Present and Future Computing Requirements

Imaging and Calibration of Mantle Structure at Global and Regional Scales Using Full-Waveform Seismic Tomography

Scott French

sfrench@seismo.berkeley.edu

Romanowicz Group

Berkeley Seismological Laboratory

UC Berkeley

NERSC BES Requirements for 2017

October 8-9, 2013

Gaithersburg, MD

University of California Berkeley Seismological Laboratory

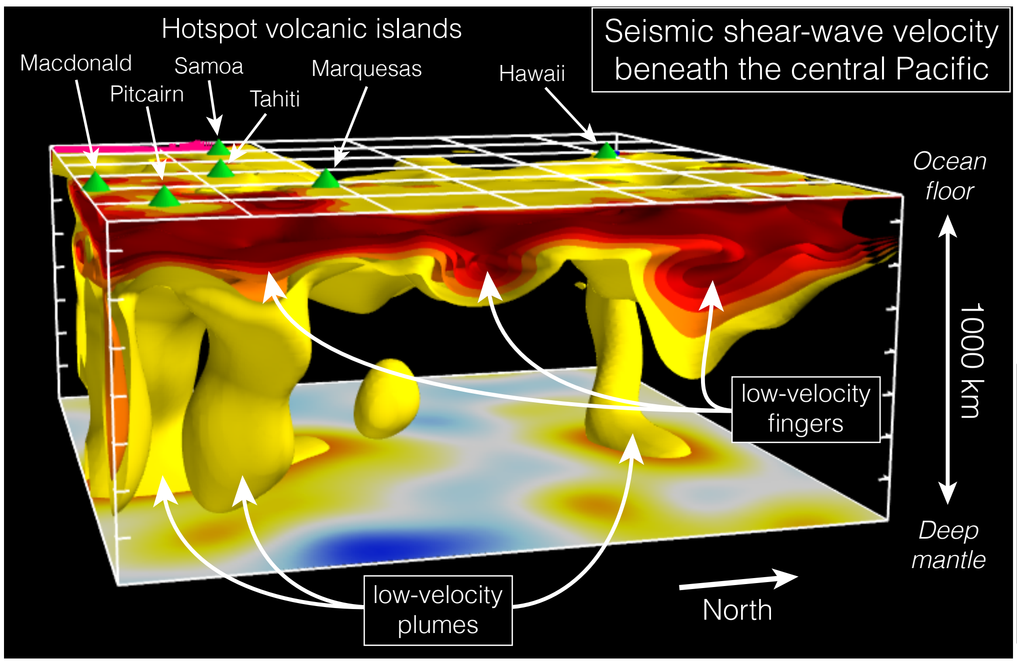
Project Description

Imaging and Calibration of Mantle Structure at Global and Regional Scales Using Full-Waveform Seismic Tomography

PI: Prof. Barbara Romanowicz

Berkeley Seismological Laboratory, UC Berkeley; *Institute de Physique du Globe de Paris*, Paris, France; *Collège de France*, Paris, France

Scientific Objectives: (1) improve our understanding of Earth's interior structure by (2) developing and applying new methods for imaging from global to regional scales



Current Focus: Imaging based on numerical simulations of seismic wave propagation; developing techniques for *simulation speedup* and rapid *model convergence*.

2017 Focus: Higher-resolution imaging, using *higher-frequency* waveform data, combined with the new methods being developed and validated at present.

Above: 3D rendering of seismic shear-wave velocity structure beneath the Central Pacific in the SEMum2 seismic model (French et al. 2013, *Science*).

Computational Strategies

Full-waveform Seismic Inversion

Minimize: $\chi(\mathbf{m}) = \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2 + \|\mathbf{m} - \mathbf{m}_p\|_{\mathbf{C}_\mathbf{m}^{-1}}$

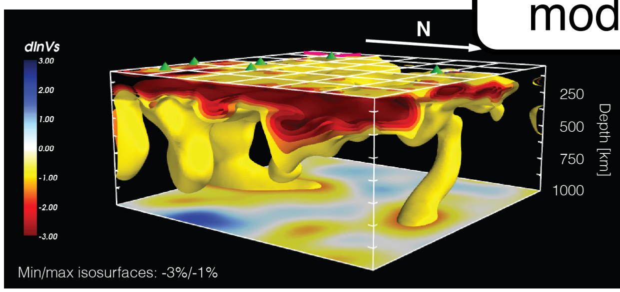
Seismic model

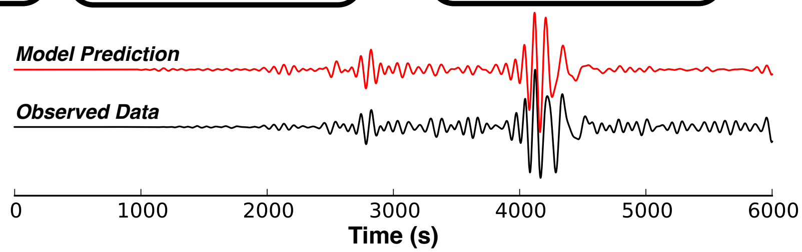
Observed waveforms

Predicted waveforms

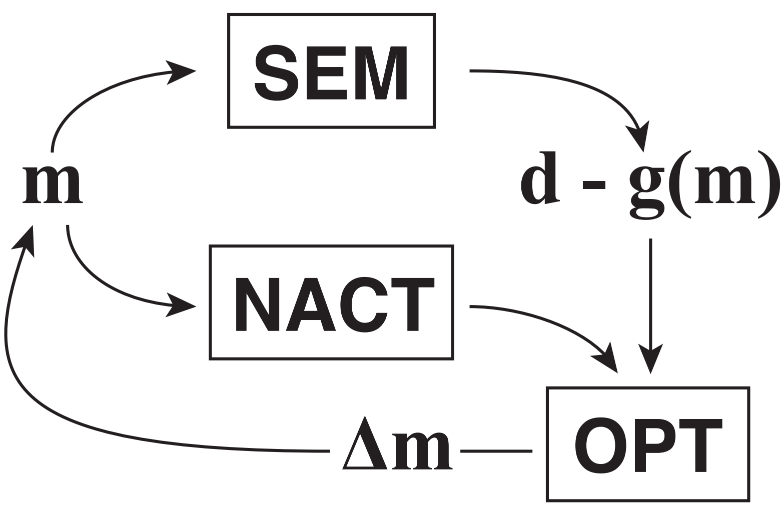
Prior Information

$\mathcal{N}(\mathbf{m}_p, \mathbf{C}_\mathbf{m})$





Solved iteratively



Computational Strategies

Full-waveform Seismic Inversion

Minimize: $\chi(\mathbf{m}) = \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2 + \|\mathbf{m} - \mathbf{m}_p\|_{\mathbf{C}_\mathbf{m}^{-1}}$

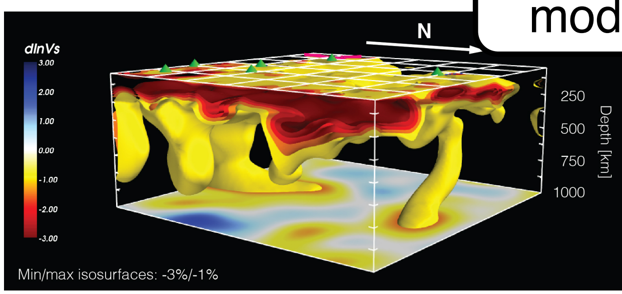
Seismic model

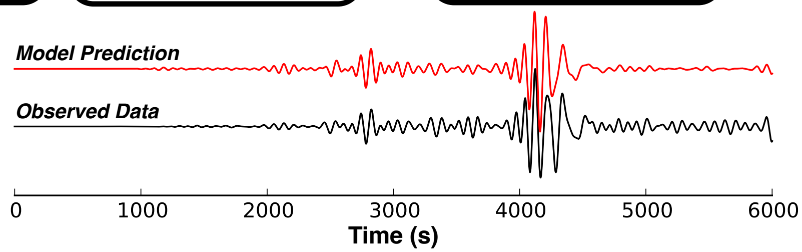
Observed waveforms

Predicted waveforms

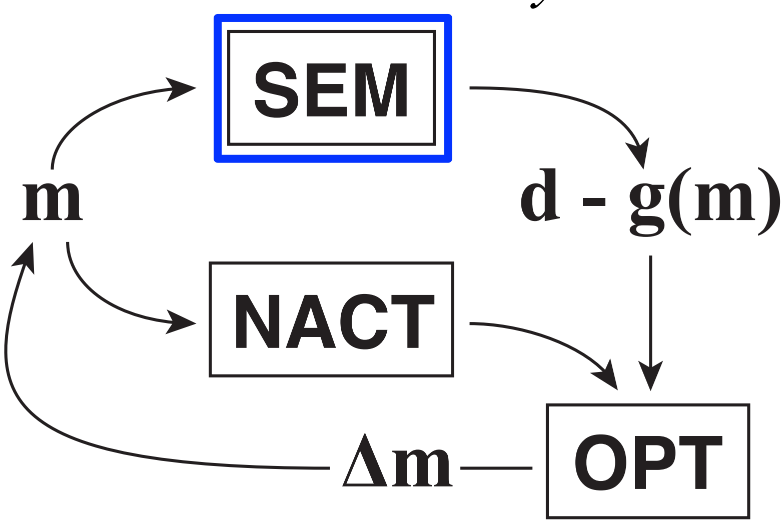
Prior Information

$\mathcal{N}(\mathbf{m}_p, \mathbf{C}_\mathbf{m})$





Solved iteratively



Simulation Phase

Spectral finite-element method (SEM)

Global and regional-scale simulation of seismic wave propagation in complex 3D seismic models

- High-order, matrix-free formulation; 50-120M degrees of freedom
- Fortran 90 / MPI
- > **80%** of allocation

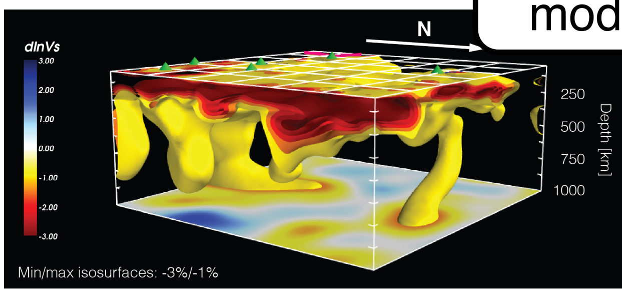
Computational Strategies

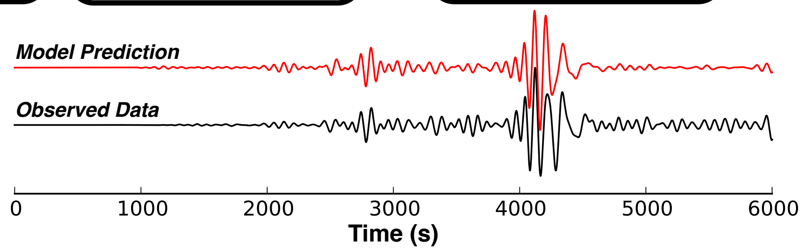
Full-waveform Seismic Inversion

Minimize: $$\chi(\mathbf{m}) = \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2 + \|\mathbf{m} - \mathbf{m}_p\|_{\mathbf{C}_\mathbf{m}^{-1}}$$

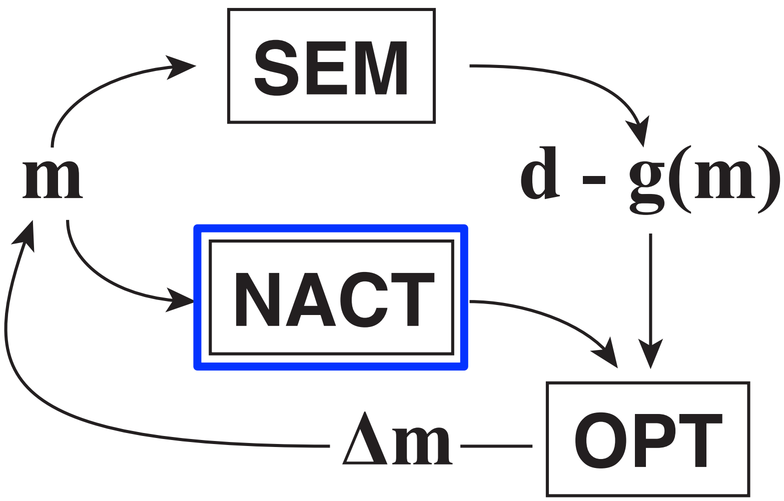
Seismic model — Observed waveforms — Predicted waveforms — Prior Information

$\mathcal{N}(\mathbf{m}_p, \mathbf{C}_\mathbf{m})$





Solved iteratively



Assimilation Phase I

Normal-mode coupling theory (NACT)

Physics-based estimation of gradient and Hessian for iterative seismic-model optimization

- Non-linear asymptotic coupling theory (NACT: Li & Romanowicz, 1995)
- C / MPI + OpenMP
- **< 10%** of allocation

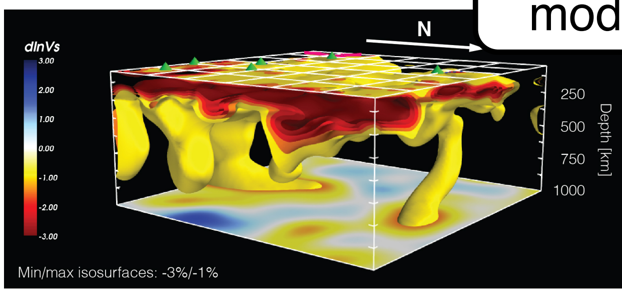
Computational Strategies

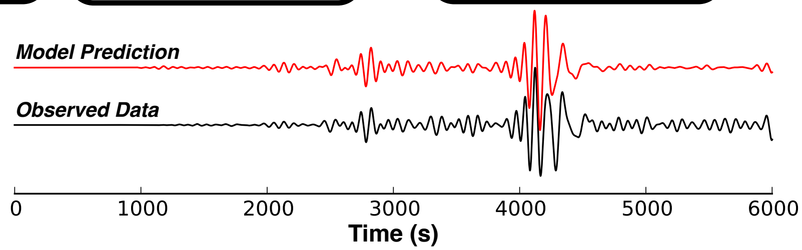
Full-waveform Seismic Inversion

Minimize: $\chi(\mathbf{m}) = \|\mathbf{d} - \mathbf{g}(\mathbf{m})\|_2^2 + \|\mathbf{m} - \mathbf{m}_p\|_{\mathbf{C}_\mathbf{m}^{-1}}$

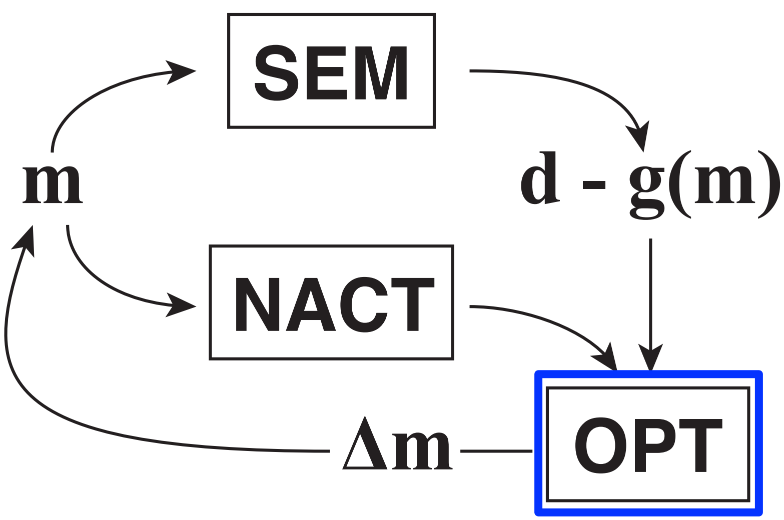
Seismic model | Observed waveforms | Predicted waveforms | Prior Information

$\mathcal{N}(\mathbf{m}_p, \mathbf{C}_\mathbf{m})$





Solved iteratively



Assimilation Phase II

Model optimization (quasi-Newton)

Minimize difference between seismic data and SEM predictions (assimilate simulation output)

- Generalized least-squares; 100-200K parameter dense linear system
- C / MPI + ScaLAPACK
- < **10%** of allocation

Current HPC Usage I

Facilities

- Current production platform: NERSC Hopper (Edison soon)
- NERSC is currently our only compute facility (NISE award)
  - **2012**: 2M compute hours
  - **2013**: 3M compute hours

Resources

- **Compute**
  - **SEM**: 200-300 simulations / model iteration
    - 150-300 cores / simulation – strong scaling: PE vs. wall time (5-8 hrs)
  - **Assimilation**: 20-30 runs / model iteration
    - 500+ cores / run – wall time (1-5 hrs)
- **I/O**
  - **SEM** (file-per-process)
    - **R:** 14-30GB mesh files **W:** 3-5GB / checkpoint (2-3x)
  - **Assimilation** (single-process aggregation *and* MPI-IO)
    - **R:** 10-20GB / run **W:** 100GB / run (2TB+ total)
- **Memory**
  - Upper bound: 1-2GB / core for both (closer to 1GB)

Current HPC Usage II

Scheduling / Workflow Considerations

- SEM simulations aggregated into 2-3K core production runs
- Workflow is ***episodic*** in nature
    - Typically, 3+ inversion *iterations* per year
    - Pauses for off-line analysis required (convergence? new data?)

Software / Library Requirements

- Minimal: MPI, ScaLAPACK, Optimized BLAS

Additional Services / Infrastructure

- Analyses offsite (simulation output, seismic model)
- Very little data transfer (< 1TB / AY)

Storage Resources

- Max scratch utilization: 3-4TB (assimilation phase, once per iteration)
- HPSS used only for heavily post-processed simulation output
    - At present: ~100GB
    - Expected to double by end of 2013

Predicted 2017 HPC Requirements I

Allocation Request

- Estimate: *at least* 25M conventional (Hopper-equivalent) compute hours
- Driving Factors
  - **Higher frequency** SEM simulations ($O(f^4)$)
  - **SEM+adjoint** 3 x number of simulations (current work: reduce to 2 x)

Resources

- **Compute**
  - **SEM+adjoint**: 2 x 300+ simulations / model iteration
    - 300-500 cores / simulation
  - **Assimilation**: 20-30 runs / model iteration
    - 500-1000 cores / run
- **I/O**
  - **SEM+adjoint** (file-per-process *and* MPI-IO)
    - **R/W:** 500GB time-history of checkpoints
    - **R:** 30-50GB mesh files **W:** 10-20GB / checkpoint (2-3x)
  - **Assimilation** (single-process aggregation *and* MPI-IO)
    - **R:** 100GB / run **W:** 0.5-1TB / run (10-20TB+ total)
- **Memory**
  - Still 1-2GB / core; Large shared-memory nodes (100GB+)

Predicted 2017 HPC Requirements II

Scheduling / Workflow Considerations
- Workflow still *episodic* (pauses for off-line analyses)
- Anticipate more *iterations* per year due to higher throughput (10+)

Software / Library Requirements
- Additions – Compiler support / libraries for:
  - Heterogeneous architectures (next slide)
  - PGAS languages
    - Considering UPC for next-generation assimilation codes

Additional Services / Infrastructure
- Analyses will remain offsite (simulation output, seismic model)
- Still fairly little data to transfer (1TB+ / AY)

Storage Resources
- Max scratch utilization: 20TB (assimilation phase, once per iteration)
- Typical: 10TB+ with 20 SEM+adjoint simulations in progress
- HPSS still used only for post-processed simulation output
  - Anticipate 0.5TB+ archived output by 2017

New Architectures

Current Status

- Some success in seismic-modeling community on porting high-order matrix-free finite element computations specifically to GPUs (Target: ORNL Titan)
- Efforts currently in general planning stage (isolating kernel computations)
- Some design choices in current code will help (element coloring / assembly)

By 2017 …

- Functioning port of our SEM code that can use GPU/MIC resources
    - Possibly merge with community SEM supporting GPU?
- But this depends on knowing *which technology* to target …

We need guidance with …

- What architectures should we have in mind?
- What programming models will be supported at the compiler level?
    - Directive-based? (OpenACC, OpenMP?)
    - Language extension / library? (CUDA Fortran, Cilk?)
- What libraries for common tasks will be available with GPU/MIC support?

Summary

Impact of Improved NERSC Resources

Answer fundamental questions about the dynamics of Earth's interior, while developing tools for seismic imaging that can be reused at a range of scales.

Critical Needs and Recommendations

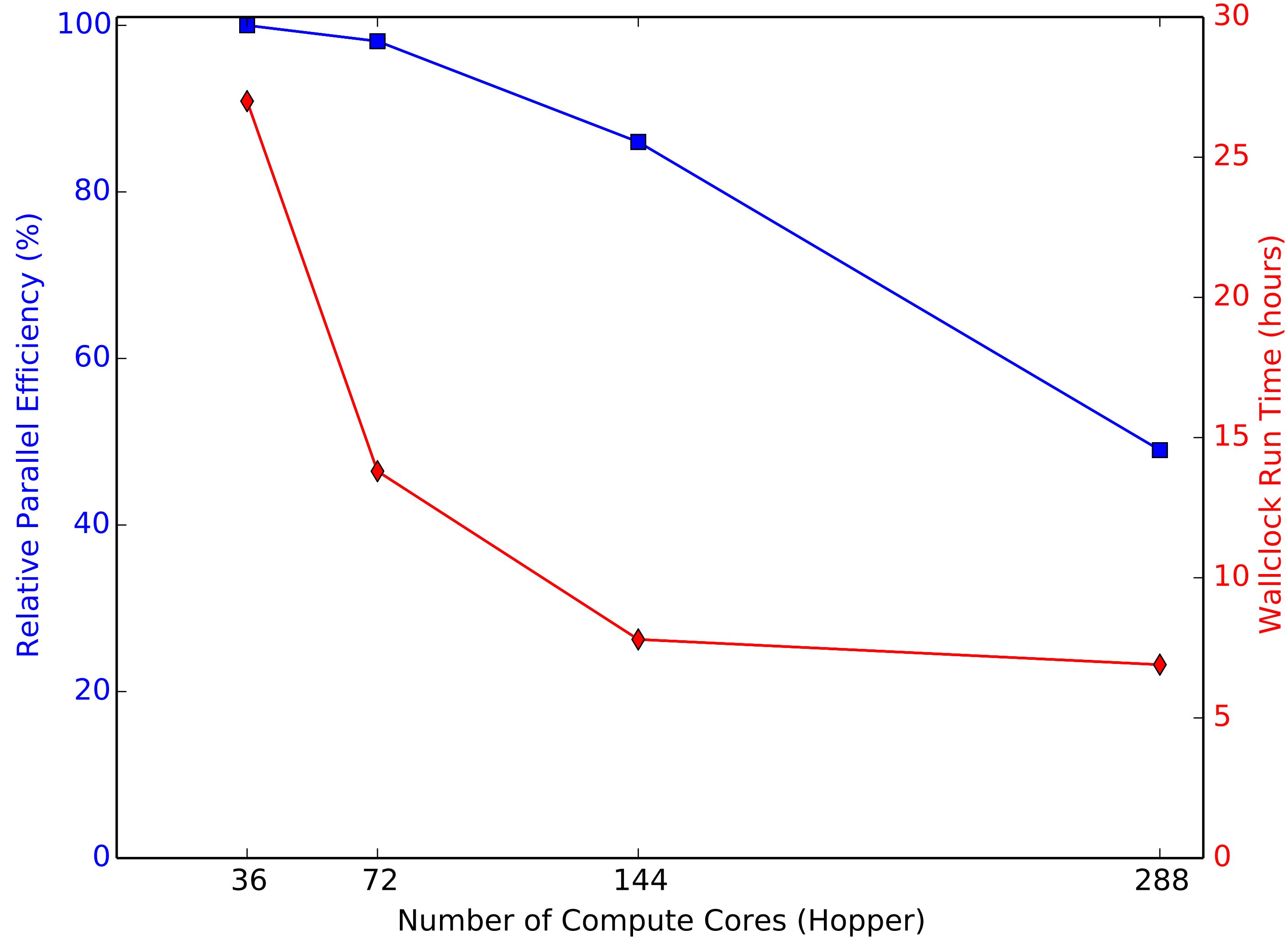
- **Resources**
  - High-performance I/O subsystem supporting scratch FS for SEM+adjoint
  - More shared memory per node (100GB+) for assimilation phase
- **Guidance and Services**
  - Heterogeneous architectures (GPU? MIC? Programming model?)
    - Early evaluation and assessment? How much lead time?
  - I/O performance and tuning
    - Best practices for new system (NERSC has been great on this)
- **Scheduling and Reliability**
  - Workflow is episodic; Contention w/ allocation reduction schedule
  - Use case: 1000s of semi-independent simulations
    - Management requires prediction/reasoning about wall-clock times
    - Non-determinism can be difficult: I/O? Interconnect? Node health?
      - Tools for monitoring, assessment, diagnosis?

Thank you!

Questions?

Extra Slides

Example: Efficiency tuning (2013)



Above: Tradeoff between (relative) parallel efficiency, $E_P = T(P_0)\, P_0 \,/\, T(P)\, P$, and wall-clock time $T(P)$ at fixed problem size (strong scaling) for a range of core counts P.