Accelerating Seismic Wave Propagation Simulations for Future Earthquake Science

Y. Cui[1], J. Zhou[1], *E. Poyraz*[1], S. Callaghan[2], D. J. Choi[1], P. J. Maechling[2], T. Jordan[2]

[1]University of California, San Diego

[2]University of Southern California

Blue Waters Symposium May 21-22 2013

NCSA, Illinois

Outline

- Petascale Research in Earthquake System Science PRAC
- SCEC NEIS-P2 Milestones and overview
- Motivation
- Details

SCEC PRAC Project

SCEC researchers are using Blue Waters to simulate a wide spectrum of earthquake and engineering processes including

- dynamic simulations of fault ruptures,
- seismic wave propagation simulation,
- probabilistic seismic hazard calculations, and
- structural response to high frequency ground motions.

To perform physically realistic earthquake and engineering simulations, and complete these simulations within a reasonable time, SCEC's numerical modeling codes must be capable of sustained petaflops performance

SCEC NEIS-P2 Activities

SCEC received NEIS-P2 support to improve the code performance on Blue Waters Cray XE6/XK7 system. Our NEIS-P2 goals were to

- improve scaling applications to large core counts on general-purpose CPU nodes and large scale storage,
- effectively use accelerators,
- use accelerated nodes as a whole with CPUs and GPUs in a single, coordinated simulation and
- enhance application flexibility for more effective, efficient use of systems.

SCEC NEIS-P2 activities were led by Y. Cui at San Diego Supercomputer Center (SDSC). His team includes SDSC, UC San Diego, and SCEC researchers.

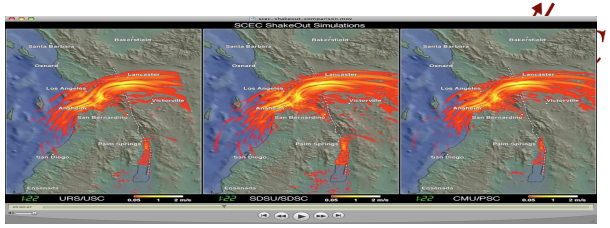
SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems – co-scheduling
   Topology-awareness for AWP-GPU – initial work
5. Preparing production runs for AWP-GPU

AWP-ODC



- Started as personal research code (Olsen 1994)
- 3D velocity-stress wave equations

$$\partial_t v = \frac{1}{\rho}\nabla\cdot\sigma \qquad \partial_t \sigma = \lambda(\nabla\cdot v)\mathbf{I} + \mu(\nabla v + \nabla v^{\mathrm{T}})$$

  solved by explicit **staggered-grid 4th-order FD**

- Memory variable formulation of **inelastic relaxation**

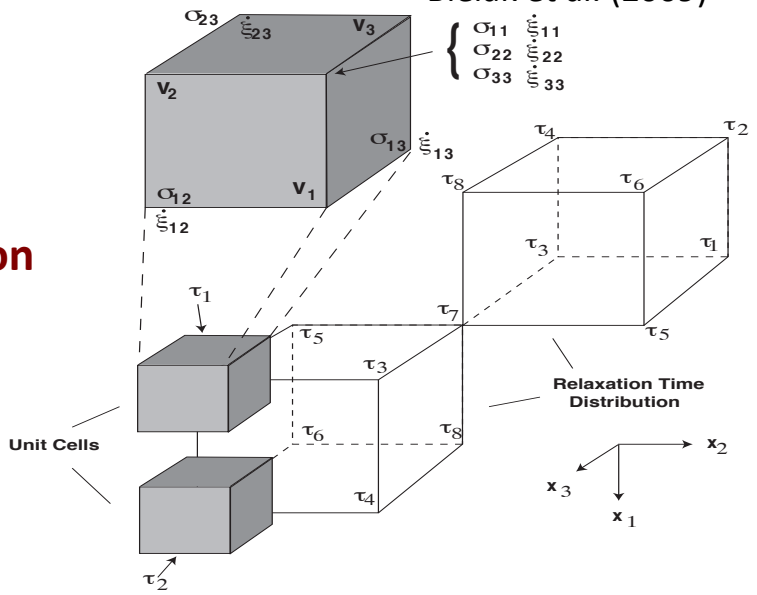
$$\sigma(t) = M_u\left[\varepsilon(t) - \sum_{i=1}^{N}\varsigma_i(t)\right] \qquad \tau_i\frac{d\varsigma_i(t)}{dt} + \varsigma_i(t) = \lambda_i\frac{\delta M}{M_u}\varepsilon(t)$$

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

  using coarse-grained representation (Day 1998)

- **Dynamic rupture** by the staggered-grid split-node (SGSN) method (Dalguer and Day 2007)

Bielak et al. (2009)



Inelastic relaxation variables for memory-variable ODEs in AWP-ODC

- Absorbing boundary conditions by **perfectly matched layers** (PML) (Marcinkovich and Olsen 2003) and **Cerjan et al.** approach

Flops to Bytes Ratio of AWP-ODC Kernels

Three most time consuming Kernels	Reads	Writes	Flops	Flops/ Bytes
Velocity Comp.	51	3	86	0.398
Stress-1 Comp.	85	12	221	0.569
Total	136	15	307	0.508

Flops to Bytes Ratio of AWP-ODC Kernels

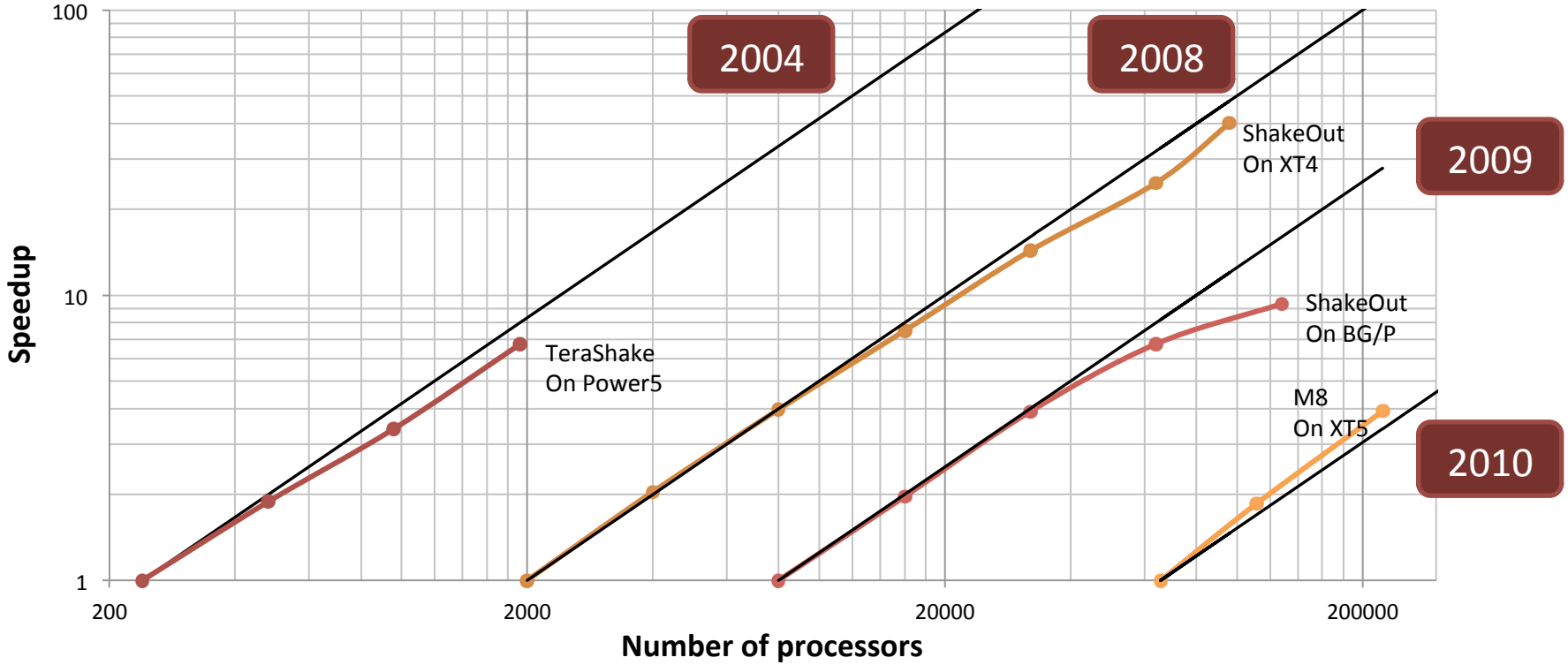
				Flops/ Bytes
Velocity Comp.	31		86	0.398
Stress-1 Comp.	85	12	221	0.569
Total	136	15	307	0.508

Memory bounded

Since Flops/bytes << machine balance (Fermi ~9, Kepler ~20)

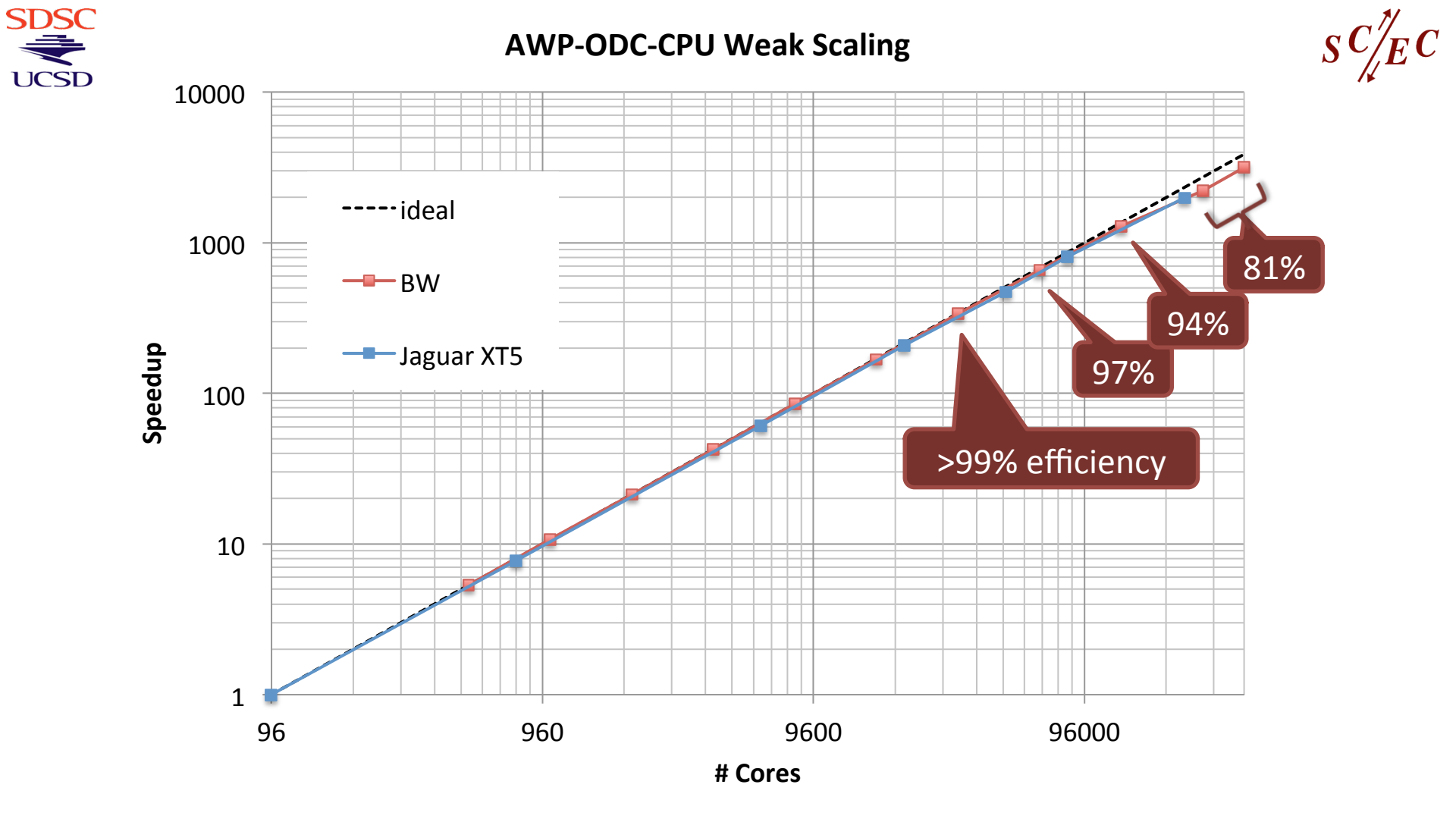
Machine balances are from Igor Podladtchikov, Spectraseis Inc, GTC`13
<http://on-demand.gputechconf.com/gtc/2013/presentations/S3176-Memory-Bound-Wave-Propagation.pdf>

Performance of SCEC Large-scale Earthquake Simulations (2004-2010)



SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

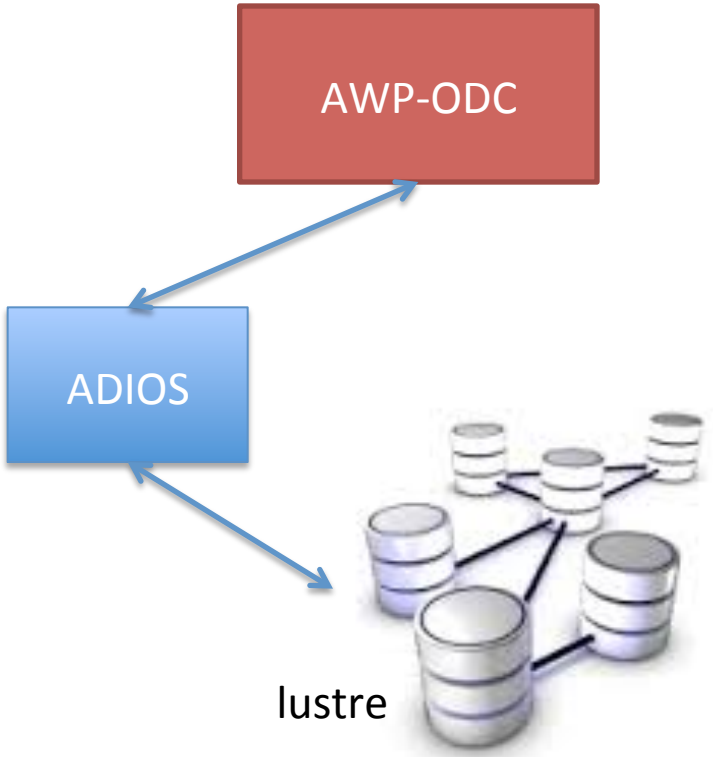

AWP-ODC-CPU Weak Scaling
10000
1000
100
10
1
Speedup
96
960
9600
96000
Cores
ideal
BW
Jaguar XT5
81%
94%
97%
>99% efficiency

SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

ADIOS Checkpointing

- Problems at M8: system instabilities, 50 TB simulation data
- Chino Hills 5Hz simulation is used to validate:
- Mesh size: 7000x5000x2500
- #cores: 87,500 on Jaguar
- WCT: 3 hr
- Total timesteps: 40K
- 3.3 TB simulation data
- ADIOS saved checkpoints at 20Kth timestep and validated the outputs at 40Kth timestep
- Avg. I/O performance: 22.5 GB/s (compared to 20 GB/s with MPI-IO)



Support by Norbert Podhorszki, Scott Klasky, Qing Liu, ORNL

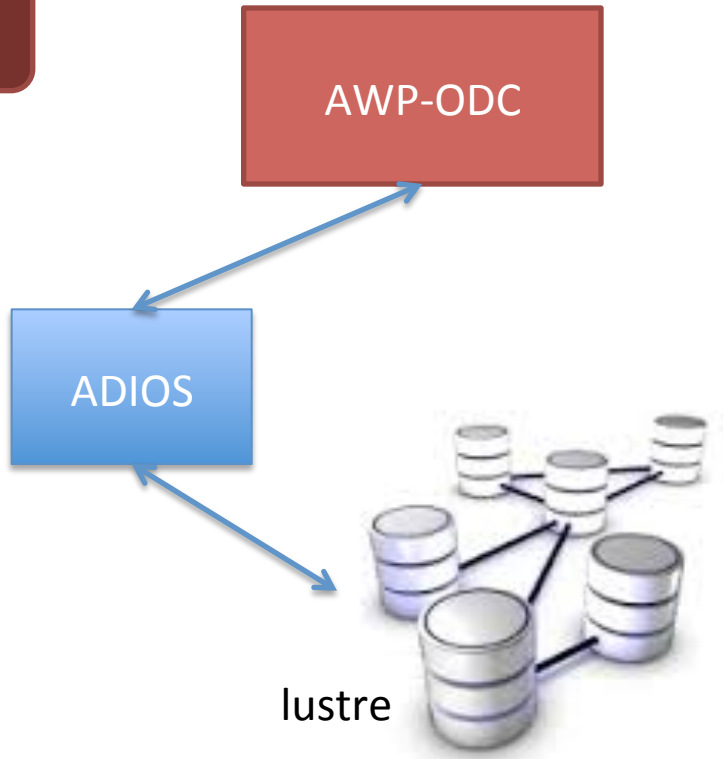
ADIOS Checkpointing

- Problems at M8: system
  simulation data
- Chino Hills 5Hz simulation is used to validate:
  00
- WCT: 3 hr
- Total timesteps: 40K
- 3.3 TB si
- ADIOS s
  nestep and
  validate
  ep
- Avg. I/O performance: 22.5 GB/s (compared to 20 GB/s with MPI-IO)

+ modularity

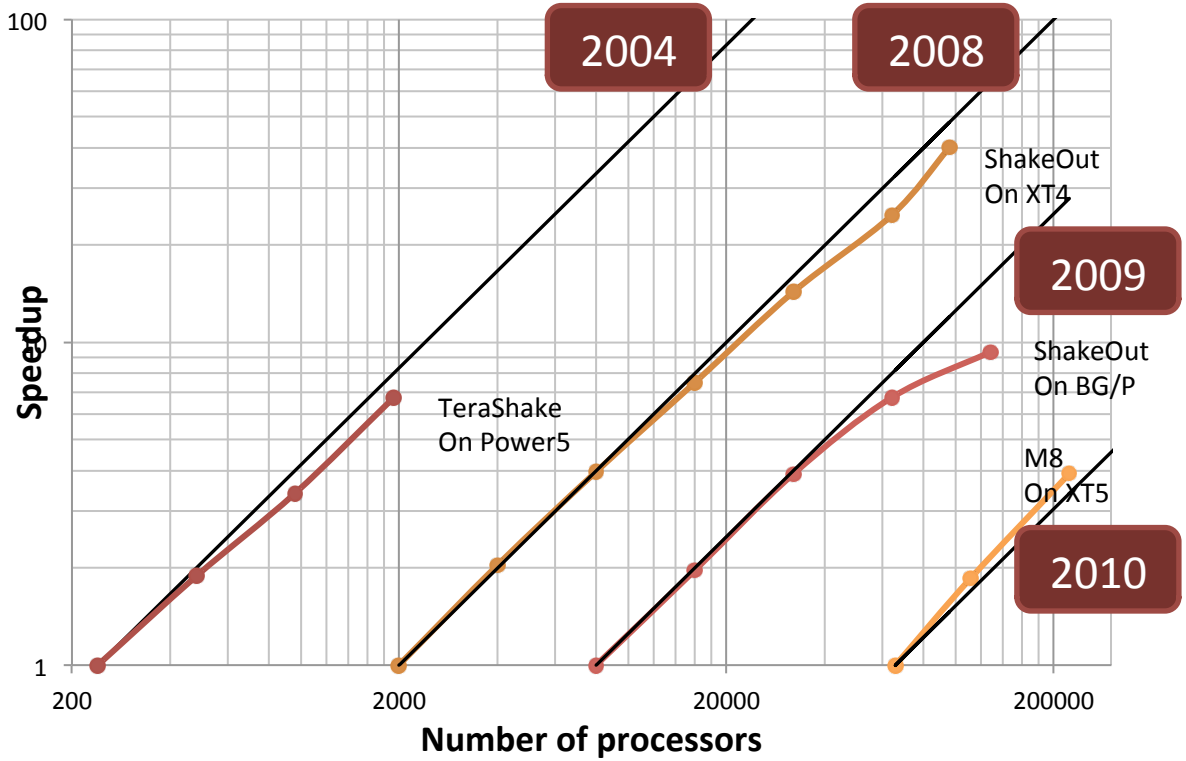
+ IO performance

+ technology-independent



Support by Norbert Podhorszki, Scott Klasky, Qing Liu, ORNL

Why GPUs?


2004
2008
2009
2010
2013
2015
2020
ShakeOut
On XT4
ShakeOut
On BG/P
M8
On XT5
TeraShake
On Power5
100
1
Speedup
200
2000
20000
200000
Number of processors

SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6

   Initial design of CUDA-MPI based AWP-GPU

2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)

   Memory and communication/computation optimizations of AWP-GPU

3. Implementation of MPI-IO on AWP-GPU

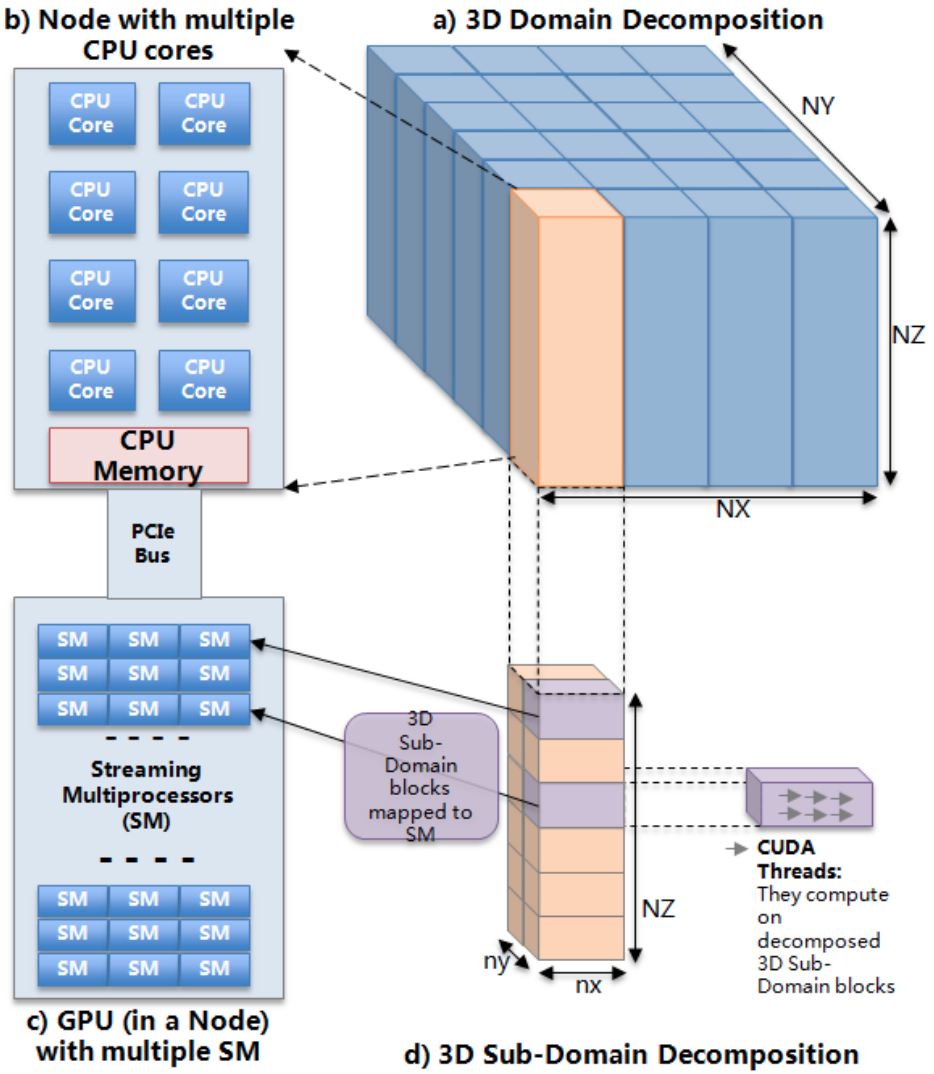
4. Optimization of AWP-GPU for hybrid systems

   Topology-awareness for AWP-GPU

5. Preparing production runs for AWP-GPU

Decomposition on CPU and GPU

- Two-layer 3D domain decomposition on CPU-GPU based heterogeneous supercomputers: X&Y decomposition for CPUs and then Y&Z decomposition for GPU SMs.



Decomposition on CPU and GPU

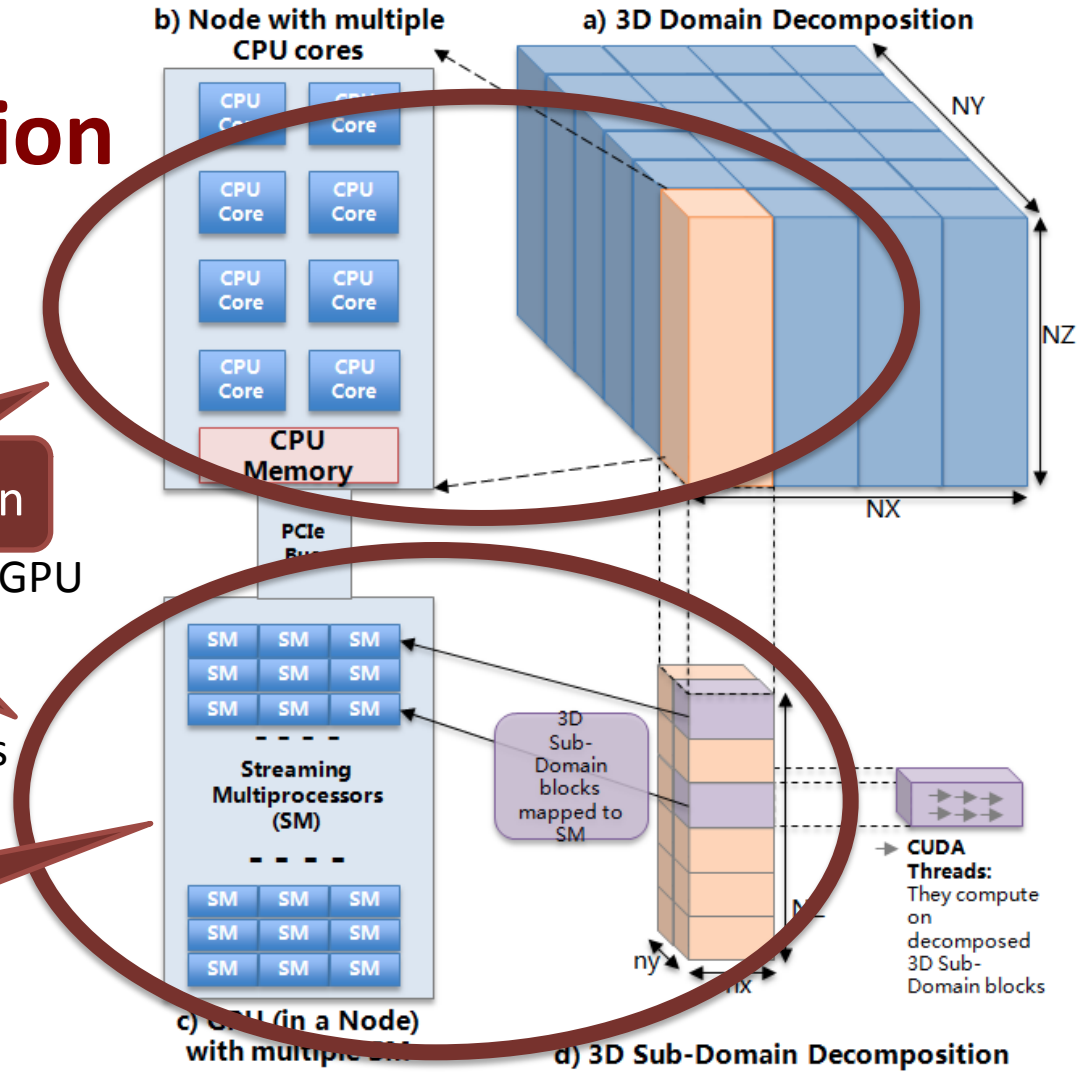
- Tw
  decomposition on CPU-GPU

  decomposition for CPUs and then Y&Z

X&Y decomposition

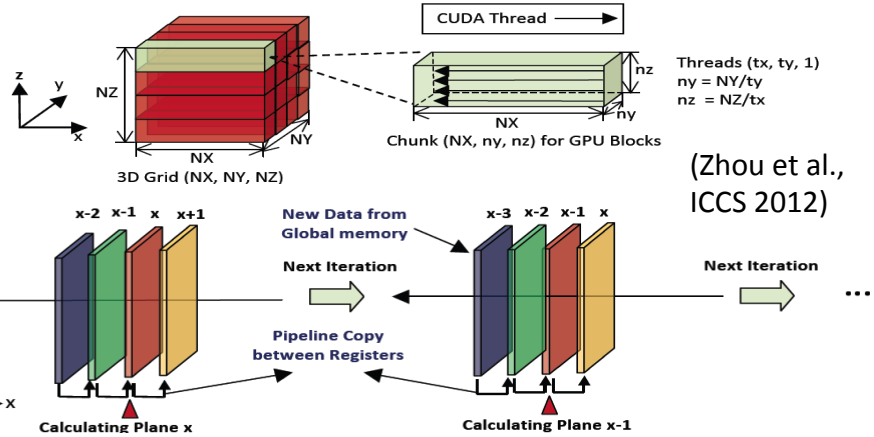
Y&Z decomposition

Computation on GPU only

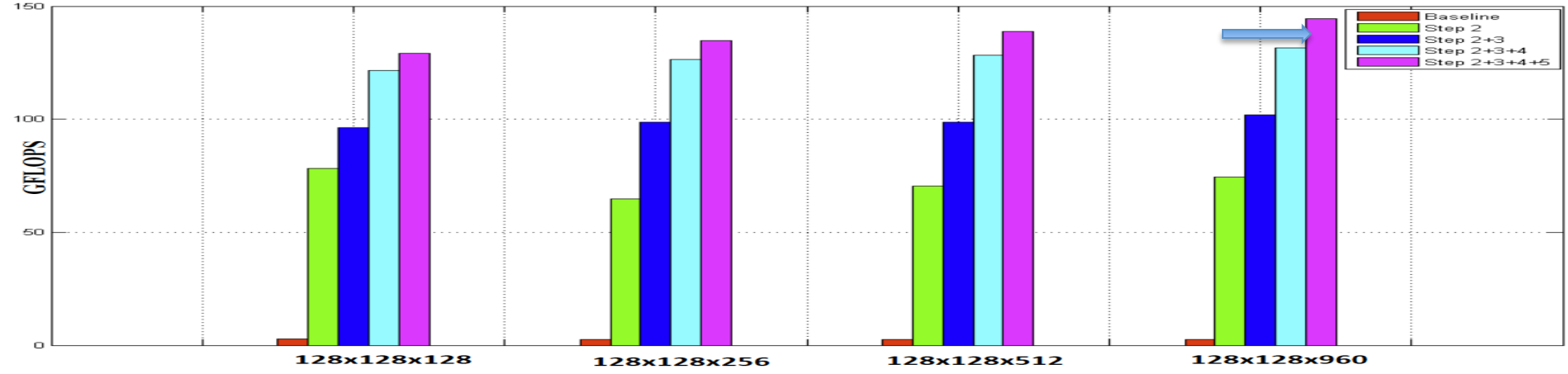


Single-GPU Optimizations

- ✓ **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 et al., ICCS 2012)

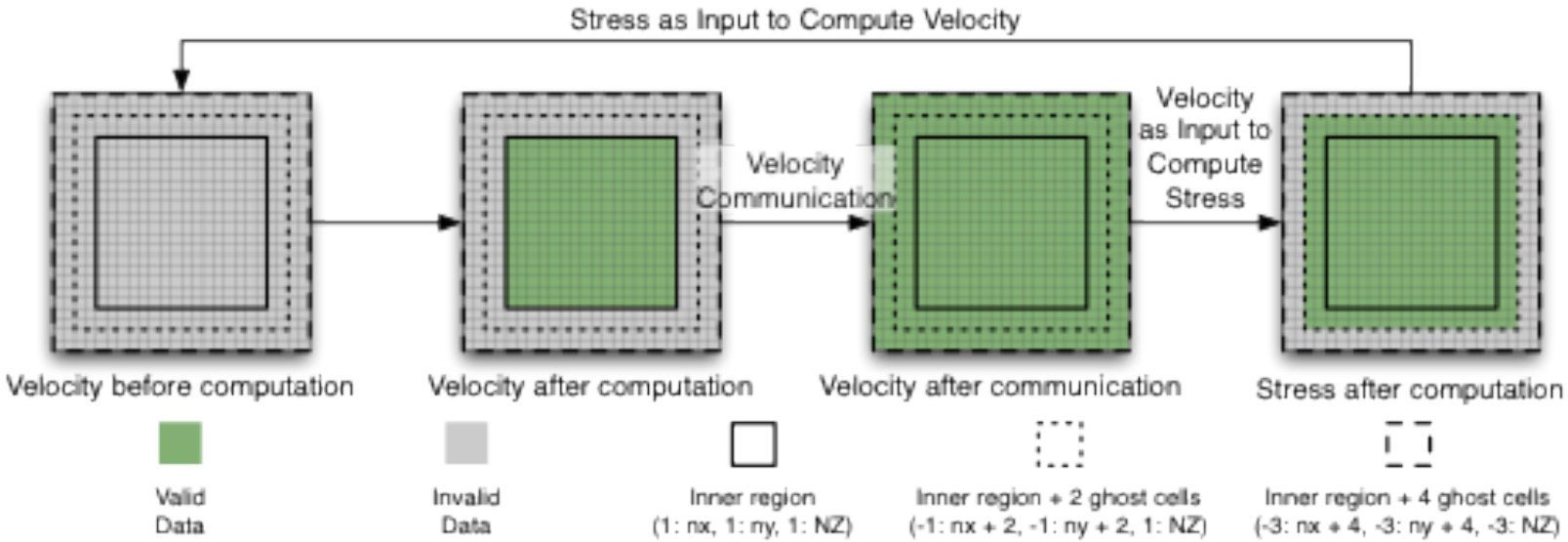


SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

Communication Reduction

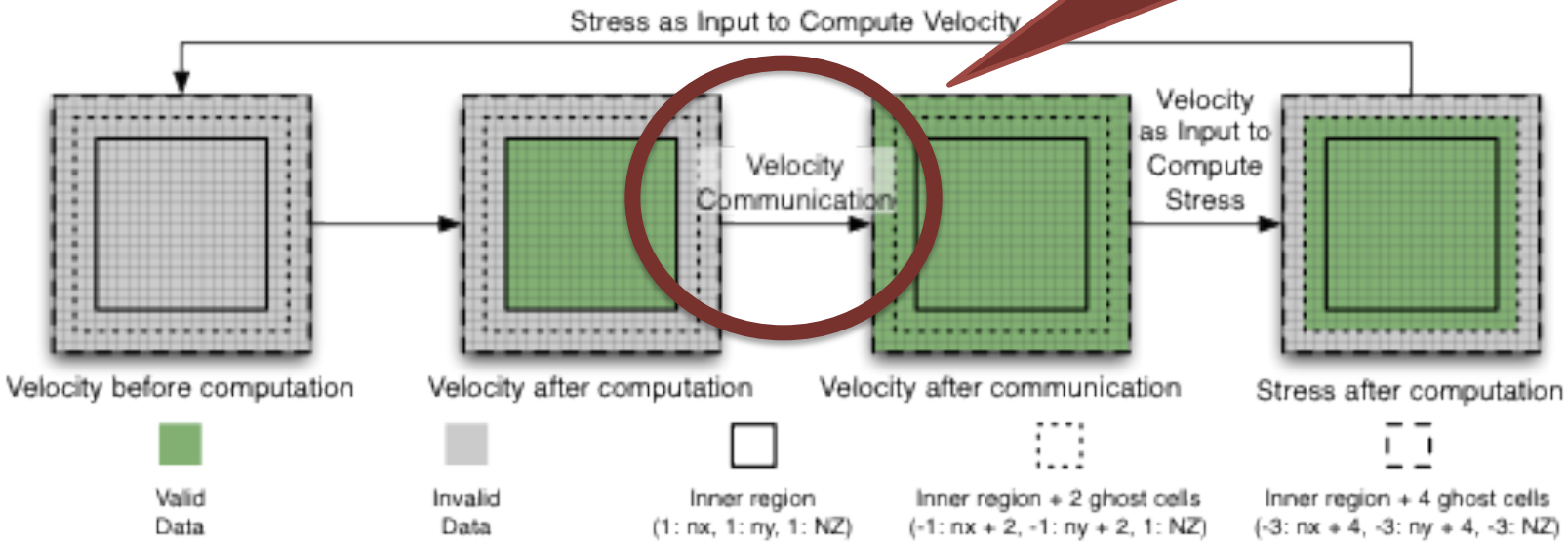
- Extend ghost cell region with two extra layers and compute rather than communicate for the ghost cell region for stress variables.
- View from top (there is no communication in Z direction):



Communication Reduction

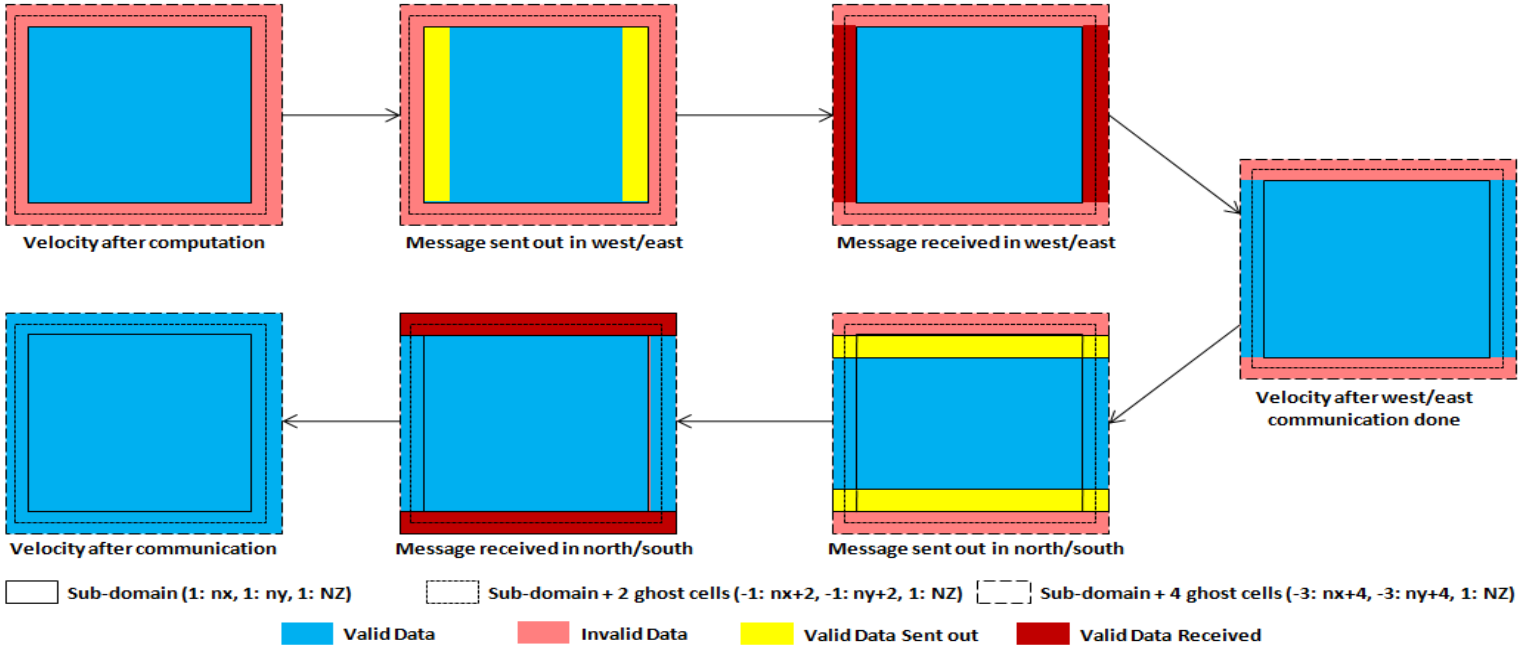
- Extend ghost cell region with two extra layers and c
  communicate for the ghost cell region for stress va
- View from top (there is no communication in Z dire

Communicate only velocity



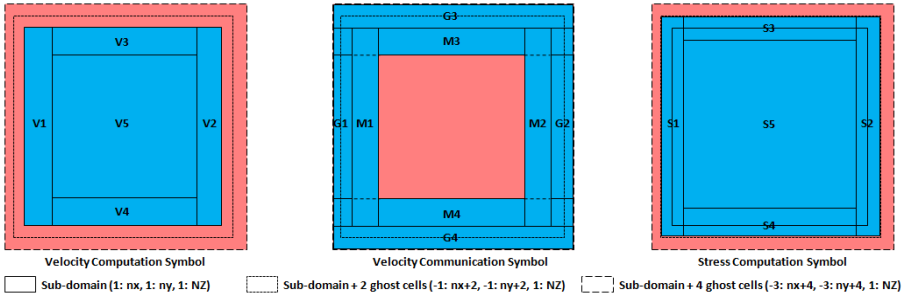
In-order Communication

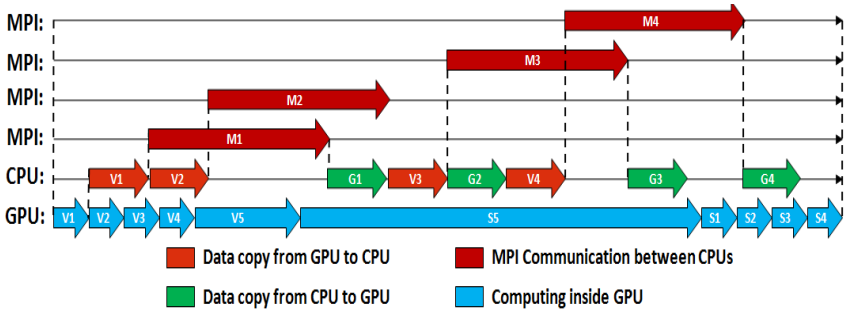
- Communicate in-order for the corners
- This way we do not communicate corners explicitly (save #MPI comms)

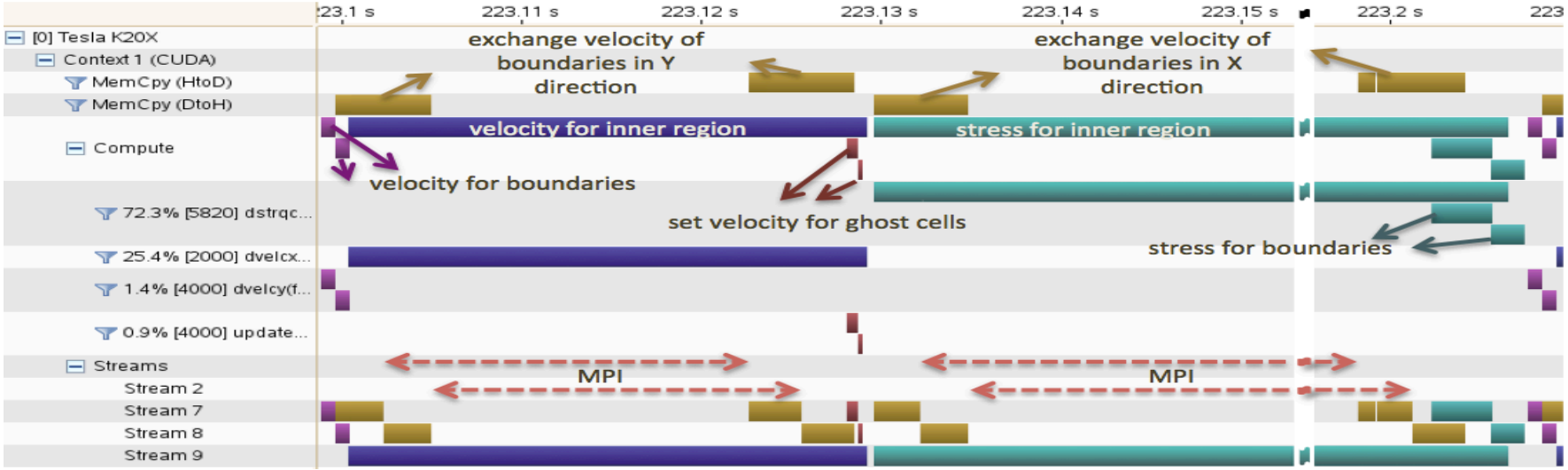


(Zhou, J. et al., ICCS 2013)

Computation and Communication Overlapping



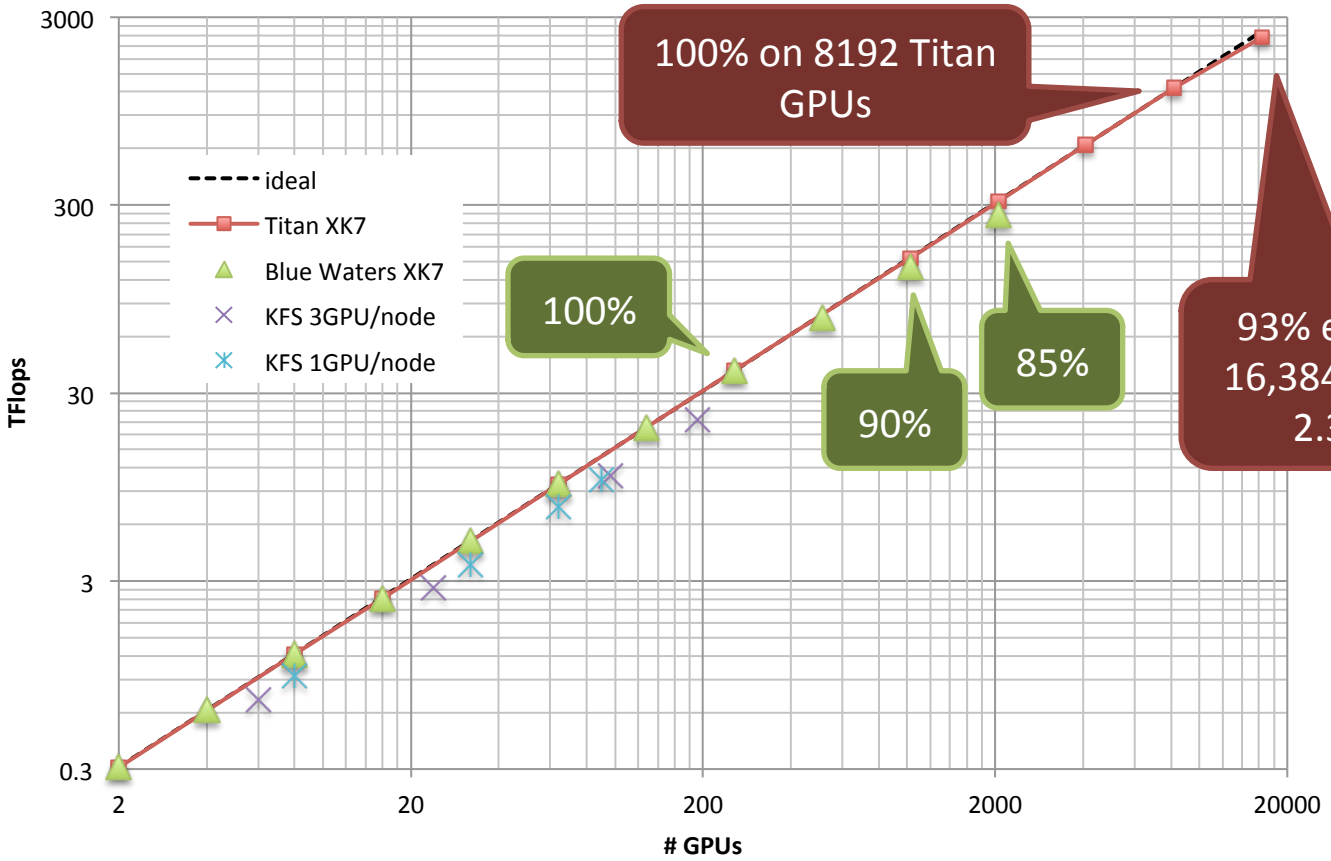




SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

Parallel Multi-GPU Performance

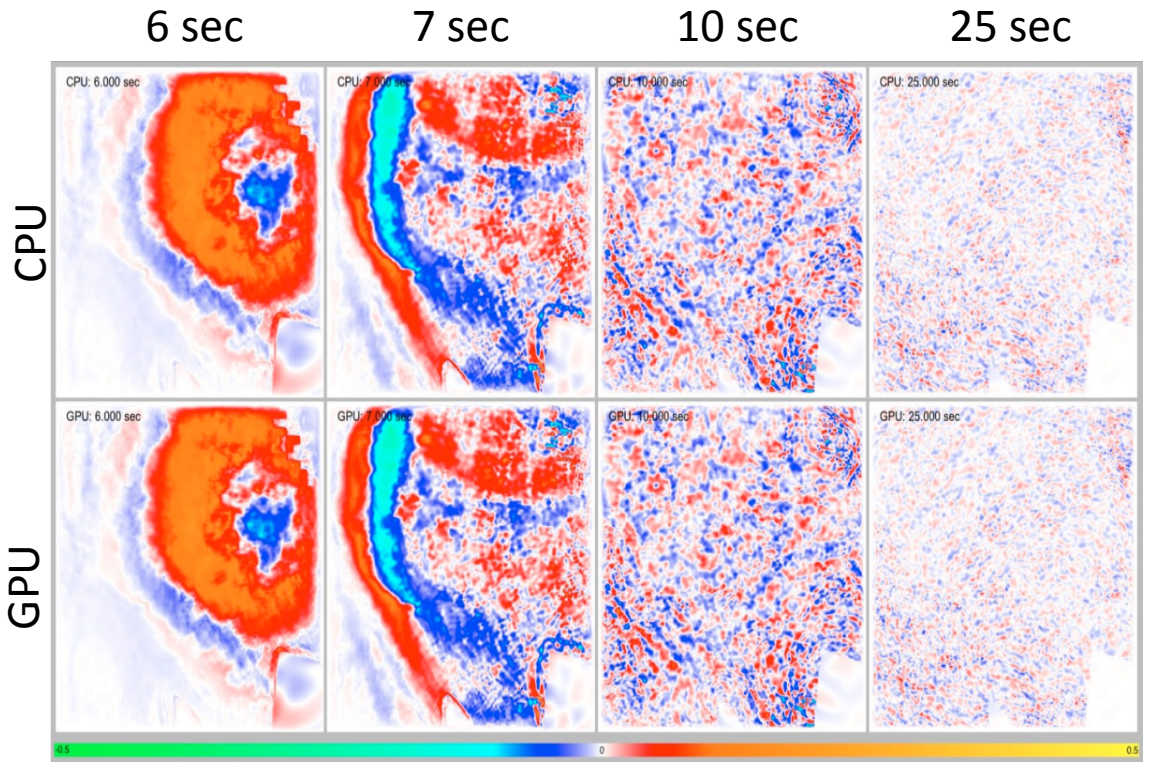


- Weak scaling
- 160x160x2048 sub-domain size

- AWP-GPU on XK7 is 5.2X faster than CPU-only usage of XK7.

Verification

- 2008 Mw5.4 Chino Hills 2.5Hz earthquake simulation performed on Keeneland KFS using 128 GPUs for a mesh of 1024x1024x1024.
- Comparison of visualizations of surface velocity in the X direction in units of m/s.
- The AWP-CPU code was extensively validated.



Science: K. B. Olsen, W. Savran, SDSU; P. J. Maechling and T. Jordan, USC
Keeneland Support: J. Vetter and team
Simulation: E. Poyraz, J. Zhou and Y. Cui, SDSC
Visualization: E. Poyraz,
Map image: Google map provided by T. Scheitlin of NCAR

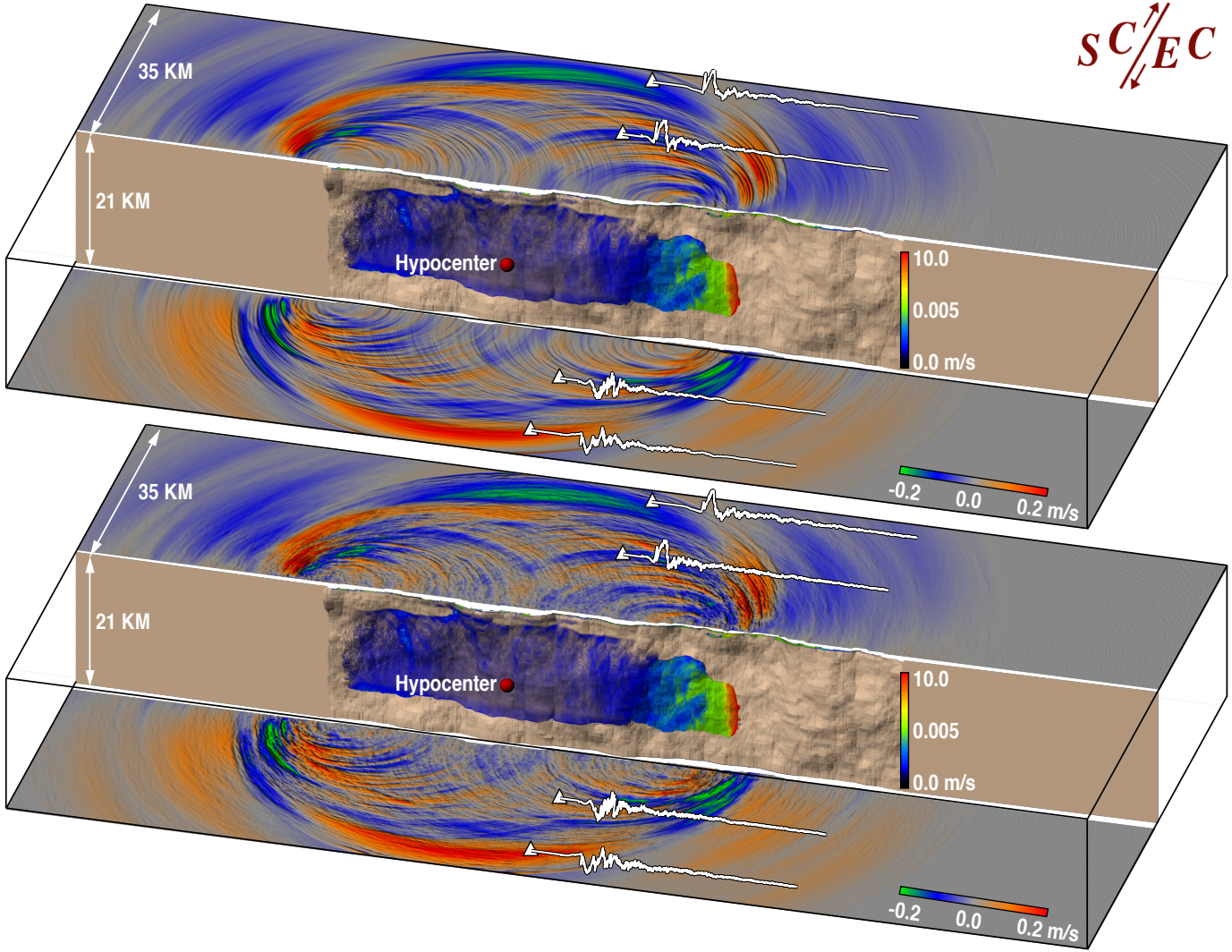
Next Movie Simulated the Same Dataset Wave Propagation for the Mw5.4 Chino Hills, CA, Earthquake, including a Statistical Model of Small-Scale Heterogeneities

Visualization: Tim Scheitlin and Perry Domingo of NCAR
Map image: supplied by Google map
Simulation: Efecan Poyraz and Yifeng Cui of SDSC
Science: Kim Olsen, William Savran, Phillip Maechling and Thomas Jordan of SCEC
Supercomputer: Yellowstone, NCAR


Alhambra
Temple City
El Monte City
Rosemead
El Monte
Vincent
Covina
San Dimas
La Verne
Claremont
Silver Lake
Monterey Park
Garvey
South El Monte
Bassett
West Covina
Los Angeles
Pomona Fwy
Avocado Heights
Valinda
Walnut Valley
Westmont
Pomona
Diamond Bar
East Los Angeles
Montebello
La Puente
South San Jose Hills
Walnut
Rolling Ridge
South Figueroa Corridor
Central Alameda
Commerce
Hacienda Heights
Chino Hills
Huntington Park
Bell
Pico Rivera
Whittier
Rowland Heights
Valle Lindo
Los Serranos
Bell Gardens
East Whittier City
Fullerton
South Gate
Santa Fe Springs
Firestone Scout Reservation
Downey
South Whittier
La Habra
Lynwood
La Mirada
Brea
Chino Hills State Park
Century Fwy
Norwalk
Santa Ana Fwy
Willowbrook
Compton
West Compton
Yorba Linda
Bellflower
Hamilton
Placentia
Gardena Fwy
Buena Park
Fullerton
North Long Beach
Goodyear Blimp Base Airport
Cerritos
Artesia Fwy
Northwest Anaheim
Northeast Anaheim
Riverside Fwy
Anaheim Hills
Lakewood
La Palma
Lakewood Mutual
Carson
Bixby Knolls
Hawaiian Gardens
Southwest Anaheim
Anaheim
South of Conant
Southeast Anaheim
Villa Park
Southeast Torrance
Plaza
Cypress
Signal Hill
West Side
Stanton
Los Altos
Los Alamitos AAF
Orange
Wilmington
Eastside
Garden Grove
Long Beach
Belmont Heights
NWS Seal Beach
Marina
North Tustin
Naples
Westminster
Santa Anita Park
Santa Ana
Tustin

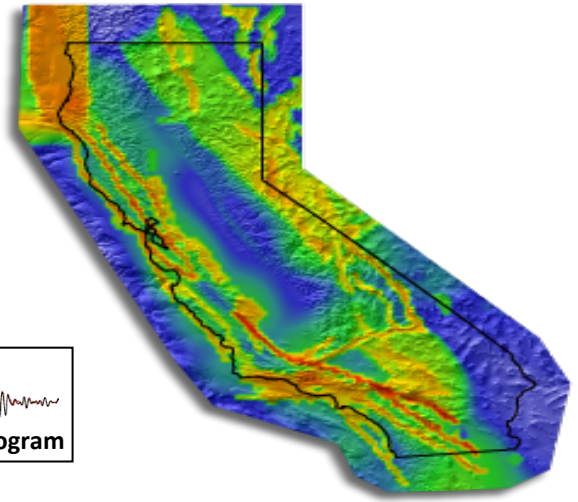
High-F

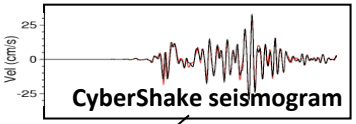
- 10Hz Rock simulation – small on Blue Waters (figure), large on Titan
- Small: 7700x3700x1280 using 1375 nodes with 20m spacing.
- Large: 20800x10400x2048 (443 billion) or 416x208x41 km (20m spacing) using 16640 nodes. 5.5 hours for 170 seconds of simulation.

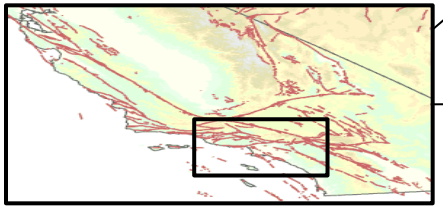


SCEC CyberShake Statewide Hazard Model

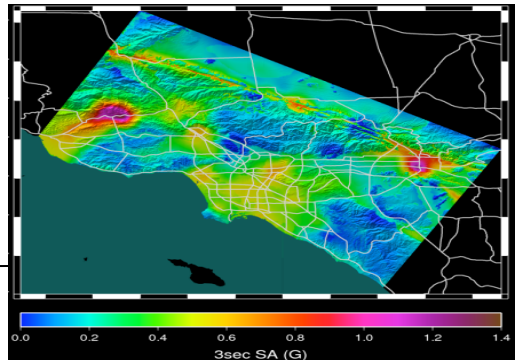
- California statewide wave propagation-based PSHA map up to 1 Hz planned
- 4240 sites, f < 1.0 Hz to be computed using AWP-Graves and AWP-ODC codes
  - 662 million CPU hrs
  - 3.6 billion jobs
  - 48 PB of total output data
  - 2.6 PB of stored data
  - 77 TB of archived data







(Source: SCEC)



LA region

CyberShake1.0 hazard map PoE = 2% in 50 yrs

AWP-GPU-based CyberShake SGT calculator: AWP-SGT

CyberShake 3.0	CPU (XE6)	GPU (XK7)	CPU+GPU** (XK7)
Number of nodes	400	400	400
Wall-clock-time per site (hr)	**10.36**	**2.80**	**2.80**
SUs charged	662M	168M	168M
Saved SUs		**494M**	**578M**

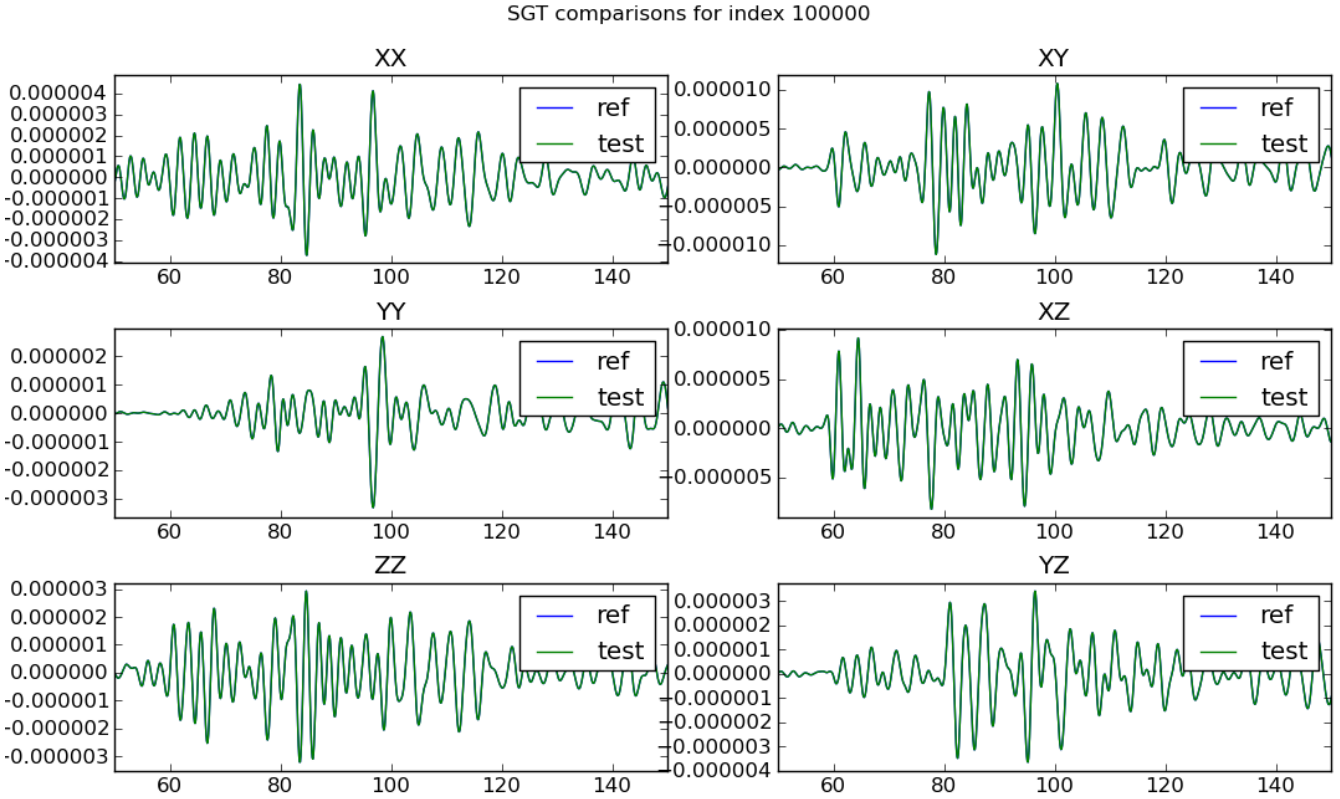
3.7X speedup

* Assuming 5000 site and two strain Green tensor runs per site

** Estimated involving CPU-only post-processing: 6.2M rupture variation calculations per site

Verification of AWP-SGT

- Mesh of 3100x2100x200 using SCEC Community Velocity Model
- On Blue Waters 50 GPUs for 35 mins, 20K timesteps.
- Reference is tested many times before.



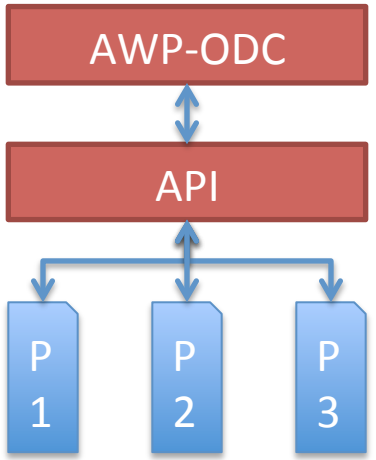
SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

Hybrid Approaches

- Co-scheduling of AWP-SGT
  - Co-schedule AWP-SGT wave propagation simulation (parallel multi-GPU code) and reciprocity-based seismogram and intensity computations (many single-CPU jobs)
  - Run multiple MPI jobs on compute nodes using Node Managers (MOM)
  - Will be able to hide half of 1000 CPU-hr post-processing jobs with 60 GPU-hr AWP-SGT simulation using other available 15 CPU cores on XK7 (1 CPU is used by the GPU code)
- AWP-API lets individual pthreads make use of CPUs: post-processing (Vmag, SGT, seismograms), statistics (real-time performance measuring), adaptive/ interactive control tools, visualization
  - Output writing is introduced as a pthread that uses the API

CyberShake 3.0	GPU (XK7)	CPU +GPU** (XK7)
Number of nodes	400	400
WCT per site (hr)	**2.80**	**2.80**
SUs charged	168M	168M
Saved SUs	**494M**	**578M**

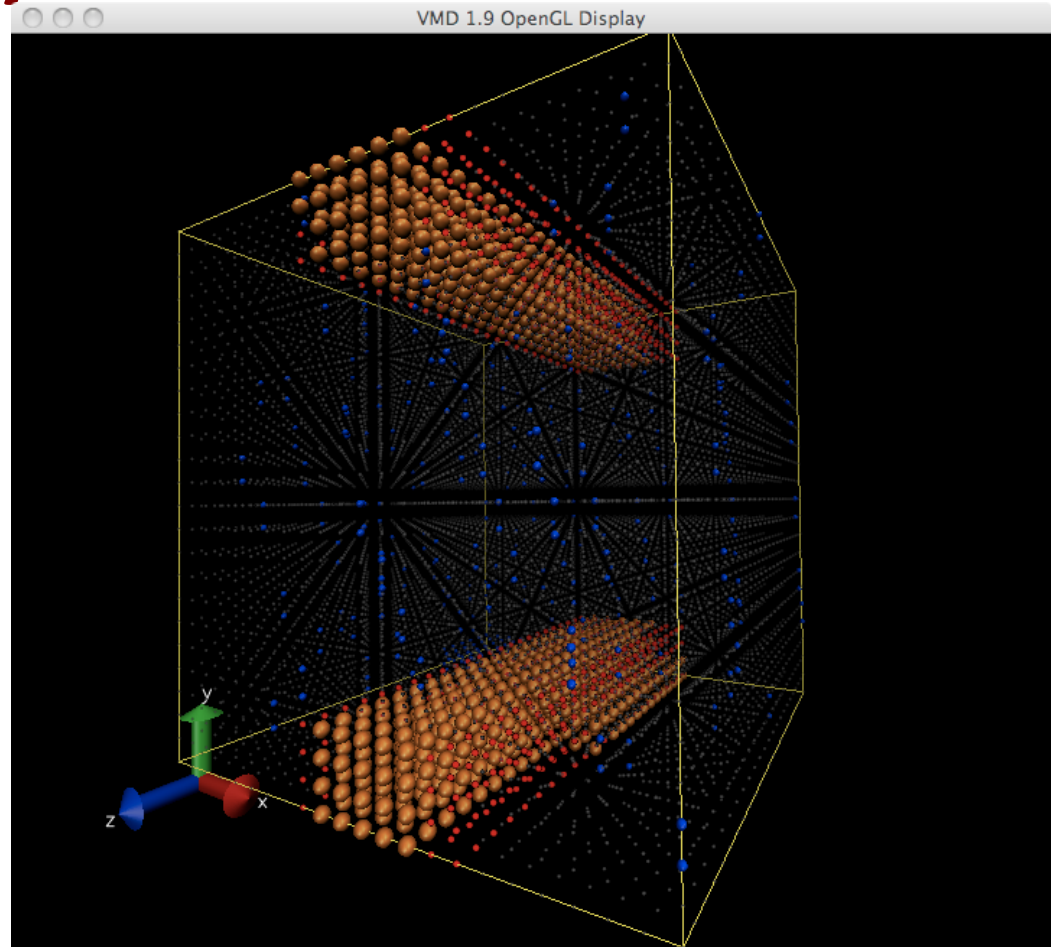


SCEC NEIS-P2 Milestones:

1. Benchmark of AWP-CPU on Blue Waters XE6
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing)
   Memory and communication/computation optimizations of AWP-GPU
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems
   Topology-awareness for AWP-GPU
5. Preparing production runs for AWP-GPU

Topology-Awareness

- Top-aware applications?
- Already tried in AWP-CPU (localized MPI rank distribution)
- Titan topology: 25x16x24 (at the time the largest 3D block was 20x16x19=12160 nodes), slower bandwidth in Y direction.
- Blue Waters network is shared with XE nodes, hence more contention.
- Observed ~5% improvement in performance using more sub-domains in X direction than Y both on Titan and Blue Waters.



Support by Matthew Norman, Ramanan Sankaran, ORNL and John Levesque, Robert Fiedler, Cray

Image and support by Jay Alamada, Gregory Bauer, Omar Padron at NCSA

Conclusions

1. Benchmark of AWP-CPU on Blue Waters XE6: 94% efficiency up to 131K cores and 81% at 373K cores.
   Initial design of CUDA-MPI based AWP-GPU
2. Fault tolerance capability of AWP-CPU (ADIOS checkpointing): IO performance beats MPI-IO with 22.5 GB/s.
   Memory and communication/computation optimizations of AWP-GPU: Perfect scaling up to 8K nodes and 93% efficiency up to 16K nodes achieving 2.33 Pflops. 5.2X speedup is achieved compared to CPU-only usage of XK7 nodes.
3. Implementation of MPI-IO on AWP-GPU
4. Optimization of AWP-GPU for hybrid systems: AWP-SGT will save 578M SUs using co-scheduling (494M SUs without co-scheduling) to make use of all the available resources on hybrid systems.
   Topology-awareness for AWP-GPU: May improve scaling performance for the system-scale runs.
5. Preparing production runs for AWP-GPU: AWP-SGT is developed for CyberShake and achieved 3.7X speedup compared to CPU code. Verification of AWP-GPU is done on Keeneland with Chino Hills 2.5Hz ground motion simulations. 10Hz Rock simulation is done on Titan and Blue Waters and achieved sustained petaflops on Titan.

Acknowledgements

- **The GPU implementation team:**
  - **Jun Zhou, Efecan Poyraz, Dongju Choi and Yifeng Cui**
- **This work is supported by:**
  - XSEDE ECSS Advanced Support for Applications OCI-1053575
  - NCSA Direct PRAC Support Funding
  - SCEC 2012 Core Program, Geoinformatics (EAR-1226343) and SEISM (OCI-1148493).
- **Implementation and tests were carried out on six M2050 and M2090 GPUs donated by NVIDIA. Benchmarks are performed on XSEDE Tesla M2090 Keeneland, NCSA Blue Waters Tesla K20X, NCCS TitanDev M2090 and Titan K20X systems**
- **Thanks to:**
  - **Kim Olsen, San Diego State University**
  - **Philip Maechling and Thomas Jordan, USC/SCEC**
  - Carl Ponder, Cyril Zeller, Stanley Posey, NVIDIA
  - Sreeram Potluri, Karen Tomko and DK Panda, OSU
  - Amit Chourasia, SDSC
  - Jeffrey Vetter, Mitch D. Horton, Graham Lopez, Richard Glassbrok, Dick Glassbrook, NICS/Georgia Tech Keeneland Team
  - Jay Alameda, Gregory Bauer, Omar Padron, NCSA