

Low-ordered Orthogonal Voxel Finite Element with INT8 Tensor Cores for GPU-based Explicit Elastic Wave Propagation Analysis

Tsuyoshi Ichimura¹, Kohei Fujita¹, Muneo Hori², and Maddegedara Lalith¹

¹ Earthquake Research Institute and Department of Civil Engineering, The University of Tokyo, Japan

² Research Institute for Value-Added-Information Generation, Japan Agency for Marine-Earth Science and Technology, Japan

Abstract. Faster explicit elastic wavefield simulations are required for large and complex three-dimensional media using a structured finite element method. Such wavefield simulations are suitable for GPUs, which have exhibited improved computational performance in recent years, and the use of GPUs is expected to speed up such simulations. However, available computational performance on GPUs is typically not fully exploited, and the conventional method involves some numerical dispersion. Thus, in this paper, we propose an explicit structured-mesh wavefield simulation method that uses INT8 Tensor Cores and reduces numerical dispersion to speed up computation on GPUs. The proposed method was implemented for GPUs, and its performance was evaluated in a simulation experiment of a real-world problem. The results demonstrate that the proposed method is 17.0 times faster than the conventional method.

Keywords: Explicit Wave Propagation Analysis · Finite Element · Tensor Core · GPU.

1 Introduction

Explicit simulation of elastic wavefields, which can be computed without solving matrix equations, is often used to evaluate the dynamic response of large and complex three-dimensional media through sequential analysis and to estimate internal structures through optimization via many simulations. Thus, further speedup is desired for conducting such large and many-case analyses. In this paper, we focus on increasing the speed of explicit elastic wavefield simulations with structured finite elements, which are suitable for generating numerical models. In the standard finite element method, the generation of the finite element model can be a bottleneck in the analysis process. However, finite element models can be generated automatically for the structured finite element method by using cubic elements that are finer than the elements used in standard finite element analysis and allowing some approximation of geometry, thereby making it suitable for analyzing large and complex media. Thus, it is used in seismic analysis [1, 2] and the estimation of the internal structure of structures [3]. There

is a strong need for a method that can be executed many times in a short period because detailed simulations and model optimizations are desired in such analyses.

Explicit wavefield simulations using structured finite element methods, where continuous memory access is dominant, are well-suited for processing on GPUs and are expected to be executed quickly on recent GPUs with improved computational performance. However, the lumped mass matrix approximation is applied to the low-order structured finite elements that are conventionally used in such analyses, and this results in numerical dispersion, which increases computational cost due to the need to use smaller element sizes. In addition, some GPUs are equipped with acceleration mechanisms, e.g., Tensor Cores [4], and their use in deep learning and other applications is progressing. However, the use of Tensor Cores in physics-based simulations is limited because their high-speed operations are predicated on transforming operations into a somewhat special form, and they are only used in some cases (e.g., to obtain a coarse solution in implicit simulations [5, 6]. See [7] for a review of mixed-precision algorithms for linear numerical algebra in general). In particular, unlike the implicit solution method, the explicit solution method does not allow for processes that refine the coarse solution. Therefore, the development of such algorithms is challenging because the results must be equivalent to those of ordinary FP64 calculations while effectively using Tensor Cores. However, Tensor Cores have high computational performance; thus, further increases in processing speed can be expected if the hardware can be exploited effectively.

Therefore, this paper proposes an explicit structured finite element elastic wavefield simulation method that uses INT8 Tensor Cores, which is more accurate and faster than the conventional method. The remainder of this paper is organized as follows. In Section 2, we present an explicit structured finite element wavefield simulation method that solves the problems of conventional methods and is suitable for INT8 Tensor Core computation. Section 3 describes the implementation of the proposed method on a GPU and presents the details of the performance of the proposed method using a realistic analysis model. We show that the proposed method can realize an analysis that yields results that are equivalent to those of the conventional method at higher speed. Finally, the paper is concluded in Section 4.

2 Low-ordered Orthogonal Voxel Finite Element with INT8 Tensor Cores

In this study, we target the governing equation of a linear dynamic elastic body:

$$\rho \ddot{\mathbf{u}} - (\nabla \cdot \mathbf{c} \cdot \nabla) \cdot \mathbf{u} = \mathbf{f}, \quad (1)$$

where ρ , \mathbf{u} , \mathbf{c} , and \mathbf{f} denote the density, displacement, elasticity tensor, and body force, respectively. In addition, (\cdot) and ∇ denote the temporal and spatial differential operators, respectively. In standard finite element analysis, the target

domain is decomposed into small elements e , and the basis functions ϕ^α are defined on these elements. By substituting $\mathbf{u} = \sum \mathbf{u}^\alpha \phi^\alpha$ into the functional in Eq. (1) and taking the stationary condition, the following linear equation is obtained for the unknown coefficients \mathbf{u}^α :

$$\mathbf{K}\mathbf{u} + \mathbf{M}\ddot{\mathbf{u}} = \mathbf{F}. \quad (2)$$

Here, \mathbf{u} is a vector assembled using \mathbf{u}^α . \mathbf{K} and \mathbf{M} are a stiffness matrix and mass matrix assembled using the element stiffness matrix \mathbf{K}_e and the element mass matrix \mathbf{M}_e obtained for each element, respectively.

In a typical structured finite element analysis, the domain of interest is divided into cubic subdomains of lengths ds per side, and these are used as elements (hereafter referred to as voxel elements). As a result, a finite element model can be generated easily by simply covering a three-dimensional region with a structured grid of constant width ds and evaluating the physical properties of each element. Thus, it is frequently used for the static and dynamic analysis of regions with complex geometry and property distributions or for geometry optimization. However, the object is modeled with cubes of uniform size; thus, it is necessary to reduce ds to improve the accuracy of modeling geometry, which frequently results in a problem with large degrees of freedom. When using such voxel elements, low-order elements are frequently used, i.e., it is common to use basis functions

$$\phi^\beta = -\frac{1}{8}(r_1 + \bar{r}_1)(r_2 + \bar{r}_2)(r_3 + \bar{r}_3) \quad (\beta = 1, 2, \dots, 8)$$

for node β , which is often used for hexahedral elements (Figure 1 shows the definition of the local coordinates, local node numbering, and the definition of $\bar{r}_1, \bar{r}_2, \bar{r}_3$). However, in this case, the element mass matrix

$$\mathbf{M}_e = \rho(\phi^\beta \phi^{\beta'})_e,$$

becomes nondiagonal, which necessitates solving a large-scale matrix equation to obtain \mathbf{u} in Eq. (2). Note that $(\)_e$ indicates volume integration in each element. To solve this problem, an approximation that concentrates the off-diagonal terms to the diagonal terms (lumped mass matrix approximation) is frequently applied, which leads to a diagonal mass matrix \mathbf{M}_e^{ap} , where the off-diagonal terms are approximated as 0, and the diagonal terms are approximated as $\rho/8(1)_e$. Consequently, for example, if we approximate the acceleration using the central difference method with a time-step width of dt , Eq. (2) at time step it becomes

$$\mathbf{K}\mathbf{u}^{it} + \mathbf{M}^{ap}(\mathbf{u}^{it+1} - 2\mathbf{u}^{it} + \mathbf{u}^{it-1})/dt^2 = \mathbf{F}^{it}, \quad (3)$$

which enables explicit computation of \mathbf{u}^{it+1} . This lumped matrix approximation is widely used when solving dynamic problems because eliminating the need to solve large matrix equations reduces the analysis cost significantly. This analysis method is referred to as VFEM in this paper.

In VFEM, the lumped mass matrix approximation causes numerical dispersion, which is particularly noticeable when the distance of the propagation

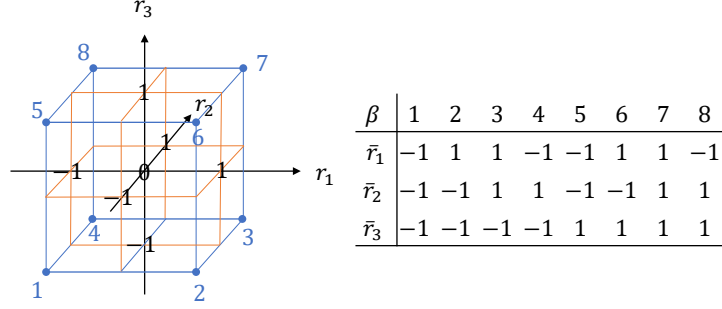


Fig. 1: Definition of a voxel element. For a cube of size ds in each direction, a local coordinate $r_1 r_2 r_3$ is defined ($-1 \leq r_1 \leq 1$, $-1 \leq r_2 \leq 1$, $-1 \leq r_3 \leq 1$). The local node number ranging from 1–8 is defined on each node in the element. ϕ^β is the basis function for node β .

of waves increases. To suppress such numerical dispersion, a VFEM using an orthogonal basis function has been proposed for the voxel elements [8]. Using the orthogonal basis functions, the mass matrix becomes diagonal without any approximation to the voxel elements. We formulate OVFEM using the Hellinger-Reissner functional:

$$J(\mathbf{u}, \sigma) = \int_V \int_T -\frac{1}{2} \rho \dot{\mathbf{u}} \cdot \dot{\mathbf{u}} - \frac{1}{2} \sigma : \mathbf{c}^{-1} : \sigma + \sigma : (\nabla \otimes \mathbf{u}) - \mathbf{u} \cdot \mathbf{f} \, dv dt. \quad (4)$$

Here, σ denotes stress, which can be expressed as $\sigma = \sum \sigma^\beta \psi^\beta$ using the basis function ψ^β and unknown coefficients σ^β . By taking the stationary condition of Eq. (4) for \mathbf{u}^α and σ^β , a discretized form of the governing equation can be obtained in the form of Eq. (2). Here, \mathbf{K} and \mathbf{M} become matrices assembled using the element stiffness matrix

$$\mathbf{K}_e^o = (\psi^\beta \nabla \phi^\alpha)_e^t \cdot (\psi^\beta \psi^{\beta'})_e^{-1} \mathbf{c} \cdot (\psi^\beta \nabla \phi^\alpha)_e, \quad (5)$$

and the element mass matrix

$$\mathbf{M}_e^o = \rho (\phi^\beta \phi^{\beta'})_e, \quad (6)$$

respectively. While the same local coordinates $-1 \leq r_1 \leq 1$, $-1 \leq r_2 \leq 1$, $-1 \leq r_3 \leq 1$ and node placement of VFEM (Fig. 1) are used, the following Heaviside function H is used in OVFEM:

$$\phi^\beta = H(\bar{r}_1 r_1) H(\bar{r}_2 r_2) H(\bar{r}_3 r_3) \quad (\beta = 1, 2, \dots, 8), \quad (7)$$

$$\psi^1 = 1, \psi^2 = r_1, \psi^3 = r_2, \psi^4 = r_3, \psi^5 = r_1 r_2, \psi^6 = r_2 r_3, \psi^7 = r_1 r_3. \quad (8)$$

As shown in the definition of Eq. (7), the element mass matrix of Eq. (6) becomes diagonal, and a stable solution is obtained because the discontinuous functions H are used together with smooth polynomial functions. Thus, the use of \mathbf{M}_e^o

allows us to solve Eq. (2) explicitly without approximation. OVFEM is expected to be more accurate than VFEM due to the suppression of numerical dispersion by the absence of the lumped mass approximation. This indicates that OVFEM may enable a coarser discretization size compared to VFEM, which can reduce the analysis costs of three-dimensional dynamic problems considerably.

Some GPUs are equipped with acceleration mechanisms that can perform certain forms of matrix products at high speed. For example, the A100 80 GB PCIe GPU [9] has Tensor Cores with peak performance of 624 TOPS for 8-bit integer matrix operations, which is significantly higher than the 9.7 TFLOPS peak performance of its FP64 cores. Therefore, effective use of Tensor Cores is expected to speed up analysis, and Tensor Cores are being used in deep learning and other applications. However, to apply Tensor Core operations efficiently, it is necessary to formulate an algorithm such that the amount of matrix operations of a particular form is dominant while guaranteeing the accuracy of the operations, which limits the applicability to only a few physics-based simulations. In particular, with the implicit method, Tensor Cores can be used as a substitute for part of the analysis by using Tensor Cores to obtain a coarse solution and then refining it to reduce the analysis cost; however, it is difficult to apply such a refining process to the explicit method. Thus, the part of the algorithm to be computed by Tensor Cores must demonstrate accuracy that is equivalent to that of FP64, and the difficulty in algorithm development lies in the fact that speedup must be achieved in consideration of the computing performance, data transfer, and other factors.

Therefore, we propose TCOVFEM, which is an OVFEM transformed to take advantage of the INT8 Tensor Core arithmetic performance with guaranteed arithmetic accuracy. Generally, solids can be treated as a linear isotropic dynamic elastic body in the most basic approximation; thus, the formulation is presented using a linear isotropic dynamic elastic body as an example. This method can be applied to the ultrasonic wave propagation analysis shown in the application examples presented in this paper and to the analysis of wavefields in other linear isotropic elastic bodies. The key is determining how to make $\mathbf{K}\mathbf{u}$, which has the highest analysis cost in Eq. (2), suitable for Tensor Core operations while guaranteeing FP64 accuracy. First, we consider the computation of $(\psi^\beta \nabla \phi^\alpha)_e$ among the operations given in Eq. (5). Note that there is arbitrariness in the selection of ψ^β in Eq. (8); thus, we select a polynomial that follows the deformation mode of the cube, which leads to an expression within the range of low-precision numbers by proper normalization. In contrast, the computation of $(\psi^\beta \psi^{\beta'})_e^{-1} \mathbf{c}$ requires care. Generally, \mathbf{c} does not fall within the range of low-precision numbers (even if it is linear isotropic); thus, it must be treated appropriately. In this section, we focus on the independence of the properties of linear isotropic elastic bodies and derive a formulation that falls within the range of 8-bit integers. For linear isotropic elastic bodies, \mathbf{c} can generally be expressed by the product of the volume modulus κ , the shear modulus G , and a constant that fits into low-precision arithmetic. Here, by defining $\bar{\mathbf{D}}^\kappa = 3 \frac{\partial (\psi^\beta \psi^{\beta'})_e^{-1} \mathbf{c}}{\partial \kappa} ds^3$ and $\bar{\mathbf{D}}^G = 3 \frac{\partial (\psi^\beta \psi^{\beta'})_e^{-1} \mathbf{c}}{\partial G} ds^3$ as the contribution of κ and G to $(\psi^\beta \psi^{\beta'})_e^{-1} \mathbf{c}$, a

constant 24×48 matrix

$$\mathbf{K}_e^{\text{INT8}} = (\mathbf{K}_e^\kappa, \bar{\mathbf{K}}_e^G),$$

where

$$\begin{aligned} \mathbf{K}_e^\kappa &= 256/3 \bar{\mathbf{B}}^t \bar{\mathbf{D}}^\kappa \bar{\mathbf{B}} ds^4, \\ \bar{\mathbf{K}}_e^G &= 128 \bar{\mathbf{B}}^t \bar{\mathbf{D}}^G \bar{\mathbf{B}} ds^4 - 128 \mathbf{I}, \end{aligned}$$

can be constructed. Here, $\bar{\mathbf{B}}$ is an equivalent strain displacement matrix and \mathbf{I} is a 24×24 identity matrix. Using this transformation, the core kernel, which has the highest analysis cost in finite element analyses, can be computed as follows:

$$\mathbf{K}_e^o \mathbf{u}_e = 1/256 \kappa ds (\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_e + 256/3G/\kappa \mathbf{u}_e). \quad (9)$$

Here, $\bar{\mathbf{u}}_e = (\mathbf{u}_e, 2/3G/\kappa \mathbf{u}_e)$.

In Eq. (9), we convert most of the analysis cost into operations that fall within the 8-bit integer range; however, to extract the full performance of Tensor Cores, it is necessary to implement Tensor Core computations in consideration of both the operation speed and the data transfer cost. Here, we present a concrete implementation of the fast Tensor Core-based computation of the dense matrix-vector product $\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_e$. Here, $\mathbf{K}_e^{\text{INT8}}$ is a 24×48 constant matrix with values in the 8-bit integer range (-128 - 127), and $\bar{\mathbf{u}}_e$ is a vector with a length of 48 in FP64. Note that conversion between floating-point numbers and integers incurs a large cost; thus, we consider fast computation algorithms that minimize the number of variable conversions while reducing the data transfer cost of the computation.

First, to save the number of significant digits during integer arithmetic expansion, the element right-hand side vector is scaled as follows:

$$\bar{\mathbf{u}}_{es} = \frac{1}{s_e} \bar{\mathbf{u}}_e. \quad (10)$$

Here, $s_e = \max_{i=1,2,\dots,48} |\bar{u}_{ei}|$, where \bar{u}_{ei} is the i -th component of $\bar{\mathbf{u}}_e$. This scales each component of $\bar{\mathbf{u}}_e$ to be within ± 1 . Then, $\bar{\mathbf{u}}_{es}$ is expanded with the product of FP64 constants times integer value vectors:

$$\bar{\mathbf{u}}_{es} = \sum_{i=1}^N \frac{1}{a^i} \bar{\mathbf{u}}_{es\text{INT}(i)}, \quad (11)$$

where N is the number of integer expansion stages, and $\bar{\mathbf{u}}_{es\text{INT}(i)}$ is a vector comprising 48 integers:

$$\bar{\mathbf{u}}_{es\text{INT}(1)} \stackrel{\leftarrow}{\text{INT}} a \bar{\mathbf{u}}_{es}, \quad (12)$$

$$\bar{\mathbf{u}}_{es\text{INT}(2)} \stackrel{\leftarrow}{\text{INT}} a \{ a \bar{\mathbf{u}}_{es} - \bar{\mathbf{u}}_{es\text{INT}(1)} \}, \quad (13)$$

$$\bar{\mathbf{u}}_{es\text{INT}(3)} \stackrel{\leftarrow}{\text{INT}} a \{ a^2 \bar{\mathbf{u}}_{es} - a \bar{\mathbf{u}}_{es\text{INT}(1)} - \bar{\mathbf{u}}_{es\text{INT}(2)} \}. \quad (14)$$

...

Here, \llcorner_{INT} denotes truncation to integer values. Using these integer vectors, the element matrix-vector product is computed as follows:

$$\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_e = s_e \sum_{i=1}^N \frac{1}{a^i} \mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{e\text{sINT}(i)}. \quad (15)$$

The above Eqs. (11)–(15) can be expanded and calculated with integers with an arbitrary number of bits. For example, using $a = 2^7$, all integer operations can be performed with 8-bit integers. However, in this case, Eqs. (12)–(14) must be calculated for the N stages, which requires many conversions between floating-point variables and integers, and, therefore, reduces the computational performance. For example, $N = 8$ stages are required to realize 56-bit accuracy which is equivalent to FP64's 52 fraction bits, which requires many conversion operations. Thus, in this study, the number of conversions between the FP64 variables and integers is reduced using 64-bit integers to perform the conversion only once and by using a hierarchical expansion in which the 64-bit integer calculation is further expanded using 8-bit integer operations. In particular, Eqs. (11)–(12) are expanded using 64-bit integers with $a = 2^{56}$ and $N = 1$, and the main computation $\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{e\text{sINT}(1)}$ (denoted $\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{e\text{sINT64}(1)}$) is expanded using $M = 8$ -stages of computations using 8-bit integers. Here, using $b = 2^7$, $\bar{\mathbf{u}}_{e\text{sINT64}(1)}$ is expanded as

$$\bar{\mathbf{u}}_{e\text{sINT64}(1)} = \sum_{j=1}^M b^{j-1} \bar{\mathbf{u}}_{e\text{sINT8}(1,j)} \quad (16)$$

using 8-bit integers, and $\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{e\text{sINT64}(1)}$ is computed as

$$\sum_{j=1}^M b^{j-1} \mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{e\text{sINT8}(1,j)}. \quad (17)$$

Figure 2 shows the computation flow of the matrix-vector products when directly converting data between FP64 and INT8 variables and when hierarchical FP64-INT64-INT8 conversion is used. While the data conversion and computation are looped N times in the direct method, the proposed hierarchical method performs the data conversion outside of the M loop. While the total computation accuracy is the same when $N = M$, the number of conversions between the integer and FP64 variables can be reduced by N -fold when using the hierarchical method.

Tensor Core operations are very fast when applied to matrix-matrix products of specific sizes; thus, we must transform the computation pattern of the matrix-vector products to matrix-matrix products suited for Tensor Cores. In addition, Tensor Core operations are very fast; thus, the cost of other operations must be suppressed to reduce time costs substantially. In particular, the data transfer from global memory and the data transfer between shared memory becomes a bottleneck. In the following, we describe how we map the matrix-vector computation to Tensor Cores and how we circumvent the data transfer bottlenecks.

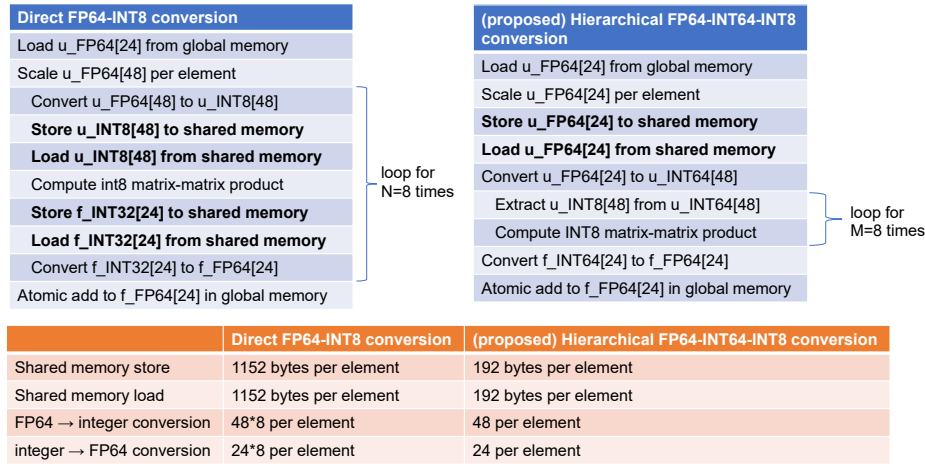


Fig. 2: Computation of matrix-vector products using INT8 Tensor Cores with direct FP64-INT8 conversion and the proposed hierarchical FP64-INT64-INT8 conversion. Here, 32 elements are computed using 32 threads per thread block.

- Transformation of matrix-vector products to matrix-matrix products: While $\mathbf{K}_e^{\text{INT8}} \bar{\mathbf{u}}_{es\text{INT8}(1,j)}$ is a matrix-vector product of size $(24 \times 48) \times 48$, we can compute 32 elements in a single thread block as a matrix-matrix product of size $(24 \times 48) \times (48 \times 32)$. We compute this using INT8 Tensor Cores with $(n, m, k) = (8, 32, 16)$: $C(8 \times 32: \text{INT32}) = A(8 \times 16: \text{INT8}) \times B(16 \times 32: \text{INT8})$. As shown in Fig. 3, a $(24 \times 48) \times (48 \times 32)$ matrix-matrix multiplication is decomposed into $3 \times 3 = 9$ parts, and each part is computed using $(8 \times 16) \times (16 \times 32)$ matrix-matrix multiplication. Here, the matrix $\mathbf{K}_e^{\text{INT8}}$ is constant in the j -loop; thus, we can load the A fragment once and keep it in registers.
- Elimination of shared memory loads/stores by direct addition of results to global memory: As the register mapping of memory fragments for the matrices in Tensor Core operations is complex, an API is provided to exchange the input and output data between threads via shared memory. However, when this technique is used for the small-scale matrix-matrix multiplications targeted in this study, it frequently results in shared memory bottlenecks, which makes it difficult to fully utilize the available performance of the Tensor Cores. Thus, following the method presented in the literature [6], we skip the remapping process using shared memory and directly output the data to global memory. As shown in Fig. 4, rather than remapping the outputs from the Tensor Core operations and having each thread add element-wise results to the global memory, the results are added directly to the designated components in the global memory. Here, the scaling coefficient s_e for each element is reflected in this procedure by sharing s_e among the threads. Simi-

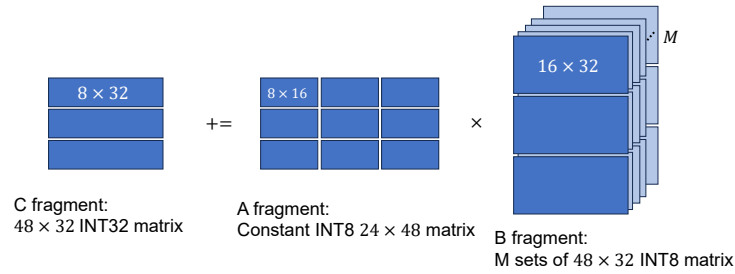


Fig. 3: Decomposition of $(24 \times 48) \times (48 \times 32)$ matrix-matrix product into nine $(8 \times 16) \times (16 \times 32)$ matrix-matrix products. Here, the A fragment can be reused throughout the M sets of computations. Note that the results in the 32-bit integer C fragment are flushed every $M = 3$ stages to an INT64 buffer to avoid overflow.

larly, we can load the B fragments directly from the registers after conversion from INT64 value inputs by proper element mapping.

- Improving cache reuse when accessing global memory variables: The matrix-vector product calculation requires reading the nodal data \mathbf{u}_e from global memory and adding the nodal result \mathbf{f}_e to the global vector \mathbf{f} . In this study, 32 elements belonging to a single thread block are arranged in the x -direction to facilitate sequential cache access in the x -direction, and then each thread block performs calculations continuously in the z -direction. This allows the reuse of nodal data (\mathbf{u} and \mathbf{f}) in cache, which results in an approximately fourfold reduction in the volume of global memory accesses compared to the case when elements are ordered randomly.

By employing a hierarchical method that reduces the number of conversions between the floating-point and integer variables, as well as using the data access reduction methods described above, we can expect high increases in speed compared to the standard FP64 computations without reduction in computational accuracy.

3 Numerical Experiment

In the following, the efficacy of the proposed TCOVFEM is demonstrated through ultrasonic analysis using the model illustrated in Fig. 5, which is based on the literature in [10]. The model is a rebar with a radius of 15 mm embedded in concrete. An impulse force with a center time of 4.096×10^{-4} s and a center frequency of 112.5 kHz with a bandpass of 100–125 kHz is applied in the z -direction from the input point, and its response is calculated and stored for 8.192×10^{-4} s at the observation points shown in Table 1. The size of the model is $324 \times 128 \times 384$ mm and assumes a cylindrical rebar of radius 15 mm centered at $x = 160$, $z = 100$ mm that is penetrating the model in the y -direction.

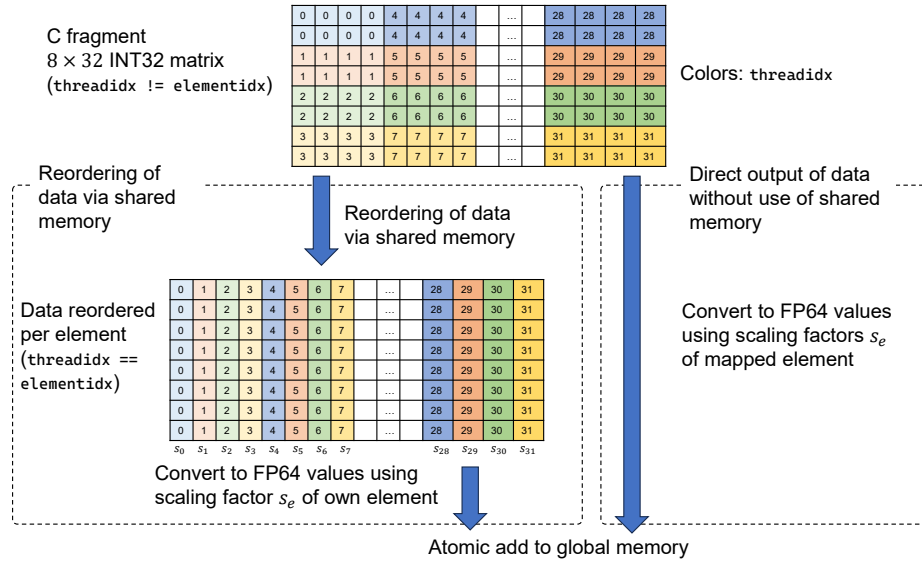


Fig. 4: Elimination of shared memory loads/stores by direct addition of results to global memory. Note that one out of the three C fragments is shown. 32 elements are computed using 32 threads per thread block.

The four corner points at the bottom ($z = 0$) of the model are fixed in three directions (x , y , and z). Rayleigh damping (100–125 kHz) is used to compute attenuation in the simulation. The following time history simulations were performed on a single A100 80 GB PCIe GPU [9], and except for the INT8 Tensor Core operation in the proposed TCOVFEM, all other operations were computed and stored using FP64 variables.

First, a reference solution was generated using an implicit finite element analysis with second-order tetrahedral elements without the lumped mass matrix approximation (hereafter referred to as CFEM), and we compared the waveforms at the observation points to confirm the accuracy of the proposed TCOVFEM. As shown in Fig. 5, the finite element mesh used for the reference solution based on CFEM is sufficiently fine. For example, the area near the rebar is discretized with an element size of approximately 1 mm (i.e., the nodal spacing becomes approximately 0.5 mm as second-order tetrahedral elements are used). We used a model with 2-mm voxel elements for the proposed TCOVFEM, and to compare the accuracy with the conventional VFEM method, we conducted VFEM analysis for cases with seven element sizes (2.0, 1.5, 1.33, 1.2, 1.09, 1.0, and 0.5 mm). Table 2 shows the error from the reference solution at all observation points, which is defined as follows for each analysis case.

$$Err = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\sum_{j=1}^{n_t} (u_{i,j}^{obs} - u_{i,j}^{ref})^2}{\sum_{j=1}^{n_t} (u_{i,j}^{ref})^2} \quad (18)$$

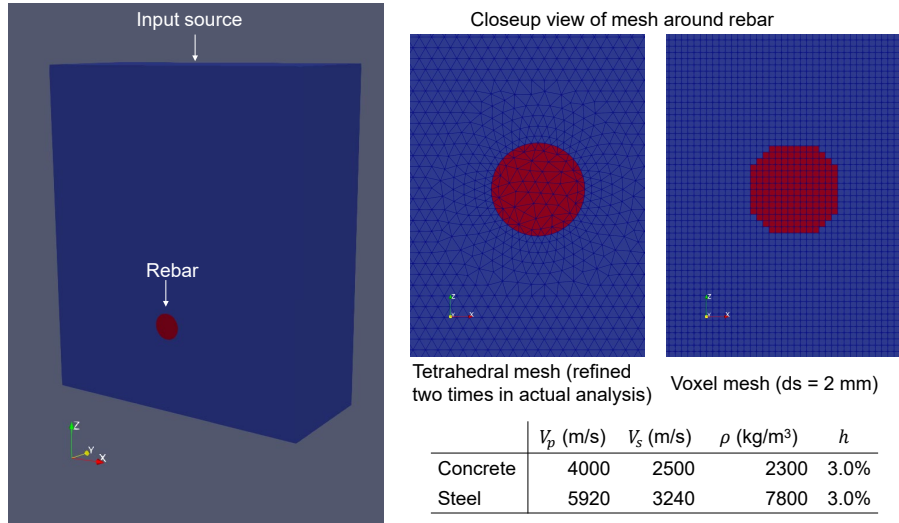


Fig. 5: Model used for the numerical experiment.

Here, n_c denotes the number of observation channels (eight observation points $\times (x, y, z)$ -components = 24 channels), n_t denotes the number of time steps, and $u_{i,j}^{\text{obs}}$ and $u_{i,j}^{\text{ref}}$ denote the computed and reference waves at time step j for channel i , respectively. The results demonstrate that the proposed TCOVFEM can compute results with high accuracy compared to the conventional method. The error of CFEM using a coarser mesh (the length of an element edge is twice that of the mesh used in the reference mesh) was 0.00022, which confirms that the reference solution computed using CFEM with the fine mesh is converged sufficiently. A comparison of the proposed TCOVFEM with the conventional VFEM indicates that the proposed method is more accurate than the VFEM in the case with the same discretization width. Here, the proposed TCOVFEM with $ds = 2$ mm achieved accuracy that is equivalent to the VFEM with $ds = 1.2$ mm. Figure 6, which visualizes the response in the y - and z -directions at observation point 1, also confirms that the proposed TCOVFEM with $ds = 2$ mm is considerably more accurate than VFEM at $ds = 2$ mm (even when viewed at individual observation points).

Next, we review the effect of INT8 Tensor Core operations on calculation accuracy. Here, we compare the accuracy when computing Eq. (9) without Tensor Cores using FP128, FP64, and FP32 operations, and computing Eq. (9) with INT8 Tensor Cores. Specifically, the calculations for multiplying the global stiffness matrix of the $ds = 2$ mm cubic element model by random vectors with double precision values were computed with FP128, FP64, and FP32 operations, and these calculations were compared with the results computed using INT8 Tensor Core operations. Note that FP128 has 112 fraction bits, FP64 has 52 fraction bits, and FP32 has 23 fraction bits. In contrast, as the resolution of

Table 1: Coordinates of the input source and observation points.

	x (mm)	y (mm)	z (mm)
Source point	156	72	384
Obs. point 1	26	60	384
Obs. point 2	60	60	384
Obs. point 3	108	60	384
Obs. point 4	144	60	384
Obs. point 5	180	60	384
Obs. point 6	216	60	384
Obs. point 7	264	60	384
Obs. point 8	300	60	384

Table 2: Comparison of relative error (TCOVFEM: proposed method; VFEM: explicit standard voxel FEM).

	ds (mm)	dt (s)	Err
TCOVFEM	2.000	5.0×10^{-8}	0.03947
VFEM	2.000	5.0×10^{-8}	0.28057
VFEM	1.500	2.5×10^{-8}	0.09579
VFEM	1.333	2.5×10^{-8}	0.05743
VFEM	1.200	2.5×10^{-8}	0.03602
VFEM	1.091	2.5×10^{-8}	0.02343
VFEM	1.000	2.5×10^{-8}	0.01570
VFEM	0.500	1.25×10^{-8}	0.00104

Table 3: Comparison of computed values for one component of \mathbf{Ku} computed for a random vector \mathbf{u} on OVFEM with $ds = 2$ mm. Values in bold letters are identical to the FP128 computation results.

Computation type	Fraction bits	Value
FP128	112	507813.690592559616827867910192902549
FP64	52	507813.6905925632
FP32	23	507814.750
INT8 ($M = 4$)	28	507813.7802133318
INT8 ($M = 8$)	56	507813.6905925595

an 8-bit integer is 7 bits, the INT8 Tensor Core calculation is equivalent to 28 fraction bits for $N = 4$ stages and 56 fraction bits for $M = 8$ stages. Thus, it is expected that accuracy equivalent to that of FP32 can be obtained with $M = 4$ stages and higher than that of FP64 with $M = 8$ stages in the INT8 calculation. The results shown in Table 3 demonstrate that the accuracy of the INT8 operation increases with the number of stages, and that the INT8 operations with $M = 8$ stages are equivalent to or higher than that of FP64. In addition, the INT8 operations with $M = 4$ stages are equivalent to that of FP32. Thus, we use $M = 8$ stages in the proposed TCOVFEM throughout the rest of this paper.

Finally, we compare the elapsed time for the entire time series analysis and the matrix-vector product kernel (Table 4). In the following, we refer to \mathbf{Ku}^{it} in

$$\mathbf{u}^{it+1} \Leftarrow 2\mathbf{u}^{it} - \mathbf{u}^{it-1} + dt^2 \mathbf{M}^{-1}(\mathbf{F}^{it} + \mathbf{Ku}^{it}), \quad (19)$$

which is obtained by transforming Eq. (3), as the matrix-vector product kernel, and the rest as other computations. We first compare the proposed TCOVFEM with $M = 8$ stages (TCOVFEM INT8 ($M = 8$)) with OVFEM without Tensor Cores (OVFEM FP64), which computes Eq. (9) as is with FP64 variables on the $ds = 2$ mm model. In the model with $ds = 2$ mm, the other computations took 4.62 s. By excluding this, we can evaluate the improved calculation time of the matrix-vector product operation when using the Tensor Cores. The matrix-vector product calculation was $43.3/9.62 = 4.5$ -fold faster using the Tensor Cores,

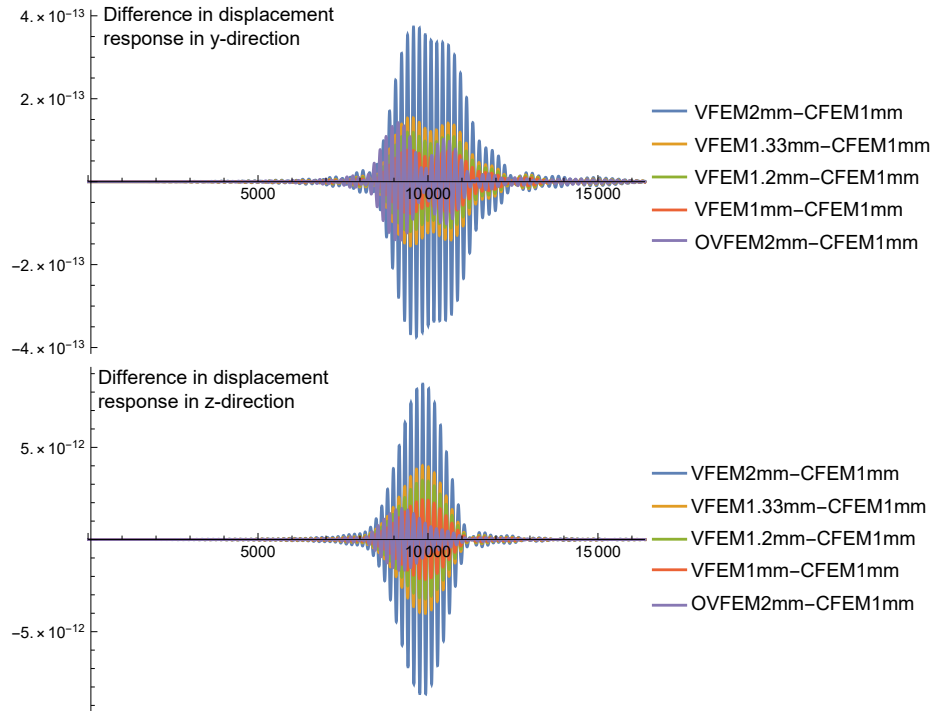


Fig. 6: Differences in time history displacement response at observation point 1 when compared with the reference solution (y - and z -directions with $dt = 5 \times 10^{-8}$ s).

which corresponds to 64.4 TOPS in computational performance. In contrast, the OVFEM matrix-vector product calculation using FP32 (without Tensor Cores), which had lower accuracy (Table 3), took 15.5 s. Thus, the proposed TCOVFEM, which demonstrates computational accuracy that is equivalent to that of FP64 computations, is 1.6 times faster than the FP32 implementation. This improved calculation time for the matrix-vector product resulted in a $48.3/14.2 = 3.39$ -fold speedup of the overall time history simulation compared to the FP64-based OVFEM without Tensor Cores. We also compared the proposed method with the conventional VFEM. Note that VFEM uses the element-by-element (EBE) method in the explicit time integration code, which calculates the matrix-vector product in an element-by-element manner without keeping the entire stiffness matrix in memory (all calculations are performed in FP64). Here, the VFEM EBE (8-point integration) implementation in the paper [11], which achieves high computational performance as a regular EBE-based VFEM, was used for comparison. As shown in Table 2, elements with $ds = 1.2$ mm are required for the VFEM to obtain accuracy that is equivalent to that of the proposed TCOVFEM with $ds = 2$ mm sized elements. Consequently, since the number of elements is

Table 4: Comparison of elapsed time for wave propagation. Here, $dt = 5 \times 10^{-8}$ s and 16,384 time steps were used for cases with $ds = 2$ mm, and $dt = 2.5 \times 10^{-8}$ s and 32,768 time steps were used for cases with $ds = 1.2$ mm.

Computation type	ds (mm)	Elapsed time (s)	
		Time-step loop	Matrix-vector product
OVFEM FP64	2.0	48.3	43.3 [2.1 TFLOPS]
OVFEM FP32	2.0	18.9	15.5 [5.9 TFLOPS]
TCOVFEM INT8 ($M = 8$)	2.0	14.2	9.62 [64.4 TOPS]
VFEM FP64	1.2	242.8	-

$(2.0/1.2)^3 = 4.63$ times greater, and the number of time steps is twice as large, the amount of operations increases such that the analysis takes 242.8 s even using this efficient VFEM implementation (Table 4). Thus, the proposed TCOVFEM is $242.8/14.2 = 17.0$ -fold faster than the conventional VFEM with equivalent accuracy in the wave calculation.

4 Concluding Remarks

We presented an explicit wavefield simulation method that uses structured finite elements with high speed and low numerical dispersion using INT8 Tensor Cores to effectively exploit the performance benefits of modern GPUs. A detailed comparison of the performance of the proposed and conventional methods on a real-world problem demonstrates that the proposed method utilizes GPU performance efficiently and that the proposed TCOVFEM (INT8 ($M = 8$)) is 17.0-fold faster than the conventional VFEM. Tensor Cores, which are an acceleration mechanism in modern GPUs, are used in deep learning and other applications due to their high computing performance. However, the application of Tensor Cores in physics-based simulations remains limited because it must be brought into a specific format and data transfer must be performed without losing computing performance. In addition, the use of Tensor Cores to obtain a coarse solution and then refine it, as in the implicit method, is difficult in the explicit method, thereby making it even more challenging. In this context, the formulation and implementation presented in this paper, which exploits the performance advantages of Tensor Cores in explicit physics-based simulations, are general and expected to provide insights into accelerating other applications. The conversion of floating-point matrix-vector computations to small integer-based matrix-matrix computations is expected to be accelerated in other GPU architectures with acceleration mechanisms for integer-based matrix-matrix multiplication. Note that the purpose of this study was to demonstrate the basic performance of the proposed method; thus, the analysis was performed on only a single GPU. However, we expect that the proposed method can easily be extended to solve large-scale problems using multiple GPUs.

References

1. Koketsu, K., Fujiwara, H., Ikegami, Y.: Finite-element Simulation of Seismic Ground Motion with a Voxel Mesh, *Pure and Applied Geophysics* 161(11):2183–2198 (2004)
2. Bielik, J., Ghattas, O., Kim, E. J.: Parallel Octree-Based Finite Element Method for Large-Scale Earthquake Ground Motion Simulation, *Computer Modeling in Engineering & Sciences*, 10(2):99–112 (2005)
3. Nakahata, K., Amano, Y., Ogi, K., Mizukami, K., Saitoh, T.: Three-dimensional ultrasonic wave simulation in laminated CFRP using elastic parameters determined from wavefield data, *Composites Part B: Engineering*, 176 (2019)
4. NVIDIA Tensor Cores. <https://www.nvidia.com/en-us/data-center/tensor-cores/>, last accessed 2024/2/28
5. Ichimura, T., Fujita, K., Yamaguchi, T., Naruse, A., Wells, J. C., Zimmer, C. J., Straatsma, T. P., Hori, T., Puel, S., Becker, T. W., Hori, M., Ueda, N.: 416-PFLOPS Fast Scalable Implicit Solver on Low-Ordered Unstructured Finite Elements Accelerated by 1.10-ExaFLOPS Kernel with Reformulated AI-Like Algorithm: For Equation-Based Earthquake Modeling, SC19 (the International Conference for High Performance Computing, Networking, Storage, and Analysis), Research Poster (2019)
6. Yamaguchi, T., Fujita, K., Ichimura, T., Naruse, A., Wells, J. C., Zimmer, C. J., Straatsma, T. P., Hori, M., Maddegadara, L., Ueda, N.: Low-Order Finite Element Solver with Small Matrix-Matrix Multiplication Accelerated by AI-Specific Hardware for Crustal Deformation Computation, In Proceedings of the Platform for Advanced Scientific Computing Conference (PASC’20). Association for Computing Machinery, 16:1–11 (2020)
7. Higham, N.J. and Mary, T.: Mixed precision algorithms in numerical linear algebra, *Acta Numerica*, 31, pp. 347–414. doi:10.1017/S0962492922000022 (2022)
8. Ichimura, T., Hori, M., Wijerathne, M. L. L.: Linear finite elements with orthogonal discontinuous basis functions for explicit earthquake ground motion modeling, *International Journal for Numerical Methods in Engineering*, 86:286–300 (2011)
9. NVIDIA Ampere architecture whitepaper. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf>, last accessed 2024/2/27
10. Dinh, K., Tran, K., Gucunski, N., Ferraro, C. C., Nguyen, T.: Imaging Concrete Structures with Ultrasonic Shear Waves—Technology Development and Demonstration of Capabilities. *Infrastructures*, 8(53) (2023)
11. Ichimura, T., Fujita, K., Hori, M., Maddegadara, L., Ueda, N., Kikuchi, Y.: A Fast Scalable Iterative Implicit Solver with Green’s function-based Neural Networks, *IEEE/ACM 11th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA)* (2020)