i[i_t] + \sum_{j_t=0}^{N_{resp}} G_i I_{scint,i}[j_t] I_{resp}[i_t - j_t] \frac{\Delta I_{resp}}{\Delta t}$$

(4.6)

where $n_i[i_t]$ is the simulated noise, $i_{resp}[i_t]$ is the unit-normalized response model, $N_{resp}$ is the number of bins in the response model, and $G_i$ is the individual photosensor gain with units of [ADC time / PE]. Noise is generated individually on each light detector by using a user-specified fast-fourier transform of the noise distribution to generate random phase sinusoids with appropriate amplitudes. Figure 11 demonstrates the simulation sequence on an example muon decay event with non-trivial time structure.

Each time profile is segmented into individual light triggers using a direct threshold on the sum of multiple light detectors. A serial algorithm loops over each light detector group and identifies each threshold crossing, accounting for dead-time between triggers. Because the timing of later triggers depends on the timing of earlier triggers, this algorithm cannot be parallelized. However, the trigger simulation only contributes $\sim 3\%$ to the overall simulation runtime and scales with the number of time ticks simulated, thus the use of a serial algorithm for the trigger generation is not a significant contributor to the overall simulation time at the DUNE Near Detector regime of $< 100$ triggers per spill per light detector group.

The final ADC waveform is then linearly interpolated into the desired sampling frequency and ADC resolution. Trigger information and waveforms are saved into dedicated datasets within the same output HDF5 file as the charge simulation.

4.4 Truth propagation

True particle energy deposition information can be propagated through this chain to provide the true track segment id and incident photoelectrons contributing to each ADC sample. However for
Figure 11. Example simulated muon decay and subsequent light simulation sequence. The region highlighted in figure 11(a) corresponds to the location of the light detector simulated in figure 11(b) and 11(c).

Typical light detectors with $O(\text{cm})$ spatial resolution and relatively broad response times $O(\mu\text{s})$, most energy deposits originating from an interaction overlap. Thus, the number of truth entries needed scales as $(dN_{\text{segments,typ}}/dV) \times (N_{\text{samples}}) \times (N_{\text{photosensors}}) \times (\text{visible volume})$. Using typical numbers of $O(500/\text{event m}^3)$, $O(100)$, $O(100)$, and $O(10 \text{ m}^3)$, respectively, tracking the true track segment id and intensity at each sample necessitates $O(5 \times 10^7)$ truth entries, using a minimum of $O(400 \text{ MB/event})$ assuming each record consists of a 32-bit integer and 32-bit floating point number. Since the truth information for an event is highly correlated in time, we opt to by default only preserve the true number of photoelectrons and arrival time per track segment per light detector, reducing
the required space by a factor of $N_{\text{samples}}$ and runtime by a factor of $N_{\text{segments}}/\text{tick}$. We however maintain the option to track the full truth information as needed by setting a configuration flag.

5 Profiling

The larnd-sim code has been profiled using the NVIDIA Nsight™ Systems and the NVIDIA Nsight™ Compute tools v2021.3.0 with an input dataset of 250 simulated cosmic rays. The analysis shows that the algorithm is well optimized for GPUs. The charge simulation is responsible for around 70% of the processing time. Of this 70%, 97.5% of the time is used by the induced current CUDA kernel described in section 3.3.2. This kernel, in turn, spends 99.9% of the processing time for computing operations. The time spent for memory transfers, including also the initial time to allocate the dataset on the GPU memory, can be considered negligible.

A Roofline [39] analysis performed on the induced current CUDA kernel proves that the algorithm is compute-bound, so the time needed to complete its task is determined principally by the computing speed. Figure 12 shows that kernel performance for 10 cosmic-ray events with a total of 5013 active pixels, which is a typical amount of data processed for each GPU iteration, is close to the peak FLOP/s limit (in orange) and far from the memory bandwidth limit (in blue) of the GPU being used.

Overall, for a 0.5 m$^3$ LArTPC operated with a surface-level cosmic ray rate, the inclusion of the light simulation increases the runtime, the memory usage, and the overall data volume of the output by 30%, 15%, and 20%, respectively.

![Figure 12](image.png)

**Figure 12.** Roofline plot for the induced current CUDA kernel (black square) on a GPU node of the NERSC Cori supercomputer, equipped with 8 NVIDIA® Tesla® V100 (Volta) GPUs. The kernel performances are far from the memory-bound region (in light blue) and close to the peak FLOP/s limit (in orange) of the GPU being used.
6 Simulation of a cosmic-ray sample and data comparison

A single-phase LArTPC called the ArgonCube Module-0 Demonstrator was operated in spring 2021 at the University of Bern as a prototype for the ND-LAr. The detector is divided into two functionally identical TPC drift regions, sharing a central high-voltage cathode that provides the drift electric field. Each anode is equipped with an array of $2 \times 4$ LArPix tiles of size $31.038 \text{ cm}^2 \times 31.038 \text{ cm}^2$. A total of 4900 pixel pads of size $0.4434 \text{ cm}^2 \times 0.4434 \text{ cm}^2$ are etched on the side of each tile facing the active volume. Each tile is instrumented by 100 ASICs placed on the opposite side of the board. The LArTPC drift length is 30.27 cm.

A dataset of cosmic-ray events was acquired with an applied electric field of $0.50 \text{kV/cm}$. The detector performances are analyzed in detail in ref. [40].

Here we will compare the cosmic-ray data acquired with this experimental setup to a sample generated by the larnd-sim software. As a first step a dataset of cosmic rays is produced using the CORSIKA cosmic-ray generator v7.7410 [41], using FLUKA 2011 for the low-energy hadronic interaction [42]. Then, the passage of the cosmic rays through the Module-0 geometry is simulated using edep-sim, a C++ wrapper around G/e.pc/a.pc/n.pc/t.pc4. The output of edep-sim is passed to larnd-sim, which produces a HDF5 sample in the same format of the Module-0 data.

Both the simulation and the data samples are passed to a reconstruction script, which takes as input an array of three-dimensional hits, one for each ADC count. The $xy$ coordinates are given by the position of the pixel center on the tile, and the $z$ coordinate is given by the corresponding clock tick, using the inverse formula of eq. (3.3). The hits are clustered together using the DBSCAN algorithm [43]. Hits grouped in the same cluster are fed to a principal component analysis (PCA), which returns three-dimensional segments that we define as reconstructed tracks. Figure 13 shows an example of a simulated cosmic ray crossing the cathode plane after the reconstruction stage, with the reconstructed track and the corresponding hits.

To improve the system self-trigger stability, 7.8% of the pixels in Module-0 were disabled due to an understood grounding issue, mostly near the tiles edges. The same pixels were also disabled in the simulation. Thus, reconstructed tracks may show artificial gaps due to the presence of disabled channels. Also, cathode-piercing tracks can be reconstructed as separated tracks, due to the non-null cathode thickness. In order to reconstruct the original cosmic-ray trajectories, we iteratively stitch together pairs of reconstructed tracks, based on their relative directions and positions at the cathode. To set the threshold for these metrics, a subsample of data events are reconstructed a second time after randomly translating and masking hits that fall into the disabled regions. We require that 95% of the gaps produced in this procedure are correctly identified and resolved.

Module-0 was operated principally in two channel threshold regimes, low and high, which correspond to thresholds of around 6000 and 12000 $e^-$, respectively. Here we compare the low-threshold data to the simulated sample, which was produced with a discrimination threshold roughly equivalent to the low regime one.

The amount of charge deposited per unit length ($dQ/dx$) is measured both in data and simulation using the reconstructed tracks, which are required to have at least 10 associated hits. In the simulation, the signal amplitude of the CSA is calibrated using a fixed gain of 250 $e^-/\text{mV}$. The value of the gain in the data is obtained by performing a template fit of the $dQ/dx$ distribution with the simulated one, shown in figure 14. Here, the $dQ$ is the sum of the hits charge associated to the reconstructed track.
Figure 13. The ADC counts are transformed into three-dimensional hits by a reconstruction script. The hits are grouped in clusters and fed to a principal component analysis, which returns a reconstructed track. This plot shows, clockwise from top left, the hits for a simulated cosmic ray on the $z\psi$ and $x\psi$ planes, a three-dimensional event display, and the hits charge as a function of time.

Figure 14. $dQ/dx$ measured in data and simulation for reconstructed tracks with at least 10 associated hits. The distributions have been normalized to unity.
and \( dx \) is the length of the reconstructed track. The best-fit value of \( 219 \text{ e}^- \text{mV} \) is compatible with the value of \( 221 \text{ e}^- \text{mV} \), obtained with a dedicated stand-alone test of the LArPix ASIC CSA.

Reconstructed tracks can be sub-divided into segments of variable length, from 10 to 400 mm. The deposited charge \( dQ \) is the sum of the hit charges belonging to the same segment and \( dx \) is the length of the segment. The distributions have been fitted with a Gaussian-convoluted Moyal function [44].

Figure 15 shows the \( dQ/dx \) data and simulated distributions, as well as their respective fits. The \( dQ/dx \) distributions, normalized to the number of tracks, are in good agreement with data and simulation for all segment lengths.

**Figure 15.** \( dQ/dx \) measured for segments of different lengths for low threshold runs (white dots) and a sample of simulated cosmic rays (red line). The distributions have been normalized to unity and fitted with a Gaussian-convoluted Moyal function for data (solid blue line) and simulation (dashed green line). The legend shows the most probable value (MPV) for each fit.
7 Conclusions

We show that it is possible to implement the simulation of a pixelated LArTPC using highly-parallelized GPU algorithms. The algorithms are written in Python and they are compiled on the GPU using the Numba just-in-time compiler.

The software provides an end-to-end microphysical simulation for light readout and pixelated charge readout. Table 1 shows a summary of the speed-up achieved by the GPU implementation for various steps of the simulation.

Table 1. Summary of speed-up factors achieved by the GPU implementation for various simulation steps.

<table>
<thead>
<tr>
<th>Calculation</th>
<th>Loop over</th>
<th>Quantity</th>
<th>GPU speed-up factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recombination factor</td>
<td>Segments</td>
<td>$10^5$</td>
<td>$\times 3$</td>
</tr>
<tr>
<td>Induced current</td>
<td>Pixels</td>
<td>$10^3$</td>
<td>$\times 7314$</td>
</tr>
<tr>
<td>Charge electronics response</td>
<td>Pixels</td>
<td>$10^3$</td>
<td>$\times 985$</td>
</tr>
<tr>
<td>Light time profile</td>
<td>Time ticks</td>
<td>$10^5$</td>
<td>$\times 228$</td>
</tr>
<tr>
<td>Scintillation profile</td>
<td>Time ticks</td>
<td>$10^3$</td>
<td>$\times 568$</td>
</tr>
<tr>
<td>Light electronics response</td>
<td>Time ticks</td>
<td>$10^3$</td>
<td>$\times 1883$</td>
</tr>
</tbody>
</table>

In general, trivial algorithms (such as the calculation of the recombination factor) require a large amount of input data before the GPU implementation starts being faster. This is expected, since there are launch and execution overheads associated with CUDA kernels. For more complex algorithms, such as the simulation of the charge readout, the GPU generally starts being many orders of magnitude faster than the CPU with only a few tens of simulated pixels in the detector. In particular, the calculation of the induced current, which is the most computationally expensive task, is around four orders of magnitude faster for $O(10^3)$ pixels. Likewise, the simulation of convolution-heavy light signals is faster on all event sizes tested, typically achieving three orders of magnitude improvement. With the acceleration achieved, we are able to simulate the charge (light) signals on $O(10^6)$ ($O(10^3)$) individual channels with sub-microsecond (nanosecond) resolution and in timescales reasonable for large Monte Carlo simulations.

As an example, the full simulation of $10^4$ cosmic rays in a ton-scale LArTPC takes in total approximately 450 s, 100 s for the simulation of the passage of the initial particles through matter with edep-sim, and the remaining 350 s for the detector simulation with larnd-sim. Highly-parallelized simulation algorithms such as the one described in this document could be adapted to speed up the simulation of the DUNE Far Detector as well, which will have $O(10^6)$ readout channels [45].

We have also performed the comparison of a simulated sample of cosmic rays with the data acquired by a prototype LArTPC equipped with a large-scale pixelated readout system, called the ArgonCube Module-0 Demonstrator. Charge distributions in data and simulation are in general good agreement.

The Module-0 LArTPC was moved from Bern to Fermilab in October 2021, where it will be tested on the NuMI neutrino beamline with three other identical modules, currently under construction [40], forming the ND-LAr 2x2 prototype. The production of a dataset of accurately simulated neutrino interactions will be of fundamental importance for the analysis of this measurement. larnd-sim will also be used to simulate the detector response of the full ND-LAr detector and has been included...
into the DUNE Offline Computing Conceptual Design Report [46]. The DUNE collaboration is already using larnd-sim on the NERSC Perlmutter supercomputer, which shares a similar GPU node architecture with Cori, to perform ND-LAr 2x2 prototype simulations. A workflow chaining together edep-sim, larnd-sim and a machine learning-based reconstruction stage is currently under development.

The choice of a high-level and largely supported programming language such as Python can enable the implementation of a fully differentiable simulator in the near future, by exploiting libraries commonly used in the machine learning and artificial intelligence fields (TensorFlow [47], JAX [48]). A differentiable model will be able to use a gradient-based optimization, such as gradient descent, to automatically infer the detector simulation input and the detector physics model parameters. This technique has already been actively explored for the simulation of the photon propagation in a LArTPC [49].

Acknowledgments

This document was prepared by the DUNE collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

This work was supported by CNPq, FAPERJ, FAPEG and FAPESP, Brazil; CFI, IPP and NSERC, Canada; CERN; MSMT, Czech Republic; ERDF, H2020-EU and MSCA, European Union; CNRS/IN2P3 and CEA, France; INFN, Italy; FCT, Portugal; NRF, South Korea; CAM, Fundación “La Caixa”, Junta de Andalucía-FEDER, MICINN, and Xunta de Galicia, Spain; SERI and SNSF, Switzerland; TÜBİTAK, Turkey; The Royal Society and UKRI/STFC, United Kingdom; DOE and NSF, United States of America. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

We are also thankful to Max Katz from NVIDIA® and Daniel Margala from NERSC for the precious suggestions.

References


The DUNE collaboration


1 Corresponding author

1 Abilene Christian University, Abilene, TX 79601, U.S.A.
2 University of Albany, SUNY, Albany, NY 12222, U.S.A.
3 University of Amsterdam, NL-1098 XG Amsterdam, The Netherlands
4 Antalya Bilim University, 07190 Dizemba/Ankara, Turkey
5 University of Antananarivo, Antananarivo 101, Madagascar
6 Universidad Antonio Nariño, Bogotá, Colombia
7 Argonne National Laboratory, Argonne, IL 60439, U.S.A.
8 University of Arizona, Tucson, AZ 85721, U.S.A.
9 Universidad Nacional de Asunció, San Lorenzo, Paraguay
10 University of Athens, Zografou GR 157 84, Greece
11 Universidad del Atlántico, Barranquilla, Atlántico, Colombia
12 Augustana University, Sioux Falls, SD 57197, U.S.A.
13 University of Basel, CH-4056 Basel, Switzerland
14 University of Bern, CH-3012 Bern, Switzerland
15 Beykent University, Istanbul, Turkey
16 University of Birmingham, Birmingham B15 2TT, United Kingdom
17 Università del Bologna, 40127 Bologna, Italy
18 Boston University, Boston, MA 02215, U.S.A.
19 University of Bristol, Bristol BS8 1TL, United Kingdom
20 Brookhaven National Laboratory, Upton, NY 11973, U.S.A.
21 University of Bucharest, Bucharest, Romania
22 University of California Berkeley, Berkeley, CA 94720, U.S.A.
23 University of California Davis, Davis, CA 95616, U.S.A.
24 University of California Irvine, Irvine, CA 92697, U.S.A.
25 University of California Los Angeles, Los Angeles, CA 90095, U.S.A.
26 University of California Riverside, Riverside CA 92521, U.S.A.
27 University of California Santa Barbara, Santa Barbara, California 93106 U.S.A.
28 California Institute of Technology, Pasadena, CA 91125, U.S.A.
29 University of Cambridge, Cambridge CB3 0HE, United Kingdom
Universidade Estadual de Campinas, Campinas — SP, 13083-970, Brazil
Università di Catania, 2 — 95131 Catania, Italy
Universidad Católica del Norte, Antofagasta, Chile
Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, RJ 22290-180, Brazil
IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
CERN, The European Organization for Nuclear Research, 1211 Meyrin, Switzerland
Institute of Particle and Nuclear Physics of the Faculty of Mathematics and Physics of the Charles University, 180 00 Prague 8, Czech Republic
University of Chicago, Chicago, IL 60637, U.S.A.
Chung-Ang University, Seoul 06974, South Korea
CIEMAT, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, E-28040 Madrid, Spain
University of Cincinnati, Cincinnati, OH 45221, U.S.A.
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav), Mexico City, Mexico
Universidad de Colima, Colima, Mexico
University of Colorado Boulder, Boulder, CO 80309, U.S.A.
Colorado State University, Fort Collins, CO 80523, U.S.A.
Columbia University, New York, NY 10027, U.S.A.
Comisión Nacional de Investigación y Desarrollo Aeroespacial, Lima, Peru
Centro de Tecnologia da Informacao Renato Archer, Campinas, SP — CEP 13069-901, Brazil
Central University of South Bihar, Gaya, 824236, India
Institute of Physics, Czech Academy of Sciences, 182 00 Prague 8, Czech Republic
Czech Technical University, 115 19 Prague 1, Czech Republic
Dakota State University, Madison, SD 57042, U.S.A.
University of Dallas, Irving, TX 75062-4736, U.S.A.
Laboratoire d’Annecy de Physique des Particules, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LAPP-IN2P3, 74000 Annecy, France
Daresbury Laboratory, Cheshire WA4 4AD, United Kingdom
Dordt University, 700 7th St NE, Sioux Center, IA 51250, U.S.A.
Drexel University, Philadelphia, PA 19104, U.S.A.
Duke University, Durham, NC 27708, U.S.A.
Durham University, Durham DH1 3LE, United Kingdom
University of Edinburgh, Edinburgh EH8 9YL, United Kingdom
Universidad EIA, Envigado, Antioquia, Colombia
Eötvös Loránd University, 1053 Budapest, Hungary
ETH Zurich, Zurich, Switzerland
Faculdade de Ciências da Universidade de Lisboa — FCUL, 1749-016 Lisboa, Portugal
Universidade Federal de Alfas, Poços de Caldas — MG, 37715-400, Brazil
Universidade Federal de Goias, Goiania, GO 74690-900, Brazil
Universidade Federal do ABC, Santo André — SP, 09210-580, Brazil
Universidade Federal do Rio de Janeiro, Rio de Janeiro — RJ, 21941-901, Brazil
Fermi National Accelerator Laboratory, Batavia, IL 60510, U.S.A.
University of Ferrara, Ferrara, Italy
University of Florida, Gainesville, FL 32611-8440, U.S.A.
Florida State University, Tallahassee, FL, U.S.A.
Fluminense Federal University, 9 Icarai Niterói — RJ, 24220-900, Brazil
Università degli Studi di Genova, Genova, Italy
Georgian Technical University, Tbilisi, Georgia
University of Granada & CAFPE, 18002 Granada, Spain
Gran Sasso Science Institute, L’Aquila, Italy
Laboratori Nazionali del Gran Sasso, L’Aquila AQ, Italy
University Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, 38000 Grenoble, France