Performance Evaluation of Acoustic Material Implementation in SeisSol Using Roofline Model

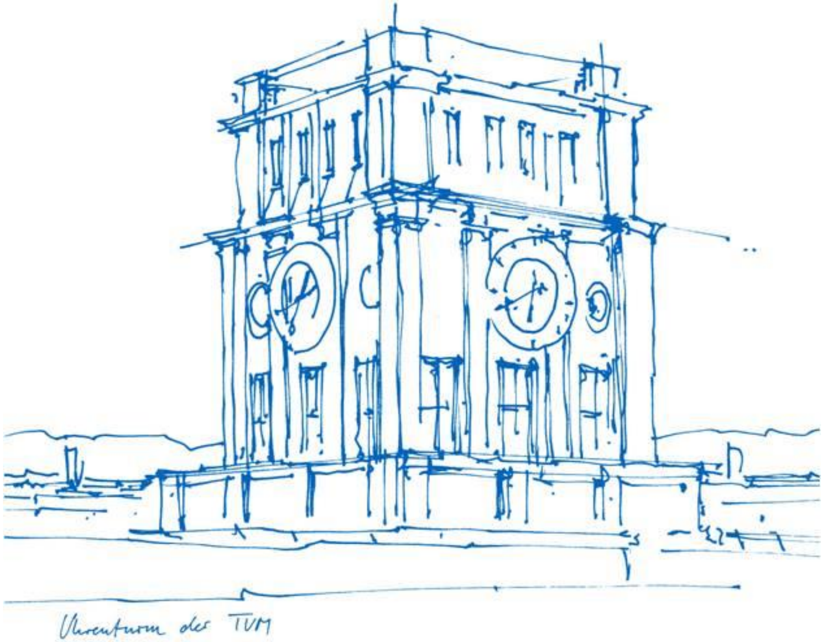
Jinwen Pan

Technical University of Munich

TUM School of Computation, Information and Technology

Chair of Scientific Computing

Garching, 27. November 2024



Content

- Physical Models
- Roofline Model
- Experimental Analysis
- Conclusion and Future Work

SeisSol

An open-source software for the simulation of seismic wave phenomena and earthquake dynamics

- Arbitrary high-order DERivative Discontinuous Galerkin method (ADER-DG)
  - ADER temporal discretization
  - DG spatial discretization

- Models of medium materials for seismic wave propagation
  - Elastic: isotropic and anisotropic
  - Poroelastic
  - Viscoelastic
  - Off-fault plastic
  - ...
  - How about acoustic? Water, air, oil?

  The ocean covers approximately 70% of Earth's surface!

Elastic Wave Model

Consider an infinitesimal cube near any point within a material:

Symmetric strain matrix: $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix}$, symmetric stress matrix: $\boldsymbol{\sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$

The generalized Hooke's law relates stress and strain (under small perturbation assumption):

$\sigma_{ij} = \lambda\delta_{ij}\sum_{k=1}^{3}\varepsilon_{kk} + 2\mu\varepsilon_{ij}$, derivative with respect to time: $\frac{\partial\sigma_{ij}}{\partial t} = \lambda\delta_{ij}\sum_{k=1}^{3}\frac{\partial v_k}{\partial x_k} + \mu(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i})$

Newton's second law:

$\frac{\partial}{\partial t}\int_{\Omega}\rho\frac{\partial \boldsymbol{u}}{\partial t}\,dV = \int_{\partial\Omega}\boldsymbol{T}(\boldsymbol{n})\,dS + \int_{\Omega}\boldsymbol{f}\,dV$, differential form: $\rho\frac{\partial v_i}{\partial t} = \sum_{j=1}^{3}\frac{\partial\sigma_{ij}}{\partial x_j} + f_i$

Finally, $\frac{\partial \boldsymbol{q}}{\partial t} + \boldsymbol{A}(\boldsymbol{x})\frac{\partial \boldsymbol{q}}{\partial x_1} + \boldsymbol{B}(\boldsymbol{x})\frac{\partial \boldsymbol{q}}{\partial x_2} + \boldsymbol{C}(\boldsymbol{x})\frac{\partial \boldsymbol{q}}{\partial x_3} = 0$ with $\boldsymbol{q} = (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{23}, \sigma_{13}, v_1, v_2, v_3)^T$

$$\hat{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{n}) = n_1\boldsymbol{A}(\boldsymbol{x}) + n_2\boldsymbol{B}(\boldsymbol{x}) + n_3\boldsymbol{C}(\boldsymbol{x}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -n_1(\lambda+2\mu) & -n_2\lambda & -n_3\lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & -n_1\lambda & -n_2(\lambda+2\mu) & -n_3\lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & -n_1\lambda & -n_2\lambda & -n_3(\lambda+2\mu) \\ 0 & 0 & 0 & 0 & 0 & 0 & -n_2\mu & -n_1\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -n_3\mu & -n_2\mu \\ 0 & 0 & 0 & 0 & 0 & 0 & -n_3\mu & 0 & -n_1\mu \\ -\frac{n_1}{\rho} & 0 & 0 & -\frac{n_2}{\rho} & 0 & -\frac{n_3}{\rho} & 0 & 0 & 0 \\ 0 & -\frac{n_2}{\rho} & 0 & -\frac{n_1}{\rho} & -\frac{n_3}{\rho} & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{n_3}{\rho} & 0 & -\frac{n_2}{\rho} & -\frac{n_1}{\rho} & 0 & 0 & 0 \end{pmatrix}$$

(neglecting f_i)

Acoustic Wave Model

Hydrostatic model: $\frac{\partial p_0}{\partial x_1} = \frac{\partial p_0}{\partial x_2} = \frac{\partial p_0}{\partial t} = \frac{\partial \rho_0}{\partial x_1} = \frac{\partial \rho_0}{\partial x_2} = \frac{\partial \rho_0}{\partial t} = 0$, $\boldsymbol{v_0}(\boldsymbol{x}, t) = 0$

Small perturbation: $(\boldsymbol{v}(\boldsymbol{x}, t), \rho(\boldsymbol{x}, t), p(\boldsymbol{x}, t))^T = (\boldsymbol{v_0}(\boldsymbol{x}), \rho_0(\boldsymbol{x}), p_0(\boldsymbol{x}))^T + (\boldsymbol{v'}(\boldsymbol{x}, t), \rho'(\boldsymbol{x}, t), p'(\boldsymbol{x}, t))^T$

Mass conservation: $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0$ -> $\frac{\partial p'}{\partial t} + K \sum_{k=1}^{3} \frac{\partial v'_k}{\partial x_k} - \rho_0 g v'_3 = 0$, bulk modulus: $K = \rho_0 \frac{dp}{d\rho}$

Momentum conservation: $\frac{\partial \rho \boldsymbol{v}}{\partial t} + \nabla \cdot (\boldsymbol{v} \otimes \rho \boldsymbol{v} + \boldsymbol{I} p) - \boldsymbol{f} = 0$ -> $\rho_0 \frac{\partial \boldsymbol{v'}}{\partial t} + \nabla \cdot (\boldsymbol{I} p') + \frac{\rho_0 g p'}{K} \boldsymbol{e_3} = 0$

Finally, neglecting the gravitational acceleration g :

$\frac{\partial \boldsymbol{q^{ac}}}{\partial t} + \boldsymbol{A^{ac}}(\boldsymbol{x}) \frac{\partial \boldsymbol{q^{ac}}}{\partial x_1} + \boldsymbol{B^{ac}}(\boldsymbol{x}) \frac{\partial \boldsymbol{q^{ac}}}{\partial x_2} + \boldsymbol{C^{ac}}(\boldsymbol{x}) \frac{\partial \boldsymbol{q^{ac}}}{\partial x_3} = 0$, with $\boldsymbol{q^{ac}} = (p', v'_1, v'_2, v'_3)^T$

$$\hat{\boldsymbol{A}}^{\boldsymbol{ac}}(\boldsymbol{x}, \boldsymbol{n}) = n_1 \boldsymbol{A^{ac}}(\boldsymbol{x}) + n_2 \boldsymbol{B^{ac}}(\boldsymbol{x}) + n_3 \boldsymbol{C^{ac}}(\boldsymbol{x}) = \begin{pmatrix} 0 & n_1 K & n_2 K & n_3 K \\ \frac{n_1}{\rho_0} & 0 & 0 & 0 \\ \frac{n_2}{\rho_0} & 0 & 0 & 0 \\ \frac{n_3}{\rho_0} & 0 & 0 & 0 \end{pmatrix}$$

Elastic vs Acoustic

When $\mu = 0$, the elastic model reduces to the acoustic model: $\boldsymbol{q} = (\sigma_{ii}, \sigma_{ii}, \sigma_{ii}, 0, 0, 0, v_1, v_2, v_3)^T$

Physical explanation of fluids: pressure is isotropic, no shear strain (stress)

Acoustic materials are a subset of elastic materials: $\sigma_{ii} = -p$

Current acoustic SeisSol: solving 9 PDEs for 4 DOFs (degrees of freedom)

But solving 4 PDEs are enough!

Kernels can be reused!

Performance comparison of elastic (with $\mu = 0$) and acoustic models

Target Application: SeisSol-Proxy

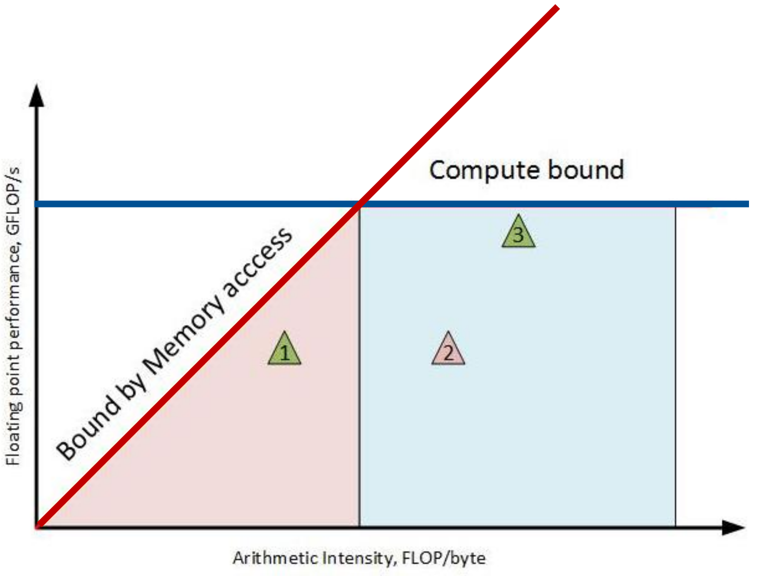
Proxy or mini applications are typically simple and lightweight but capture the core computational characteristics of the main simulation program.

- C++ OpenMP program run on a single node (all cores without SMT) of CoolMUC-2 at LRZ
- Parameters specified when building: physical model, accuracy order, host architecture (Haswell), numerical precision (double)
- $\mu = 0$ is hard-coded in the source code of the elastic model (only for SeisSol-proxy)
- `$ ./<proxy_executable> <num_cells> <num_timesteps> <kernel>`
  - `<kernel> = all` is executed `<num_timesteps> = 100` times on `<num_cells>` cells in parallel.
  - Iterations -> OpenMP threads: static scheduling
  - OpenMP threads -> hardware threads: one-to-one, pinned by LIKWID
  - First touch to avoid NUMA issue and load LLC for more accurate timing
  - Non-zero FLOP vs **hardware FLOP**

Roofline Model

A visual model used to analyze and optimize the performance of compute-intensive applications, which helps identify performance bottlenecks by comparing computational performance (FLOPS) with data throughput (memory bandwidth).

- Arithmetic (operational, computational) intensity:
  - Large value suggests compute-bound scenarios
  - Small value suggests memory-bound scenarios
- Components:
  - Peak performance (FLOPS)
  - Maximum data throughput (memory bandwidth)
  - Application points (arithmetic intensity, FLOPS)

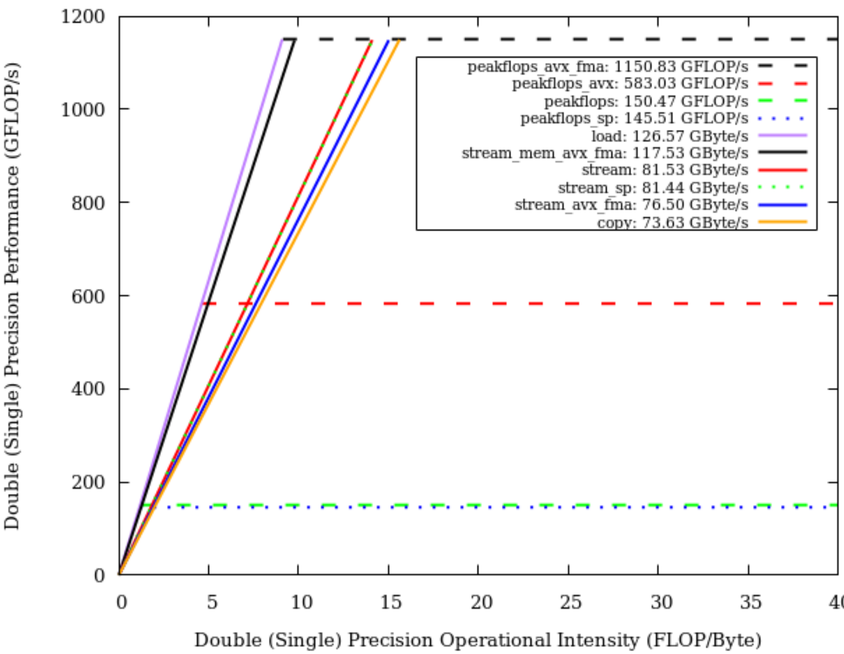


Roofline Models: Different Becnmarks

- Why benchmarks?
  - No full parallelization and vectorization
  - Ideal IPC
  - Thermal constraints

- Peak FLOPS
  - Precision
  - Vectorization
  - FMA

- Maximum memory bandwidth
  - `load (scalar=A[i])`
  - `copy (A[i]=B[i])`
  - `stream (A[i]=B[i]+scalar*C[i])`
  - `[_mem]`: non-temporal write

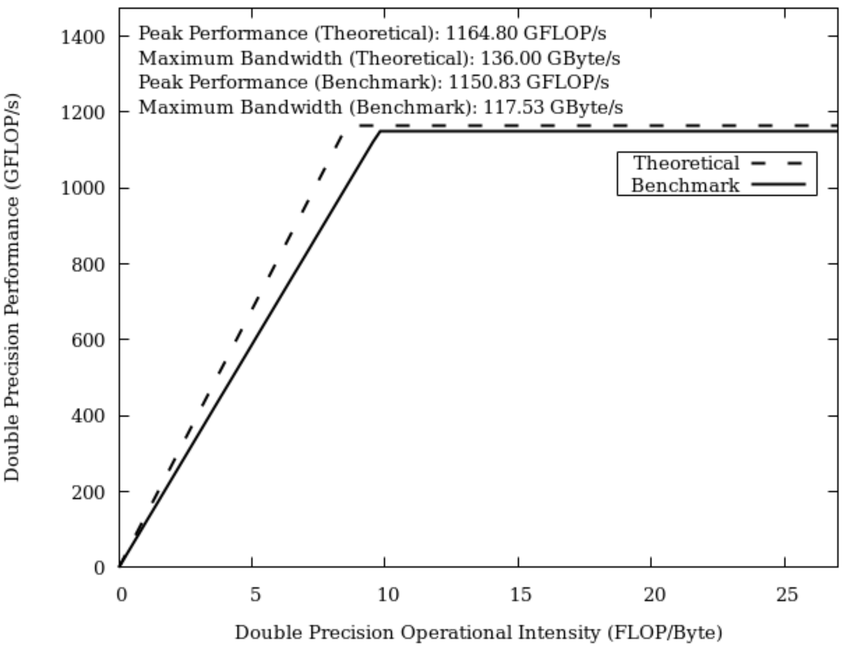


Roofline Models: Theoretical vs Benchmark

- Which benchmarks?
  - `peakflops_avx_fma`
  - `stream_mem_avx_fma` (read:write = 2:1)

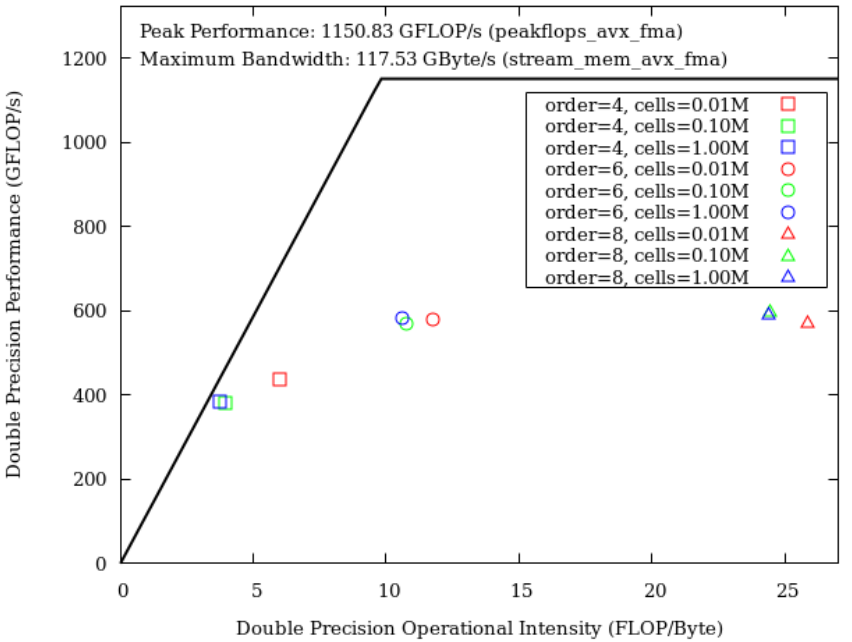
- Theoretical peak FLOPS:
  - Cores
  - Instruction width
  - FMA
  - IPC
  - Maximum clock frequency

- The theoretical maximum memory bandwidth is obtained from Intel's documentation.

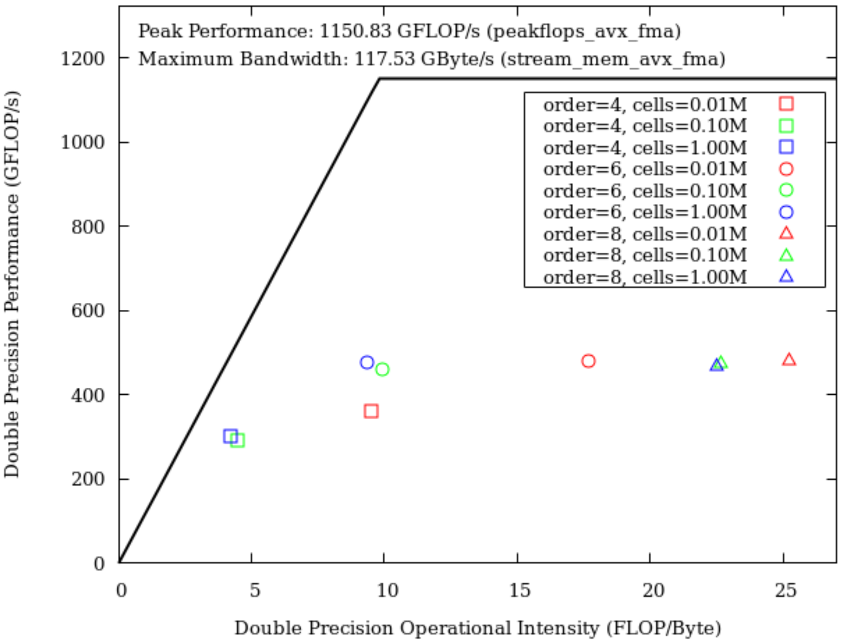


Roofline Models: Elastic vs Acoustic Materials

Elastic ($\mu = 0$):



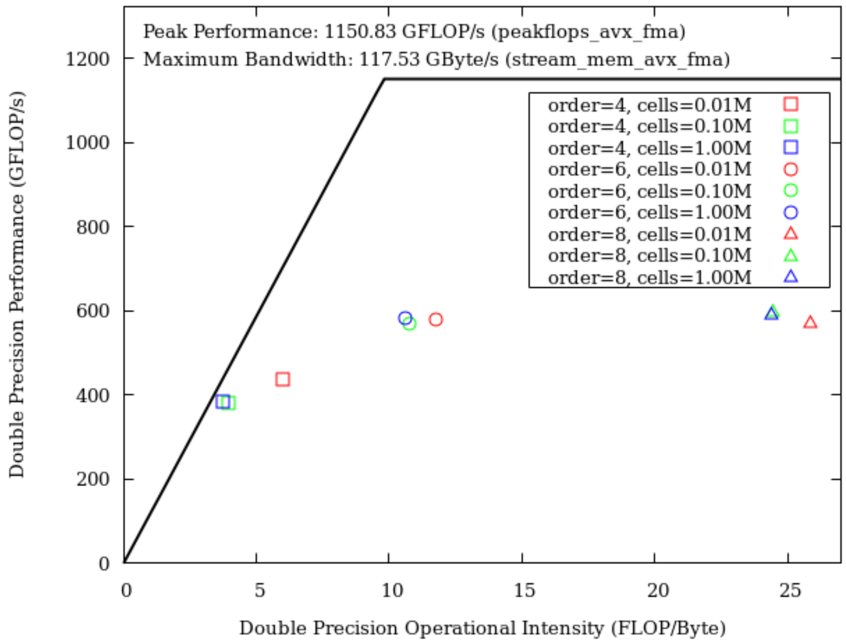
Acoustic:



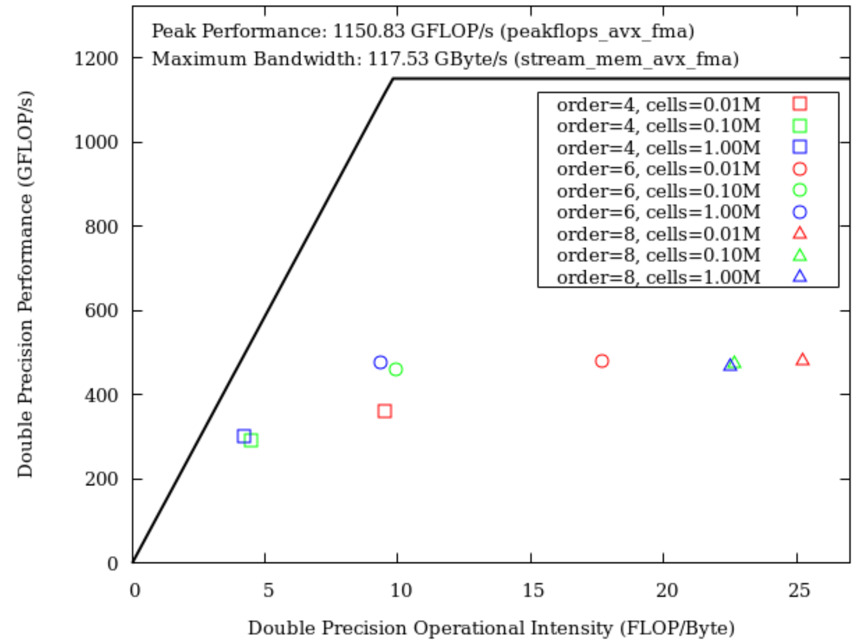
- Is the application compute-bound or memory-bound?
  - Determined by the operational intensity
  - Theoretically, the operational intensity is determined by the code uniquely even without running the program

Roofline Models: Elastic vs Acoustic Materials

Elastic ($\mu = 0$):



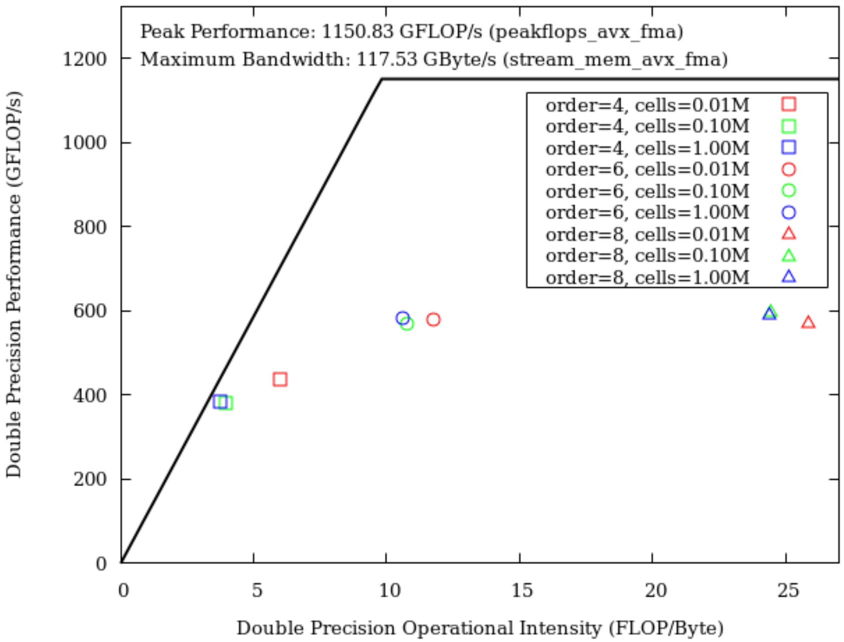
Acoustic:



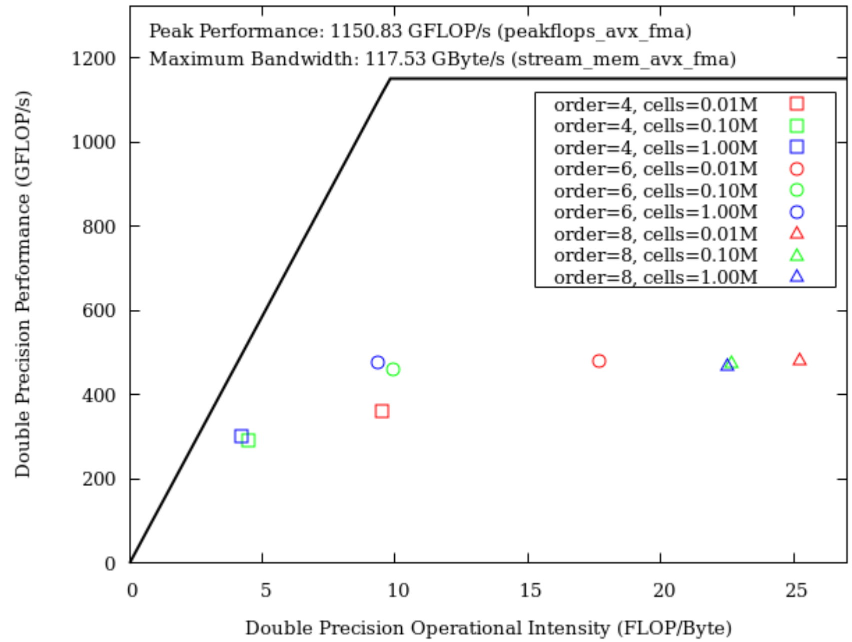
- What is the potential maximum performance for the application, and how to optimize?
  - Moving right: optimizing memory access patterns, reducing unnecessary memory accesses, making better use of caching, efficiently utilizing the cache hierarchy, or compressing data to reduce the amount of data transferred

Roofline Models: Elastic vs Acoustic Materials

Elastic ($\mu = 0$):



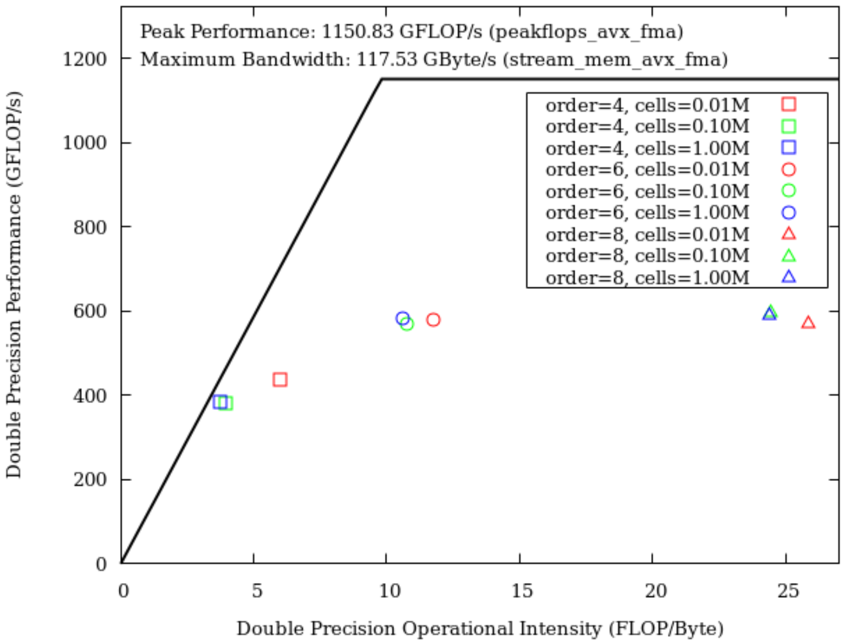
Acoustic:



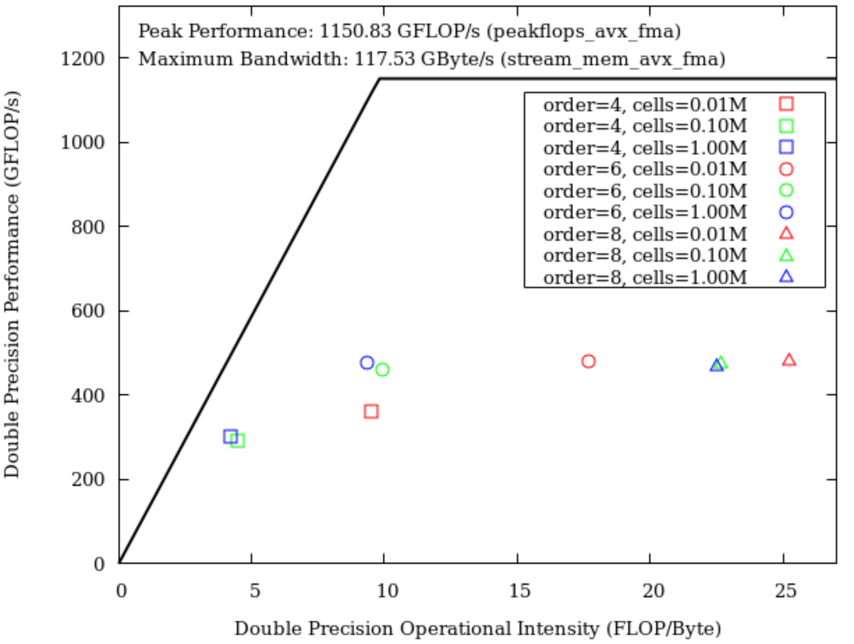
- What is the potential maximum performance for the application, and how to optimize?
  – Moving upward: increasing parallelism, vectorization, using more efficient mathematical algorithms, or leveraging hardware features such as SIMD and FMA instructions to improve computational efficiency

Roofline Models: Elastic vs Acoustic Materials

Elastic ($\mu = 0$):



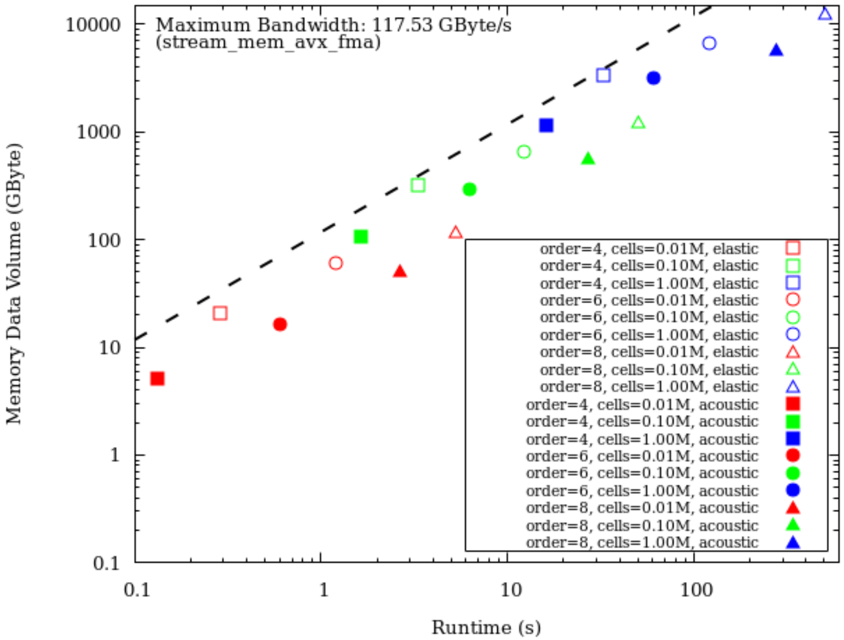
Acoustic:



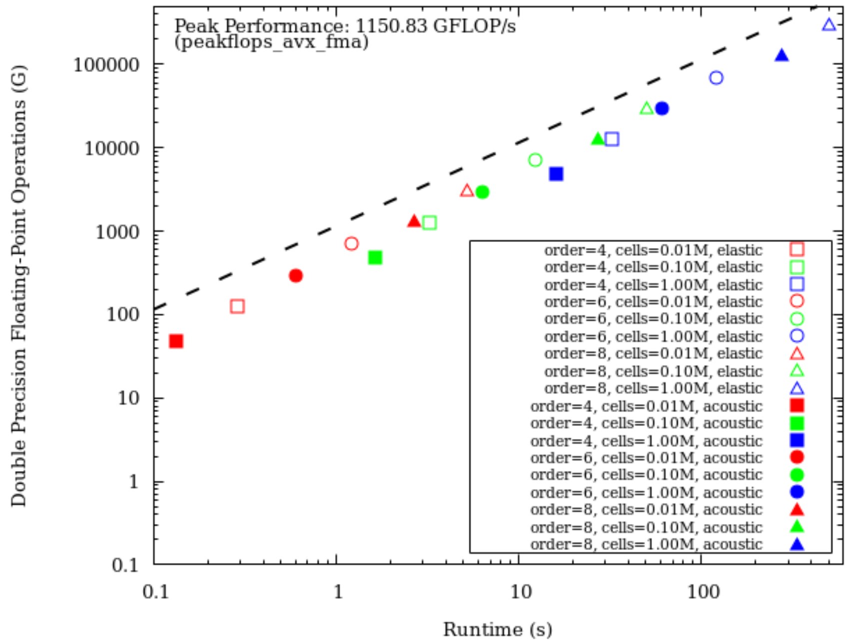
- What is the potential maximum performance for the application, and how to optimize?
  - Upgrading the hardware: memory or processor

Behind Roofline Models

Memory Data Volume vs Runtime:



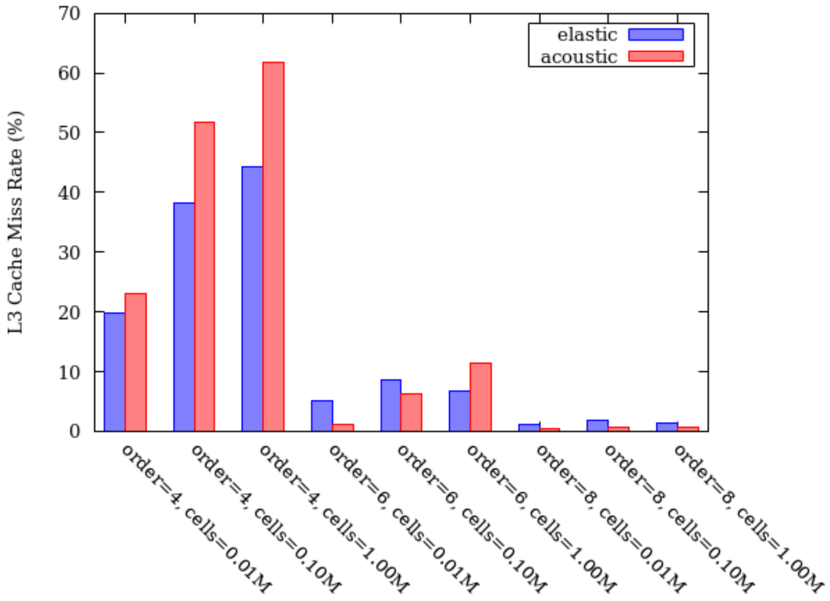
GFLOP vs Runtime:



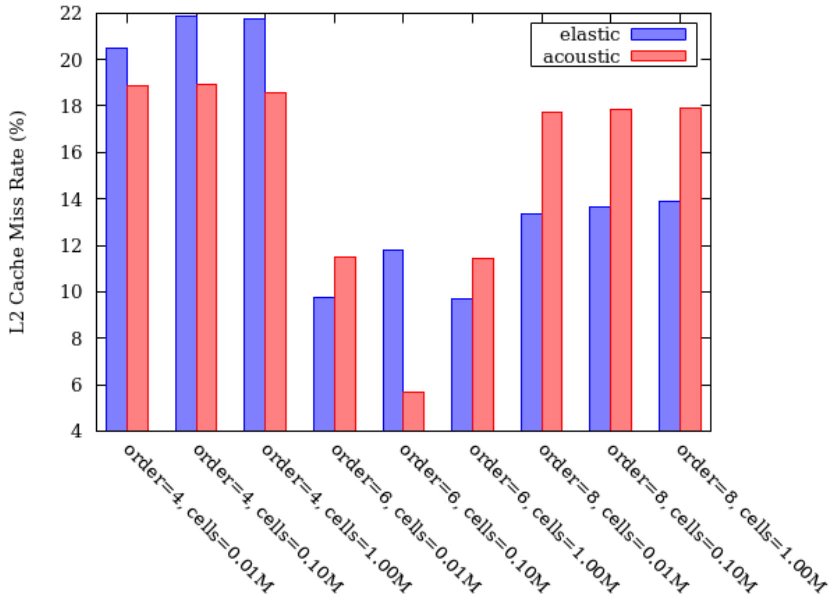
- Memory data volume and runtime are both reduced by approximately half.
- Both FLOPS and memory bandwidth decrease (operational intensity = FLOPS / bandwidth).

L3 and L2 Cache Miss Rates

L3 Cache:



L2 Cache:



Conclusion

- The performance bottlenecks have not significantly changed.
- The memory data volume and runtime have been reduced by approximately half.
- The cache miss rate shows increases and decreases under different test conditions.
- Limitations of roofline model:
  - Simplified assumptions, such as ideal memory access patterns and computational models
  - Missing subtle performance characteristics like memory access latency, cache effects, or thread contention
  - Depending on accurate performance measurements and system parameters
  - Static model
  - Applicability

Future Work

- Implementing the 4-DOF acoustic model for the full version of SeisSol is feasible!
- Performance modeling for distributed memory systems -- 3D “roofline” model?
- Roofline model for heterogeneous computing systems (e.g., CPU+GPU)?

References

Main references for the presentation:

- L. D. S. Krenz, A Fully Coupled Model for Petascale Earthquake-Tsunami and Earthquake-Sound Simulations. PhD thesis, Technical University of Munich, 2024.
- R. J. LeVeque, Finite Volume Methods for Hyperbolic Problems. Cambridge Texts in Applied Mathematics, Cambridge: Cambridge University Press, 2002.
- W. A. Strauss, Partial Differential Equations: An Introduction. New York: John Wiley & Sons Inc, 2nd edition ed., 2007.
- https://seissol.readthedocs.io/en/latest/index.html
- https://doku.lrz.de/linux-cluster-10745672.html
- https://github.com/RRZE-HPC/likwid/wiki/Tutorial%3A-Empirical-Roofline-Model

The full list is presented in my report.

THANKS

Questions? Comments?