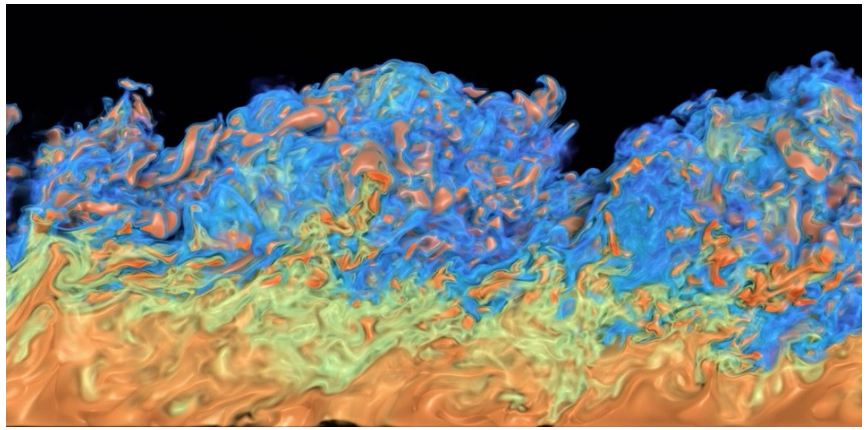
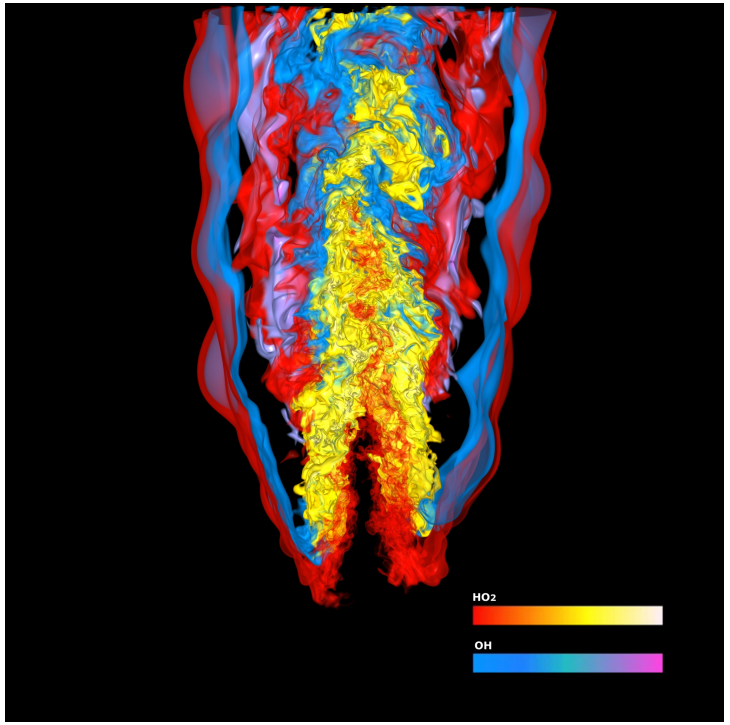
S3D: A Combustion Use Case, DNS with a Surrogate Time-Dependent Reduced Order Model



Temperature in a NH3/H2/N2-Air Premixed Flame in a Shear Layer at 10 atm

Jacqueline Chen
Senior Scientist
Combustion Research Facility
Sandia National Laboratories
jhchen@sandia.gov

Legion Retreat
7 December, 2022
Stanford University



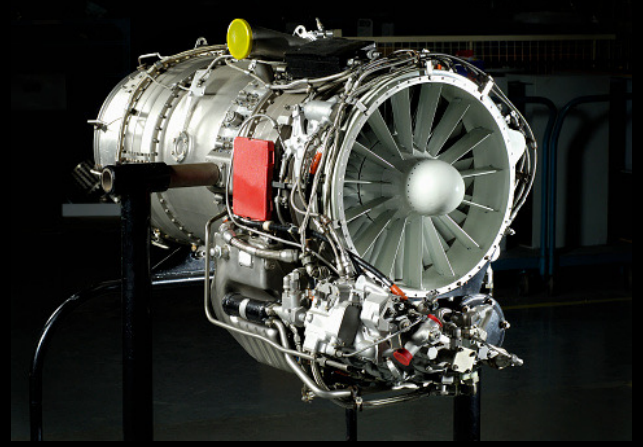
HO_2 and OH in a lifted H2/Air Jet Flame

Office of Science
U.S. Department of Energy

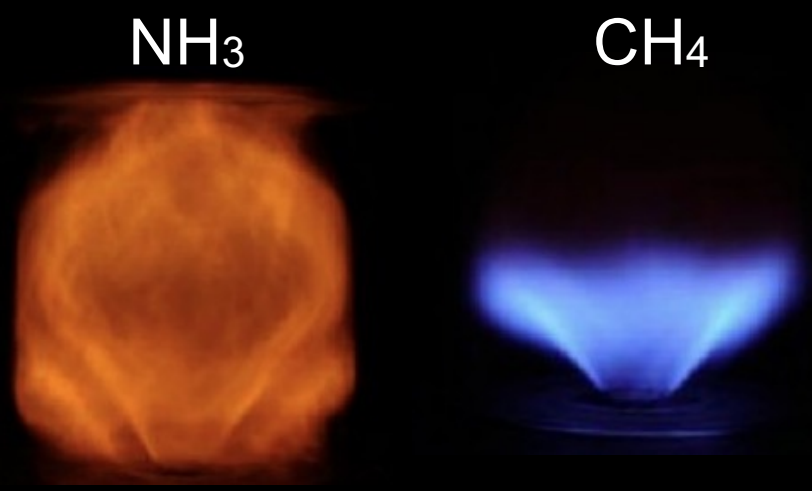
Office of Science
INCITE
Advancing America's Science and Industrial Competitiveness

Sandia National Laboratories

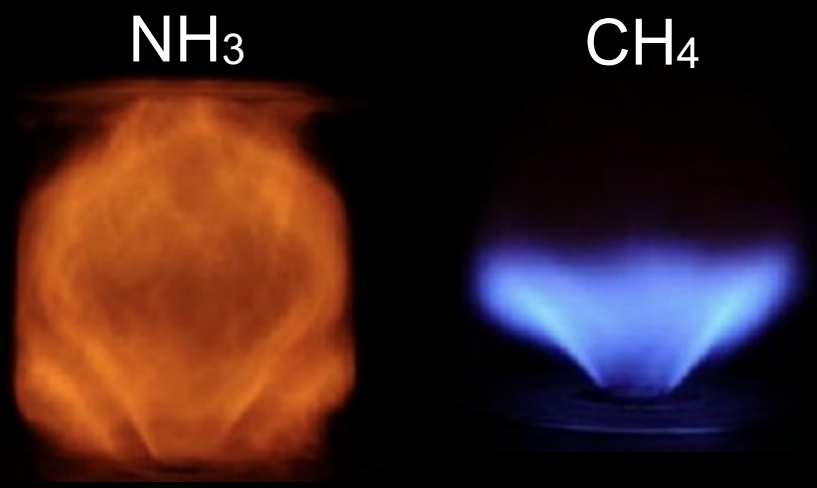
Ammonia/hydrogen has the potential to be an alternative **zero-carbon** fuel for hard-to-electrify sectors: long-haul shipping, power generation and agriculture



However, neat ammonia has a low reactivity compared to, e.g., natural gas (CH_4)



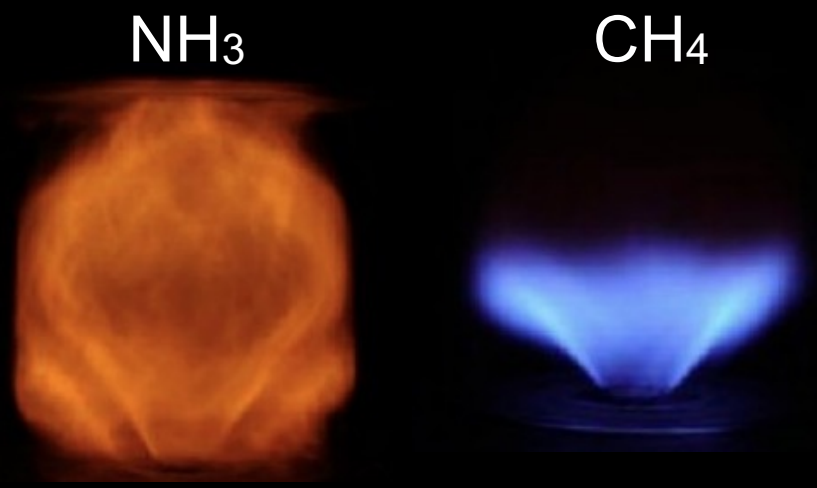
However, neat ammonia has a low reactivity compared to, e.g., natural gas (CH_4)



Valera-Medina et al., PECS, 2018.

The reactivity can be improved by adding hydrogen or partial cracking of ammonia to hydrogen and nitrogen

However, neat ammonia has a low reactivity compared to, e.g., natural gas (CH_4)



Valera-Medina et al., PECS, 2018.

The reactivity can be improved by adding hydrogen or partial cracking of ammonia to hydrogen and nitrogen

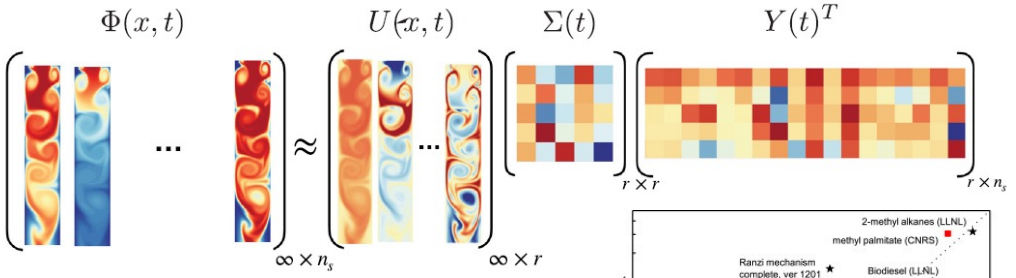
Combustion behavior of ammonia/hydrogen blends is poorly understood, especially behavior at high pressure

The convergence of exascale computing and data science

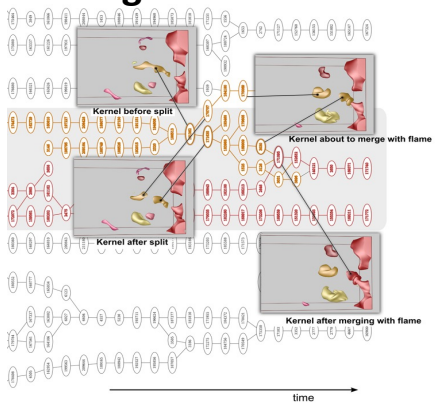
HPE/AMD OLCF-5 (Frontier), 2 Exaflop 2022-2023

Intel/HPE ALCF-3 (Aurora), 2 Exaflop, 2023-2024

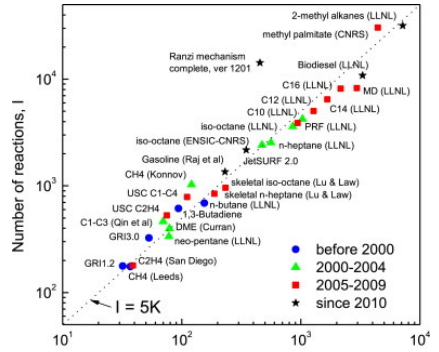
Reduced Order Surrogate Models

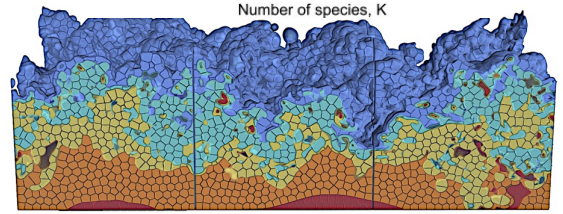


Merge Trees

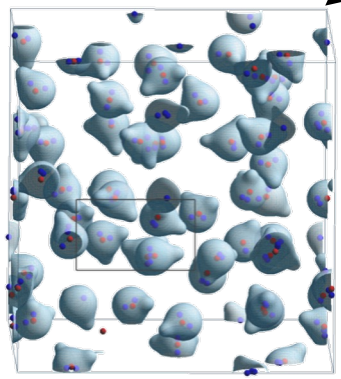


In Situ Feature Detection/Tracking, Statistical Aggregation and Viz

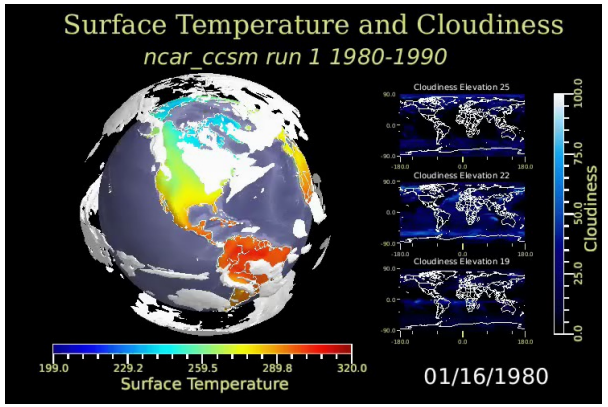




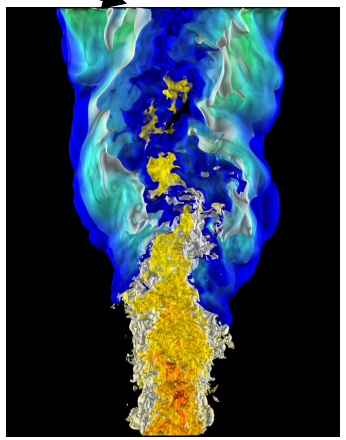
Voronoi tessellation



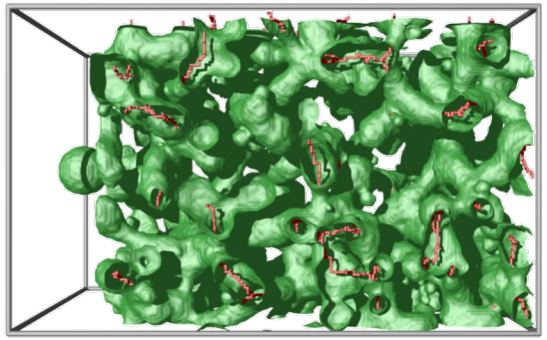
Molecular Dynamics



Climate Modeling



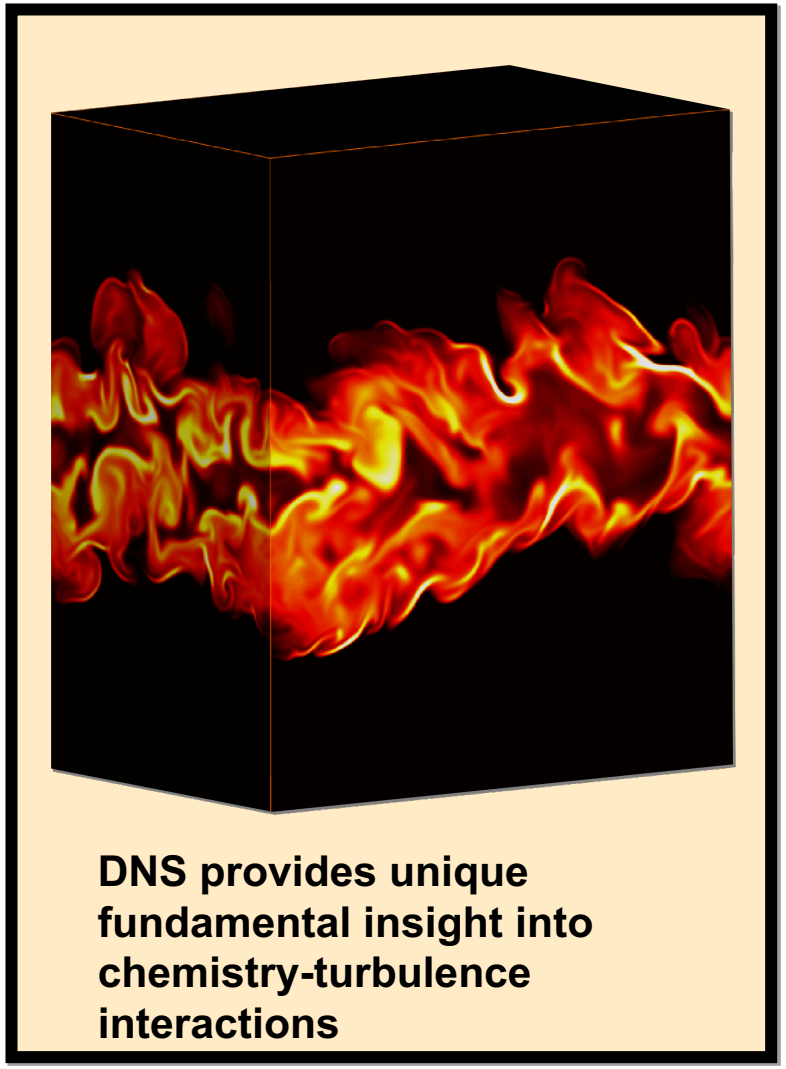
Combustion Modeling



Material Sciences

Direct Numerical Simulation of Turbulent Combustion: S3D

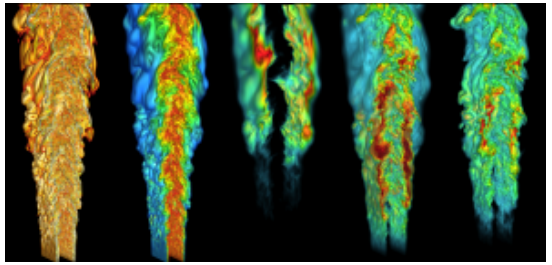
- Solves compressible reacting Navier-Stokes, total energy and species continuity equations
- High-order finite-difference methods
- Detailed reaction kinetics and molecular transport models
- Lagrangian particle tracking (tracers, spray, soot)
- Shock capturing
- Multi-temperature method (nonequilibrium)
- *In situ* analytics and visualization
- Geometry using immersed boundary method and multi-block approach
- Refactored for heterogeneous architectures using dynamic task based programming model (Legion/Regent)

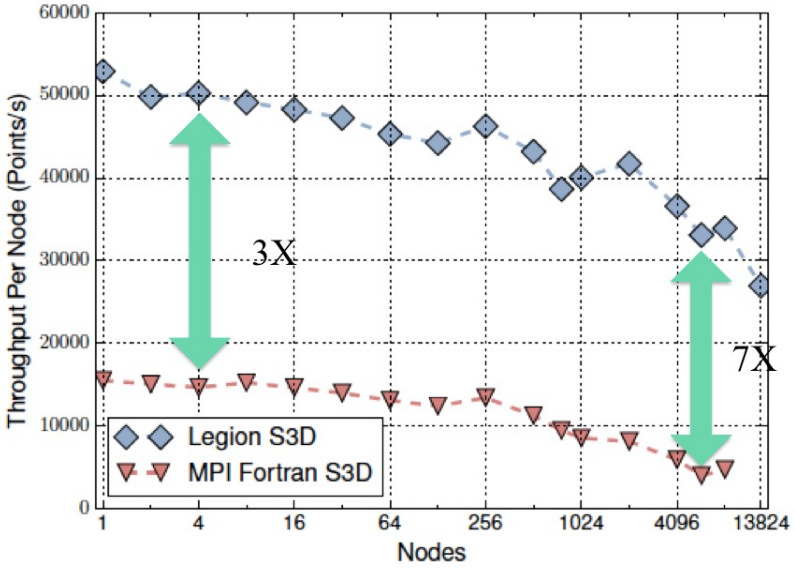


Chen *et al.*, *Comp. Sci. Disc.*, 2009

Legion Programming System applied to S3D

- A data-centric parallel programming system
- A programming model for heterogeneous, distributed machines
  - Automates many aspects of achieving high performance, such as extracting task- and data-level parallelism
  - Automates details of scheduling tasks and data movement (*performance optimization*)
  - Separates the specification of tasks and data from the mapping onto a machine (*performance portability*)
- Legion application example: S3D
  - Production combustion simulation
  - Written in ~200K lines of Fortran
  - Direct numerical simulation using explicit methods





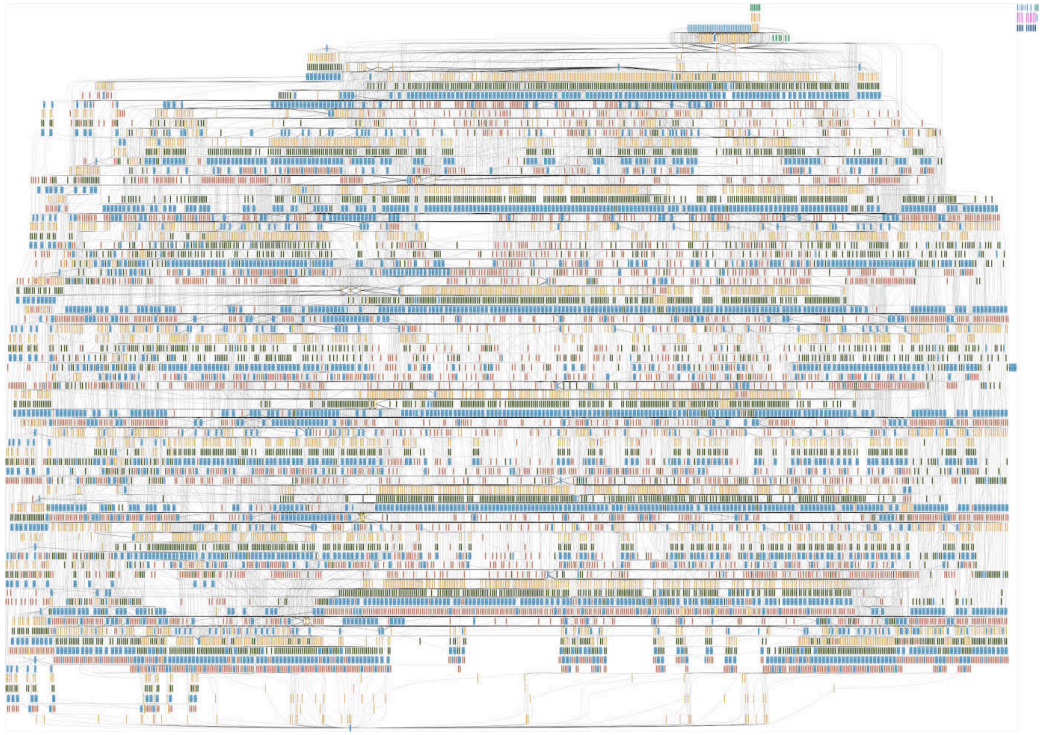
S3D performance Legion vs. MPI

S. Treichler et al., "S3D-Legion: An Exascale Software for Direct Numerical Simulation (DNS) of Turbulent Combustion with Complex Multicomponent Chemistry," in press, CRC Book on Exascale Scientific Applications: Programming Approaches for Scalability Performance and Portability, 2017.

Regent applied to Legion S3D

- Regent is a programming language for task-based implicit parallelism
- Regent programs are composed of tasks (functions that perform computations on regions eligible for parallel execution) and logical regions (hierarchical collections of structured objects represented as cross-product of an index space and field space)
- Regent programs execute with sequential semantics with no explicit synchronization and a type system ensures program correctness on parallel/distributed machines
- Regent programs are transformed into efficient implementations for Legion
- Regent employs several novel compiler optimizations to minimize the dynamic overhead of the runtime system, especially at large node counts

Slaughter et al. 2015



Dependence Graph of S3D

ECP EXASCALE COMPUTING PROJECT

Pressure Effects on Turbulent Premixed Lean Ammonia/Hydrogen Flames

Martin Rieth[2], Andrea Gruber[3], Myoungkyu Lee[2], Forman Williams[4], Jacqueline Chen[2]

Visualization: Tyson Neuroth[1], Qi Wu[1], Kwan-Liu Ma[1]

[1] University of California, Davis, [2]Sandia National Laboratories, [3]SINTEF Energy Research, [4]Univerisity of California at San Diego

Rieth et al. Comb Flame 2022

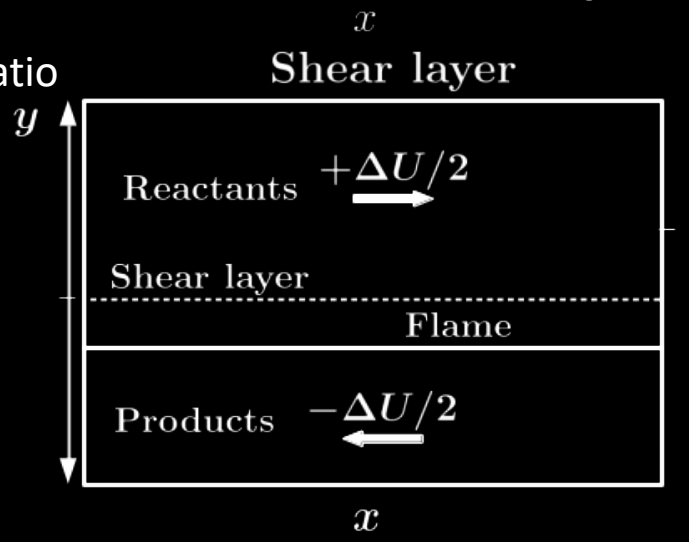
Pressure effects on ammonia/hydrogen premixed flames: direct numerical simulations in a temporal planar shear layer configuration

Simulation Parameters

- Lean premixed $NH_3/H_2/N_2$-air (40%/45%/15% vol) with equivalence ratio of 0.45 at **1 and 10 atm**
- Reactants are preheated to 750 K
- Flames are nominally in the broken reaction zone regime (Karlovitz # (Ka)>600, turbulent Re # (Re_t)>1000)
- Simulations are designed such that normalized parameters (Re_t, Ka, Damköhler #) are consistent between 1 and 10 atm
- Growing grid size with 11 B grid points on the final grid
- 19 species chemical mechanism (Jiang et al., 2020)
- DNS using S3D-Legion run on OLCF's 200 Pflop Summit supercomputer on ~1000 nodes using a DOE INCITE allocation

More information: Rieth et al., *Comb. Flame*, 2022

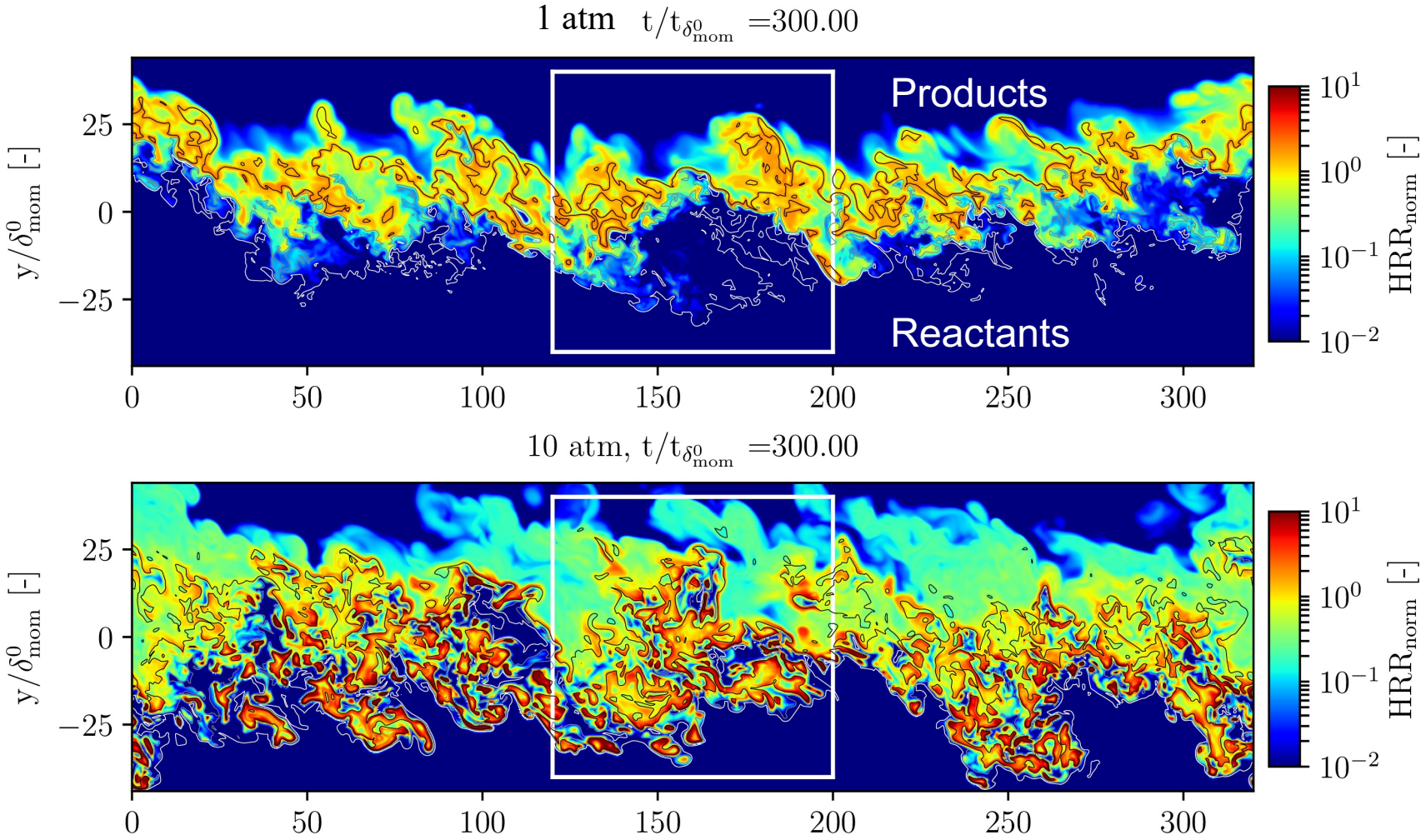
Simulation setup



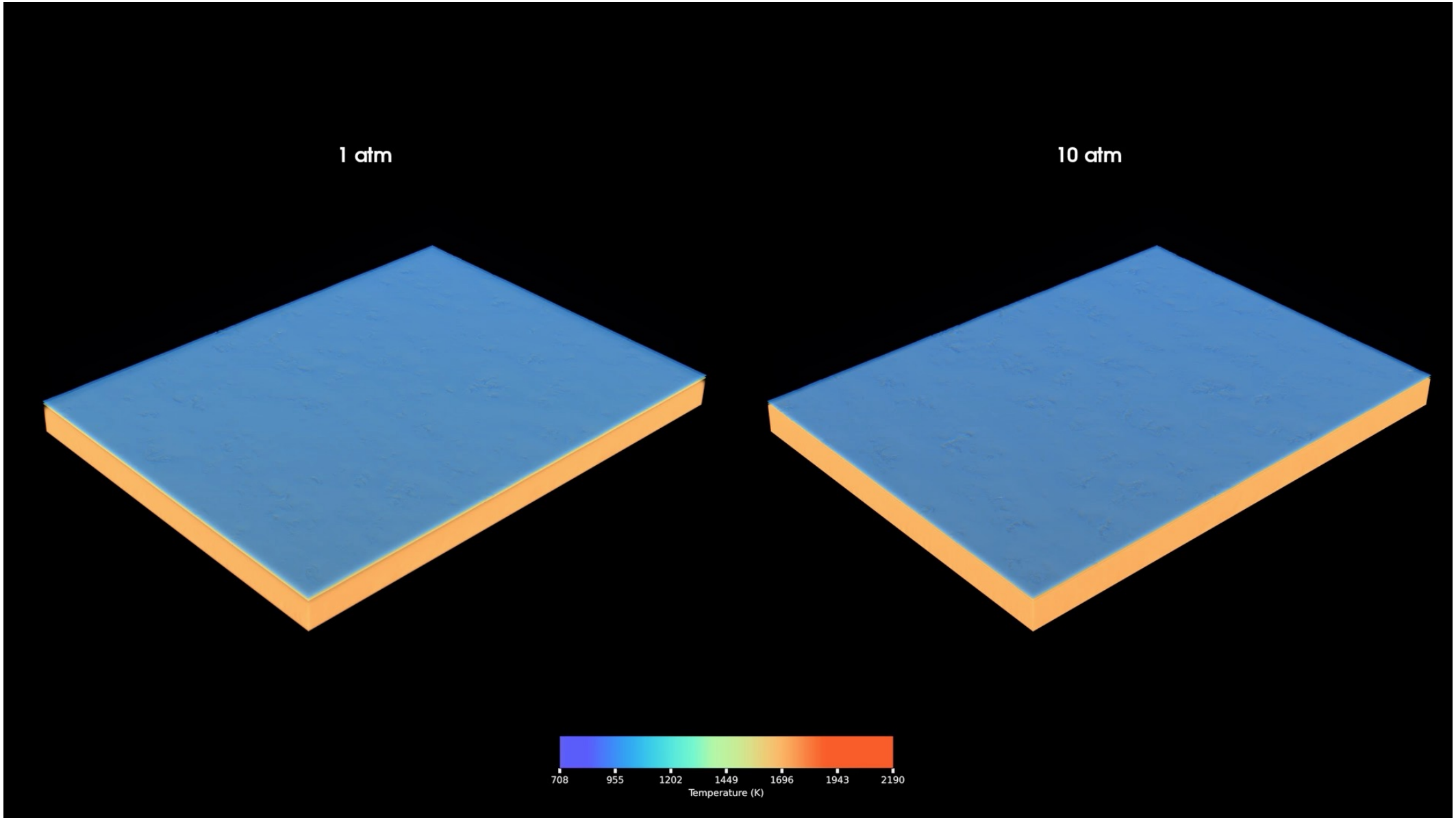
$-10\delta^0_{\mathrm{mom}}$ at $t = 100t_{\delta^0_{\mathrm{mom}}}$

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING

Heat release rate from $NH_3/H_2/N_2$/air premixed flame in turbulent mixing layer at 1 and 10 atm



Iso-lines correspond to temperatures of 800 (white), 1000(grey) and 1500 K (black)

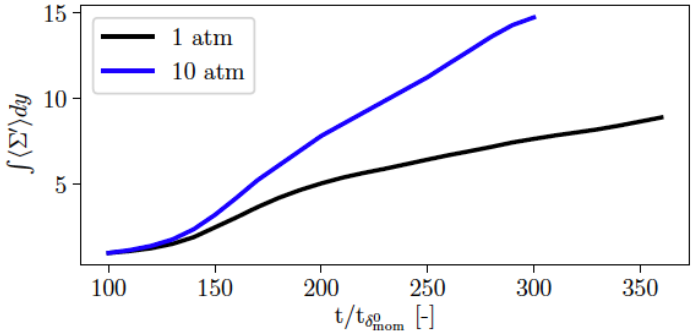


- Strong shear-driven turbulence wrinkles and disrupts the flame front
- More wrinkling and cellular structures appear at 10 atm, disrupted preheat zone at 1 atm
- Strong super-adiabaticity at 10 atm, larger fluctuations in equivalence ratio

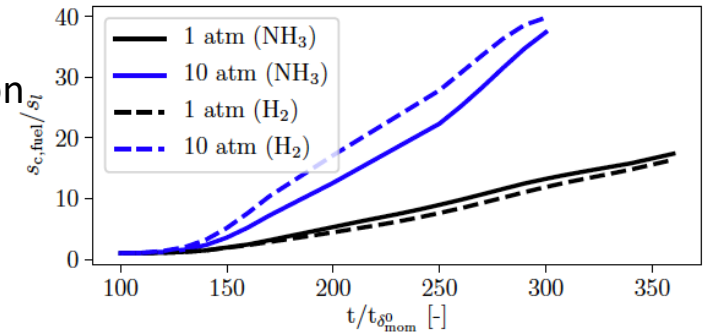
Thermo-diffusive instabilities and preferential diffusion

Flame Surface Area and Burning Rate

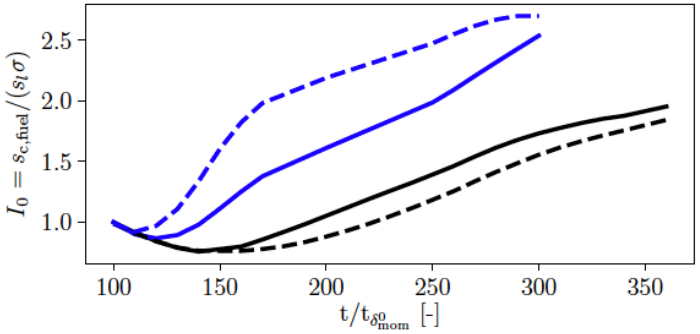
Flame surface density



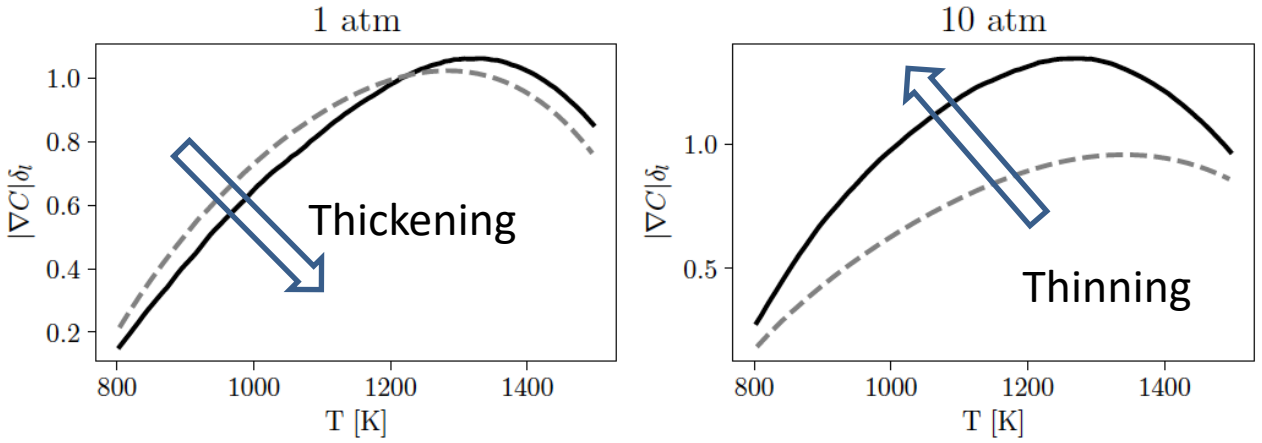
Fuel consumption speed



Burning Intensity



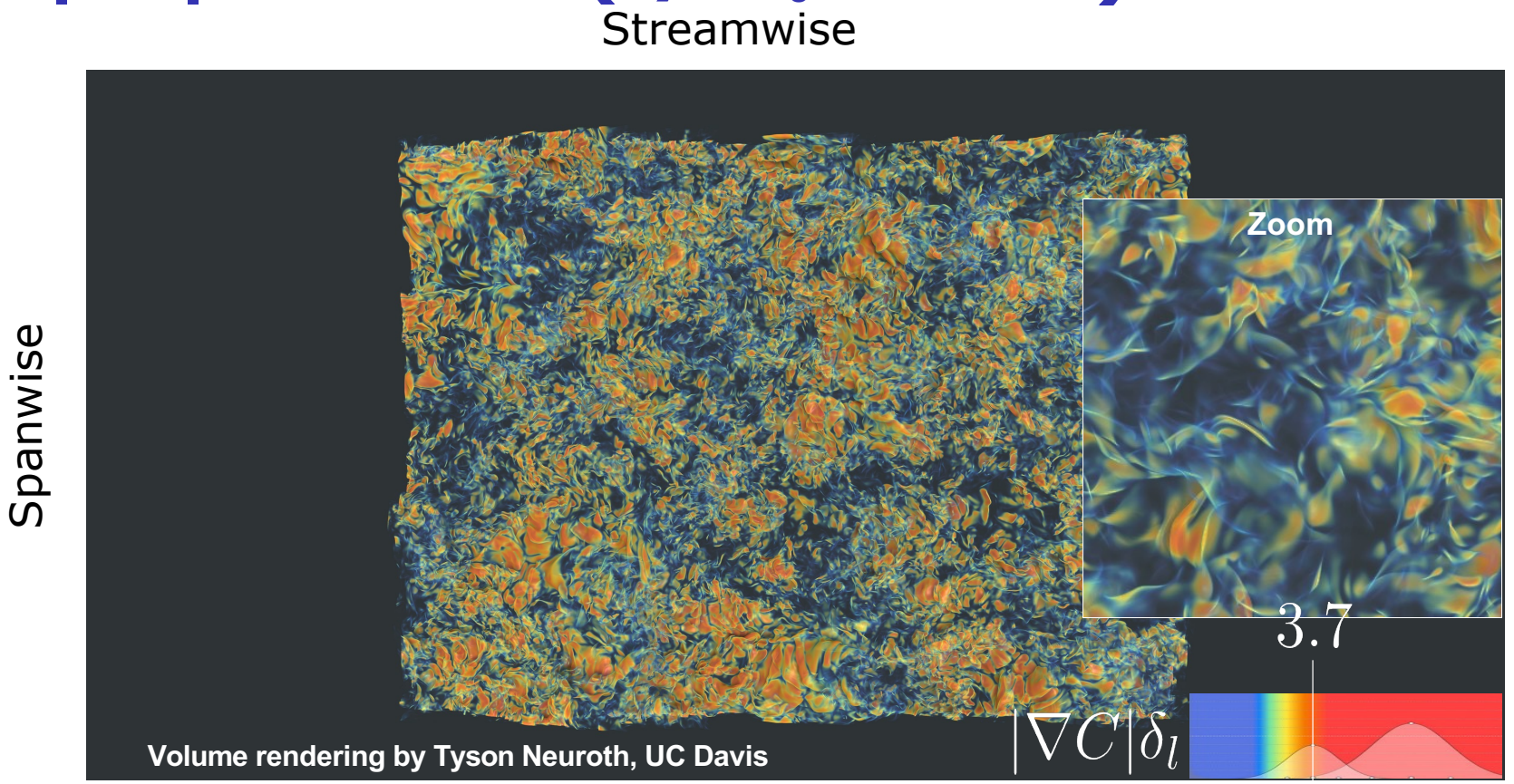
Flame thickness



Black lines represent DNS solution at $t/t_{\delta^0_{mom}} = 300$. Grey dashed lines represent laminar flame solution.

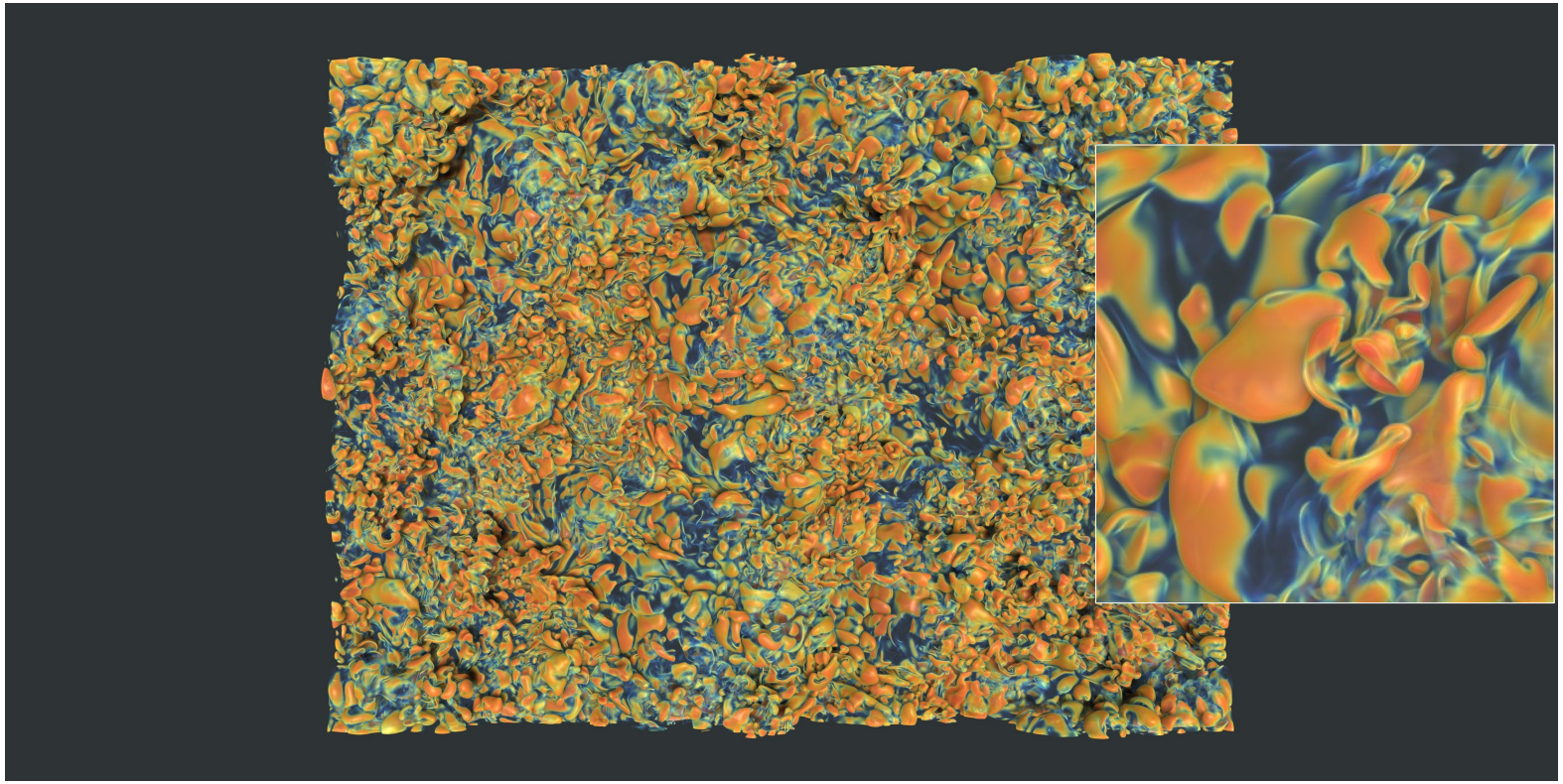
- Faster increase of flame surface density ($\int \langle \Sigma' \rangle dy$), overall fuel consumption speed ($s_{c,fuel}/s_l$) and burning intensity (I_0)
- Thinner flames at 10 atm (larger $|\nabla C|\delta_l$)

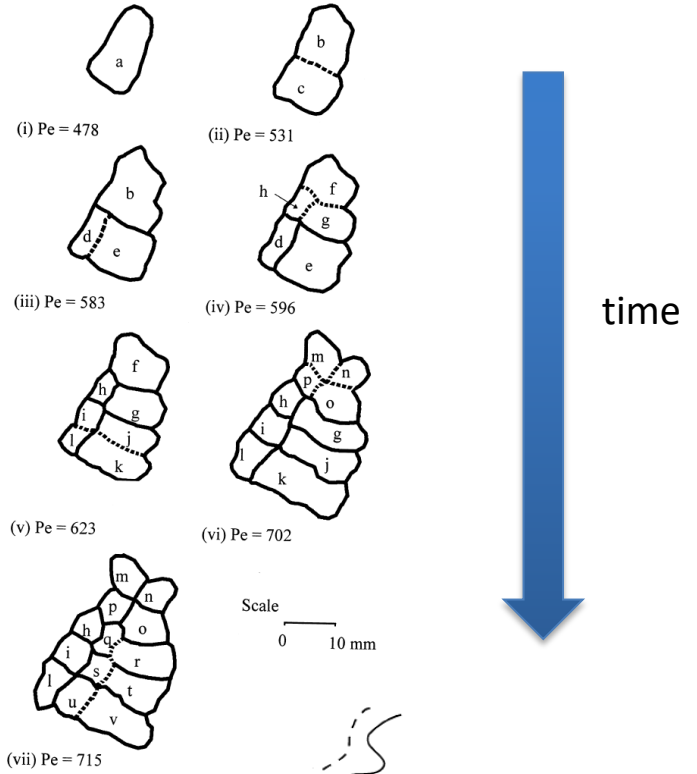
Volume Rendering of Fine-Grained Flame Surface Density $|\nabla C|$ at 1 atm ($t/t_{\delta_0} = 300$)



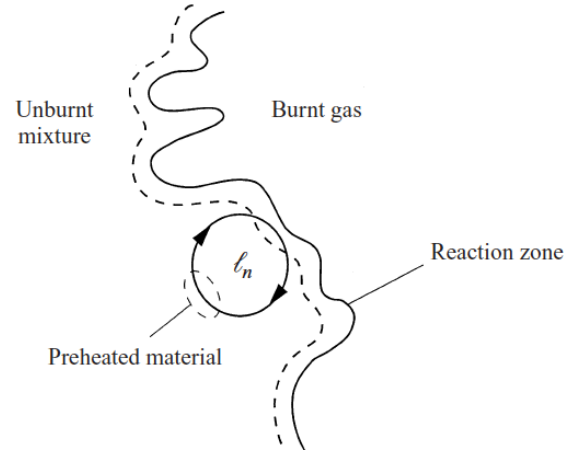
At 1 atm, turbulence disrupts the flame structure to a large extent, but some cellular flame structures can be observed

What are the dynamics of cellular flame structure evolution for thermo-diffusively unstable ammonia/hydrogen/air flames and how does this affect flame surface area growth?





How does the flame structure change with pressure conditioned on distance function or reaction progress?



Summary

- Performed DNS of lean (ϕ=0.45) $NH_3/H_2/N_2$-air flames at 1 and 10 atm in intense sheared turbulence

- Despite nominally similar ‘turbulence-flame’ interaction parameters, both flames behave very differently:
  - 1 atm flame exhibits significant preheat zone broadening and strongly disrupted flame surface
  - 10 atm flame features cellular structure, strong super-adiabaticity, faster flame surface area growth, faster fuel consumption and thinner flames despite the high Karlovitz number

- Thermo-diffusive instabilities need to be accounted for in modeling (also shown by $NH_3/H_2/N_{air}$ and CH_4-air flame comparison)

In Situ Time-dependent reduced-order modeling of transient combustion phenomena

Swapnil Desai[1], Jacqueline H. Chen[1], Kisung Jung[1], Seshu Yamajala[2], Hessam Babaee[3]

[1]Sandia National Laboratories

[2]SLAC

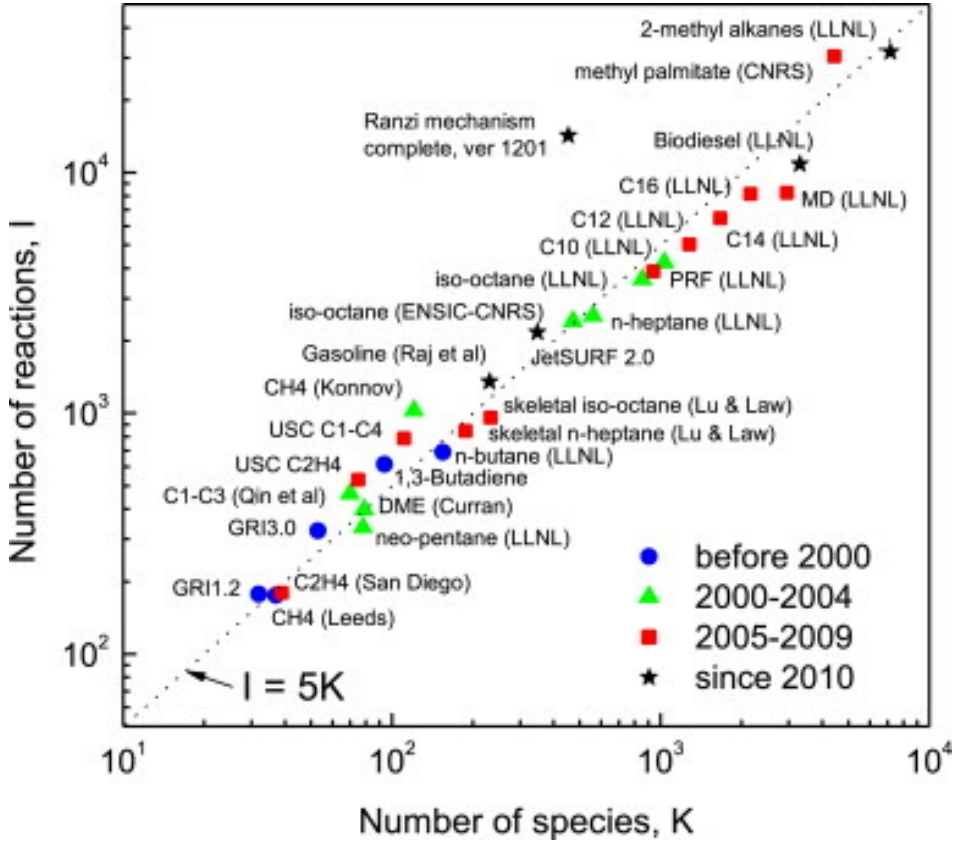
[3]University of Pittsburgh

Motivation

Direct Numerical Simulation (DNS)

Advantages:
- **Full access to time resolved 3D fields**

Disadvantages:
- **Huge computing and memory requirements**

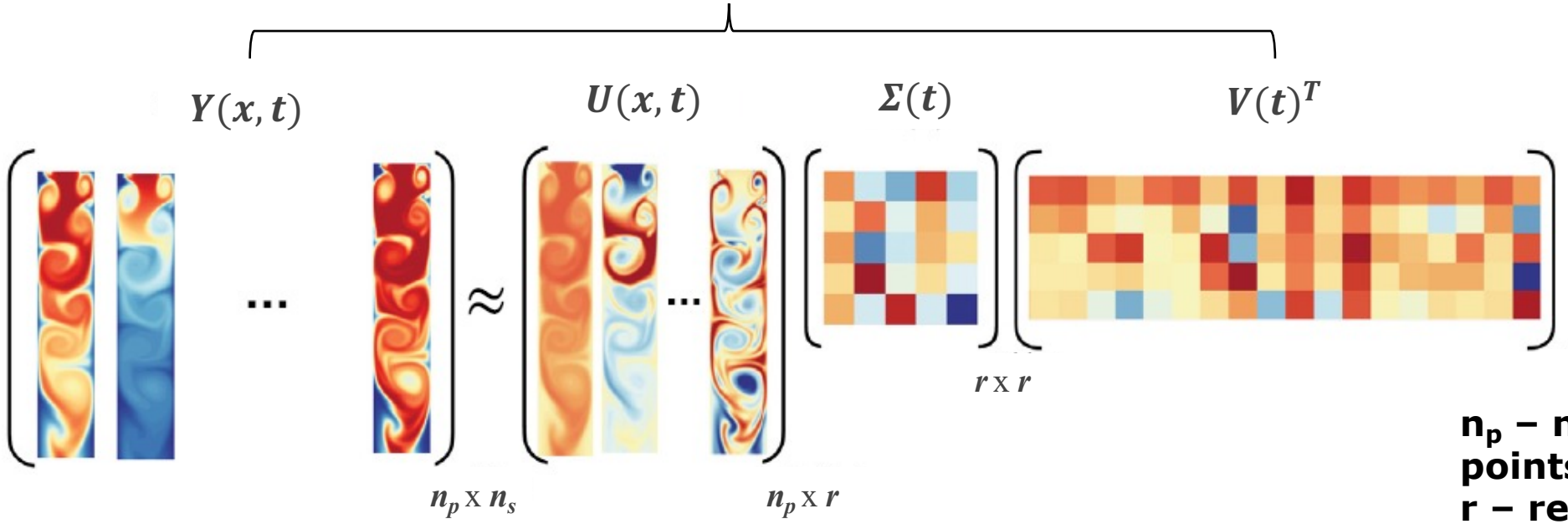


Solutions:
- Use of skeletal / reduced kinetic mechanisms
  - Mechanism size for gasoline / diesel surrogates still upwards of 100 species and 1000 reactions
- Principal component analysis (PCA) based reduced-order modeling (ROM)
  - Find low-dimensional manifold in composition space as a surrogate for full system dynamic of the reacting flow, exist due to inherent correlations of chemical species
  - Need for high fidelity training data

1: Egolfopoulos et al., PECS, 2014

Dynamically bi-orthonormal (DBO) decomposition based ROM

Singular Value Decomposition (SVD)



n_p – number of grid points
r – reduction size
n_s - number of species ($r \ll n_s$), includes the bath gas

Idea: Obtain on-the-fly low-rank decomposition of species transport equation[1,2]

Advantages:

- No need to generate training data
- Extract instantaneous correlations between different species on the fly and directly from the species transport equation
- No need to store the entire species vector
- Potential to scale linearly w.r.t. data size and low-rank-r without requiring to solve large scale optimization problem

1: Patil et al., JCP 2020
2: D. Ramezanian et al., Comp. Met. Appl. Mech. & Eng., 2021

DBO: Math

$$\frac{\partial Y_i}{\partial t} = -\frac{1}{\rho}\left(u\frac{\partial Y_i}{\partial x} + v\frac{\partial Y_i}{\partial y} + w\frac{\partial Y_i}{\partial z}\right) - \frac{1}{\rho}\left(\frac{\partial J_{i,x}}{\partial x} + \frac{\partial J_{i,y}}{\partial y} + \frac{\partial J_{i,z}}{\partial z}\right) + \frac{W_i\dot{\omega}_i}{\rho}$$ (Modified Governing equation)

$$J_{i,\alpha} = -\rho D_i^{mix}\left[\nabla_\alpha Y_i - Y_i W \sum_{j=1}^{N_s}\frac{\nabla_\alpha Y_j}{W_j}\right]$$

Evolve the governing equation for few time steps (~1000) with DNS so that rank (i.e. r) based on instantaneous-PCA increases and then switch to DBO

$$\frac{\partial U}{\partial t} = \prod_{\perp U}\left[-\left(u\frac{\partial U}{\partial x} + v\frac{\partial U}{\partial y} + w\frac{\partial U}{\partial z}\right) + \left(\frac{\partial J_{i,x}V}{\partial x} + \frac{\partial J_{i,y}V}{\partial y} + \frac{\partial J_{i,z}V}{\partial z}\right)\Sigma^{-1} + \frac{W_i\dot{\omega}_i}{\rho}V\Sigma^{-1}\right] \rightarrow PDE$$

$$\frac{d\Sigma}{dt} = \left\langle U, u\frac{\partial U}{\partial x} + v\frac{\partial U}{\partial y} + w\frac{\partial U}{\partial z}\right\rangle\Sigma + \left\langle U, \frac{\partial J_{i,x}V}{\partial x} + \frac{\partial J_{i,y}V}{\partial y} + \frac{\partial J_{i,z}V}{\partial z}\right\rangle + \left\langle U, \frac{W_i\dot{\omega}_i}{\rho}V\right\rangle \rightarrow ODE$$

$$\frac{dV}{dt} = \prod_{\perp V}\left[\left\langle\frac{\partial J_{i,x}}{\partial x} + \frac{\partial J_{i,y}}{\partial y} + \frac{\partial J_{i,z}}{\partial z}, U\right\rangle + \left\langle\frac{W_i\dot{\omega}_i}{\rho}, U\right\rangle\right]\Sigma^{-T} \rightarrow ODE$$

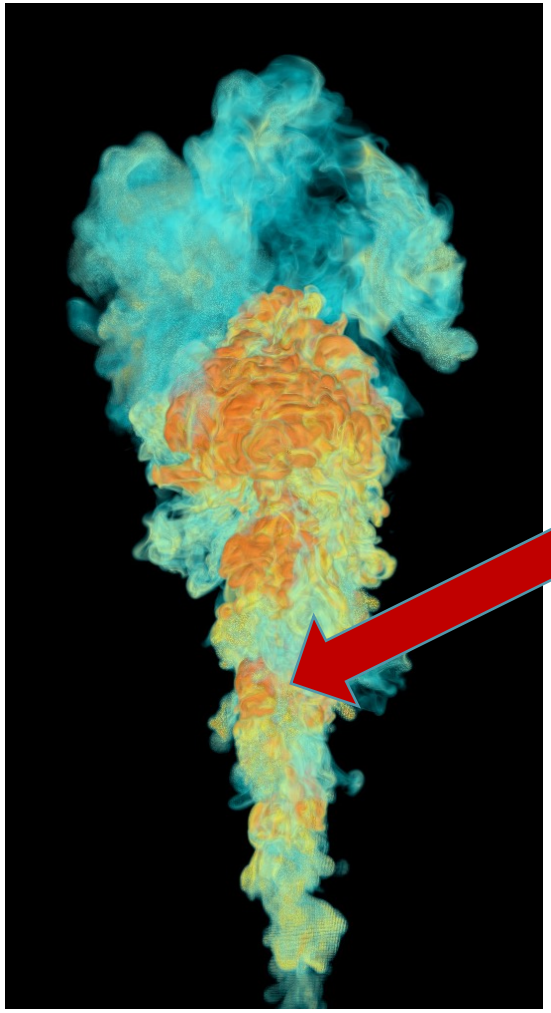
D. Ramezanian et al., Comp. Met. Appl. Mech. & Eng., 2021

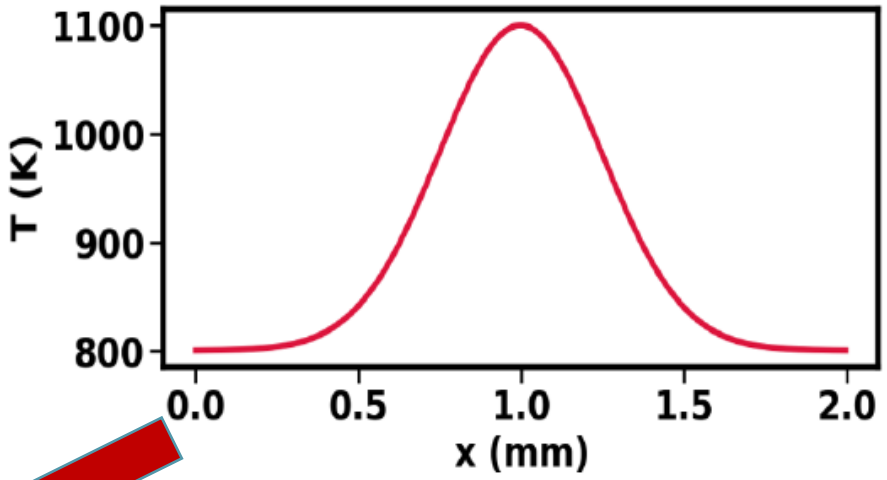
$$\prod_{\perp U} f = f - U\langle U, f\rangle \quad \text{and} \quad \prod_{\perp Y} z = z - YY^T z$$

$$< U, f >= \sum_{p=1}^{p=m} matmul(U^T, f),$$

$$m = number\ of\ processors$$

Canonical test case: 1D hotspot of lean n-heptane/air





Pele DNS of turbulent multi-injection autoignitive diesel jet

T_{mean} = 800 K, T_{peak} = 1100 K, P = 50 bar, phi = 0.4

52 species n-heptane air mechanism[1]

S3D[2]: 8th order finite difference operator, 4th order, 6-stage explicit Runge-Kutta time integrator, LAPACK for linear algebra operations

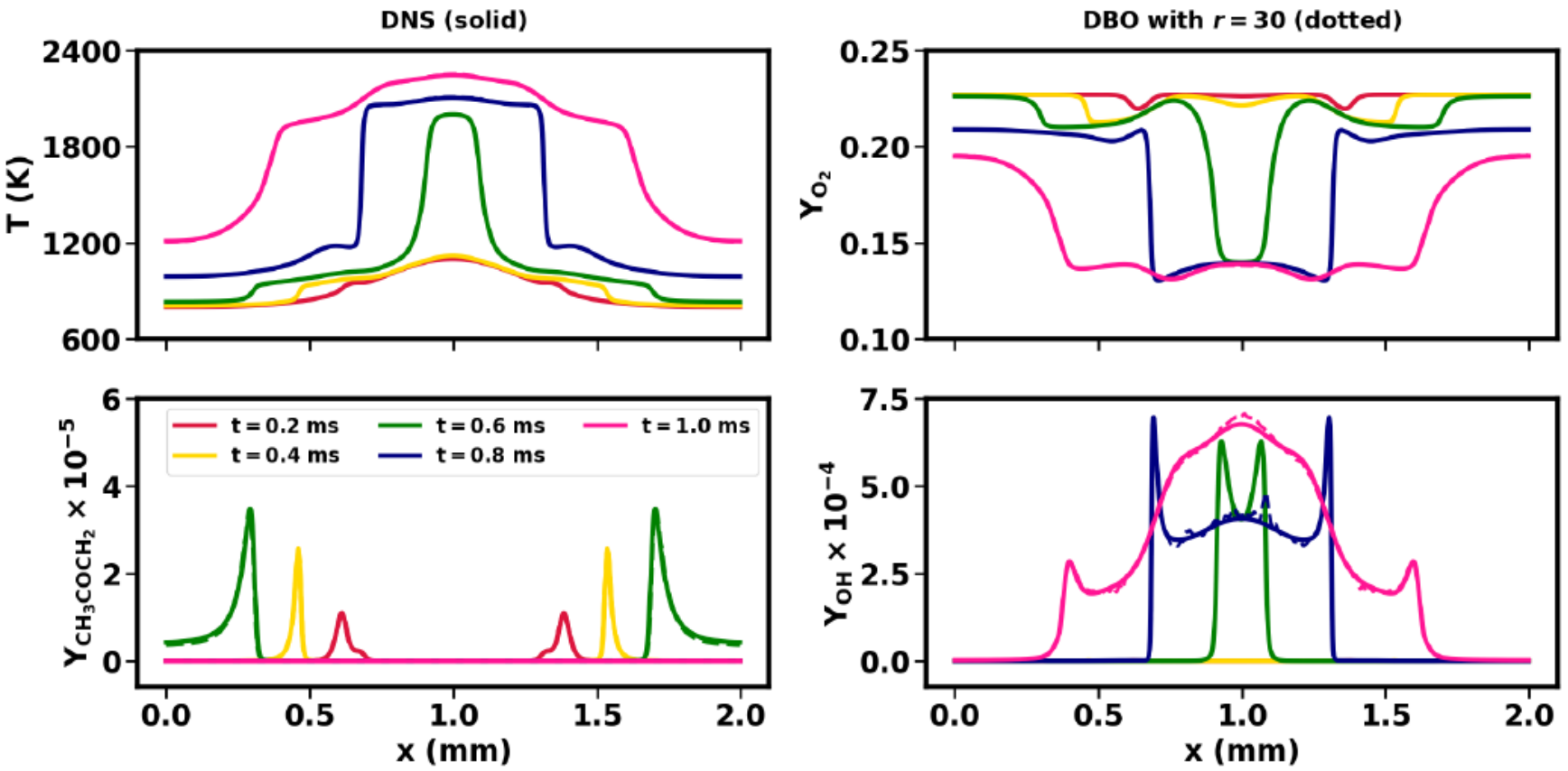
2 micron grid resolution, 1 ns time step size

Periodic boundaries at x = 0 and x = 2 mm

DBO started at t = 0.2 ms and ran for additional 0.8 ms.

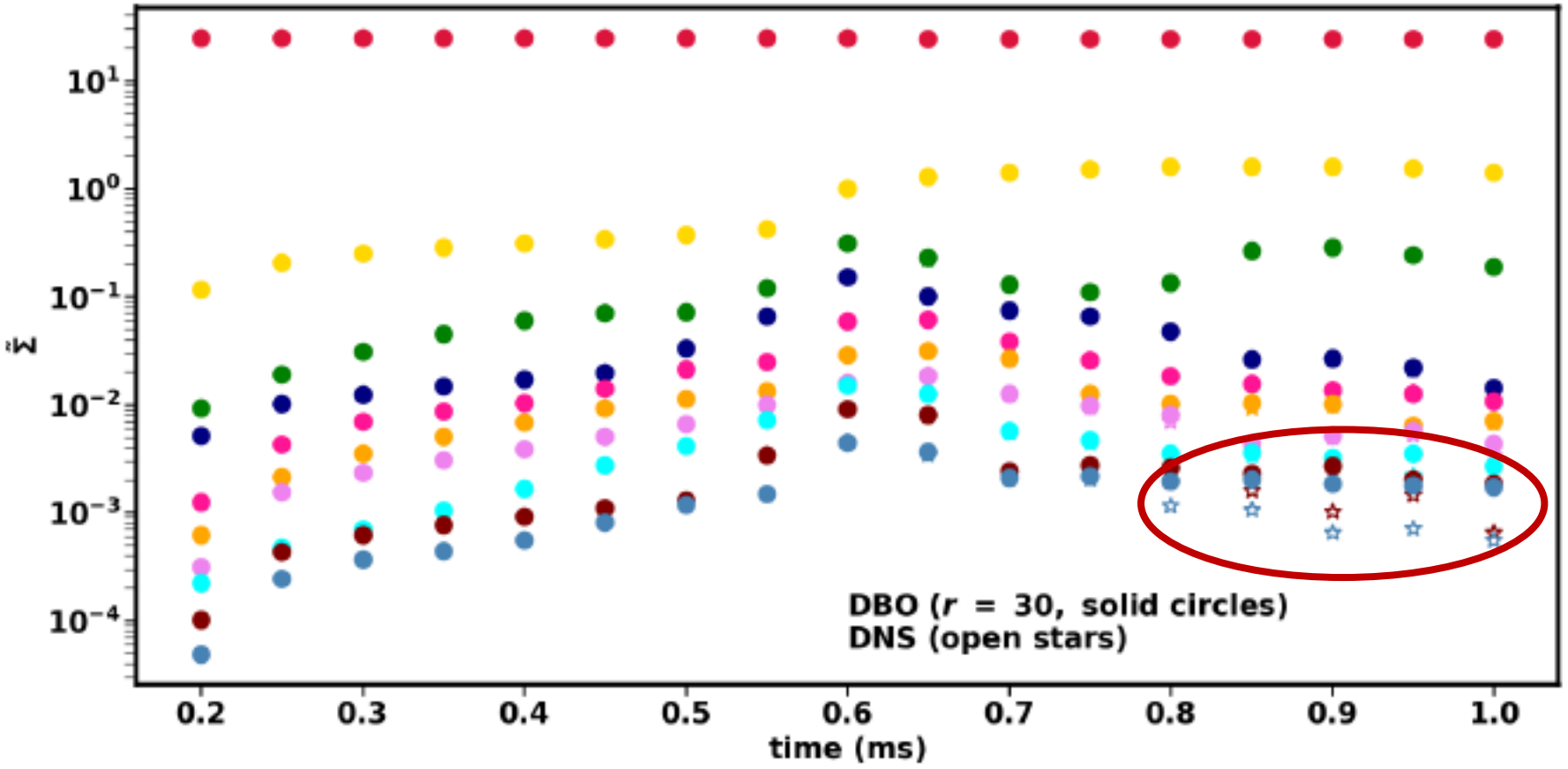
1: Lu et al., Combust. Flame, 2009
2: Chen et al., Comp. Sci & Disc. 2009.

Transient auto-ignition: DNS versus DBO



- DBO simulation performed with r = 30
- Low temperature chemistry identified by CH_3COCH_2 radical
- High temperature chemistry identified by OH radical
- DBO able to capture the entire transient ignition process: low-temperature ignition → cool flame propagation → high-temperature ignition → hot flame propagation
- Differences in the OH profiles appear around 0.8 ms

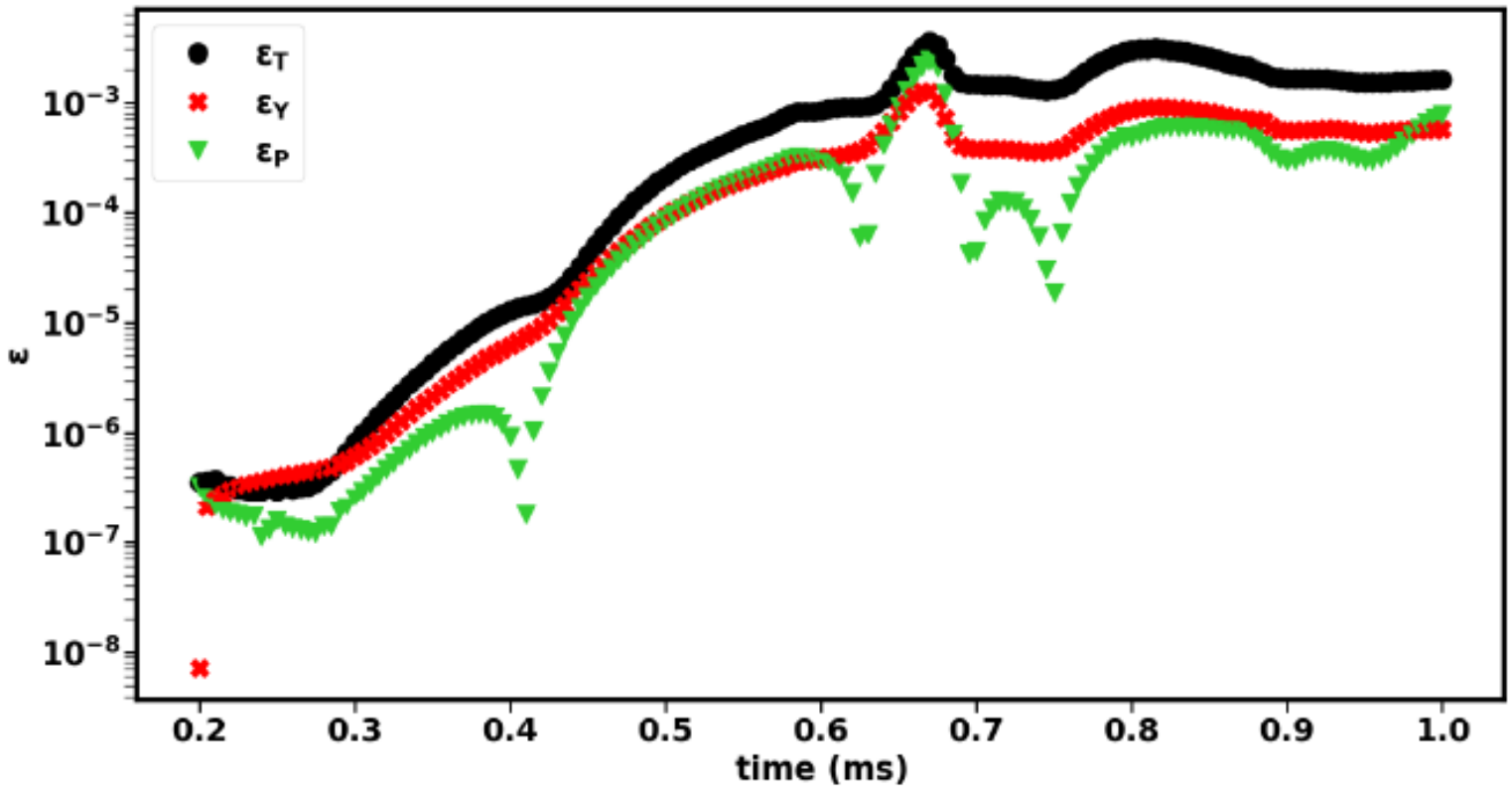
Temporal evolution of covariance matrix: DNS (Instantaneous-PCA) versus DBO



- DBO accurately captures first 10 singular values until t = 0.7 ms beyond which the accuracy of DBO degrades
- Since r < n_s, errors accumulate over time due to lost interactions of DBO modes with unresolved modes
  - Affects the modes with smaller singular values more intensely[1]
  - Spurious oscillations appear in mass fraction profiles at later times

1: D. Ramezanian et al., Comp. Met. Appl. Mech. & Eng., 2021

DBO approximation error



$$\epsilon(t) = \|Y(x,t) - U(x,t)\Sigma(t)V(t)^T\|_F / \|Y(x,t)\|_F$$

$\| \|_F \rightarrow$ Frobenius Norm

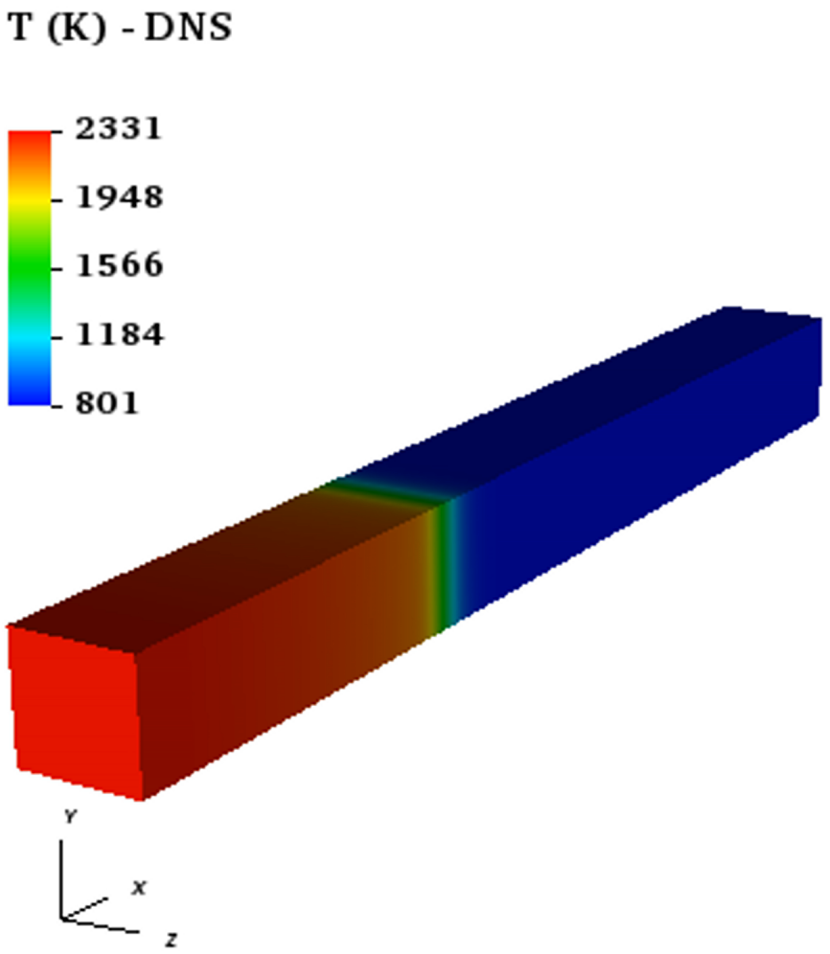
Full rank solution

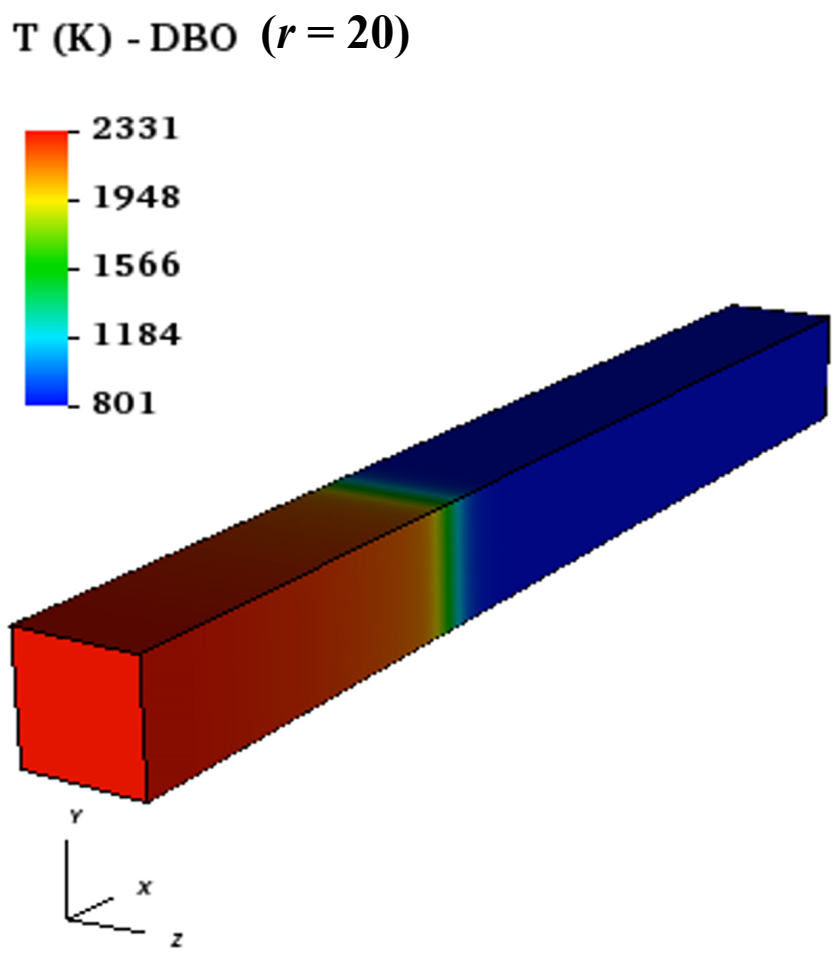
DBO reconstruction

Legion S3D-DBO

- DBO implementation in S3D ported to Legion/Regent
- All kernels ported to GPU
- Performed weak scaling experiments on Perlmutter
  - 160x64x64 Local Grid
  - Using Shell-D surrogate mechanism for gasoline (39 species)
  - 4 ranks/node (1 gpu/rank)
  - Scaled from 1 – 64 nodes

Test Case: 3D Flame Propagation at Constant Pressure





Initial Condition:

T_u = 800 K, P = 1 bar, phi = 1.0

L_x x L_y x L_z = 0.44 x 0.044 x 0.044 cm^3

10-micron grid resolution

39-species gasoline surrogate mechanism

Left X-Boundary: Adiabatic wall

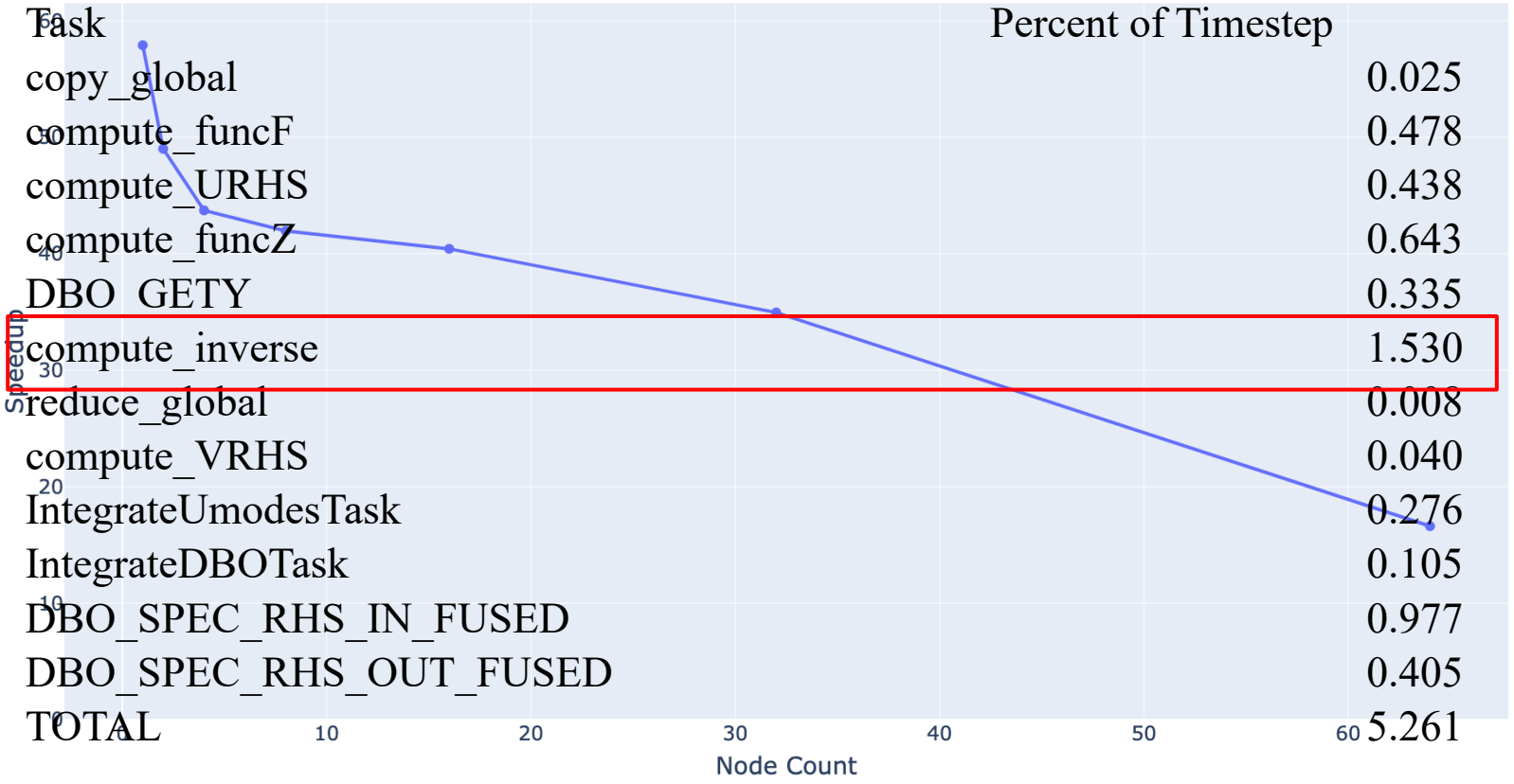
Right X-Boundary: Constant pressure outlet

Periodic in Y and Z

DBO-ROM with *r*=20 started at t=0

Weak Scaling Performance of DBO-ROM

Speed-up in S3D-Regent V/s S3D-Fortran

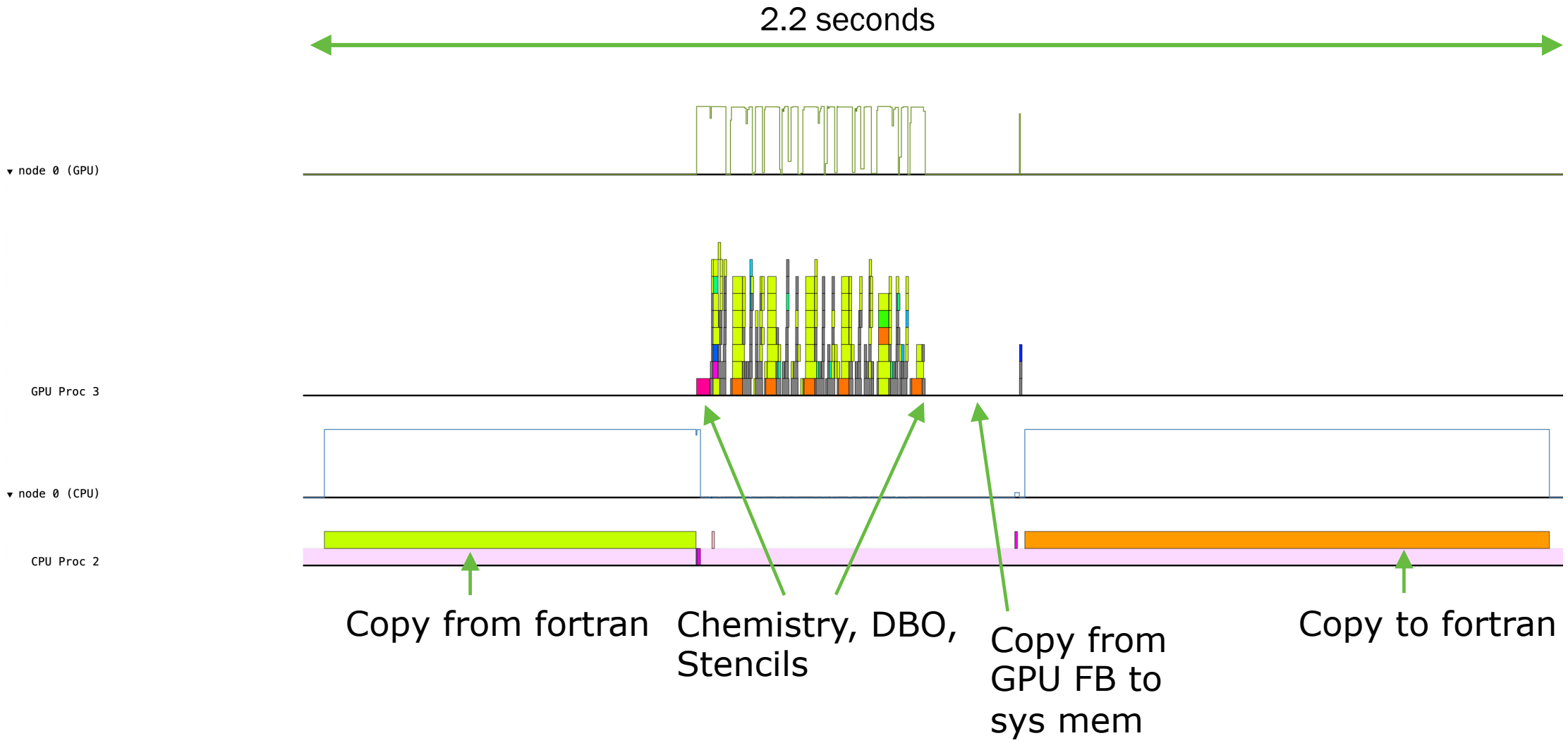


Task	Percent of Timestep
copy_global	0.025
compute_funcF	0.478
compute_URHS	0.438
compute_funcZ	0.643
DBO_GETY	0.335
compute_inverse	1.530
reduce_global	0.008
compute_VRHS	0.040
IntegrateUmodesTask	0.276
IntegrateDBOTask	0.105
DBO_SPEC_RHS_IN_FUSED	0.977
DBO_SPEC_RHS_OUT_FUSED	0.405
TOTAL	5.261

New DBO tasks take around 5% of total execution of timestep

- DBO implementation ported to Legion/Regent
- Performed weak scaling experiments on Perlmutter at NERSC
  - 160x64x64 Local Grid
  - 39 species gasoline surrogate mechanism
  - 4 ranks/node
  - 1 gpu/rank
  - Scaled from 1 – 64 nodes

Legion Prof Profile



Summary

- DBO based ROM of compressible reacting flows demonstrated to capture strongly transient phenomena such as multi-stage autoignition in domains with non-trivial boundary conditions
- Depending on the problem configuration and the chemical mechanism, reduction in composition space observed to be approximately between 25% and 50%
  - Effective reduction possible even in heavily reduced chemical mechanisms of gasoline and diesel surrogates!

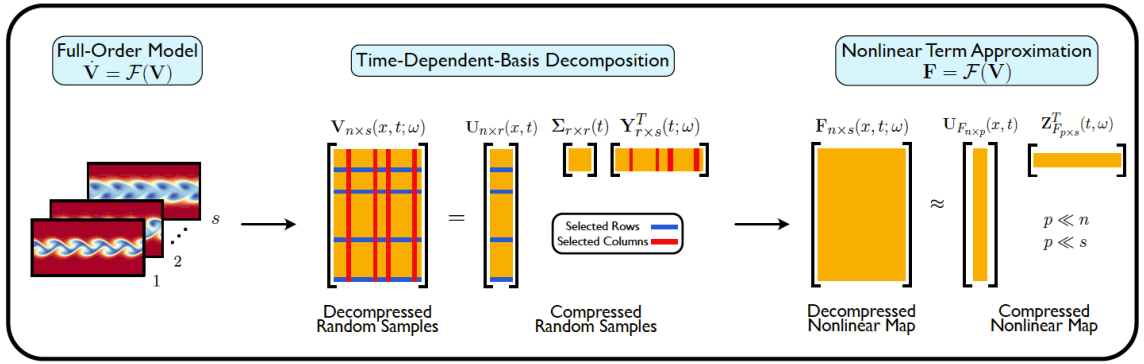
Planned Next Steps

- Perform additional cases with open/inflow/wall boundaries in multiple dimensions and with turbulence
- Extend current DBO-ROM implementation to enable:
  1) *Sparse* sampling using DEIM in DBO-ROM simulations
  2) *Dynamic* rank adaptivity, switching to full-rank simulation if needed

Future work—Speed up DBO in GPU environment

Sparse sampling

- The computation of the current DBO model requires decompressing $\mathbf{Y} = \mathbf{U\Sigma V^T}$ and applying the nonlinear map, $\mathcal{F}(\mathbf{Y})$, every time step
- In a sparse sampling framework, $\mathcal{F}(\mathbf{Y})$ is represented by the compressed form, $\mathcal{F}(\mathbf{Y}) \approx \mathbf{U}_F \boldsymbol{Z}_F^T$, where $\mathbf{U}_F$ and $\boldsymbol{Z}_F^T$ are defined by sampling the *p*-numbers of rows and columns of $\mathbf{U}$ and $\boldsymbol{V}^T$



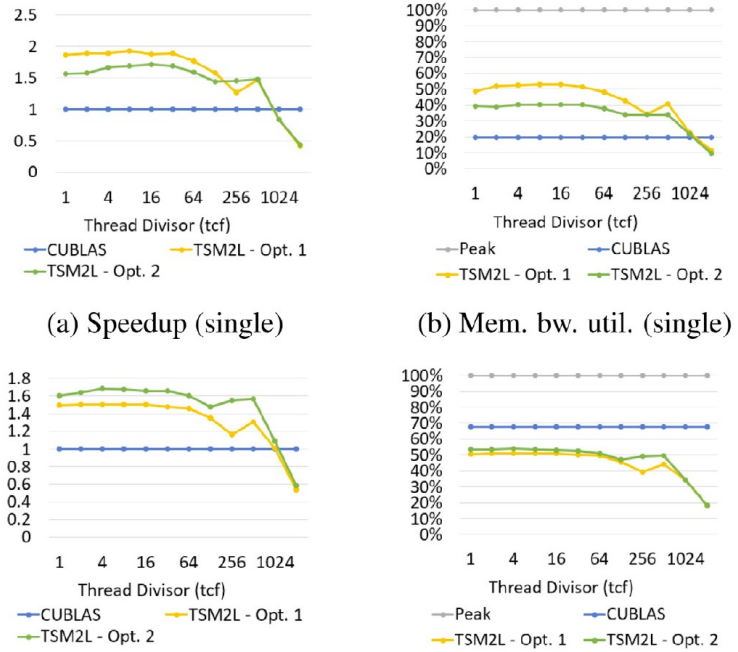
Schematic of the sparse sampling algorithm[1)]

Rank adaptation

- The rank of the ROM, r, is adjusted with time depending on the complexity of the domain

Linear algebra optimization

- General matrix-matrix multiplication (GEMM) in BLAS library often performs poorly for tall-and-skinny matrices, such as $\mathbf{U}_F$ and $\boldsymbol{Z}_F^T$ in the DBO model
- An optimization algorithm of the computation of GEMM with tall-and-skinny matrix is needed on the GPU platform to further speed up the DBO model.



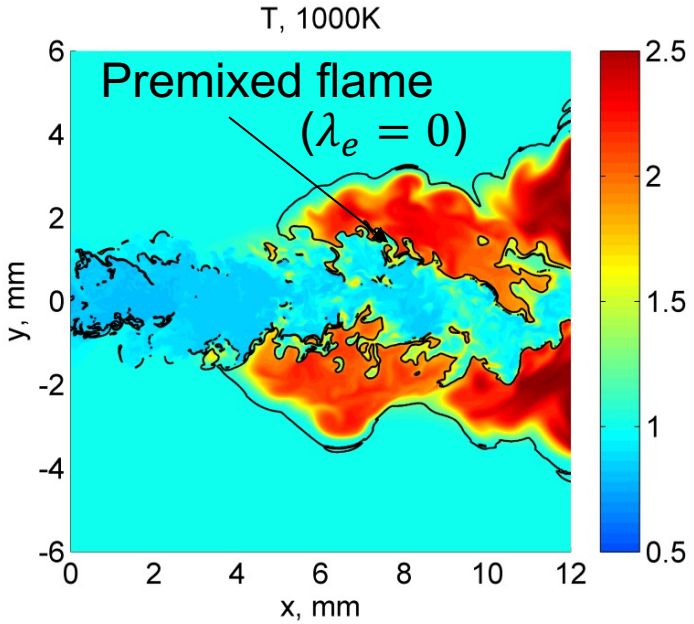
(a) Speedup (single)　(b) Mem. bw. util. (single)

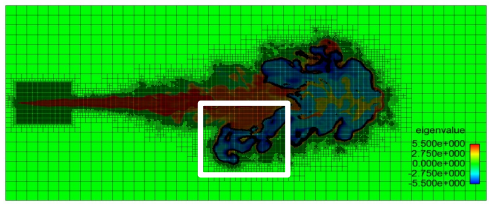
Performance of TSM2X algorithm for a skinny matrix, $10^7 \times 16 \times 16$ [2)]

1) Naderi *et al.* arXiv:2207.10656 ,　2) Rivera *et al.* JPDC 151 (2021) 70-85

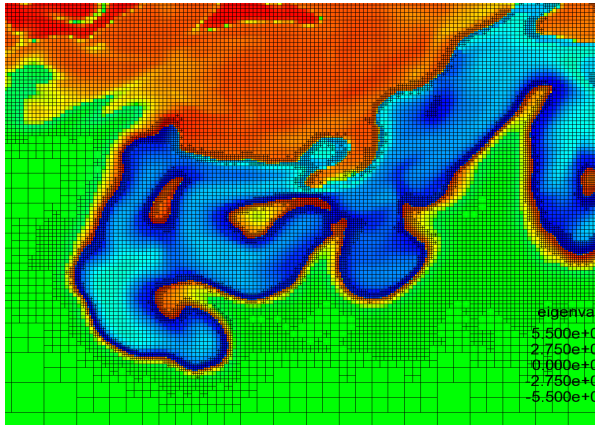
In-Situ Chemical Explosive Mode Analytics (CEMA)

- CEMA: eigenvalue solve on the reaction rate Jacobian to determine the mode of combustion





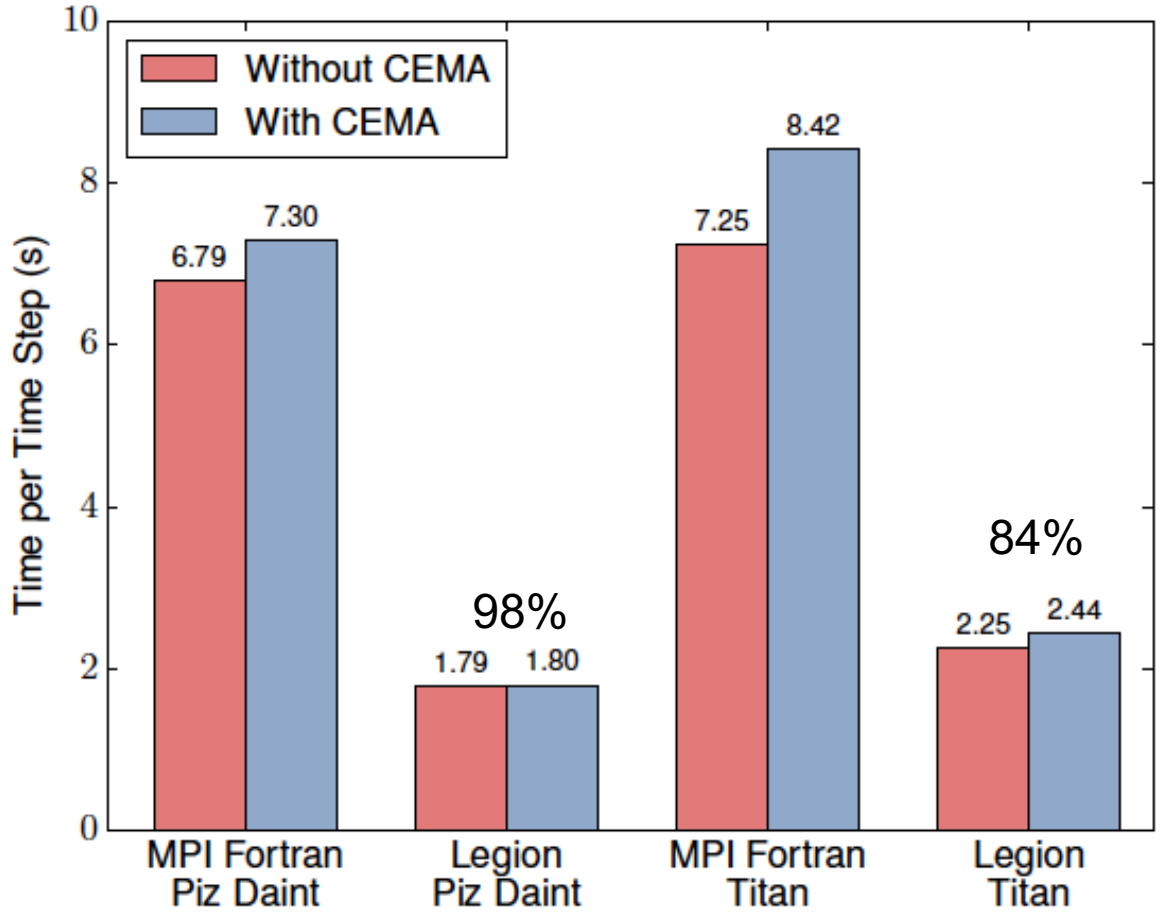
CEMA-based AMR to capture premixed fronts $(\lambda_e = 0)$



- Run CEMA at each time step as a diagnostic to steer mesh refinement
- CEMA computation takes longer than a single explicit RK stage (6 stages/timestep)
- Dividing CEMA across RK stages and interleaving with other computation so as not to impact other critical operations would be hard to schedule manually
- Asynchronous task execution, schedule CEMA on CPU resources
- Interoperate Fortran CEMA with Legion code – took a day to implement

S. Treicher et al.

Execution Overhead of In-situ Analytics (CEMA)



In situ Chemical Explosive Mode Analysis (CEMA) within the S3D Combustion Application – analysis results drive application parameters dynamically.

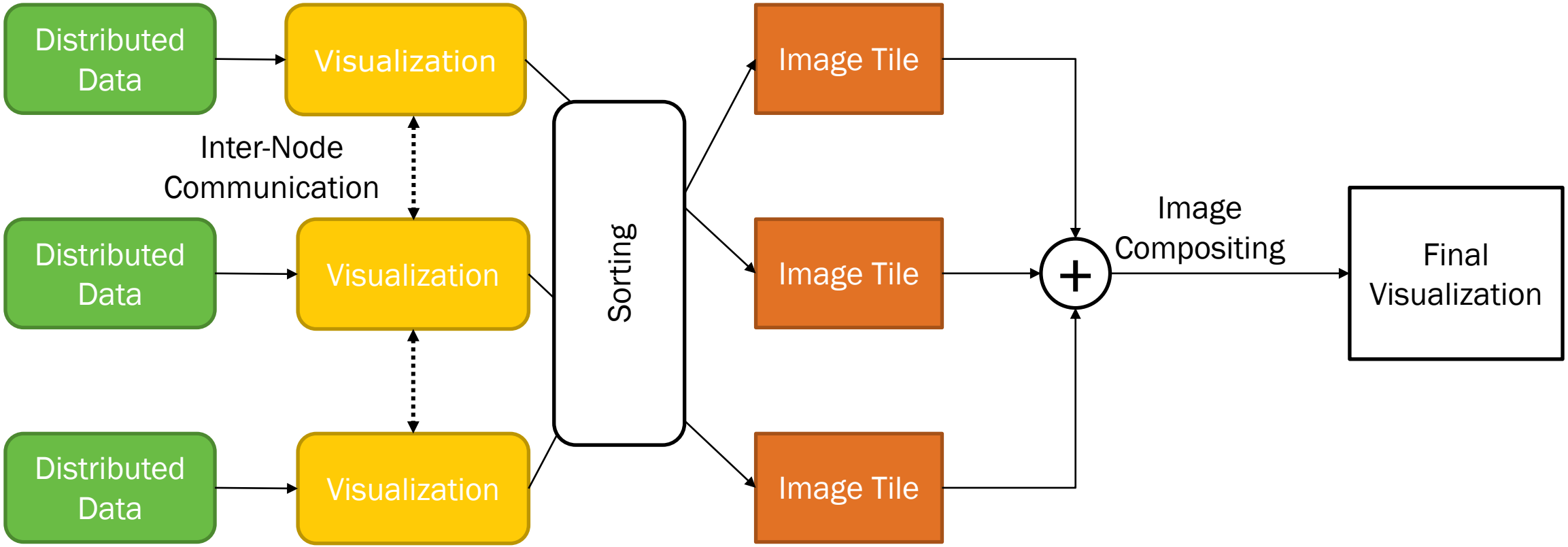
Reduced overhead of CEMA by a factor of 10 via adaptive/dynamic scheduling of analysis operations (**overhead now less than 2% of overall execution time**).

Runtime system analyzes and allows mapping of dynamic application workloads – e.g. infrequent, changing and data-dependent analysis and visualization operations.

S. Treichler et al.

Coupling Visualization Pipeline with S3D MPI only

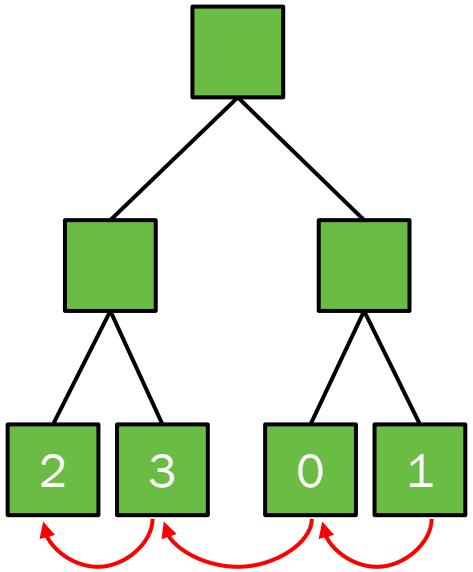
- Traditional Scientific Visualization Pipeline (Sort-Last)



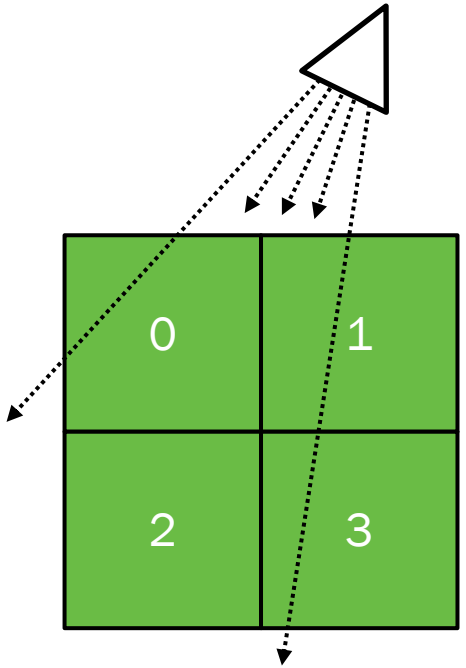
Coupling Visualization Pipeline with S3D Legion/Regent

- Sort-Last Image Compositing in Regent

Simulation Partitions



KD-Tree over Simulation Partitions



Generate the Image Compositing Partition by Traversing the KD-Tree

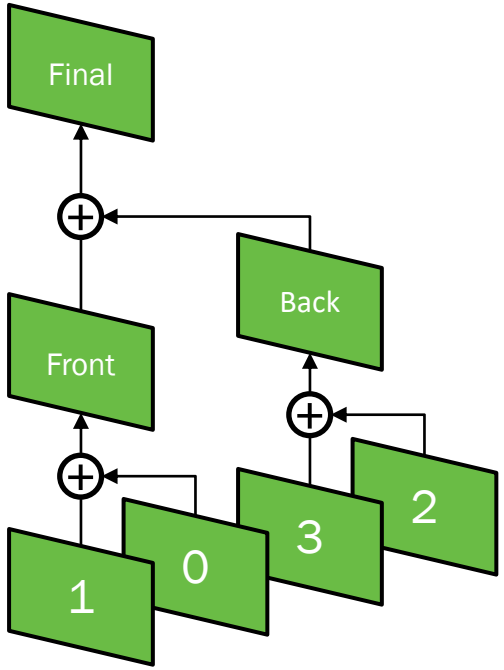
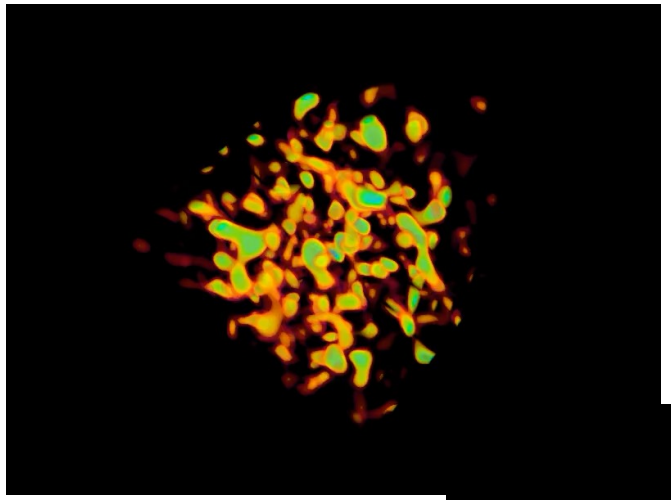


Image Compositing via Binary Reduction

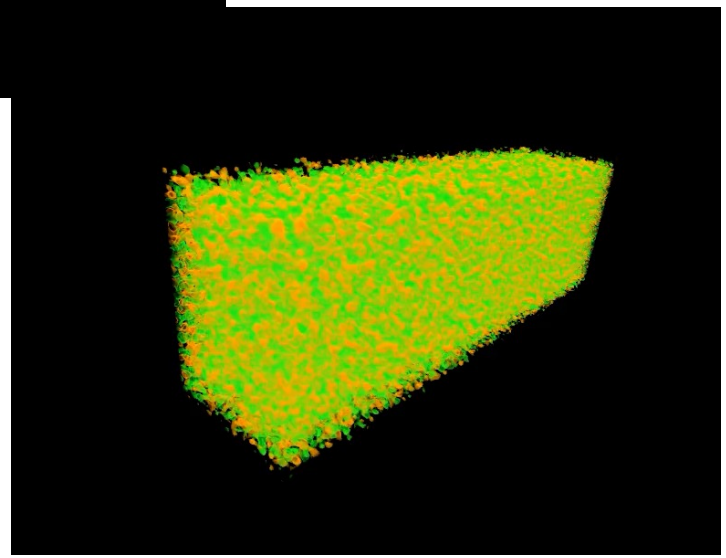
Qi Wu, Kwan-Liu Ma UC Davis, Seshu Yamajala, SLAC

Coupling Visualization Pipeline with Legion/Regent

- Integration of Volume Rendering

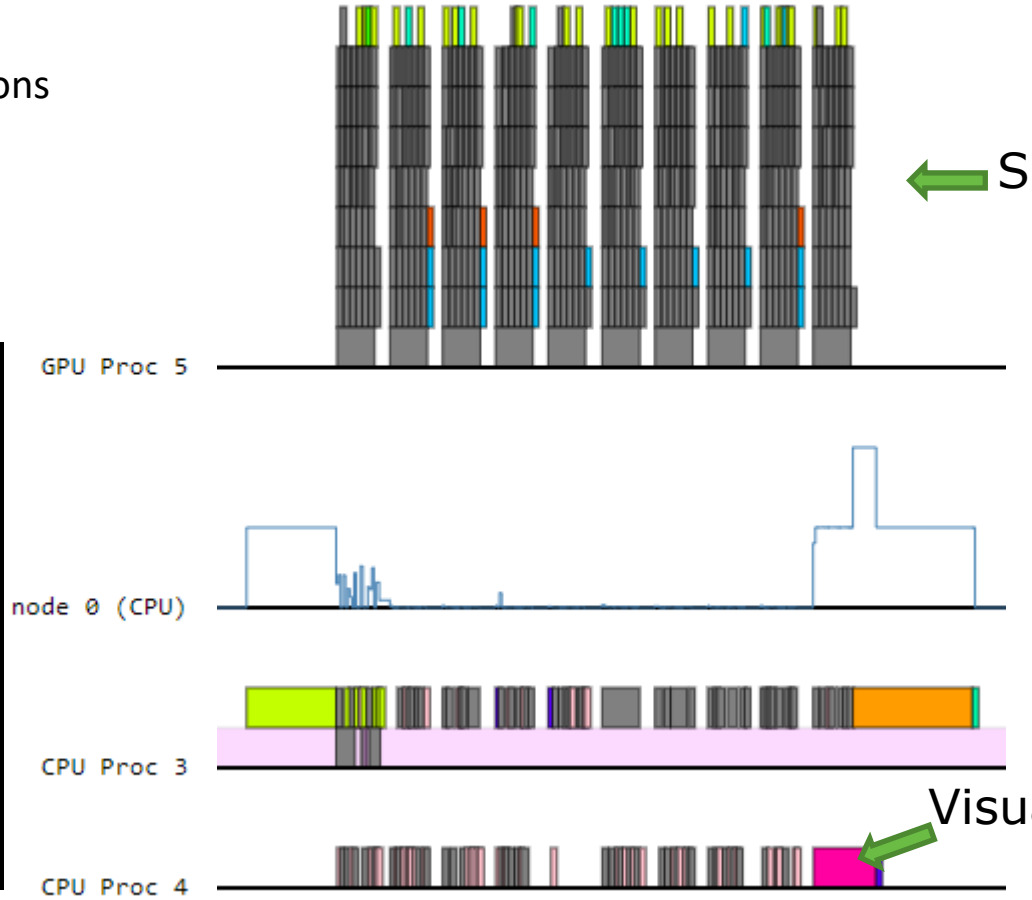


Temperature fluctuations



Spanwise velocity

Execute visualization every 10th step



Qi Wu, Kwan-Liu Ma UC Davis, Seshu Yamajala, SLAC

Summary: Task-Based Systems Enable Complex Workflows at Extreme Scale

Goal: Future workflows will combine simulation, in situ ROMS, sensitivity analysis, uncertainty quantification, analytics and visualization at extreme scale on heterogeneous machines-> need for productivity, composability, resilience, interoperability and scalability

Pros:

- Scientists write sequential code
  - Run in parallel
  - And distributed
  - And GPU
- No synchronization bugs
- Automatically asynchronous, automatic data movement
- Performance portability (mapper utility)

Cons:

- More explicit about data partitioning, tasks

  For the runtime to orchestrate everything you need to be more explicit in describing what you're doing

Questions: jhchen@sandia.gov

Acknowledgments:
DOE Basic Energy Sciences, Gas Phase Chemical Physics

DOE Office of Science Distinguished Sciences Fellows Award

ASCR Project on Collaborative In Situ Visual Analytics Technologies for Extreme Scale Combustion Simulations

ASCR Exascale Computing Project

DOE INCITE, ALCC, and NERSC ERCAP Awards

SINTEF BIGH2 and AMAZE projects