Regional Scale Earthquake Simulations on OLCF Titan and NCSA Blue Waters

Yifeng Cui, SDSC
- a SCEC collaboration
SCEC Collaborators

Thomas Jordan        Kim Olsen        Steve Day           Christine Goulet      Philip Macheling

Computing Resources
NSF Blue Waters, DOE INCITE (Titan, Mira), NSF XSEDE (Stampede, Comet)

Grants/Awards
NSF SI2-SSI, PRAC, SCEC/USGS and XSEDE
Keck Foundation, Intel, NVIDIA
The goal of operational earthquake forecasting is to provide the public with authoritative information on the time dependence of regional seismic hazards

- Thomas H. Jordan
Why choose AWP-ODC?

https://github.com/HPGeoC/awp-odc-os

- Started as personal research code (Olsen 1994)
- 3D velocity-stress wave equations solved by explicit staggered-grid 4th-order FD
- Memory variable formulation of inelastic relaxation using coarse-grained representation (Day 1998)

\[
\sigma(t) = M_u \left[ \varepsilon(t) - \sum_{i=1}^{N} \zeta_i(t) \right] \quad \tau_i \frac{d\zeta_i(t)}{dt} + \zeta_i(t) = \lambda_i \frac{\delta M}{M_u} \varepsilon(t)
\]

\[
Q^{-1}(\omega) \approx \frac{\delta M}{M_u} \sum_{i=1}^{N} \frac{\lambda_i \omega \tau_i}{\omega^2 \tau_i^2 + 1}
\]

solve for \( \frac{\lambda_i}{M_u} \) using linear least squares to fit a target \( Q(f) \) (Withers et al., 2015)

\[
Q(f) = Q_0 \cdot \left( \frac{f}{f_0} \right)^\gamma
\]

- Dynamic rupture by the staggered-grid split-node (SGSN) method (Dalguer and Day 2007)
- Absorbing boundary conditions by PML (Marcinkovich and Olsen 2003) and Cerjan et al. (1985)
**SCEC Computational Pathways**

1. **Earthquake Rupture Forecast**
   - **FM**
   - **DM**
   - **ERM**
   - **PM**

2. **Ground Motions**
   - **AWP**
   - **NSR**

3. **Structural Representation**
   - **DFR**
   - **KFR**
   - **AWP**
   - **F3DT**

4. **Other Data**
   - Geology
   - Geodesy

---

**1 Million core-hours per day!**

**TACC Stampede**
- UCERF3
  - Uniform California Earthquake Rupture Forecast (UCERF3)

**NCSA Blue Waters**
- CyberShake 14.2 seismic hazard model for LA region

**OLCF Titan**
- Dynamic rupture model of fractal roughness on SAF

**ALCF Mira**
- Full-3D tomographic model CVM-S4.26 of S. California
OLCF Titan and NCSA Blue Waters

OLCF’s “Titan” Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla

SYSTEM SPECIFICATIONS:
- Peak performance of 27.1 PF (24.5 & 2.6)
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU (32 GB)
  - NVIDIA Tesla “K20x” GPU (6 GB)
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
Inference Spiral of Earthquake Prediction

- Earthquake system science requires an iterative, computationally intense process of model formulation and verification, simulation-based predictions, validation against observations, and data assimilation to improve the model.

- As models become more complex and new data bring in more information, we require ever increasing computational resources.

(Source: Thomas Jordan, SCEC)
Validating New Physics for High Frequency

**Seismic band**

- **low-order free oscillations**
  - 1000 s (0.001 Hz)
- **mantle waves**
  - 100 s (0.01 Hz)
- **crustal waves**
  - 10 s (0.1 Hz)
- **basin waves**
  - 1 s (1 Hz)
- **strongly scattered waves**
  - 0.1 s (10 Hz)

**Frequency**

- **Period**
  - 1000 s
  - 100 s
  - 10 s
  - 1 s
  - 0.1 s

**Earthquake engineering band**

- **physics-based deterministic**
- **CyberShake 0.5 Hz**

**SCEC simulations**

- 2014
- 2018

**Must validate new physics**

- **fault roughness**
- **near-fault plasticity**
- **frequency-dependent attenuation**
- **Topography**
- **small-scale near-surface heterogeneity**
- **near-surface nonlinearity**

**Physics-based**

- **deterministic**
- **stochastic**

**Empirical**

- **stochastic**

(Withers, 2015)
Validating New Physics for High Frequency

Seismic band

- **low-order free oscillations**
  - period: 1000 s
  - frequency: .001 Hz
- **mantle waves**
  - period: 100 s
  - frequency: .01 Hz
- **crustal waves**
  - period: 10 s
  - frequency: .1 Hz
- **basin waves**
  - period: 1 s
  - frequency: 1 Hz
- **strongly scattered waves**
  - period: 0.1 s
  - frequency: 10 Hz

**Earthquake engineering band**
- **physics-based deterministic**
  - SCEC simulations 2014
- **empirical stochastic**
  - CyberShake 0.5 Hz

Must validate new physics

- **fault roughness**
- **near-fault plasticity**
- **frequency-dependent attenuation**
- **Topography**
- **small-scale near-surface heterogeneity**
- **near-surface nonlinearity**

(Cui et al., 2013)

SCEC simulations 2018

physics-based deterministic

5 Hz

physics-based stochastic

empirical stochastic
Validating New Physics for High Frequency

Fault roughness
Near-fault plasticity
Frequency-dependent attenuation
Topography
Small-scale near-surface heterogeneity
Near-surface nonlinearity

SCEC simulations 2014
Physics-based deterministic
CyberShake 0.5 Hz
Empirical stochastic

SCEC simulations 2018
Physics-based deterministic
5 Hz
Physics-based stochastic
Empirical stochastic

Earthquake engineering band
Physics-based deterministic

(Roten et al., 2016)

SCEC
2014

Seismic band

Low-order free oscillations
Mantle waves
Crustal waves
Basin waves
Strongly scattered waves

Period

1000 s
100 s
10 s
1 s
0.1 s

Frequency

0.001 Hz
0.01 Hz
0.1 Hz
1 Hz
10 Hz

Tall buildings
Houses
Stiff structures

(Cui et al., SC'13, image by Chourasia)
Nonlinear Material Response

In the fault damage zone

- caused by high stresses at rupture front
- leads to flower-like damage zone around the fault

In shallow sedimentary deposits

- caused by hysteretic stress-strain relationship in soft soils
- leads to a reduction in amplification

\[ A(f) = \frac{\mathcal{F}(S(t))}{\mathcal{F}(R(t))} \]

Bonilla et al. (2011)

(Roten et al., 2014)

IBRH16 Vs30 = 626 m/s

Bonilla et al. (2011)

Borehole response

Frequency (Hz)
Return map algorithm in AWP-ODC

Mean stress:
\[ \tau_m = \frac{1}{3}(\sigma_{11} + \sigma_{22} + \sigma_{33}) = \frac{I_1}{3} \]

Stress deviator:
\[ s_{ij} = \tau_{ij} - \tau_m \delta_{ij} \]

Second invariant of stress deviator:
\[ J_2 = \frac{1}{2} \sum_{i,j} s_{ij} s_{ji} \]

Drucker-Prager yield stress:
\[ Y(\tau) = \max\left(0, c \cos \varphi - (\tau_m + P_f) \sin \varphi\right) \]

Drucker-Prager yield function:
\[ F(\tau) = \sqrt{J_2(\tau)} - Y(\tau) \]

Yield factor \( r \):
\[ r = \frac{Y(\tau_{\text{trial}})}{\sqrt{J_2(\tau_{\text{trial}})}} \]

Adjusted stress:
\[ \tau_{ij} = \tau_{m} \delta_{ij} + r s_{ij}^{\text{trial}} \]

Yield factor \( r \) with viscoelastic relaxation time \( T_v \):
\[ r = \frac{Y(\tau_{\text{trial}})}{\sqrt{J_2(\tau_{\text{trial}})}} + \left(1 - \frac{Y(\tau_{\text{trial}})}{\sqrt{J_2(\tau_{\text{trial}})}}\right) \exp \frac{-\Delta t}{T_v} \]

<table>
<thead>
<tr>
<th>method</th>
<th>CPU time</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elastic</td>
<td>0.176</td>
<td>100%</td>
</tr>
<tr>
<td>Individual interpolation (EP1)</td>
<td>0.676</td>
<td>384%</td>
</tr>
<tr>
<td>Yield factor interpolation (EP2)</td>
<td>0.290</td>
<td>165%</td>
</tr>
</tbody>
</table>

Linear + 11 variables \(-\rightarrow\) 5 var \(-\rightarrow\) 3 var

(Barall 2014, Roten et al., 2014, 2015, 2016)
Why choose GPU?

Create a SCEC broadband CyberShake hazard model for all of California

Computational requirements for 1400 sites across California

The statewide CyberShake hazard model will comprise 1.8 billion seismograms

**The CS14.2 study launched on Blue Waters in 2014,**
0.5 Hz deterministic, 2 components

- XE6/XK7 nodes used: 1620, or 49,280 cores
- Jobs submitted: 31,463
- Number of tasks: 470 million
- Storage used: 57 TB
- Allocation hours: 16 M (CPUs + GPUs)

**The CS 15.4 study on BW and Titan in 2015,** 1.0 Hz deterministic, 2 components

- XK7 nodes used: 13,500
- Jobs to submit: 4,372
- Number of tasks: 575 million
- Storage used: 446 TB
- Allocation hours: 13.5 M (GPUs) + 14 M (CPUs)

**The entire CA CS study in plan,**
1.5 Hz deterministic + stochastic, 3 components

- Turnaround: 16 days
- XK7 nodes to use: 17,400
- Jobs to submit: 51,000
- Number of tasks: 1.73 billion
- Storage used: 8 PB
- Allocation hours: 160 million (GPUs)

Go Green
# Flops to Bytes Ratio of AWP-ODC Kernels

<table>
<thead>
<tr>
<th>Three most time consuming Kernels</th>
<th>Reads</th>
<th>Writes</th>
<th>Flops</th>
<th>Flops/ Bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity Comp.</td>
<td>51</td>
<td>3</td>
<td>86</td>
<td>0.398</td>
</tr>
<tr>
<td>Stress-1 Comp.</td>
<td>85</td>
<td>12</td>
<td>221</td>
<td>0.569</td>
</tr>
<tr>
<td>Total</td>
<td>136</td>
<td>15</td>
<td>307</td>
<td>0.508</td>
</tr>
</tbody>
</table>

(Barba & Yokota, SIAM News, 46/6, 2013)
GPU Code: Decomposition on CPU and GPU

- Two-layer 3D domain decomposition on CPU-GPU based heterogeneous supercomputers
  - first step X&Y decomposition for CPUs
  - second step Y&Z decomposition for GPU SMs
Single-GPU Optimizations

✓ Step 2: GPU 2D Decomposition in y/z vs x/y
✓ Step-3: Global memory Optimization
  Global memory coalesced, texture memory for six 3D constant variables, constant memory for scalar constants
✓ Step-4: Register Optimization
  Pipelined register copy to reduce memory access
✓ Step-5: L1/L2 cache vs shared memory
  Rely on L1/L2 cache rather on-chip shared memory

(Zhou, J et al., ICCS 2012)
Blue Waters PAID Project

- **Optimization strategies**
  - Increasing occupancy to hide memory latency
  - Reducing redundant halo accesses by using texture cache combined with register queues

- **Velocity Kernel**
  - Increasing the block size
  - More register queueing, allow read-only array accesses to use texture cache

- **Stress Kernel**
  - Shared memory to optimize accesses to velocity arrays
  - Texture cache with register queuing
  - 65% in DRAM throughput and access fraction of 80% after optimization

- **Plasticity Kernel**
  - Optimized block size
  - Sufficient number of active threads
  - Read-write array accesses replaced with register queues
  - 75% in DRAM throughput and access fraction of almost 100% after optimization

### Execution Time (speedup) unit: ms

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Optimized</th>
<th>Multiple Stream</th>
<th>DRAM Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>dstrqc</td>
<td>65.56</td>
<td>58.957 (x1.11)</td>
<td>58.957</td>
<td>65%</td>
</tr>
<tr>
<td>drprecpc _calc</td>
<td>27.604</td>
<td>18.795 (x1.47)</td>
<td>18.795</td>
<td>75%</td>
</tr>
<tr>
<td>dvelcx</td>
<td>21.295</td>
<td>20.09 (x1.06)</td>
<td>20.09</td>
<td>65%</td>
</tr>
<tr>
<td>other kernels</td>
<td>4.472</td>
<td>4.472</td>
<td>1.922</td>
<td>--</td>
</tr>
<tr>
<td>MPI</td>
<td>6.972</td>
<td>6.972</td>
<td>overlapped</td>
<td>--</td>
</tr>
<tr>
<td>data transfer</td>
<td>10.513</td>
<td>10.513</td>
<td>overlapped</td>
<td>--</td>
</tr>
<tr>
<td>full iteration</td>
<td>136.416</td>
<td>119.799 (x1.13)</td>
<td>99.764 (x1.37)</td>
<td>--</td>
</tr>
</tbody>
</table>

A collaboration with Prof. Wen-mei Hwu of UIUC IME team and Dr. Peng Wang of NVIDIA
Communication Reduction

- Extend ghost cell region with two extra layers and compute rather than communicate for the ghost cell region updates before stress computation.
- The 2D XY plane represents the 3D sub-domain, as no communication in Z direction is required due to 2D decomposition for GPUs.

(Zhou et al., 2013)
## GPU-GPU Communication

<table>
<thead>
<tr>
<th>Communication</th>
<th>Velocity</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Message size</td>
</tr>
<tr>
<td>Before Comm Reduction</td>
<td>4</td>
<td>$6^*(nx+ny)^*NZ$</td>
</tr>
<tr>
<td>After Comm Reduction</td>
<td>4</td>
<td>$12^*(nx+ny+4)^*NZ$</td>
</tr>
</tbody>
</table>
Computing and Communication Overlapping

AWP-ODC-GPU Main Loop:

Do

T = timestep 0 to timestep N:

Pre-Post MPI_IRecv waiting for V1, V2, V3, and V4 of (vx, vy, vz).

Compute V1 for (vx, vy, vz) in GPU

Compute V2/V3 for (vx, vy, vz) in GPU and initiate the transfer of V1 from GPU to CPU

Compute V4/V5 for (vx, vy, vz) in GPU

Compute S5 for (xx, yy, zz, xy, yz, xz) in GPU

Wait for V1/V2 data transfer done and then initiate MPI_ISend for M1/M2

Wait for M1, received MPI message and initiate the transfer of G1 from CPU to GPU

Wait for G1 data transfer done and then initiate the transfer of V3 from GPU to CPU

Wait for V3 data transfer done and then initiate MPI_ISend for M3

Wait for M2, received MPI message and initiate the transfer of G2 from CPU to GPU

Wait for G2 data transfer done and then initiate the transfer of V4 from GPU to CPU

Wait for V4 data transfer done and then initiate MPI_ISend for M4

Wait for M3/M4, received MPI message and initiate the transfer of G3/G4 from CPU to GPU

Compute the rest of stress computation S1-S4 for (xx, yy, zz, xy, yz, xz) in GPU

End Do

(Cui et al., 2013)
Multi-streaming for Computing/Communication Overlap

- Multi-streaming technique to overlap communication with computation
- Small kernels can be optimized by GPU job scheduler
- Linear implementation of AWP-ODC-GPU achieves a parallel efficiency of 100% with 16,384 XK7 nodes
Two-phase I/O Model

• Parallel I/O
  • Read and redistribute 6.9 TB inputs
    – Contiguous block read by reduced number of readers
    – High bandwidth asynchronous point-to-point communication redistribution
  • Aggregate and write
    – Temporal aggregation buffers
    – Contiguous writes
    – Throughput
    – Overlap of source inputs with GPU computation

(Poyraz et al., 2014)
(Roten et al., 2016)
(Cui et al., 2010)
Initialize simulation

Initialize modules

Start computation on GPU

Specified time step?

yes

no

Copy velocity data and signal modules

More time steps?

yes

no

Finalize

Is the signal received?

yes

no

Calculate SGT

Is it time to write out?

yes

no

Write out - MPI-IO

Heterogeneous Computing: API for Pthreads

individual Pthreads make use of CPUs: post-processing

- Vmag, seismograms
- Adaptive/interactive control tools
- Output writing
- Statistics
- In-situ viz
- SGT calculations

(Cui et al., 2013)
AWP-ODC Weak Scaling

Number of XK7/XE6 nodes vs. TFLOPS

<table>
<thead>
<tr>
<th># nodes</th>
<th>Default</th>
<th>Topaware</th>
<th>Speedup</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>920</td>
<td>0.111</td>
<td>0.104</td>
<td>6.1%</td>
<td>93.9%-&gt;100%</td>
</tr>
<tr>
<td>1840</td>
<td>0.111</td>
<td>0.104</td>
<td>5.9%</td>
<td>94.1%-&gt;100%</td>
</tr>
<tr>
<td>3680</td>
<td>0.111</td>
<td>0.104</td>
<td>5.8%</td>
<td>94.2%-&gt;100%</td>
</tr>
</tbody>
</table>

99.5% parallel efficiency

100% parallel efficiency
## AWP-ODC Performance

<table>
<thead>
<tr>
<th>Device</th>
<th>GHz</th>
<th>Mem bwth GB/s</th>
<th>GDDR5/4 (GB)</th>
<th>TFLOP/s (SP)</th>
<th>AWP MLUPS (SP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2090</td>
<td>1.3</td>
<td>177</td>
<td>6</td>
<td>1.33</td>
<td>361</td>
</tr>
<tr>
<td>K20X</td>
<td>0.73</td>
<td>250</td>
<td>6</td>
<td>3.95</td>
<td>552</td>
</tr>
<tr>
<td>Titan-X</td>
<td>1.5</td>
<td>480</td>
<td>12</td>
<td>10.60</td>
<td>1143</td>
</tr>
<tr>
<td>KNL 7210</td>
<td>1.3</td>
<td>460</td>
<td>16</td>
<td>5.32</td>
<td>1110</td>
</tr>
<tr>
<td>ES-2680v3</td>
<td>2.5</td>
<td>120</td>
<td>128</td>
<td>0.48</td>
<td>131</td>
</tr>
</tbody>
</table>

* Nonlinear code
OLCF Summit in 2018

• Hybrid CPU/GPU system delivered in 2017
  – Multiple IBM POWER9 processors and multiple NVIDIA Volta GPUs
  – 3,400 nodes
  – Over 40 TF peak performance
  – More than 512 GB of combined DDR4 and high bandwidth memory
  – Non-blocking fat tree, dual rail EDR-IB (23 GB/s)

• NVLink
  – 160GB/s per GPU bidirectional to Peers
  – 5x-12x PCI-e Gen3 Bandwidth
  – Load/Store access to Peer Memory

• HBM (Stacked) Memory
  – 4x higher bandwidth, ~1 TB/s
  – 3x larger capacity
  – 4x more energy efficient per bit
Extreme-scale Earthquake computing

Sustained SCEC measured performance for a single milestone capability simulation
Summary

- **Science-driven earthquake computational requirements beyond petascale**
  - Large ensembles of CyberShake runs stretch HPC resources across the board

- **Major algorithmic advances needed to engage computing at extreme scale**
  - Accuracy through advanced physics such as near-surface heterogeneities, frequency-dependent attenuation, fault roughness, plasticity, topography
  - Efficiency through scaling and advanced algorithms e.g. ADER-DG, SpecFEM3D
    - Bader talk on Tuesday, 12:25-13:00
    - Komatitsch talk on Wednesday, 11:30-12:05

- **Exascale challenges on heterogeneous systems**
  - Significant investment needed, MPI+X, in re-writing and re-designing algorithms to manage hierarchical parallelism at nodes, cores and threads level, with data locality, heterogeneity and reliability
  - Dealing with billion-way concurrency, strong + weak scaling, and decreased memory bandwidth
  - Inexactness computing for reduced energy consumption that can tolerate a degree of inaccuracy
  - Time-to-solution and energy consumption are the final measures
HPGeoC Supports Earthquake Simulations

Supported by

Dr. Dmitry Pekurovsky
Dr. Yifeng Cui
Dr. Alexander Breuer
Dr. Dawei Mu
Dr. Daniel Roten
Amit Chourasia
Josh Torbin
Hui Zhou, CEA, visiting
Marcus Noack, SRL, visiting
Thank You!