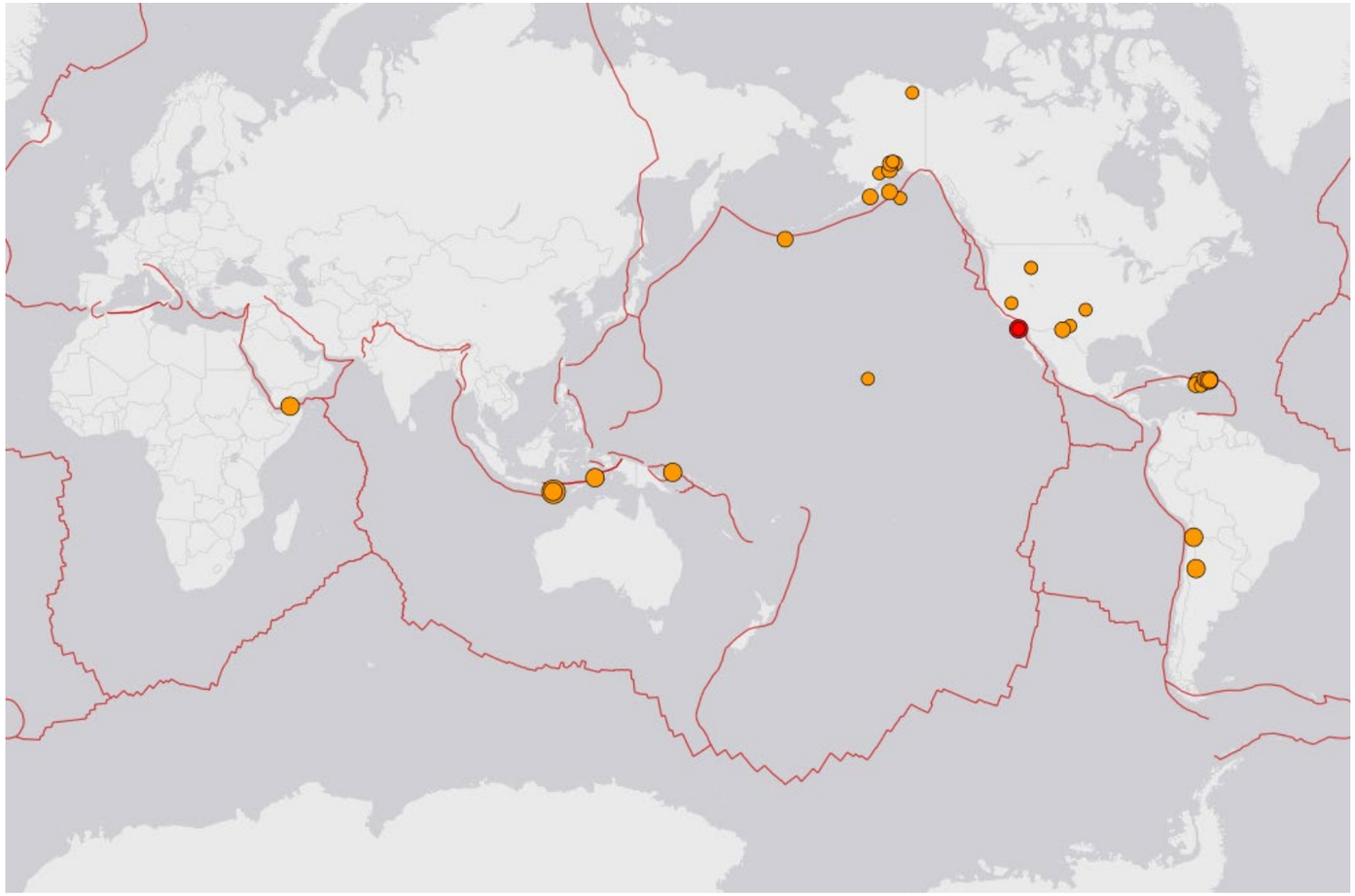
SCEC

HPC Challenges in Seismology: There's a Whole Lotta Shakin' Goin' On

Scott Callaghan
scottcal@usc.edu

Thursday, July 13, 2023
2023 IHPCSS

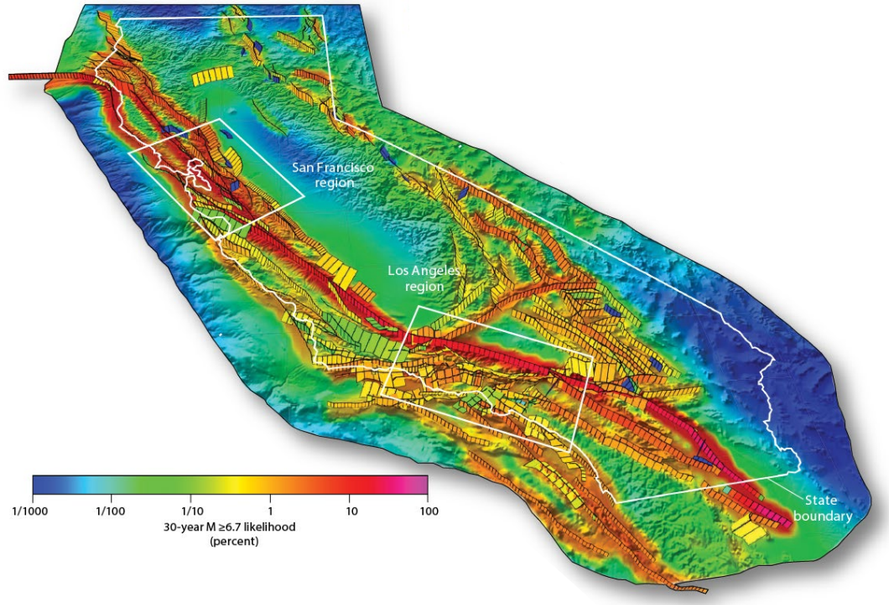
- In 2023 so far:
  - M7-8: 13
  - M8+: 0
- Potential for large societal impact
  - 2023 Turkey (M7.8 & M7.5)
  - 2021 Haiti (M7.1)
  - 2011 Tohoku (M9.1)
- So far, prediction is elusive

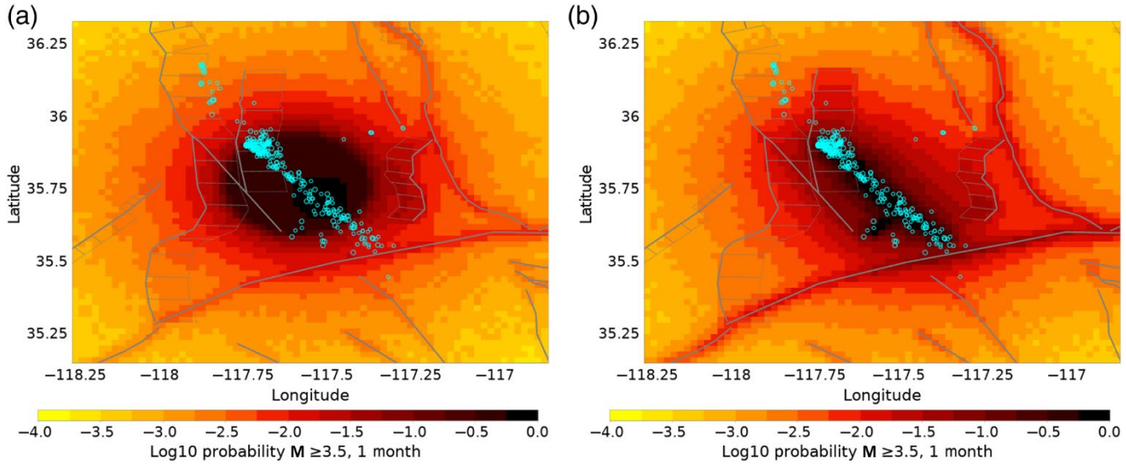


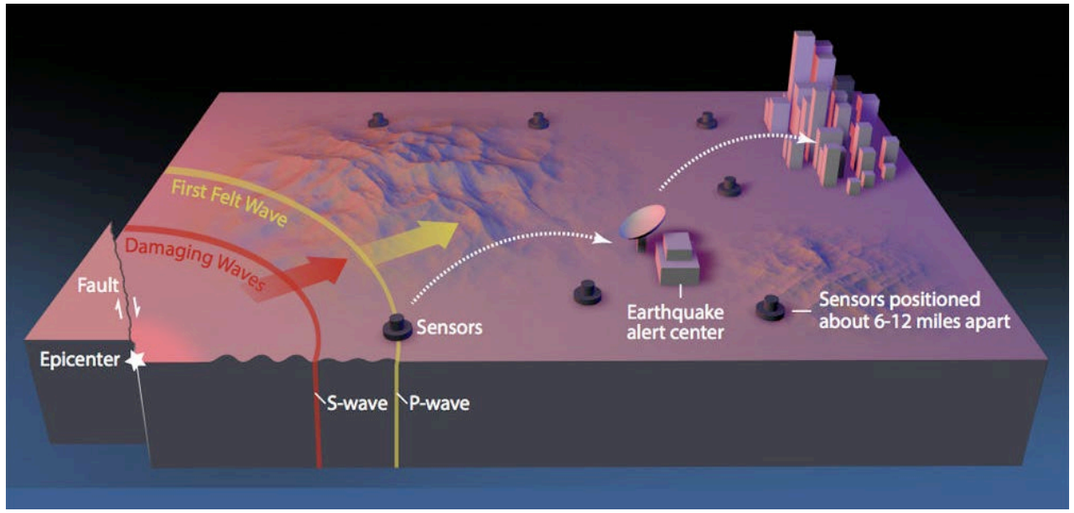
35 earthquakes, M2.5+, last 24 hrs, max M5.0

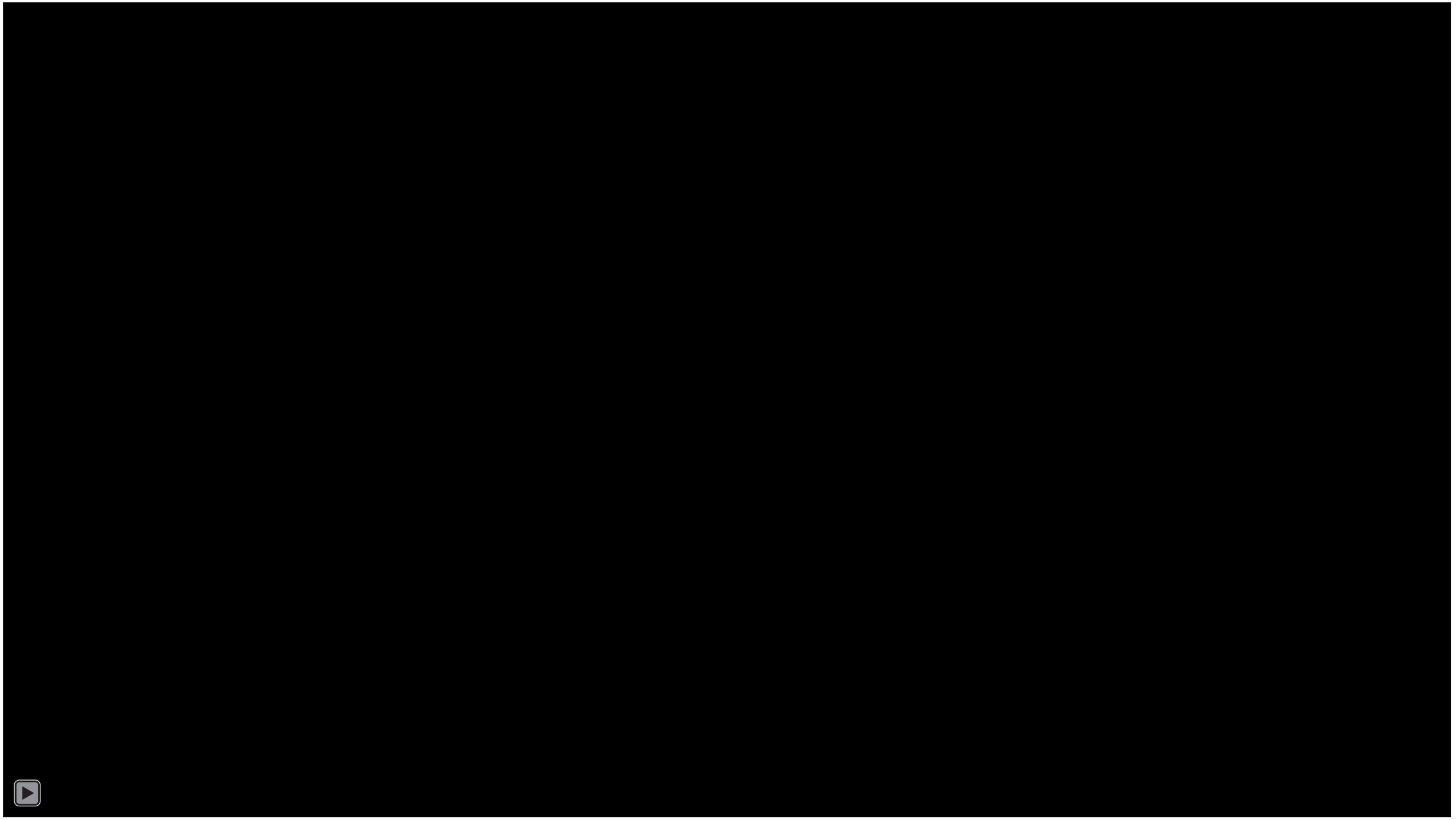
Earthquake Forecasting

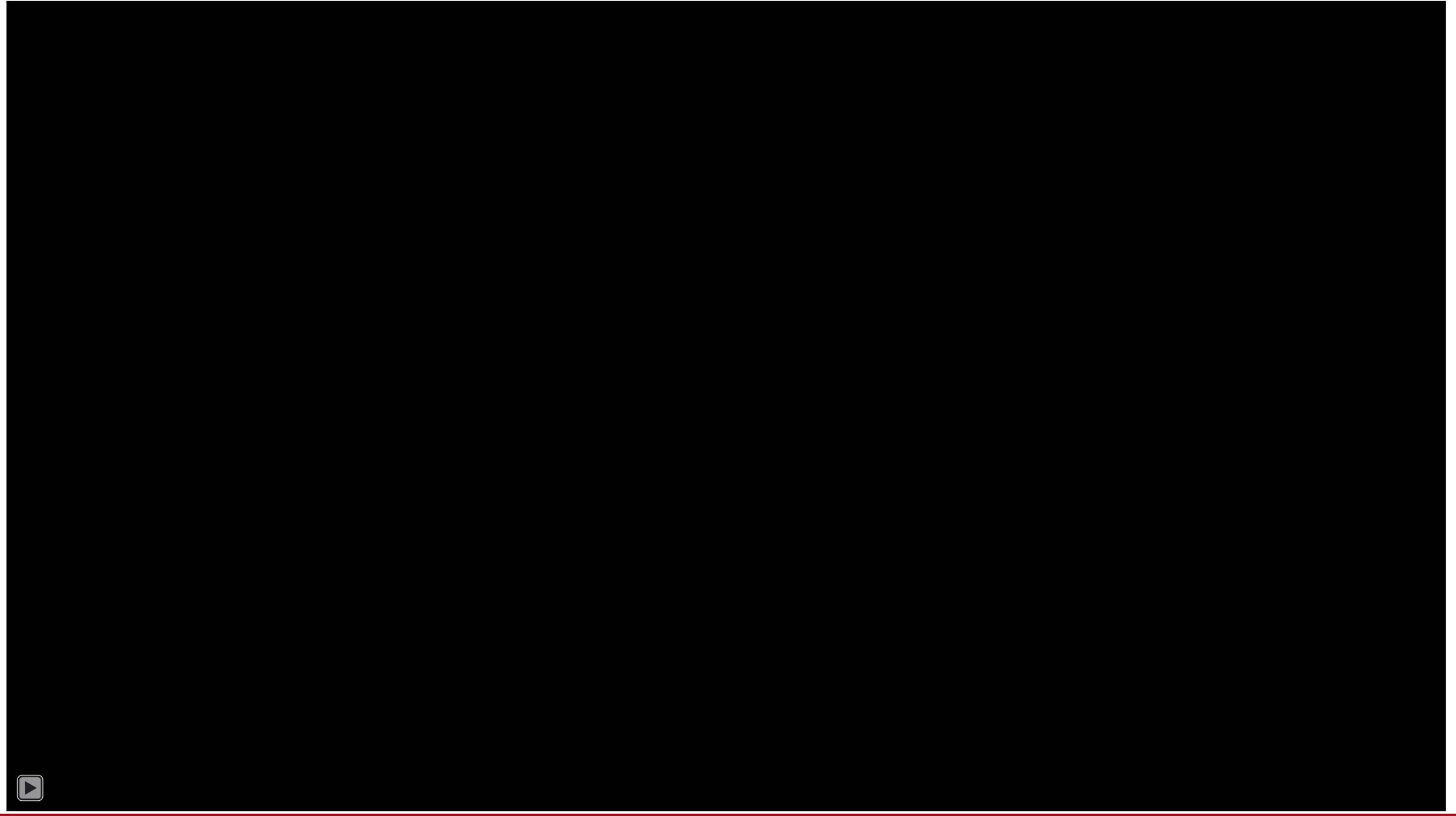
- If we can't predict, what can we do?
- Earthquake forecasts
  - Long-term
  - Short-term
- Earthquake early warning
- Simulations of individual events





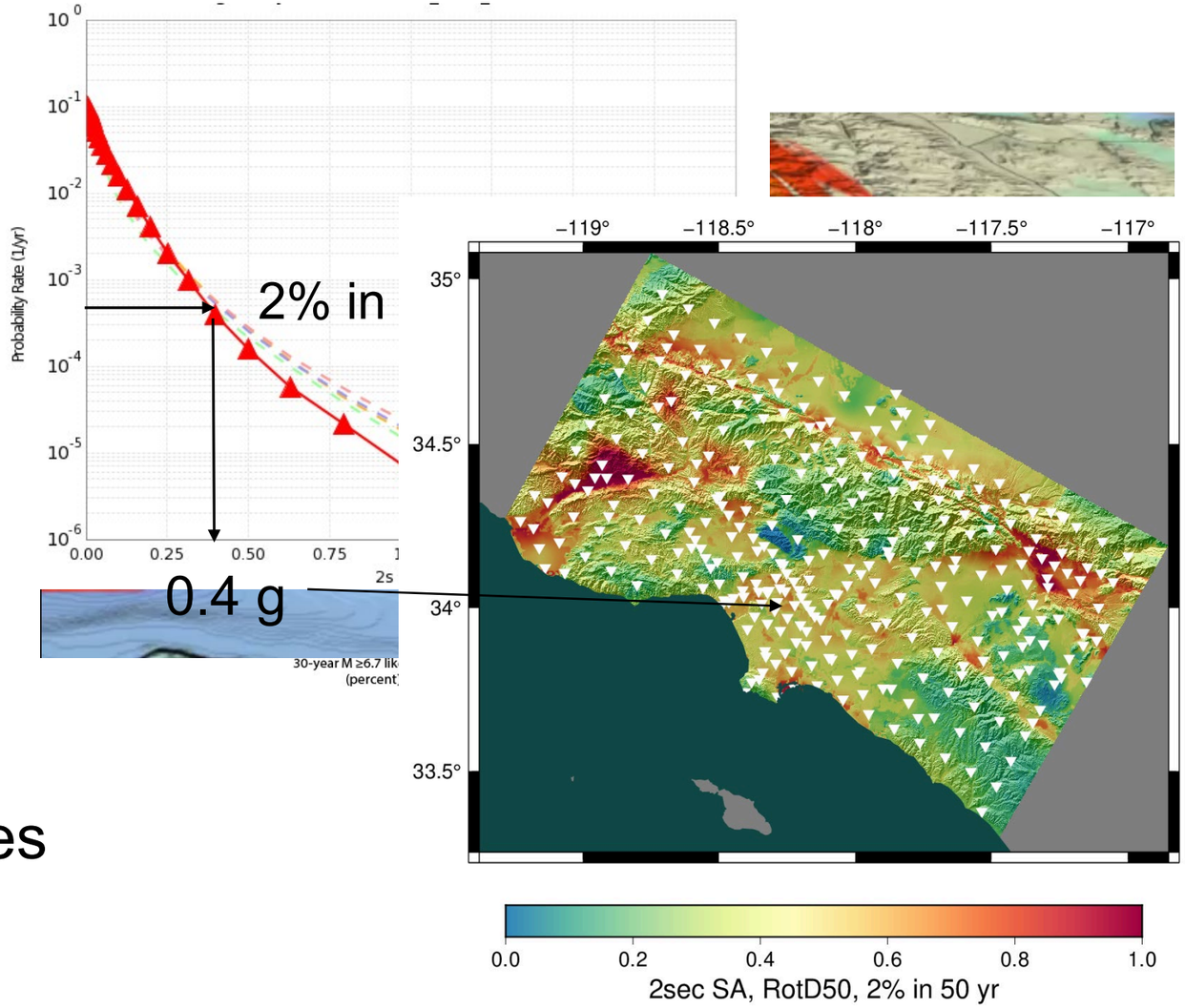






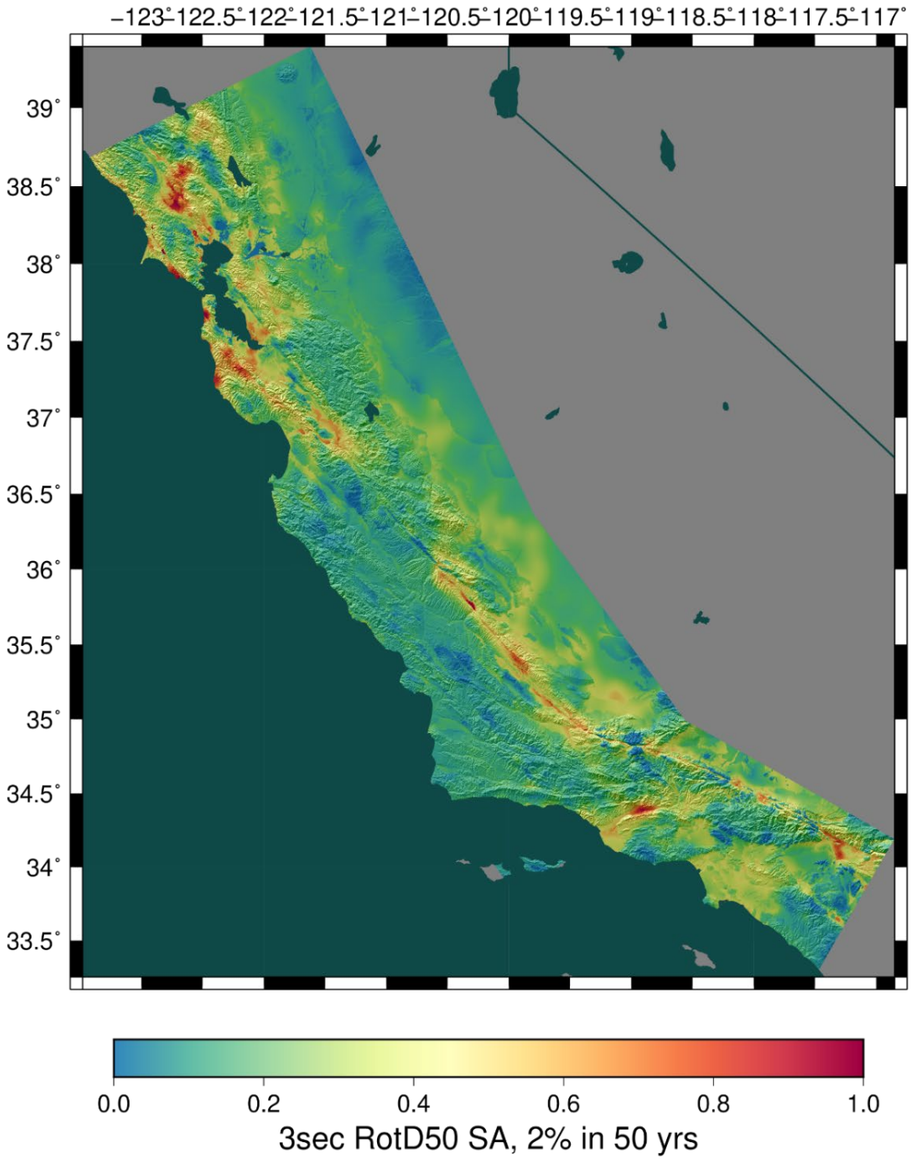
Seismic Hazard

- What ground motions can I expect in the next 50 years?
  - Building engineers
  - Insurance companies
  - Disaster planners
- Answered via Probabilistic Seismic Hazard Analysis (PSHA)
  1. Get a list of all possible earthquakes
  2. Determine how much shaking each earthquake causes
  3. Combine shaking with earthquake probability to generate hazard estimates



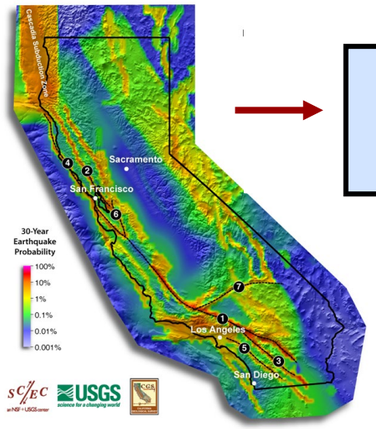
CyberShake platform

- CyberShake was developed by SCEC to perform 3D physics-based PSHA
- Uses wave propagation simulations to determine shaking from each earthquake
- To reduce computational cost, utilizes reciprocity
  - 2 simulations per site rather than 1 per event
  - (# of sites) << (# of events)
- Shaking measures derived from seismograms
- Shaking measures combined with probabilities for site and regional hazard



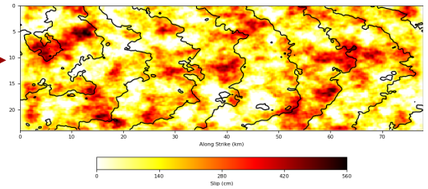
Parallel Jobs

Serial Jobs



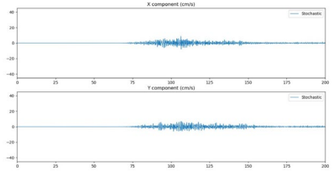
Uniform California Earthquake Rupture Forecast

Rupture generator



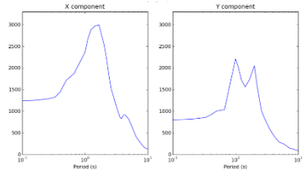
600,000+ events in memory

High-frequency seismogram synthesis

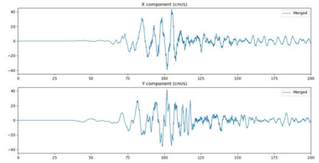


1-50 Hz seismograms
190 GB

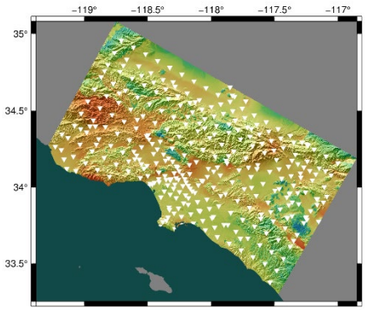
Merge and combine



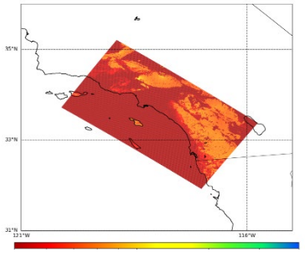
Period-dependent shaking measures
<1 GB



0-50 Hz broadband data products

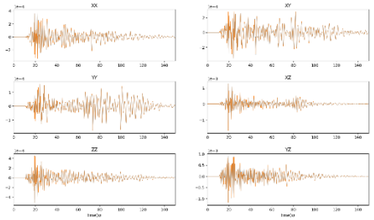


Velocity mesh generator



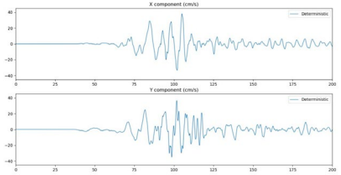
Velocity Mesh
300 GB

Wave propagation code (4th order FD, GPU)



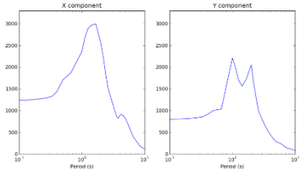
Strain Green Tensors
2 x 750 GB

Low-frequency seismogram synthesis



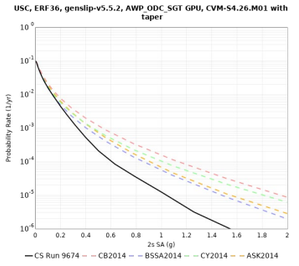
0-1 Hz low-frequency seismograms
40 GB

Intensity measures

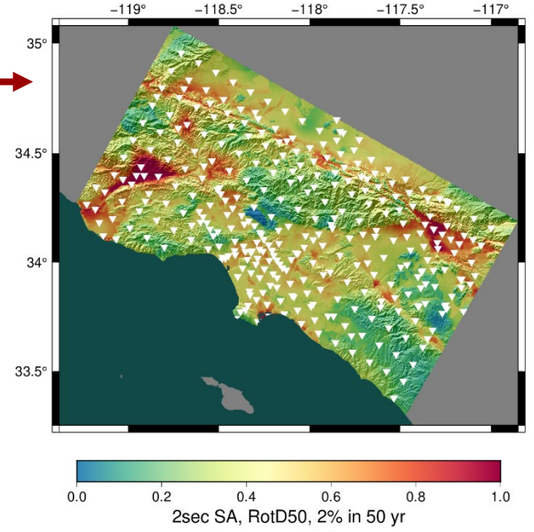


Period-dependent shaking measures
<1 GB

Aggregate data products



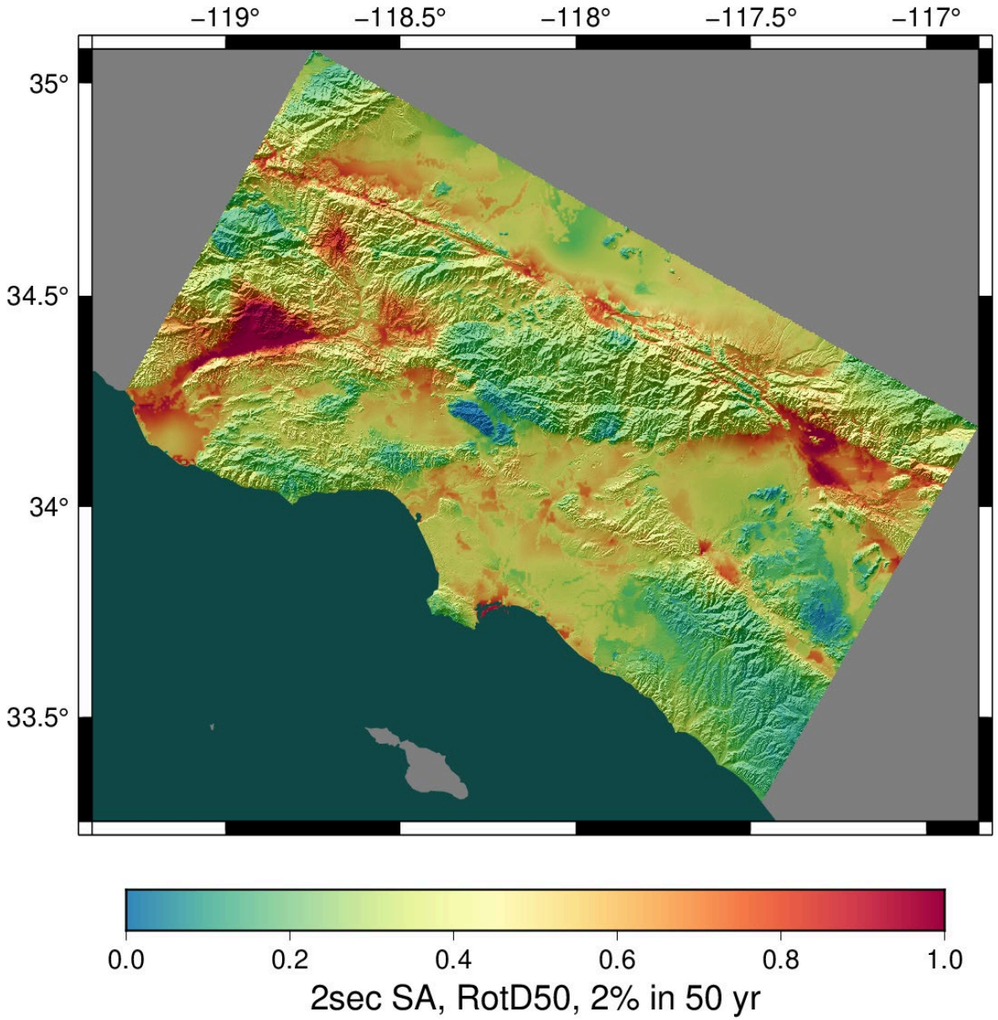
Hazard Curve



Hazard Map

Recent Results

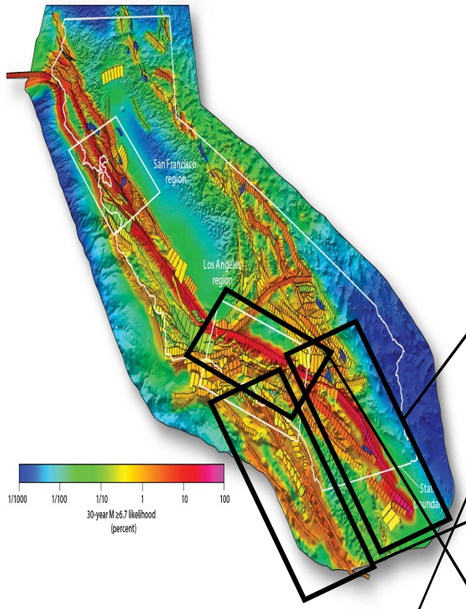
- Regional hazard calculation for Southern California
- 95 days of around-the-clock execution
- Used 772,000 node-hours on *Summit*
  - Peak of 73% of the system (3382 nodes)
- 26 million computational tasks run in 28,130 workflow jobs
- 2.5 PB of total data
- Staged 74 TB / 19 M files to long-term storage



Resolved Technical Challenges – pre & post processing

- Additional codes not included in the flow chart
  - Create configuration files
  - Write metadata needed to parse data files
  - Reorder files to be fast in time instead of fast in space
- As we scaled up, each of these became a bottleneck
  - Parallelized them with straightforward MPI
    - Manager process reads in configuration information
    - Broadcasts information to other processes to figure out their work
    - Processes do their work
    - Depending on size of output, either manager aggregates and writes output, or MPI collective I/O
- This simple parallelism reduced runtimes back to noise

Resolved Technical Challenges – Intermediate Files



Strain Green Tensors

4.5 GB

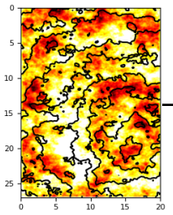
Extracted SGTs, rupture 1

Extracted SGTs, rupture 2

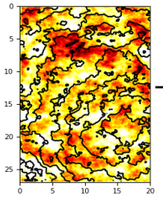
…

Extracted SGTs, rupture N_r

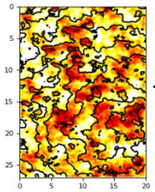
480 TB



Event 1



Event 2



Event N_e

300 GB

Seismogram, event 1

Seismogram, event 2

…

Seismogram, event N_e

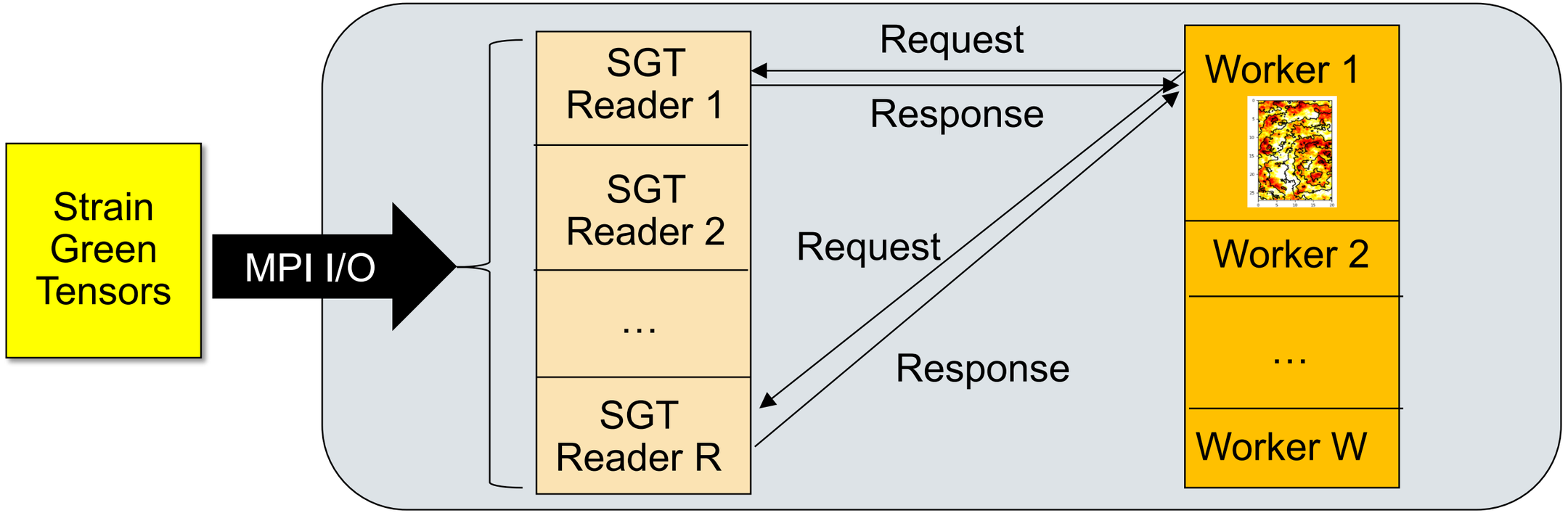
40 GB

0.5 Hz simulation

1.5 Hz simulation

Resolved Technical Challenges – Intermediate Files

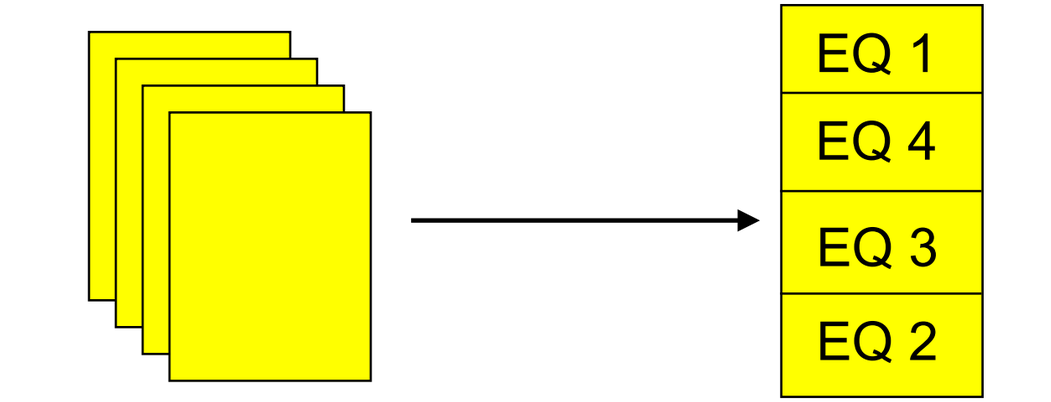
- Rewrote this stage into MPI manager-worker job



- Reduced I/O by 99%
- Enabled 1 Hz calculations

Resolved Technical Challenges – Number of Files

- Aggregated files to reduce total number by factor of 85



- Output files relatively small: ~10 Mbps/sec write
- Output data sent to manager process, then written out
  - Reduces filesystem load
  - Reduces challenge of synchronizing files

Resolved Technical Challenges – File Integrity

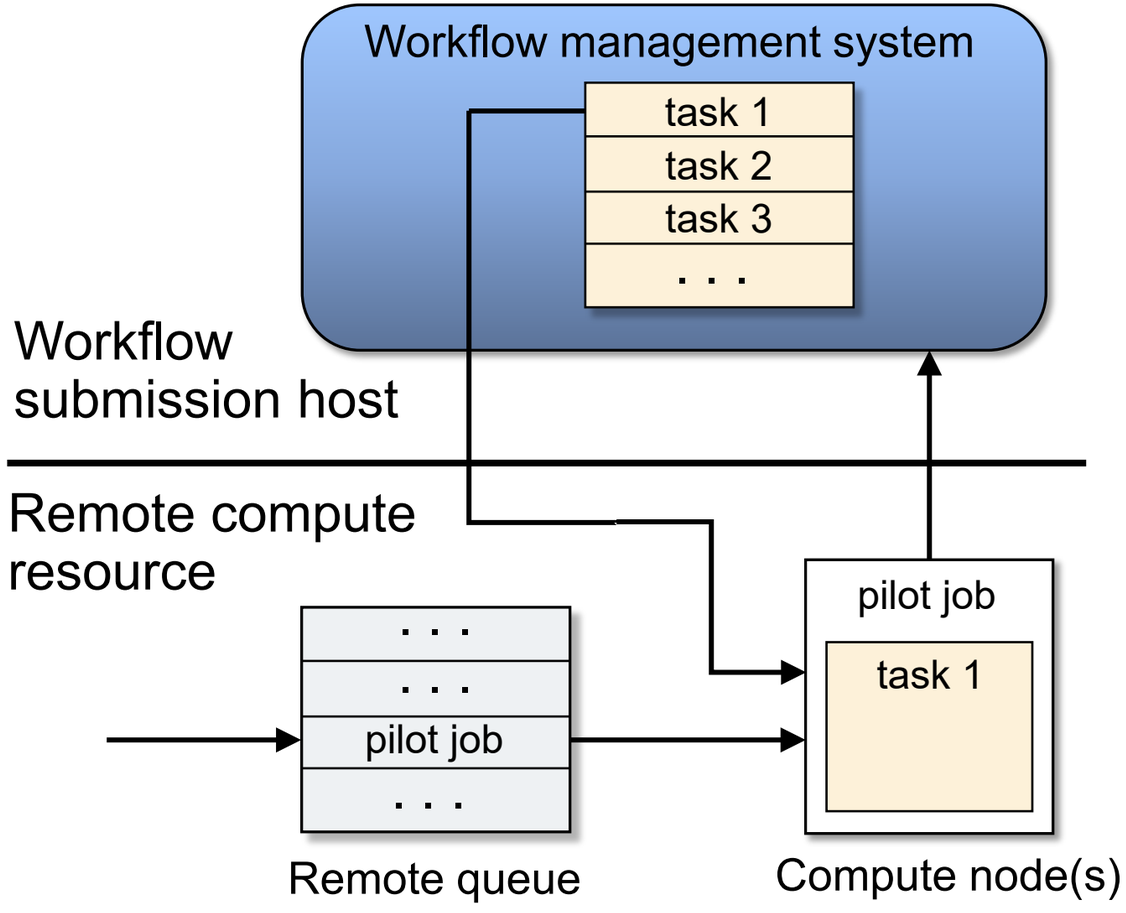
- Found occasional issues with files
  - Sometimes subtle bugs in the code
  - Sometimes filesystem hiccups
  - Sometimes problems in file transfer (less of an issue now)
- Added sanity checks to our workflow
  - Job to check number of files, NaNs, size of files, etc.
  - Calculate MD5 sums for files which will be archived
- In current study, encountered new exciting errors
  - Problem with OS flushing to disk?
  - Additional checks needed

Resolved Technical Challenges – Automation + 2FA

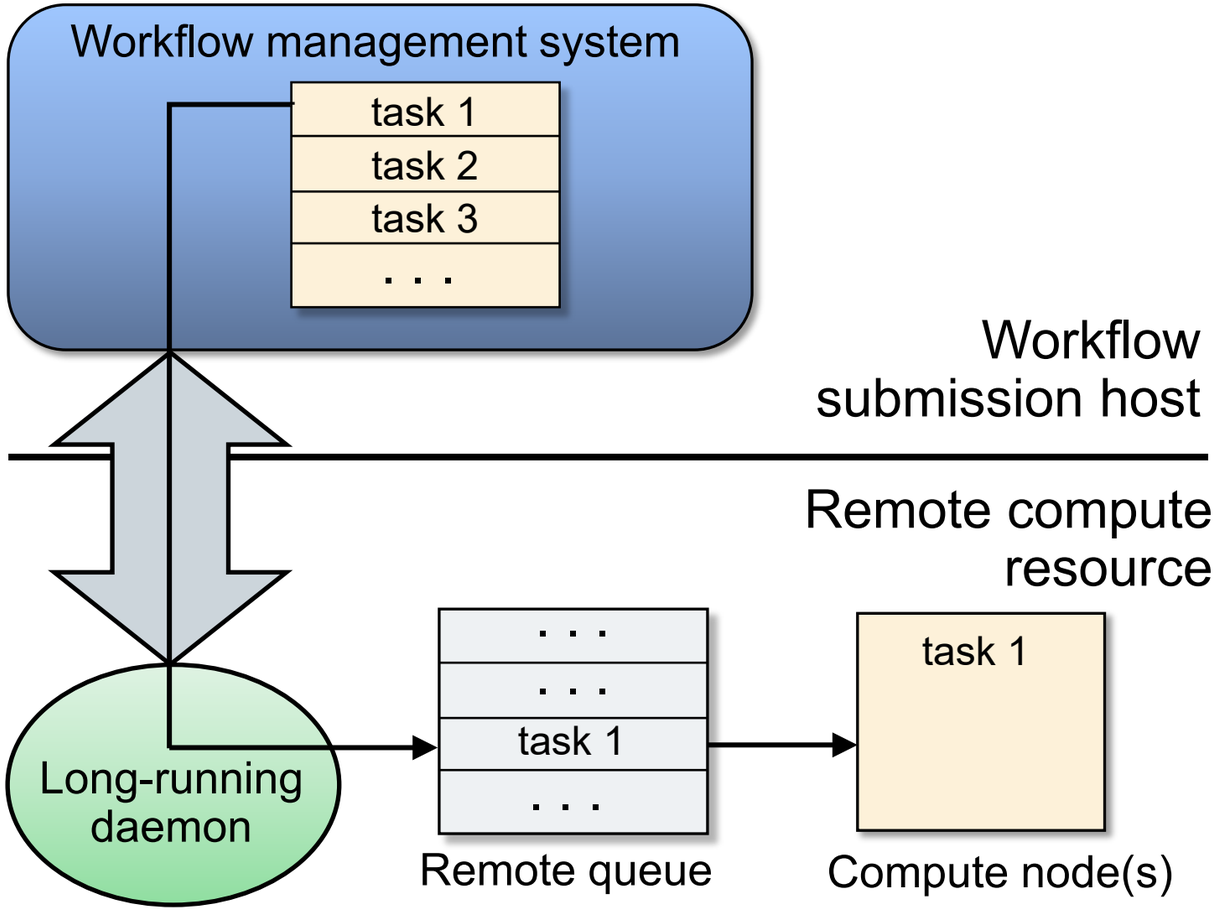
- Automated job submission is required
  - Most recent CyberShake study ran ~26 million computational tasks
- How do you submit jobs in an automated fashion with 2FA?
  - (Without making the system administrators very angry)
- Workflow tools provide two solutions:
  - Push-based (get work first, then find them nodes)
  - Pull-based (get nodes first, then give them jobs)
- We use both approaches in CyberShake

Resolved Technical Challenges – Automation + 2FA

- Pull-based: get resources first, then ‘pull’ work onto them



- Push-based: jobs sent when ready
  - Daemon required to set up connection

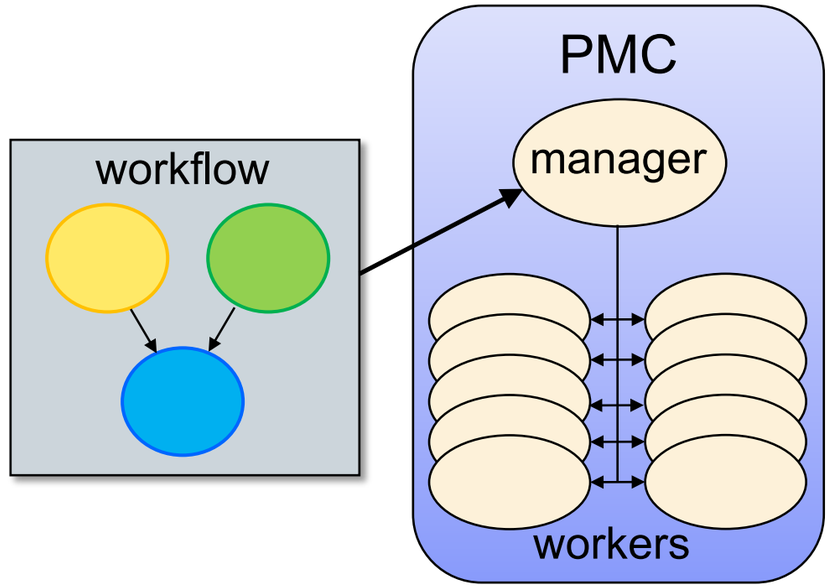


Resolved Technical Challenges – Job Throughput

- Large parallel tasks
  - Our target systems prefer large jobs
  - Using pull-based approach, request 1000 nodes
    - Run 10 100-node tasks on these nodes
  - Better throughput than 10 individual jobs
- Small serial tasks
  - Can't put them in the queue individually
  - Use Pegasus-MPI-Cluster, a workflow tool
  - Runs workflow tasks inside MPI job
  - Great for high throughput or self-contained workflow
  - Push-based

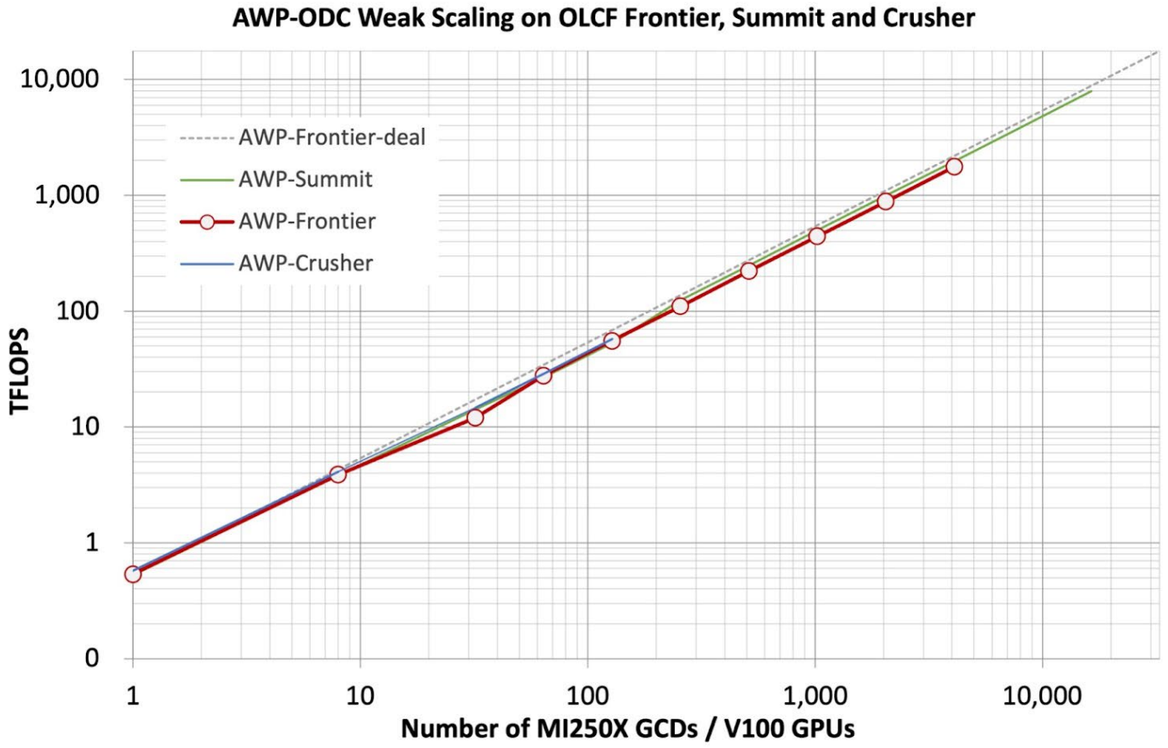
Bin	Node Range	Aging Boost
1	2765 – 4608	15 days
2	922 – 2674	5 days
3	92 – 921	0
4	46 – 91	0
5	1 – 45	0

OLCF *Summit* Scheduling Policy



What about the wave propagation code?

- Collaborators continue to enhance with new physics
  - Planning to migrate to faster discontinuous mesh version
  - Will also enable higher-frequency simulations
- Optimizing for new architectures
  - Currently being ported to HIP for AMD GPUs
  - Successfully tested last week on almost full-system *Frontier*
- Value of collaborations and leaning on expertise

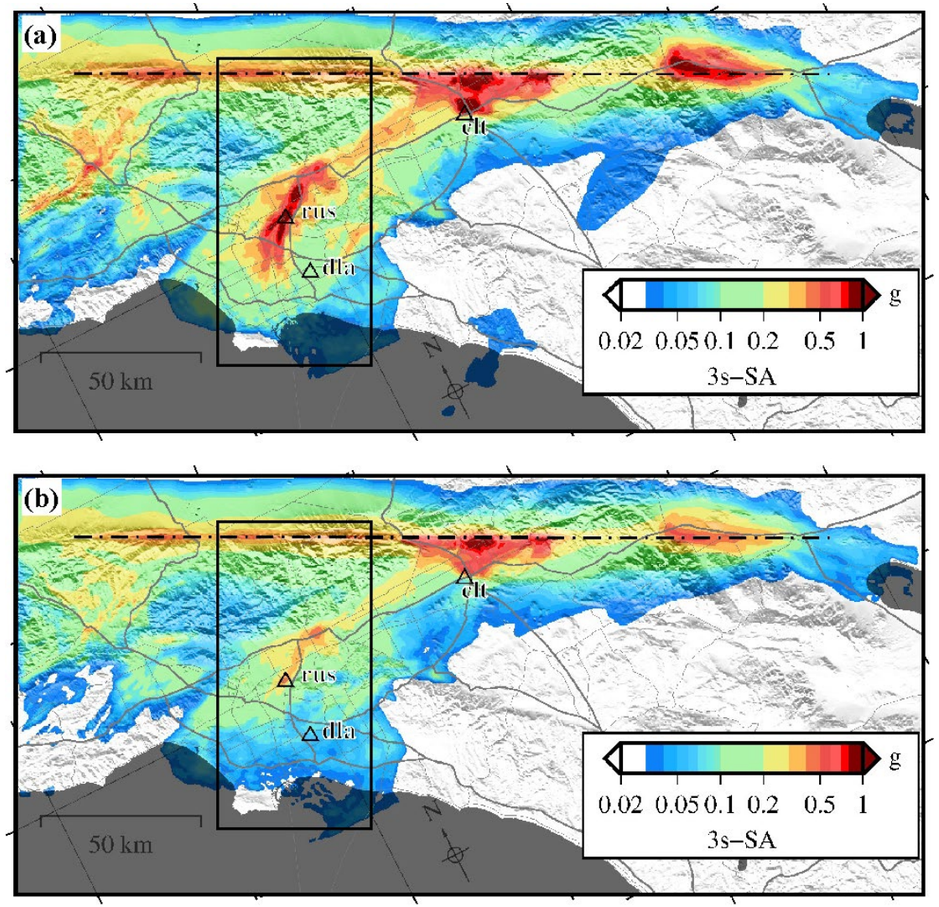


What are my current challenges?

- Improve access to output data
  - So people don't have to email me for the seismograms!
  - Data access tool to be released shortly
  - Seismograms delivered through Globus Collection
- Better database solution
  - We insert shaking metrics into database to support data products
  - Upwards of 12 billion rows (3 TB) for most recent study
  - MySQL performance is poor
  - Looking for alternative approaches
- Reduce storage footprint of seismograms
  - ~200 TB of seismograms from CyberShake various studies
  - Looking at lossy compression: what metrics are needed to evaluate?

What are my future challenges?

- Nonlinearity
  - Rocks don’t always exhibit linear response
  - Current reciprocal approach is linear
  - Will need to combine some nonlinear simulations with reciprocity
- Improve reproducibility
  - A few parameters are still hard-coded
  - Common configuration file to track parameters
- GPU versions of additional codes
  - Some systems are requiring GPU codes
  - GPU versions give us greater flexibility (and hopefully greater performance)



Closing Thoughts

- Your time and resources are limited
  - Evaluate what's interfering with getting science results
  - What will get you the best payoff?
  - May not be typical code optimization
- Incremental improvement is still improvement
- If you're in an earthquake:

Thanks!


USGS
science for a changing world


National
Science
Foundation


PG&E


U.S. DEPARTMENT OF
ENERGY


SCEC
AN NSF+USGS CENTER


ISI
Information Sciences Institute


Pegasus


OAK RIDGE
National Laboratory
LEADERSHIP
COMPUTING
FACILITY


USC
Advanced Research Computing
Enabling scientific breakthroughs at scale