2005 Joule Software Effectiveness Study of S3D Applied to the INCITE Goal

J. H. Chen, R. Sankaran, E. R. Hawkes and J. H. Sutherland
Sandia National Laboratories, Livermore, CA

and

D. Skinner
NERSC, Lawrence Berkeley National Laboratories, Berkeley, CA.

Executive Summary

In anticipation of performing the FY05 INCITE goal, the first Direct Numerical Simulation of a 3D turbulent nonpremixed flame with detailed chemistry aimed at understanding extinction and reignition mechansisms, we have successfully optimized key kernels in the DNS code, S3D, on NERSC's IBM SP, Seaborg, and on ORNL's Cray-X1, Phoenix. This document provides an overview of the INCITE goal, the mathematical formulation and numerical implementation, the rationale behind the selection of the physical configuration and parameters, and code optimization on two platforms. While optimization on Seaborg resulted in modest gains, vectorization of S3D on the Phoenix resulted in a ten-fold increase in performance efficiency. S3D was found to scale well on both platforms.

1 Introduction

The objective of this document is to provide an overview of the code optimization and performance studies leading up to and including the FY05 INCITE goal as part of the 2005 Joule Software Effectiveness Study. The INCITE goal is to simulate the first 3D turbulent nonpremixed H2/CO/N2-air flame with detailed chemistry aimed at studying the mechanisms of extinction and reignition. The grid number planned is 0.2 billion. To our knowledge, this would be the largest DNS of a turbulent flame with detailed chemistry performed to date. A state-of-the-art parallel 3D Direct Numerical Simulation (DNS) solver for turbulent reacting flows, S3D, is used to perform the simulations [Chen et al. 2005]. This code was developed with support from the BES Chemical Sciences over the past decade, and more recently, has been extended to include multi-physics and numerical improvements by the SciDAC project High-Fidelity Terascale Simulations of Turbulent Combustion. S3D is a F90/F77 code that is extensible and scales well to over 4000 IBM SP processors using MPI for scaleable parallelism. The code solves the compressible Navier-Stokes, total energy, and species continuity conservation equations in multi-dimensions using high-order spatial finite difference discretization and high-order Runge-Kutta explicit/implicit temporal integration on a uniform or stretched mesh. This document provides the background and motivation for the INCITE goal, the specific research objectives, a description of the mathematical formulation of S3D, the computational approach, physical configuration, and simulation parameters, and S3D code improvements on NERSC's IBM SP and ORNL's CrayX1.

2 Background and Motivation

In many practical combustors the fuel and air are not premixed. For example, this is the case in aircraft applications where fuel and oxidizer are segregated for safety reasons, and in direct injection internal combustion engines, for reasons related to efficiency gains. Therefore, many fundamental combustion questions revolve around the issue of rapid mixing of the reactants which is desirable to maximize heat release rates, thus enabling smaller combustion chamber volumes, and minimizing the production of pollutants. The disadvantage of enhanced mixing rates is that, above a critical value, local extinction or even destabilization of the entire flame can occur. Extinction is dependent mainly upon the balance between local mixing and chemical rates, which depend, in turn, upon the local fuel-air composition and temperature. Extinction adversely affects efficiency, pollutant generation and safety. If extinguished pockets of unburned mixture fail to reignite during a given combustion residence time, then reactions are quenched and unburnt fuel is emitted out of the combustor. If extinguished pockets are abundant due to excessive turbulent strain, and reignition is slow, combustion may cease altogether. In an aircraft, of course, this would be catastrophic.

DNS of turbulent reacting flows has long been a useful, yet computationally limited tool to gain fundamental insight into the physics of turbulence-chemistry interactions [Chen and Im 2000, Hilbert et al. 2004, Poinsot et al. 1996, Vervisch and Poinsot 1998]. These interactions reflect the coupling between fluid dynamics, chemistry, and molecular transport in reacting flows. Even within the continuum assumption, where the Navier-Stokes equations are valid for a large class of flows, the range of length and time scales (over 10 decades) may impose prohibitive requirements on high-fidelity, fine-grained simulations, such as DNS, which are beyond present computer capabilities. Therefore, it is generally accepted that the primary simulation tools for design and optimization of combustion devices will remain limited to two coarse-grained approaches, Reynolds-Averaged Navier Stokes (RANS) simulations and Large-Eddy Simulations (LES). In RANS, the Navier-Stokes (N-S) equations are solved for ensemble mean quantities; while LES is based on the spatial filtering of these equations such that a range of small length and time scales, notably those scales where strong turbulence-chemistry coupling occurs, is not resolved in the simulations. The unclosed terms that result from averaging or filtering of non-linear terms in the incompressible N-S equations require additional modeling. In reacting flows, additional modeling is required for transport and chemical source contributions in the species and energy equations that reflect highly non-linear phenomena. Therefore, the correct representation via modeling of the small scale mixing and reaction interactions is crucial to the successful prediction of efficiency, stability, and emissions in practical devices. Principal approaches to non-premixed combustion modeling include those based on the mixture fraction, e.g. steady [Peters 2000] or unsteady flamelets [Pitsch and Steiner 2000], and conditional moment closure (CMC) [Klimenko and Bilger 1999], those based on the solution of the transport equation for the joint probability density function (PDF) [Pope 1985], and an approach based on a statistical one-dimensional description of turbulence [Kerstein 1999].

As a result of its technological importance, extinction and reignition, and other finite-rate phenomenon in nonpremixed combustion has received considerable attention recently, due in part to a well-documented series of experiments on turbulent jet flames that exhibit local unsteadiness and extinction. The resulting library of flame data has been an invaluable benchmark for the advancement of fundamental understanding and model validation of nonpremixed turbulent combustion in an international collaboration among experimental and computational researchers referred to as the Turbulent Nonpremixed Workshops (http://www.ca.sandia.gov/TNF/abstract.html). Several groups in this forum have demonstrated reasonable success in modeling nonpremixed flames without extinction. However, there are still limitations and uncertainties in these models in their ability to describe extinction and reignition, and other important finite-rate combustion phenomena. For example, in the ODT model, only one mode of reignition is possible due to its intrinsic one-dimensional nature. Multi-dimensional ignition modes would need to be included in this model empirically - the same comment applies to flamelet approaches. A key limitation in the transported PDF approach is the choice of relevant mixing time scale(s) in the presence of finite-rate chemistry, and the influence of preferential diffusion among species on this selection. CMC approaches may require additional conditioning variables in order to predict extinction and reignition. In summary, current modeling approaches would benefit greatly from more detailed characterization of the dynamics of extinction/reignition and other finite-rate effects in turbulent nonpremixed combustion.

In recent years, the rapid growth of computational capabilities has presented both opportunities and challenges for high-fidelity simulations of turbulent combustion flows. Realistic simulations that address complex multi-physics interactions, such as the so-called turbulence-chemistry interactions in combustion flows, have become accessible through the growth of processor speed, computer memory and storage, and significant improvements in computational algorithms and chemical models. While the opportunity exists for gleaning fundamental physical insight into fine-grained chemistry-flow interactions in simplified two-dimensional physical configurations (see the review in [Hilbert et al. 2004]), it remains a formidable challenge to directly simulate three-dimensional turbulent flames with detailed chemistry. In the past several years, the advent of terascale computers in the U.S. and in Japan, has made it possible to begin to study fundamental issues such as flame stabilization and extinction in three-dimensional laboratory configurations with multi-step chemistry using the DNS approach [Mizobuchi et al. 2002, 2004 and Pantano 2004]. These simulations are costly, requiring several million cpu-hours on a terascale computer and between 20 and 100 million grid points. While costly, three-dimensional turbulent direct numerical simulations with detailed chemistry enable both turbulence dynamics and chemical reaction to be accurately represented concurrently, thus opening new realms of possibility for the understanding of turbulence-chemistry interactions and the development of models.

3 Research Goals

The primary objective of the proposed study is to perform a three-dimensional turbulent direct numerical simulation of a nonpremixed H2/CO/N2- air flame with detailed chemistry. This simulation, the first in a series of different Reynolds numbers, will be targeted at providing fundamental insight into key outstanding issues related to modeling of turbulent nonpremixed combustion: extinction and reignition, flow and flame unsteadiness, the correlation of strain rate and scalar dissipation rate, differential diffusion of species, and turbulent mixing in a finite-rate chemical environment. Through collaboration with experimentalists and modelers in the TNF Workshop, we also plan to gather statistics required to further improve or validate different modeling approaches. In the following subsections the specific objectives of the proposed work are presented.

3.1 Extinction and Reignition Dynamics and Statistics

In the absence of significant preferential diffusion effects, models for non-premixed combustion are presently capable of representing with reasonable accuracy turbulent flows without extinction. However the inclusion of extinction and re-ignition remains a challenge. There is a need to provide fundamental information regarding the mechanisms of extinction and re-ignition in a turbulent environment. These processes are likely quite dependent on finite rate, complex chemistry interactions with turbulence and turbulent mixing. The proposed DNS will make a new contribution to this understanding by including detailed chemistry (i.e. capable of representing fully burning and igniting chemical states), heat release, and realistic thermo-chemical properties. Previous studies have typically either used reduced chemistry with heat release [Pantano 2004], reduced chemistry without heat release and artificial adjustment of rate constants [Kronenburg and Papoutsakis 2004] or global one-step chemistry without heat release [Sripakagorn et al. 2004, Mitarai et al. 2003a,b].

Unlike previous DNS studies, the proposed physical configuration with detailed chemistry will permit reignition to occur by either autoignition via chemical chain-branching or by flame propagation (either normal or tangential to the stoichiometric surface of the extinguished flame.) Statistics regarding the occurrence of the different modes of reignition as a function of key flow and flame parameters will be obtained.

3.2 Differential Diffusion

Several nonpremixed combustion models parameterize the thermochemical state with a conserved scalar known as the mixture fraction. The mixture fraction is conserved if the species are assumed to have equal mass diffusivity. Hydrogen and hydrocarbons exhibit a wide spectrum of mass diffusivities distinct from thermal diffusivity. Fast-diffusing, chemically crucial intermediates like H atom are mobile and can segregate from other species and heat. Recently, [Sutherland et al. 2004] have proposed a method of quantifying the degree of differential diffusion (DD) in a flame. The proposed DNS will be used to begin to understand the significance of differential diffusion, as a function of local mixing and reactive conditions, on finite-rate phenomena such as extinction and reignition. For example, we will seek a conserved scalar definition that is least sensitive to DD and a relevant combustion progress variable definition, that together, may allow us to parameterize extinction and reignition processes in a real flame. In the context of LES, the DNS data will be spatially-filtered to assess the relative importance of DD to convection and sub-filter terms on the filtered mixture fraction equation.

3.3 A-priori Model Development and Validation

3D simulations with complex chemistry, and turbulence parameters within the realm of moderate Reynolds number flames will be of significant interest to the combustion modeling community. After an initial investigation of the data, our plan is to invite members of the modeling community to employ the data set either through collaboration or by sharing our data. Several of the key modeling issues that can be addressed with the proposed DNS are outlined below.

For approaches based on the mixture fraction or other conditioning variables, it will be possible to assess the magnitude of the conditional fluctuations, leading to understanding of the extent to which the thermo-chemical state can be parameterized by a reduced set of variables, and, building on our previous work, developing alternative choices for conditioning variables [Sutherland et al. 2004]. Furthermore it will be possible to assess the degree to which these variables can be predicted from the resolved or mean flow, for example using presumed forms of the PDF.

Flamelet approaches may be evaluated and improved by a better understanding of unsteady, multidimensional and differential diffusion effects, and of course extinction and reignition. For flamelet approaches it will be possible by Lagrangian tracking of fluid or flame elements to identify and understand effects of unsteadiness, including extinction and re-ignition, providing valuable information for recently developed approaches to account for these effects [Pitsch et al. 2003, Mitarai et al. 2004]. Multi-dimensional effects can similarly be identified using our parallel surface-based post-processing algorithms [Hawkes and Chen 2004]. Statistics of the conditional average of the scalar dissipation or its PDF can be obtained from the DNS.

CMC approaches will benefit from a better understanding of the magnitude of the conditional fluctuations and the conditional scalar dissipation [Klimenko and Bilger 1999], particularly the differences between extinguished and fully burning regions. In the case of extinction, CMC may require a second conditioning variable, and the DNS can be used to provide information on the selection of the second conditioning variable and the doubly-conditional scalar dissipation rate [Kronenburg and Papoutsakis 2004]. Differential diffusion represents a challenge for CMC [Kronenburg and Bilger 2001 a,b], and DNS can contribute to its development. Recently the CMC approach has been applied to LES, where it is noted that a full implementation of the CMC on the LES grid may be prohibitive due to the introduction of the additional mixture fraction dimension. However researchers [Bushe and Steiner 1999, Navarro-Martinez and Kronenburg 2004] have argued that the spatial variations of the conditional averages may be significantly less than for unconditional quantities, potentially allowing the use of a coarser grid for LES simulations. It will be possible to verify the validity of these assumptions, and any dependence on filter size.

For transported PDF approaches, the main closure problem is for scalar mixing [Pope 1985]. Presently PDF approaches do not distinguish between mixing of conserved and reacting scalars. However reacting scalars can potentially have very different mixing characteristics, particularly where there are different diffusivities and extinction [Mitarai et al. 2003]. Also there may be different mixing characteristics of the conserved scalar in extinguished and burning regions. It will be possible to make a contribution to the understanding of these issues through study of the DNS database.

The disadvantage of the ODT model is that it allows only limited mechanisms for reignition due to its intrinsic one-dimensional nature. For example, triple flame or edge flame propagation which requires nonaligned gradients of the mixture fraction and progress variable can not occur along a one-dimensional domain and would need to be empirically modeled [Hewson and Kerstein 2002]. The DNS could assess the importance of the different reignition modes, and provide clues towards empirical models for edge and triple flame propagation.

3.4 DNS Benchmark for comparison

This use of the data is inspired by the highly successful TNF Workshop, in which experimental benchmark flames were developed and significant progress was made in model development through the provision of a collaborative framework for comparison of modeled and measured results. Typically DNS is not used in this way, rather it is normally performed and exploited by a single or limited number of research groups seeking to advance a particular modeling strategy. However, there are good reasons to suggest that this could be a very profitable use of DNS data. While DNS is necessarily performed with only a limited range of length scales, the ambiguities present in an experiment, in terms of comparisons of modeled and measured results, simply do not exist. The thermo-chemistry, boundary and initial conditions are all completely known, and there is negligible measurement error. Furthermore in the DNS we will have access to time dependent three-dimensional fields, which greatly enhances possibilities for evaluation of models beyond the typical comparisons of simple point-wise averages. It is proposed that after an initial investigation of the results, the data set will be introduced to and shared with the community under the framework of the TNF workshop. Our target for this would be the forthcoming TNF workshop in 2006.

3.5 LES Connection

The proposed work is planned in conjunction with a simultaneous LES effort at Sandia led by J. Oefelein. In contrast to DNS, LES allows a high fidelity representation of the large scales that are strongly influenced by the geometry of the problem, and small scales are modeled. The LES and DNS efforts are complementary and allow us to span the range of scales that exist in a laboratory flame. It is planned to exploit the LES to provide more realistic boundary and initial conditions for DNS, and for the DNS to provide improved sub-grid scale models for the LES.

3.6 Experimental Connection

We plan to use DNS data to determine the effect of measurement uncertainties, for example, due to photon shot noise, on measurements of the scalar dissipation rate. Comparison between experiment and DNS will be achieved through spatial filtering and by modeling shot noise in the DNS data and comparing with the raw DNS and experimental data. Similar comparisons have already been made with large-eddy simulation of a jet flame; however, in the LES approach the full spectrum of mixing and combustion is not resolved on the grid, but rather modeled through flamelets [Geyer et al 2004]. Disparities in such a comparison may not entirely be attributable to shot noise. We will further examine the adequacy of the OH radical to extract flame normal vectors required for multi-dimensional scalar dissipation rate measurements [Karpetis and Barlow 2004].

4 Computational Approach

The simulation will be performed using Sandia's massively parallel direct numerical simulation code, S3D. This code solves the full compressible reacting Navier-Stokes, total energy, species and mass continuity equations coupled with detailed chemistry. It is based on a high-order accurate, non-dissipative numerical scheme. It has been used extensively to investigate fundamental turbulence-chemistry interactions in combustion topics ranging from premixed flames [Hawkes and Chen, 2004, Chen and Im 2000], autoignition [Sankaran *et al.* 2004, Echekki and Chen 2002], to nonpremixed flames [Mahalingam et al 1995, Sutherland et al. 2004]. Time advancement is achieved through a six-stage, fourth-order explicit Runge-Kutta (R-K) method [Kennedy et al 2000], spatial differencing is achieved through high-order (eighth-order with tenth-order filters) finite differences on a Cartesian, structured grid [Kennedy and Carpenter 1994], and Navier-Stokes Characteristic Boundary Conditions (NSCBC) [Poinsot and Lele 1991, Sutherland and Kennedy 2004] were used to prescribe the boundary conditions. The equations are solved on a conventional structured mesh, and scaleable parallelism is achieved through MPI.

This computational approach is very appropriate for the problem selected. The coupling of high-order finite difference methods with explicit R-K time integration make very effective use of the available resources, obtaining spectral-like spatial resolution without excessive communication overheads and allowing scalable parallelism. An alternative strategy that could have been employed is the use of Adaptive Mesh Refinement (AMR). While AMR is attractive for many combustion problems, it is very doubtful that this approach would result in computational savings in the present case. AMR is most efficiently applied in cases where there is a large disparity between flame and turbulence length scales. In the present case, however, the flame and turbulence length scales are overlapping, and thus the region in which a fine grid is required occupies a large proportion of the computational domain. In addition, AMR has not yet been demonstrated to scale up to the large number of processors required for a calculation of this magnitude.

5 Physical Configuration

DNS of a 3D turbulent nonpremixed CO/H2/N2-air jet flame with detailed chemistry will be performed. The kinetic mechanism employed for CO/H2 oxidation includes 12 species and 33 reactions [Li et al]. The physical configuration chosen corresponds to a temporally-evolving jet flame. In the temporal configuration an inner turbulent fuel core flows within quiescent air, and these streams are separated by reacting mixing layers under the influence of significant mean shear. The configuration results in similar but not identical turbulent structures to those observed in a spatially evolving planar jet, with an observation window that moves with the mean jet velocity. This configuration was selected rather than the spatially-evolving jet because it allows for more significant flame-turbulence interaction within a given computational domain with wider separation in mixing scales than previously possible [Pantano 2004, Sripakagorn 2004], thereby potentially creating a more wrinkled flame surface through intense turbulent mixing. More wrinkling may lead to a greater probability of different modes of reignition to occur than previously studied, such as by flame propagation in a direction normal to the stoichiometric surface or by self-ignition of a fluid parcel following heat conduction from neighboring flame or product gases. Pantano focused primarily on reignition via edge flame propagation tangent to the stoichiometric surface [Pantano 2004]. The other modes rely on high strain to reduce the separation distance between burning and quenched regions in a wrinkled flame, in the direction normal to the stoichiometric surface, so that heat conduction and radical diffusion can reignite the mixture.

??? include a schematic diagram here?

To ramp up to the large INCITE calculations, many test calculations ranging from three to six million grid points and one production calculation with forty million grid points have been performed. The production calculation was performed on the MPP2 Linux cluster at PNNL and run on 480 processors there. Figure 1 shows a volume rendering of the hydroxyl radical mass fraction at an instant at approximately 4 flow-through times into the simulation. The field shows a complex three-dimensional structure with areas of low OH. Figure 2 shows a volume rendering of the local scalar dissipation rate. Areas of high scalar dissipation, highlighted in red, exist in highly localized sheet-like structures, which also correspond to low OH values. This calculation reveals significant three-dimensional flame structure largely generated by vortex stretching induced by the mean shear, and localized regions of extinction followed by reignition. Relative to this run, we expect to increase both Reynolds numbers and the amount of extinction for the INCITE calculation.

6 Physical and Numerical Parameter Selection Considerations

The Direct Numerical Simulation of a reacting jet with extinction and re-ignition is a formidable task, both in terms of the computational cost and in finding the optimal physical parameter space. The numerical and physical parameter selection is closely intertwined and must be discussed together.

The computational cost of this explicit code can be easily estimated as the product of the total grid number, the total number of time steps and the cost per grid point and time step. It is essential to resolve both the small scale chemical and fluid mechanical timescales as well as provide enough large scale structures to allow for normal flow development and adequate statistics. There are limitations on the total number of grid points and on the total computational cost.

Selection of the physical parameters is determined mainly by the Reynolds number, which it is desirable to maximize, and the characteristics of extinction and reignition. First, an appropriate amount of extinction must be obtained. This is governed largely by the Damköhler number, defined as the ratio of a characteristic turbulence timescale to a characteristic flame timescale. Low Damköhler numbers are required to obtain extinction. Second, realistic reignition modes must be obtained. This implies that mixing rates must relax in order to allow reignition, which occurs naturally in the plane jet configuration, but must occur within a reasonable computational time. We believe that the importance of different reignition modes is governed by the ratio of the turbulence intensity to the edge flame speed, introducing a further parameter of importance.

The number of grid points is determined by simultaneous considerations of the Reynolds number,

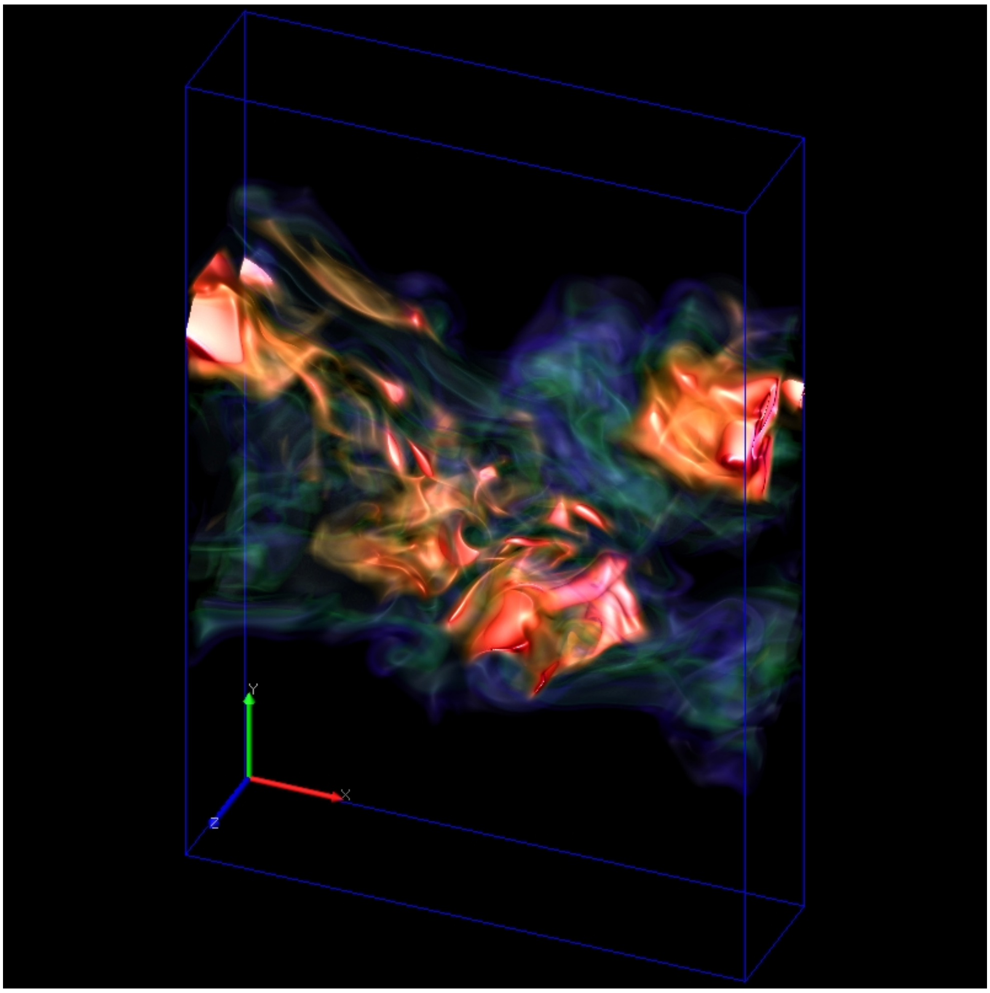

Figure 1: Hydroxyl radical in a turbulent jet flame

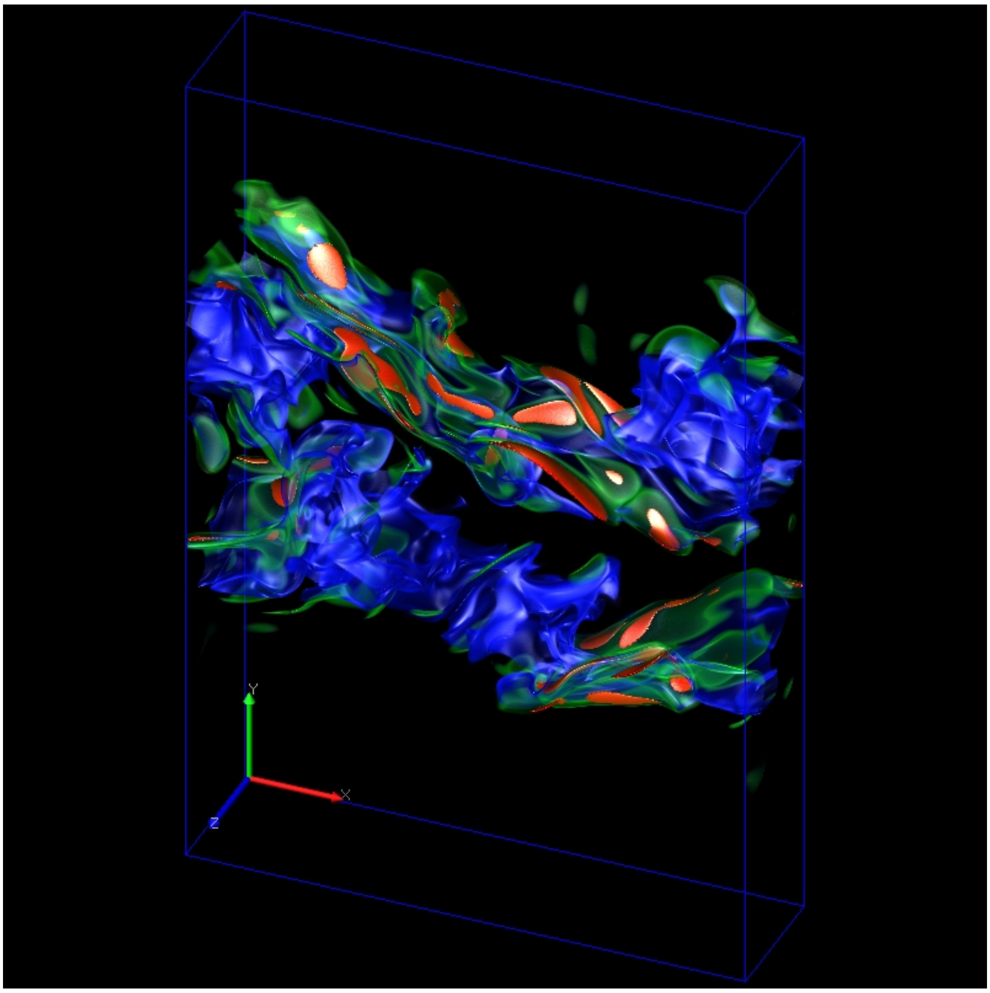

Figure 2: Scalar dissipation rate in a turbulent jet flame

Damköhler number, and resolution required by the chemistry. It is well known that the resolution requirements for cold flow DNS scale with $Re^{9/4}$. The Reynolds number together with an adequate number of large scale structures implies a certain grid resolution, and the Damköhler number together with an adequate range of mixture fraction scales also implies a resolution. Reynolds number scales with the jet velocity and height, while Damköhler number is proportional to the height and inversely proportional to the jet velocity. For a given chemistry, there may be only a small range of relevant parameter space that is accessible, even on terascale computers, with a given number of grid points. Although we are still in the process of parameter selection, we estimate the grid spacing to be 15 microns in each direction. These estimates are based on two- and three-dimensional turbulent simulation tests accounting for the local structure of highly strained, extinguishing flames, which require approximately three times more resolution than what would be estimated from steady extinction conditions. Our target grid number is 0.2 billion grid points, allowing a moderate Reynolds number.

For the compressible code, assuming uniform grid and that the acoustic CFL criterion controls the time-step, the cost for simulation of a single transient jet time is proportional to $N\frac{A}{M}N_H$ where N is the total grid number, M is the Mach number (the ratio of the jet velocity to the sound speed), N_H is the number of grid points resolving the jet height, and A is the factor by which the time step must be smaller than the acoustic CFL stability limit. Many such transient times must be run over the course of a given simulation. To reduce the overall computational cost the Mach number will be maximized while ensuring that the flow is essentially incompressible. Other parameters will be selected in order to maximize the Reynolds number for the given computational effort and to give the desired levels of extinction.

7 Formulation

The DNS solves a coupled system of time varying partial differential equations (PDEs) governing the conservation of mass, momentum, and energy, and species continuity. These governing equations are outlined in §7.1. The PDEs are supplemented with additional constitutive relationships, such as the ideal gas equation of state, and models for reaction rates, molecular transport, and thermodynamic properties.

7.1 Governing Equations

The equations governing reacting flows may be written in conservative form as

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= -\nabla_\beta \cdot (\rho \mathbf{u}_\beta), && (1)\\
\frac{\partial (\rho \mathbf{u}_\alpha)}{\partial t} &= -\nabla_\beta \cdot (\rho \mathbf{u}_\alpha \mathbf{u}_\beta) + \nabla_\beta \cdot \boldsymbol{\tau}_{\beta\alpha} - \nabla_\alpha p + \rho \sum_{i=1}^{N_s} Y_i \mathbf{f}_{\alpha i}, && (2)\\
\frac{\partial (\rho e_0)}{\partial t} &= -\nabla_\beta \cdot [\mathbf{u}_\beta(\rho e_0 + p)] + \nabla_\beta \cdot (\tau_{\beta\alpha} \cdot \mathbf{u}_\alpha) \\
&\quad -\nabla_\beta \cdot \mathbf{q}_\beta + \rho \sum_{i=1}^{N_s} Y_i \mathbf{f}_{\alpha i} \cdot (\mathbf{V}_{\alpha i} + \mathbf{u}_\alpha), && (3)\\
\frac{\partial (\rho Y_i)}{\partial t} &= -\nabla_\beta \cdot (\rho Y_i \mathbf{u}_\beta) - \nabla_\beta \cdot (\rho Y_i \mathbf{V}_{\beta i}) + W_i \dot{\omega}_i, && (4)
\end{aligned}$$

where ∇_β is the gradient operator in direction β, Y_i is the mass fraction of species i, W_i is the molecular weight of species i, $\tau_{\beta\alpha}$ is the stress tensor, $\mathbf{f}_{\alpha i}$ is the body force on species i in direction

α, $\mathbf{q}_\beta$ is the heat flux vector, $\mathbf{V}_{\beta j}$ is the species mass diffusion velocity, $\dot{\omega}_i$ is the molar production rate of species i, and e_0 is the specific total energy (internal energy plus kinetic energy),

$$e_0 = \frac{\mathbf{u}_\alpha \cdot \mathbf{u}_\alpha}{2} - \frac{p}{\rho} + h, \tag{5}$$

and h is the total enthalpy (sensible plus chemical). Throughout this document, α, β, γ will indicate spatial indices and i, j, will indicate species indices unless stated otherwise. Repeated spatial indices imply summation. For example, in cartesian coordinates,

$$\nabla_\beta \cdot (\rho \mathbf{u}_\beta) = \frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} + \frac{\partial(\rho w)}{\partial z},$$

where u, v, and w are the velocity components in the x, y, and z directions, respectively. Only $(N_s - 1)$ species equations are solved because the sum of the N_s species equations yields the continuity equation. The mass fraction of the last species is determined from the constraint

$$\sum_{i=1}^{N_s} Y_i = 1. \tag{6}$$

Assuming an ideal gas mixture, the equation of state is given as

$$p = \frac{\rho R_u T}{W}, \tag{7}$$

where R_u is the universal gas constant and W is the mixture molecular weight given by

$$W = \left(\sum_{i=1}^{N_s} Y_i / W_i \right)^{-1} = \sum_{i=1}^{N_s} X_i W_i. \tag{8}$$

The species mass fractions (Y_i) and mole fractions (X_i) are related by

$$\frac{Y_i}{X_i} = \frac{W_i}{W}. \tag{9}$$

Relevant thermodynamic relationships between enthalpy and temperature for an ideal gas mixture include

$$h = \textstyle\sum_{i=1}^{N_s} Y_i h_i, \qquad h_i = h_i^0 + \int_{T_0}^{T} c_{p,i}\, \mathrm{d}T,$$

$$c_p = \textstyle\sum_{i=1}^{N_s} c_{p,i} Y_i, \qquad c_p - c_v = R_u / W.$$

where h_i is the enthalpy of species i, h_i^0 is the enthalpy of formation of species i at temperature T_0, and c_p and c_v are the isobaric and isochoric heat capacities, respectively.

7.2 Constitutive Relationships

The stress tensor, species diffusion velocities, and heat flux vector in equations (2)-(4) are given by [?, ?, ?]

$$\tau_{\beta\alpha} = \tau_{\alpha\beta} = \mu \left[\nabla_\alpha \mathbf{u}_\beta + \nabla_\beta \mathbf{u}_\alpha \right] - \delta_{\alpha\beta} \left(\frac{2}{3}\mu - \kappa \right) \nabla_\gamma \cdot \mathbf{u}_\gamma, \tag{10}$$

$$\mathbf{V}_{\alpha i} = \frac{1}{X_i} \sum_{j=1}^{N_s} \frac{Y_j}{X_j} D_{ij} \mathbf{d}_{\alpha j} - \frac{D_i^T}{\rho Y_i} \nabla_\alpha (\ln T), \tag{11}$$

$$\mathbf{q}_\alpha = -\lambda \nabla_\alpha T + \sum_{i=1}^{N_s} h_i \mathbf{J}_{\alpha i} - \sum_{i=1}^{N_s} \frac{p}{\rho Y_i} D_i^T \mathbf{d}_{\alpha j}. \tag{12}$$

where μ is the mixture viscosity, κ is the bulk viscosity, D_{ij} are the *multicomponent* diffusion coefficients, D_i^T is the thermal diffusion coefficient for species i, λ is the thermal conductivity, $\mathbf{J}_{\alpha i} = \rho Y_i \mathbf{V}_{\alpha i}$ is the species diffusive flux, and $\mathbf{d}_{\alpha i}$ is the diffusion driving force for species i in direction α, given by [?, ?, ?]

$$\mathbf{d}_{\alpha i} = \underbrace{\nabla_\alpha X_i}_{1} + \underbrace{(X_i - Y_i)\nabla_\alpha(\ln p)}_{2} + \underbrace{\frac{\rho Y_i}{p}\left[\mathbf{f}_{\alpha i} - \sum_{j=1}^{N_s} Y_j \mathbf{f}_{\alpha j}\right]}_{3}. \tag{13}$$

The driving force vector involves thermodynamic forces generated by gradients in concentration (term 1), gradients in pressure (term 2) also called “barodiffusion,” and due to any body force such as an electrical or gravitational field (term 3). Equation (13) allows for the possibility that the force on each species, $\mathbf{f}_{\alpha i}$, is different, though in the case of a gravitational field, $\mathbf{f}_{\alpha j} = \mathbf{g}_\alpha$, and term 3 is identically zero. In the following sections, we will consider the fluxes given in equations (10)-(12) in more detail.

7.2.1 Stress Tensor

For monatomic gases, κ is identically zero, and it is often neglected for polyatomic gases as well [?, ?, ?]. It will be neglected in all discussion herein. This simplifies (10) to

$$\tau_{\beta\alpha} = \tau_{\alpha\beta} = \mu\left(\nabla_\alpha \mathbf{u}_\beta + \nabla_\beta \mathbf{u}_\alpha - \frac{2}{3}\delta_{\alpha\beta}\nabla_\gamma \cdot \mathbf{u}_\gamma\right), \tag{14}$$

which is the form that will be used for this work.

7.2.2 Mass Diffusion Flux

It should be noted that all diagonal components of the multicomponent diffusion coefficient matrix (D_{ii}) are identically zero [?]. Also, the diffusive fluxes and driving forces for all species must sum to zero,

$$\sum_{i=1}^{N_s} \mathbf{J}_{\alpha i} = \sum_{i=1}^{N_s} \rho Y_i \mathbf{V}_{\alpha i} = 0, \qquad \sum_{i=1}^{N_s} \mathbf{d}_{\alpha i} = 0. \tag{15}$$

Equation (11) is often approximated as [?, ?, **?**, ?]

$$\mathbf{V}_{\alpha i} = -\frac{D_i^{\text{mix}}}{X_i}\mathbf{d}_{\alpha i} - \frac{D_i^T}{\rho Y_i}\nabla_\alpha \ln T, \tag{16}$$

where D_i^{mix} are “mixture-averaged” diffusion coefficients given in terms of the binary diffusion coefficients ($\mathcal{D}_{ij}$) and the mixture composition as

$$D_i^{\text{mix}} = \frac{1 - X_i}{\sum_{j\neq i} X_j / \mathcal{D}_{ij}}, \tag{17}$$

where the binary coefficient matrix is symmetric ($\mathcal{D}_{ij} = \mathcal{D}_{ji}$), and the diagonal elements are zero ($\mathcal{D}_{ii} = 0$). If we assume that body forces act in the same manner on all species and baro-diffusion (term 2 in (13)) is negligible, then (13) becomes $\mathbf{d}_{\alpha i} = \nabla_\alpha X_i$. If we further neglect the Soret effect, (the second term in (11) and (16)), then (16) reduces to

$$\mathbf{V}_{\alpha i} = -\frac{D_i^{\text{mix}}}{X_i}\nabla_\alpha X_i, \tag{18}$$

which, using (9), can be expressed in terms of mass fractions as

$$\begin{aligned}
\mathbf{V}_{\alpha i} &= -\frac{D_i^{\text{mix}}}{Y_i}\left[\nabla_\alpha Y_i + \frac{Y_i}{W}\nabla_\alpha W\right] \\
&= -\frac{D_i^{\text{mix}}}{Y_i}\left[\nabla_\alpha Y_i - Y_i W \sum_{j=1}^{N_s} \frac{\nabla_\alpha Y_j}{W_j}\right]. \qquad (19)
\end{aligned}$$

Studies on the effects of thermal diffusion suggest that the Soret effect is much more important for premixed flames than for nonpremixed flames, the Dufour effect is of little importance in either premixed or nonpremixed flames [**?**].

7.2.3 Heat Flux

The heat flux vector is comprised of three components representing the diffusion of heat due to temperature gradients, the diffusion of heat due to mass diffusion, and the Dufour effect [?, **?**, ?, ?, ?]. In most combustion simulations, the Dufour effect is neglected, and (12) may be written as

$$\mathbf{q}_\alpha = -\lambda \nabla_\alpha T + \sum_{i=1}^{N_s} h_i \mathbf{J}_{\alpha i}. \qquad (20)$$

All treatment here will be restricted to adiabatic systems. While it is certainly true that radiation and other heat-loss mechanisms are important in many combustion applications, this complication will not be considered here.

8 Code Overview

The structure and flow of the code S3D is described in this section and illustrated in figure 3. The S3D code is structured to either execute in the run mode or postprocessing mode. In the run mode, the code integrates the governing equations forward in time based on a case specific initialization of the primitive variables. In this mode, all required operations are directed by the routine solve-driver. In postprocessing mode the code executes with the same processor topology as in the run mode but all required operations are directed by the routine post-driver.

After the initialization of the primitive variables for each time step the convective, diffusive and chemical terms in the conservation equations are updated, once for each of the six stages of the fourth-order accurate explicit Runge-Kutta time advancement solver. The main kernels in this solver where over 95% of the computation occurs are given below:

- Chemistry - Computes chemical reaction rate source terms for species equations. The chemical kinetics data is preprocessed and the code to compute the reaction rates, named as "getrates", is generated by the Chemkin compatible preprocessing utility Autogetrates package . The routines are packaged in a separate module which acts as an interface to the code and abstracts the actual implementation of the reaction rates computation. This will allow us to use different versions of the getrates subroutine targeted at different platforms.

- Transport - Computes molecular transport properties for the species. The properties computed include the viscosity, thermal diffusivity and species mass diffusivities. The code is linked with the transport library which is part of the standard Chemkin suite.

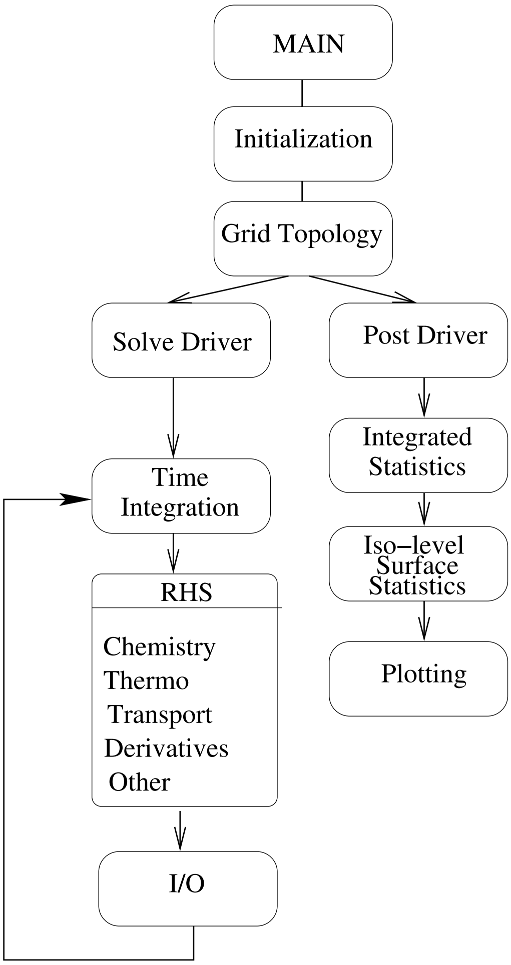

Figure 3: Program flow diagram for S3D

- Thermodynamics - Computes the thermodynamic properties such as enthalpy and specific heats of the mixture. The thermodynamic data are given in the Chemkin compatible format and are preprocessed through the chemkin interpreter. Rather than directly evaluate the properties using the chemkin routines, the code employs a tabulation and lookup strategy.
- Derivatives - Computes the spatial derivatives of the primitive and conserved variables using higher order finite difference operators. The code uses non-blocking sends and receives to exchange the data at the processor boundaries among different processors.
- Other RHS - The right hand side of the time advance equation involves all of the above mentioned operations and the convection terms. These terms are summed up according to the governing equations. All operations involved in this procedure are lumped into the Other RHS module for accounting purposes.
- Time Integration - Advances the solution in time using a 4th order accurate Runge-Kutta scheme. This module also includes an error controller which routinely checks for the time accuracy of the solution and adjusts the time step to achieve the desired error tolerances.

There are four input parameter files necessary to initialize and control the run parameters in S3D. These include:

- s3d.in - All major run parameters are set in this file such as mode, grid dimension parameters (grid points and processors in each direction), run time parameters for IO, geometry parameters (run title, number of dimensions, boundary conditions, and turbulence initialization), physical parameters (domain size, Re, Pr), numerical parameters (order of spatial derivatives, order of RK scheme, order of spatial filtering, frequency to perform various diagnostics), and reference values required for non-dimensionalization.
- Controller.in - Sets information for the timestep controller such as the initial timestep, number of digits of temporal accuracy, minimum timestep, etc.
- Lewis.in - Species Lewis numbers for each species in the reaction mechanism.
- Chem.asc - contains thermodynamic and kinetic information for all species and reactions in Chemkin format.
- Active.in - Generated by S3D at run-time. It is read periodically by S3D during execution and allows the user to steer the computation by changing the parameters in the file such as terminate-status, ending timestep, frequency to write restart files, how often in seconds to write files, how often to write diagnostic information, etc.

9 S3D Software Performance and Improvements

The scientific benefit of the INCITE DNS calculations will be maximized when the code is tuned to require the least amount of computational time per step and grid point. Within the fixed INCITE allocation, this could allow an increased grid number and/or a longer physical simulation time. Such increases would help to achieve the scientific goals by allowing higher Reynolds numbers, a greater sample of turbulent structures from which to take statistics, and/or a more complete temporal development of the turbulent flame. Therefore, the INCITE team, together with NERSC consultant David Skinner, are working towards understanding and improving the performance of S3D on the Seaborg IBM SP platform. Considering that the code has been run on this platform for several years, improvements are expected to be incremental. Also, the INCITE team has been working on porting the code to the Pheonix Cray X1 architecture at ORNL. Due to the substantially different vector architecture, much more significant gains have already been achieved, and further gains may be forthcoming.

The performance of the computational implementation of the DNS software can be measured in terms of the following metrics: (i) Computational time required for a given problem size or larger problem size for a given computational effort (ii) Communication overhead and scaling of a parallel computation over several hundred to several thousand processors (iii) The maximum problem size that can fit onto a machine given the system memory limitations. Performance evaluation and improvements have been divided into the following areas:

- scalar performance evaluation and improvements
- evaluation of parallel scaling and communication overheads
- evaluation of memory limitations

9.1 Test problem description

For the scalar profiling and the parallel scaling studies a pressure wave problem on 40^3 grid points per processor was chosen for simulation. The tests were conducted using detailed CO-H_2 chemistry. The choice of problem size and chemical complexity is representative of the work load associated with the INCITE run. The problem size per processor chosen here, permits the INCITE run to be completed on 1000 to 2000 processors on the IBM SP at NERSC. The initial condition consists of a gaussian temperature profile centered in the domain with non-reflecting boundary conditions. When integrated in time, the initial temperature non-uniformity gives rise to pressure waves and spreading of the temperature profile.

9.2 Scalar profiling and performance improvements

The scalar performance was measured by evaluating the computational time per time step and grid point. The most CPU intensive sections of the code were identified by profiling the code execution on various platforms including Linux Clusters, IBM SP, and CrayX1. Rather than describe exhaustively the code by subroutine, these sections were grouped according to a modular physics based decomposition of the computation as described in section 8.

9.2.1 Performance on Seaborg

The profiling on Seaborg was done using "Xprofiler" which is a GUI based performance profiling tool distributed as part of the IBM Parallel Environment for AIX. It was used to graphically identify the functions which are the most CPU intensive in S3D. It provides results in the form of a call tree as well as a flat profile that details the time spent in each routine. The results of the profiling tool are analyzed and the time spent in each of the subroutines is lumped into one of the several modules described earlier. The profiling results of the original code are shown in Fig. 4. The figure shows a breakdown of the time spent per time-step and grid point for the code evaluated from scalar runs on Seaborg at NERSC. As expected, the code spends most of its time in the chemistry, transport and thermodynamic modules, in that order. Changes were implemented in these three modules as described below.

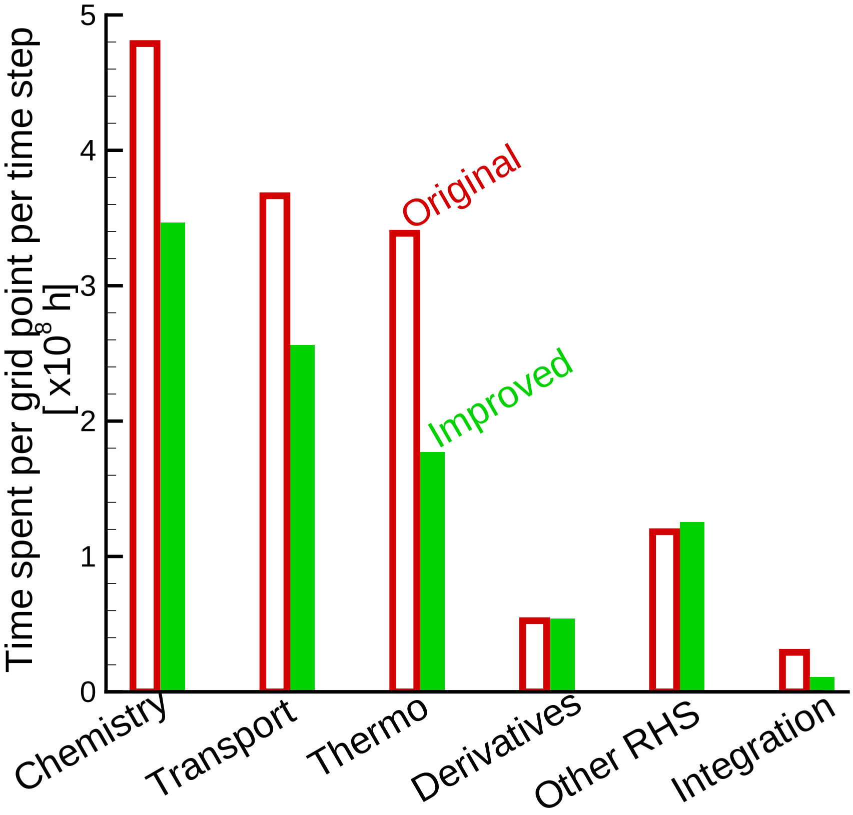

Figure 4: Performance improvements on Seaborg IBM SP at NERSC

1. Most sections of the code use non-dimensional form of the variables and equations to minimize the truncation error. However the transport and chemistry modules are interfaced with the CHEMKIN libraries, which use dimensional quantities in order to be able to use the standard chemical and transport properties databases. The profile showed that a considerable amount of time was being spent by the code in converting the relevant variables between dimensional and non-dimensional units. Several sections of the code were rewritten to minimize the unit conversions and reuse some of the converted data when available. This resulted in around 8% improvement in performance.

2. The computation of chemistry and transport properties involved calls to the exponential and logarithm mathematical functions. To minimize the cost of computing these mathematical functions the code was linked to an accelerated math library written in POWER3 assembly

code that is available on Seaborg. The Mathematical Acceleration SubSystem (MASS) consists of libraries of tuned mathematical intrinsic functions, which offer improved performance over the standard mathematical library routines at the expense of not being as accurate in some cases. The use of these libraries led to a performance improvement of around 10%.

3. The evaluation of the thermodynamic properties involves evaluating polynomials of up to 7th degree in temperature and are very expensive to calculate. Therefore the current thermodynamics module tabulates most of these properties as a function of the temperature. Extracting the properties from this table instead of computing them has proved to be an effective strategy in minimizing the cost of computation. However, only one property, namely the Gibbs energy, continued to be computed instead of tabulated. This property is used in computing the equilibrium constants of reversible reactions in the chemistry module. In the improved version of the code the Gibbs energy of each species is tabulated as a function of temperature, like other thermodynamic properties. The tabulation strategy led to savings in CPU time of around 8%.

The profiling results obtained from the improved version of the code are shown in fig. 4. After these improvements, the code as a whole spends 26% less time on a single processor run. The scalar computing cost was lowered from 1.5x10^{-7}hours/gridpoint/timestep to 1.1x10^{-7} hours/gridpoint/timestep.

Future scalar performance improvements on Seaborg may be possible. These will focus firstly on the evaluation of reaction rates and transport properties. The use of vectorized exponential functions in the MASS library may lead to further gains in the reaction rate evaluation. In transport property evaluations, reorganization of loops in legacy code may lead to more effective compiler optimizations. Using the xprofiler tool, several instances of unnecessary repetition of dynamic allocation of temporary data-structures in chemistry and thermo kernels have been identified. Elimination of these may result in small gains.

9.2.2 Porting the code to Cray X1

The profiling on Phoenix was done using CrayPat. The CrayPat suite of tools do not require the code to be recompiled in order to perform the profiling. The "pat_build" utility is used to instrument the compiled executable with tracing and polling calls and produce a modified executable. The report produced on execution is analyzed using the "pat_report" utility. The improved version obtained on Seaborg as a result of the optimization exercise is used as the starting point for the CRAY experiments. The scalar execution time on Cray was 1.8e-7 hrs/gridpoint/timestep. The profile of executing this code is shown in Fig. 5. It it seen that the code spends a disproportionately long time in the chemistry and thermodynamic modules. These modules were recompiled with the -rm option to obtain a detailed listing file that shows the optimizations performed by the compiler and the sections of code that it was not able to optimize. It was found that the compiler was not able to vectorize the chemistry module. Hence a large portion of the time was spent in evaluating the scalar version of the exponent function. Similarly the thermodynamic module had several functions that involved type constructs and allocatable data objects. The Cray compiler was not able to inline these functions and as a result their callers were not being vectorized.

The problem was rectified by rewriting a significant part of the chemistry module in a form suitable for vectorization. In particular, instead of computing the reaction rates at every grid point separately, a new routine was written to compute the reaction rates all over the domain. This modification made it possible for the compiler to invoke the vector exponent which resulted in a significant improvement. The issue with the thermchem module was resolved by replacing

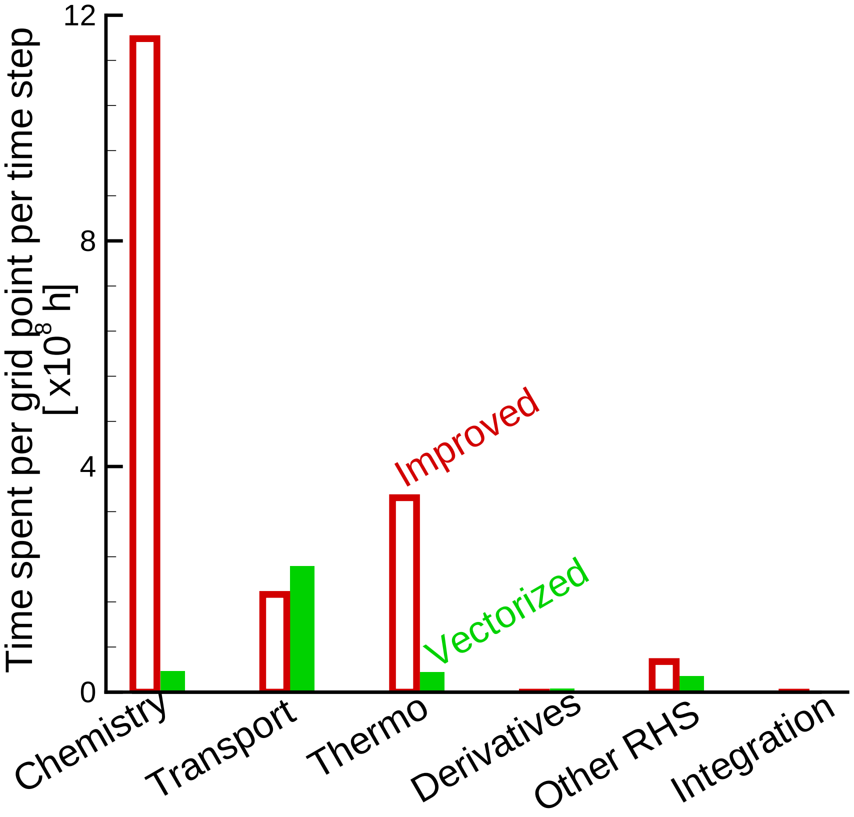

Figure 5: Performance improvements on Pheonix Cray X1 at ORNL

some of the type constructs and allocatable arrays with generic typed variables and static arrays. Furthermore some parts of the code were manually inlined to assist the vectorization.

As a result of these modifications the compute cost went down to 3.55x10^{-8} hours/gridpt/timestep. As seen in Fig. 5 the cost of computing the chemistry and thermodynamics modules are reduced to insignificant levels as a result of the vectorization.

Further improvements on the Cray X1 architecture are likely. The primary candidate is the evaluation of transport properties, which does not vectorize well due to the present structure of loops.

10 Evaluation of parallel scaling and communication overheads

S3D is a mature DNS code and has been demonstrated to scale very well up to 4000 IBM SP processors. Also, the inter-process communications are minimal and exist only between the nearest neighbors in the processor domain topology. Hence, communication was expected not to be a bottleneck for the performance of the code.

10.1 Scaling on Seaborg

A series of comparison runs of S3D on 1, 8, 64, 256, and 512 processors of increasing total problem size proportional to the number of processors were done on Seaborg. This comparison was between the S3D code as it started out on Seaborg and the code as of Q1 2005. The Integrated Performance Monitoring (IPM) tool was deployed to analyze the communication overhead and scaling performance. IPM is a portable profiling infrastructure which provides users with a concise report on the execution of parallel jobs. The IPM reports, generated by David Skinner, are available at: `http://www.nersc.gov/~dskinner/tmp_s3d/`. The files are named s3d_orig_N or s3d_inciteQ1_N, where “orig” refers to the original code, “inciteQ1” refers to the new code, and N is the number of tasks for the particular run.

A great deal of information is contained in the IPM reports. The main points are summarized as below.

- The code is scaling well. Figure 6 shows the total wall-clock time versus the number of tasks. Aside from one outlier at 256 tasks the new code shows consistently better performance approaching 14%.
- The code scaling is not communication bound showing only 10-20% communication time.
- The communication topology looks to be well blocked as seen in the lower part of the IPM reports. There may be a more effective task ordering that could lead to more SMP vs. switch traffic, but gains are not expected to be significant.
- The amount of time spent in MPI_Barrier is sometimes an appreciable fraction of the communication time. This suggests that load balance may be improved to some small extent. The first step toward this has been completed, namely removing most of the unnecessary barriers. While this does not improve performance directly it does expose the tasks which are blocked. Discussions have been started about how the layout of processes assigned to Seaborg nodes might be optimized.

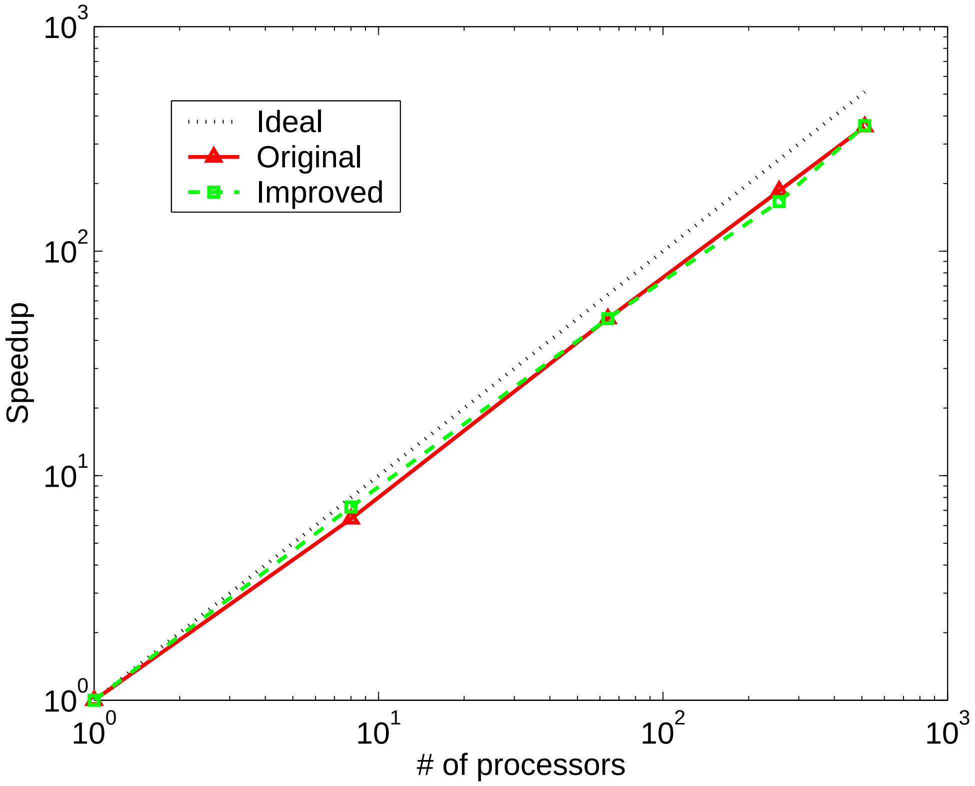

Figure 6: Code scaling on IBM SP

10.2 Scaling on Cray X1

Preliminary scaling tests on 1,8,64, and 256 processors have been performed on the Cray X1. Figure 7 shows the speed-up of the code against the number of processors. As in the Seaborg tests, the total problem size is increased proportional with the number of processors. The code is also scaling reasonably well on this platform. The appearance of better scaling of the “improved” code is due to the higher ratio of computation to communication per processor. Further work is needed to evaluate whether any scaling improvements can be obtained on this architecture.

11 Evaluation of memory limitations

Due to the large size of the INCITE calculation, it is necessary to allow a larger problem size per task than usual. On Seaborg, it was initially found that the problem size was memory limited to approximately 90000 grid points per task. However, this difficulty was quickly remedied by utilizing the appropriate compiler flag, -bmaxdata, allowing the use of up to nearly 3 million grid points per task.

In order to decide the precise parameter space for the INCITE calculation, it is necessary to run many smaller test calculations, which are in themselves computationally very demanding. The INCITE team is employing a new local CRF Opteron Linux cluster with Infiniband switches for the test calculations. On this machine, which has a very high processor speed, the code was found to be memory limited. The memory usage was analyzed using a tool named Valgrind. Based on the results of the analysis, several unnecessary arrays were eliminated and different sections of the

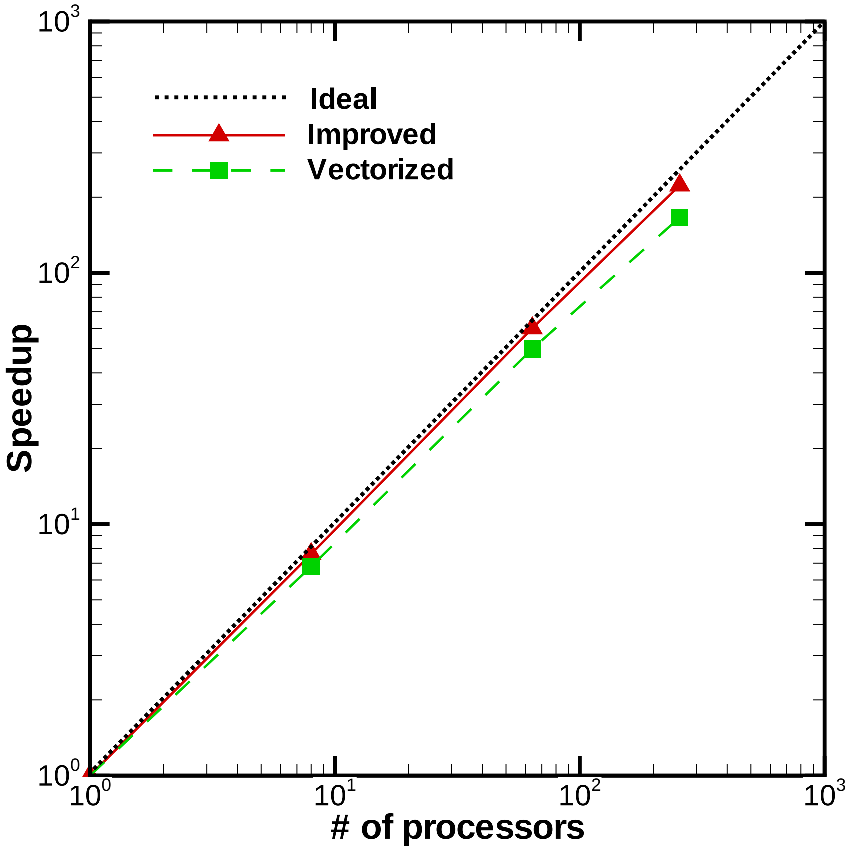

Figure 7: Code scaling on Cray X1

code were made to share the same memory space for their operations. Such improvements reduced the memory footprint of the code by approximately 25-30 %, allowing a equivalent increase in the size, and therefore relevance, of the test problems.

References