SCEC Statewide California Earthquake Center

MPI-based Optimizations for AWP-ODC Seismic Simulation Code

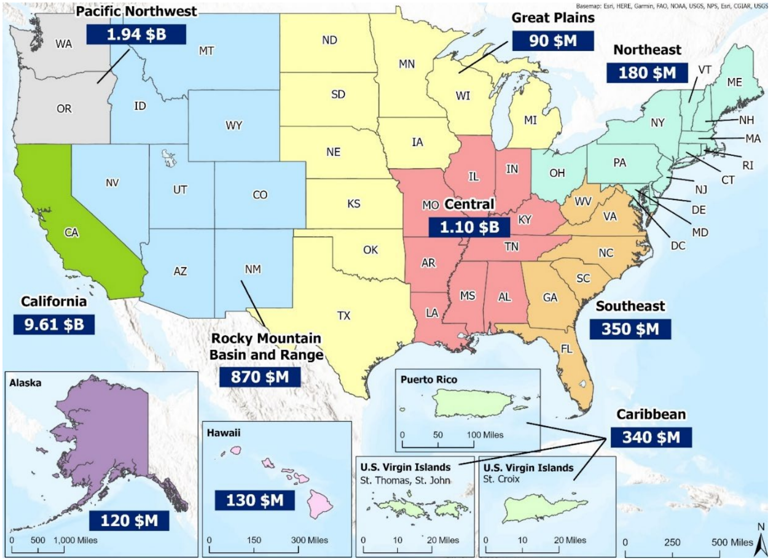
Scott Callaghan (Statewide California Earthquake Center)
Yifeng Cui (San Diego Supercomputing Center)

November 19, 2024

scottcal@usc.edu
yfcui@sdsc.edu

Earthquake Simulations

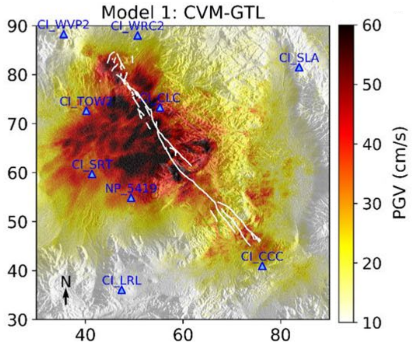
- Earthquakes cause major human impacts
  - Haiti (2021)
  - Türkiye doublet (2023)
  - 10 M7+ earthquakes worldwide in the past year
  - Annualized loss in the US is $14.7 billion/year
- Large earthquakes in well-instrumented areas are rare
  - Difficult to collect useful data
- Experiments only possible on small scales
- Simulation and modeling critical to test hypotheses and improve preparedness



US annualized earthquake loss by region (source: FEMA)

Need for HPC

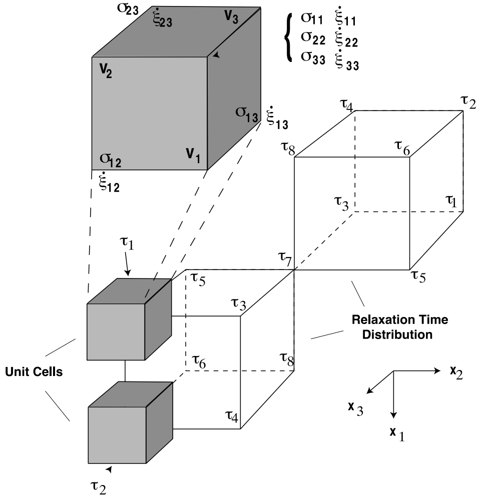
- Wave propagation simulations capture ground motion from earthquakes
- Different kinds of buildings are affected by different frequency motion
  - Resonance is approximately 10/(building height in floors) Hz
- Computational cost scales as the 4th power of frequency
- Other components increase cost too
  - Nonlinear response
  - Topography
- Higher-performing codes can yield more accurate, broadly useful results

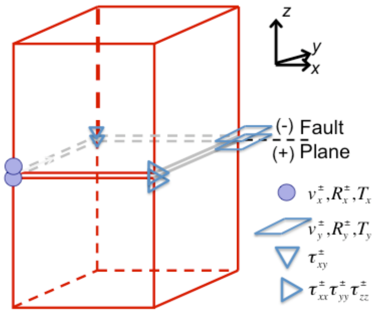


Simulated peak velocity for 2019 M7.1 Ridgecrest earthquake

AWP-ODC

- **A**nelastic **W**ave **P**ropagation – **O**lsen, **D**ay, and **C**ui
- Started as personal research code
- 3D velocity-stress wave equations solved by explicit staggered-grid 4th order finite difference
- Displacement nodes split at fault surface: explicitly discontinuous displacement & velocity
- Absorbing boundary conditions by perfectly matched layers
- Supports dynamic rupture simulations as well



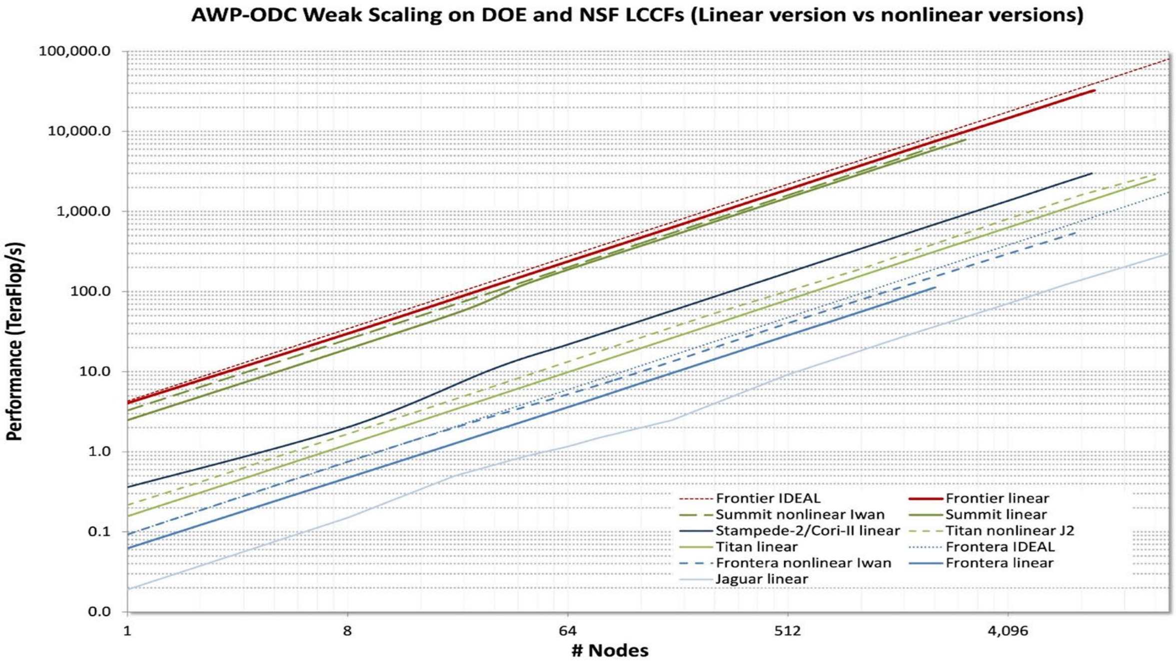


Variables:

$V_i^\pm$	split-node particle velocities
τ_{ij}	stresses
$T_i^\pm$	split-node traction (no jump)
$R_i^\pm$	stress divergence terms

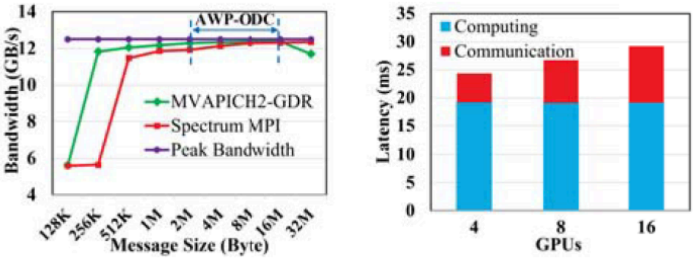
Scalability

- AWP-ODC scales well on leadership-class systems



Motivation for MPI Compression

- Collaboration with DK Panda team at OSU (IPDPS’21 Best Paper Finalist)
- Each AWP-ODC process communicates with (up to) 6 neighbors
- Significant communication times at large scale
- Inter-node bandwidth is often saturated
- Disparity between intra-node and inter-node GPU communication bandwidth impedes efficient scalability



(a) Inter-node D-D Bandwidth (b) AWP-ODC time breakdown

Fig. 2. Motivating Example: production-quality and optimized CUDA-Aware MPI libraries can saturate IB EDR network while the communication time remains a significant bottleneck for HPC applications e.g. AWP-ODC. The message range for AWP-ODC is 2M to 16M as shown in Figure (a).

(Q. Zhou et al. IPDPS’21)

On-the-fly Compression on GPUs

- Designed on-the-fly message compression schemes in MVAPICH2-GDR
- Messages are compressed and combined, then sent and decompressed
- Two GPU compression algorithms integrated into MVAPICH-GDR:
  - MPC: lossless
  - ZFP: lossy
- Overlap compression/ decompression kernels

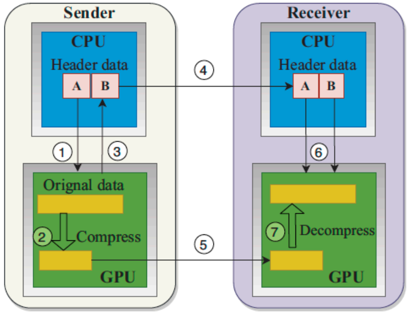


Fig. 4. Data flow of GPU communication with compression. There are seven steps: 1) Launch compression kernel with control parameters 2) Run compression kernel on GPU 3) Returned compressed size 4) Send header data with RTS packet 5) Send compressed GPU data 6) Launch decompression kernel with header data 7) Run decompression kernel to restore the data.

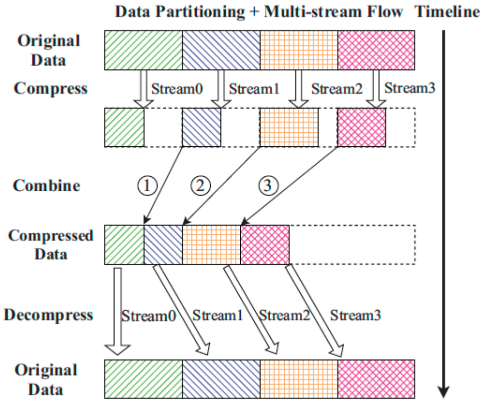
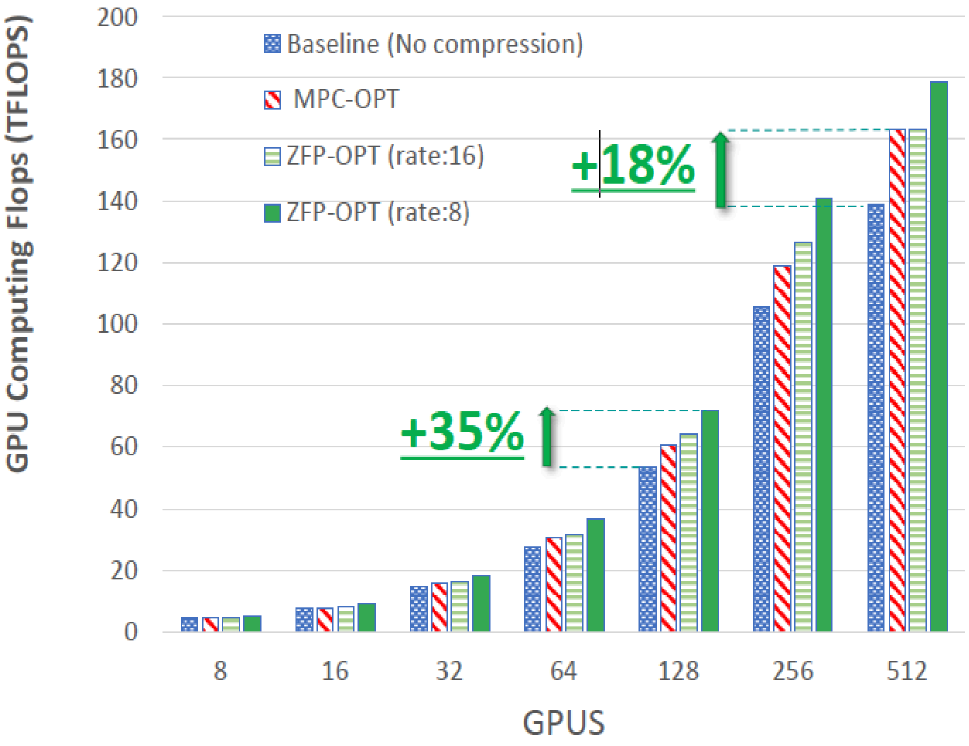


Fig. 7. Data partitioning and multi-stream flow for MPC.

(Q. Zhou et al. IPDPS'21)

MPI Compression Performance Results

- MPC (lossless)
    - 18% increase in flops
    - 15% reduction in runtime
- ZFP (lossy)
    - 35% increase in flops
    - 26% reduction in runtime
- Latency reduced up to 85%



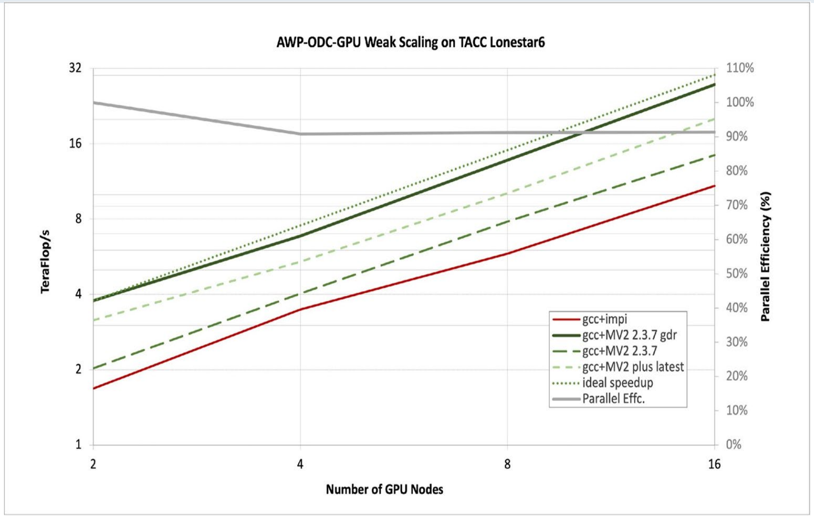
(Q. Zhou et al. IPDPS'21)

Performance Comparisons

- Best performance with MVAPICH2 + GDR + compression (at right)
- ~2x speedup with CUDA-aware MPI + compression (below)
- Promising results with Grace Hopper on TACC *Vista*

AWP-ODC	K20X	KNL7250	V100 (NVLink)	A100 (NVLink)	A100 (PCIe)	A100 (PCIe+Opt)	H100 (PCIe)	H100 (PCIe+Opt)	MI250X (Slingshot)	GH200
MLUPS**	**552**	**1092**	**1598**	**1937**	**896**	**2009**	**3713**	**5145**	**1711**	**8480**
Speedup	**1x***	**1.98x**	**2.89x**	**3.51x**	**1.62x**	**3.64x**	**6.72x**	**9.32x**	**3.10x**	**15.36x**

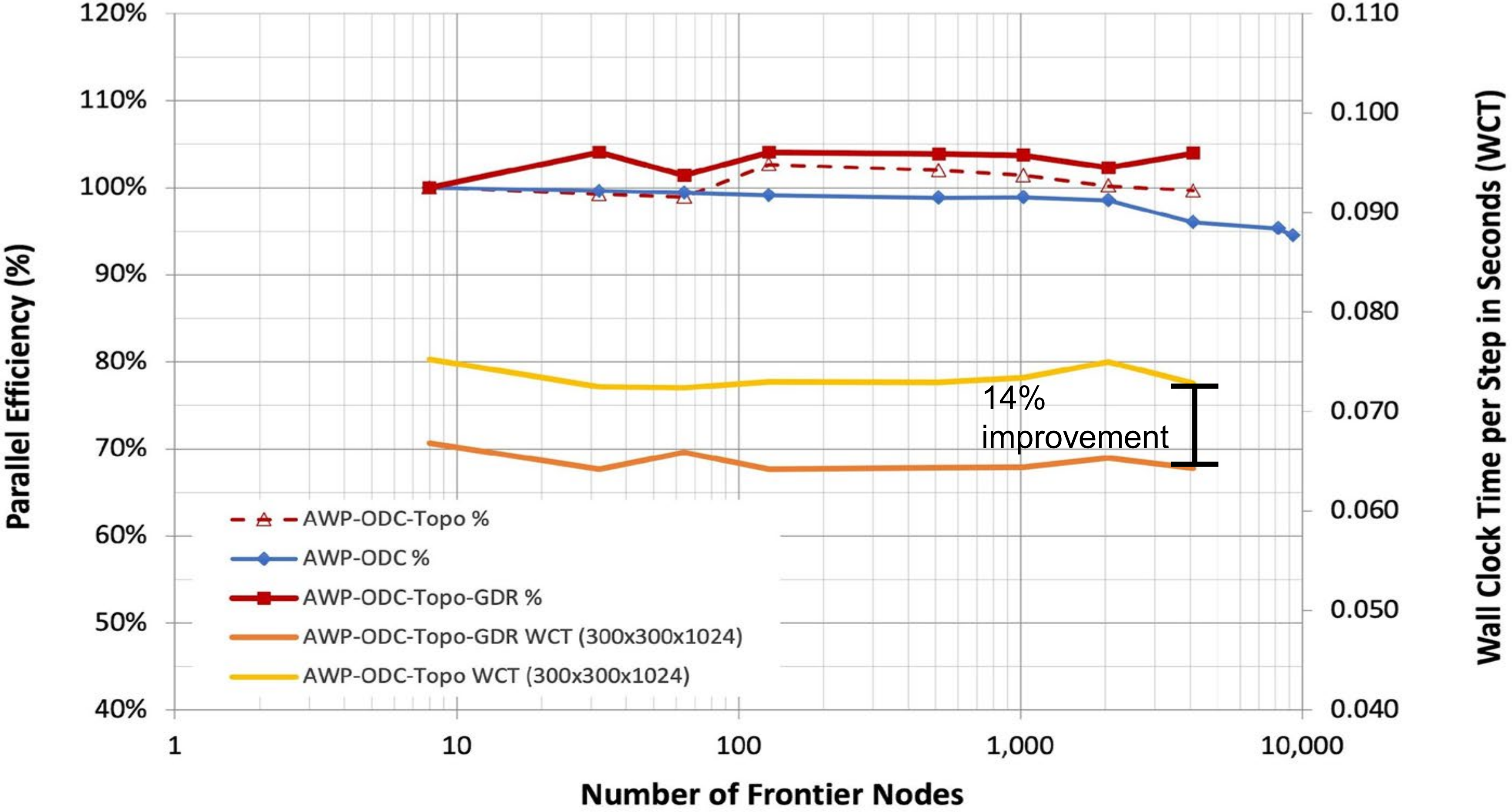
*** 160x160x2048 per GPU configuration ** Millions of lattice point update completed per second**



Lonestar6	mvapich2-2.3.7			mvapich2-2.3.7-gdr			mvapich2-2.3.7-gdr-compresson		
a100	gcc11.2.0			gcc11.2.0			gcc11.2.0		
nodes	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.
2	2.0250	0.0488	100.0%	2.2960	0.0399	100.0%	3.7710	0.0261	100.0%
4	4.0270	0.0494	99.4%	4.5260	0.0436	98.6%	6.8510	0.0288	90.8%
8	7.8250	0.0510	96.6%	9.3250	0.0425	101.5%	13.7560	0.0288	91.2%
16	14.4130	0.1543	89.0%	17.1360	0.0460	93.3%	27.5580	0.0288	91.3%
	impi19.0.9			mvapich2-plus-3.0a2			mvapich2-plus-latest		
	gcc11.2.0			gcc11.2.0			gcc11.2.0		
	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.	Tflop/s	sec/step	parall eff.
2	1.6800	0.0585	100.0%	2.391	0.0411	100.0%	3.151	0.0311	100.0%
4	3.4800	0.0572	103.6%	4.579	0.0431	95.8%	5.399	0.0366	85.7%
8	5.8170	0.0686	86.6%	7.796	0.0509	81.5%	10.136	0.0391	80.4%
16	10.8380	0.0737	80.6%	15.214	0.0523	79.5%	20.097	0.0395	79.7%

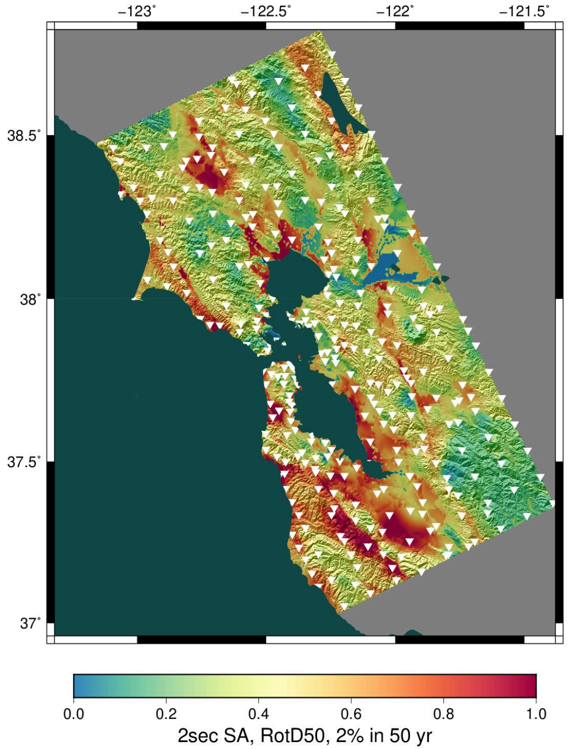
48%-64% improvement using on-the-fly MPC compression over GDR

AWP-ODC-Topo w/ and w/o ROCm-Aware on Frontier



Applications of AWP-ODC

- SCEC CyberShake project uses simulations to improve seismic hazard models in California
  - AWP-ODC used to run wave propagations
  - Ran code 945 times on *Frontier* to create hazard map
  - Represents best available science
- In 2025, code will be used to improve input velocity models and run high-resolution nonlinear scenarios
- Planning to use optimized code for capability runs on LLNL El Capitan



Physics-based seismic hazard map for Northern California

Acknowledgments

HPGeoC Team


Yifeng Cui


Akash Palla


Arnav Talreja


Daniel Roten

Collaborators


Philip Maechling


Scott Callaghan


Kim Olsen


Lars Koesterke

NOWLAB Team


DK Panda


Hari Subramoni


Qinghua Zhou


Lang Xu

Computing Allocation OLCF DD, TACC LSCP and CSA, ACCESS Delta, SDSC Expanse, AMD AAC, DOE INCITE & ALCC

Funding NSF CSSI, LCCF/CSA, NSF/USGS SCEC Core, SDSC