SAND2016-5706C

Towards Exascale Simulation of Turbulent Combustion

Jacqueline H. Chen

Sandia National Laboratories

jhchen@sandia.gov

2016 ISC High Performance Computing Conference

June 19-23, 2016

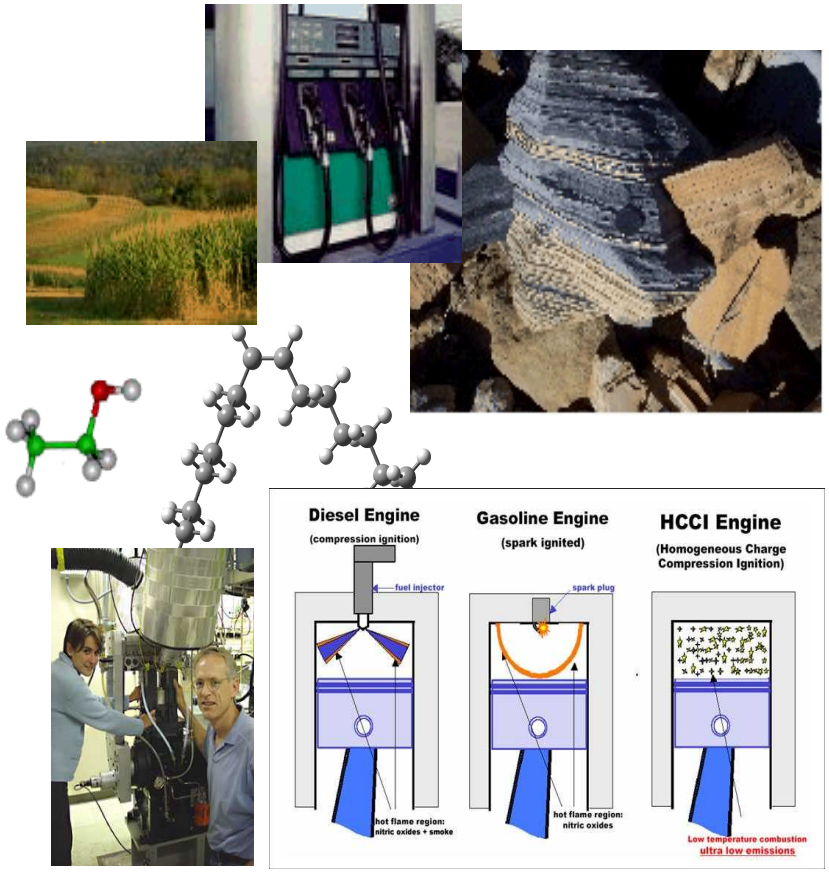
Frankfurt, Germany

Sponsored by the Division of Chemical Sciences, Geosciences and Biosciences, the Office of Basic Energy Sciences and the Office of Advanced Scientific Computing Research, the US Department of Energy (DOE)

Exascale Combustion

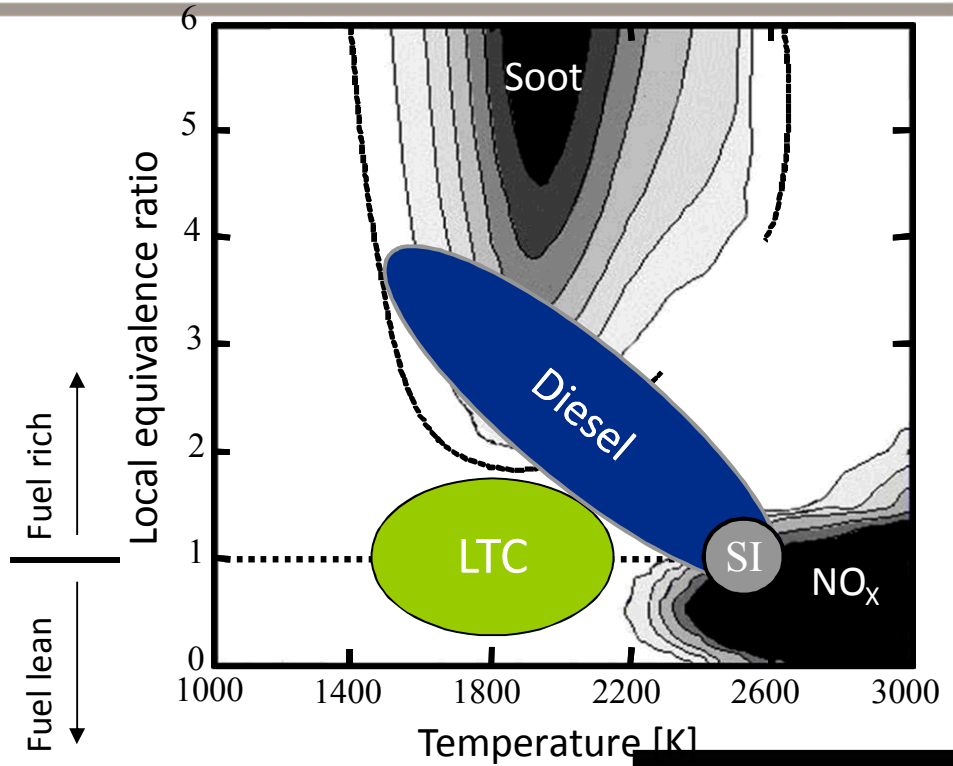
- Predict behavior of new fuels in different combustion scenarios at realistic pressure and turbulence conditions
  - Develop more efficient combustors
  - Optimize co-design of fuels and engines
  - Extend the longevity of fossil fuel and reduce CO2 and other emissions
  - Government mandates (CAFE standard of 54 mpg by 2025 and 80% reduction in GHG emissions by 2050)

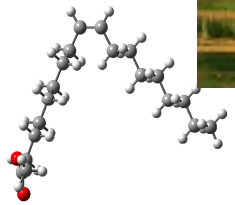
- High-fidelity direct numerical simulation methodologies
  - sufficient chemical fidelity to differentiate effects of fuels where there is strong coupling with turbulence
  - uncertainties in thermo-chemical properties and physics models (spray, soot, radiation)
  - complex flows due to compression by a piston, swirl, bluff-bodies, cavities

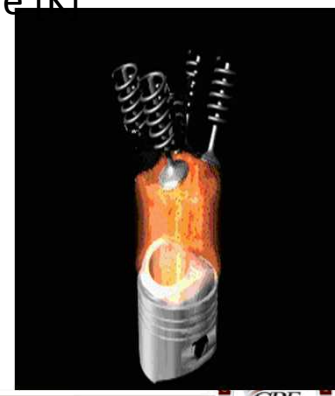


Fundamental Turbulence-Chemistry Interactions Motivated by Advanced Engines and Gas Turbines

- Higher fuel efficiency and lower emissions driving combustion towards more **dilute, fuel lean, partially-premixed conditions**
- New **mixed-mode combustion regimes**
- Strong **sensitivities to fuel chemistry**

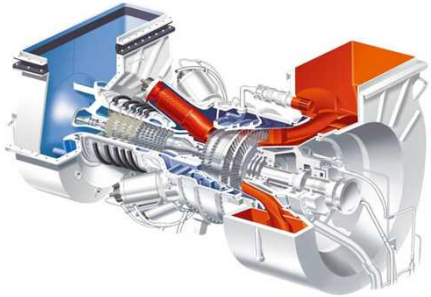


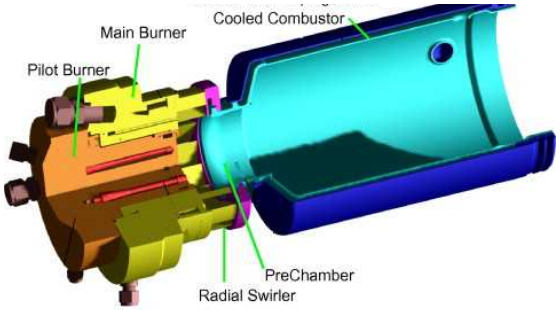


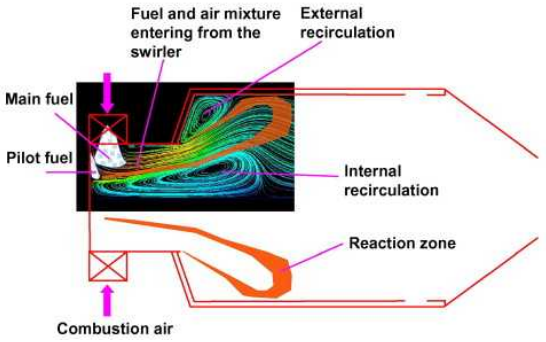


Fundamental Turbulence-Chemistry Interactions Motivated by Advanced Engines and Gas Turbines

- **Turbulent Lean Premixed (LPM) Combustion** with composition & enthalpy stratification relevant to state-of-the-art staged combustors for operational and fuel flexiblility
- LPM combustion: standard technique employed by GT manufacturers (*GE, Siemens*)
  - **reduced emissions (CO & NO) & improved efficiency**
- Combustion under intense turbulence & strain **(complex flow fields)**
  - recirculation regions induce mixing of reactants with hot products (stratification)
  - Swirling flows
- Desirable conditions but can lead to operability issues;
  - **flame blow-off**, dilution and radical depletion of the reaction zone, **flashback**, **thermoacoustics**
- Modeling challenges for LES and RANS:



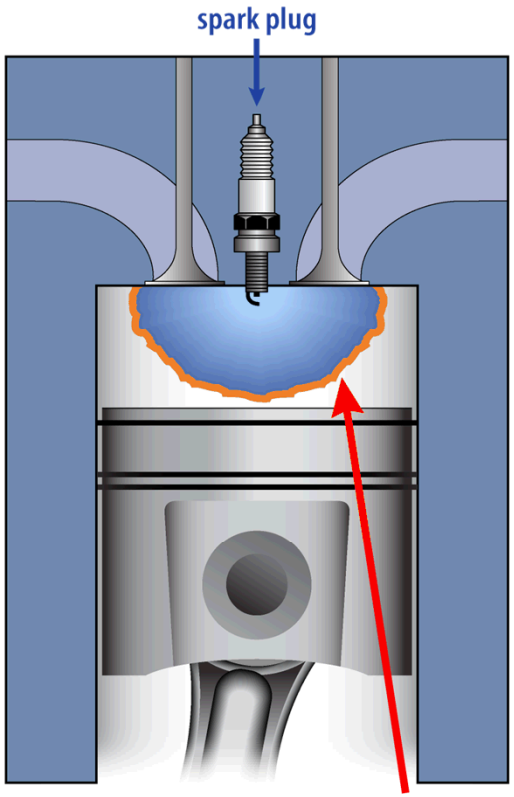




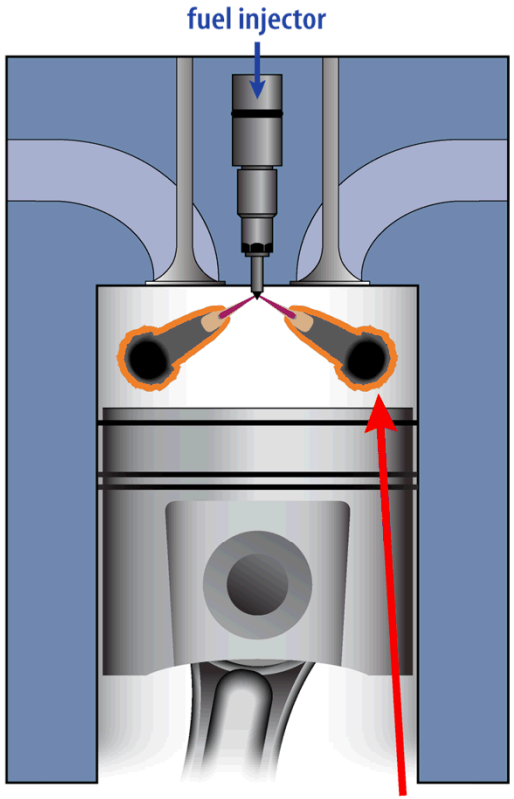
Siemens DLE combustor, Liu & Sanderson 2013
Courtesy of Siemens Industrial Turbomachinery Ltd.

Comparison of Engines

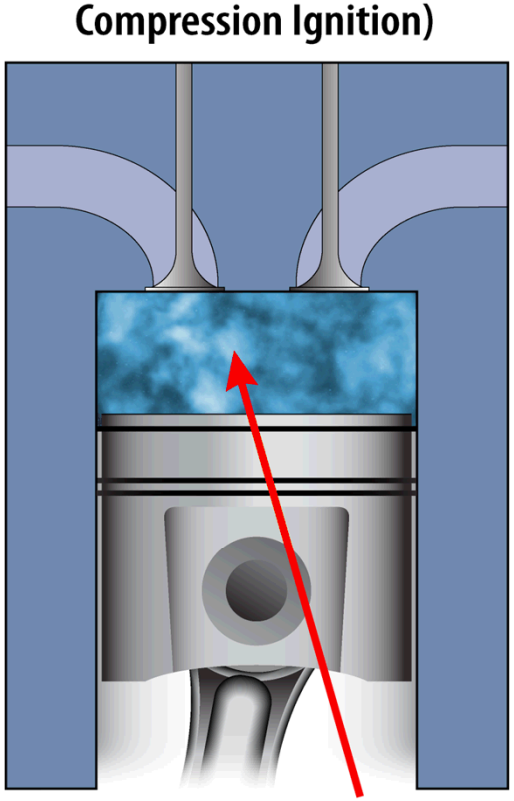
Gasoline Engine
(Spark Ignition)



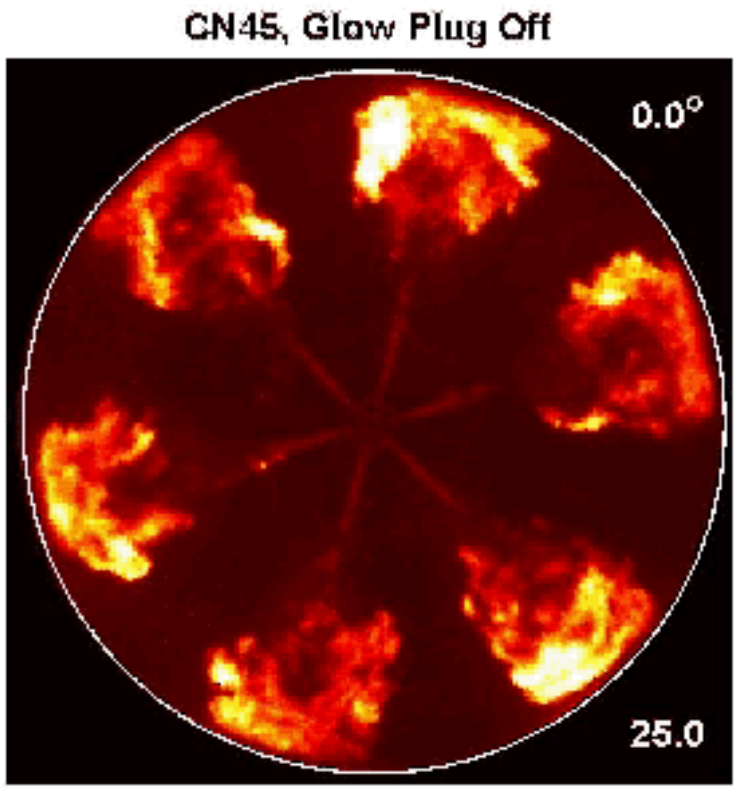
Diesel Engine
(Compression Ignition)



HCCI Engine
(Homogeneous Charge Compression Ignition)



IC Engine Combustion is Multi-physics Multi-scale



Diesel Engine Autoignition, Soot Incandescence
Chuck Mueller, Sandia National Laboratories

Large range of length and time scales

- In-cylinder geometry (cm)
- Turbulence-chemistry (microns-cm)
- Soot inception (nanometer)

Chemical complexity

- large number of species and reactions (100's of species, thousands of reactions)

Multi-Physics complexity

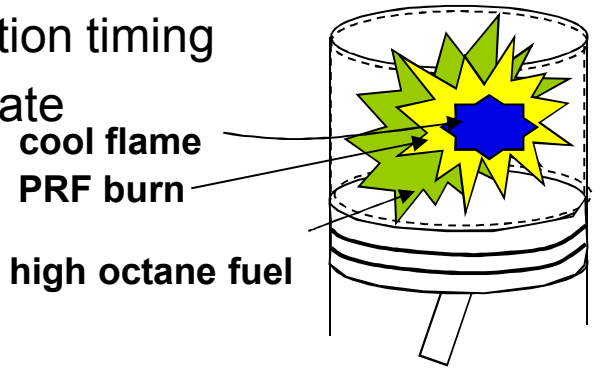
- multiphase (sprays, gas phase, soot)
- thermal radiation

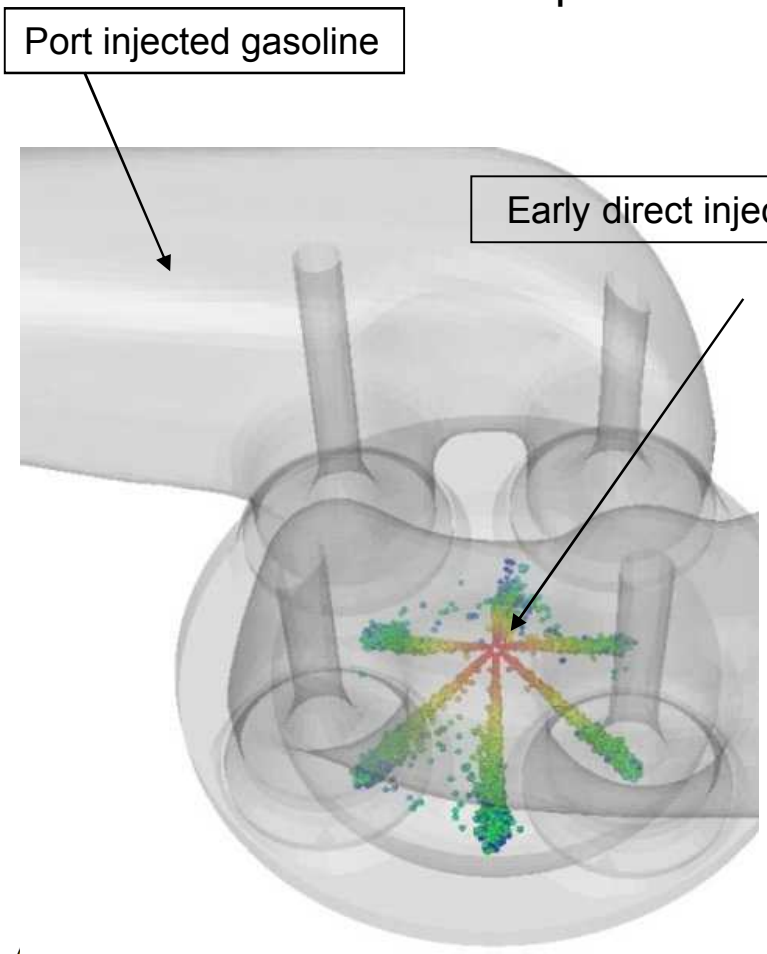
All these are tightly coupled

Exascale Target: Dual Fuel RCCI combustion

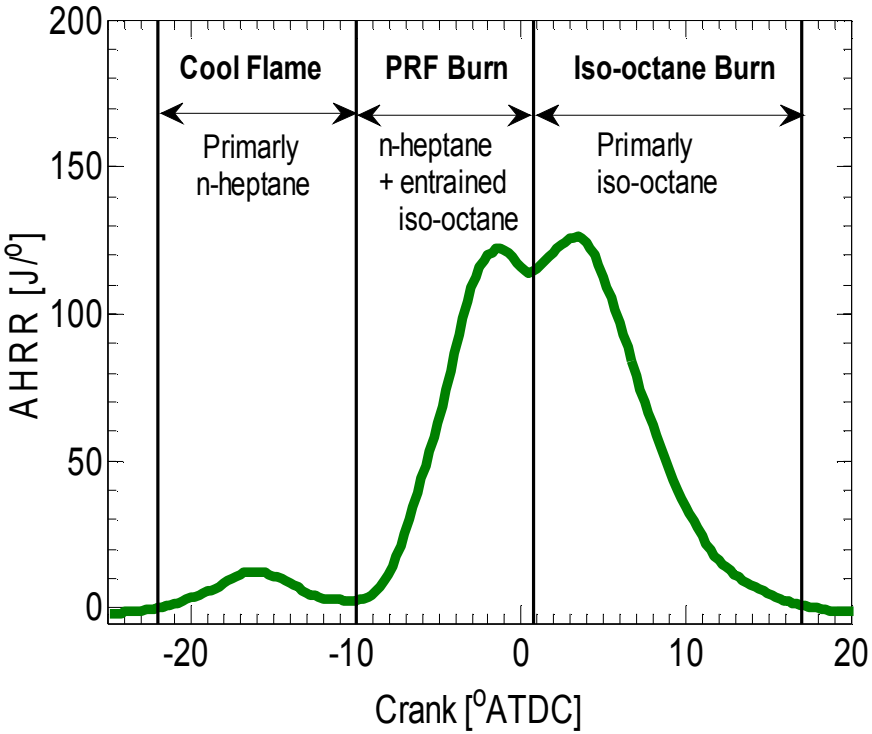
- Reactivity Controlled Compression Ignition
- Optimized in-cylinder fuel blending of high cetane diesel with high octane gasoline: control phasing (ignition timing relative to piston motion) and combustion rate

RCCI



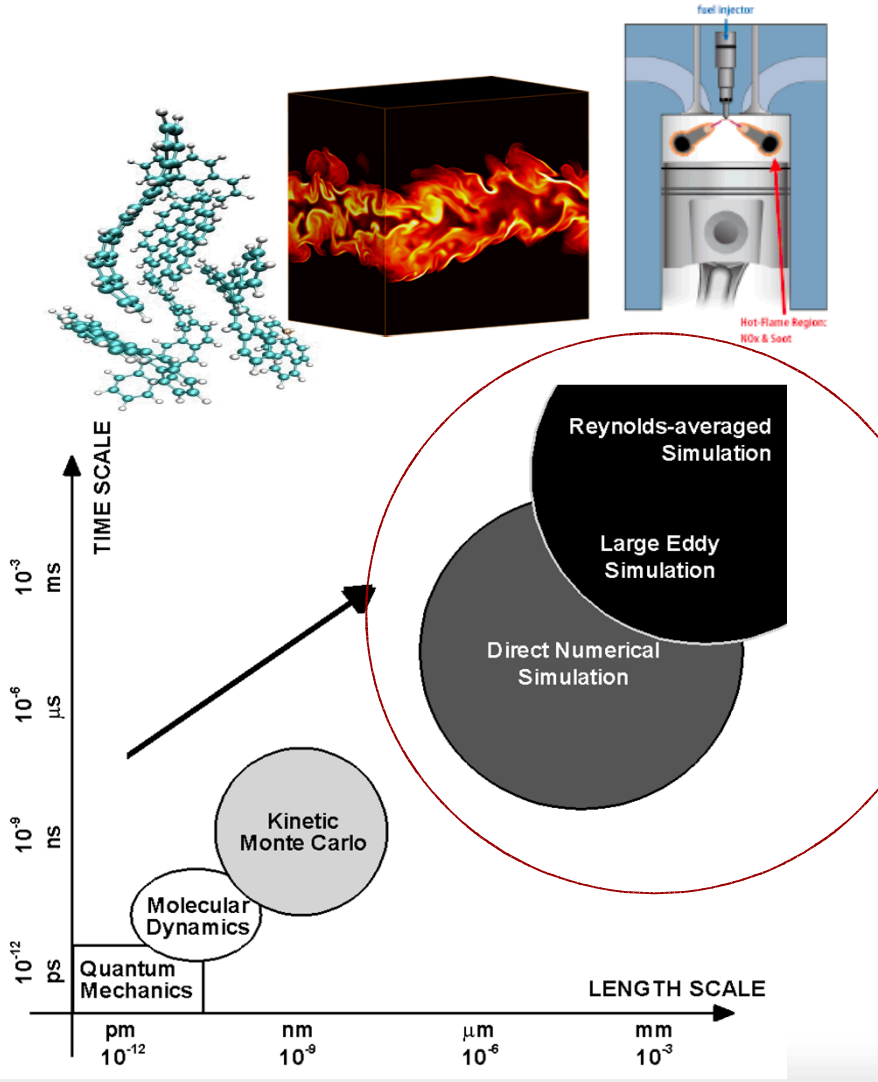


Control of combustion duration by ratio of fuels



Multi-scale Modeling of Combustion Processes

- Multi-scale modeling describes combustion processes, from quantum scales up to device-level, continuum scales
- Multi-scale Strategy:
  - Use exascale computing power to perform direct simulation at the atomistic and fine-continuum scales (~4 decades in scales)
  - Develop new parameterizations that will enable bootstrapping information upscale



Continuum Approaches: DNS – LES – RANS

- **Reynolds Averaged Navier-Stokes (RANS)**
  - No attempt to resolve any scales
  - Inexpensive, not very accurate

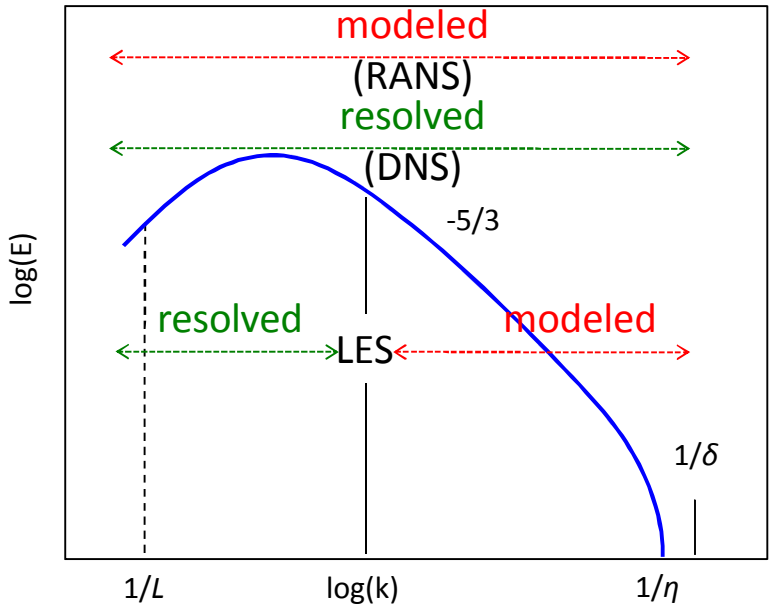
Large Eddy Simulation (LES)

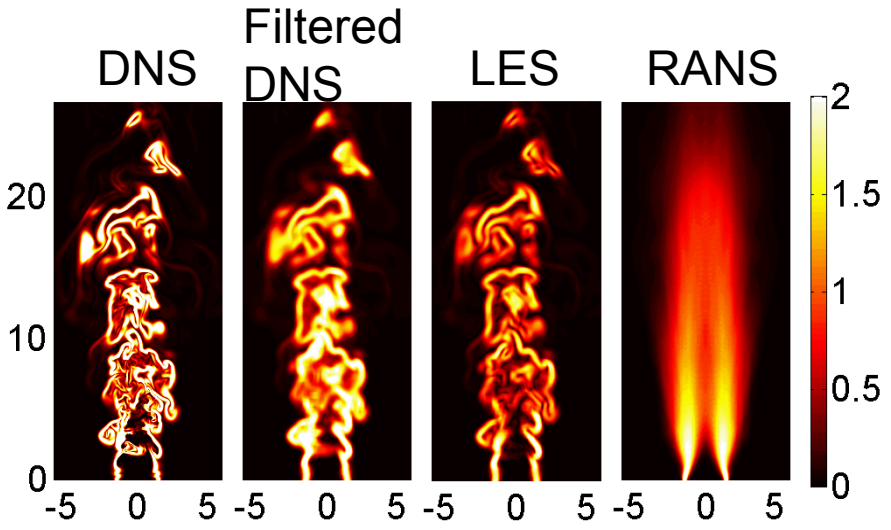
- Energy containing scales are resolved
- Model subgrid physics

Direct Numerical Simulation (DNS)

- All continuum scales are resolved
- Most expensive

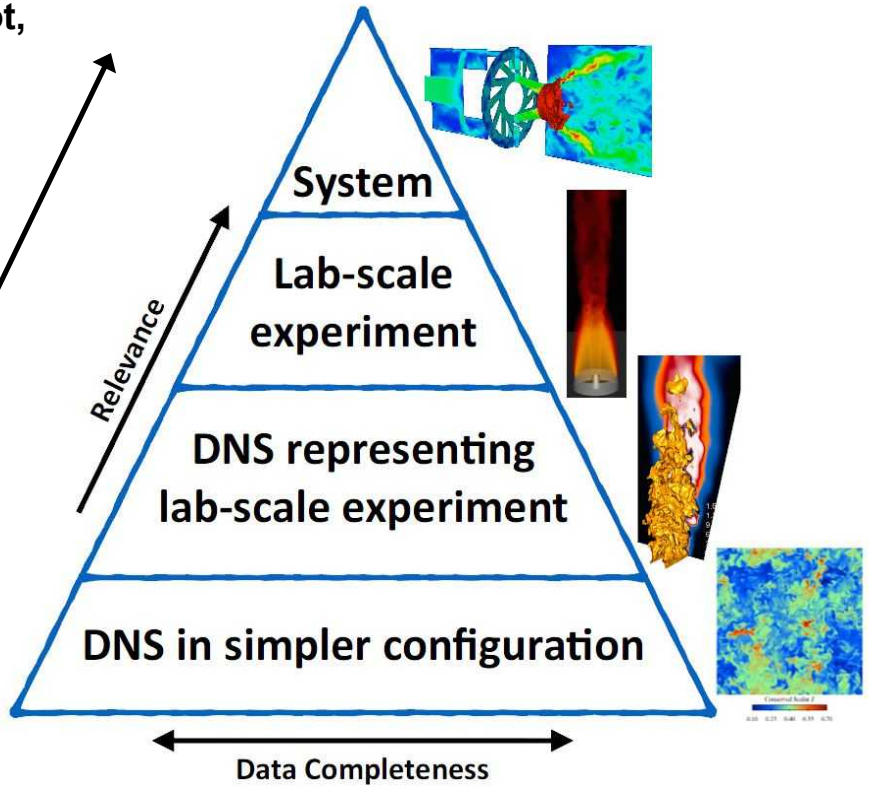
Accuracy
Expense





Multi-Scale Validation Approach

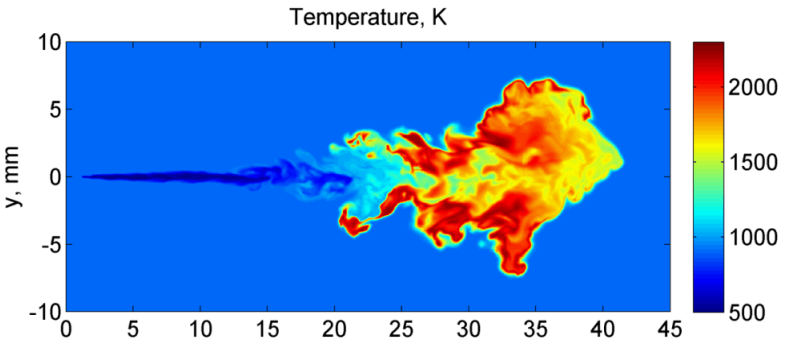
Exascale – high pressure, high Re, complex chemistry O(100 species), complex geometry, spray, soot, radiation



Petascale – moderate pressure, moderate Re, complex chemistry O(40 species)

Multi-Regime Combustion

- Technology is pushing combustion into regimes with strong turbulence-chemistry interactions (high Karlovitz number, low Damköhler number)
- Applications operate at high pressure and under preheated conditions (750-1100 K) where hydrocarbon fuels exhibit two-stage ignition
- Staged combustors, product recirculation, cavity driven flows, pilots provide stratification of heat and species



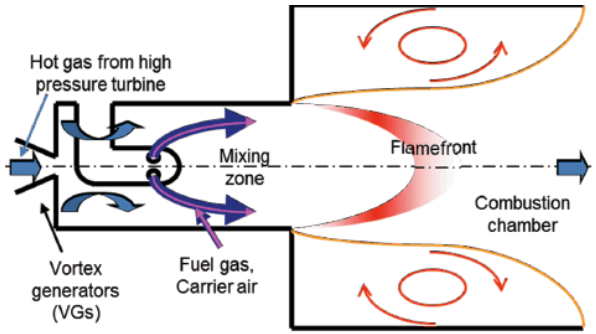
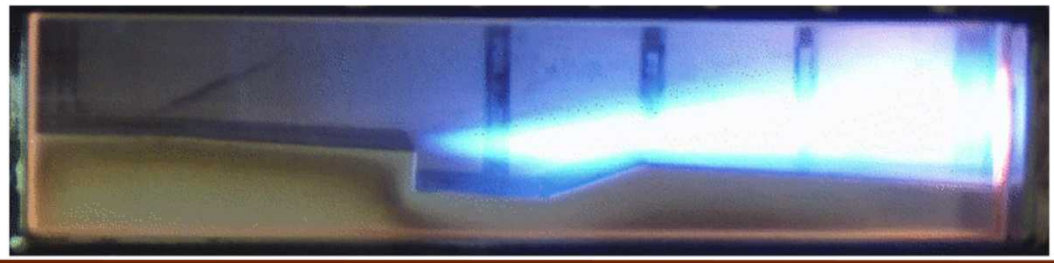
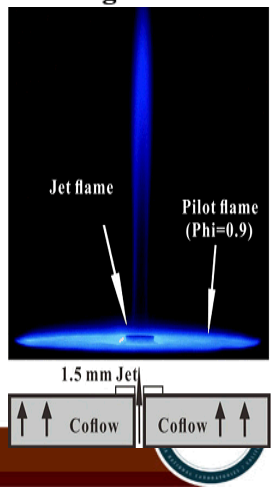


Figure 1: SEV "reheat" combustor configuration

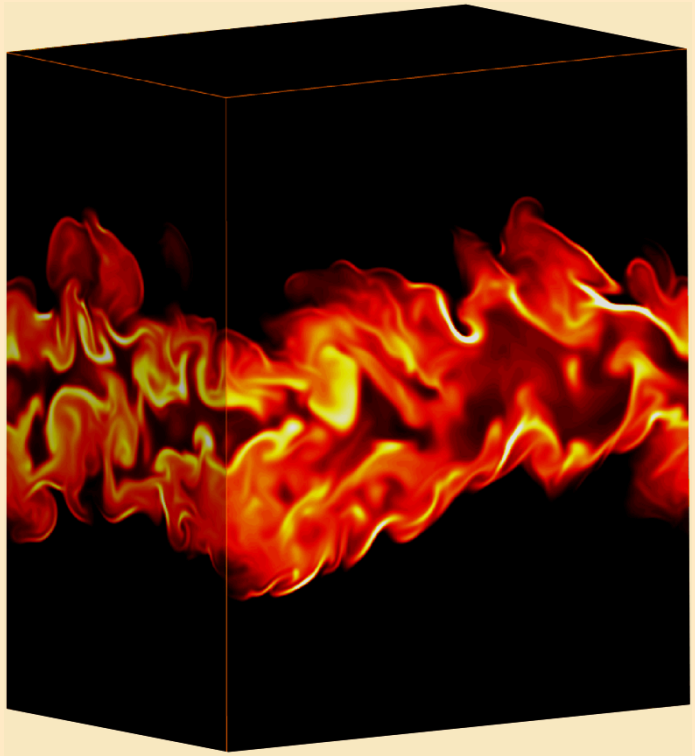




Direct Numerical Simulation Code – S3D

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

- **Used to perform first-principles-based DNS of reacting flows**
- **Solves compressible reacting Navier-Stokes equations**
- **High-fidelity numerical methods**
- **Detailed reaction kinetics and molecular transport models**
- **Ported to all major petascale platforms (heterogeneous cpu-gpu)**
- **Particle and flame element tracking**
- **In situ analytics and visualization**
- **Refactored for multi-threaded, many core heterogeneous architectures**



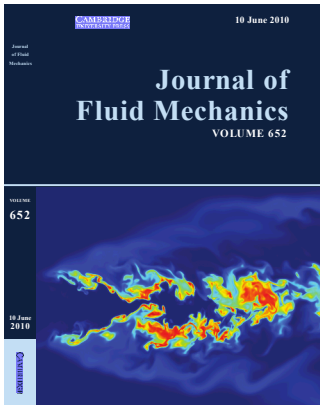
DNS provides unique fundamental insight into the chemistry-turbulence interaction

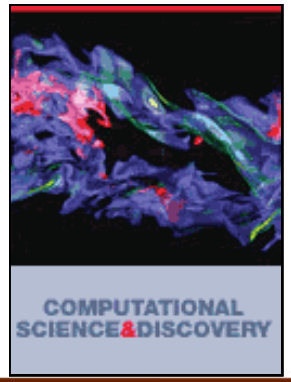
Petascale High Performance Computing on DOE Leadership Class Machines

Cray XK7, ORNL 27 Pflop

Blue Gene/Q, Argonne, 10 Pflop

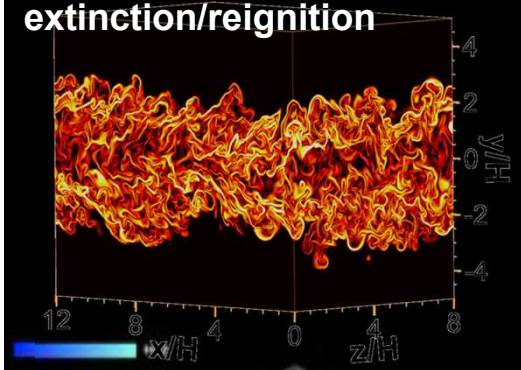
- Petascale computing for scientific discovery
- DOE INCITE Awards – large computing allocations

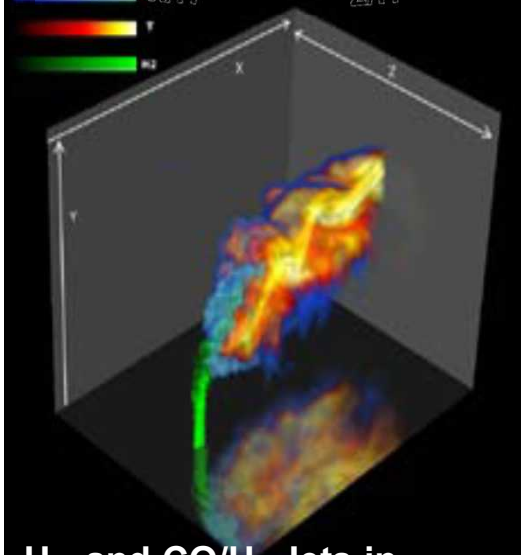




DNS Benchmarks for Model Development

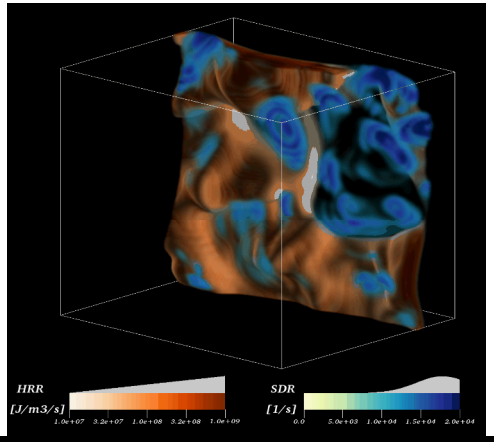
CO/H_2 C_2H_4 and DME jet flames extinction/reignition

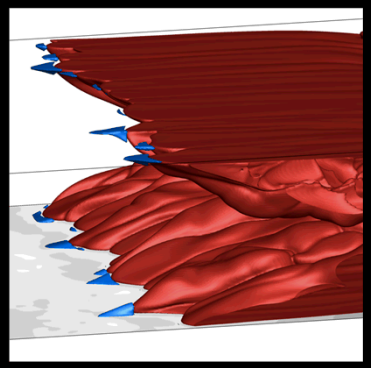




H_2 and CO/H_2 Jets in Crossflow

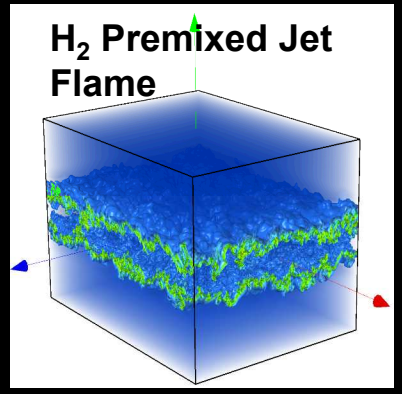
Turbulent Counterflow H_2/air Stratified Flames

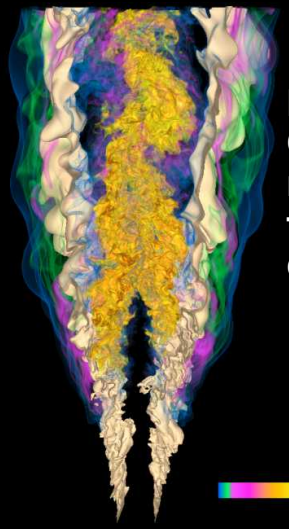




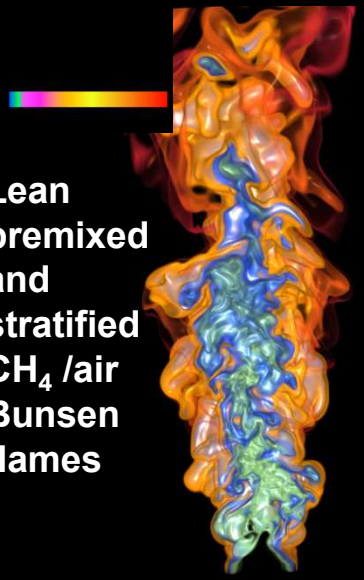
H_2/air Boundary Flame-Layer Interaction

H_2 Premixed Jet Flame





Lifted H2 and C_2H_4 , DME, and n-dodecane jet flames in hot coflow



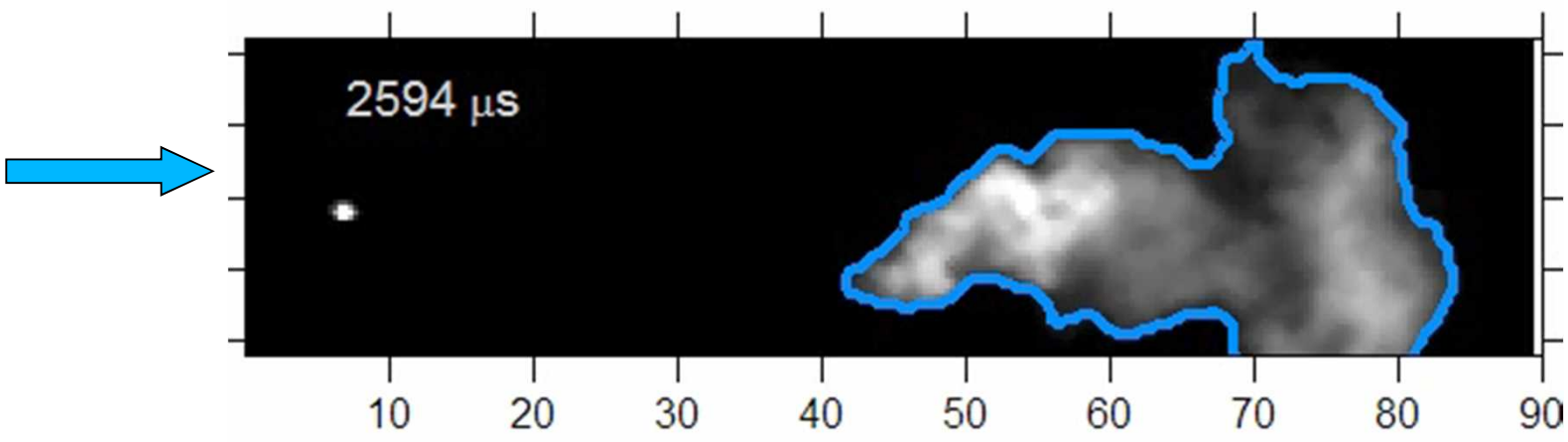
Lean premixed and stratified CH_4 /air Bunsen flames

Petascale DNS of Turbulence-Chemistry Interactions Relevant to Compression Ignition

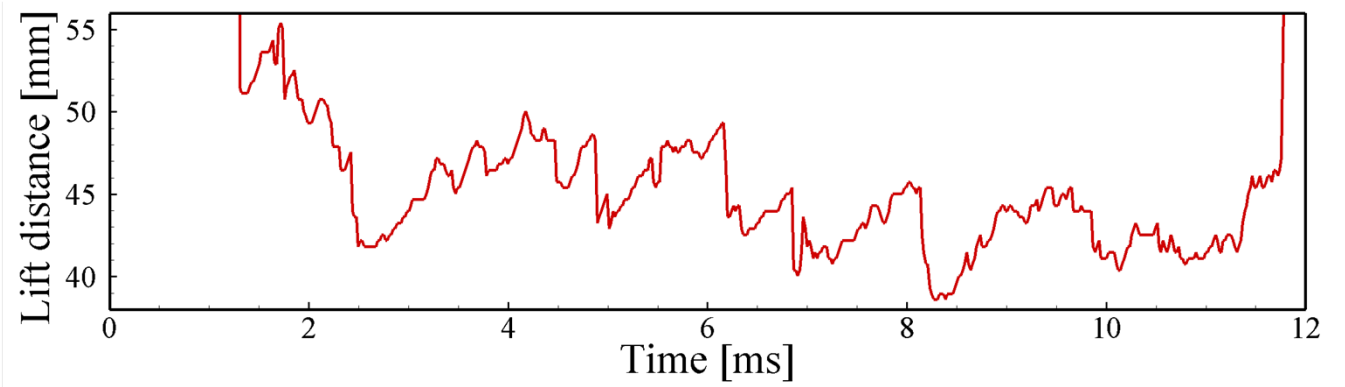
- Lifted Flame Stabilization in Laminar and Turbulent Jet Flames with Multi-Stage Ignition
- Low-Temperature Autoignition in a Diesel Jet
- Propagation mode in Reactivity Controlled Compression Ignition (RCCI) with PRF gasoline blend of iso-octane and n-heptane

Motivation: Understanding Stabilization of Lifted Flames in Heated Coflow

What is the role of ignition in lifted flame stabilization?



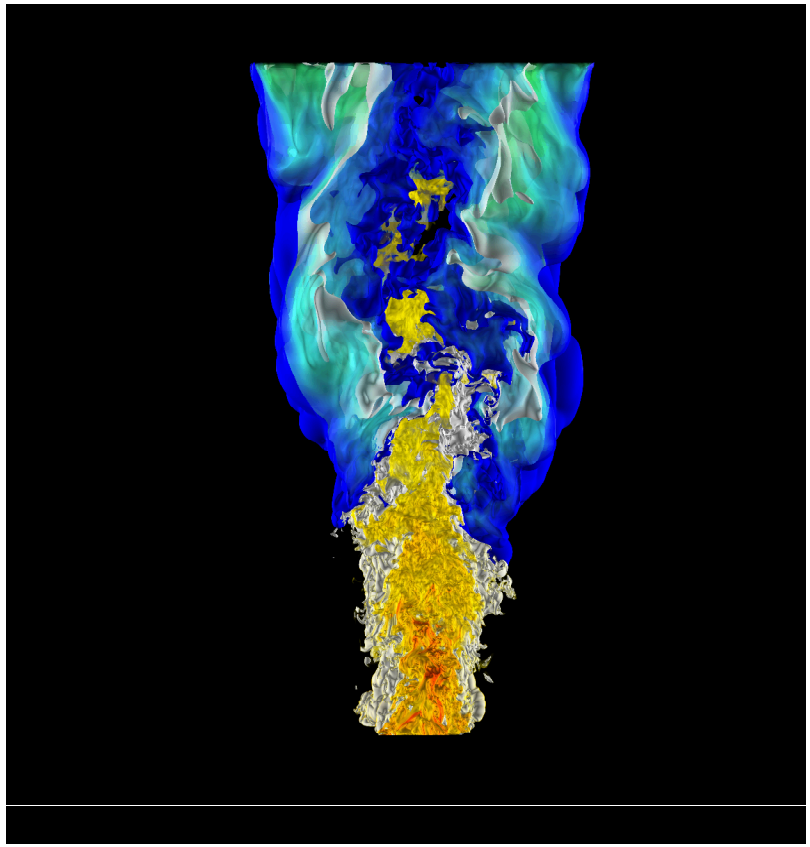
Chemiluminescence from diesel lift-off stabilization for #2 diesel, ambient 21% O_2, 850K, 35 bar courtesy of Lyle Pickett, SNL



Lifted DME Jet Flame in Heated Coflow at 5 atm

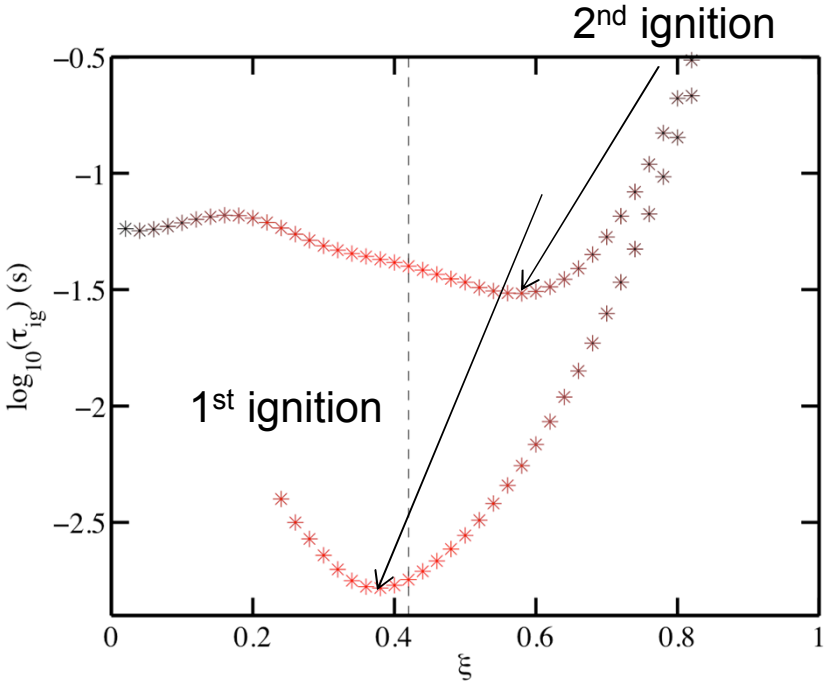
(Minamoto and Chen, 2016 *Combustion and Flame*)

- 11,700 jet Reynolds number
- Turbulent Reynolds number of 1430
- 5 atm (NTC and low temperature heat release, LTHR)
- DME reduced chemical model with 30 species (Bhagatwala et al. 2014) based on Zhao&Dryer



	$\tilde{u}$ (m/s)	u' (m/s)	l_E (mm)	λ (μm)	η (μm)	Re_{l_E}	Re_λ	Da	Ka
$y = 0$	104	27.4	0.85	46.4	3.6	1430	79	0.81	48
$\tilde{\xi} = \xi_{st}$	19.4	19.8	0.64	48.3	4.7	715	54	0.85	34

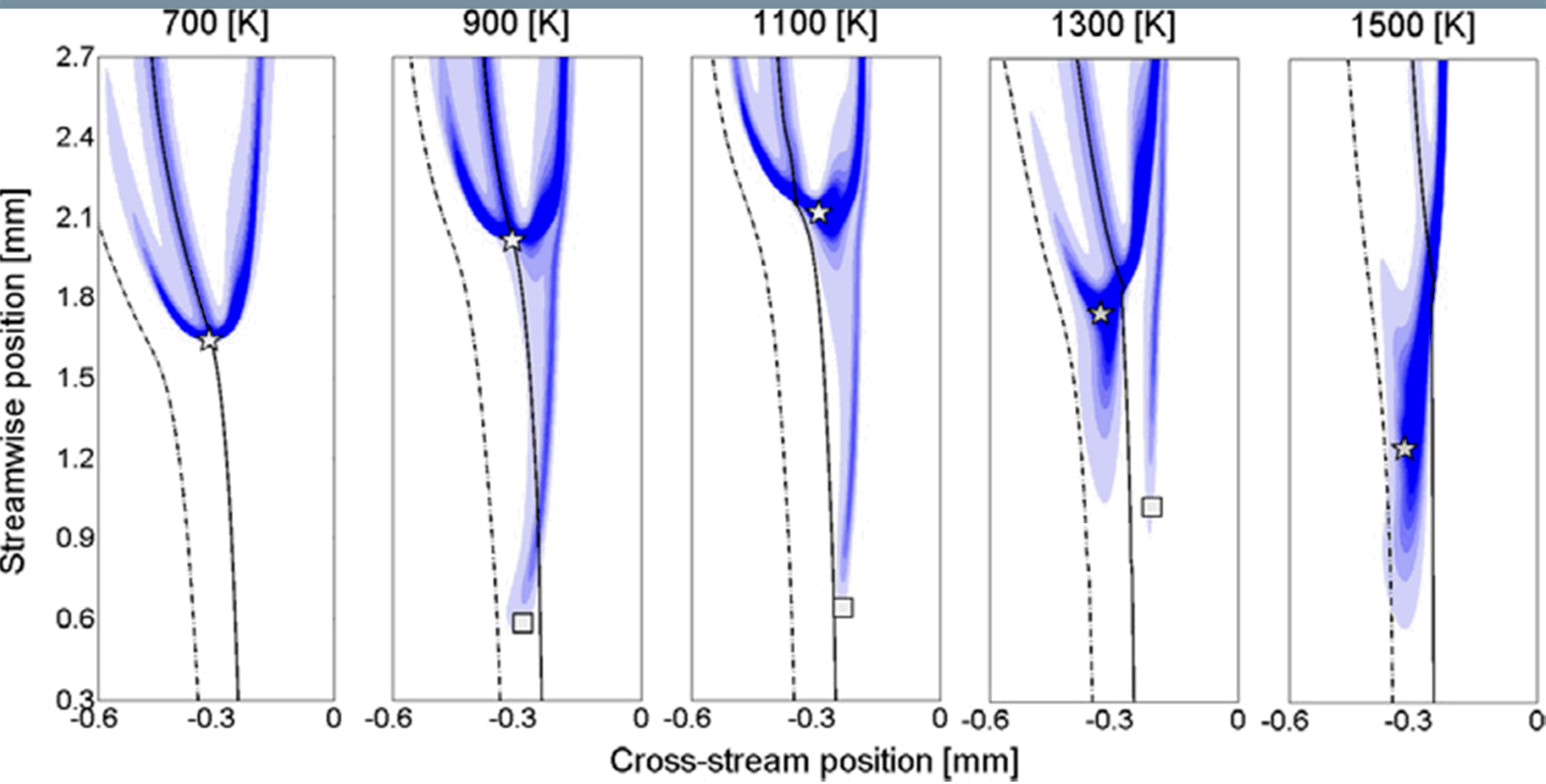
Negative Temperature Coefficient (NTC) & Two-stage Ignition in Dimethyl Ether (DME) at 5 atm



Fuel stream: 0.1 DME+0.9 N_2 (500 K)

Oxidizer stream: 0.21 O_2 + 0.79 N_2 (1000 K)

DNS of a Laminar DME Jet Flame at 40 atm - Polybrachial Structure (Krisman et al. 2015)



Heat release rate, * denotes stabilization point, square denotes Low temperature ignition, black line is stoichiometric condition

Laminar Lifted DME Jet Flame at 40 atm

Krisman et al. 2015

"Pentabrachial flame structure"

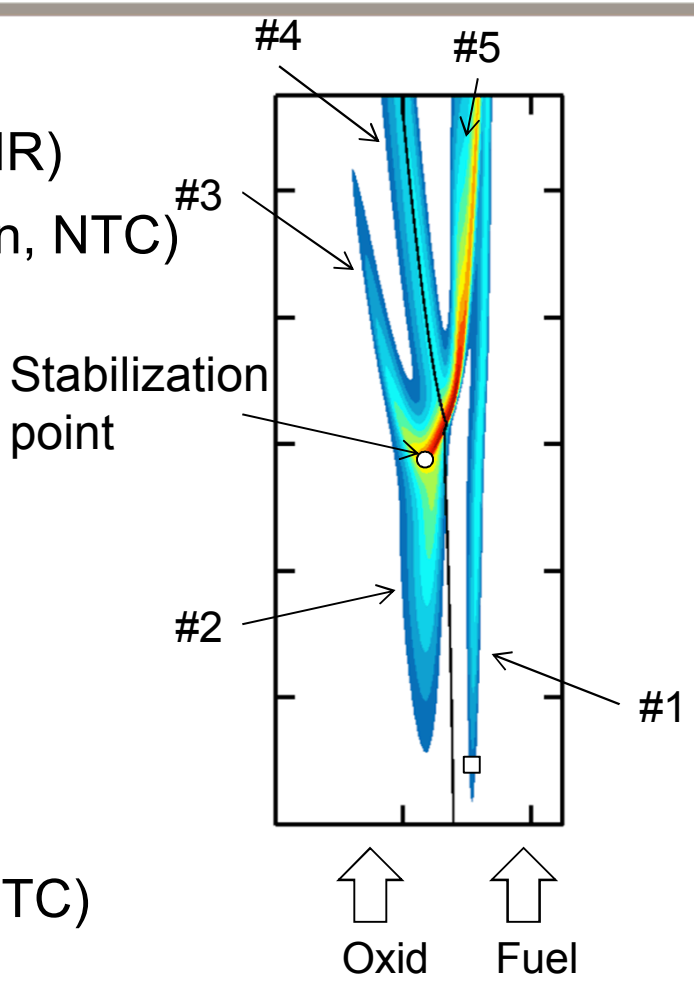
#1. Low-temperature reaction (LTHR)

#2. High-temperature reaction (lean, NTC)

#3. Lean premixed flame

#4. Diffusion flame

#5. Rich premixed flame



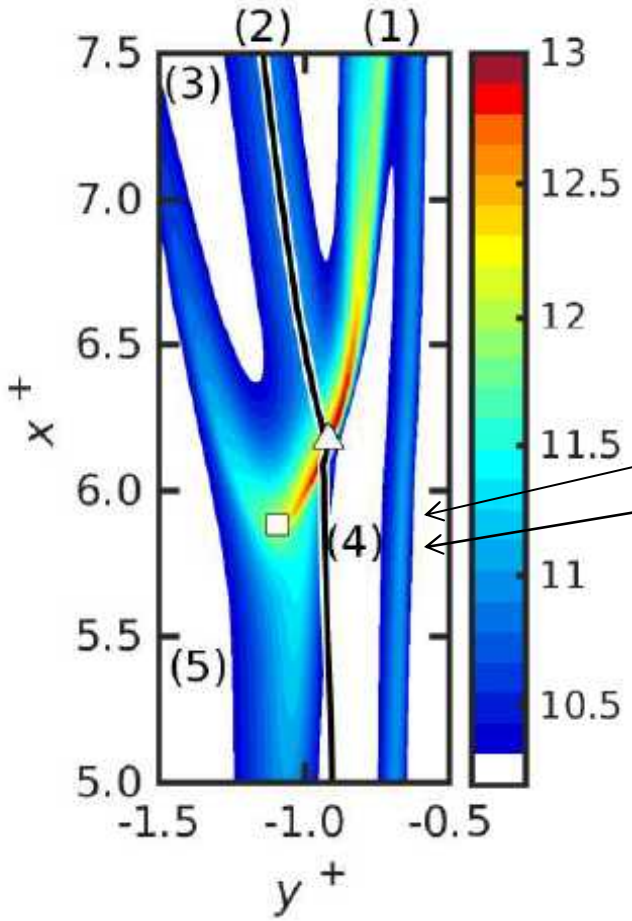
Objectives: What are the characteristics of a lifted jet flame in the presence of:

- Sheared Turbulence
- Mean velocity gradient
- Negative Temperature Coefficient Regime (NTC)
- Low temperature heat release (LTHR)

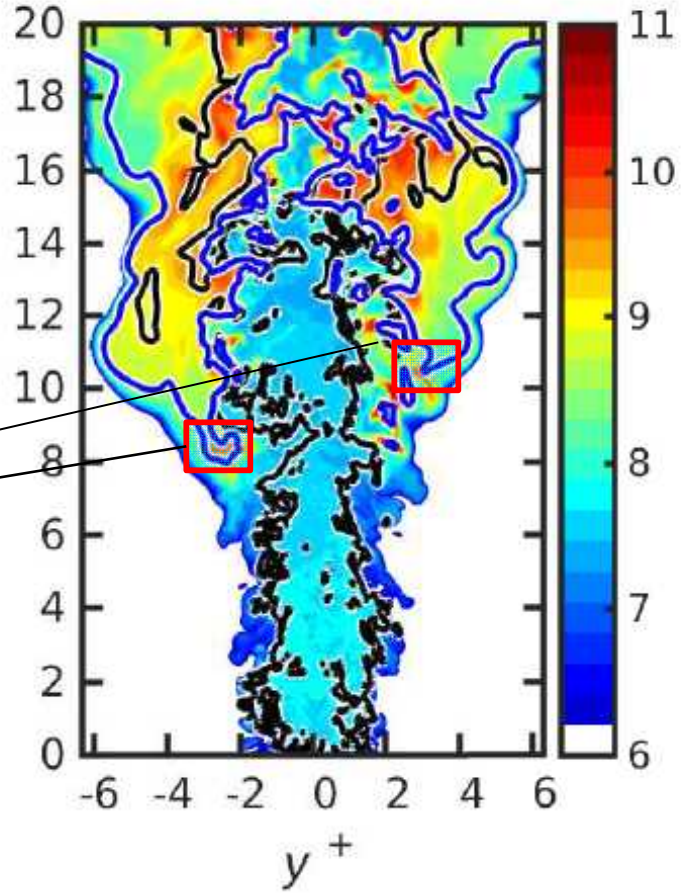
Fuel stream: 0.3 DME + 0.7 N_2 (400 K)

Oxid stream: 0.21 O_2 + 0.79 N_2 (1300 K)

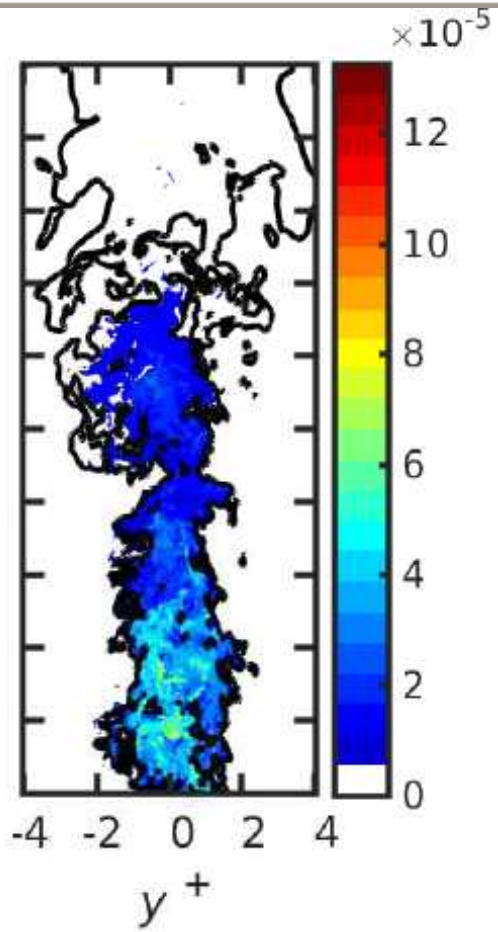
Laminar and Turbulent DME Flame Structure



Laminar pentabrachial flame, Log (heat release rate)

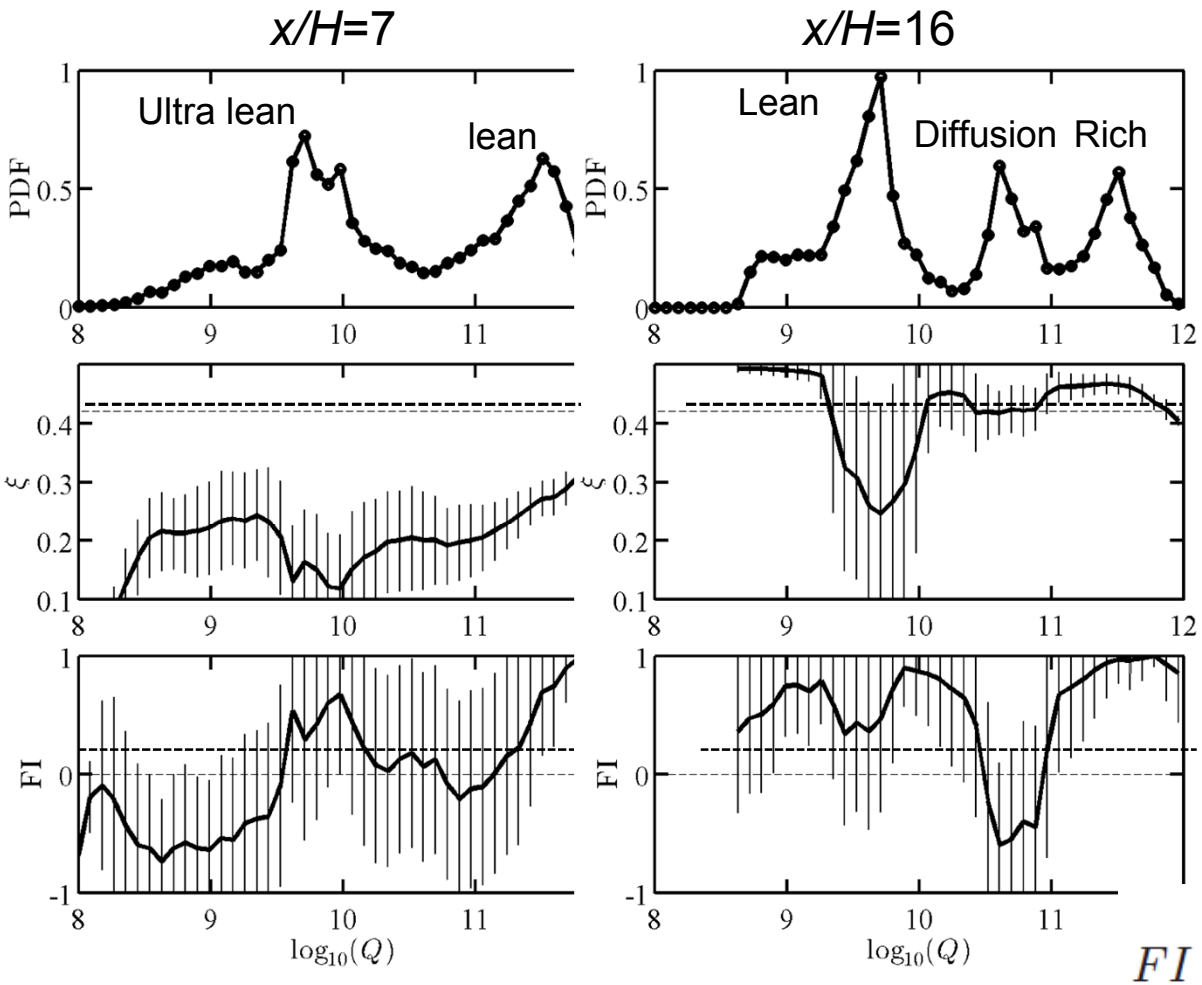


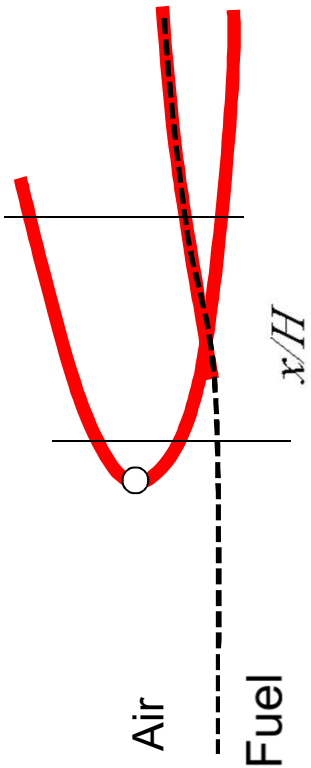
Log of heat release

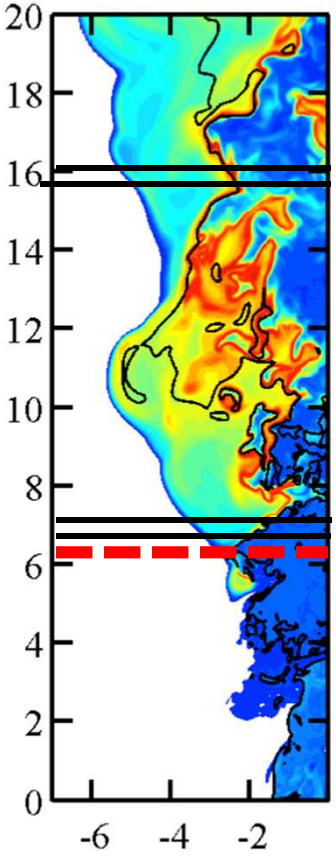


CH3OCH2O2

Downstream Flame Branches in Turbulent Flame





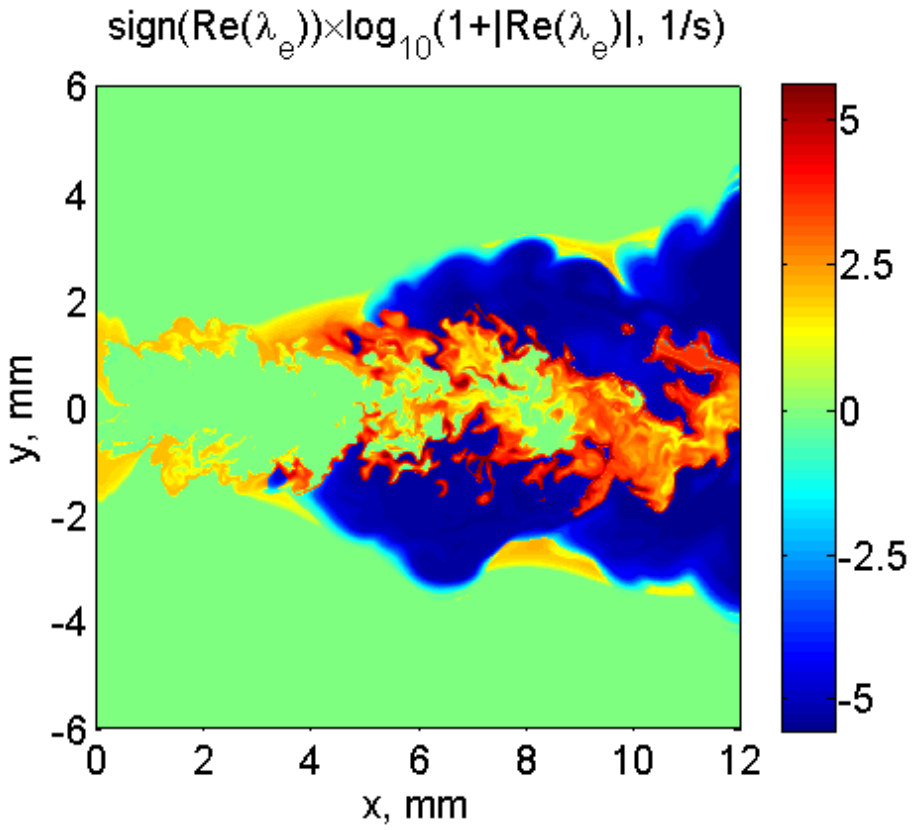


$$FI = \frac{|\nabla Y_F| \cdot |\nabla Y_O|}{|\nabla Y_F||\nabla Y_O|}$$

Premixed +

Nonpremixed -

Structure of the Lifted DME Jet Flame Visualized by CEMA



- Important flame features involved
  - A non-premixed flame kernel
  - Lean premixed flamelets
  - Rich premixed flame fronts in the broken reaction zones regime (can be important for soot modeling)
  - A mixing layer with fresh mixtures (auto-igniting)
  - Pockets of cool flame

Positive eigenvalue, λ_{exp}, of Jacobian J_ω indicates the chemical explosive mode

$$\mathbf{J}_\omega = \frac{d\boldsymbol{\omega}}{d\mathbf{y}}$$

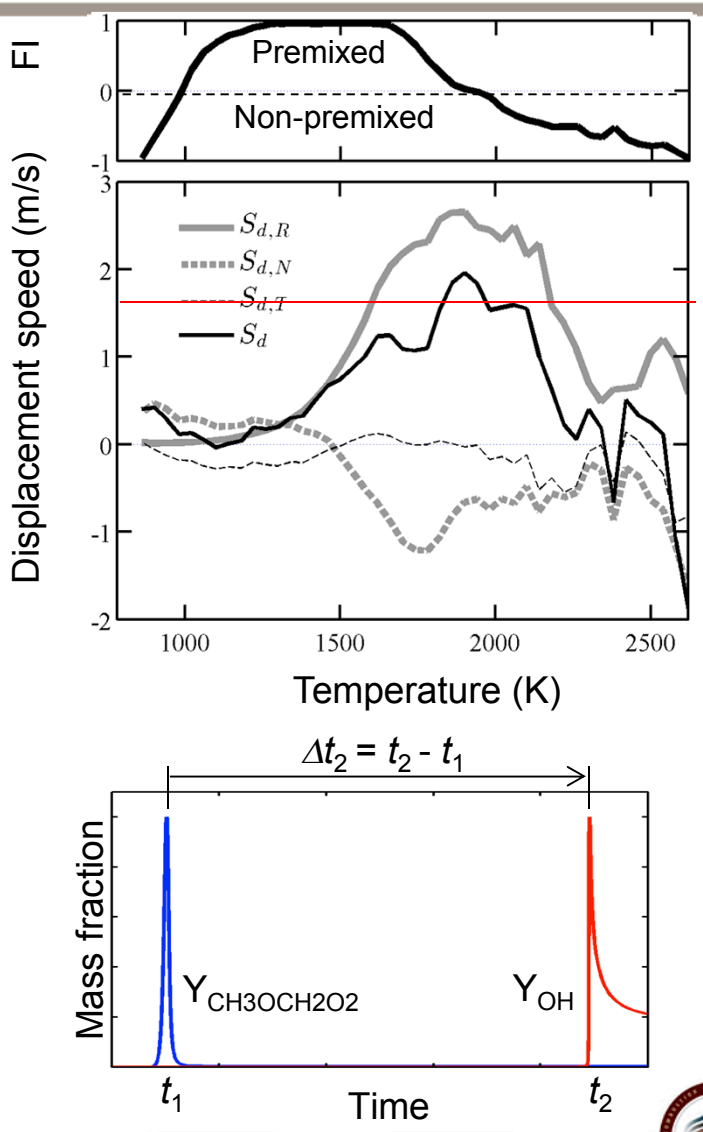
Flame Displacement Speed Along ξ_{st}

$$S_d = \frac{Q}{\rho_0 c_p |\nabla T|} \Big\} S_{d,R} + \underbrace{\frac{\mathbf{n}\cdot\nabla(\lambda\,\mathbf{n}\cdot\nabla T)}{\rho_0 c_p |\nabla T|}}_{S_{d,N}} + \underbrace{\frac{\lambda}{\rho_0 c_p}\nabla\cdot\mathbf{n}}_{S_{d,T}} - \underbrace{\frac{\rho}{\rho_0 c_p}\nabla T\cdot\left(\sum_1^N c_{p,k} D_k \frac{W_k}{W}\nabla X_k\right)}_{S_{d,cp}\to 0}$$

Ruetsch et al. 1995

$S_{ref} = \sqrt{\frac{\rho_0}{\rho_1}} S_L$ (ρ_0, ρ_1: unburnt and burnt densities on ξ_{st})

$\xi = \xi_{st}$	Flame speed (m/s)	S_{ref} (m/s)
No radicals	0.56	1.06
$t = t_1$	0.75	1.43
$t = 0.01\Delta t_2 + t_1$	0.79	1.50
$t = 0.5\Delta t_2 + t_1$	0.90	1.71



Summary of Lifted DME Jet Flame

- The simulated turbulent lifted DME flame shows a similar flame structure and dynamics to the conventional lifted flames without NTC or LTHR, and to the laminar lifted DME flames.
  - Two upstream branches (NTC, LTHR)
  - Triple flame (downstream)
  - Lean stabilization point (NTC)

- Radicals and heat produced at the upstream high temperature flame branch are unlikely to influence overall flame behavior.

- Radicals and heat produced during the first stage ignition enhances the laminar flame speed by around 1.7 in an average sense, leading to a higher triple flame propagation speed.

Turbulent n-C12H26 / air mixing layer autoignition at 25 bar – Borghesi and Chen

- **Pressure:** 25 bar
- **Air stream:** 15% X_{O2}+85% X_{N2}, T=960 K
- **Fuel stream:** *n*-dodecane, T=450 K
- **Kinetics:** 35-species non-stiff reduced
- **Fuel jet velocity:** 21 m/s, Re_j = 7000
- **Setup:**
  - Grid: 1200 x 1500 x 1000 nodes
  - Resolution: 3 micron in fine-grid region
  - Dimensions: 3.6 mm x 4.0 mm x 3.0 mm
  - BCs: X and Z periodic, Y NSCBC outflows

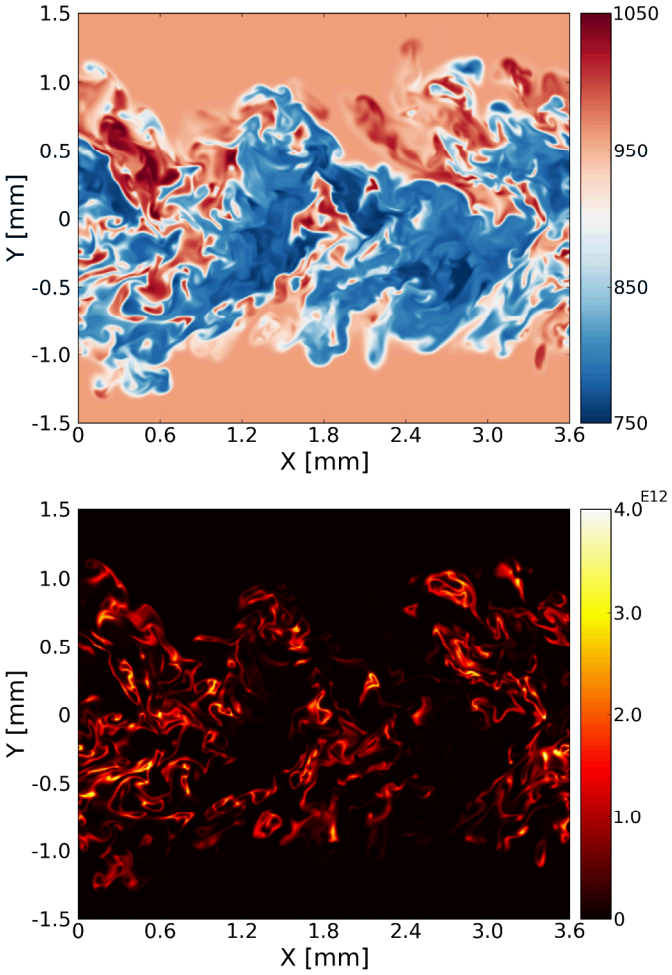
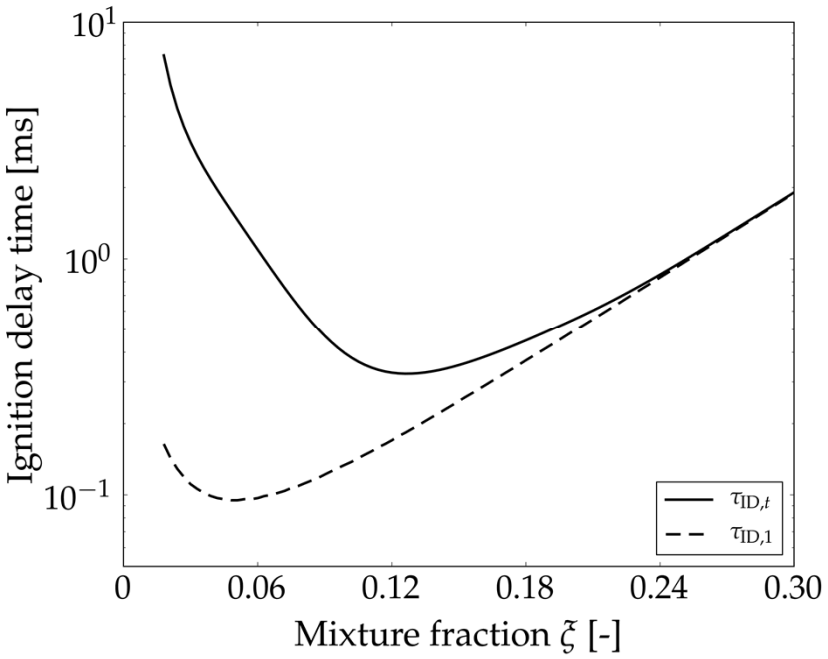
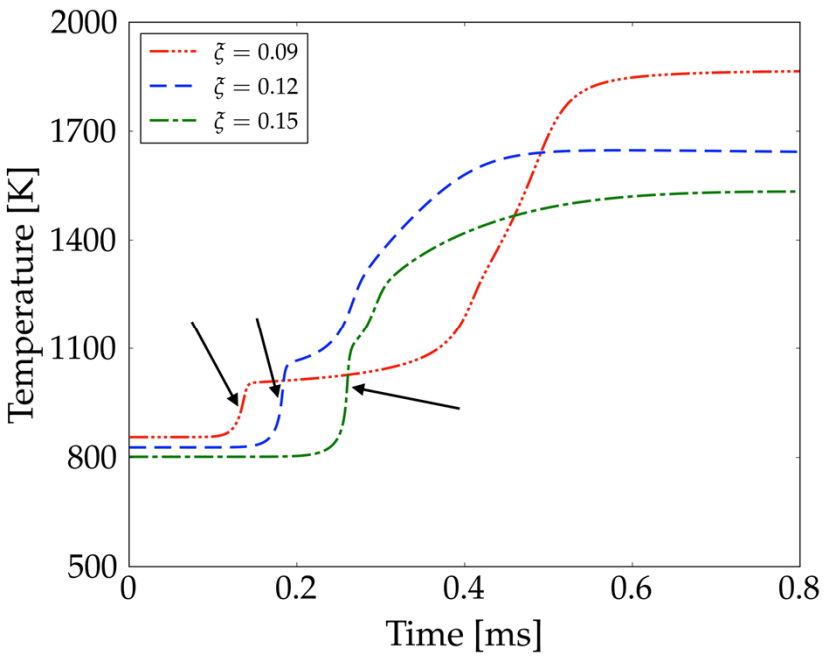


Figure: temperature and HRR at t=0.17 ms after start of reactions

Homogeneous low- and high-T ignition delay





- Low- and high-T fastest ignition occurring at different ξ values;
- Low-T ignition criterion based on inflection point of T(t) curve.

Flamelet simulations

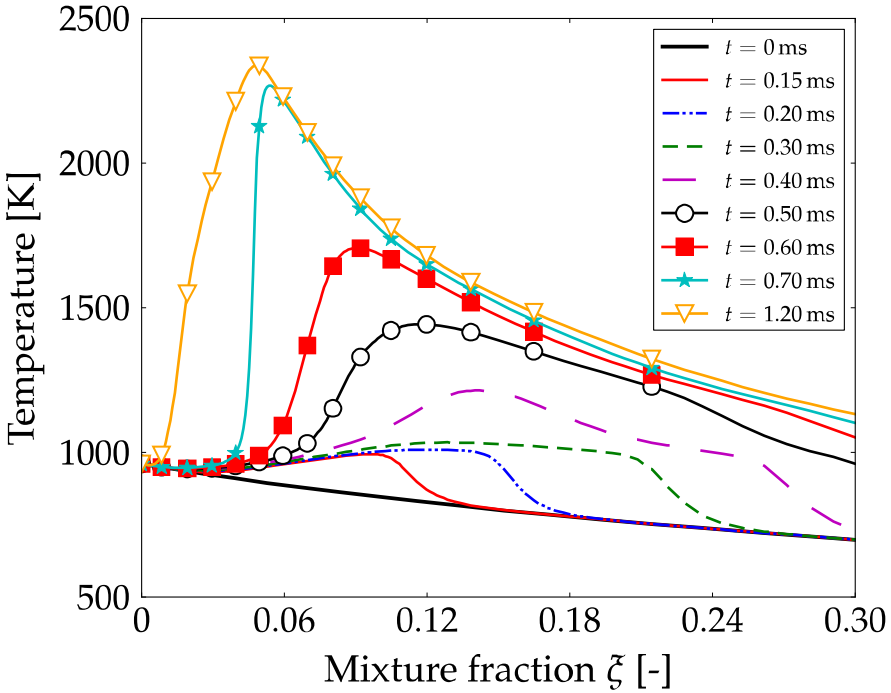


Figure: flamelet solution corresponding to N=10 s^{-1}. Flamelet thermodynamics conditions as in DNS

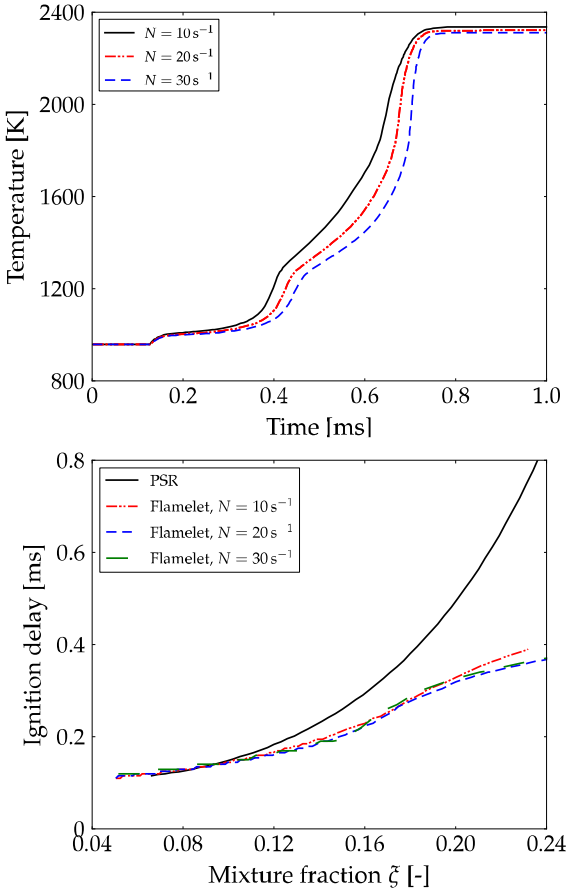


Figure: maximum flamelet temperature and low-T ignition delay for different *N* values

Low-T Ignition in Turbulent Mixing Layer

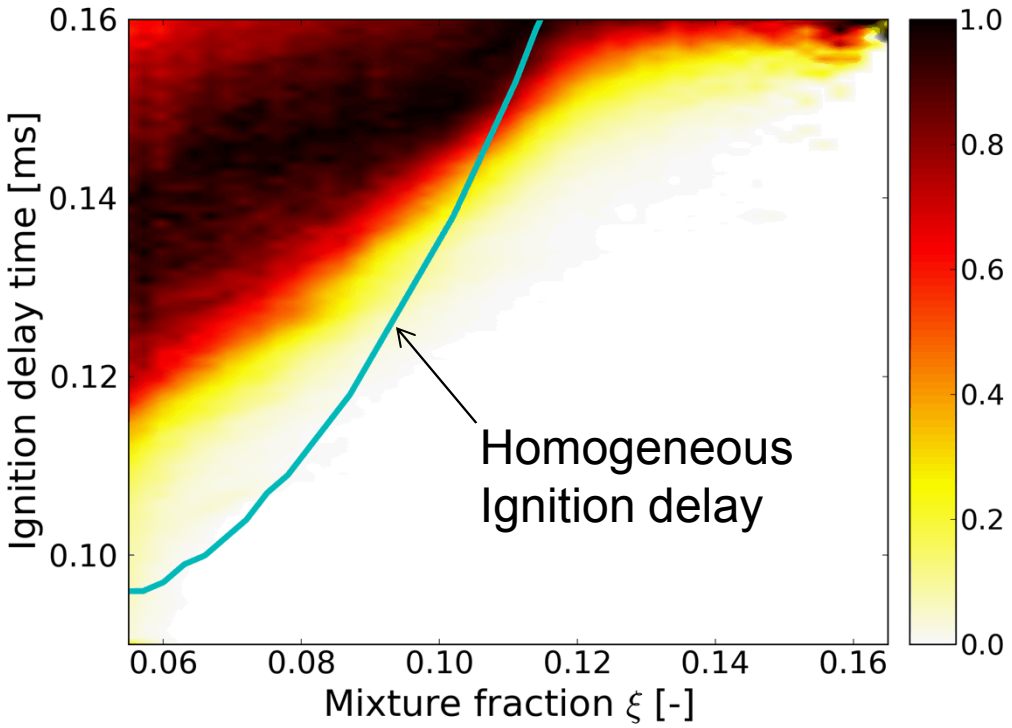
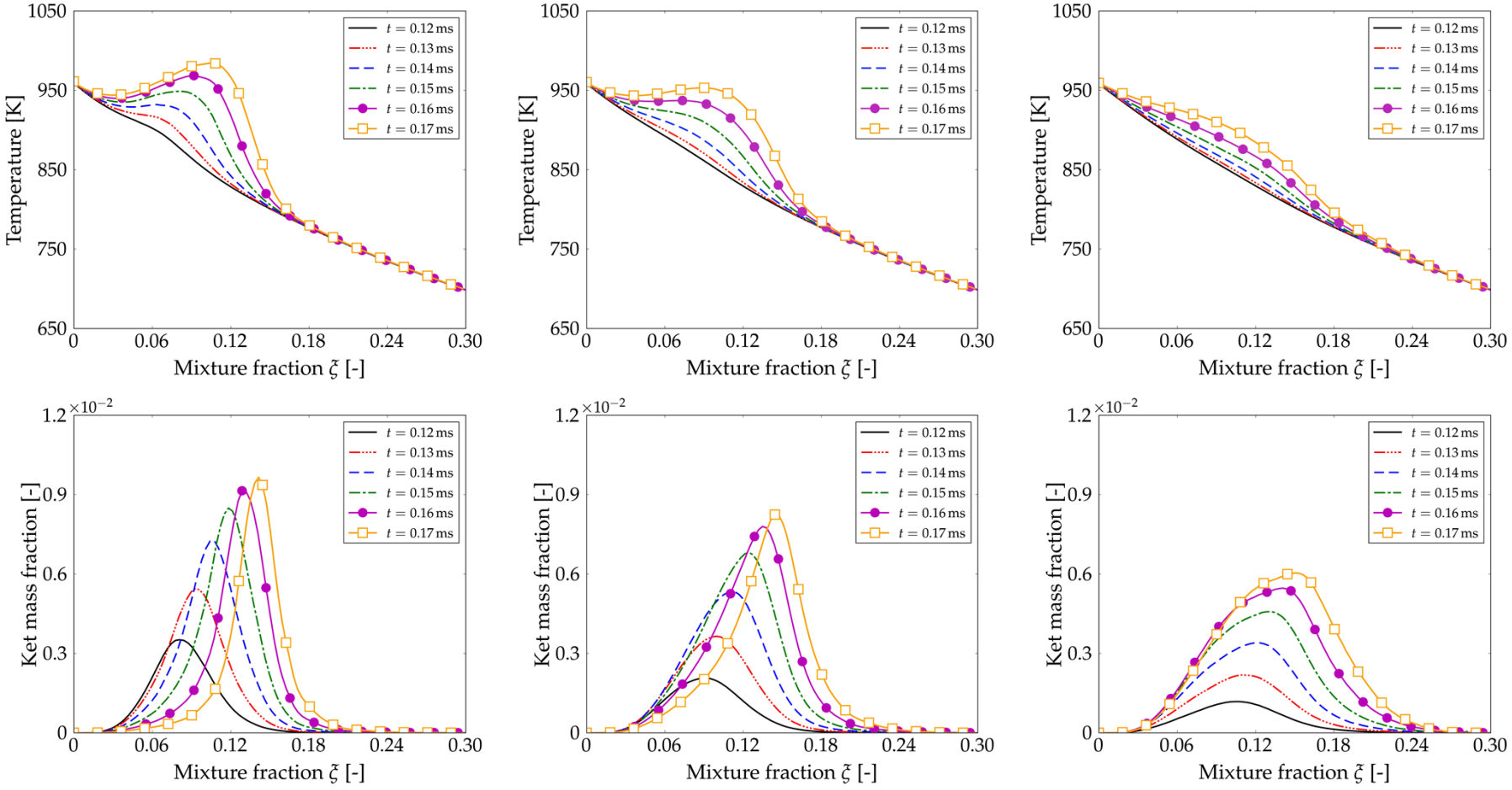


Figure: fraction of computational cells undergoing low-T ignition as a function of elapsed simulation time. Cyan line corresponds to ignition delay from PSR simulations

Temporal evolution of conditional means



increasing scalar dissipation

Beyond Petascale - Why do we need exascale?

- Real-world turbulent combustion consists of phenomena occurring over a wide range of scales that are closely coupled
  - More grid points needed to resolve larger dynamic range of scales (higher Reynolds number, pressure)
  - More time steps or bigger domain needed for larger statistical ensembles
  - Complex laboratory configurations
- More complex fuel blends require larger number of equations per grid point (O(100) species, O(1000) reactions)
- More complex multi-physics(spray, soot, radiation)
- In situ uncertainty quantification with adjoint sensitivity – reverse causality – uncertainties in chemical inputs
- In situ analytics/visualization

Changing HPC Landscape – Why Co-Design?

Old Constraints

- **Peak clock frequency:** as primary limiter for performance improvement
- **Cost:** *FLOPs* are biggest cost for system: *optimize for compute*
- **Concurrency**: Modest growth of parallelism by adding nodes
- **Locality***: MPI+X model (uniform costs within node & between nodes)*
- **Memory Scaling:** maintain byte per flop capacity and bandwidth
- **Uniformity**: Assume uniform system performance

New Constraints

- **Power:** primary design constraint for future HPC system design
- **Cost:** Data movement dominates: optimize to minimize data movement
- **Concurrency:** Exponential growth of parallelism within chips
- **Locality***:* must reason about data locality and possibly topology
- **Memory Scaling**: Compute growing 2x faster than capacity or bandwidth, no global hardware cache coherence
- **Heterogeneity:**Architectural and performance non-uniformity increase

- Future algorithms, programming environments, runtimes, hardware need to:
  - Express data locality (sometimes at the expense of FLOPS) and independence
  - Allow expression of massive parallelism
  - Minimize data movement and reduce synchronization
  - Detect and address faults

ExaCT Vision and Goal

- Goal of combustion exascale co-design is to consider all aspects of the combustion simulation process from formulation and basic algorithms to programming environments to hardware characteristics needed to enable combustion simulations (including in situ UQ and analytics) on exascale architectures
  - Interact with vendors to help define hardware requirements, computer scientists on requirements for programming environment and software stack, and applied mathematics community locality-aware algorithms for PDE's, UQ, and analytics
  - High-fidelity block structured adaptive mesh refinement (AMR) with embedded UQ and in situ analytics

- Combustion is a surrogate for a much broader range of multiphysics computational science areas

Programming Environment Critical to Performance

Effective use of exascale hardware will require programming environment that effectively maps algorithms to hardware

- Driven by programmability of combustion applications and characterization of algorithms on different designs of architectures
  - Simplify programming to express locality and independence
  - Automate discovery of parallelism and hide latencies
  - Simplify programming of extensible workflows, block-structured PDE's, analytics, UQ for performance, scalability, portability, and productivity on heterogeneous architectures

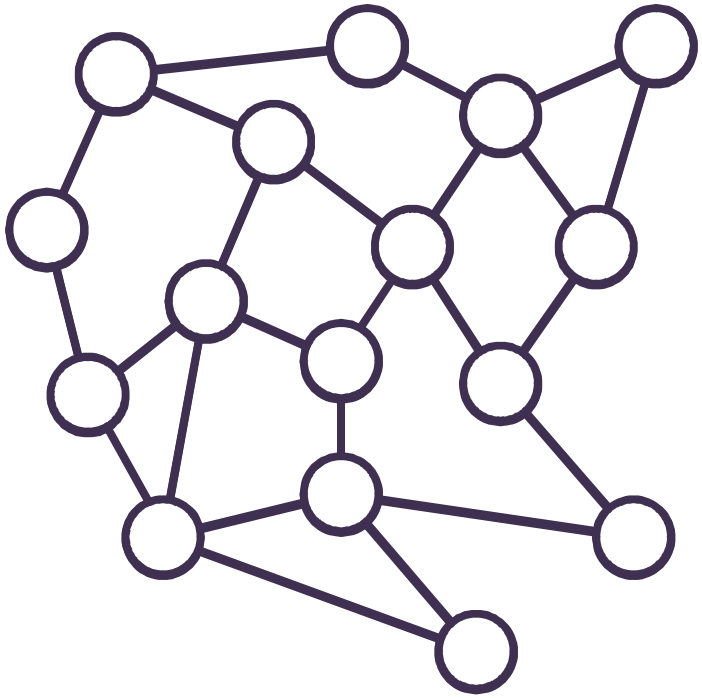
Legion Programming Model

http://legion.stanford.edu

- **Motivation**:
  - Performance-oriented, portable, productive
- **Model**:
  - Capture the structure of program data
  - Decouple specification from mapping to memory hierarchy and processors
  - Automate: data movement, parallelism discovery, synchronization, hiding long latency
  - Support for task- and data-parallel forms
- **Approach:**
  - Proxy applications reveal little detail of impact a programming model has at the scale of a full application – therefore we ported a full application to Legion.

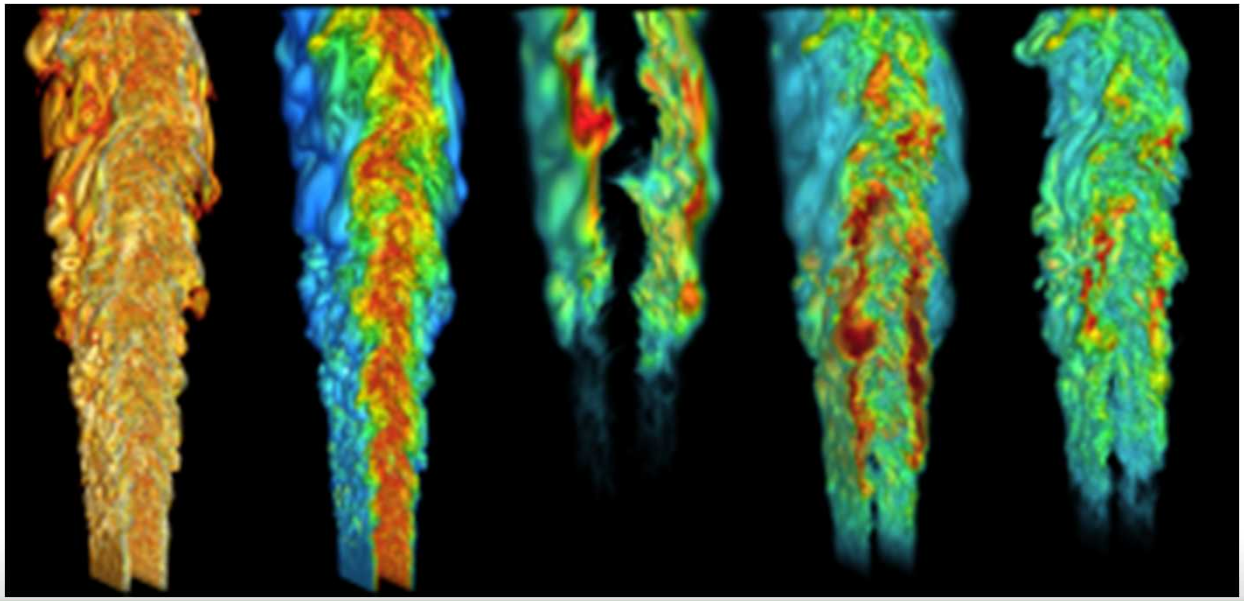
Describing Data with Regions

- A *region* is a (typed) collection of data
  - E.g., nodes of a mesh
- Regions can be *partitioned* into subregions
- Tasks declare
  - (sub)regions they use
  - How they use them
    - Read, write, reduce



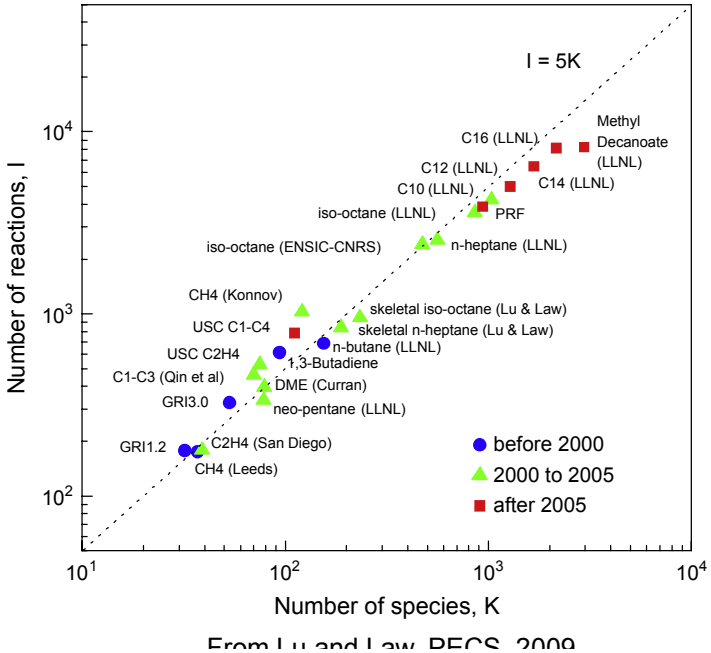
S3D

- Production combustion simulation
- Written in ~200K lines of Fortran
- Direct numerical simulation using explicit methods

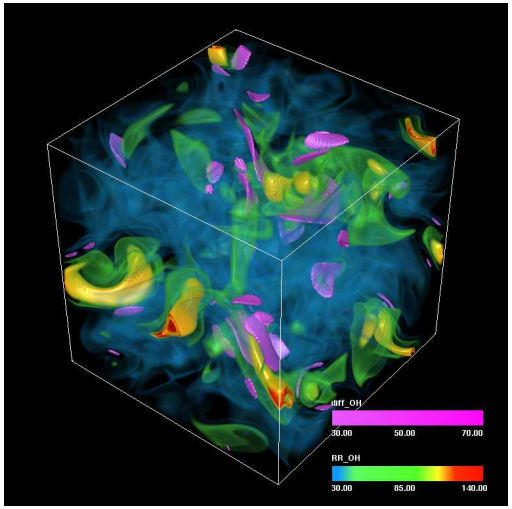


S3D Versions

- Supports many chemical mechanism
  - DME (30 species)
  - Heptane (52 species)
- Fortran + MPI
  - Vectorizes well
  - MPI used for multi-core
- "Hybrid" OpenACC
  - Recent work by Cray/Nvidia/DOE
- Legion interoperates with MPI



From Lu and Law, PECS, 2009



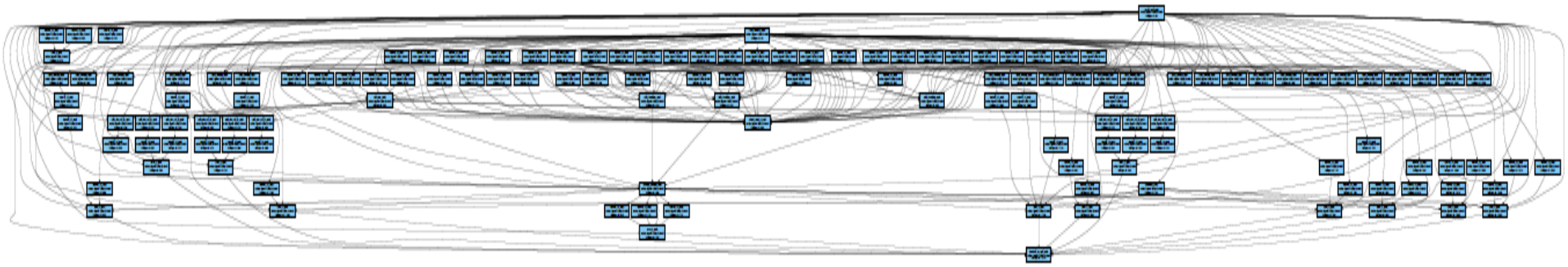
Recent 3D DNS of auto-ignition with 30-species DME chemistry (Bansal *et al*. 2011)

Parallelism in S3D

- Data is large 3D cartesian grid of cells
- Typical per-node subgrid is 48^3 or 64^3 cells
  - Nearly all kernels are per-cell
  - Embarrassingly data parallel
- Hundreds of tasks
  - Significant task-level parallelism
- Except…
  - Computational intensity is low
  - Large working sets per cell (1000s of temporaries)
  - Performance limiter is data, not compute

S3D Task Parallelism

- One call to Right-Hand-Side-Function (RHSF) as seen by the Legion runtime
  - Called 6 times per time step by Runge-Kutta solver
  - Width == task parallelism
  - H2 mechanism (only 9 species)
  - Heptane (52 species) is significantly wider
- Manual task scheduling would be difficult!



Dependence graph of tasks for RHS for hydrogen

S3D Legion GPU Performance

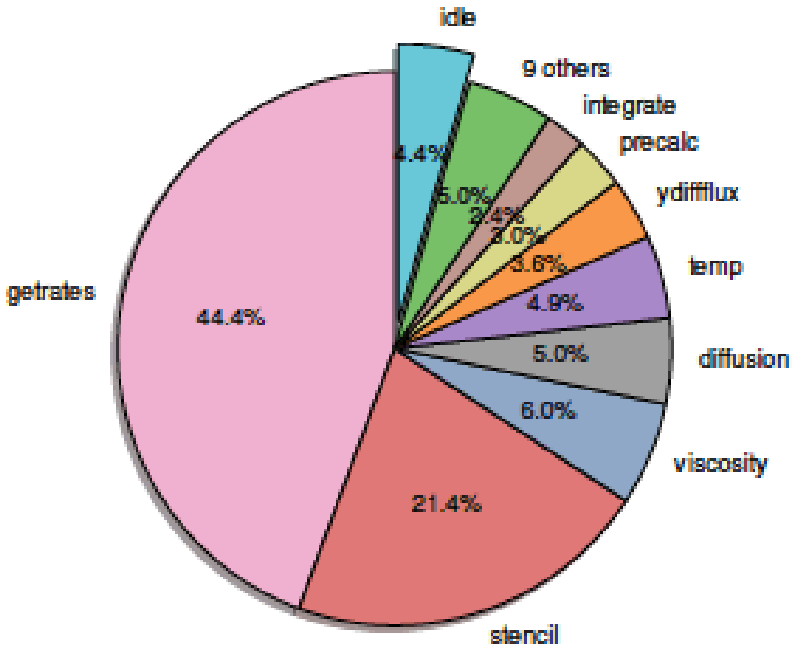


Figure 7: GPU Usage by Task on Titan

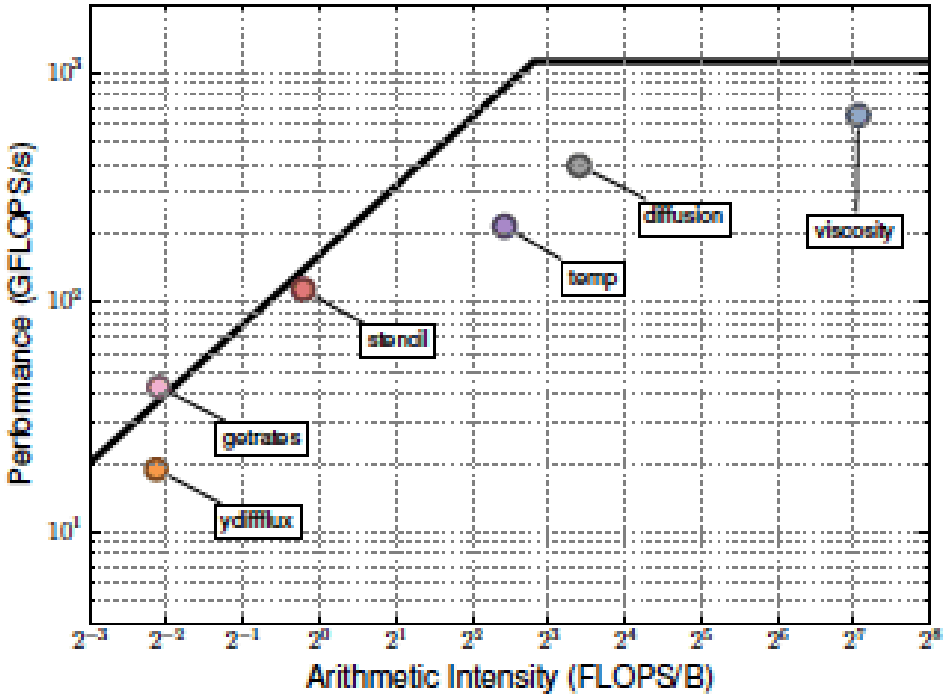
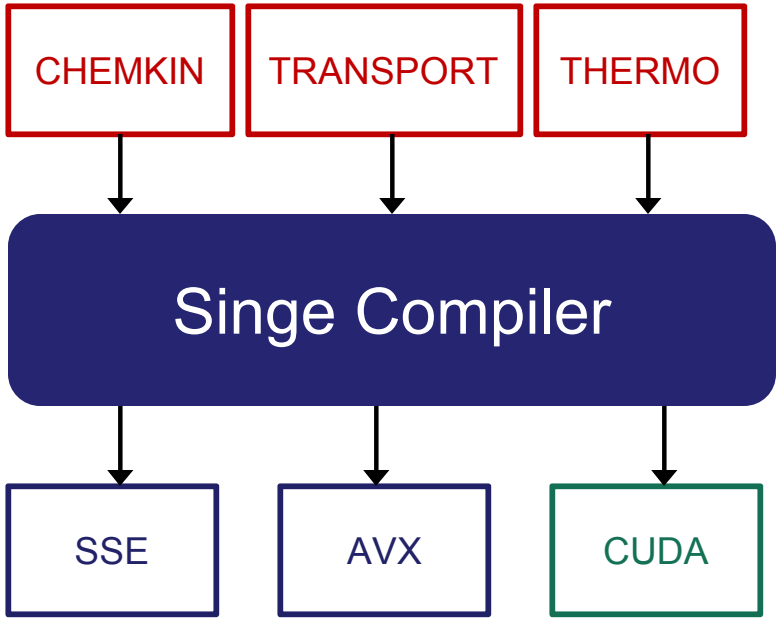


Figure 8: Roofline Analysis of Key GPU Kernels

Leaf Tasks

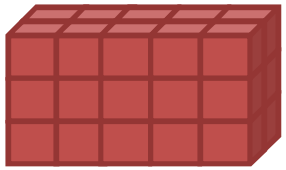
- Legion treats tasks as black boxes
  - Doesn't care how tasks are written
- Still need fast leaf tasks for computationally expensive chemistry, diffusion, viscosity
  - For CPUs & GPUs
  - For multiple mechanisms
- Singe* is a DSL compiler for chemistry kernels

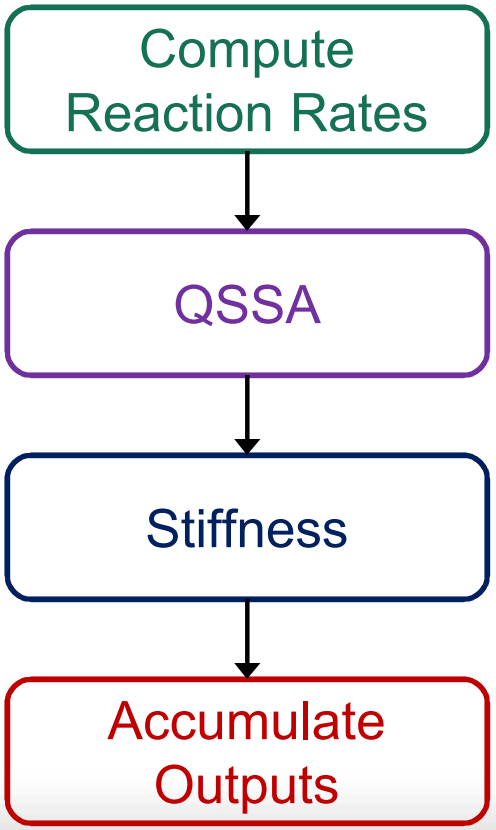


*Bauer et al. PPoPP'14

Combustion Challenges

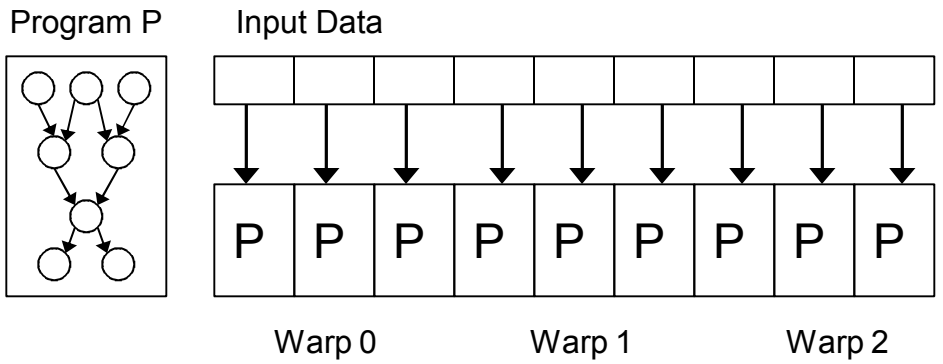
- GPU programming models emphasize data parallelism
  - Not always the best choice for performance
- Large working sets (per point)
  - PRF chemistry needs 1722 double precision reaction rates (per point)
  - GPU register file only store 128 per thread
- Multi-phase computations
  - Fissioned kernels limited by memory bandwidth, slow



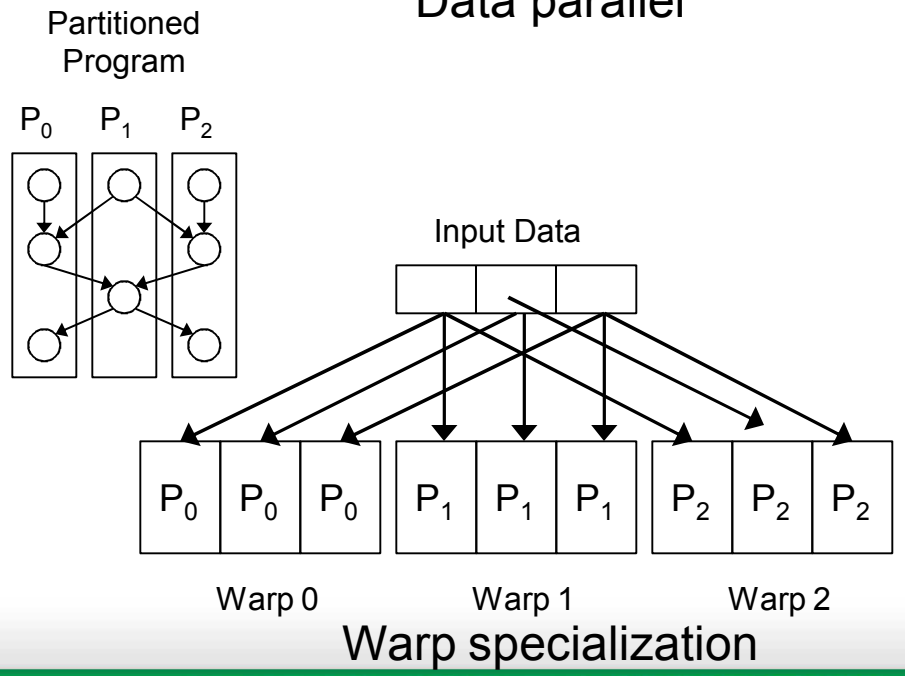


Warp Specialization

- Leverage knowledge of underlying hardware
  - GPUs execute warps: streams of 32-wide vector instructions
  - All threads in warp execute the same program (data parallel unit)



- Each warp can run different computation
  - Generate code that specializes each warp,(leverage task parallelism)
  - Different warps do different computations on the same data
  - Allows much better use of memory while keeping processors busy
  - Fit large working sets on chip



Performance Results

- Chemistry Kernel
  - All Singe kernels significantly faster than current production versions
  - Warp specialized SINGE code is up to 3.75 times faster than previously optimized data-parallel CUDA kernels

- Multi-Node Heterogeneous Testbeds S3D Legion:
  - Keeneland: 128 nodes, 16 Sandy Bridge cores (24 GB RAM), 3 Fermis (6 GB RAM each)
  - Titan: 18K nodes, 16 Interlagos cores(32GB), 1 Kepler K20X GPU (6GB), Cray Gemini interconnect (2nd on Top500)
  - Piz Daint: 5272 nodes, 8-core Sandy Bridge-EP CPU (32GB), 1 Kepler K20X (6GB) GPU, and Cray Aries interconnect (6th on Top500)

S3D Legion Performance on Titan – Weak Scaling

48^3 PRF 116 species

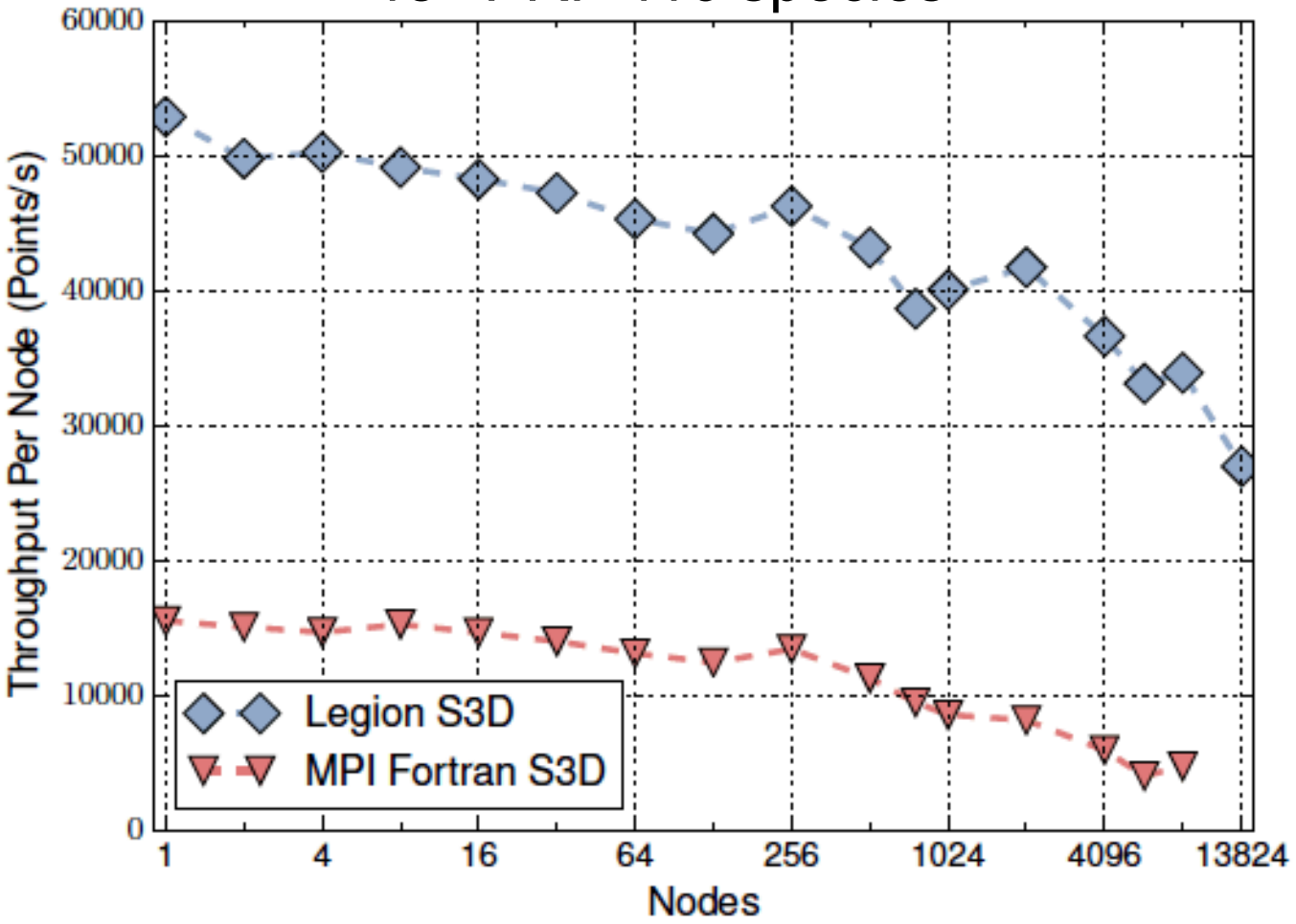


Figure 9: Weak Scaling of PRF on Titan

Legion S3D Execution with In-situ Analytics on a Titan node

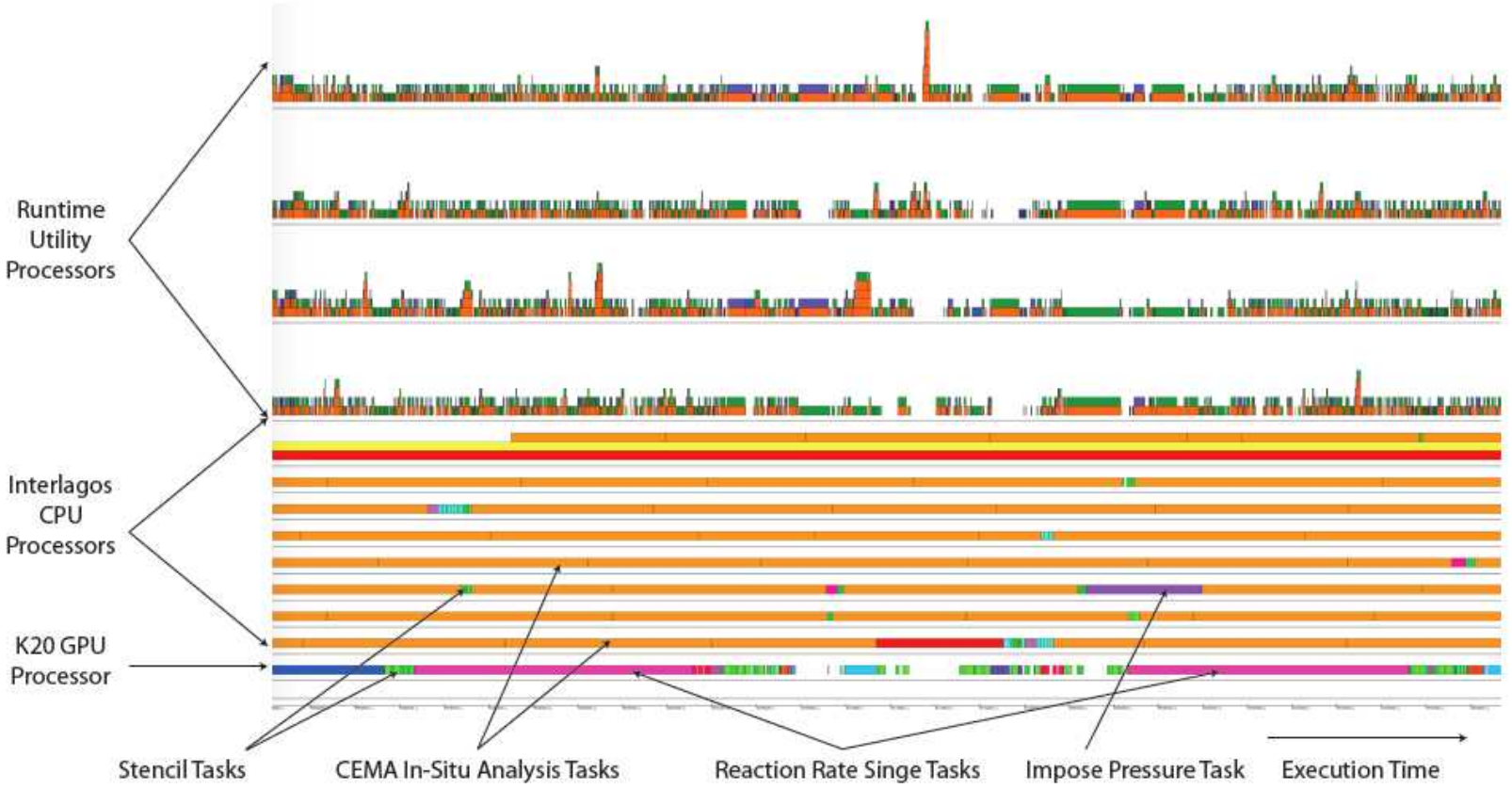


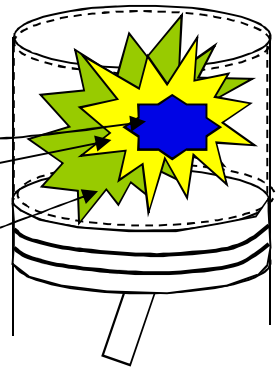
Figure 13: Example All-GPU Mapping Strategy with CEMA In-Situ Analysis for S3D on a Titan Node.

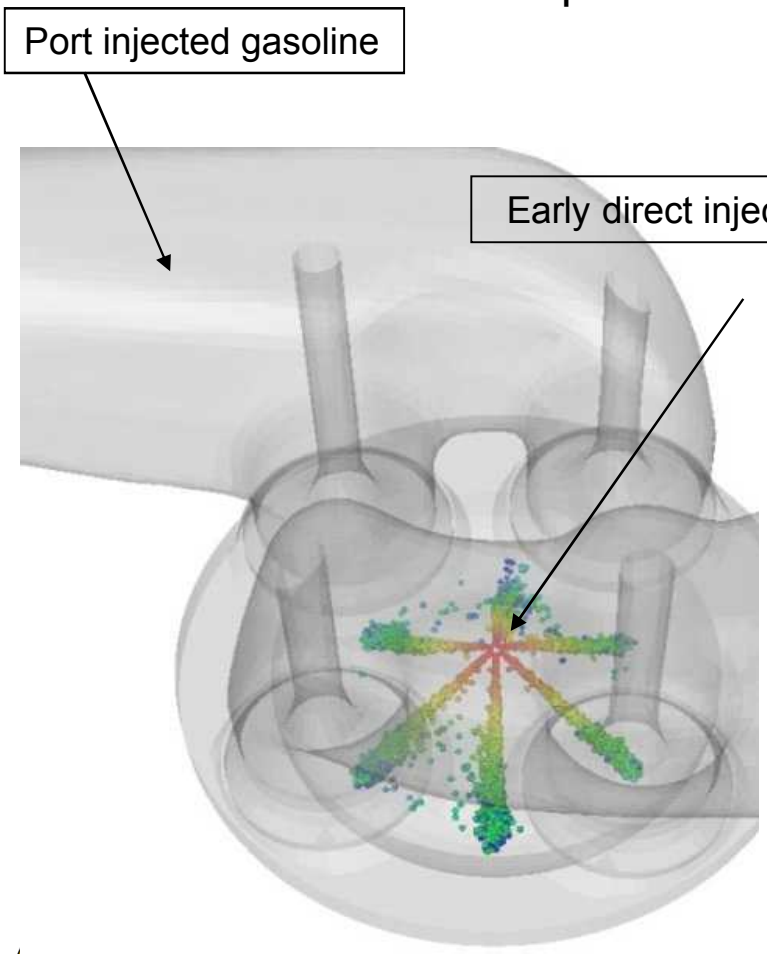
Dual Fuel RCCI combustion – controlled HCCI

Reactivity Controlled Compression Ignition

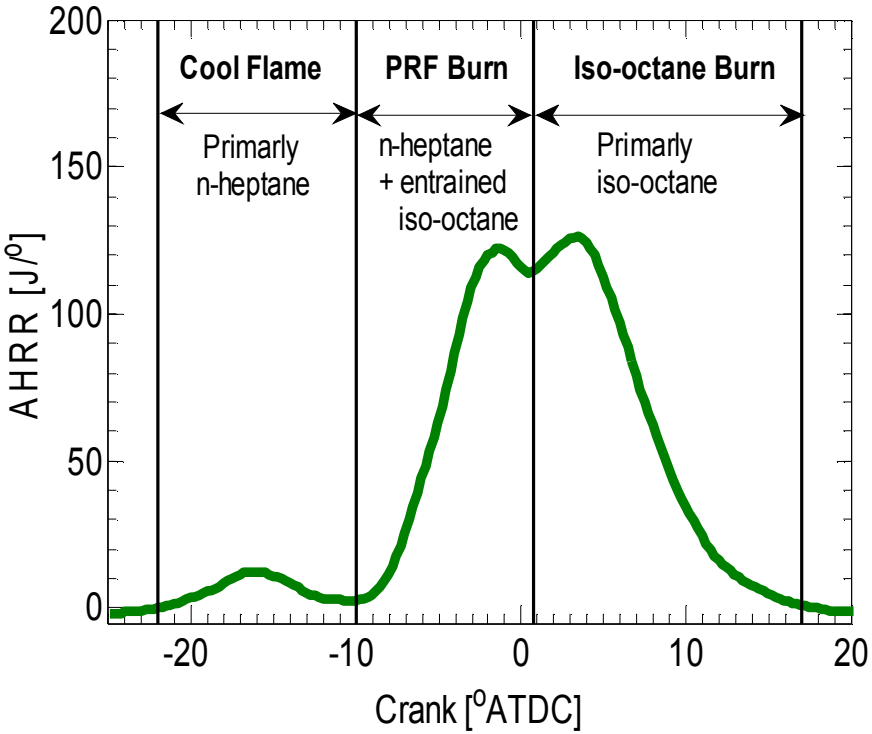
Optimized in-cylinder fuel blending of high cetane diesel with high octane gasoline: control phasing (ignition timing relative to piston motion) and combustion rate

RCCI

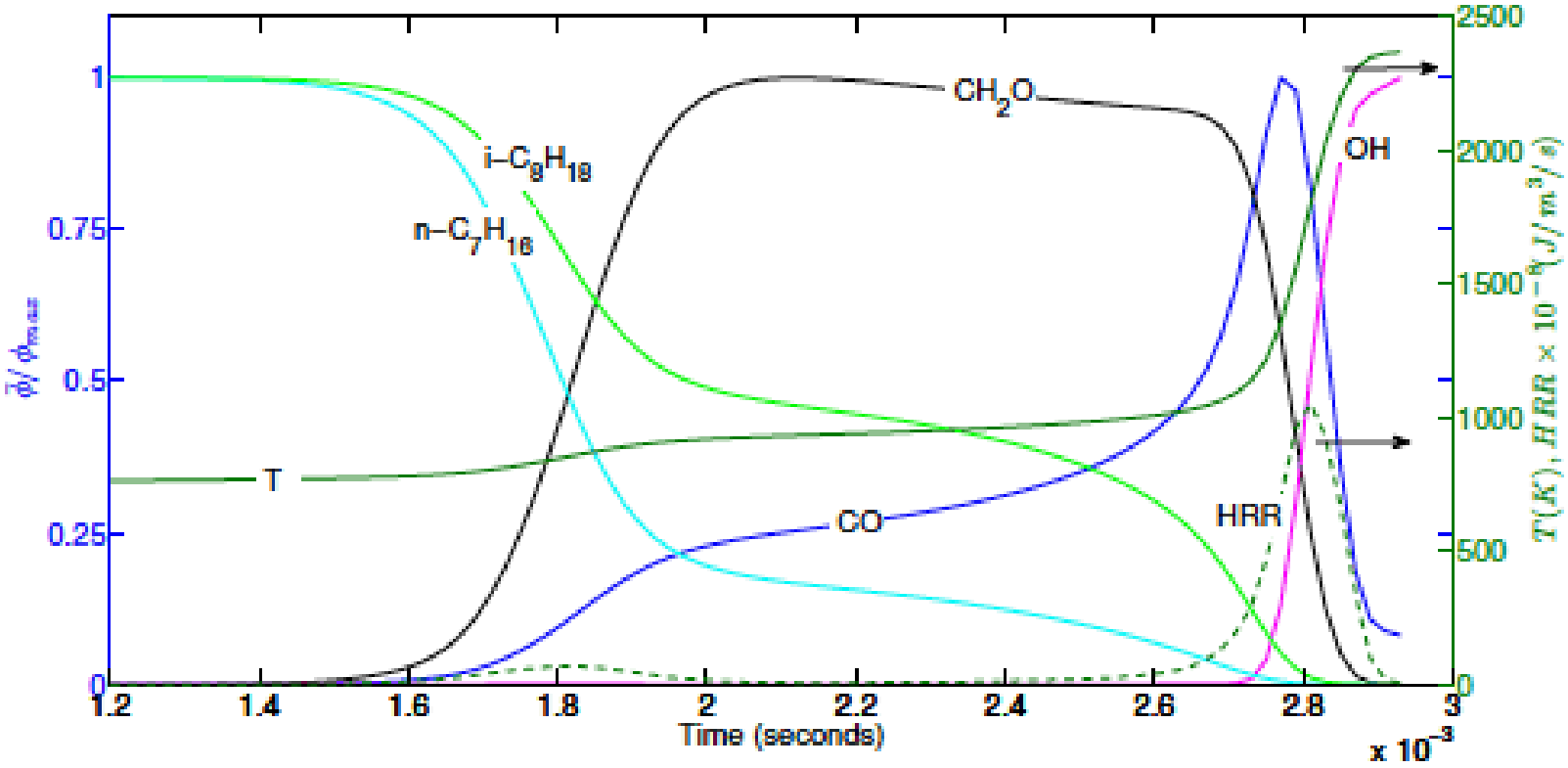




Control of combustion duration by ratio of fuels



Legion S3D: Reactivity Controlled Compression Ignition Primary Reference Fuel Gasoline Blend



RCCI Combustion Mode: Premixed Flames or Spontaneous Autoignition

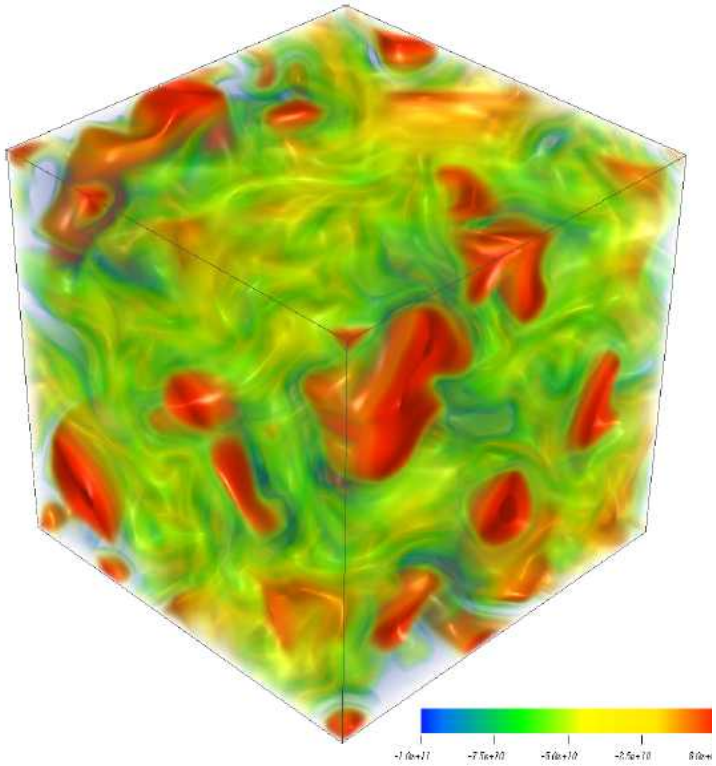


Figure 3: Volume rendering of the heat release rate at the time corresponding to 50% of total heat release. Values are in $J/m^3/s$.

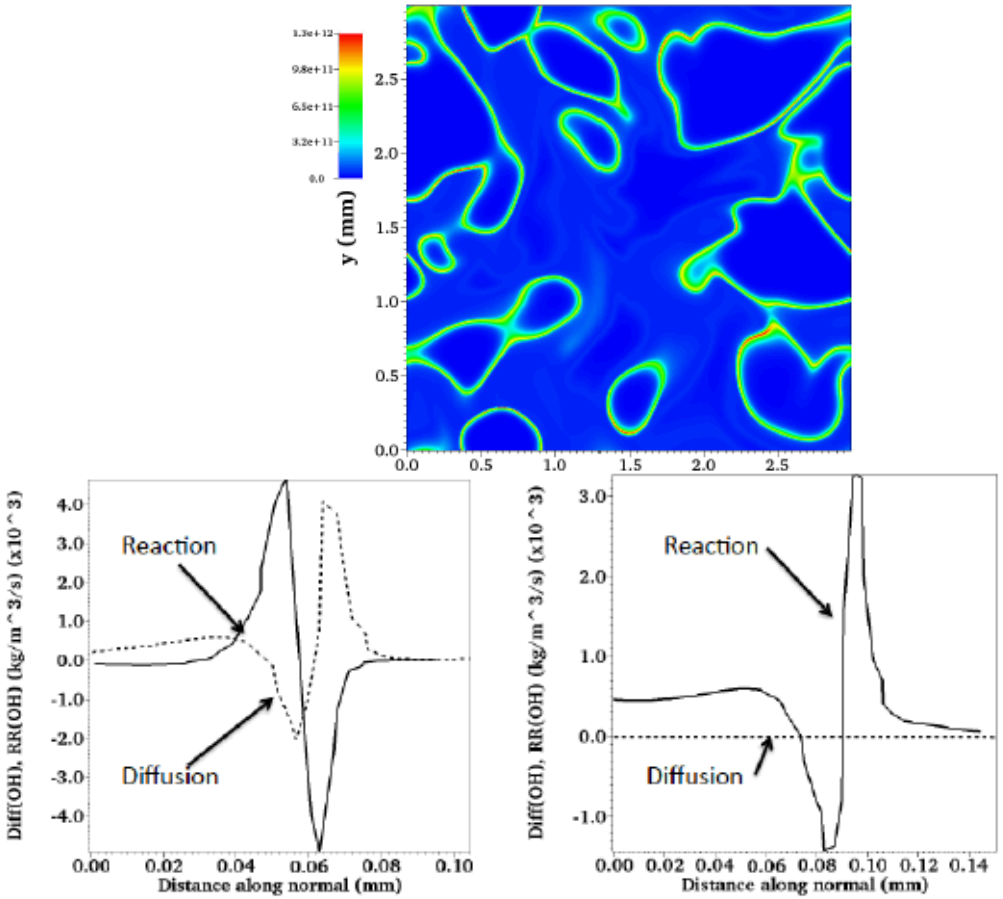


Figure 6: Comparison of the diffusion rate and reaction rate of the OH radical along a flame surface (left) and an ignition front (right).

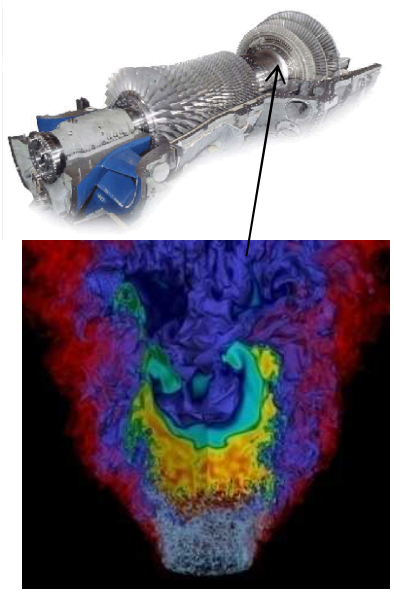
High cetane fuel promotes premixed flame propagation

Legion S3D Lessons Learned

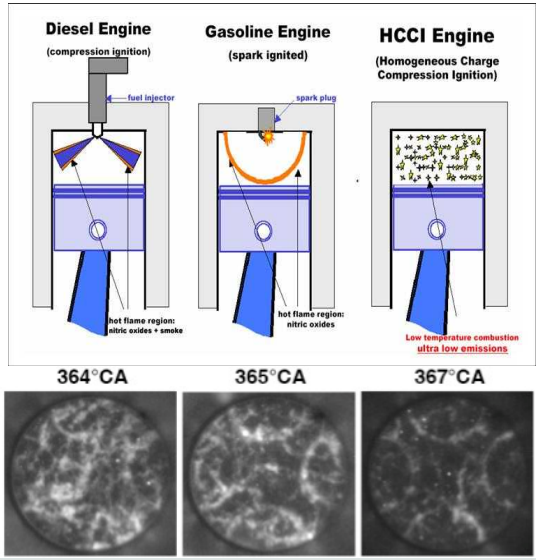
- Legion
  - S3D shows potential of data-centric, task-based models
  - Enables new simulation capabilities (physics, and in situ analytics)
  - Code is easier to modify and maintain
    - Ports are just new mappings, easy to tune for performance
    - New functionality usually just means new tasks
    - Legion will figure out the dependences and scheduling
- Co-Design
  - The Legion/S3D experience is a tribute to co-design
  - Computer and computational scientists worked closely
  - Major progress on important problems resulted

Exascale Targets: Science at Relevant Conditions

- **Reactivity Controlled Compression Ignition (RCCI engine combustion)** – 'Chemical' engine with high diesel-like efficiency without NOx and soot, tailor the charge stratification through direct injection of high-cetane fuel to control combustion phasing, burn rate, and soot formation/oxidation



- **Staged Combustors for Operational/Fuel Flexibility in Gas Turbines**-swirl Injector spray combustion in staged gas turbines with lean premixed combustion, flame stabilization, nitric oxide emissions, thermo-acoustics and hydrogen-enriched natural gas



- **Include UQ with respect to chemistry and transport properties**