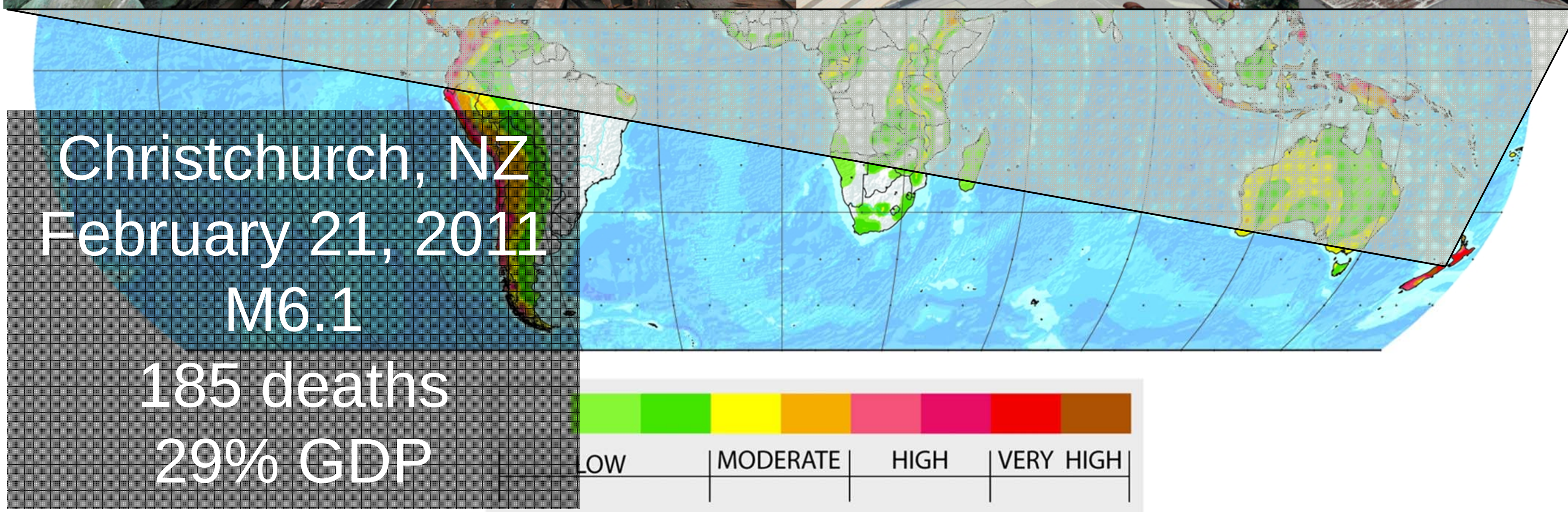
How’s It Shakin’?
Simulating Earthquakes with HPC

International HPC Summer School
June 3, 2014

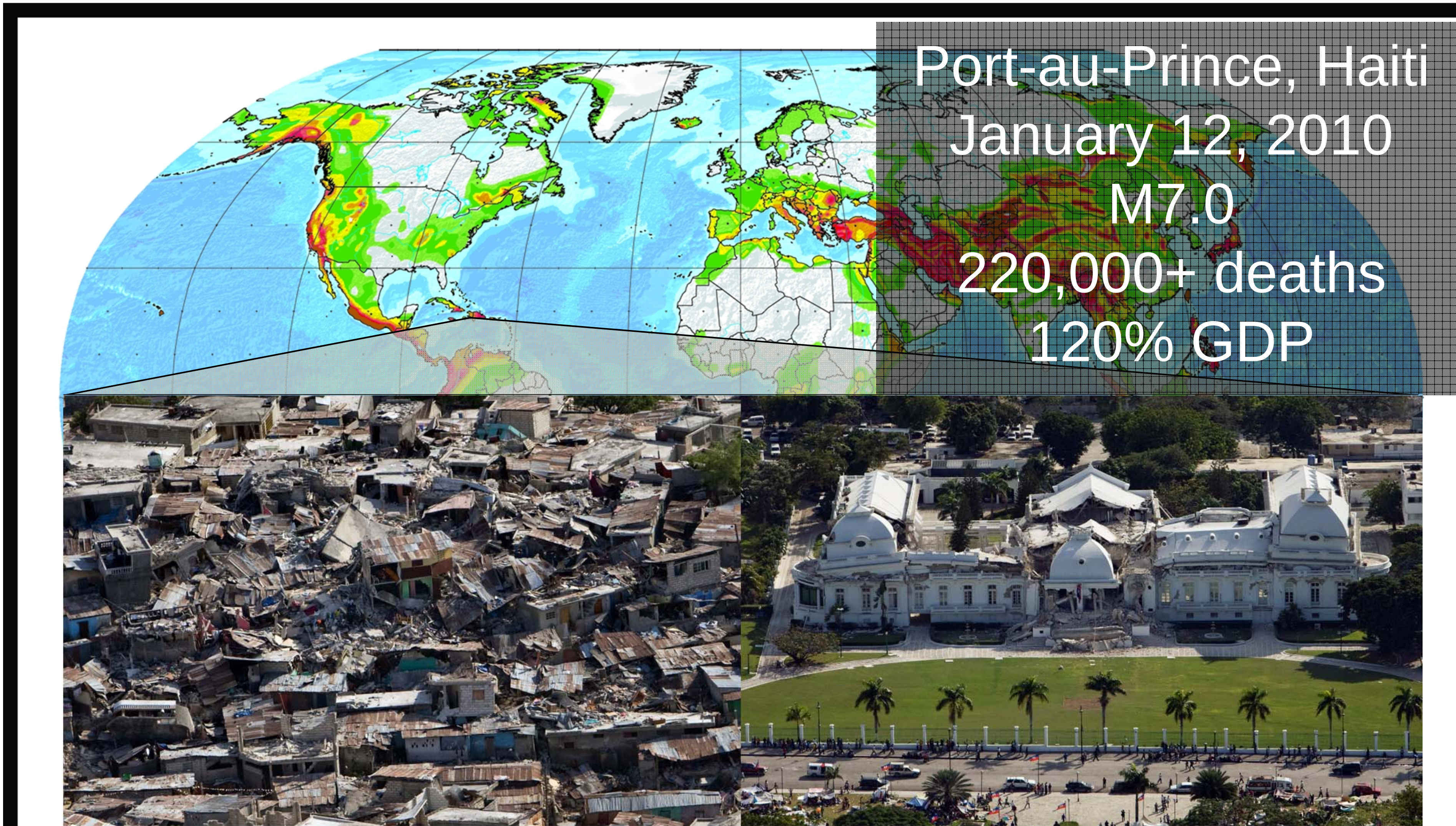
Scott Callaghan
Southern California Earthquake Center
University of Southern California
scottcal@usc.edu

SCEC
an NSF+USGS center



Christchurch, NZ
February 21, 2011
M6.1
185 deaths
29% GDP

Earthquakes are worldwide



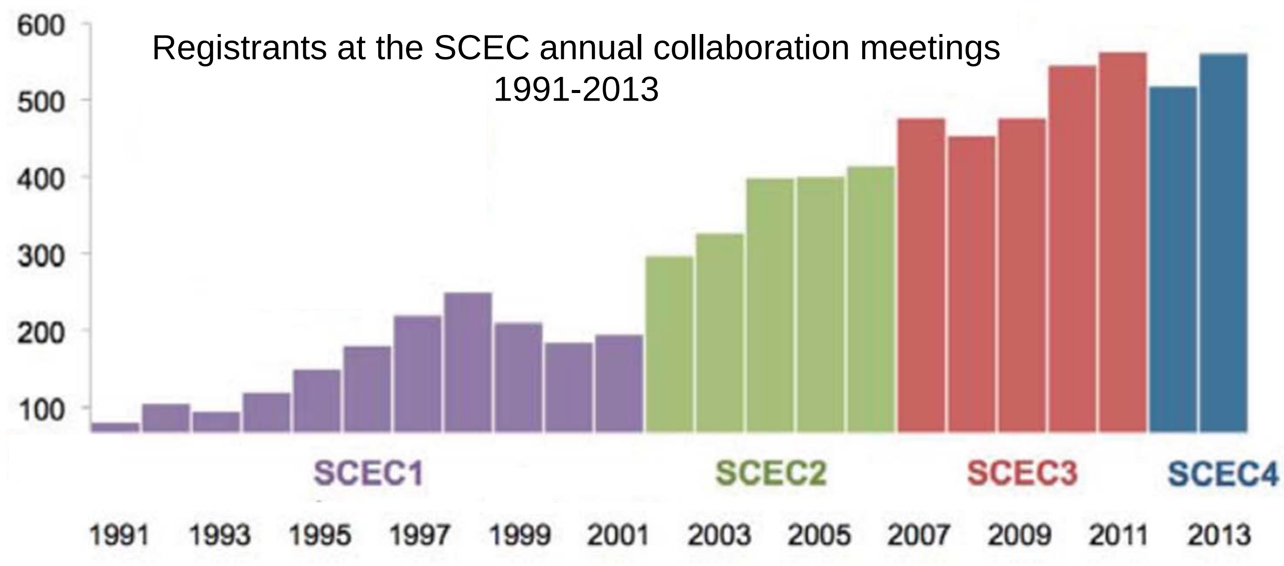
Earthquakes are worldwide

Why such variable impact?

- Earthquakes span a huge range of scales
  - Each magnitude point is 10x displacement, 32x energy
  - M9.0 (Tohoku, Japan) has 790x motion, 23000x energy compared to M6.1 (Christchurch, New Zealand)
- Earth structure strongly affects ground motion
  - Maximum Tohoku ground motion: 2.7g
  - Maximum Christchurch ground motion: 2.2g
  - Christchurch earthquake shallow, in basin
- Building types and infrastructure affect human impact

Southern California Earthquake Center

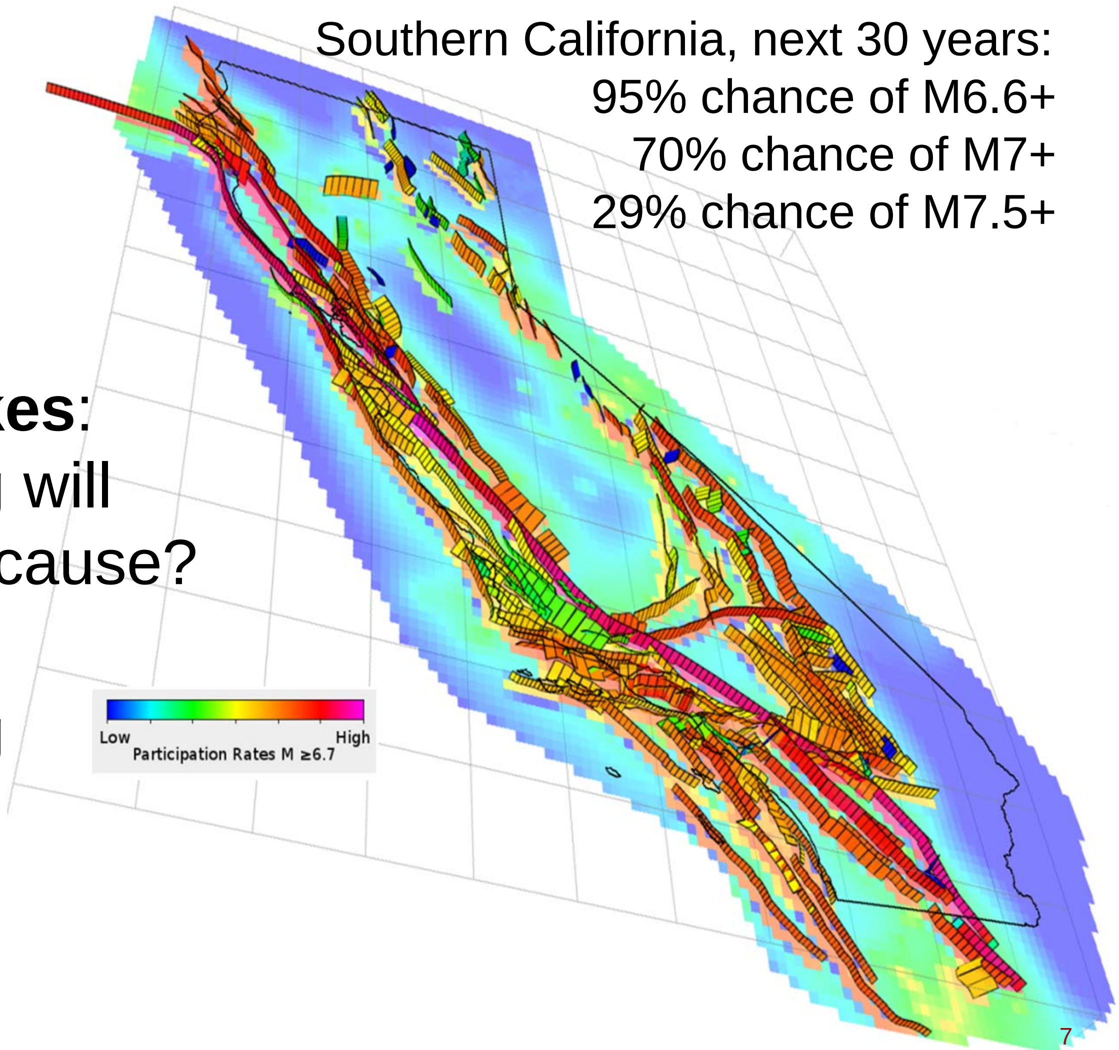
- Collaboration of 600+ scientists at 60+ institutions
- Major missions:
  - Gather field and experimental data
  - Integrate "ground truth" with simulation results
  - Communicate understanding to society



SCEC Simulations

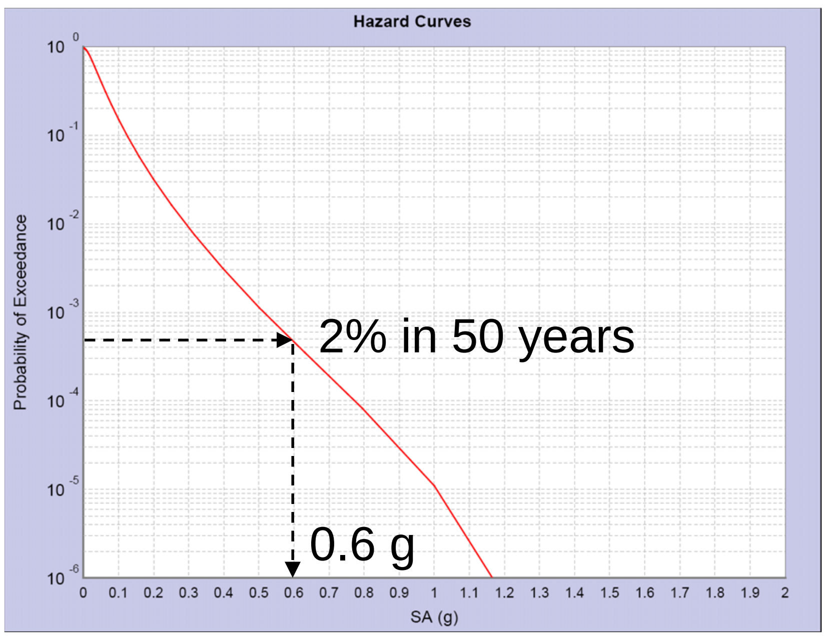
- Focus on Southern California
- Two main types of SCEC HPC projects
  - **Scenario earthquakes**: What kind of shaking will this *one earthquake* cause?
  - **Seismic hazard:** What kind of shaking will this *one location* experience?

Southern California, next 30 years:
95% chance of M6.6+
70% chance of M7+
29% chance of M7.5+

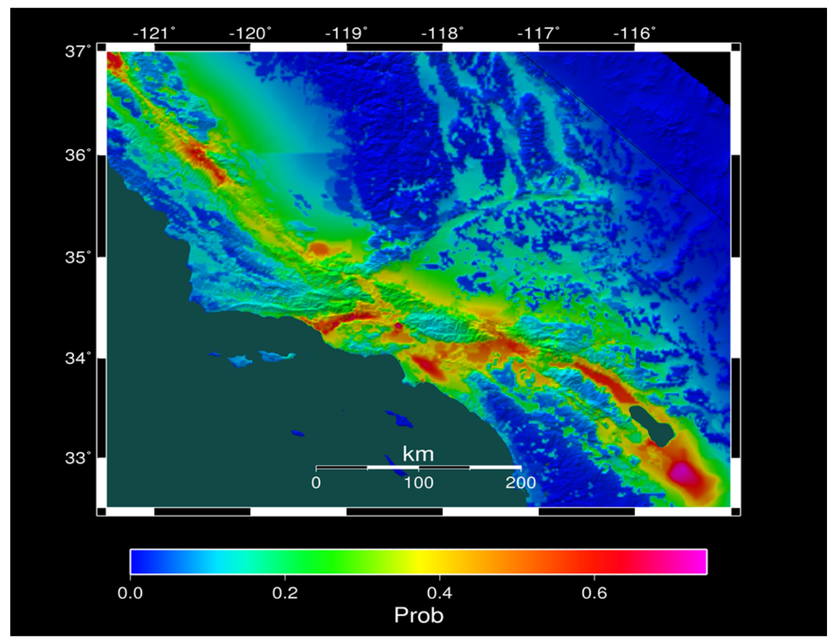


Seismic Hazard Analysis

- What will peak ground motion be over the next 50 years?
  - Used in building codes, insurance, planning
  - Answered via Probabilistic Seismic Hazard Analysis (PSHA)
  - Communicated with hazard curves and maps



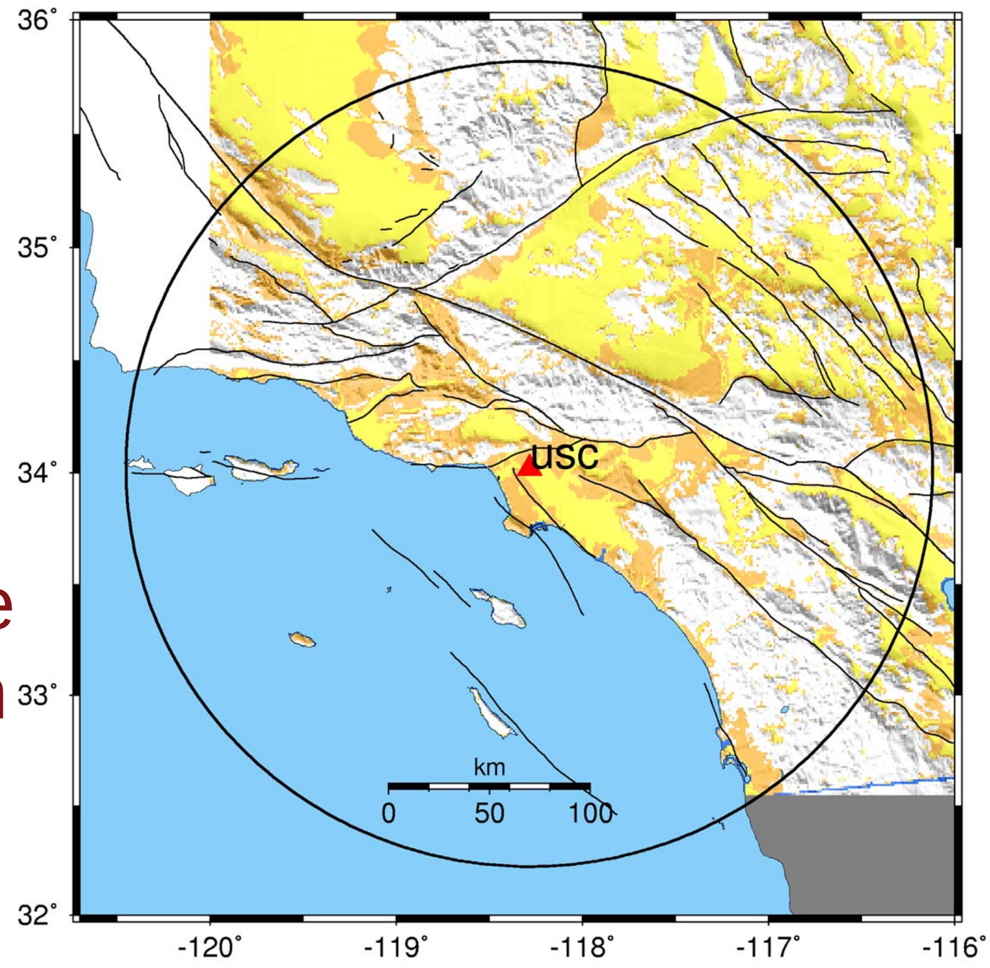
Hazard curve for downtown LA



Probability of exceeding 0.1g in 50 yrs

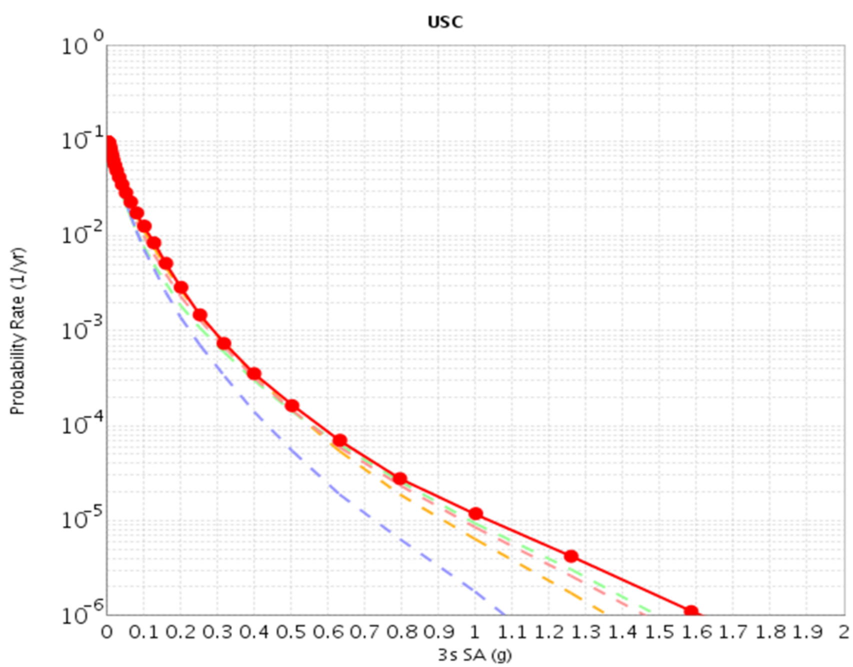
How to Calculate Seismic Hazard

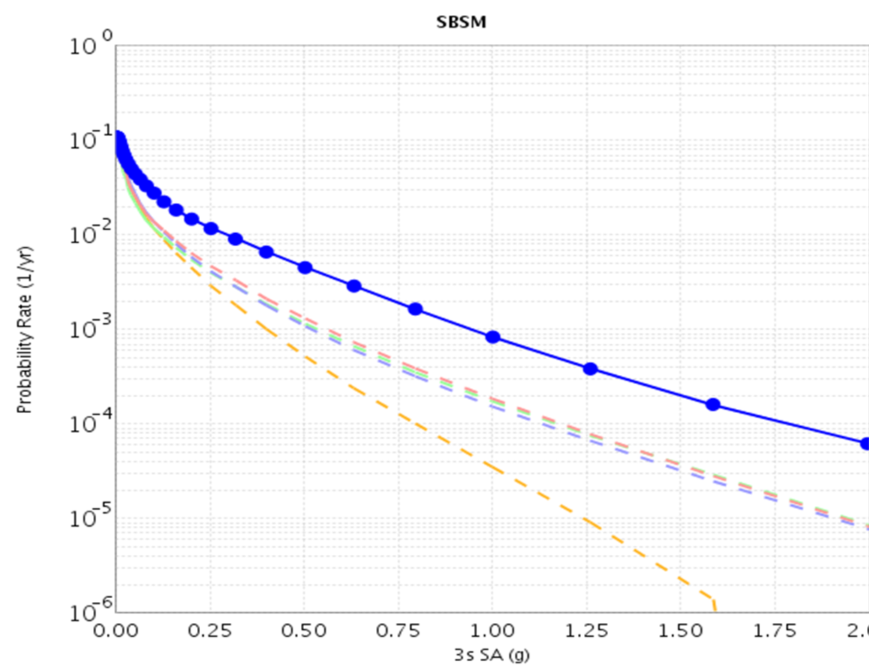
1. Pick a location of interest.
2. Determine what future earthquakes might happen which could affect that location.
3. Estimate the magnitude and probability for each earthquake using "earthquake rupture forecast"



4. Determine the shaking caused by each earthquake at the site of interest.
    - Two different strategies, each with pros and cons
5. Combine the levels of shaking with probabilities to produce a hazard curve.

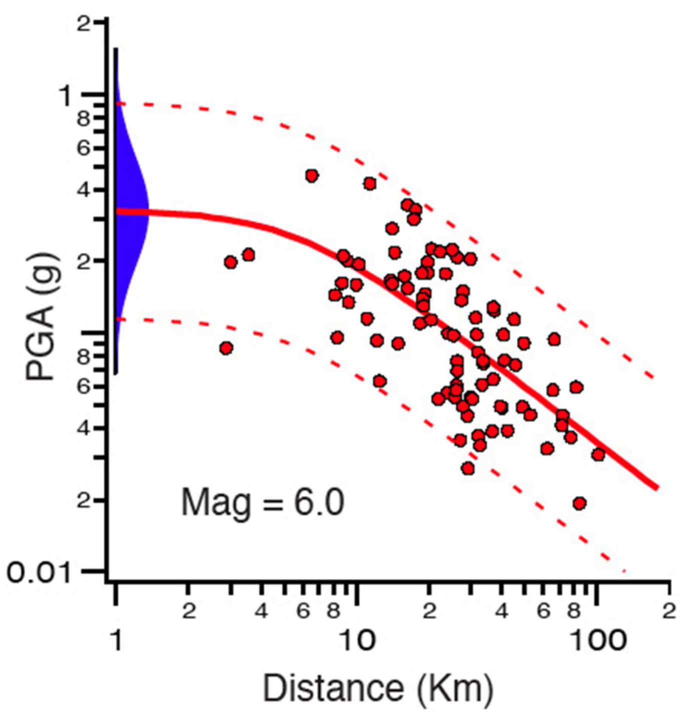
Repeat for many locations for a hazard map.

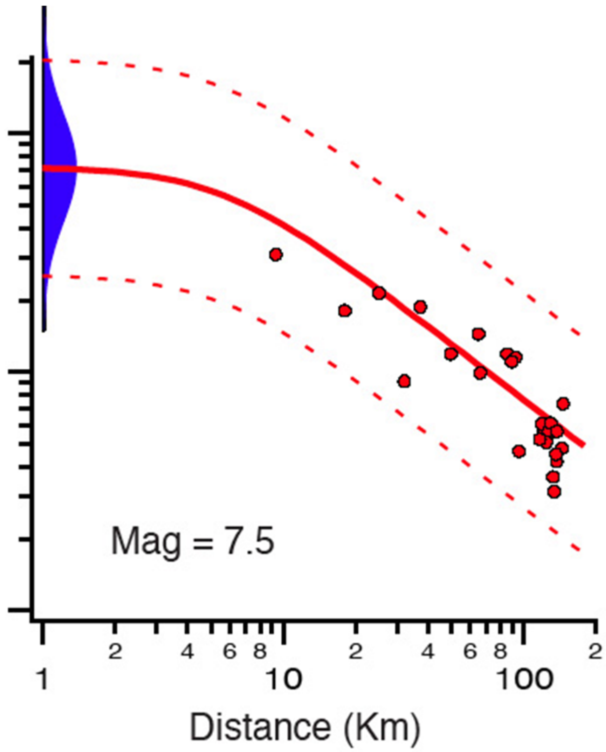


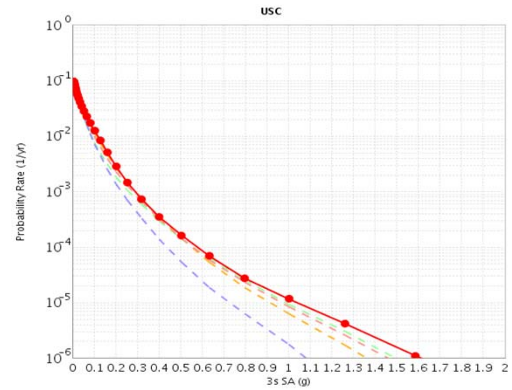


Option 1: Attenuation Relationships

- Extrapolate from historical data
- Based on what magnitude, how far
- Very quick, but (too?) simple.

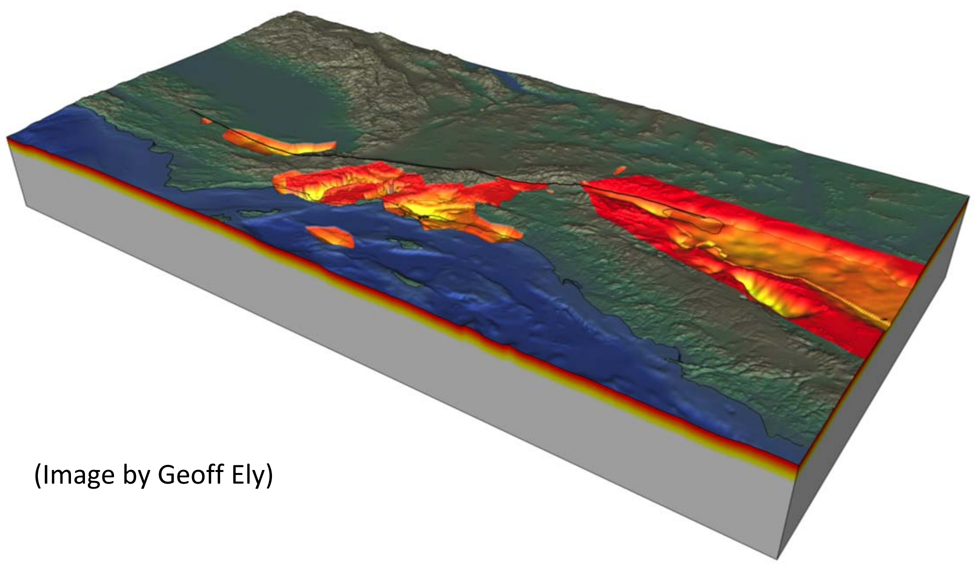






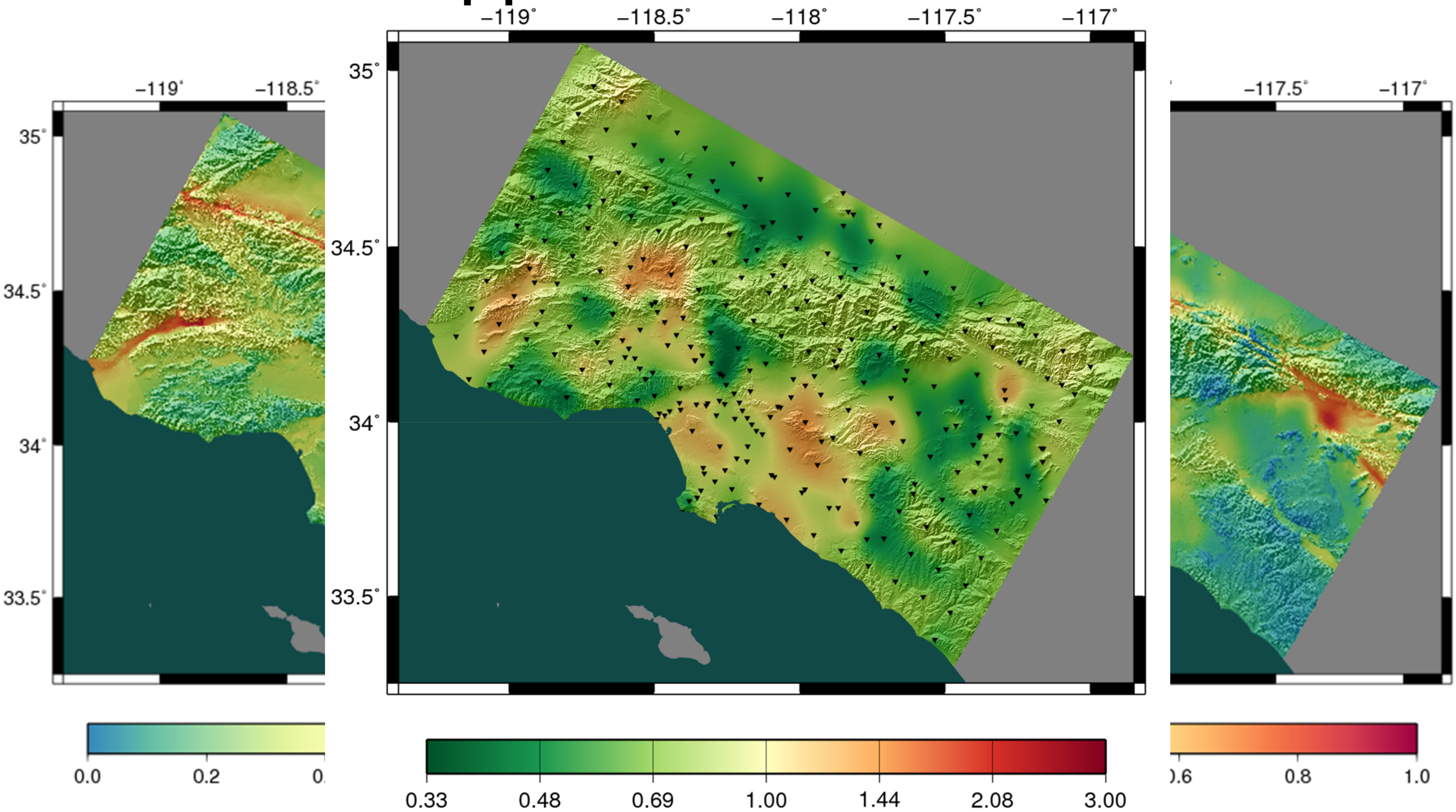
Option 2: Physics-Based Approach

- Alternatively, we can use a physical approach to simulate each earthquake
- SCEC does this in the “CyberShake” project
- Requires HPC: more expensive than attenuation approach



(Image by Geoff Ely)

Does the approach make a difference?



Attenuatio

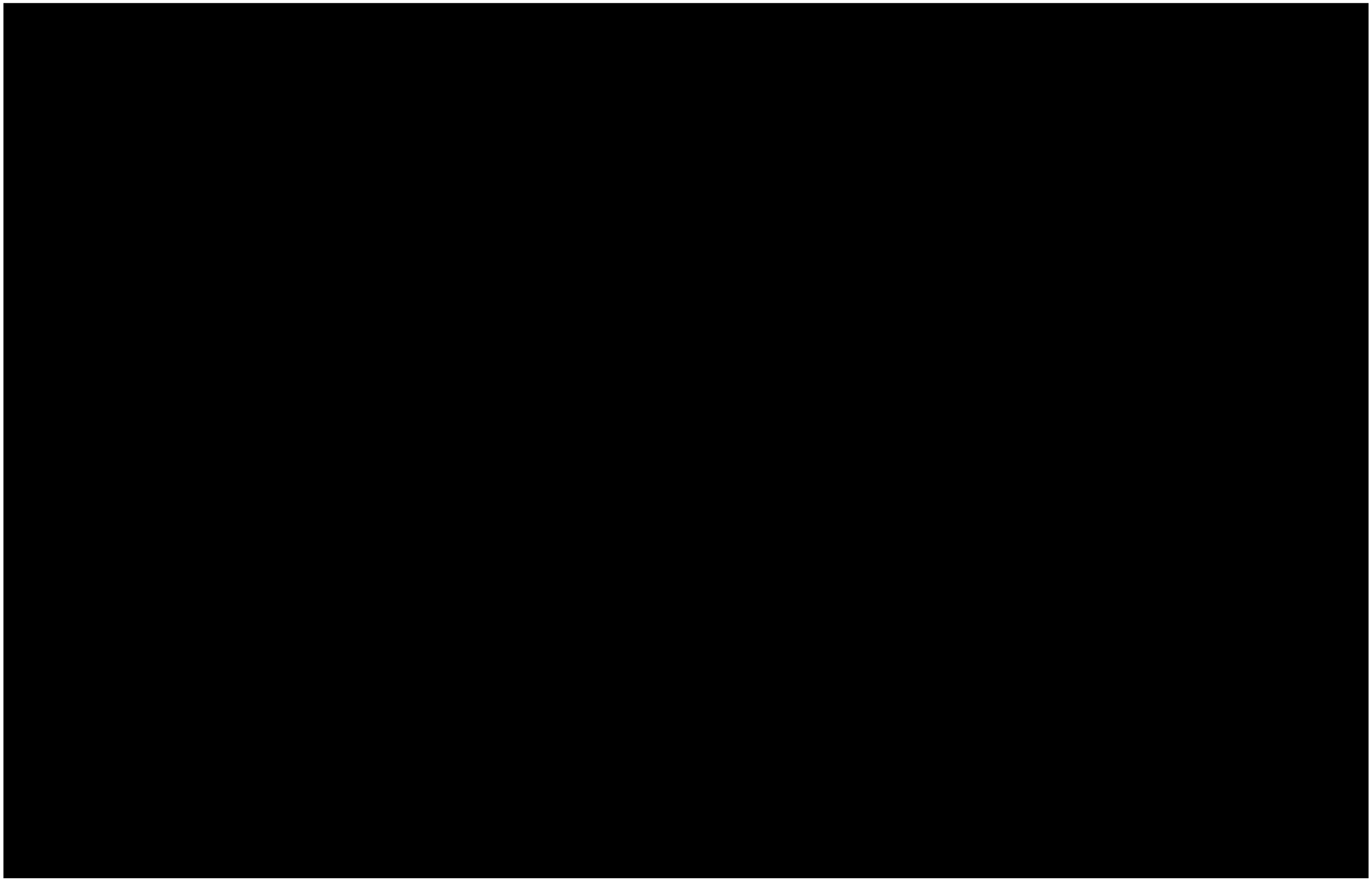
Higher Attenuation

Higher CyberShake

azard Map

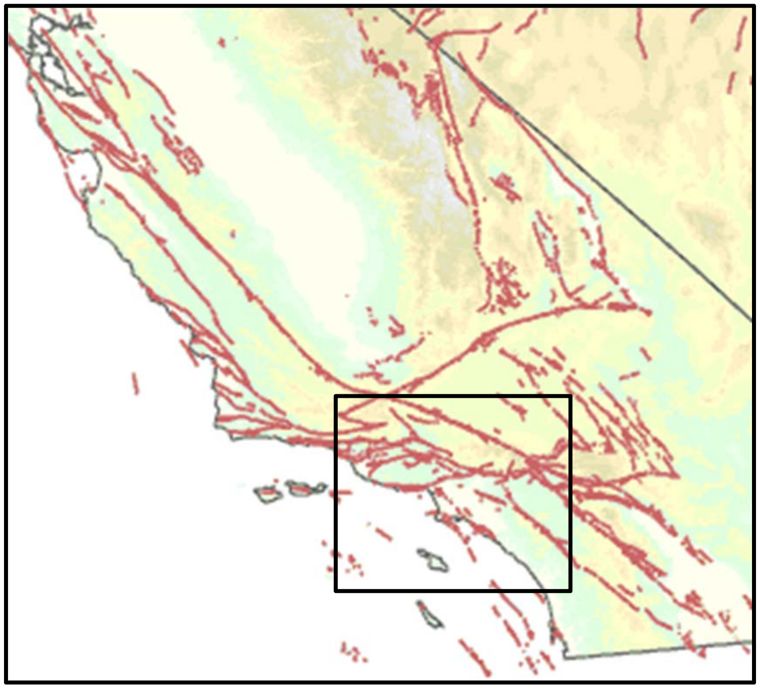
Simulation Results (N->S)

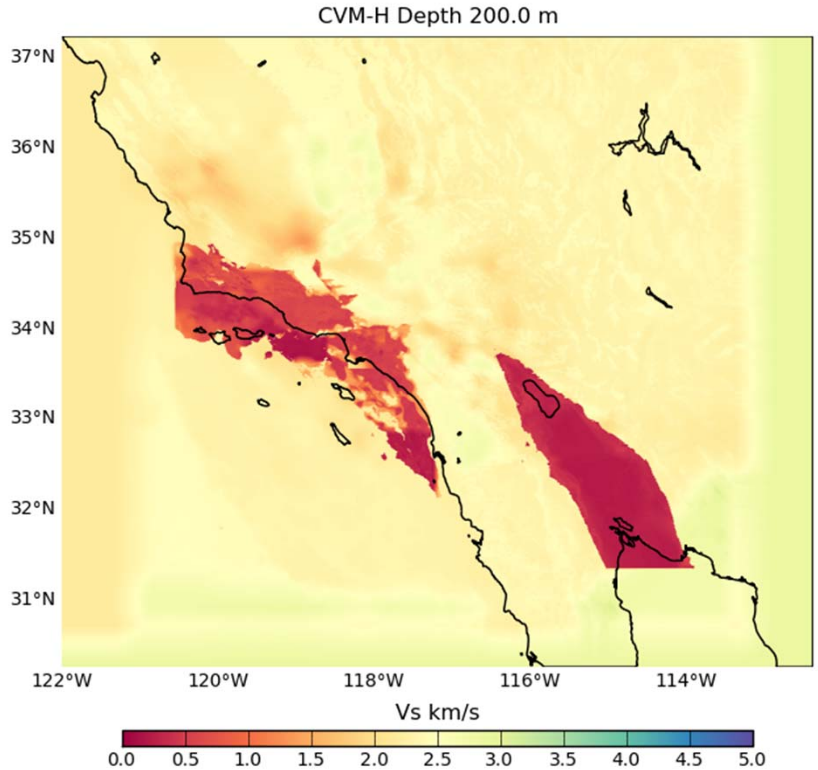
Simulation Results (S->N)



Physics-based CyberShake approach

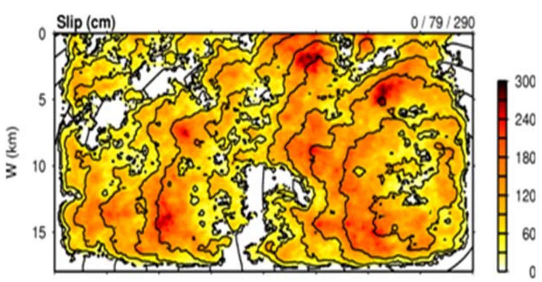
- Wave propagation simulation
  - Create 1.5 billion point mesh with material properties
  - Generate Strain Green Tensors across volume
  - Parallel, ~8,000 CPU-hrs

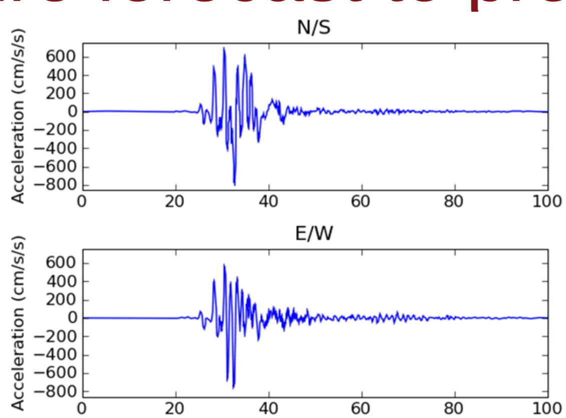


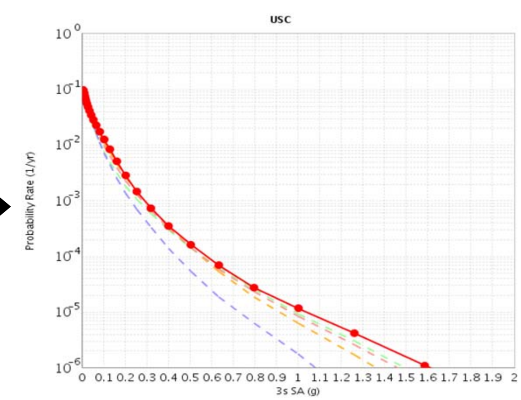


Post-Processing

- Individual earthquake contributions
  - Use “seismic reciprocity” to simulate seismograms for each of ~415,000 earthquakes
  - Loosely-coupled, short-running serial jobs
  - From each seismogram determine peak shaking (“peak spectral acceleration”)
- Combine the levels of shaking with probabilities from earthquake rupture forecast to produce hazard curve

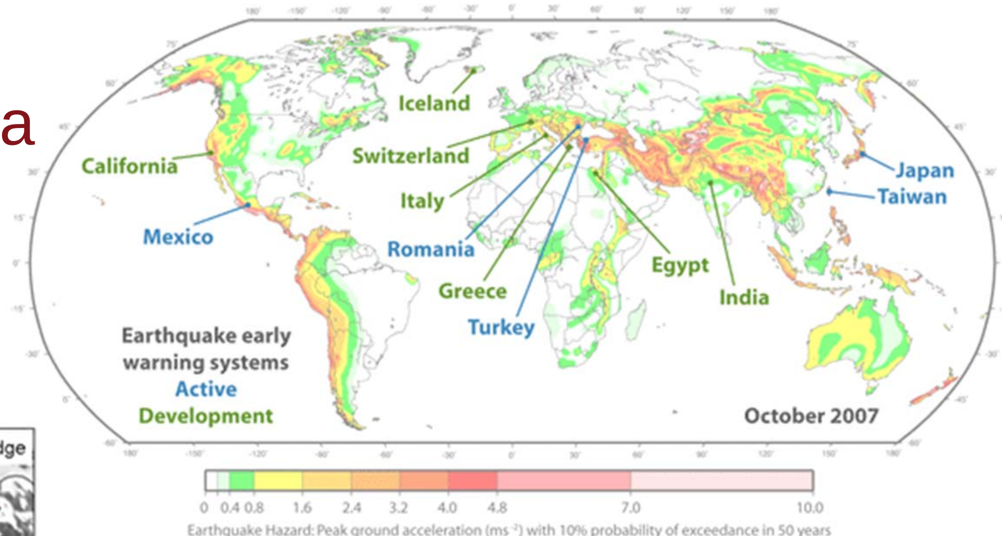


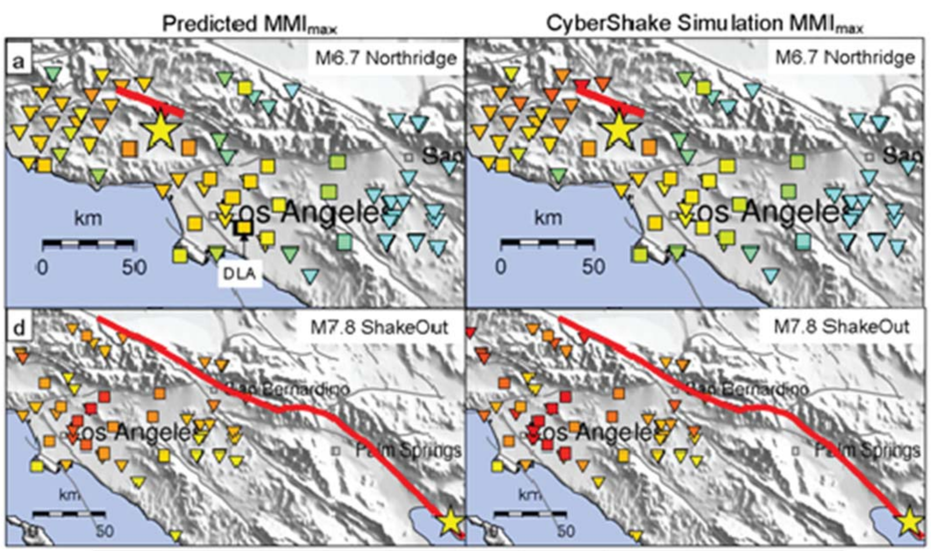




Earthquake Early Warning

- Detect earthquake at distance from populated area
- Send signal ahead
- Up to 60 sec warning

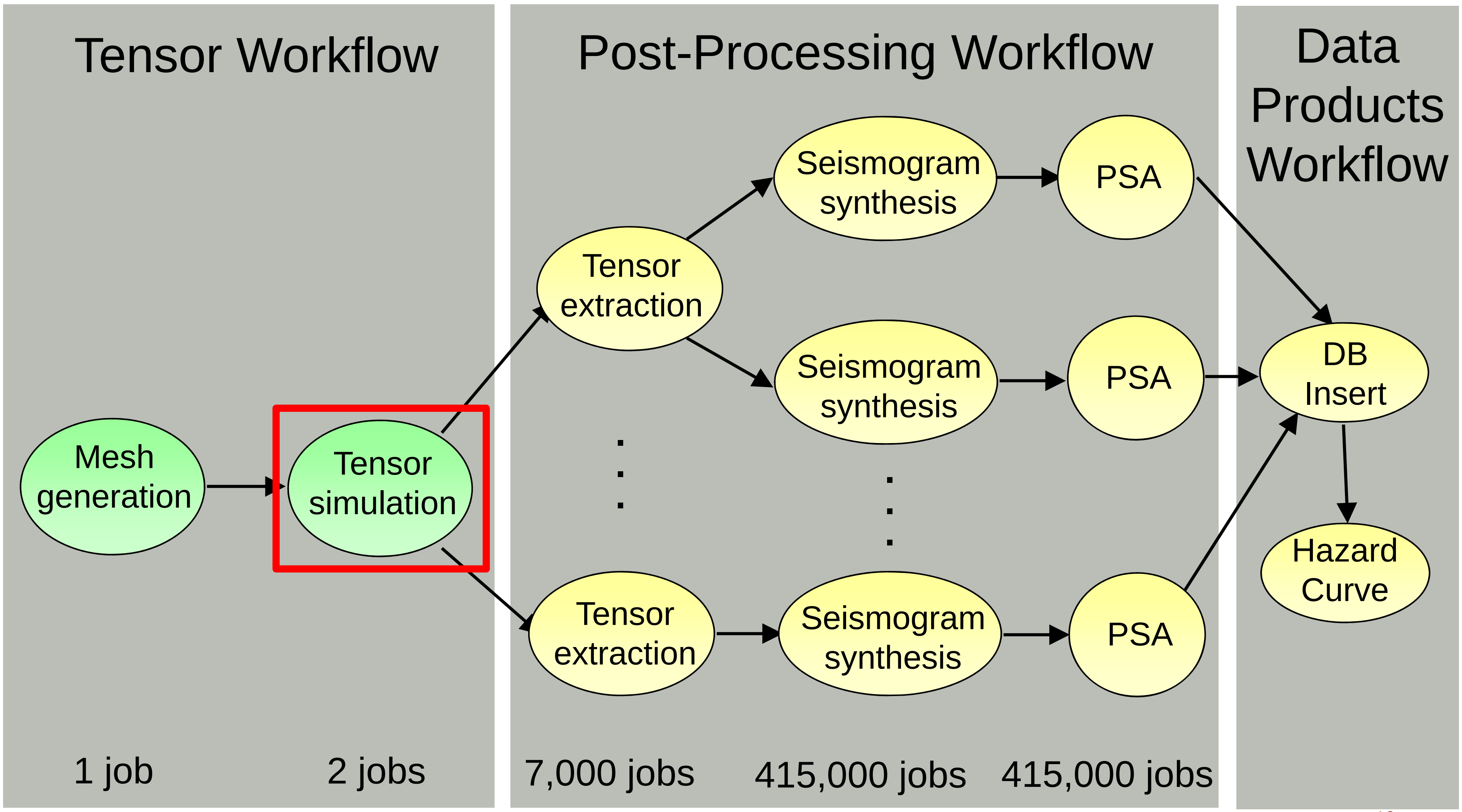




PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Mod./Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<0.05	0.3	2.8	6.2	12	22	40	75	>139
PEAK VEL.(cm/s)	<0.02	0.1	1.4	4.7	9.6	20	41	86	>178
INSTRUMENTAL INTENSITY	I	II–III	IV	V	VI	VII	VIII	IX	X+

- CyberShake seismograms used for training and testing machine learning algorithm
- Influenced EEW system for California

CyberShake workflows



Challenge 1: Strain Green Tensor Code

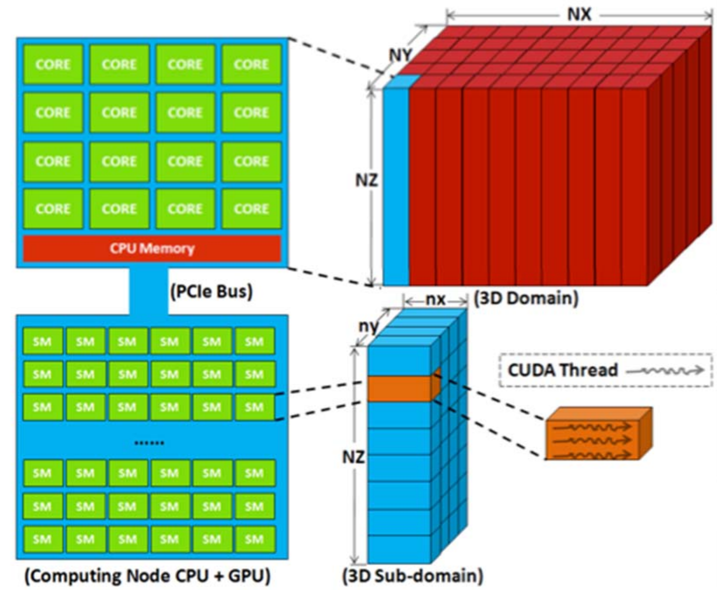
- 4^{th} order, staggered-grid, finite difference code
- 85% of CyberShake CPU-hours
- Used same SGT code since 2007
  - Readable, easy to use, interfaces with other software
  - Scaling limitations
    - Writes file-per-core, merged in post-processing
    - Synchronous MPI communication
    - Little single-core optimization
- Moved to alternative SCEC community code, AWP-ODC
  - Optimized starting in 2004

AWP-ODC enhancements

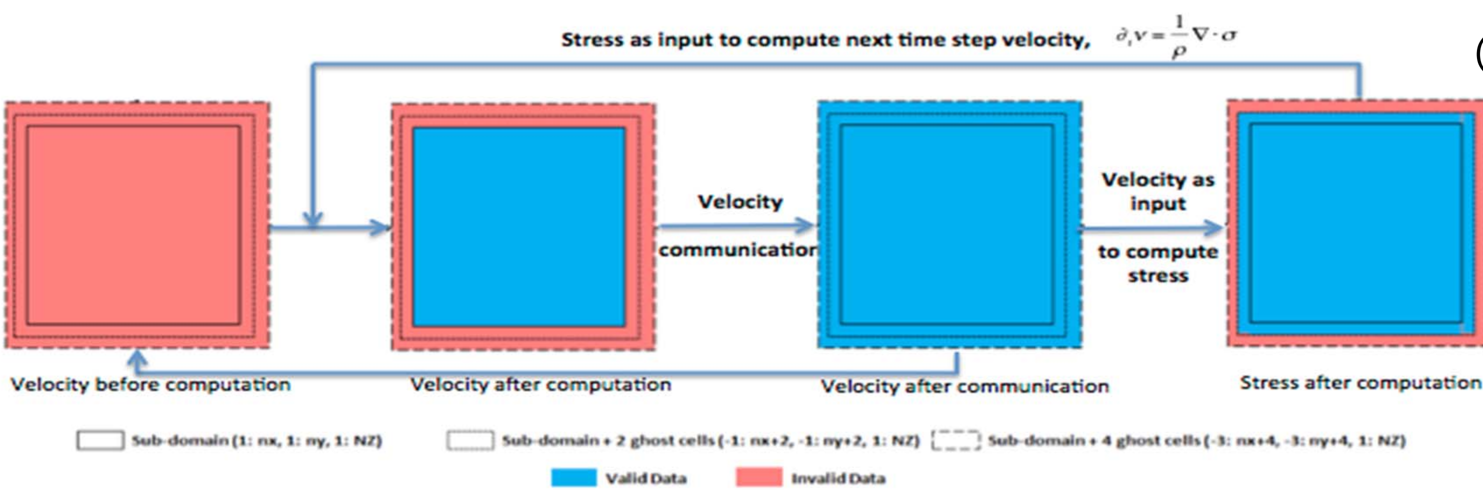
- Runtime per timestep of CPU version reduced 98%
- Two enhancements responsible for 82% of improvement
  - Asynchronous communication
  - Single-core optimization (no division, vectorization, loop unrolling, cache blocking, etc.)
- Leads to 100% weak scaling in CyberShake regime

AWP-ODC ported to GPU

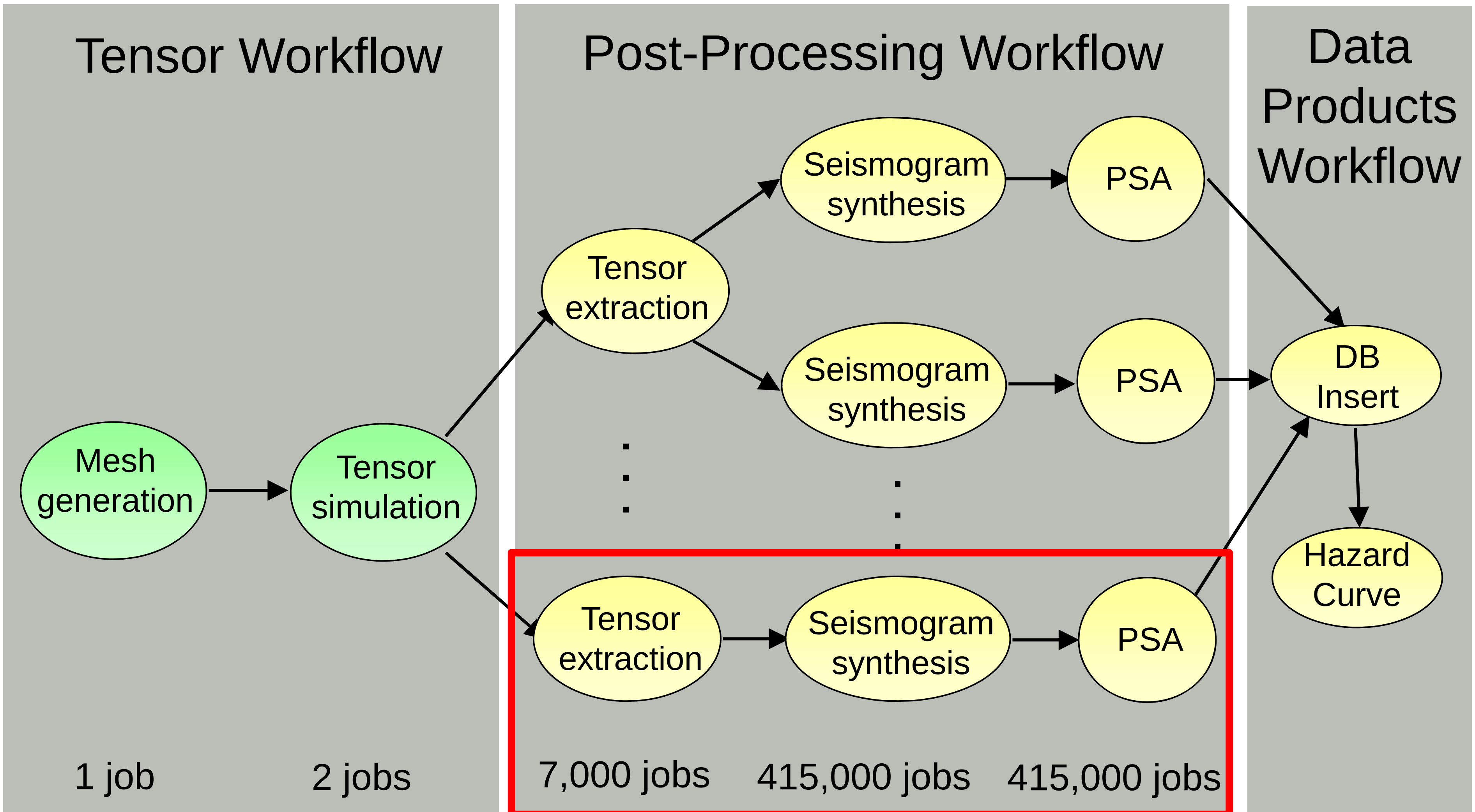
- 3D domain decomposition
  - Z-striping good for cache
  - Reduces # of neighbors
- Single-GPU optimizations
- Multi-GPU optimizations
  - Eliminate stress communication



(Zhou et al., ICCS'12)
(Cui et al., SC'13)



CyberShake workflows



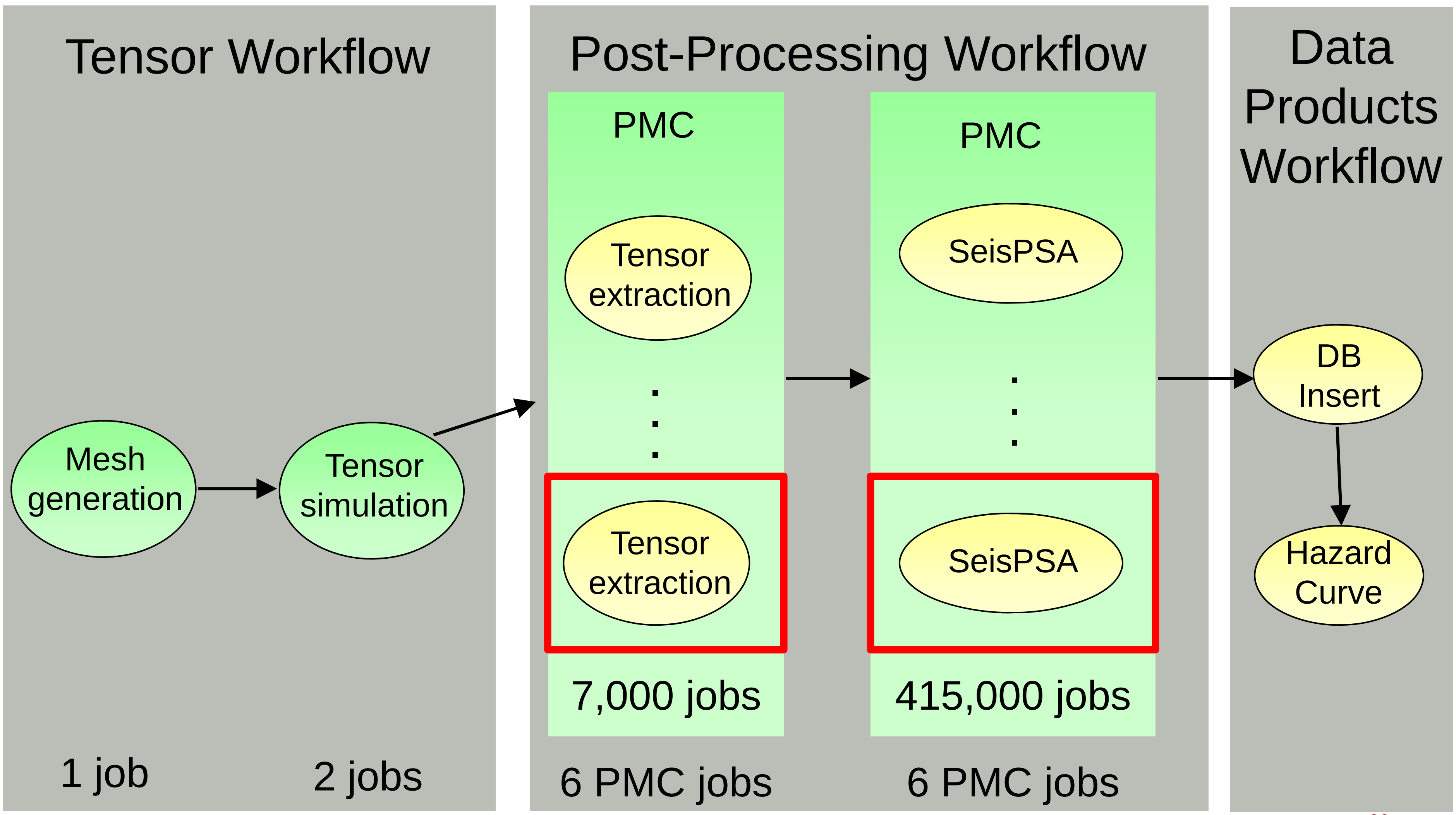
Challenge 2: High Throughput Jobs

- ~837,000 serial jobs per run
  - 0.1 to 60 sec
  - Mostly independent
- Combined seismogram and PSA jobs into one using C wrapper
  - “SeisPSA”
  - Half as many jobs
  - Also eliminates PSA job reading in seismogram file

High Throughput Scheduling

- Workflow tools required to manage the jobs
- Can't put them directly into the queue
  - Schedule can't handle millions of short jobs
  - Scheduler cycle is too slow (5+ minutes)
- Pegasus-mpi-cluster workflow tool (PMC)
- Workflow tools request chunk of cores, PMC manages task scheduling
- MPI wrapper around serial or thread-parallel jobs
  - Master-worker paradigm
  - MPI messaging has low latency

CyberShake workflows



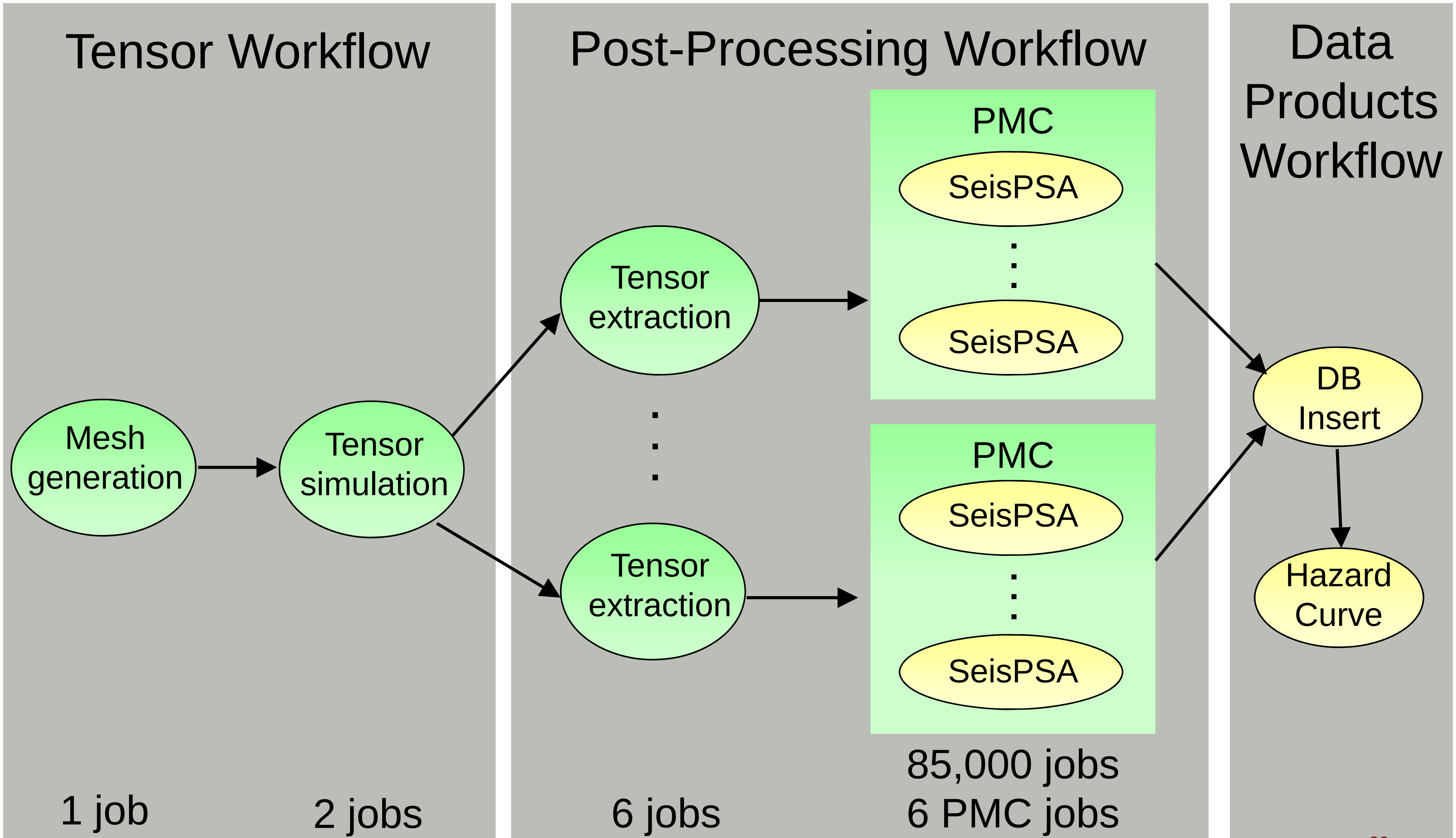
Challenge 3: I/O

- I/O load influenced by:
  - Amount of data
  - Number of reads and writes
  - Number of opens and closes
- Tensor extraction jobs
  - Read 40 GB, then write a subset
    - 40 GB x 7000 jobs = 273 TB of data read
  - Instead, restructure as MPI job
    - Read in 40 GB distributed among processors
    - Write many subsets
    - 40 GB x 6 jobs = 240 GB read = 99.9% improvement

I/O-memory-CPU tradeoff

- SeisPSA jobs
  - Read earthquake description and tensors
  - Write two files (350 bytes, 24 KB)
  - Groups of 2 to 1568 SeisPSA jobs share input files
- Reduce reads
  - Generate earthquake description on the fly from geometry
    - Exchange I/O for memory and CPU time
    - Use memcached library to explicitly cache rupture geometry
  - Batch jobs together to reuse tensors
    - Read tensors once, calculate multiple seismograms
- Reduce writes
  - Pegasus-mpi-cluster supports “pipe forwarding”
  - Workers write to pipes, master writes to 60x fewer files

CyberShake workflows



Challenge 4: Lengthy Runs

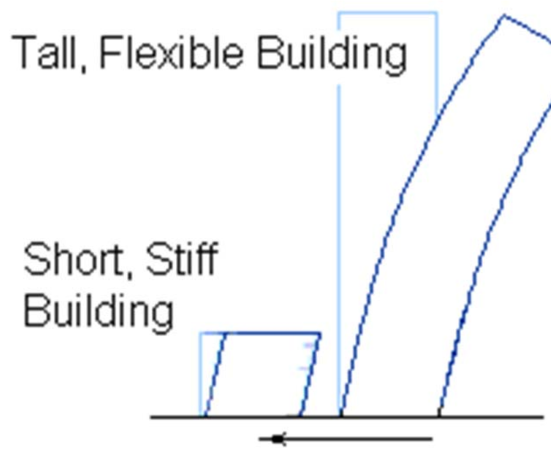
- Run 1100+ hazard curves (2+ weeks wallclock time)
- High degree of automation required
  - Workflow tools
    - No manual job submission
    - Automatic retries
  - Easy monitoring
    - Database tracks run states
    - Email notifications
  - Data products generated automatically

CyberShake Study 14.2 Metrics

- 1144 hazard curves (4 maps) on NCSA Blue Waters
- 342 hours wallclock time (14.25 days)
- 46,720 CPUs + 225 GPUs used on average
  - Peak of 295,040 CPUs, 1100 GPUs
- GPU SGT code 6.5x more efficient than CPU
- 99.8 million jobs executed (81 jobs/second)
  - 31,463 jobs automatically run in the Blue Waters queue
- On average, 26.2 workflows (curves) concurrently

Future (Science) Directions

- Higher frequencies
  - Buildings are most affected by *frequency = 10 / (height in floors)*
  - Currently at 0.5 Hz, moving to 1 Hz
  - 2x frequency -> 16x computational work



Future (Technical) Directions

- Improve post-processing
  - Currently extraction reads 40 GB of data, writes 690 GB
  - At 1 Hz, would read 1 TB, write 17 TB
  - Make extraction more clever?
  - Remove extraction entirely?
- Coscheduling
  - GPU XK nodes have 1 GPU, 16 CPUs
  - While running SGTs on GPUs, schedule post-processing to CPUs
  - Difficult to explain to workflow tools (2 workflows, 1 job)

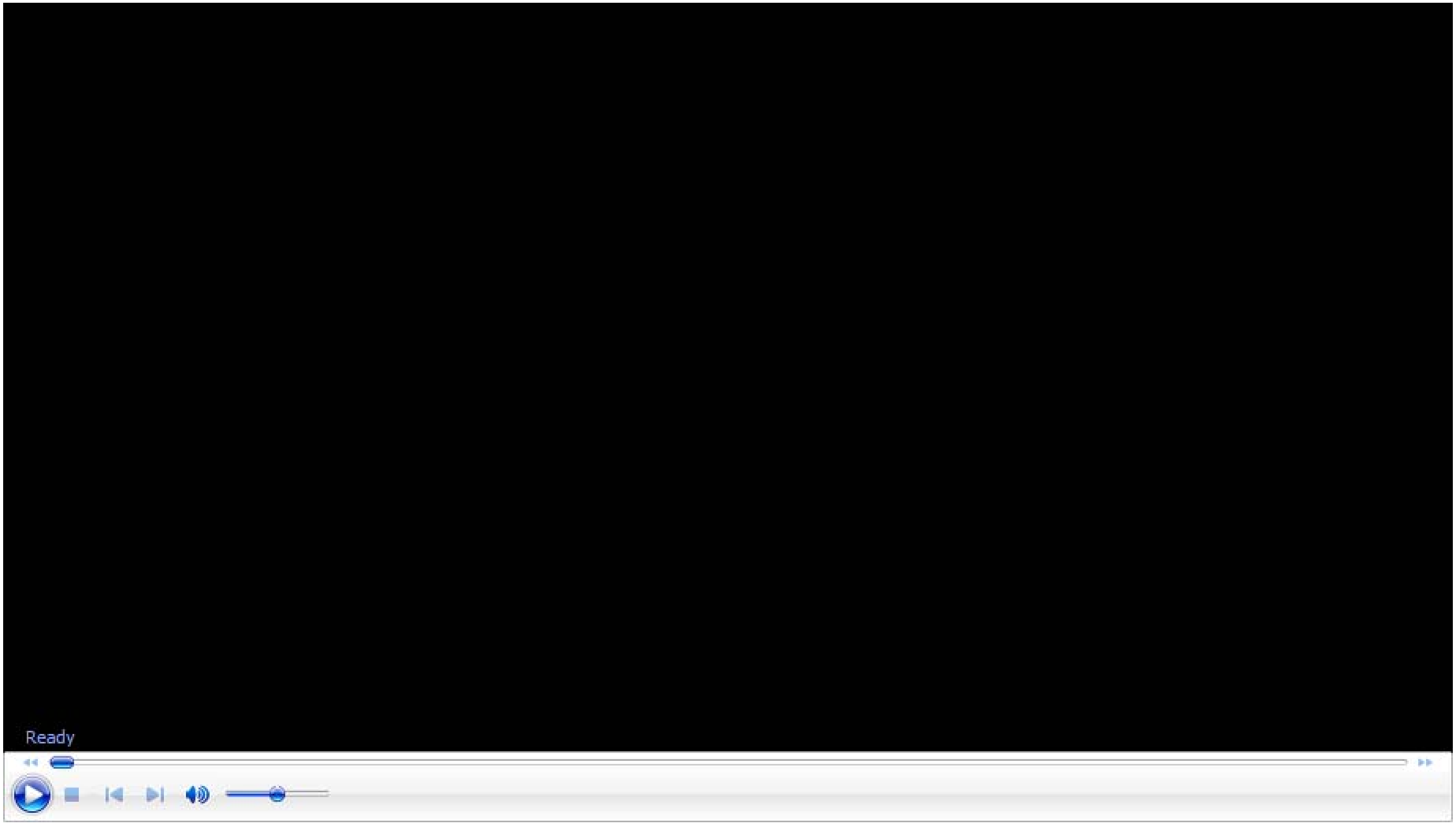
Things we wish we knew: Workflow Tools

- Help to:
  - Automate processes
  - Manage data
  - Gather metadata
  - Do more than 1 person can do manually
- CyberShake not possible without them

Verification

- Verification
  - Does this code behave as I expect it to? Was it programmed correctly?
- Construct verification test problem
  - Identify small test problem (ideally, automated)
  - Generate reference solution
  - Run after all code modifications, compare to reference
  - If discrepancy is “too much”, dive in deeper
  - Can also use to quantify impact of change

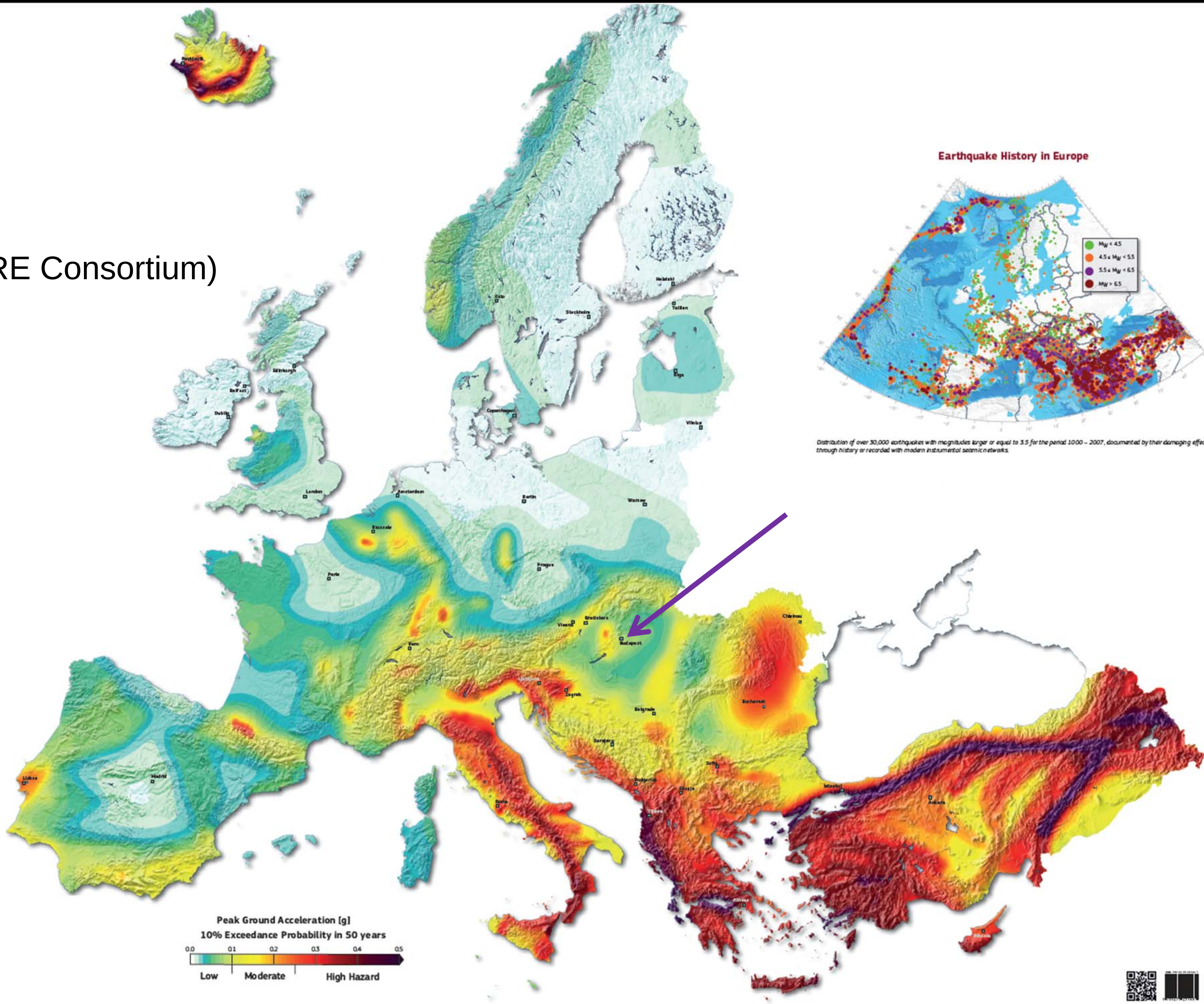
Codes Comparison



Software Engineering Best Practices

- Version control
  - Can always go back
  - Track what code was used with which simulation
- Alternate production and development cycles
  - Gives time for both optimization and science results
- Incremental improvement
  - "Premature optimization is the root of all evil."
    Donald Knuth
  - Wait until you have identified a limiting factor
  - Cost/benefit analysis: how much development time for how much performance gain?

(SHARE Consortium)



Distribution of over 30,000 earthquakes with magnitudes larger or equal to 3.5 for the period 1000 – 2007, documented by their damaging effects through history or recorded with modern instrumental seismic networks.