Reformulated AI-like Algorithm for Solving Huge Problems

**Introduction**

A better understanding of earthquake physics is a grand challenge because of the potential of large damage to the society and cities.

- A magnitude-9 earthquake is anticipated along the San Andreas Fault System, which could also be affected by the plate activity in the Cascadia Subduction Zone, where a magnitude-9 earthquake and a huge tsunami occurred in 1700.
- We expect probabilistic long-term earthquake forecasting to become possible by constructing a physics-based earthquake model with a realistic plate geometry and an assimilation of continuous data while solving the governing equations.
- The computation of governing equations with equation-based modeling considering the crust, plate, and fault geometry in high fidelity is required.
- Unstructured fictitious finite-element method is crucial for computing the visco-elastic-plastic time history on a heterogeneous 3D structure.

**Huge cost in computing the large spatial- and temporal-scale problem**

- Many case analyses for large spatial- and temporal-scale problems in high fidelity are required (10^13 km^3 km^-1 day^-1; 10^2 year duration; km-scale resolution; 10^2 iterations for assimilating data and considering uncertainty).
- The visco-elastic-plastic computation cost is equivalent to solving 10^11-12 degrees-of-freedom (DOF) elastic analysis for 10^4-10^5 cases.
- At least a 50-fold speedup is required to conduct this analysis even when using the state-of-the-art solver on full Piz Daint.
- State-of-the-art solver, a directive-based SC16 WACCPD solver [1] designed for P100 GPU based systems, was developed based on the SC14 Gordon Bell finalist solver [2].

**Developed solver attains a 75-fold speedup from the state-of-the-art solver**

- The developed solver attains a 75-fold speedup from the state-of-the-art solver on full Summit corresponded to a 75-fold speedup from the state-of-the-art solver on full Piz Daint.
- This speedup was very high considering the 215/25 = 8.6-fold difference in the FP64 system peak performance between Summit and Piz Daint.
- This speedup is expected to be enough to conduct breakthroughs in science.

**Efficient Implementation of Tensor Core**

Special care required for using Tensor Cores for small matrices in equation-based modeling

- Tensor Core is designed for large matrix-multiplication with lower precision data types.
- The reduction of data access cost and prevention of loss of accuracy are required.

1. Ensuring convergence of the solver
   - Although a low precision is allowed, a very low precision leads to preconditioner failure.
   - The values of vectors \( p_{ij} \) and \( w_{ij} \) are normalized per element to improve accuracy.

2. Efficient data mapping of small matrices
   - Frequent data movement leads to inefficiency.
   - The computation of 32 elements is subdivided into 72 Tensor Core operations for reuse of matrix many times on registers.

- The API for the Tensor Core computation requires data movement between the shared memory and the registers; thus, we compute on registers by using the PTX assembly.

**Performance Measurement**

- Performance of the matrix-vector kernel
  - The performance of the multiplication of the basis functions with complex connectivity and varying strengths.
  - The values of vectors \( p_{ij} \) and \( w_{ij} \) corresponding to the material properties of the 32 elements.

- The multiplication of the basis functions with complex connectivity and varying strengths.
- Frequent data movement leads to inefficiency.
- Same final FP64 results with those computed by the state-of-the-art solver on full Piz Daint.
- The FP64 implementation of the matrix-vector kernel.

**Summary and Future Prospects**

- An equation-based earthquake modeling algorithm is transformed to an algorithm suitable for high-performance hardware originally designed for AI.
- High performance and scalability on full Summit are achieved.
- Our approach using local and uniform expansions is applicable to other problems where the target computer architecture characteristics.
- We plan to use the developed method to enable long-term earthquake forecasting, which is expected to address earthquake disaster mitigation.

**Acknowledgments**

Our results were obtained using the Summit at Oak Ridge Leadership Computing Facility, a U.S. Department of Energy, Office of Science User Facility at Oak Ridge National Laboratory (ORNL). We thank Yukihiko Hirano (NVIDIA) for coordination of the collaborative research project. We thank Christopher B. Fussel, Don E. Maxwell, Oscar Hernandez, Scott Atchley, Veronica Melesse-Vergara (ORNL), Jeff Larkin, Stephen-Abbott (NVIDIA), Lixiao Luo (IBM), Richard Graham ( Mellanox Technologies) for generous support concerning use of Summit. We thank Takashi Kono (The University of Tokyo) for support and discussion at the Earthquake Research Institute. We thank Nozomi Tomoyuki and Hikaru Iwase (Fujitsu Limited) for support in program development. We acknowledge support from Post K computer project (Priority Issue 3 - Development of integrated simulation systems for hazards and disasters induced by earthquakes and tsunamis: proposal numbers 19017, 190077, and 190174) and Japan Society for the Promotion of Science (19H05239, 26249096, 25260898, and 17K174179). Simone Puel and Thorsten W. Becker acknowledge support from NSF-EAR 1722690.

**References**