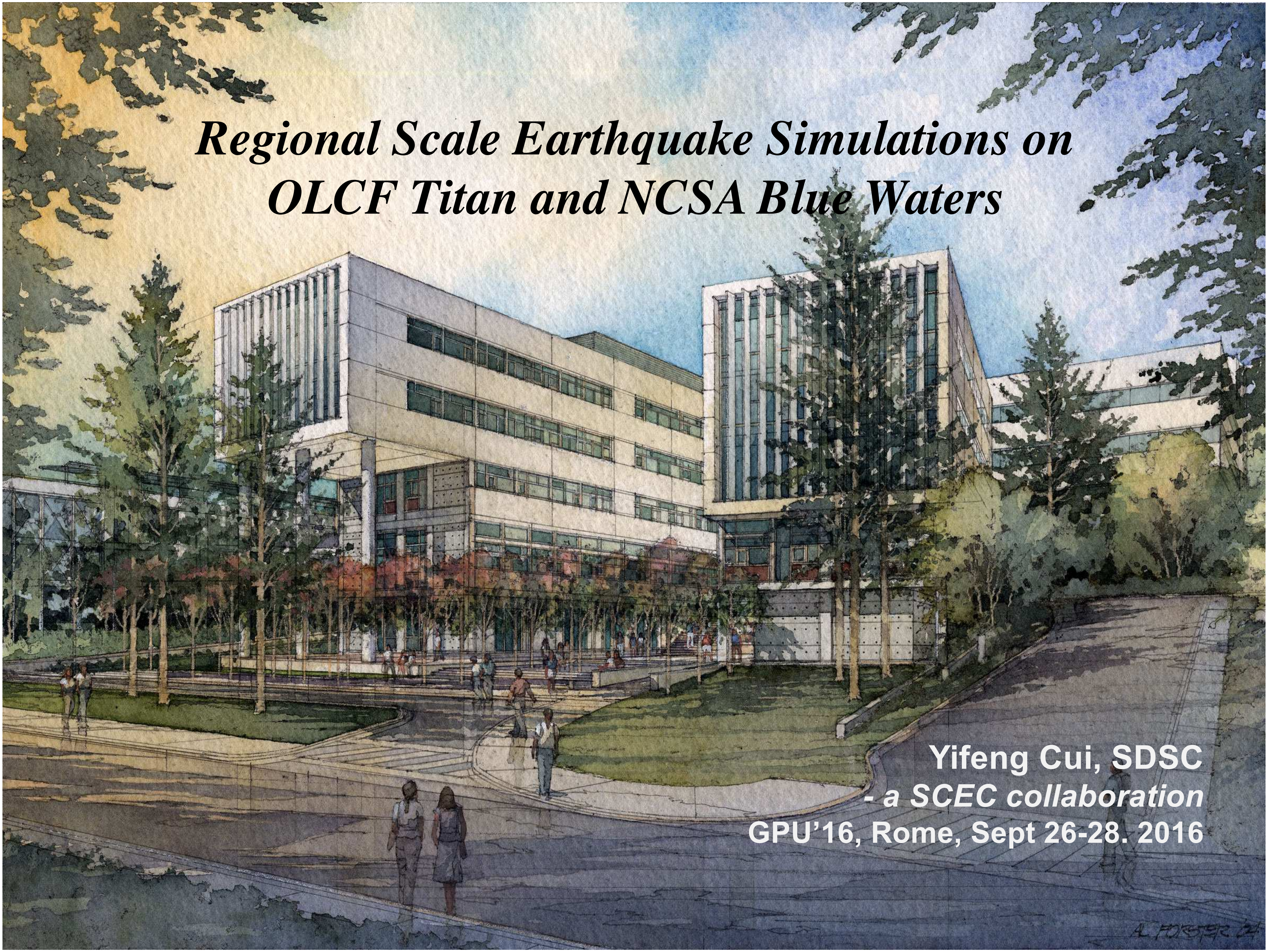


Regional Scale Earthquake Simulations on OLCF Titan and NCSA Blue Waters

Yifeng Cui, SDSC
- a SCEC collaboration
GPU'16, Rome, Sept 26-28. 2016

SCEC Collaborators

Thomas Jordan

Kim Olsen

Steve Day

Christine Goulet

Philip Macheling

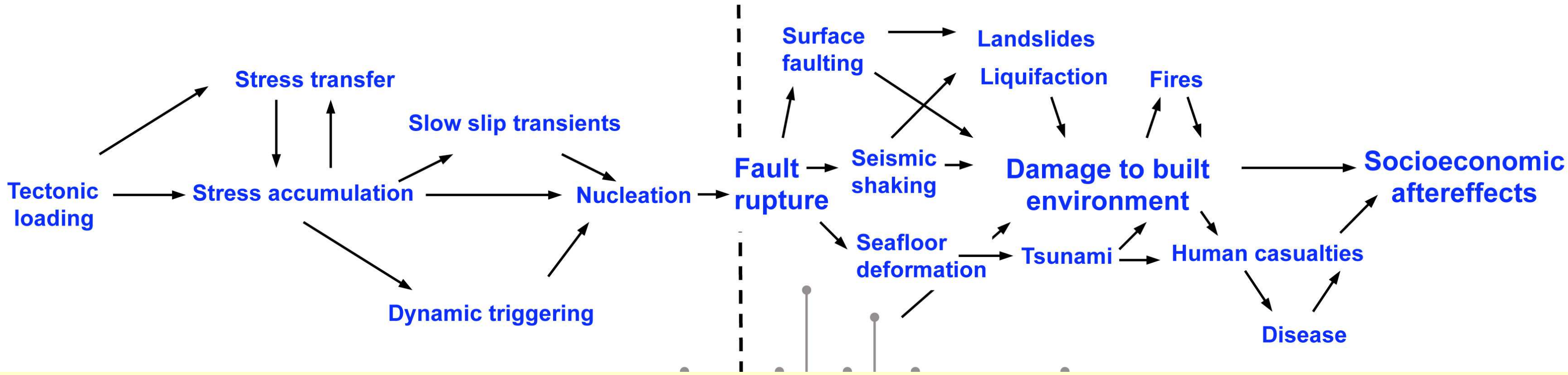
Computing Resources

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

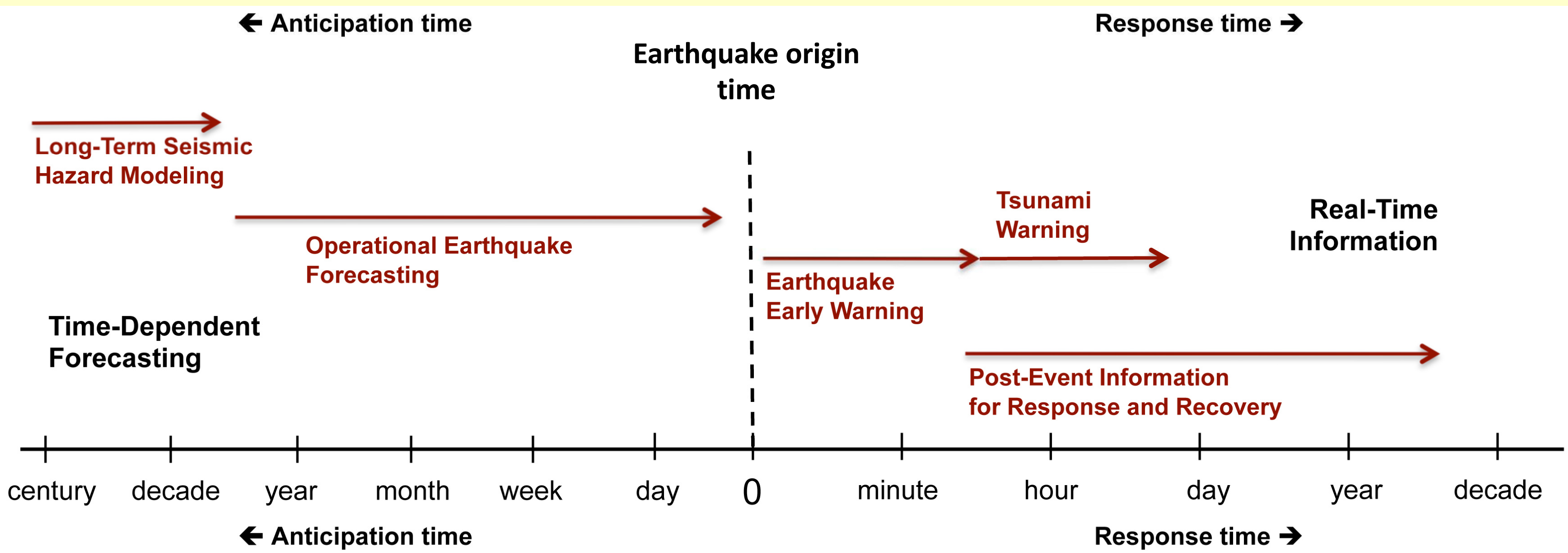
Grants/Awards

NSF SI2-SSI, PRAC, SCEC/USGS and XSEDE
Keck Foundation, Intel, NVIDIA

Prediction Problems of Earthquake System Science



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**

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

solved by explicit staggered-grid 4th-order FD

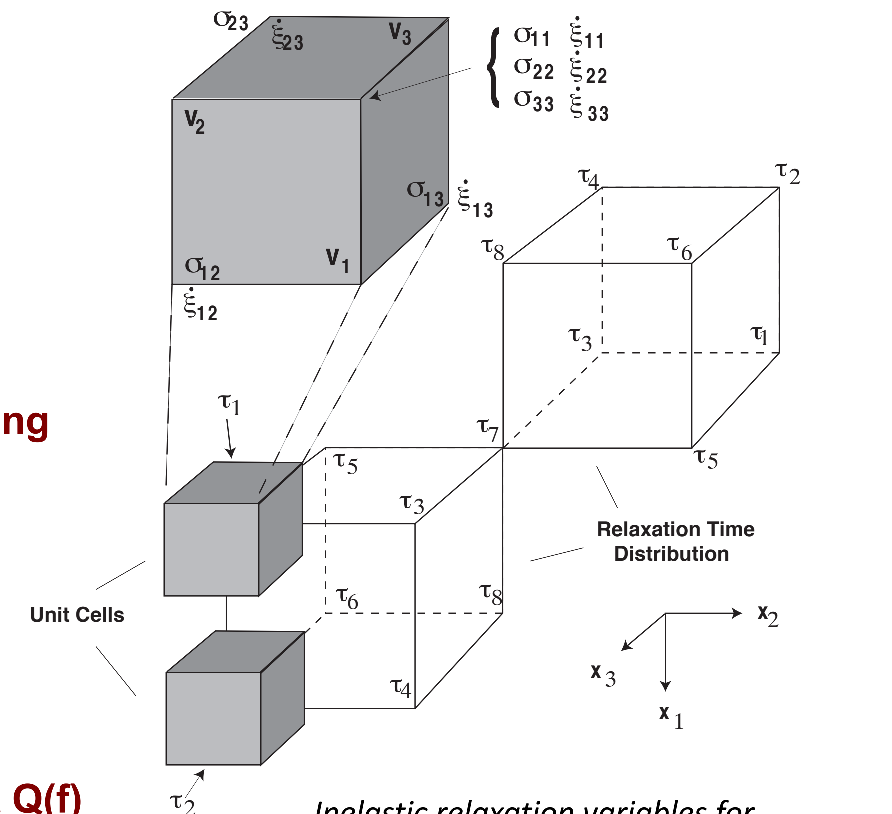
- **Memory variable formulation of inelastic relaxation using coarse-grained representation (Day 1998)**

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

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

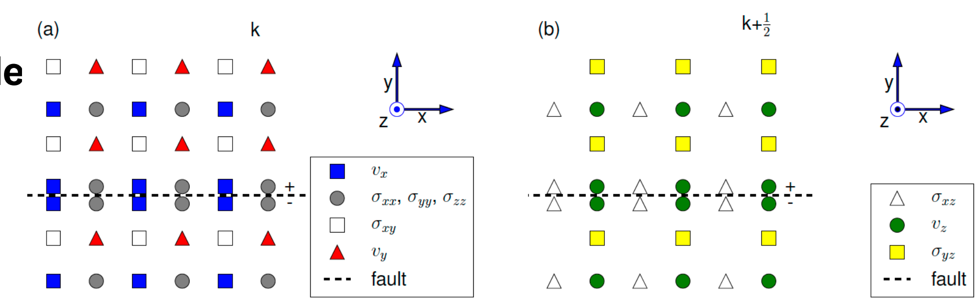
solve for $\frac{\delta M}{M_u} \lambda_i$ using linear least squares to fit a target Q(f) (Withers et al., 2015)

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



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

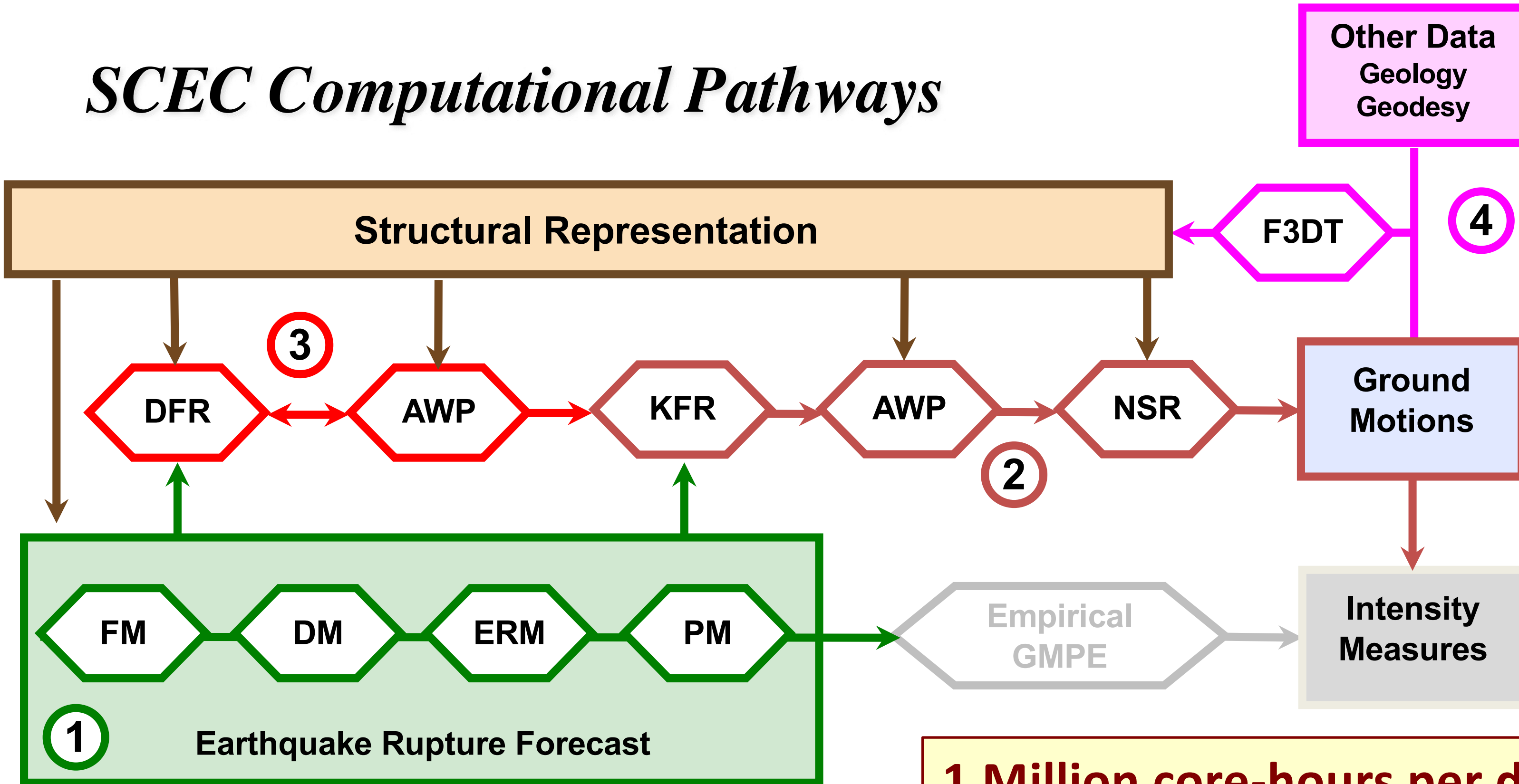
- **Dynamic rupture by the staggered-grid split-node (SGSN) method (Dalguer and Day 2007)**
- **Absorbing boundary conditions by PML (Marcinkovich and Olsen 2003) and Cerjan et al. (1985)**



HPGeoC SDSC
High Performance Geocomputing Laboratory

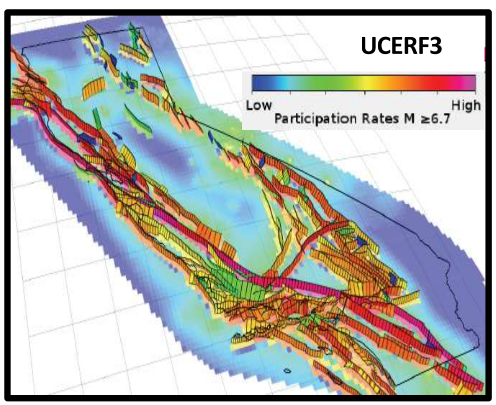
SCEC

SCEC Computational Pathways



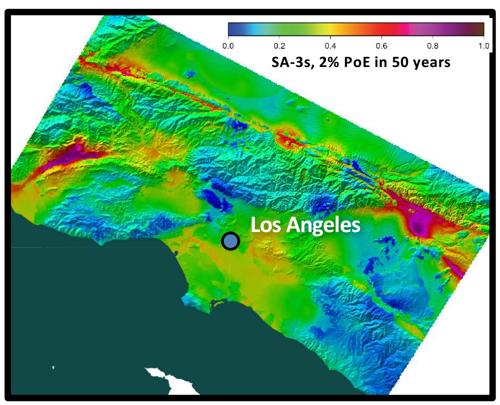
1 Million core-hours per day!

TACC *Stampede*



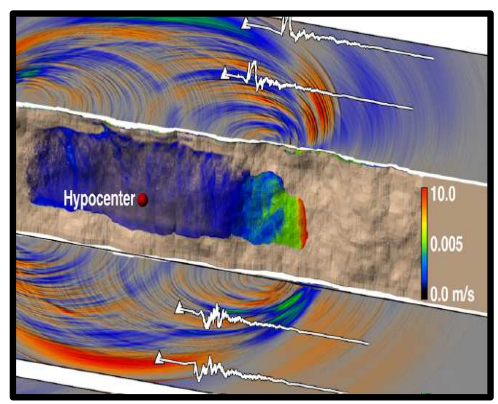
① Uniform California Earthquake Rupture Forecast (UCERF3)

NCSA *Blue Waters*



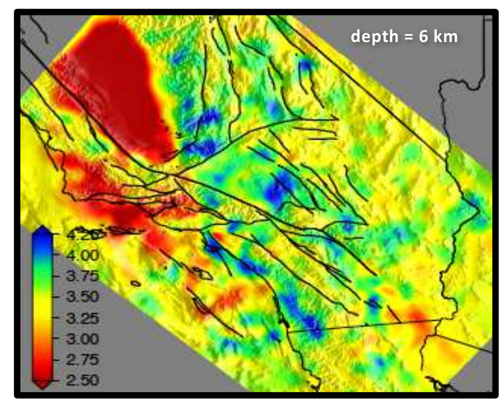
② CyberShake 14.2 seismic hazard model for LA region

OLCF *Titan*



③ Dynamic rupture model of fractal roughness on SAF

ALCF *Mira*



④ Full-3D tomographic model CVM-S4.26 of S. California

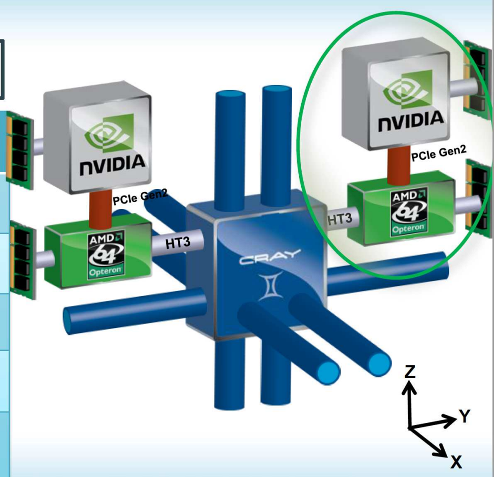
OLCF Titan and NCSA Blue Waters

BLUE WATERS
SUSTAINED PETASCALE COMPUTING
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION
CRAY

Cray XK7

Blue Waters contains 4,224 NVIDIA K20x (GK110) GPUs

XK7 Compute Node Characteristics	
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	156.8 Gflops
K20x Peak (DP floating point)	1.32 Tflops
Host Memory	32GB 51 GB/sec
K20x Memory	6GB GDDR5 capacity 235GB/sec ECC



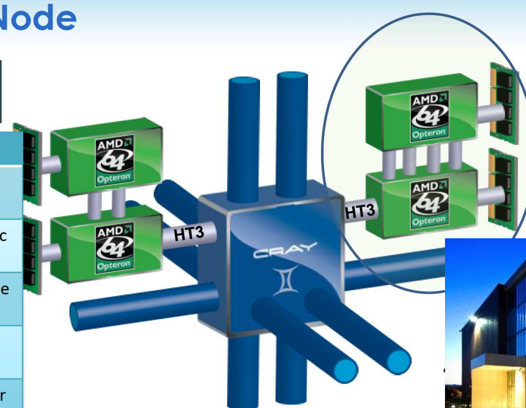
BLUE WATERS
SUSTAINED PETASCALE COMPUTING
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION
CRAY

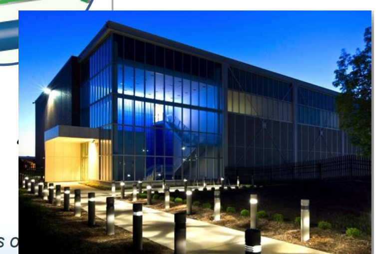
Blue Waters XE6 Node

Blue Waters contains 22,640 XE6 compute nodes

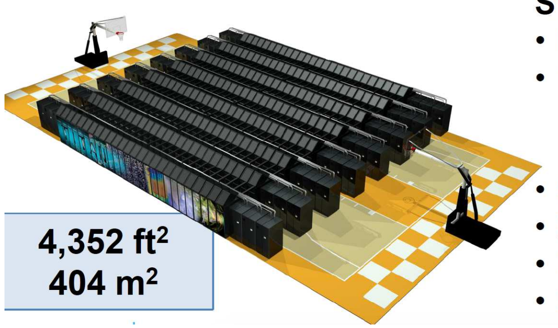
Node Characteristics	
Number of Core	16
	313 Gflops/sec
Memory Size	64 GB per node
Memory Bandwidth (Peak)	102 GB/sec
Interconnect Injection Bandwidth (Peak)	9.6 GB/sec per direction

*Each core module includes 1 256-bit wide FP unit and 2 integer units. This is o advertised as 2 cores, leading to a 32 core node.





ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla

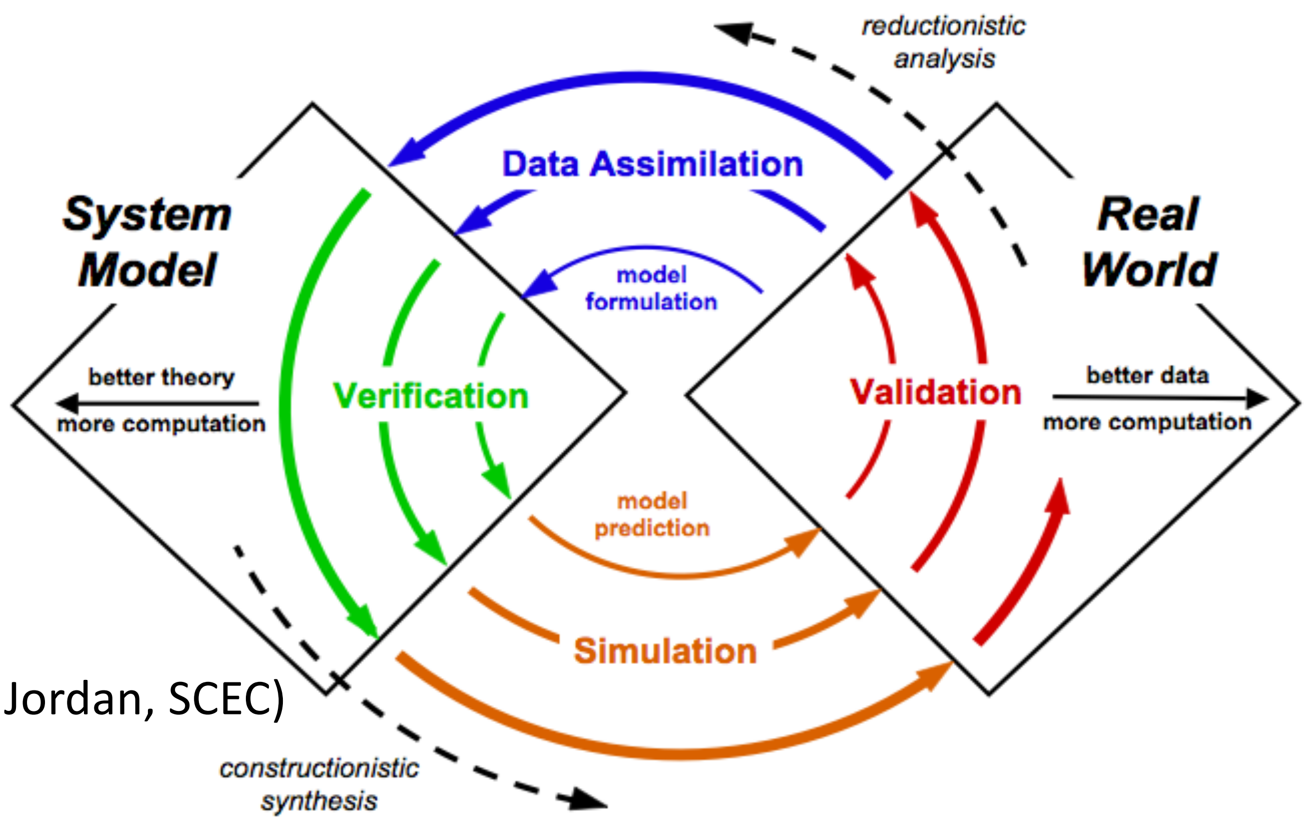


SYSTEM SPECIFICATIONS:

- 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**



(Source: Thomas Jordan, SCEC)

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

Validating New Physics for High Frequency

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	.001 Hz	.01 Hz	.1 Hz	1 Hz	10 Hz

tall buildings | *houses* | *stiff structures*

Earthquake engineering band

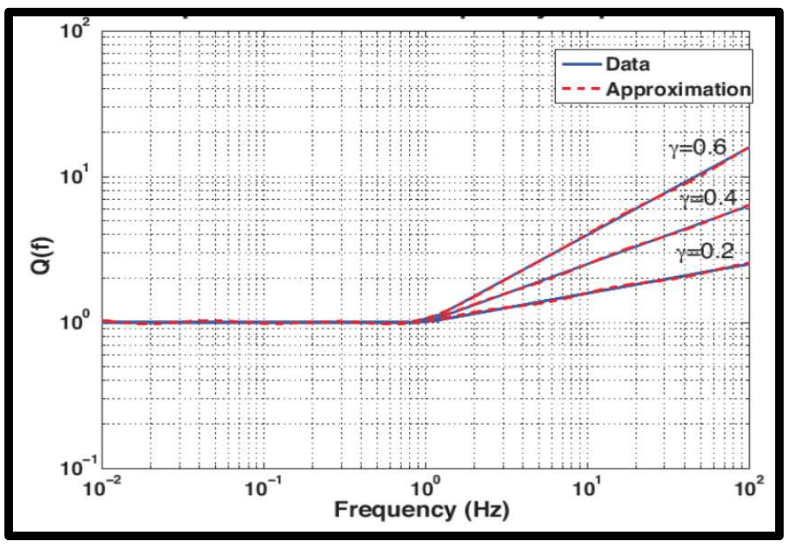
SCEC simulations 2014

physics-based deterministic

CyberShake 0.5 Hz

empirical stochastic

(Withers, 2015)



Must validate new physics

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

5 Hz

SCEC simulations 2018

physics-based deterministic

physics-based stochastic

empirical stochastic

Validating New Physics for High Frequency

Seismic band

low-order free oscillations *mantle waves* *crustal waves* *basin waves* *strongly scattered waves*

1000 s 100 s 10 s 1 s 0.1 s period

.001 Hz .01 Hz .1 Hz 1 Hz 10 Hz frequency

tall buildings *houses* *stiff structures*

Earthquake engineering band

SCEC simulations 2014

physics-based deterministic

CyberShake 0.5 Hz

empirical stochastic

(Cui et al., 2013)

Must validate new physics

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

5 Hz

SCEC simulations 2018

physics-based deterministic

physics-based stochastic

empirical stochastic

HPGeoC High Performance Geocomputing Laboratory SDSC

SCEC

Validating New Physics for High Frequency

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	.001 Hz	.01 Hz	.1 Hz	1 Hz	10 Hz

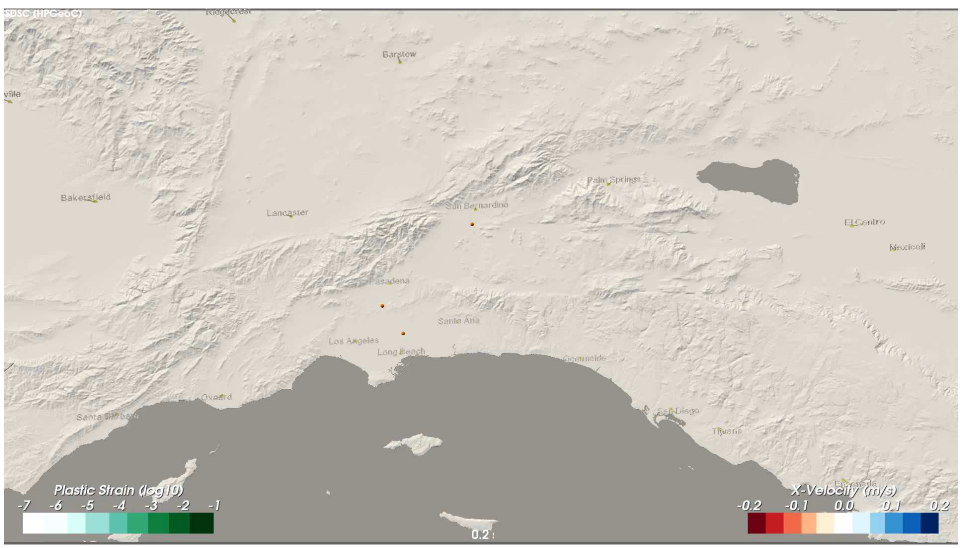
tall buildings — *houses* — *stiff structures*

Earthquake engineering band

SCEC simulations 2014

physics-based deterministic → CyberShake 0.5 Hz - - - - *empirical stochastic*

(Roten et al., 2016)



Must validate new physics

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

5 Hz

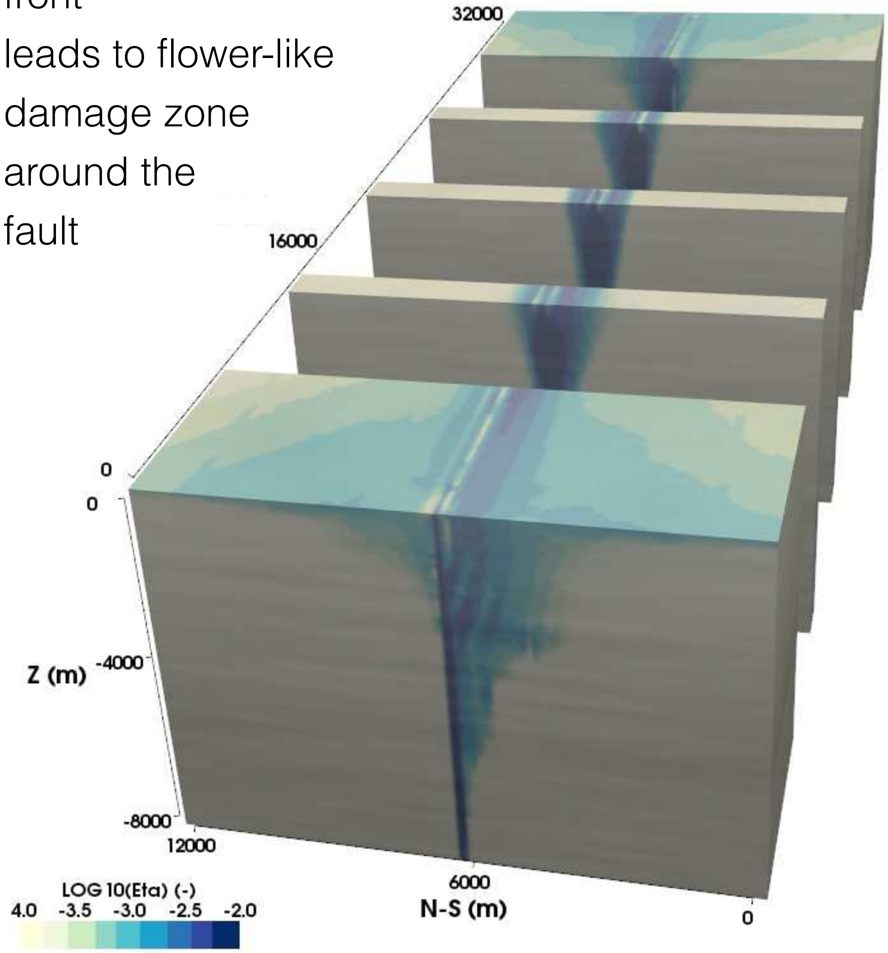
SCEC simulations 2018

physics-based deterministic → *physics-based stochastic* - - - - *empirical stochastic*

Nonlinear Material Response

In the fault damage zone

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

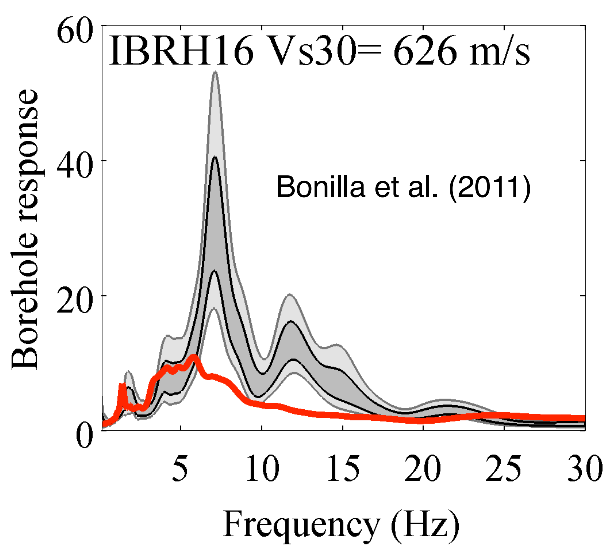


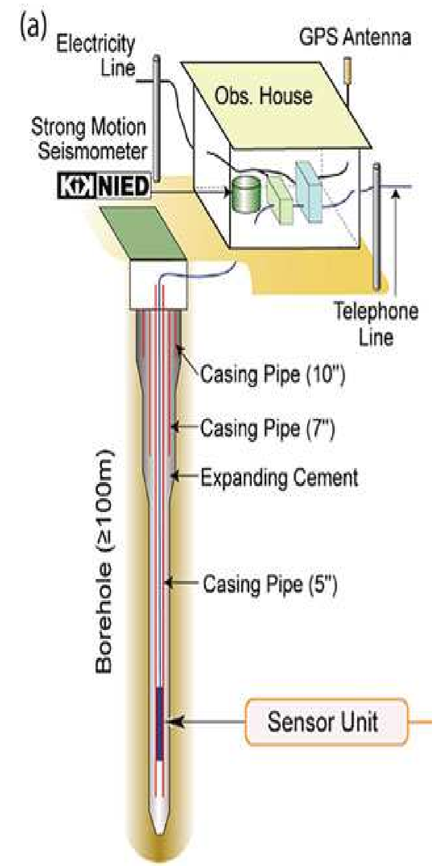
(Roten et al., 2014)

In shallow sedimentary deposits

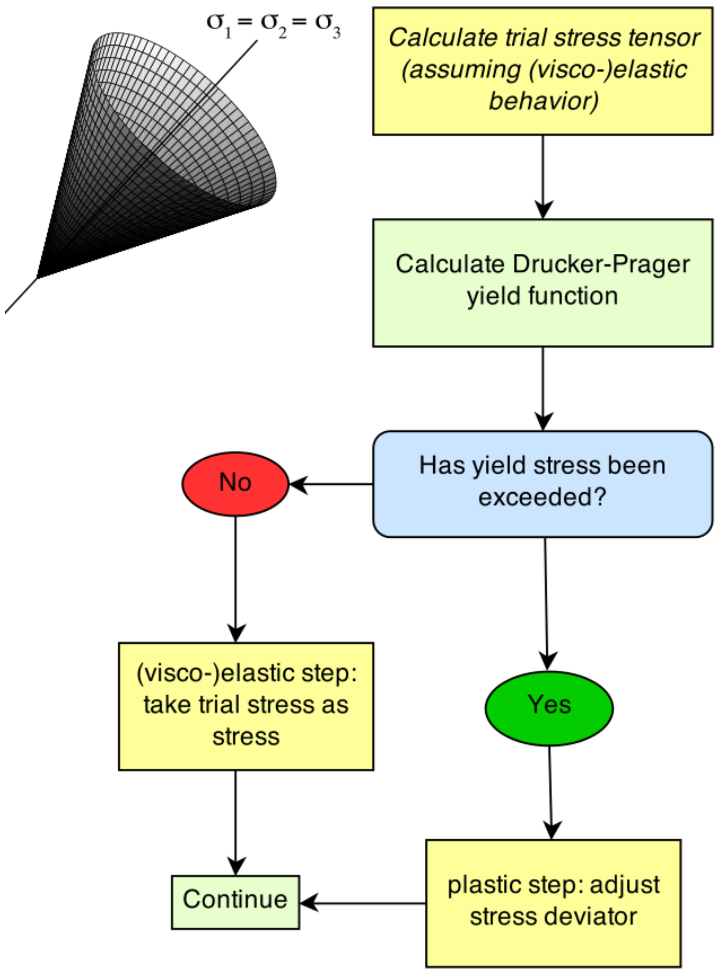
- caused by hysteretic stress-strain relationship in soft soils
- leads to a reduction in amplification

$$A(f) = \frac{\mathcal{F}\,(S(t))}{\mathcal{F}\,(R(t))}$$





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:

$$\tau_{ij} = \tau_m^{\text{trial}} \delta_{ij} + r s_{ij}^{\text{trial}}$$

Yield factor r with viscoelastic relaxation time T_v:

$$r = \frac{Y(\tau^{\text{trial}})}{\sqrt{J_2(\tau^{\text{trial}})}} + \left(1 - \frac{Y(\tau^{\text{trial}})}{\sqrt{J_2(\tau^{\text{trial}})}}\right) \exp \frac{-\Delta t}{T_v}$$

method	CPU time	Normalized
Elastic	0.176	100%
Individual interpolation (EP1)	0.676	384%
Yield factor interpolation (EP2)	0.290	165%

(Barall 2014, Roten et al., 2014, 2015, 2016)

Linear + 11 variables -> 5 var -> 3 var

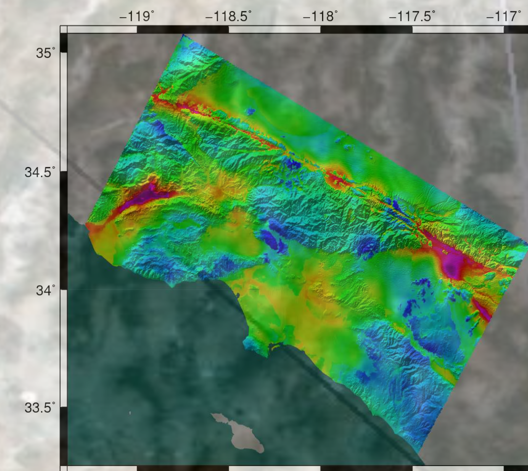
HPGeoC SDSC
HIGH PERFORMANCE GEOCOMPUTING LABORATORY

SCEC

Why choose GPU?

Create a SCEC broadband CyberShake hazard model for all of California

Computational requirements for 1400 sites across California



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)**

The statewide CyberShake hazard model will comprise 1.8 billion seismograms

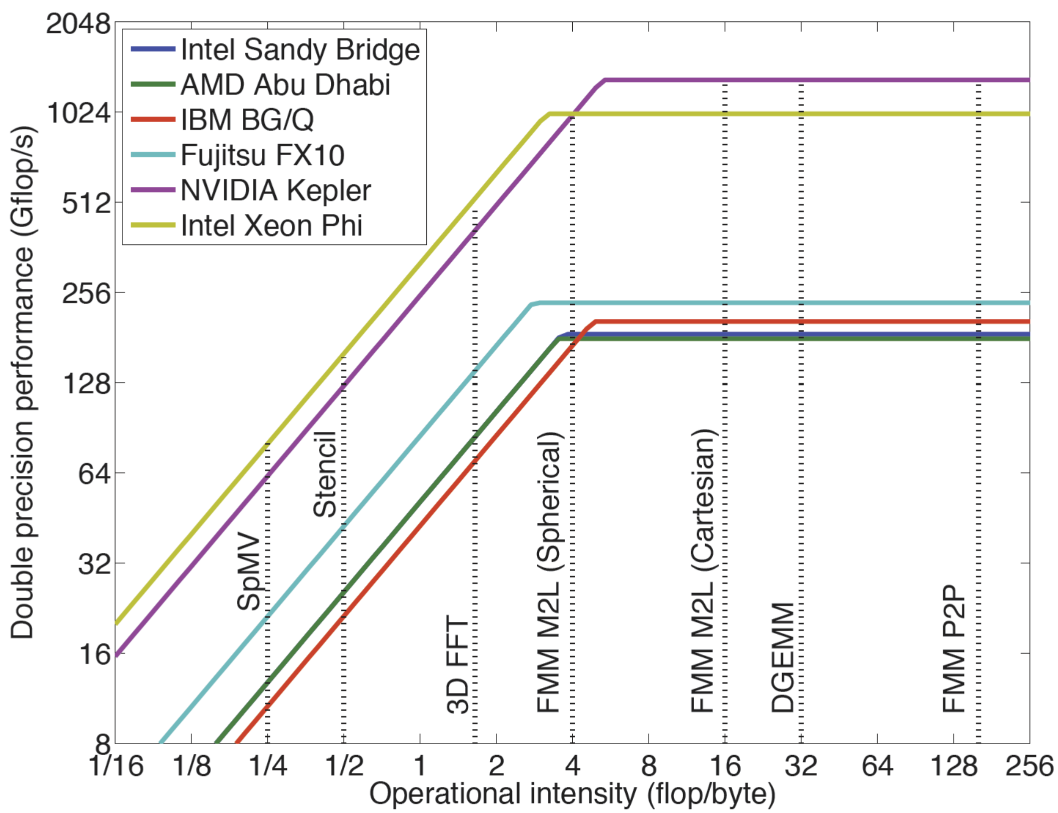
Go Green

HPGeoC SDSC
HIGH PERFORMANCE GEOCOMPUTING LABORATORY

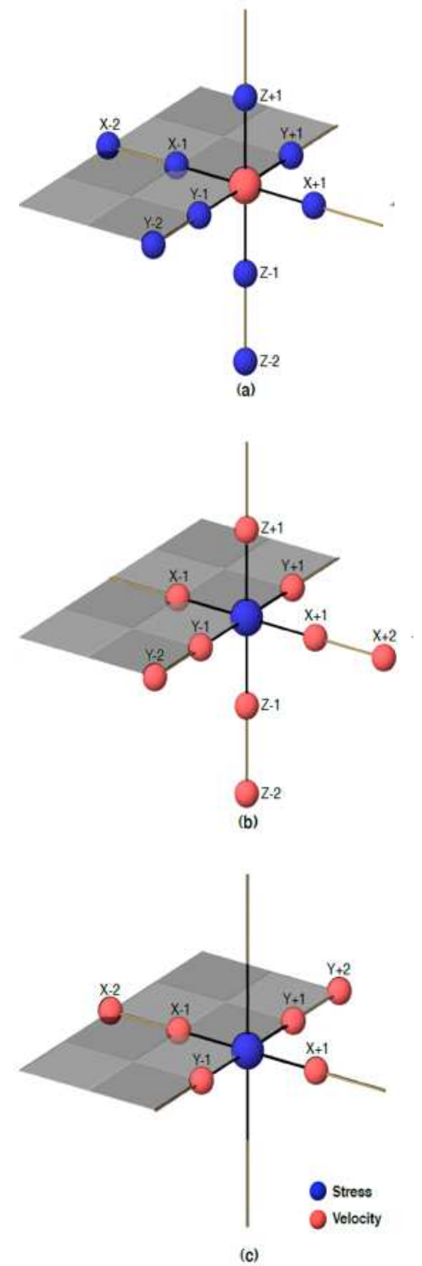
SCEC

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

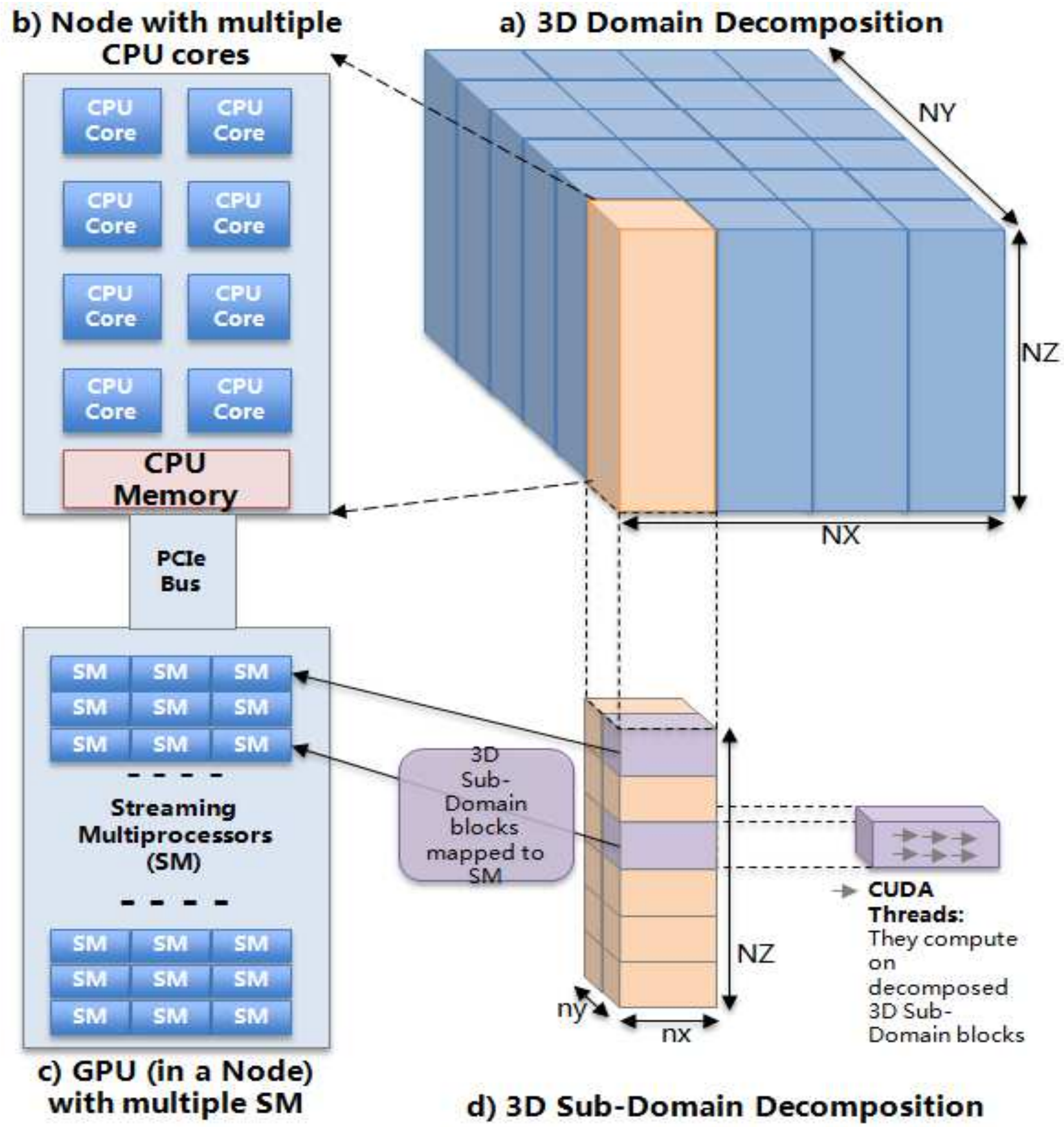


(Barba & Yokota, SIAM News, 46/6, 2013)



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**



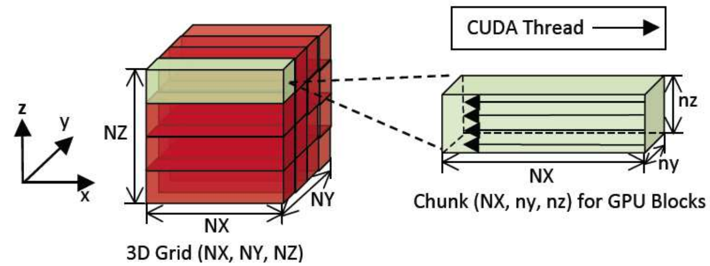
Single-GPU Optimizations

- ✓ **Step 2: GPU 2D Decomposition in y/z vs x/y**
- ✓ **Step-3: Global memory Optimization**

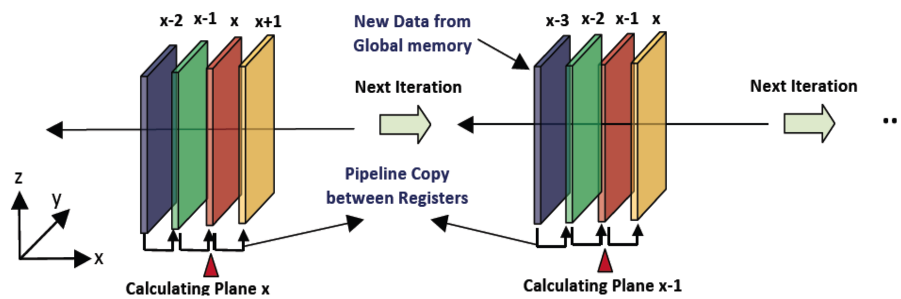
  Global memory coalesced, texture memory for six 3D constant variables, constant memory for scalar constants
- ✓ **Step-4: Register Optimization**

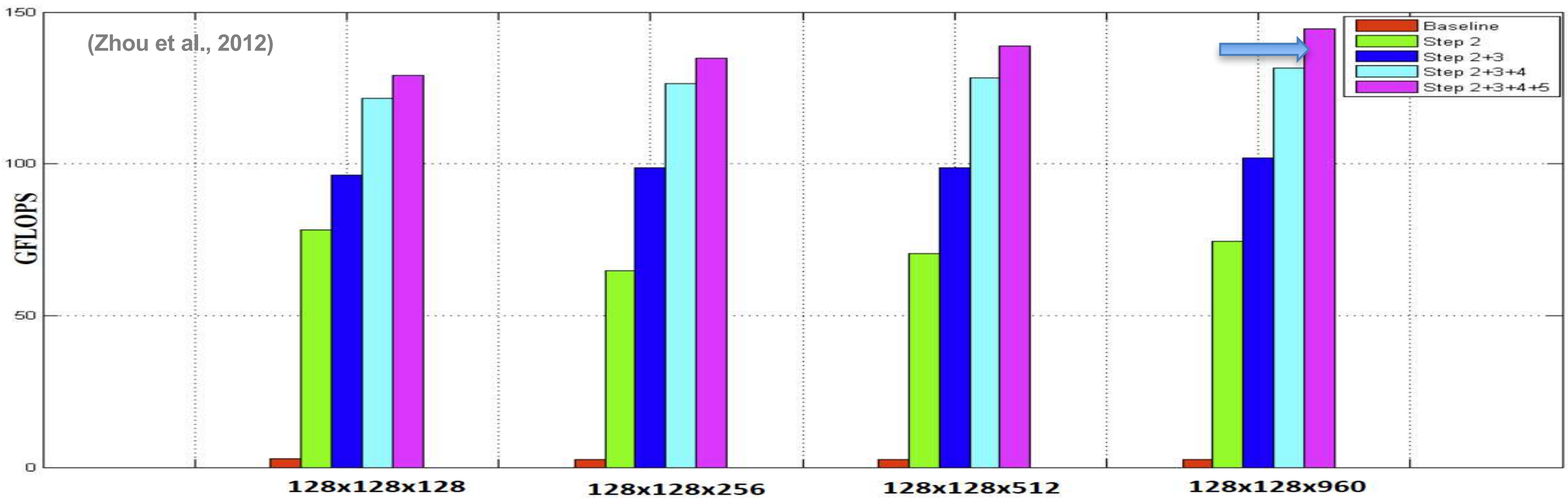
  Pipelined register copy to reduce memory access
- ✓ **Step-5: L1/L2 cache vs shared memory**

  Rely on L1/L2 cache rather on-chip shared memory



(Zhou, J et al., ICCS 2012)





HPGeoC SDSC
High Performance Geocomputing Laboratory

SCEC

Blue Waters PAID Project

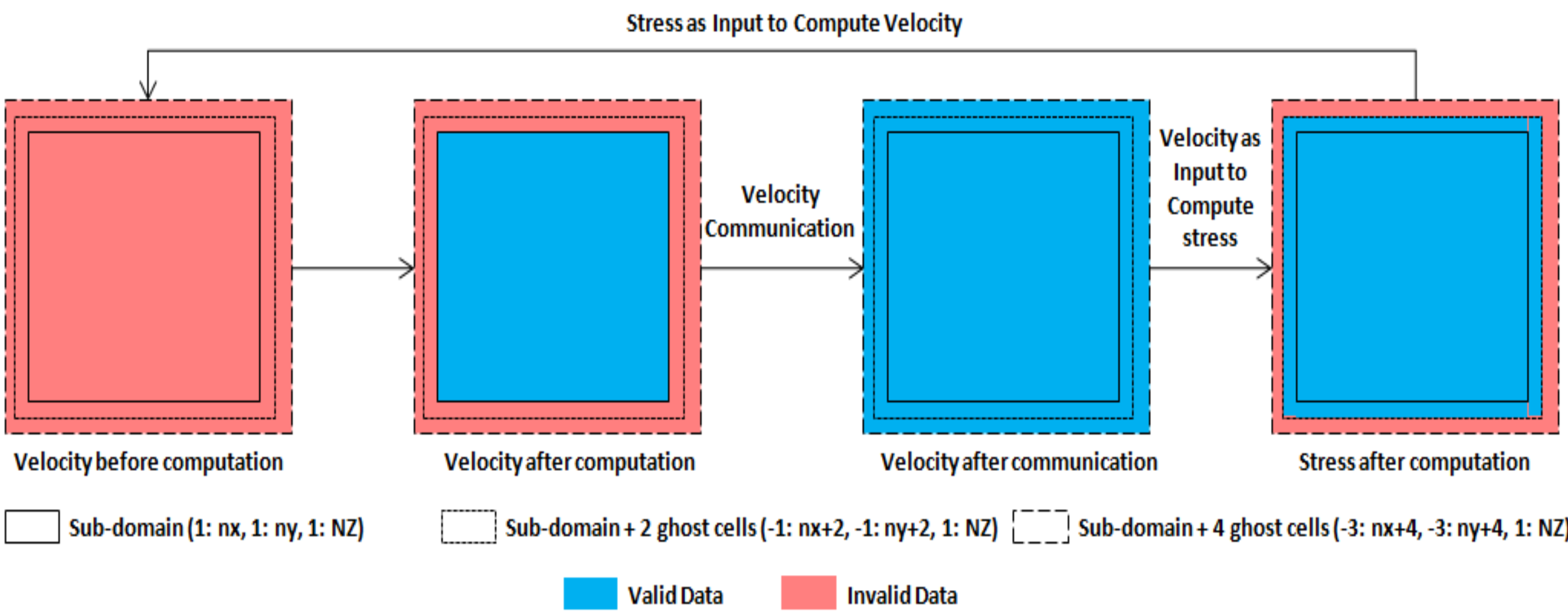
- **Optimization strategies**
  - Increasing occupancy to hide memory latency
  - Reducing redundant halo accesses by using texture cache combined with register queues
- **Velocity Kernel**
  - Increasing the block size
  - More register queueing, allow read-only array accesses to use texture cache
- **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			
	Baseline	Optimized	Multiple Stream	DRAM Bandwidth
dstrqc	65.56	58.957 (x1.11)	58.957	65%
drprecpc_calc	27.604	18.795 (x1.47)	18.795	75%
dvelcx	21.295	20.09 (x1.06)	20.09	65%
other kernels	4.472	4.472	1.922	--
MPI	6.972	6.972	overlapped	--
data transfer	10.513	10.513	overlapped	--
full iteration	136.416	119.799 (x1.13)	99.764 (x1.37)	--

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 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.**

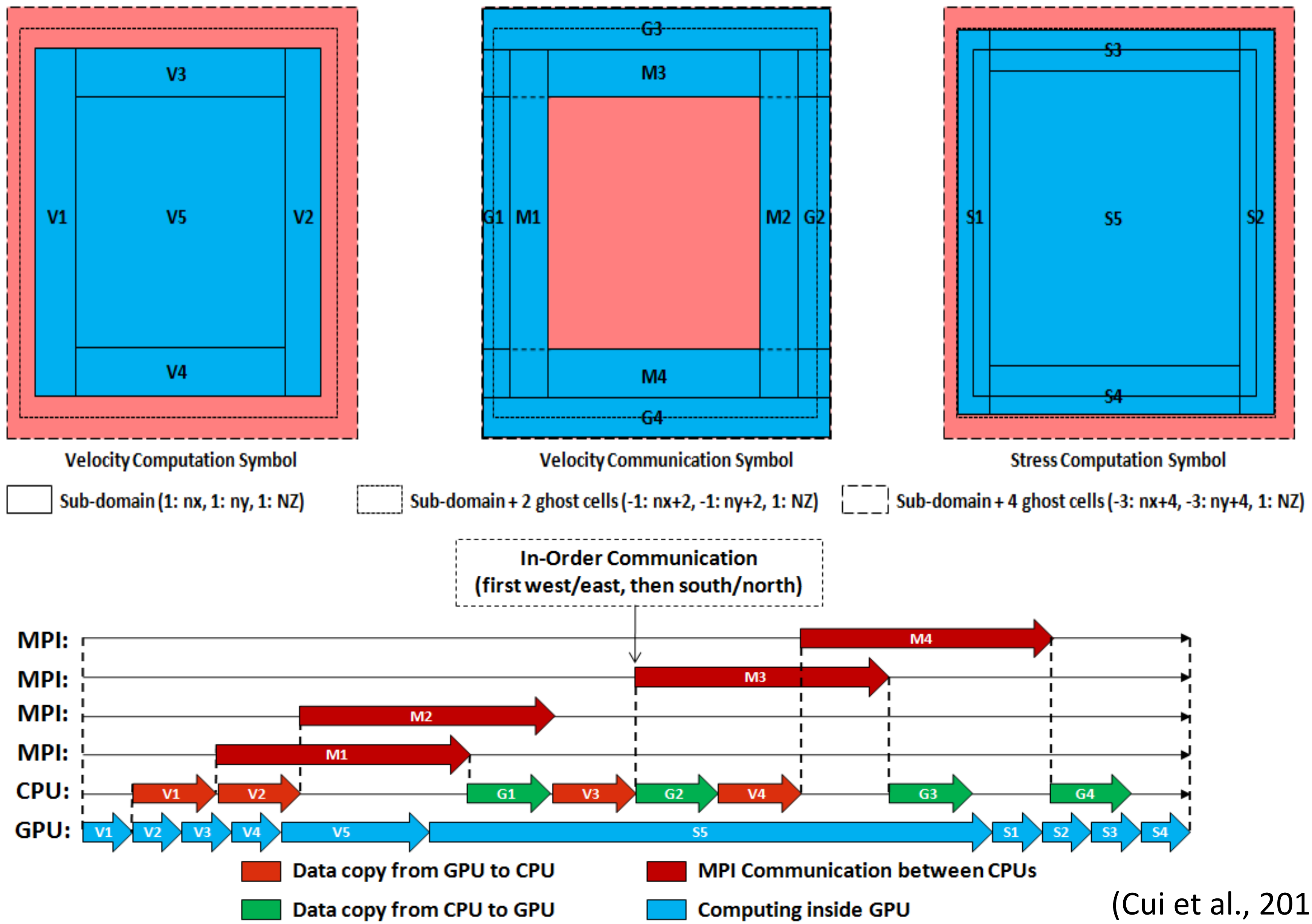


(Zhou et al., 2013)

GPU-GPU Communication

Communication	Velocity		Stress	
	Frequency	Message size	Frequency	Message size
Before Comm Reduction	4	6*(nx+ny)*NZ	4	12*(nx+ny)*NZ
After Comm Reduction	4	12*(nx+ny+4)*NZ	**No communication**	

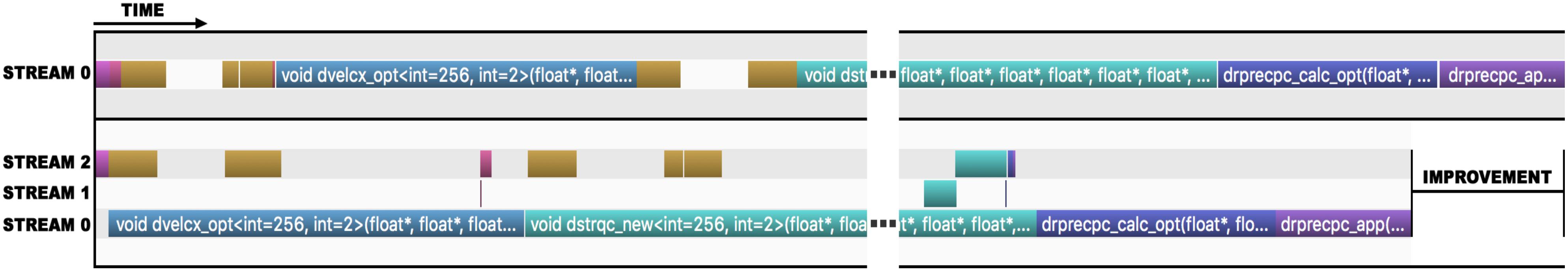
Computing and Communication Overlapping



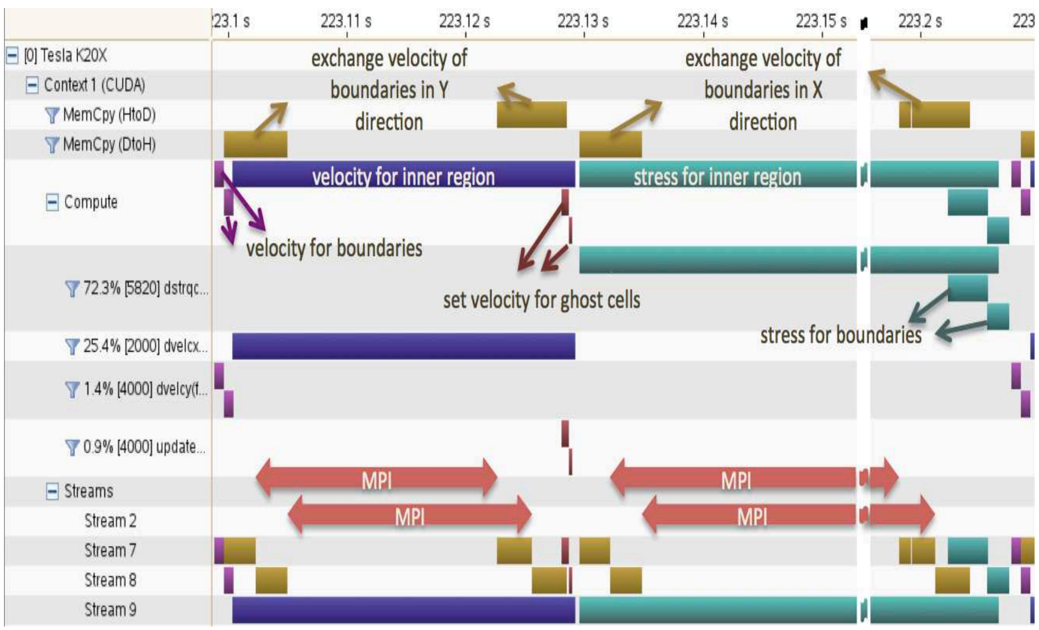
(Cui et al., 2013)

Multi-streaming for Computing/Communication Overlap

Nonlinear code



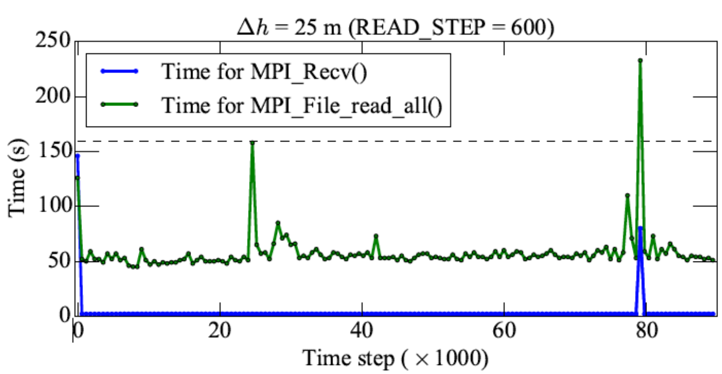
Linear code



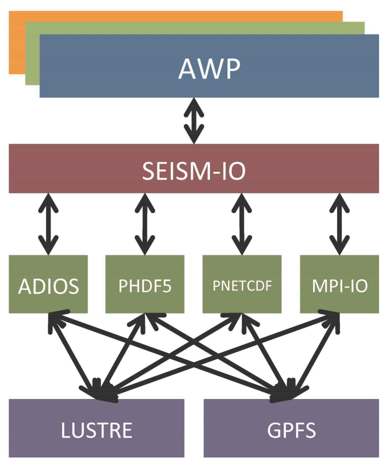
- **Multi-streaming technique to overlap communication with computation**
- **Small kernels can be optimized by GPU job scheduler**
- **Linear implementation of AWP-ODC-GPU achieves a parallel efficiency of 100% with 16,384 XK7 nodes**

Two-phase I/O Model

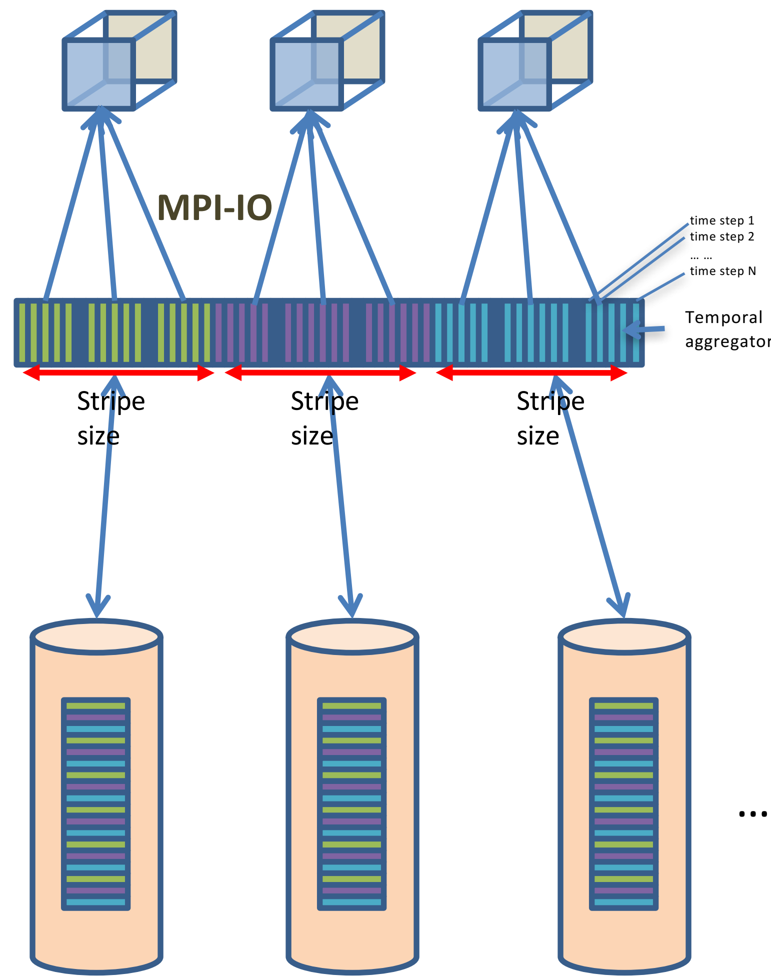
- **Parallel I/O**
  - **Read and redistribute 6.9 TB inputs**
    - Contiguous block read by reduced number of readers
    - High bandwidth asynchronous point-to-point communication redistribution
  - **Aggregate and write**
    - Temporal aggregation buffers
    - Contiguous writes
    - Throughput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



(Roten et al., 2016)



(Poyraz et al., 2014)



OSTs

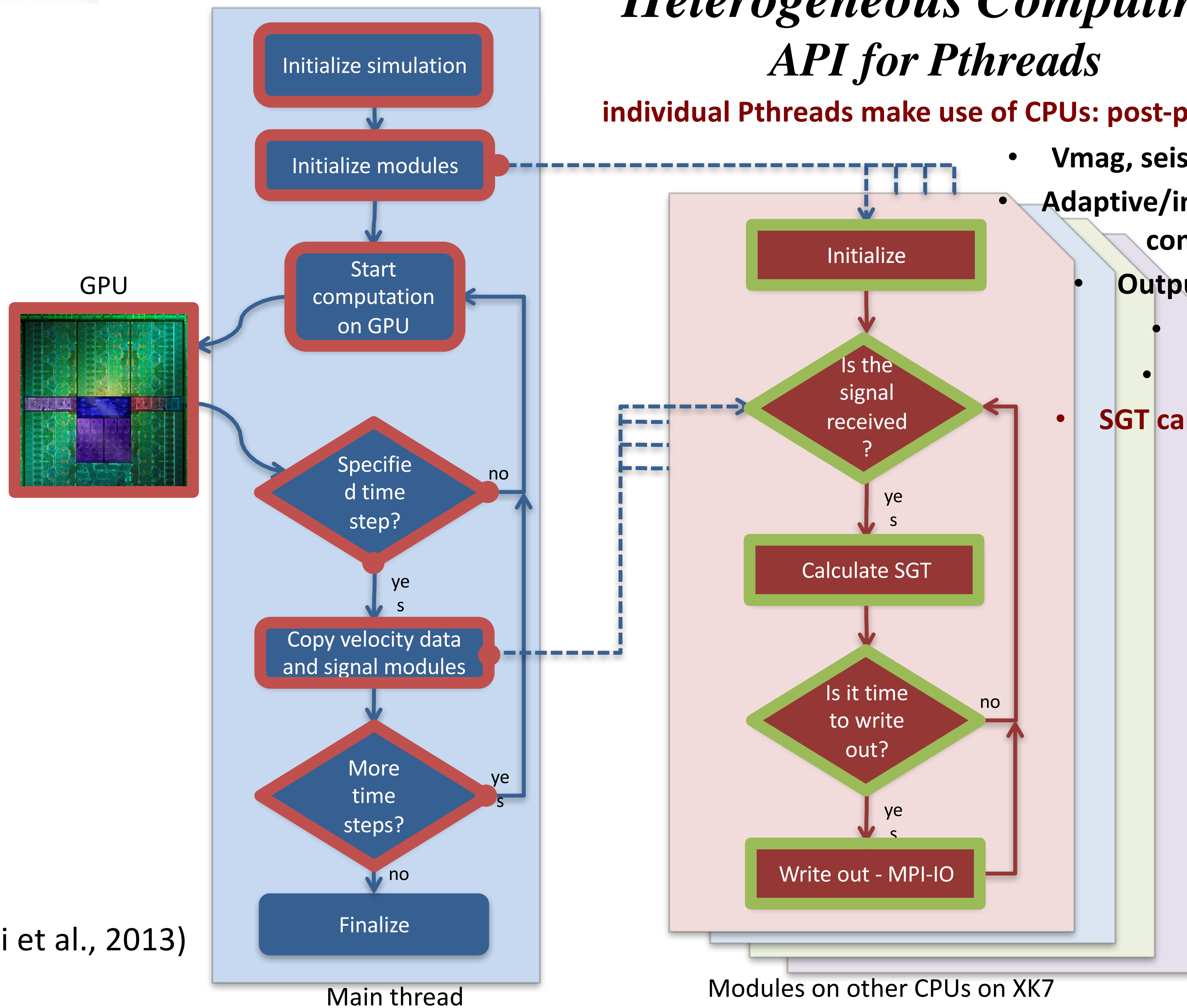
(Cui et al., 2010)

HPGeoC SDSC
High Performance Geocomputing Laboratory

Heterogeneous Computing: API for Pthreads

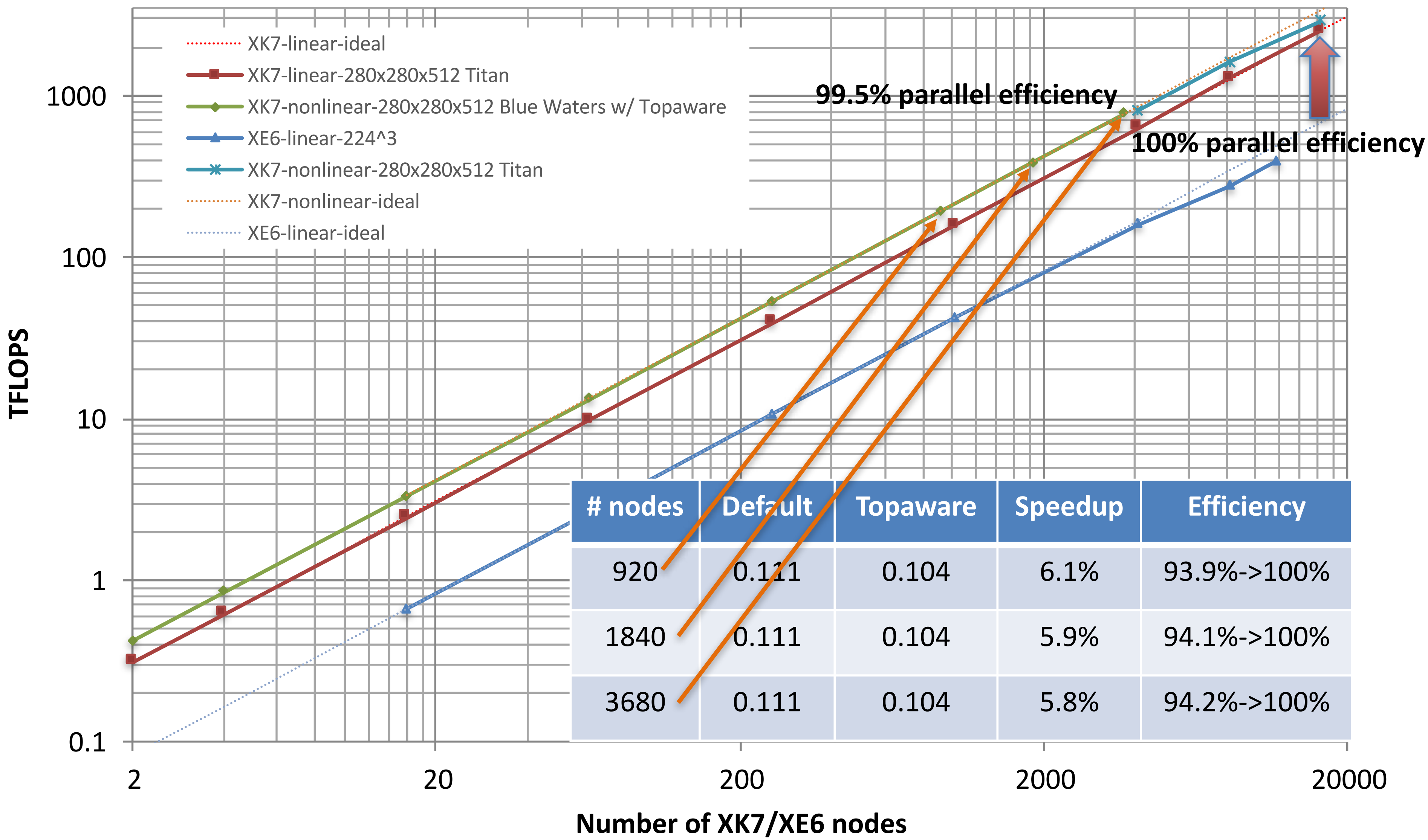
individual Pthreads make use of CPUs: post-processing

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



(Cui et al., 2013)

AWP-ODC Weak Scaling



# nodes	Default	Topaware	Speedup	Efficiency
920	0.111	0.104	6.1%	93.9%->100%
1840	0.111	0.104	5.9%	94.1%->100%
3680	0.111	0.104	5.8%	94.2%->100%

AWP-ODC Performance

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

*** Nonlinear code**

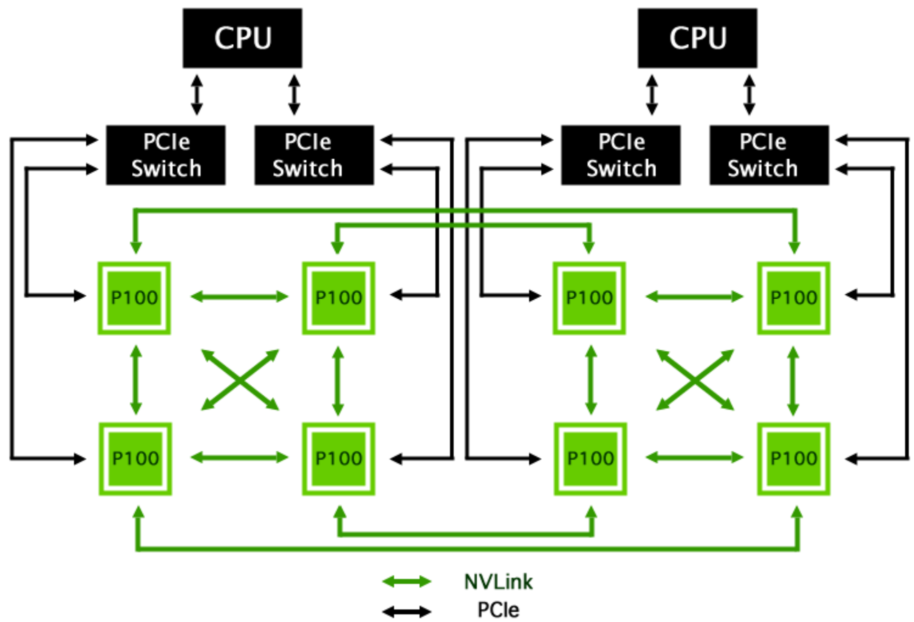
HPGeoC High Performance Geocomputing Laboratory SDSC

SCEC

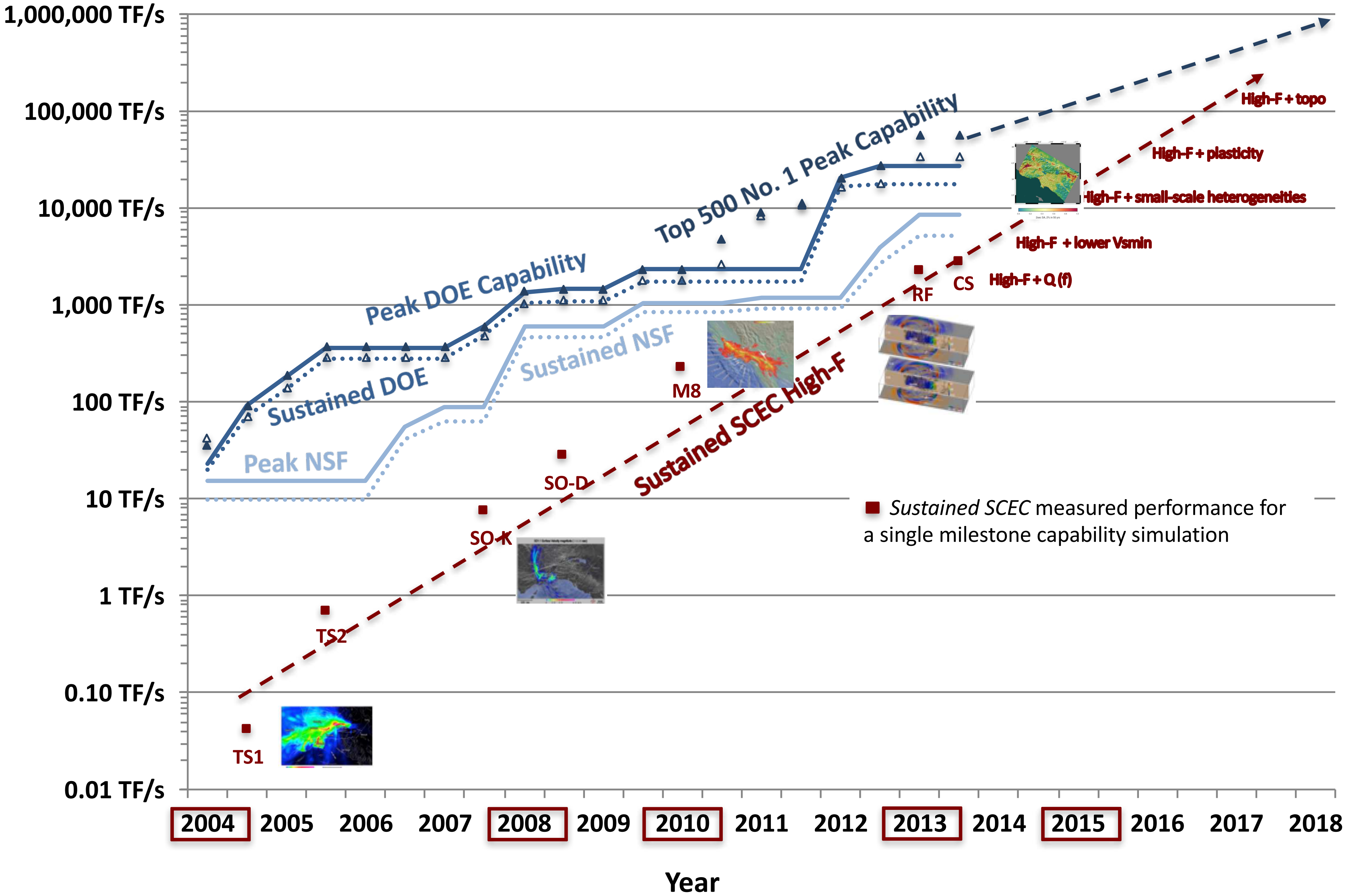
OLCF Summit in 2018

- **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

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY



Extreme-scale Earthquake computing



HPGeoC SDSC
High Performance Geocomputing Laboratory

SCEC

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 SDSC
HIGH PERFORMANCE GEOCOMPUTING LABORATORY

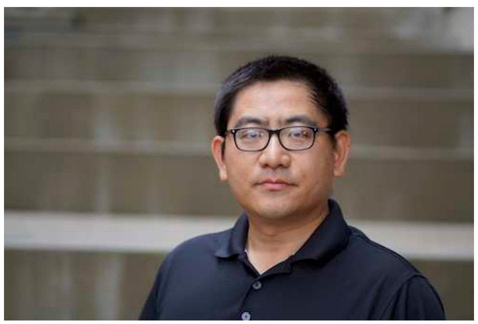
SCEC

HPGeoC Supports Earthquake Simulations

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

Supported by

SCEC

NSF

USGS
science for a changing world

intel

XSEDE
Extreme Science and Engineering Discovery Environment

NVIDIA

W.M. KECK FOUNDATION

Thank You!