Leveraging Data Analysis to Improve Simulations

Plus Additional Musings on Exascale Simulation

Ryan G. McClarren

Texas A&M

The current high-fidelity simulation paradigm

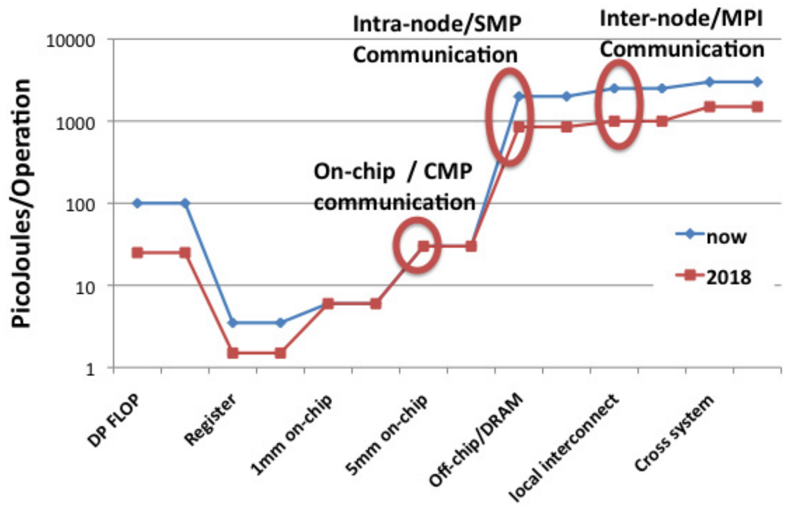
- Think about what simulation(s) to run, count on being able to investigate results *after* the simulations.
- Data is large but manageable.
- Select features that are important for analysis after the collection of data.
- For solution verification, uncertainty quantification, and other situations where an ensemble of calculations is needed
  - ⇒ Post-processing and feature extraction from stored results.
- Focus on what the particulars of the system/experiment/ phenomenon are
  - ⇒ Less so on what to do with the results afterwards

What will Exascale Data Analysis Look Like

What will Exascale Data Analysis Look Like

What will Exascale Data Analysis Look Like

- I don't know for sure.
- The data generated will be large and generated with velocity.
- It is very likely that it will be difficult, if not impossible, to
  - ⇒ Transmit the data
  - ⇒ Compute complex functions, transformations to the data
  - ⇒ Store the data
- Part of this is due to power



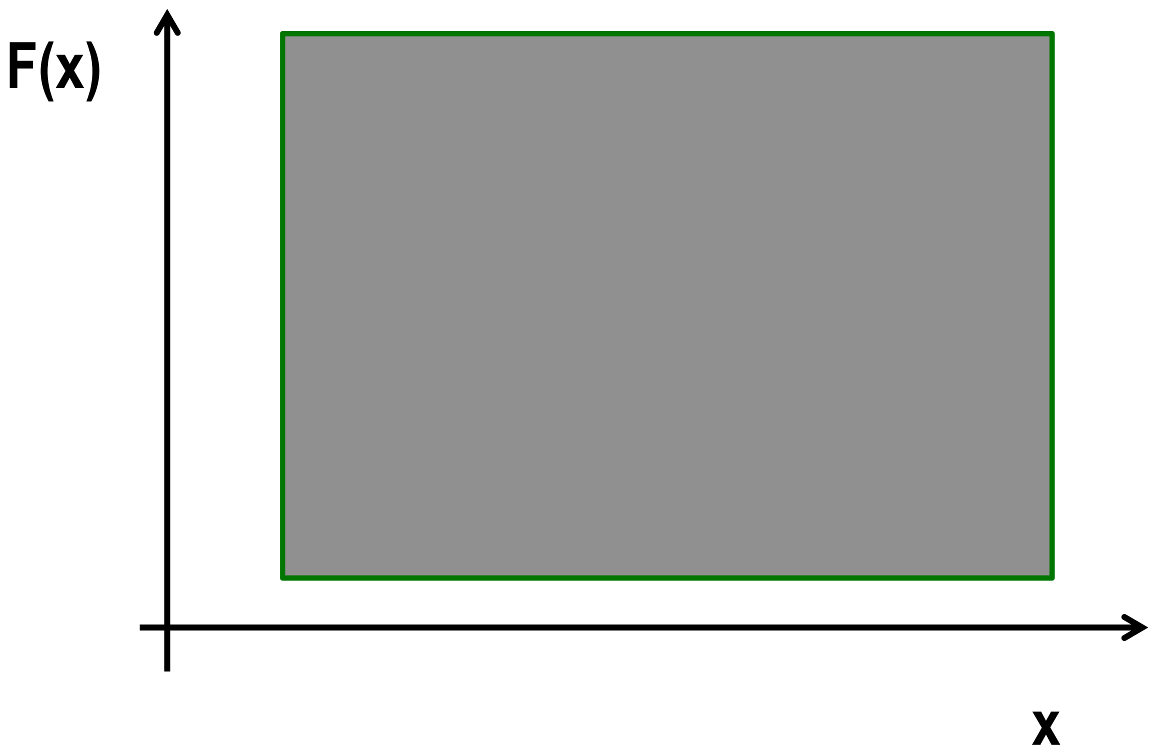
Shalf, J., Dosanjh, S., & Morrison, J. (2011). Exascale Computing Technology Challenges. *Lecture Notes in Computer Science* (Vol. 6449, pp. 1–25)

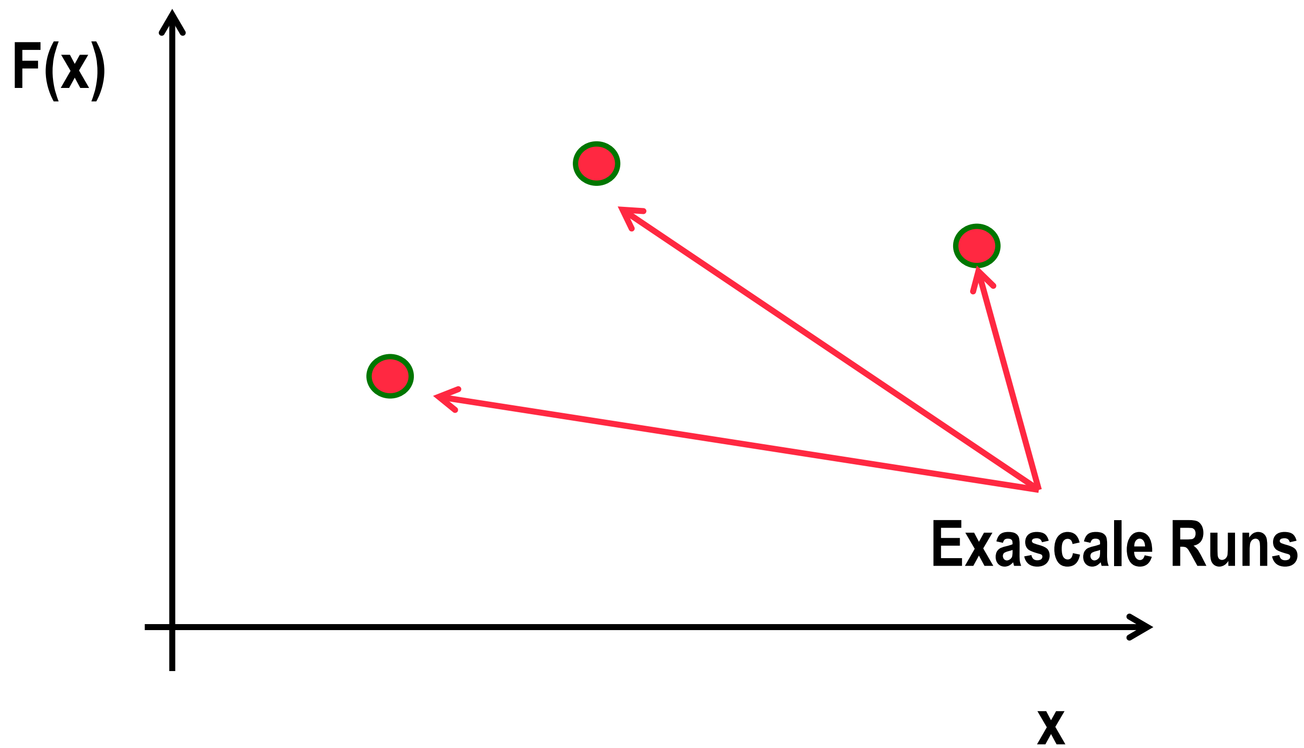
What types of data analysis infrastructure will we need?

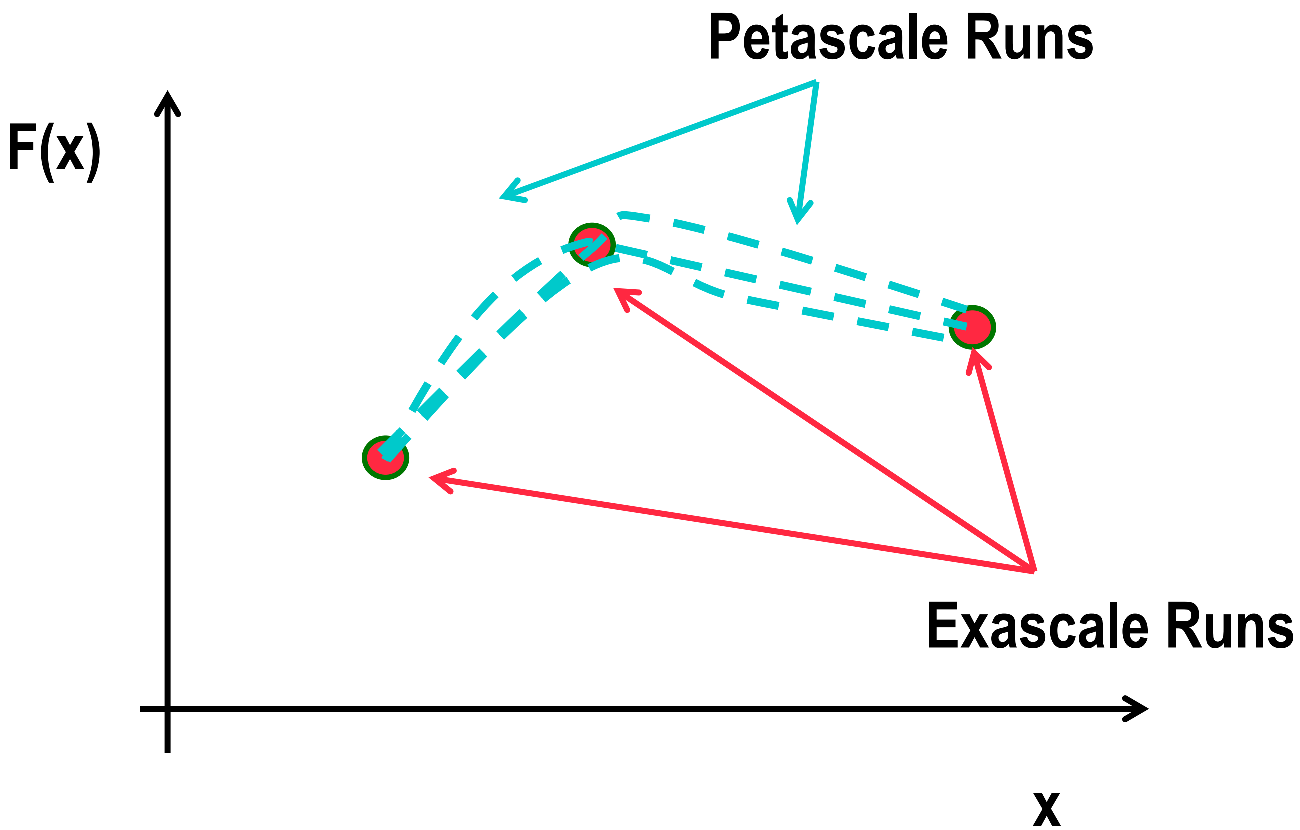
- The data analysis will need to be moved closer to the computation.
- Code-user will need to decide beforehand:
  - ⇒ What metrics to calculate
  - ⇒ What analysis do we need to do
  - ⇒ What features do we require to do the analysis
  - ⇒ What visualizations do we want (and what is the resolution of those visualizations)
  - ⇒ What is the coarsest granularity I need the data?
- Will have to know what is interesting before you do the simulation.
- We will also have to consider predictive models and analysis modalities that do not require having all of the data at once.
  - ⇒ Analysis is not a post-process anymore
- Could allow for interesting benchmarking of models
  - ⇒ Build and Test a surrogate model at every step of the simulation

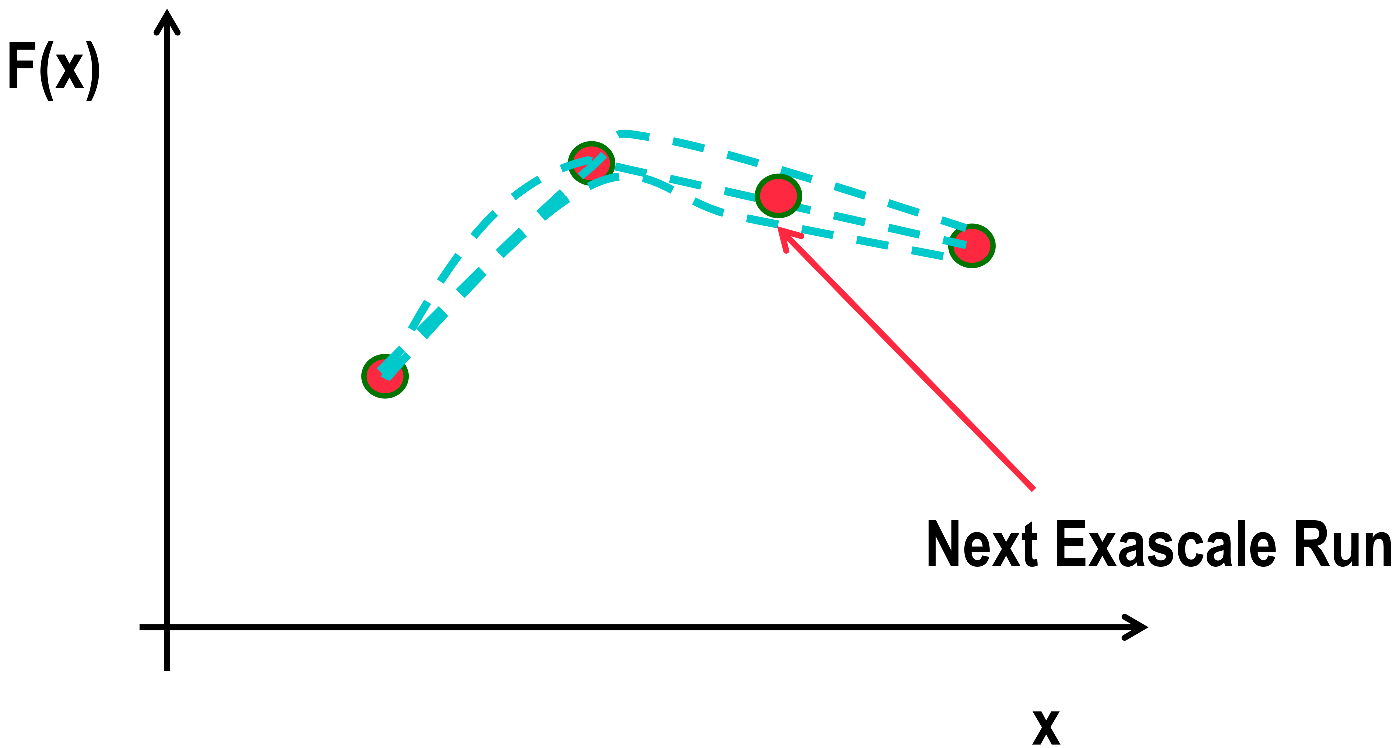
There is Potential Multi-fidelity Models

- True exascale simulations will still be rare. This coupled with the lack of post-simulation data exploration will keep petascale simulations valuable.
- Sometimes we can formulate the lower fidelity models so that when informed by the high fidelity model, it gives the same result.
  - $\Rightarrow F_{\text{highres}}(x) = F_{\text{lowres}}(x, \theta)$
  - $\Rightarrow \theta(F_{\text{highres}})$
- Common examples of this type of low fidelity model (oftentimes called closures)
  - ⇒ Variable Eddington factor derived from transport solution particle transport
  - ⇒ Equations of state informed by a kinetic model
- Even if the model is only approximate, but informed by the high fidelity simulation, it can improve the workhorse calculations.
- This can be further enhanced by statistical models that help bridge the fidelity gap.


F(x)
x


F(x)
Exascale Runs
x

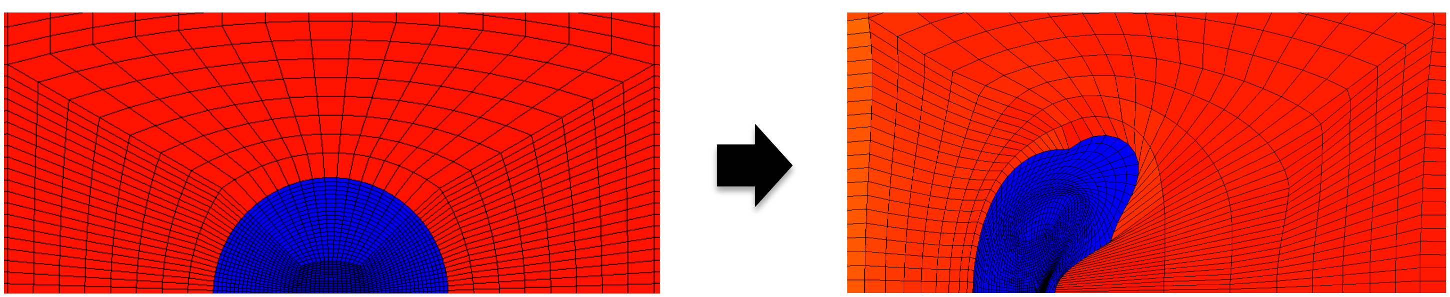

Petascale Runs
F(x)
Exascale Runs
x


F(x)
Next Exascale Run
x

STEERING (ENABLING) SIMULATION WITH PREDICTIVE ANALYTICS

Arbitrary Lagrangian-Eulerian (ALE) Hydrodynamics

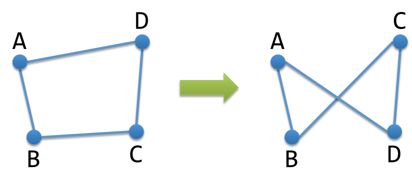
- In simulating hydrodynamics, especially where multiple materials are present, the arbitrary Lagrangian-Eulerian (ALE) method is a widely used method.
- The method combines the two approaches
  - $\Rightarrow$ Allows the mesh to move with the flow (Lagrangian)
    - Preserves numerical interfaces
  - $\Rightarrow$ Keep the mesh fixed (Eulerian)
    - Numerical diffusion in solution
- Combine the two by evolving solution with a moving mesh and performing an Eulerian relaxation step



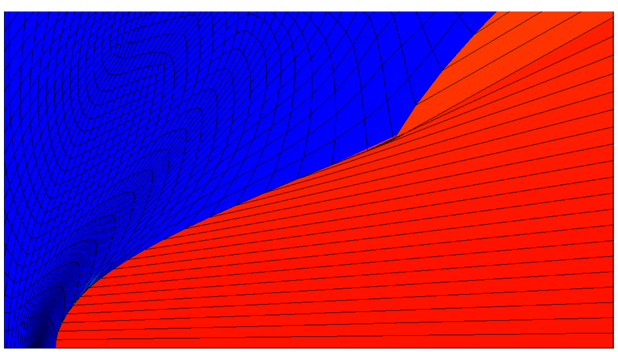
Figures from LLNL-PRES-660220

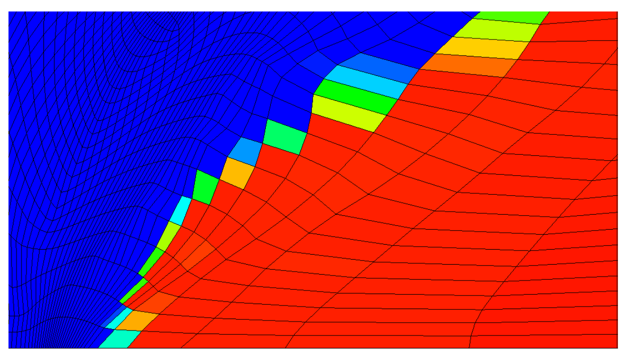
The Problem with ALE

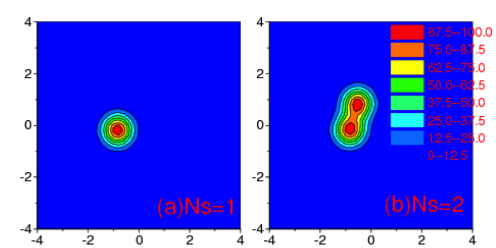
- Ideally, one would evolve the simulation without any relaxation to preserve material interfaces.
- This can lead to "mesh tangling" that crashes the simulation.



- Therefore, the relaxation is used to prevent this sort of tangling.
- Over-relaxation can lead to numerical errors and loss of accuracy.







Figures from LLNL-PRES-660220

Human Relaxers

- The way the amount of relaxation is typically chosen is by hand, by experts who have run many simulations.
- Often this is done by running the simulation until it crashes then
  - ⇒ The expert goes in and sets relaxation parameters based on
    - Mesh metrics (e.g., aspect ratio)
    - Physical parameters (e.g., pressure, temperature)
    - Gut instinct / past experience
- This works, but is not ideal.
  - ⇒ Training someone to do this is a long process
  - ⇒ Slow
- Hard to output many simulations for uncertainty quantification runsets.
- What is the uncertainty in different human relaxers?

The robots are coming for our relaxers

- What we would like to do is map the knowledge/process of the best human relaxers to a statistical model.
- Build a model to predict whether a zone will need to be relaxed based on the state of the simulation.
- The training set would be simulations of a class of problems
  - ⟹ Dependent variable is whether a zone will cause the simulation to fail.
- Ideal output would give relaxation automatically that
  - ⟹ Minimizes human interaction with simulation (fewer crashes)
  - ⟹ Minimizes the error introduced by relaxation
- Current approach uses random forests to predict the needed relaxation.
- Introduces a two new problems: feature selection and understanding of relaxation error.

Feature Selection

- Common problem: encode human decisions into a statistical model
- When a human decides relaxation parameters, several considerations are possible
  - ⇒ Time history of zone
  - ⇒ Relation to physical features in the problem (e.g., distance from shock)
- This could be problematic for large-scale simulations.
  - ⇒ At a given time step in the simulation, it can be expensive to access
    - Data from previous time steps (data can't fit in memory and could be on disk)
    - Non-local data (ask for data on another processor, across network)
- Need a balance between accessible data and useful data.
  - ⇒ How far can we get with predicting the needed relaxation?
  - ⇒ Sequential enhancement of available data as needed by
    - Accuracy and availability considerations

In situ or On-the-Fly Relaxers

- To this point the automatic relaxers are built by storing the results of many simulations, and
- Post-processing the results to extract features and then fit the models.
- In an exascale reality, we can't afford to store all of that data, load it in, build a model, …
- We want a system that learns as it goes:
  - ⇒ While a suite of simulations runs the model evolves
    - Train the model on any failures
    - Automatically roll back solution to before the failure, relax, and continue running.
  - ⇒ The initial model will be based on results from previous simulations.
- The idea is to make the creation of the statistical model a one step process, rather than separating data production and analysis.

Which is the better relaxer?

- Above I mentioned that we want our relaxation to be set to minimize errors introduced by the relaxer.
- Traditionally, the measure of relaxer efficacy is whether the code ran to completion.
- Intuitively we might look at the amount of mixing in zones, because this is introduced/enhanced by the relaxer.
- Two different approaches to this:
  - ⟹ Measure of mass fraction differences in zone (alpha)
  - ⟹ Variability of material speed of sound in zone (beta)

$$\alpha(\text{zone}) = \prod_{i}^{\text{materials in zone}} (1 - v_i)$$

where v_i is volume fraction

$$\beta = \left(\sum_{\text{all zones}} \Delta^2_{\text{sound}} \right)^{\frac{1}{2}}$$

Mass weighted	Unweighted	Volume Weighted
$\int_{\text{all zones}} \alpha \, dm$	$\sum_{\text{all zones}} \alpha$	$\int_{\text{all zones}} \alpha \, dv$

Is this a good measure of simulation accuracy?

- Test problem of a ICF capsule implosion.
- Have a human-tuned relaxer as the baseline.
- Loosen (increase the relaxation) or tighten the relaxer and look for changes in the
  - ⇒ Typical quantities of interest
    - ablation front or time of maximum density (bang time)
  - ⇒ Our mixed-related quantities
- Want to show that the mix quantities are correlated with standard QoIs

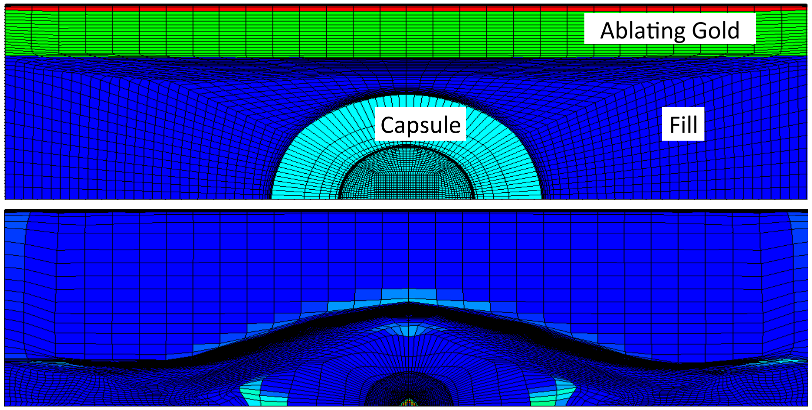


Figure 1: Ablating hohlraum and Shock at 11.5 ns (slightly before bang time)

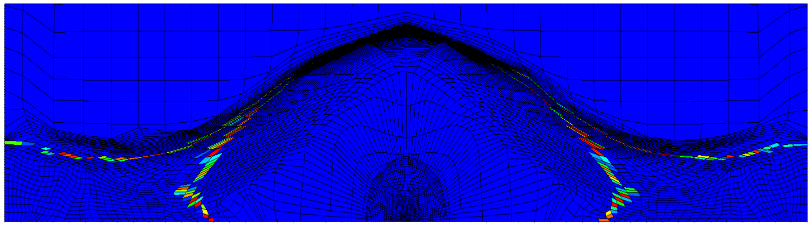
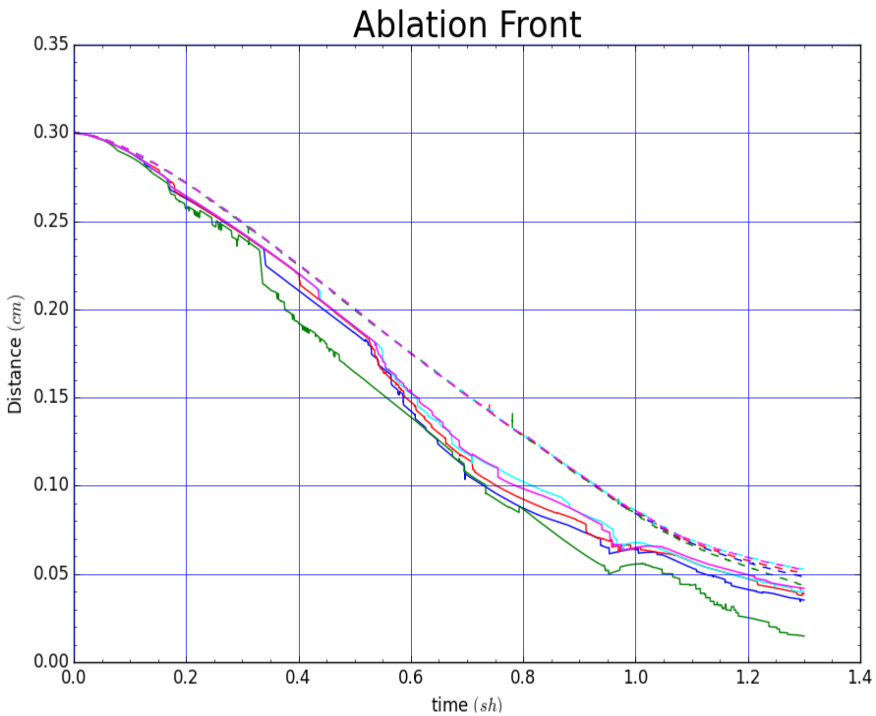
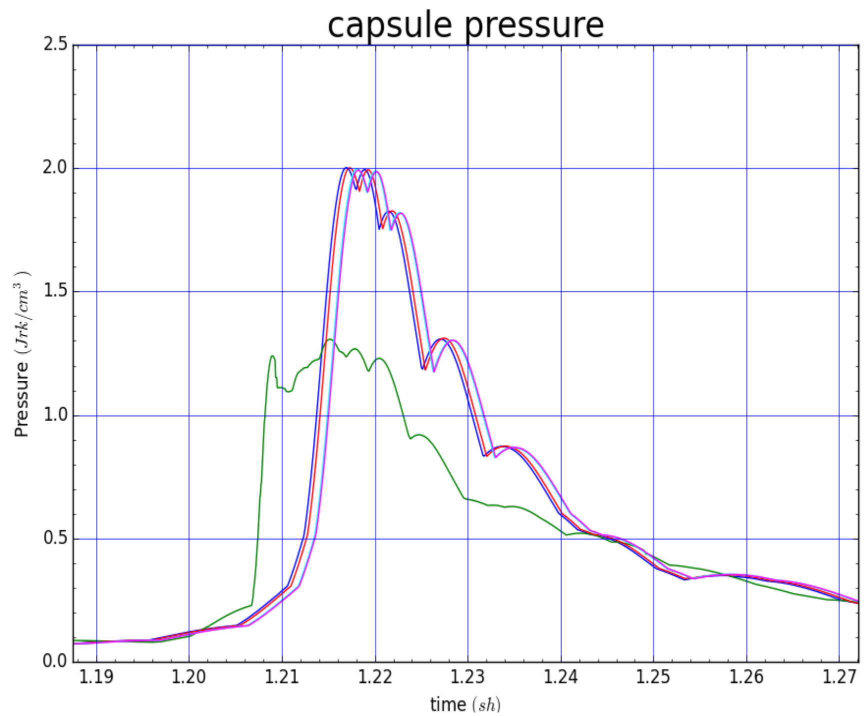


Figure 2: Mix induced numerical error at 12.5 ns (approximately bang time)

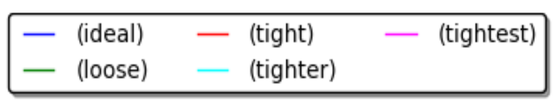
Yes, the relaxer matters



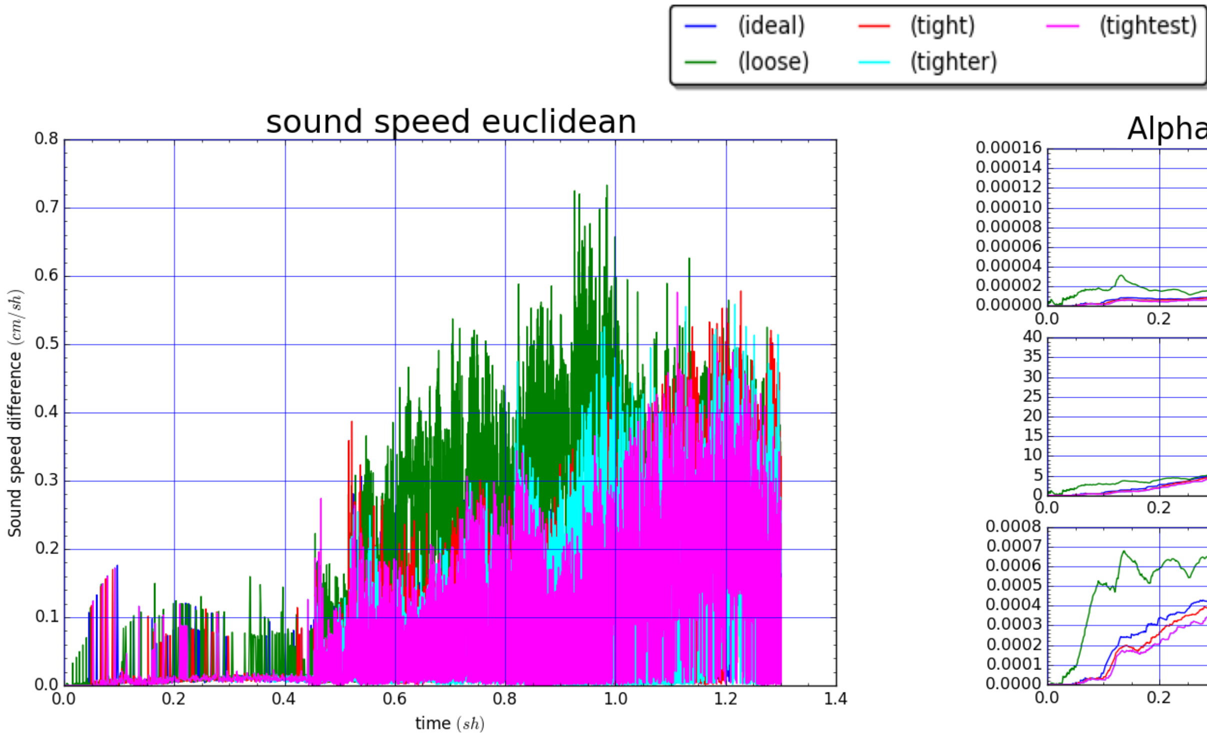


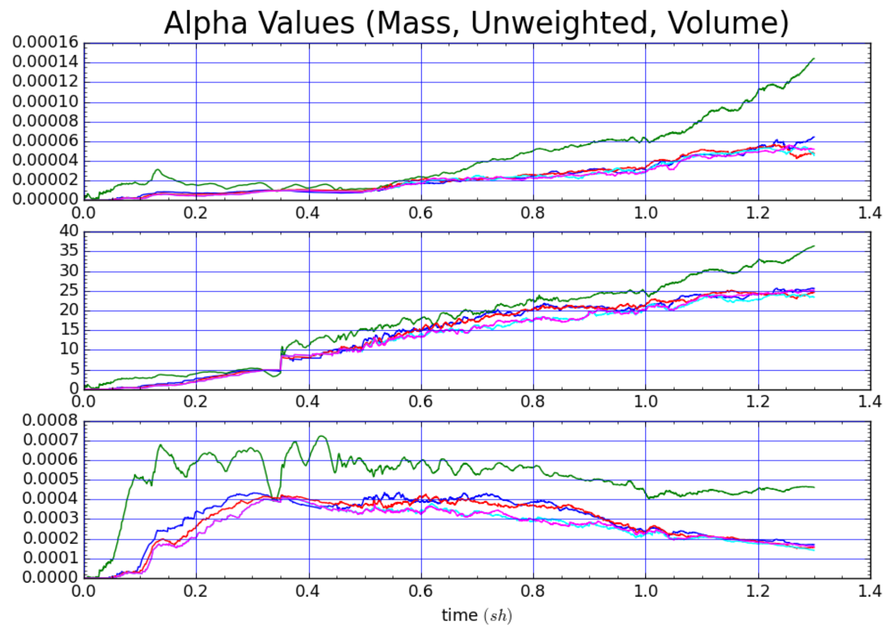
The lowest zone distance (solid) is significantly more affected then the 99% line (dashed). Increasing relaxation causes a divergence of the two solutions.

ALE relaxation shifts the pressure peak left. As settings are tightened the bang time converges.



The mix variables are a good metric for solution quality





Euclidian difference of the sound speed is increased by additional ALE. The tighter settings converge to a common value.

Mixing parameters diverge as the ranges are increased. The tighter ranges show a convergence to an ideal amount of mixing.

Validating the Mixing Metrics and the Future

- We still need to show that our solutions are correct and
- That the "ideal" relaxer is the right target to shoot for.
- Simulations are ongoing with a converged Eulerian code to show that minimizing mix metrics gives us the best solution.
- Once we are confident in the mixing metrics, we can further explore the creation of statistical models and enhancing the phase space.
- Run a variety of problems and see how transferable the results are.
- …

Coda: We live in interesting times

- Exascale will push us to think about simulation and how we consume/analyze simulations differently.
- The changes will make the way we think about multiple fidelity models.
- To get the benefits of exascale, we need to make sure that humans aren't the bottleneck.

Acknowledgments

- The "A Data Analytics Approach to Improving Simulation Workflow" LDRD team.
  - ⇒ LLNL Ming Jiang (PI), Brian Gallagher, Daniel Laney, Joshua Kallman
  - ⇒ Yuriy Ayzman, TAMU graduate student