

Using SIMD and SIMT vectorization to evaluate sparse chemical kinetic Jacobian matrices and thermochemical source terms

Nicholas J. Curtis^{a,*}, Kyle E. Niemeyer^b, Chih-Jen Sung^a

^a*Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA*

^b*School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, OR 97331, USA*

Abstract

Accurately predicting key combustion phenomena in reactive-flow simulations, e.g., lean blow-out, extinction/ignition limits and pollutant formation, necessitates the use of detailed chemical kinetics. The large size and high levels of numerical stiffness typically present in chemical kinetic models relevant to transportation/power-generation applications make the efficient evaluation/factorization of the chemical kinetic Jacobian and thermochemical source-terms critical to the performance of reactive-flow codes. Here we investigate the performance of vectorized evaluation of constant-pressure/volume thermochemical source-term and sparse/dense chemical kinetic Jacobians using single-instruction, multiple-data (SIMD) and single-instruction, multiple thread (SIMT) paradigms. These are implemented in pyJac, an open-source, reproducible code generation platform. Selected chemical kinetic models covering the range of sizes typically used in reactive-flow simulations were used for demonstration. A new formulation of the chemical kinetic governing equations was derived and verified, resulting in Jacobian sparsities of 28.6–92.0 % for the tested models. Speedups of $3.40\text{--}4.08\times$ were found for shallow-vectorized OpenCL source-rate evaluation compared with a parallel OpenMP code on an `avx2` central processing unit (CPU), increasing to $6.63\text{--}9.44\times$ and $3.03\text{--}4.23\times$ for sparse and dense chemical kinetic Jacobian evaluation, respectively. Furthermore, the effect of data-ordering was investigated and a storage pattern specifically formulated for vectorized evaluation was proposed; as well, the effect of the constant pressure/volume assumptions and varying vector widths were studied on source-term evaluation performance.

Speedups reached up to $17.60 \times$ and $45.13 \times$ for dense and sparse evaluation on the GPU, and up to $55.11 \times$ and $245.63 \times$ on the CPU over a first-order finite-difference Jacobian approach. Further, dense Jacobian evaluation was up to $19.56 \times$ and $2.84 \times$ times faster than a previous version of `pyJac` on a CPU and GPU, respectively. Finally, future directions for vectorized chemical kinetic evaluation and sparse linear-algebra techniques were discussed.

Keywords: Chemical Kinetics, SIMD, SIMT, Sparse, Jacobian

1. Introduction

As the combustion and reactive-flows community has recognized the importance of detailed chemical kinetics for predictive reactive-flow simulations [1], chemical kinetic models have grown in size and complexity to describe current and next-generation fuels relevant to transportation and power generation. For example, a recent biodiesel model [2] consists of ~ 3500 chemical species and over 17,000 reactions. Moreover, the cost of evaluating the chemical source-terms scales linearly with the size of the model, while evaluating and factorizing the chemical kinetic Jacobian respectively scale quadratically and cubically with the number of species in the model, if naively implemented via a finite-difference method [1]. These factors often prohibit using detailed chemical kinetics in practice; e.g., in a direct numerical simulation using a 22-species model, evaluating reaction rates consumed around half of the total run time [3]. In addition, most common implicit integration techniques need to evaluate and factorize the Jacobian matrix to deal with stiffness. As a result, these operations are bottlenecks when using even moderately sized chemical models in realistic reactive-flow simulations, necessitating other cost-reduction strategies [1].

A host of techniques have been developed to lessen the computational demand of chemical kinetic calculations while maintaining fidelity, falling broadly into three categories: removal of unimportant species and reactions [4–8], lumping of species with similar thermochemical properties [9–11], and time-scale methods that reduce numerical stiffness [12–15]. We refer

*Corresponding author

Email address: `nicholas.curtis@uconn.edu` (Nicholas J. Curtis)

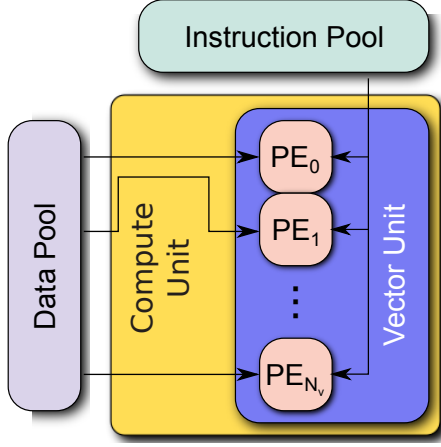
interested readers to the recent review by Turányi and Tomlin [16] for a comprehensive overview.

In addition to the previously mentioned cost reduction methods, effort has gone into improving the integration algorithms and codes that evaluate the chemical kinetics [15, 17–19]. In particular, a carefully derived analytical formulation of the Jacobian matrix can greatly increase sparsity [17] and drop the cost of Jacobian evaluation to linearly depend on the number of species in the model [1]; sparse-matrix techniques can then reduce the cost of Jacobian factorization [20]. In addition, studies have shown that Single-Instruction, Multiple-Data (SIMD) and the related Single-Instruction, Multiple-Thread (SIMT) processors can accelerate chemical kinetic simulations [18, 21–25].

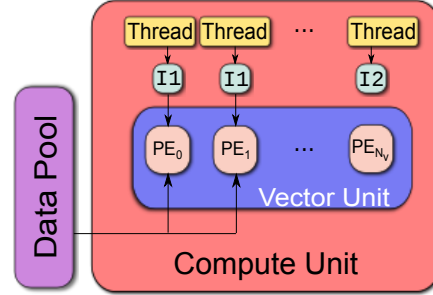
SIMD and SIMT programming are two important vector-processing paradigms used increasingly in scientific computing. Traditional multicore parallelism is now used to increase central processing unit (CPU) performance, as the exponential growth in processing power—colloquially known as Moore’s law—has slowed [26]. Recently, SIMD/SIMT processors, e.g., in the form of graphics processing units (GPUs), have gained recognition due to their increased floating operation throughput. The parallel programming standard OpenCL [27] has further enabled adoption of vector processing in scientific computing by providing a common application program interface (API) for execution on heterogeneous systems, e.g., CPU, GPU, or Intel’s Many Integrated Core (MIC) architecture. Here we will largely use OpenCL terminology to describe these processing paradigms, as it provides a convenient way to classify otherwise disparate processor types (e.g., CPUs and GPUs). However, the concepts discussed herein broadly apply to SIMD/SIMT processing.

A typical modern CPU contains multiple compute units (i.e., cores), each with specialized vector processing units capable of running SIMD instructions, as Fig. 1a depicts. A SIMD instruction uses the vector processor to execute the same floating-point operation (e.g., multiplication, division) on different data concurrently. The vector-width is the number of possible concurrent operations, typically around two to four in double precision.¹ Specialized hardware accelerators have also been developed, like Intel’s Xeon Phi co-processor (i.e., the MIC architecture), that have tens of cores with

¹OpenCL allows for use of vector-widths different from the actual hardware vector-width via implicit conversion, and may provide some performance benefit as Sec. 3.5 discusses.



(a) Schematic of SIMD processing. A single compute unit (e.g., a CPU core) contains a vector unit with N_v processing elements (PEs), together called a vector-lane. The vector unit executes a single instruction concurrently on multiple data.



(b) Schematic of SIMT processing. A single compute unit (e.g., a GPU streaming multiprocessor) contains many processing elements (PEs) and hosts many threads, each with an instruction to execute (I1, I2). Threads with the same instruction execute concurrently on multiple data while the others must wait (leading to thread divergence).

Figure 1: Simple diagrams explaining the fundamentals of the SIMD and SIMT vector-processing paradigms.

wide vector-widths (e.g., 4–8 double-precision operations). Cutting-edge and forthcoming Intel CPUs also include these wide vector-widths, like the Skylake Xeon and Cannon Lake architectures.

Modern GPUs rely on the related computing paradigm of SIMT processing, where a single compute element hosts large numbers of threads (a streaming multiprocessor in Nvidia terminology) [28]. Figure 1b depicts a SIMT compute unit, where a group of threads—typically 32, known as a warp on Nvidia GPUs—execute the same SIMT instruction on multiple data concurrently. If some threads must execute a different instruction, they are forced to wait and execute later; this may occur due to if/then branching or predication. This phenomenon, known as thread-divergence, is a key consid-

eration for SIMT processing and can cause serious performance degradation for complicated algorithms [24].

1.1. Related work

Recognizing the need to accelerate chemical-kinetic Jacobian evaluation and factorization, a number of recent works have been published on constructing analytical Jacobian matrices; although as will be discussed at the end of this section, here we offer several key improvements over past efforts. Schwer et al. [17] were among the first to recognize the critical importance of a sparse analytical Jacobian to accelerate chemical kinetic simulations. Later, Safta et al. [29] developed the **TChem** software package, which was one of the first developed that provides analytical Jacobian evaluation. However, **TChem** has several limitations, including incompatibility with modern reaction types—i.e., pressure-dependent Arrhenius (or P-Log) and Chebyshev reactions—and its lack of thread-safety to enable parallel execution [30]. Youssefi [31] explored the importance of analytical Jacobian matrices for time-scale analysis techniques as well as their effect on computational efficiency in zero-dimensional homogeneous reactor simulations. Bisetti [32] developed an isothermal, isobaric analytical Jacobian code-generation utility; this approach significantly increases Jacobian sparsity, although the chosen isothermal assumption is not typical in most combustion simulations. In the same work Bisetti also provided a novel way to compute dense matrix-vector multiplications resulting from a change of system variables without storing the full Jacobian. Perini et al. [33] developed an analytical Jacobian code for constant-volume combustion, with additional options to increase sparsity (at the expense of strict correctness) and tabulate temperature-dependent properties; they reported an 80 % speedup over a finite-difference-based Jacobian when used in a multidimensional reactive-flow simulation. Gao et al. [19] derived a sparse analytical Jacobian, but did not verify it outside the context of use with an implicit-integration technique. In addition, since the Jacobian was based on an over-constrained system [34], the effect on strict conservation of mass/energy was not studied.

Recently, some groups have developed frameworks for constructing analytical Jacobians for evaluation on modern SIMD or SIMT processors. Dijkmans et al. [35] developed a GPU-based analytical Jacobian code with optional tabulation of temperature-dependent properties, and showed speedups up to $120\times$ for zero-dimensional chemical kinetic integration with large chemical models (~ 3000 species). Bauer et al. [36] used warp-specialization

to improve GPU-vectorization over a standard data-parallel vectorization approach; they achieved speedups of up to $2.81\text{--}3.75\times$, $1.91\text{--}2.58\times$, and $1.4\text{--}1.5\times$ for evaluating viscosity, species diffusion, and chemical source terms, respectively. Niemeyer et al. [18] created and verified the open-source analytical chemical kinetic Jacobian code-generator, `pyJac`, which supports parallel execution on CPUs and SIMT execution on GPUs; `pyJac` enables a speedup of $3\text{--}7.5\times$ over a finite-difference Jacobian on the CPU.

Relevant to all of the aforementioned efforts, Hansen and Sutherland [34] explored the choice of thermochemical state vectors and the resulting effect on consistency and errors in conserved properties such as mass and energy. They also characterized how the choice of state vector affects implicit/linearly implicit integration algorithms and chemical mode analysis techniques. Overall they found that while many literature Jacobian formulations are not strictly correct or over-specified, such flaws negligibly affect Newton–Krylov methods—perhaps because the incorrect Jacobian reasonably approximates the true Jacobian. On the other hand, linearly implicit algorithms like Rosenbrock methods and analysis techniques like chemical explosive mode analysis [37] need accurate and correct Jacobians.

A number of recent works have investigated using high-performance SIMT devices like GPUs to accelerate reactive-flow and chemical kinetics simulations. Spafford et al. [3] coupled GPU-based chemical source-term evaluation with an explicit direct numerical simulation code, achieving an order of magnitude speedup compared to a CPU-based serial implementation. Shi et al. [38] combined GPU-based chemical kinetic source-term evaluation and Jacobian factorization with two implicit CPU solvers, achieving an order-of-magnitude speedup for homogeneous reactor simulations of large chemical models over a serial CPU implementation. Niemeyer et al. [39] implemented an explicit solver for non-stiff chemistry on a GPU, achieving a speedup of nearly two orders of magnitude over a sequential CPU code. Shi et al. [21] proposed a strategy for chemical-kinetic integration in three-dimensional reactive-flow simulations, where a traditional implicit integrator handled the stiffest computational cells on a CPU and a stabilized-explicit solver solved the less-stiff cells on a GPU; this hybrid solution technique performs $11\text{--}46\times$ faster than the implicit CPU solver alone for simulation of a premixed diesel engine. Le et al. [40] found a $30\text{--}50\times$ speedup for a GPU-based shock-capturing reactive-flow code as compared with a sequential CPU version of the same. Stone and Davis [41] investigated a GPU-based version of a common implicit integrator (VODE [42]), finding an order-of-magnitude speedup

over a serial CPU implementation. Niemeyer and Sung [22] developed a GPU-based stabilized explicit integrator for use with moderately-stiff chemical kinetics, achieving an order of magnitude speedup over a multithreaded VODE solver on a six-core CPU. Sewerin and Rigopoulos [23] studied a fifth-order implicit Runge–Kutta solver on both consumer-grade and high-end GPUs/CPU; the high-end GPU solver was at best $1.8 \times$ slower than the high-end CPU version running on 16 cores. Yonkee and Sutherland [43] implemented accelerated evaluations of thermodynamic parameters, multi-component transport properties, and species production rates on both the CPU and GPU, achieving speedups over serial evaluation between $8\text{--}13 \times$ on a 16-core CPU and $20\text{--}40 \times$ on the GPU. In addition, $\sim 9 \times$ and $\sim 25 \times$ speedups were achieved for the simulation of a partially premixed methanol flame for solving partial differential equations (PDE) on 16 CPU cores and the GPU, respectively. Curtis et al. [24] implemented a fifth-order implicit Runge–Kutta method [44], as well as two fourth-order exponential integration techniques [45, 46] paired with an analytical Jacobian code [18] on the GPU and CPU. The GPU-based implicit Runge–Kutta method performed equivalently to a standard implicit integrator [47] running on 12–38 CPU cores for two relatively small chemical models with an integration time step of 10^{-6} s.

In contrast, SIMD-based chemical kinetics evaluation/integration have been studied far less. Linford et al. [48] implemented a three-stage, second-order Rosenbrock integrator for atmospheric chemical kinetics on the CPU, GPU, and cell broadband engine (CBE)—a specially designed vector processor—and found speedups regularly exceeding $25 \times$ over a serial CPU implementation. Kroshko and Spiteri [49] implemented a SIMD-vectorized third-order stiff Rosenbrock integrator for atmospheric chemistry on the CBE and found a speedup of $1.89 \times$ (a parallel scaling efficiency of 94 %) over a serial version of the same code. Stone et al. [25] implemented a linearly implicit fourth-order stiff Rosenbrock solver in the OpenCL for various platforms including CPUs, GPUs, and MICs. They found that SIMD vectorization improves integrator performance over an OpenMP baseline vectorized by simple compiler hints (i.e., `#pragmas`) by $2.5\text{--}2.8 \times$ on the CPU and $4.7\text{--}4.9 \times$ on the MIC, while the GPU performs only $1.4\text{--}1.6 \times$ faster than the OpenMP baseline due to thread divergence [25].

1.2. Goals of this study

In this article we

- Derive and verify a new Jacobian formulation that greatly increases sparsity;
- Detail the implementation of cross-platform SIMD/SIMT vectorization for CPUs, GPUs, and other accelerators;
- Investigate the performance of SIMD/SIMT-vectorization for a wide range of chemical kinetic models, and compare with the previous version of our analytical chemical kinetic Jacobian code [18]; and finally
- Discuss future extensions to this work as well as several promising directions for SIMD/SIMT vectorization in reactive-flow simulations.

This work builds upon our previous analytical chemical kinetic Jacobian code, `pyJac` [18], using the new formulation, `pyJac v2`, to achieve these goals. To our knowledge is the first open-source, verified effort that vectorizes the evaluation of chemical-kinetic source terms and Jacobian matrices for any chemical model on a wide selection of platforms.

2. Methodology

2.1. Data ordering and vectorization patterns

(1,1)	(1,2)	...	(1,N)
(2,1)	(2,2)	...	(2,N)
⋮		⋱	
(K,1)	...		(K,N)

(a) A simple 2-D data array with K rows and N columns.

(1,1)	(1,2)	...	(1,N)	(2,1)	...	(K,N)
-------	-------	-----	-------	-------	-----	-------

(b) Row-major data ordering

(1,1)	(2,1)	...	(K,1)	(1,2)	...	(K,N)
-------	-------	-----	-------	-------	-----	-------

(c) Column-major data ordering

Figure 2: Simple data-layout patterns for 2-D arrays

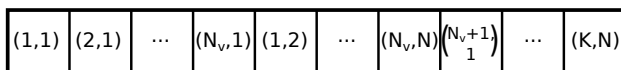
When storing arrays for a chemical kinetic model, the data-storage layout and vectorization patterns are critical to achieving high-performance code. Figure 2a depicts an example data array with K rows and N columns where index (i, j) corresponds to the i th row and j th column. For example, the concentration of species j for the i th thermochemical state would be stored in $[C]_{i,j}$ with $1 \leq i \leq N_{\text{state}}$ (the number of thermochemical states considered for evaluation) and $1 \leq j \leq N_{\text{sp}}$ (the number of species in the model). The stored concentrations would then have $K = N_{\text{state}}$ rows and $N = N_{\text{sp}}$ columns.

The “C” (C row-major) format stores the concentrations of all species for a single thermochemical condition i sequentially in memory, i.e., with $[C]_{1,1}$ in index 1 (using one-based index notation), $[C]_{1,2}$ in index 2, and so on, as shown in Fig. 2b. Conversely, in the “F” (Fortran column-major) format the concentrations of a single species j over all thermochemical states lie adjacent in memory, corresponding to storing $[C]_{1,1}$ in index 1, $[C]_{2,1}$ in index 2, and so on, as shown in Fig. 2c. This ordering strongly affects the performance of SIMD/SIMT-vectorized algorithms, as does the device (CPU, GPU, etc.) and vectorization pattern in question.

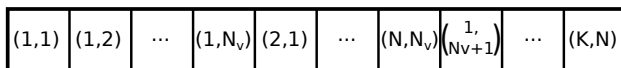
In a *shallow* SIMD/SIMT vectorization (also referred to as “per-thread” in previous works using GPUs [41]), each SIMD lane or SIMT thread in a compute unit evaluates the source terms or Jacobian for a different thermochemical state. If the data is stored in “F”-order, the SIMD lanes/SIMT thread accessing $[C]_{1,j} \dots [C]_{N_v,j}$ will load sequential locations in memory, where $[C]_{i,j}$ is the concentration of species j for state i and N_v is the SIMD vector-width or the number of threads in a SIMT warp. The first $(j + 1)$ th species concentration, $[C]_{1,j+1}$, will be N_{state} memory locations away; this increases the likelihood of cache misses on the CPU [50], but conversely well matches the pattern of coalesced memory access on the GPU [51].

In a *deep* SIMD/SIMT vectorization (also referred to as “per-block” in previous GPU works [24, 41]), a compute unit uses its SIMD lanes/SIMT threads cooperatively to evaluate the thermochemical source terms for a single thermochemical state; thus SIMD lanes loading $[C]_{1,j} \dots [C]_{1,j+N_v}$ will access sequential memory locations if the data is stored in “C”-order. Further, in “C” ordering any two species concentrations within the same thermochemical state lie at most N_{sp} locations away, with $N_{\text{sp}} \ll N_{\text{state}}$ in most cases; this greatly improved data locality increases the chances of a cache hit on the CPU, but may lead to uncoalesced memory accesses on the GPU. Deep vectorization requires synchronization between SIMD lanes/SIMT threads via

memory fences/barriers, a potentially expensive operation. In addition, deep vectorization may result in SIMD waste or SIMT thread divergence caused by different lanes/threads executing different instructions (e.g., resulting from different if/then branches). Shallow vectorization may also experience SIMD waste or SIMT thread divergence, e.g., in chemical kinetic integration due to varying internal solver time-step sizes [24]. However, in this work shallow vectorization is largely unaffected by this concern as the only major code paths that differ between vector lanes are high/low-temperature polynomial evaluations and differing pressures for P-Log reactions, which cause far fewer issues compared with differing internal ODE integration time-steps [24].



(a) Row-major, shallow-vectorized data ordering



(b) Column-major, deep-vectorized data ordering

Figure 3: Vectorized data-ordering patterns

Finally, Fig. 3 shows a vectorized data-ordering that improves the caching patterns of a shallow, “C”-ordered SIMD vectorization on the CPU (Fig. 3a) and a deep, “F”-ordered SIMT vectorization on the GPU (Fig. 3b). We accomplish this by splitting the slower-varying axis of the data array—columns for “C”-ordering, and rows for “F”-ordering—into chunks of size N_v (the SIMD vector width or SIMT warp size) and laying these data out contiguously in memory. For example, using the shallow-vectorized “C”-ordering pictured in Fig. 3a, the concentrations of species j for states i to $i + N_v$ ($[C]_{i,j}, \dots, [C]_{i+N_v,j}$) lie contiguously in memory and are followed by the concentrations of species $j + 1$ for the same states ($[C]_{i,j+1}, \dots, [C]_{i+N_v,j+1}$). This pattern ensures that any SIMD operation occurs on data contiguous in memory, which greatly improves caching and SIMD throughput; it is also similar to OpenCL’s native vector data-types, e.g., `double8` treats eight contiguous double-precision floating-point numbers as a single vector datum. Conversely, the data-ordering in Fig. 3b enables coalesced memory accesses for “F”-ordered, deep SIMT vectorization on the GPU. We will discuss the

effects of these various data-ordering and vectorization patterns on performance in Section 3.5.

2.2. Thermochemical source terms and Jacobian

This new version of `pyJac` is capable of evaluating the thermochemical source-terms for using the constant-pressure (CONP) or constant-volume (CONV) assumption² In this section, we will outline a brief summary of the system evaluated by `pyJac`; the supplemental material contains the complete—and lengthy—derivations.

The thermochemical state vector consists of the temperature, a non-constant thermodynamic state parameter (volume or pressure for CONP and CONV, respectively), and the number of moles of all species except the last species in the chemical model, typically taken to be the bath gas (e.g., N_2):

$$\Phi = \{T, V, n_1, n_2 \dots n_{N_{\text{sp}}-1}\} \quad \text{for CONP,} \quad (1a)$$

$$\Phi = \{T, P, n_1, n_2 \dots n_{N_{\text{sp}}-1}\} \quad \text{for CONV,} \quad (1b)$$

where T is the temperature, V and P the volume and pressure respectively, and n_j the number of moles of the j th species in the model (containing N_{sp} total species).

This state vector—inspired by Schwer et al. [17]—has a number of beneficial features. First, the state vector results in highly sparse chemical kinetic Jacobians, as will be detailed in Section 3.4. Second, this formulation explicitly conserves mass, because the number of moles and rate of change of the final species are calculated from the ideal gas law and conservation of mass, respectively; see the supplemental material for the full details of the governing equations. The system is not over-constrained [34] and does not require use of a more-complicated differential algebraic equation solver (as compared to an ODE integrator) for integration. Finally, the chemical kinetic Jacobian for this formulation changes relatively little between the CONP and CONV forms, making maintaining the codebase much simpler. Although most current combustion codes do not use species moles as a state variable, conversion to/from the more-common mass/mole fractions and moles is straightforward,

²Note: in this context, the “constant-pressure” and “constant-volume” assumptions refer to evaluation within a reaction sub-step in the operator splitting scheme, rather than a general constant-pressure or constant-volume reactive-flow simulation.

and the choice of variables no longer matters once inside the integration of an chemical kinetic initial-value problem (IVP).

The evolution of the thermochemical state vector is described by a set of chemical kinetic ordinary differential equations:

$$f = \frac{d\Phi}{dt} = \left\{ \frac{dT}{dt}, \frac{dV}{dt}, \frac{dn_1}{dt}, \frac{dn_2}{dt} \dots \frac{dn_{N_{\text{sp}}-1}}{dt} \right\} \quad \text{for CONP}, \quad (2a)$$

$$f = \frac{d\Phi}{dt} = \left\{ \frac{dT}{dt}, \frac{dP}{dt}, \frac{dn_1}{dt}, \frac{dn_2}{dt} \dots \frac{dn_{N_{\text{sp}}-1}}{dt} \right\} \quad \text{for CONV}. \quad (2b)$$

For both CONP and CONV, the molar source terms are [52]:

$$\frac{dn_k}{dt} = V\dot{\omega}_k \quad k = 1, \dots, N_{\text{sp}} - 1, \quad (3)$$

where $\dot{\omega}_k$ is the k th species' overall molar production rate:

$$\dot{\omega}_k = \sum_{i=1}^{N_{\text{reac}}} \nu_{k,i} R_i c_i, \quad (4)$$

$\nu_{k,i}$ is the net stoichiometric coefficient of species k in reaction i , N_{reac} is the total number of reactions, R_i is the net rate of progress of reaction i , and c_i is the pressure-dependent modification term, i.e., for third-body or falloff/chemically-activated reactions. `pyJac` is capable of evaluating all modern reaction types, e.g., P-Log and Chebyshev reactions.

The temperature source-term [52] is:

$$\frac{dT}{dt} = - \frac{\sum_{k=1}^{N_{\text{sp}}} H_k \dot{\omega}_k}{\sum_{k=1}^{N_{\text{sp}}} [C]_k C_{p,k}} \quad \text{for CONP}, \quad (5a)$$

$$\frac{dT}{dt} = - \frac{\sum_{k=1}^{N_{\text{sp}}} U_k \dot{\omega}_k}{\sum_{k=1}^{N_{\text{sp}}} [C]_k C_{v,k}} \quad \text{for CONV}, \quad (5b)$$

where H_k , U_k , $C_{p,k}$, and $C_{v,k}$ are the enthalpy, internal energy, constant-pressure specific heat, and constant-volume specific heat of species k in molar units, respectively, while $[C]_k$ is the concentration, given by

$$[C]_k = \frac{n_k}{V}. \quad (6)$$

By differentiating the ideal gas law, given by

$$PV = n\mathcal{R}T \quad (7)$$

where $\mathcal{R}$ is the ideal-gas constant in molar units, we find the volume and pressure source terms (where W_k and $W_{N_{\text{sp}}}$ are the molecular weights of species k and N_{sp} , respectively):

$$\frac{dV}{dt} = V \left(\frac{T\mathcal{R}}{P} \sum_{k=1}^{N_{\text{sp}}-1} \left(1 - \frac{W_k}{W_{N_{\text{sp}}}} \right) \dot{\omega}_k + \frac{1}{T} \frac{dT}{dt} \right) \quad \text{for CONP,} \quad (8a)$$

$$\frac{dP}{dt} = T\mathcal{R} \sum_{k=1}^{N_{\text{sp}}-1} \left(1 - \frac{W_k}{W_{N_{\text{sp}}}} \right) \dot{\omega}_k + \frac{P}{T} \frac{dT}{dt} \quad \text{for CONV.} \quad (8b)$$

`pyJac` arranges the computed Jacobian entries such that entry (i, j) corresponds to the partial derivative of the i th source-term in Eq. (2) by the j th state variable in Eq. (1):

$$\mathcal{J}_{i,j} = \frac{\partial f_i}{\partial \Phi_j}, \quad i, j = 1 \dots N_{\text{sp}} + 1. \quad (9)$$

The supplemental material for this article contains the complete derivation of the Jacobian used by `pyJac`, for interested readers.

2.3. Code generation and testing infrastructure

The new version of `pyJac` uses the Python package `loo.py` [53] for code generation, which translates pseudo-code and data to OpenCL/C code. As the name implies, `loo.py` generates code using for loops; this differs from the previous version of `pyJac` [54] that generates static code—i.e., fully unrolled loops, with thermodynamic/reaction parameters written directly in code rather than stored in arrays. In our previous work [18], this static code generation caused some issues with large file sizes, long compilation times, and even occasionally broke the `gcc` and `nvcc` compilers (the latter issue necessitated splitting the Jacobian/source-term evaluations into separate files). We will discuss the implications of this change in Section 3.5.2, where the performance of the new version of `pyJac` will be compared with the previous version.

In addition, `loo.py` allows the user to more easily make changes to the structure of the generated program, e.g., the data ordering, vectorization, and threading patterns, as well as switch the target language for code generation (and more simply extend to additional languages, e.g., CUDA). Further, `loo.py` can execute developed subroutines from Python (natively for C code, or via `PyOpenCL` [55] for OpenCL), enabling unit testing/verification

for each component of the Jacobian or source terms; the unit testing suite also helps ensure that bugs are not present in less commonly used code-paths, or are introduced by future code changes. The source terms and sub-components thereof (e.g., rates of progress, pressure-modification terms) are directly compared with Cantera [56], while the automatic differentiation code Adept [57, 58] provides reference values for Jacobian sub-components. We use the Portable OpenCL (POCL) implementation [59] and OpenMP [60] to perform OpenCL and C unit testing, respectively, on the continuous-integration framework Travis CI [61]. We will discuss verification of the complete (as opposed to the sub-component testing discussed here) generated source-terms and Jacobian codes in detail in Section 3.2.

3. Results and discussion

3.1. Testing platforms

CPU Model	Xeon X5650	E5-2690 V3
Instruction Set	SSE4.2	AVX2
Vector Width	two doubles	four doubles
Cores	2×6	2×12
Identifier	sse4.2	avx2
OpenCL Version	1.2	1.2

Table 1: The Intel CPU configurations used in this study. The vector widths are reported in (ideal) number of double operations per SIMD instruction, as this will be used in measuring SIMD efficiency; for reference, the vector widths of the **sse4.2** and **avx2** machines are 128 bit and 256 bit, respectively. The identifier field will be used as a shorthand descriptor in the performance plots to quickly identify the CPU type.

We ran the performance and verification studies for this work on a variety of CPU and GPU platforms. Table 1 shows the number of cores, vector instruction set, and model of the CPUs used in this work; each CPU had both **v16.1.1** of the Intel OpenCL runtime [62] and **v1.0** of the POCL [59] runtime installed, both enabling OpenCL **v1.2** execution. Additionally, **v5.0** of the LLVM/**clang** [63] compiler chain was installed on all machines to enable use of POCL. Table 2 lists the model, number of CUDA cores, and Nvidia driver of each GPU we used. Nvidia’s OpenCL runtime is bundled with the Nvidia driver [51], hence the driver version is used to specify the OpenCL runtime version.

Nvidia Model	Tesla C2075	Tesla K40m
Driver Version	384.81	387.26
CUDA Cores	448	2880
Identifier	C2075	K40m
OpenCL Version	1.1	1.2
Memory ³	6 GB	12 GB

Table 2: The Nvidia GPU configurations used in this study. Nvidia’s OpenCL runtime is provided with the graphics driver, rather than any specific version of CUDA. The identifier field will be used as a shorthand descriptor during analysis of the performance results.

Table 3 lists the platforms and vectorization/execution patterns that they are capable of running. The Intel and Nvidia OpenCL runtimes lack implementations of atomic operations on double-precision variables; `pyJac` currently needs these to run deep-vectorized code. On the other hand, POCL is an open-source OpenCL runtime that works on all CPU types tested here, and does implement these atomic operations. However, POCL’s implicit vectorization module—which uses the LLVM compiler [63] to translate OpenCL code to vectorized machine code—typically fails to achieve much, if any, speedup. Thus POCL is useful for verification but not necessarily for performance studies; it is noted that while POCL is currently used by `pyJac` for unit-testing purposes, it is not required to use `pyJac`. We will expand upon this discussion in Section 5 to highlight future directions.

Platform	Parallel	Shallow Vectorization	Deep Vectorization
OpenMP	✓	—	—
POCL OpenCL	✓	✓	✓
Intel OpenCL	✓	✓	—
Nvidia OpenCL	—	✓	—

Table 3: The platforms used in this study and the execution /vectorization patterns that they are capable of running.

Finally, Table 4 displays the chemical kinetic models used in this work,

³A driver implementation issue limited total memory to 4 GB and 10 GB on the C2075 and K40m GPUs, respectively [64].

as well as number of partially stirred reaction conditions (PaSR) used in the condition database for each. Our previous works describe the creation of the PaSR databases in detail works [18, 24].

Model	Number of Conditions	Reference
H ₂ /CO	900,900	[65]
GRI-Mech 3.0	450,900	[66]
USC-Mech II	91,800	[67]
iC ₅ H ₁₁ OH	450,900	[68]

Table 4: The chemical kinetic models used in this study and number of conditions in the partially stirred reactor database for each.

3.2. Source-term verification

We verified the reaction rates of progress (ROP), species production rates, and temperature rates in this study by comparing with values calculated using Cantera [56]. However, special care must be taken due to floating-point arithmetic issues.

For a direct comparison, a relative error norm of a quantity X_{ij} over all states j and reactions (or species) i was computed using the L^∞ norm:

$$E_X = \left\| \frac{|X_{ij,\text{CT}} - X_{ij}|}{10^{-10} + 10^{-6} \times |X_{ij,\text{CT}}|} \right\|_\infty, \quad (10)$$

where the CT subscript indicates values from Cantera [56].

However, computing the net ROP of reaction i for state j from the forward and reverse ROP, i.e., $R_{ij} = R'_{ij} - R''_{ij}$, can easily lose accuracy as the net ROP may be many orders of magnitude smaller than the forward and/or reverse rates—particularly near chemical equilibrium. To quantify this phenomena, we first define the error in forward ROP as

$$\varepsilon'_{ij} = |R'_{ij} - R'_{ij,\text{CT}}|, \quad (11)$$

while the error in reverse ROP, ε''_{ij} , can be defined analogously. Finally, for the reaction i^* and the state j^* that result in the largest error in net ROP, i.e., E_R , an estimate of the error attributable to floating-point error accumulation from the forward and reverse ROPs can be obtained using

$$E_\varepsilon = \frac{\max(\varepsilon'_{i^*j^*}, \varepsilon''_{i^*j^*})}{10^{-10} + 10^{-6} \times |R_{i^*j^*,\text{CT}}|}. \quad (12)$$

This estimate allows for directly comparing the error in forward or reverse ROPs with the value of the net ROP itself; the error in net ROP will be large if these are similar in magnitude.

Model	H ₂ /CO	GRI-Mech. 3.0	USC-Mech II	iC ₅ H ₁₁ OH
$E_{R'}$	1.56×10^{-8}	2.95×10^{-8}	9.42×10^{-8}	4.86×10^{-4}
$E_{R''}$	6.92×10^{-8}	6.53×10^{-8}	1.20×10^{-7}	5.07×10^{-4}
E_R	1.49×10^1	1.11×10^0	2.80×10^0	4.82×10^{-1}
E_ϵ	1.48×10^1	1.13×10^0	2.93×10^0	5.03×10^{-1}
$E_{\frac{dn}{dt}}$	2.53×10^1	2.60×10^0	7.62×10^0	1.58×10^1
$E_{\frac{dT}{dt}}$	3.94×10^5	3.35×10^8	3.95×10^6	7.11×10^7
$E_{\frac{dS}{dt}}$	3.52×10^{12}	3.46×10^{12}	3.44×10^{12}	3.38×10^{12}

Table 5: Summary of errors in rates of progress, species, temperature, and thermodynamic state-parameter rate compared with Cantera. Error statistics are based on the infinity-norm of the relative error detailed in Eq. (10) for each quantity. The “S” in $E_{\frac{dS}{dt}}$ refers to the thermodynamic state parameter, either V or P for CONP and CONV, respectively.

Table 5 compares `pyJac` v2’s source-term evaluations with Cantera’s [56] using the data set of PaSR conditions (Table 4). The forward and reverse ROPs agree closely for all models, though the error norm is ~ 3 – 4 orders of magnitude larger for the isopentanol model. This discrepancy results from differences in evaluation of P-Log reactions between `pyJac` and Cantera: `pyJac` computes the logarithm of the upper and lower reaction Arrhenius rates analytically (see supplemental material) while Cantera evaluates this term numerically. If we neglect the errors from P-Log reactions in Eq. (10), the errors for the forward and reverse ROPs fall to 5.44×10^{-8} and 1.59×10^{-7} , respectively. This discrepancy does not imply any actual error in either `pyJac` or Cantera—in fact, the error still lies well within the proscribed tolerances in Eq. (10)—but merely emphasizes how even small code changes can affect the accumulation of floating-point errors.

The error in the net ROP further underscores this point: it is ~ 3 – 9 orders of magnitude (or 7 – 9 orders of magnitude when including P-Log reaction contributions) larger than the error in forward or reverse ROP. Table 5 shows that the magnitudes of E_ϵ and E_R agree in all cases, indicating that the accumulation of floating-point error from the forward and reverse ROPs causes this large increase in error as previously discussed. The magnitudes of the errors in molar species production rate and net ROP agree, but thermody-

dynamic properties amplify the error in net species production rates and lead to high discrepancies in temperature and state-parameter rates. Again, these discrepancies in net ROP will not necessarily cause errors when integrating the chemical kinetics—either in `pyJac` or `Cantera`—as this loss of accuracy only occurs when the forward and reverse ROPs are nearly equal (i.e., near equilibrium).

3.3. Jacobian verification

As in our previous work [18], we determined Jacobian matrix correctness by comparing with that obtained by automatic differentiation of the `pyJac`-generated source term, using the `Adept` software library [57, 58]. We previously explained this choice fully [18], but broadly speaking automatic differentiation provides relatively fast, highly accurate Jacobian matrix evaluation with minimal additional programming effort. (In contrast, it is challenging to obtain robust, accurate Jacobians using finite differences.) The discrepancy between the analytical and automatic-differentiation Jacobians for thermochemical state k , denoted by $\mathcal{J}_k$ and $\hat{\mathcal{J}}_k$ respectively, is determined by the relative error Frobenius norm over all Jacobian indices i, j :

$$E_{\text{rel},k} = \left\| \frac{\hat{\mathcal{J}}_{ij,k} - \mathcal{J}_{ij,k}}{\hat{\mathcal{J}}_{ij,k}} \right\|_F. \quad (13)$$

To avoid large relative discrepancies in small nonzero Jacobian elements due to accumulation of floating-point error, the Frobenius norm of the automatically differentiated Jacobian is calculated over all thermochemical states k :

$$\mathcal{T} = \left\| \hat{\mathcal{J}} \right\|_F. \quad (14)$$

The error statistics reported in this section are then based only on matrix elements where $\mathcal{J}_{ijk} \geq \frac{\mathcal{T}}{\mathcal{C}}$, where $\mathcal{C}$ is a tunable threshold parameter; this filtered form of Eq. (13) is denoted $E_{\mathcal{C},k}$. Finally, the Frobenius norm is calculated over all the states k in the PaSR thermochemical condition database:

$$E_{\mathcal{C}} = \|E_{\mathcal{C},k}\|_F. \quad (15)$$

This error norm is quite different from the relative error Frobenius norm suggested by Anderson et al. [69] for quantifying the error of matrices in LAPACK, e.g., over Jacobian indices i, j :

$$E_{\mathcal{L},k} = \frac{\left\| \hat{\mathcal{J}}_{ijk} - \mathcal{J}_{ijk} \right\|_F}{\left\| \hat{\mathcal{J}}_{ijk} \right\|_F}$$

and states k :

$$E_{\mathcal{L}} = \|E_{\mathcal{L},k}\|_F. \quad (16)$$

In our experience, the accuracy of larger elements in a Jacobian often dominates the LAPACK error norm, while the filtered error norm can identify errors in both large and small Jacobian entries. Further, with the tunable threshold parameter $\mathcal{C}$, we can assess the error of different ranges of element sizes and isolate the effects of floating-point error. For reference, both our error norm and the LAPACK error norm will be reported.

Model	$E_{\mathcal{C}}$	$\bar{\mathcal{T}}$	$E_{\mathcal{C}=10^{20}}$	$E_{\mathcal{C}=10^{15}}$
H ₂ /CO	1.862×10^{-14}	6.431×10^{18}	1.741×10^0	4.508×10^{-5}
GRI-Mech 3.0	1.567×10^{-14}	7.783×10^{19}	3.842×10^{-7}	3.687×10^{-7}
USC-Mech II	1.137×10^{-14}	2.830×10^{21}	1.199×10^{-2}	1.983×10^{-7}
iC ₅ H ₁₁ OH	1.227×10^{-10}	2.733×10^{26}	1.363×10^{-3}	2.764×10^{-5}

Table 6: Summary of Jacobian matrix verification results. The reported error statistics are the maximum filtered relative error $E_{\mathcal{C}}$ and LAPACK error $E_{\mathcal{L}}$ over all test platforms, vectorization patterns (Table 3), CONP/CONV, and sparse/dense Jacobians. The Frobenius norm described in Eq. (14) varies slightly between the CONP and CONV cases; the reported $\bar{\mathcal{T}}$ is the average of the two, with the appropriate value used during calculations of the error statistics.

Table 6 reports the maximum $E_{\mathcal{C}}$ and $E_{\mathcal{L}}$ values over all test platforms and vectorization patterns (see Table 3), sparse and dense (see Section 3.4), as well as CONP and CONV formulations. The most stringent filtered error norm ($\mathcal{C} = 10^{20}$) ranges from 10^{-7} – 10^0 ; the largest error is for the H₂/CO model. For this model, $\mathcal{T}$ is smaller than the tolerance of 10^{20} , and hence the error norm considers Jacobian entries smaller than $\mathcal{O}(1)$. GRI-Mech 3.0 has a $\mathcal{T}$ roughly an order of magnitude larger and so the stringent error norm is significantly smaller: $\mathcal{O}(10^{-7})$. Given the intricacy of floating-point error evaluation, the use of different languages and OpenCL platforms (the effect of these differences will be explored in Appendix B), and the general complexity of `pyJac` it would be exceedingly difficult to pinpoint an exact cause for this phenomenon, as was done in Section 3.2. To ensure no bugs or errors present in the Jacobians generated by `pyJac`, relaxed filtered error norms ($\mathcal{C} = 10^{15}$) are also presented for each model in Table 6. This relaxed norm is smaller by 2–5 orders of magnitude for all models—except GRI-Mech

3.0, where the stringent case already has small error as previously discussed—which indicates that accuracy is higher when the smaller Jacobian entries are excluded. This result suggests that floating-point error accumulation controls the stringent filtered error norm.

The relative LAPACK error norm—ranging from $\sim 10^{-10}$ – 10^{-14} —re-enforces this finding, as it indicates roughly 10–14 digits of accuracy [69]. The $\text{iC}_5\text{H}_{11}\text{OH}$ model has the largest LAPACK error norm, likely due to the presence of P-Log/Chebyshev reactions and the resulting complicated derivatives with many logarithms, exponentiations, and summations. Further, the LAPACK error norm does not correlate well with the stringent filtered error norm, e.g., USC-Mech II has the smallest LAPACK error norm (1.137×10^{-14}) but the second-largest stringent filtered error norm (1.119×10^{-2}). Conversely, the model with the largest LAPACK error norm, $\text{iC}_5\text{H}_{11}\text{OH}$ has the second smallest stringent filtered error norm: $E_{\mathcal{L}} = 1.227 \times 10^{-10}$ and $E_{\mathcal{C}=10^{20}} = 1.363 \times 10^{-3}$, again suggesting that floating-point error accumulation influences the stringent error norm. These findings, along with the individual unit-testing of all chemical source-terms and Jacobian sub-components described in Section 2.3, gives high confidence in the correctness of `pyJac` v2.

3.4. Sparsity patterns

In general, the Jacobian matrices generated by `pyJac` are largely sparse with non-zero entries corresponding to species that participate in the same reaction or non-default efficiency third-body species in a reaction, with dense rows/columns corresponding to temperature and the thermodynamic state parameter. However, the explicit-mass conservation formulation of `pyJac` can introduce additional non-zero entries in two ways. First, if the last species in the model (i.e., the bath gas) participates directly in any reaction, the derivative of its forward or reverse rate of progress is non-zero with respect to all other species in the model, regardless of whether the other species participate in that reaction or not. Similarly, if the last species has a third-body efficiency not equal to the default (one), this will again create nonzero derivatives for the pressure-modification term with respect to all other species (see the supplemental material). Either case will result in a fully dense Jacobian row for all species with a non-zero net stoichiometric coefficient in such a reaction.

However, `pyJac` v2 allows the user to ignore these derivatives (via a

command-line switch) and avoid the adverse effects on Jacobian sparsity.⁵ The rationale behind this choice is that many common implicit integration techniques (e.g., CVODE [47]) used to solve chemical-kinetic initial-value problems are formulated around the assumption that the supplied Jacobian is approximate; this allows the Jacobian and its LU factorization to be reused for multiple internal integration time steps, accelerating the solution process. Such solvers do not need the exact form of the Jacobian and thus the so-called “approximate” form is preferable. Though this might be used as a crude form of preconditioning for such solvers, the primary purpose is merely to increase Jacobian sparsity; McNenly et al. [70] more thoroughly investigated preconditioners. Hence, in this section we will detail the sparsity of both forms of the Jacobian for the chemical models tested.

Figure 4 graphically represents the Jacobian sparsity of GRI-Mech 3.0. In particular we note that Fig. 4b has several rows that are no longer fully dense, as result of its approximate form; these rows correspond to species directly interacting with N_2 , largely in GRI-Mech 3.0’s nitrogen chemistry reactions. Table 7 shows the density of the exact and approximate Jacobians for all chemical kinetic models tested in this work. The smallest model, H_2/CO , is very dense with 71.4 % of the exact Jacobian entries non-zero; this drops to 56.7 % for GRI-Mech 3.0, continues to decrease to 28.2 % for USC-Mech II, and is just 11.5 % for the isopentanol model. The approximate Jacobian assumption drops the density of Jacobian by $\sim 3\text{--}7\%$ for all models.

Currently, `pyJac` can use two common sparse-matrix storage formats [71]: compressed row storage (CRS) and compressed column storage (CCS), used for “C” and “F”-ordered data respectively. For brevity we will outline only the CRS format but the CCS format is similar [71]. An $N \times N$ CRS matrix is stored using three vectors: a value vector of length N_{NZ} (the number of non-zero elements in the Jacobian) that stores the elements of the Jacobian, a row pointer vector of length N that stores the locations in the value vector that begin a row, and a column index vector length N_{NZ} that stores the column indices of the elements in the value vector.

The Jacobian access pattern used by `pyJac` is fairly irregular; for simplicity we will only discuss looping-structure of species derivatives calculations since these form the bulk of the computation and have the most challenging

⁵Alternatively, one may choose the last species as one that does not participate in any reactions.



Figure 4: A graphical representation of the sparsity pattern of the chemical kinetic Jacobian generated by `pyJac` for GRI-Mech 3.0. Black squares indicate a non-zero Jacobian entry, while white square correspond to an empty entry. The numbers indicate the index of the entry in the state vector.

Model	Exact Jacobian Density	Approximate Jacobian Density
H ₂ /CO	71.4 %	68.4 %
GRI-Mech 3.0	56.7 %	49.8 %
USC-Mech II	28.2 %	26.4 %
iC ₅ H ₁₁ OH	11.5 %	7.98 %

Table 7: The density of the exact and approximate Jacobians generated by `pyJac` for the various models studied.

Jacobian access patterns. In general, an outer loop iterates over all reactions of a certain type (e.g., falloff reactions) and calculates the relevant Jacobian subproducts—independent of any particular species—for the reaction (e.g., the derivative of the falloff pressure modification term). Two inner loops then iterate over the species in a reaction, updating the Jacobian entries for these species as appropriate. This pattern leads to fairly easily vectorizable code and efficient Jacobian evaluation, since the bulk of the computation depends only on the reaction in question, as discussed in our previous work [18]. Generally, this means that a lookup operation is required to find the sparse Jacobian index for any pair of state variables; in some cases this can be avoided, e.g., the rows corresponding to derivatives of the temperature and thermodynamic state parameter source-terms are fully

dense in `pyJac`, and hence no lookup is necessary. This lookup operation is currently implemented as a simple “for” loop, e.g., for a sparse lookup of a pair of indices (i, j) in a CRS matrix, the lookup function searches the column index vector between the values `row_pointer[i], ..., row_pointer[i+1]` for j , and returns the offset from `row_pointer[i]` (or -1 if not found). As will be explored in Section 3.5.2, this slows down sparse Jacobian evaluation, and might be improved by a static mapping of the full Jacobian indices to the sparse index (or some null value if the entry is empty). However, this would require increased constant-data usage, a limitation for OpenCL. Additionally, this might be an excellent usage of OpenCL’s Image memory type (similar to texture memory in CUDA terminology). Both of these sparse indexing techniques merit future investigation.

3.5. Performance

The performance studies in this work were run on the platforms listed in Tables 1 and 2. Run times in each case were averaged over ten runs, each using the same set of PaSR conditions used in verification. The OpenMP Jacobian/source-term kernels, as well as the OpenMP/OpenCL wrapping code (responsible for initializing/transferring memory, reading input, etc.) was compiled with `gcc v5.4.0` on the `avx2/K40m` platforms and `gcc v4.8.5` on the `sse4.2/C2075` machines. The optimization level “`-O3 -mtune=native`” was used and no “fast math” OpenCL optimizations were enabled. Additionally, the exact form of the Jacobian (as opposed the “approximate” form discussed in Section 3.4) was used in all cases. Finally, unless stated otherwise: the performance results used a single CPU core, the CONP assumption, a vector width of 8/128, and “C”/“F”-ordered data for the CPU/GPU cases, respectively; the run times reported are for the number of conditions specified in Table 6 and include data-transfer overhead to/from internal buffers used in `pyJac`. The effects of choice of vector width, data ordering, and differences between CONP and CONV evaluations on the CPU/GPU will be explored in Section 3.5.1, while parallel scaling for multiple CPU cores will be examined in Sections 3.5.1 and 3.5.2.

3.5.1. Source-term evaluation

Figure 5 explores the performance of the source-term evaluations generated by `pyJac` on the CPU test platforms listed in Table 1. Source-term evaluations—critical in their own right for direct numerical simulations of

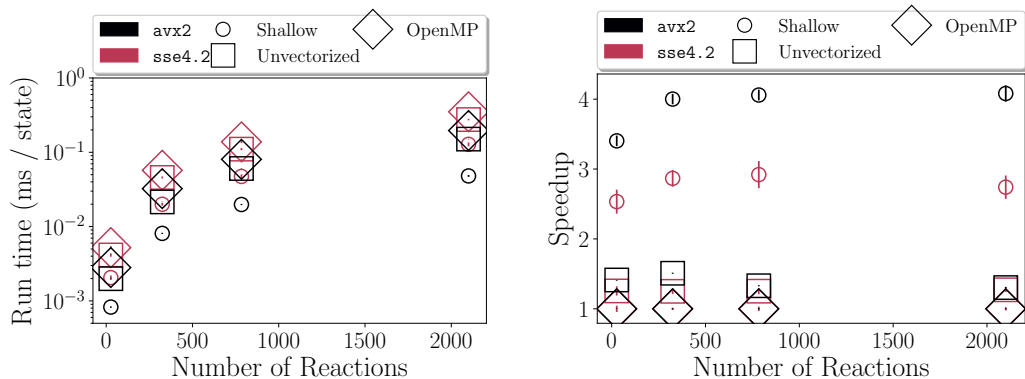
reactive-flows [3], among other applications—also provide a convenient platform to detail the effects of various code configuration options before investigating the more involved Jacobian evaluation performance.

Figure 5a shows the mean run time per initial condition for both the `avx2` and `sse4.2` CPUs, using Intel OpenCL and OpenMP. This normalization of the run time by the number of initial conditions tested is chosen to account for the varying numbers of conditions in the PaSR databases for each model (Table 4). For both CPUs, the OpenMP implementation is the slowest for all models; interestingly, the unvectorized (i.e., strictly parallel) Intel OpenCL code is slightly faster than OpenMP in all cases. As expected, the `avx2` machine is faster than the `sse4.2` CPU for the strictly parallel cases, performing $1.82\text{--}2.13\times$ and $1.72\text{--}1.85\times$ faster for the unvectorized OpenCL case and OpenMP, respectively. Additionally, the shallow-vectorized OpenCL code performs significantly faster than either the OpenMP or unvectorized OpenCL codes on both processors.

Figure 5b details the extent of this speedup; the speedup displayed is calculated per-machine, e.g., the `avx2` shallow-vectorized code speedup is relative to OpenMP on the same CPU. On both machines, the unvectorized OpenCL code is faster than the baseline parallel OpenMP code, by $1.30\text{--}1.51\times$ on the `avx2` CPU and $1.25\text{--}1.27\times$ on the `sse4.2` machine. Additionally, the shallow-vectorized OpenCL code is $2.53\text{--}2.92\times$ and $3.40\text{--}4.08\times$ faster than the OpenMP code for the `sse4.2` and `avx2` machines, respectively.

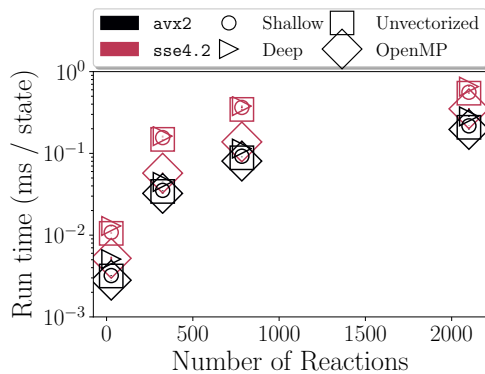
In contrast, Fig. 5c shows the mean run time per condition of deep, shallow, and unvectorized OpenCL codes using the POCL runtime, as compared with OpenMP parallelization. No speedup is achieved for either vectorization type on either CPU—indeed, the OpenMP case is faster on both CPUs, though we stress that Intel OpenCL runtime achieves vectorization using the same code and hardware. The reasons for this lack of vectorization with POCL are quite technical; however, personal communication with the developers of POCL revealed that to achieve vectorization, changes are required to both the way POCL prepares LLVM intermediate representation code, as well as improvements to LLVM’s loop-vectorizer itself [72].⁶ As will be

⁶Specifically, improvements such as ensuring LLVM recognizes uniform vectorization loop bounds (even if said bounds *are* uniform in practice), proving vector instructions’ ability to handle all edge cases identically to the corresponding scalar instruction, handling of branches and conditionals (potentially within POCL instructions themselves), and



(a) The mean run time per condition for each chemical model using the Intel OpenCL and OpenMP on both CPUs studied.

(b) The speedup achieved over the baseline OpenMP parallelization by both the unvectorized and shallow-vectorized Intel OpenCL codes; the speedup is presented on per-machine basis.



(c) The mean run time per condition of the Portable OpenCL runtime compared to OpenMP parallelization.

Figure 5: Mean run times per condition and speedups achieved by the various CPU OpenCL runtimes compared to OpenMP parallelization for each chemical model studied. The names in the legends correspond to the identifiers listed in Table 1.

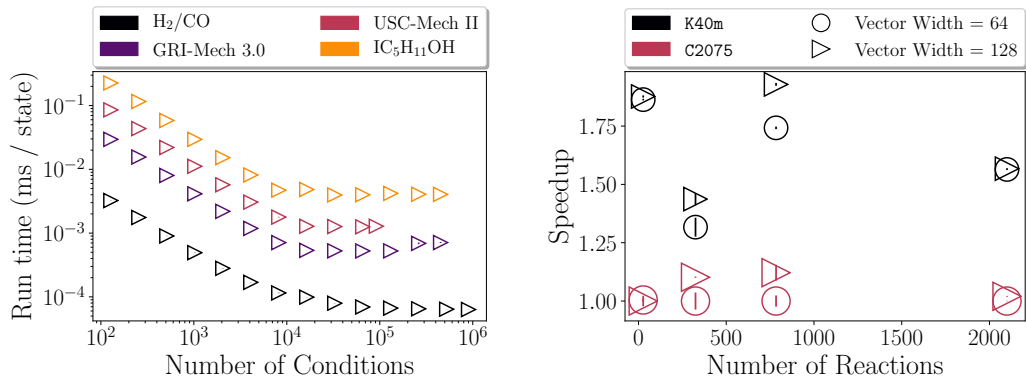
discussed in Section 4, we hope that using explicit vector types (to lessen demands on the LLVM-vectorization module) in combination with some of these changes might solve this issue, but for the moment POCL is still quite useful as a verification tool.

Figure 6 shows the performance of evaluating source terms on the GPUs listed in Table 2. Figure 6a investigates how the number of initial conditions evaluated affects the mean run time per condition on the K40m GPU; the run time decreases until around $\sim 10^4$ conditions for all chemical models, at which point the GPU becomes saturated and performance levels off. The performance plateaus slightly later for the H_2/CO model compared with the others. Figure 6b shows the speedup in source-term evaluation that the K40m GPU achieves over the C2075 GPU for the maximum number of conditions in Fig. 6a with two vector widths (i.e., GPU block size), 64 and 128. The best K40m case with a vector width of 128 is $1.40\text{--}1.88\times$ faster than the slowest case (C2075 with a vector-width of 64) depending on the chemical model in question. Figure 6b also shows that varying the vector-width minimally affects performance for most of the K40m and C2075 cases; the GRI-Mech 3.0 and USC-Mech II models show the largest improvements with a vector width of 128: $\sim 10\text{--}18\%$ for both GPUs. This is likely caused by higher occupancy on the GPUs, but it is unclear exactly how Nvidia’s OpenCL runtime balances the registers/warps per streaming-multiprocessor, as controls occupancy in CUDA [73].

Figure 7 shows how changing data-ordering patterns, the CONP or CONV formulation, and the CPU vector width affect the performance of source-term evaluation in pyJac. Per Fig. 7a, we see that the choice of CONP or CONV formulation has little to no effect on run time for OpenMP as well as the shallow-vectorized/unvectorized Intel OpenCL codes on the avx2 machine. Generally speaking, the difference between the CONP and CONV formulations only marginally affects performance regardless of CPU/GPU choice.

In contrast, Fig. 7b shows significant speedups of “C”-ordered data over “F”-ordered data on the avx2 machine; the speedup presented is calculated per language, e.g., the $1.35\text{--}2.09\times$ speedup of the “C”-ordered OpenMP implementation is relative to the “F”-ordered OpenMP baseline. Additionally, the “C”-ordered shallow-vectorization in Fig. 7b and the other shallow-vectorized CPU data shown in Sections 3.5.1 and 3.5.2 use the vectorized-

handling of memory access/vector-element extraction patterns.



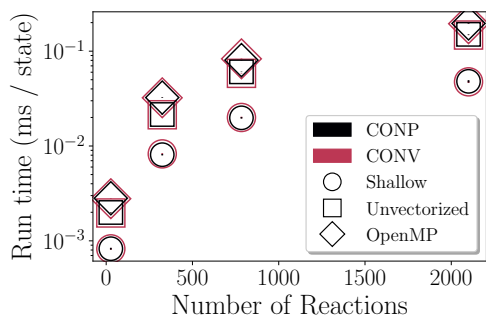
(a) The mean run time per condition for each chemical model on the K40m GPU as a function of the number of initial conditions tested. (b) The speedup achieved on the K40m versus the C2075 GPU; the names correspond to the identifiers listed in Table 2

Figure 6: pyJac source-term evaluation performance on the Nvidia GPUs

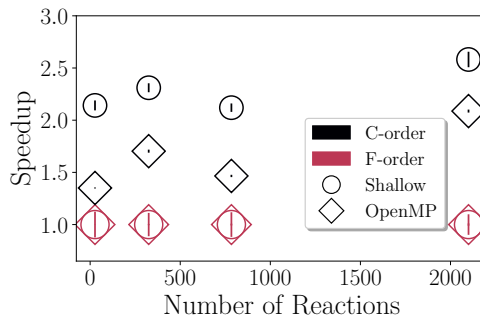
data ordering described in Section 2.1; this case achieves speedups of $2.14\text{--}2.58\times$ over the “F”-ordered shallow-vectorization, demonstrating the value of the vectorized-data ordering for CPU execution.

Figure 7c shows how the “C”- and “F”-ordering affect the performance of source-term evaluation on both GPUs, with the speedup presented per-GPU. The “C”- and “F”-ordered shallow-vectorizations perform almost equivalently on both GPUs, with less than a 10% difference in run time between data orderings. For the isopentanol model, “C”-ordered data is $\sim 1.08\times$ faster on both GPUs (while the trend is less clear for the other models). The roughly equivalent performance between the “C” and “F”-ordered approaches on GPUs counters what one might expect: typically speaking, coalesced memory access in a shallow vectorization is easier to achieve with “F”-ordering (see Section 2.1). However, the vectorized-data ordering here ensures that memory storage is aligned to the vector width and, thus, encourages coalesced accesses.

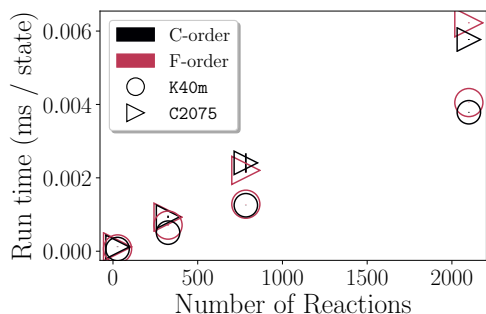
Figure 7d shows how changing vector width affects source-term evaluation performance on the avx2 CPU. The vector width of 8 performs the fastest (out of those tested), while the larger vector width of 16 is slightly slower due to increased register pressure [74]. It is unclear why the vector width of 4 results in no speedup at all (in fact, it is the slowest case). Intel’s



(a) The mean run time per condition for each chemical model using both the CONP and CONV formulations on Intel OpenCL and OpenMP on the **avx2** CPU.



(b) The speedup achieved by “C” ordering over “F” ordering for Intel OpenCL and OpenMP on the **avx2** CPU. The speedup presented is calculated per-language (OpenMP and OpenCL) to better assess the effect of the data ordering.



(c) The affect on performance of “C” vs “F”-ordering for the shallow-vectorized Nvidia OpenCL code on both GPUs with a vector-width of 128.



(d) The effect of vector-width on “C”-ordered shallow-vectorized Intel OpenCL source-term evaluations on the **avx2** CPU.

Figure 7: The effect of the CONP and CONV formulations, “C” and “F” data-ordering, and CPU vector-width on source-term evaluation performance in **pyJac**. The shallow-vectorized “C”-ordered OpenCL cases correspond to the vectorized data ordering described in Section 2.1.

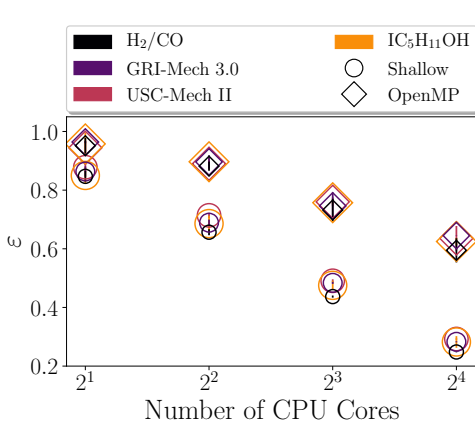
vectorization guide [75] mentions that a heuristic determines the optimal vector width (in this case, it appears from compiler output to be 8), so it is possible that using a vector width smaller than the heuristic breaks the implicit vectorizer. This issue does not occur for a vector width of 4 on the `sse4.2` CPU.

Finally, Fig. 8 displays the (strong) parallel scaling efficiency and SIMD efficiency for the CPU platforms. The strong parallel scaling efficiency is defined as

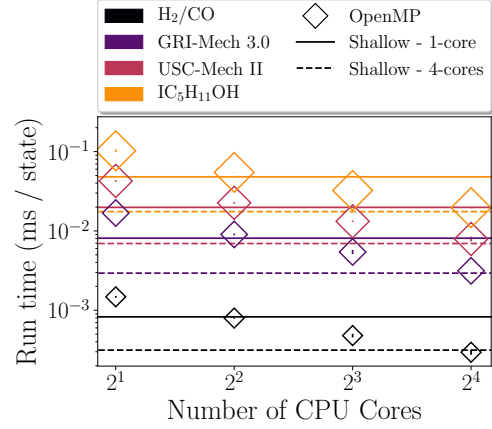
$$\varepsilon = \frac{\bar{t}_1}{N\bar{t}_N}, \quad (17)$$

where $\bar{t}_N$ is the mean run time per condition on N CPU cores and $\bar{t}_1$ the same on a single CPU core. The strong parallel scaling efficiency measures the speedup due to the use of additional CPU cores as a fraction of linear speedup; strong scaling tends to decrease with the number of processors used due to memory-bandwidth limitations and decreasing computation work allocated per CPU core [76].

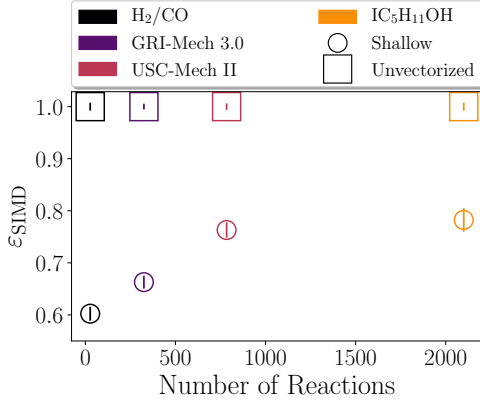
Figure 8a shows the strong parallel scaling efficiency of source-term evaluation in `pyJac` on the `avx2` machine for both the shallow-vectorized Intel OpenCL and OpenMP codes. In general, the H_2/CO mechanism has the worst scaling efficiency for both Intel OpenCL and OpenMP, likely resulting from both its relatively small size and few falloff/chemically activated reaction (in particular, the additional expensive logarithm and exponential evaluations that accompany them). As demonstrated in Appendix C, the amount of computational work required per thermochemical state plays a critical role in fully utilizing SIMD instructions/multiple threads. Additionally, OpenMP tends to scale better than the shallow-vectorized OpenCL code, e.g., ~ 0.9 and ~ 0.75 for four and eight CPU cores, respectively, compared to just 0.66–0.72 and 0.44–0.48 for OpenCL. Though not pictured (to keep the figure readable), the unvectorized Intel OpenCL code scales only slightly worse than the OpenMP code, hence the poorer scaling is unique to the shallow-vectorized code. This is due in large part to the superior performance of the shallow-vectorized OpenCL code, coupled with the relatively small amounts of work associated with source-term evaluation. To illustrate this, Fig. 8b shows the mean run time per-condition of the OpenMP source-term evaluations for 2–16 cores, compared to the shallow-vectorized OpenCL code on one (solid line) and four (dashed line) cores on the `avx2` machine. For all chemical models, the mean run time per-condition (and hence the computational work allocated per-core, one of the key-drivers of parallel scaling



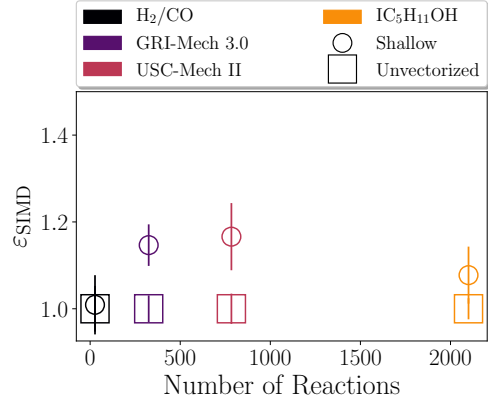
(a) The strong parallel scaling efficiency (as defined in Eq. (17)) of source-term evaluation for Intel OpenCL/OpenMP on the **avx2** machine.



(b) The mean run time per condition of OpenMP source-term evaluation on different cores compared to a shallow-vectorized Intel OpenCL evaluation on 1 (solid lines) and 4 (dashed lines) cores on **avx2** machine.



(c) The SIMD efficiency (Eq. (18)) of source-term evaluation for the Intel OpenCL runtime on a single core of the **avx2** CPU.



(d) The SIMD efficiency (Eq. (18)) of source-term evaluation for the Intel OpenCL runtime on a single core of the **sse4.2** CPU.

Figure 8: The parallel scaling efficiency and SIMD efficiency of source-term evaluation for Intel OpenCL on the **avx2** and **sse4.2** CPUs.

efficiency [76]) of OpenMP running on four cores is roughly equal to that of the shallow-vectorized OpenCL code on a single core. Similarly, OpenMP running on 16 cores is roughly equivalent to the OpenCL code on 4 cores. Therefore, a more fair comparison of parallel scaling efficiencies is to compare OpenCL running on 4 cores with OpenMP on 16; the OpenMP code’s parallel efficiency drops to ~ 0.64 for 16 cores, similar to OpenCL’s parallel scaling efficiency of 0.66–0.72 at 4 cores. Indeed, as will be seen in Section 3.5.2, sparse Jacobian evaluation—the most computationally intensive task in this work—exhibits similar strong-scaling efficiency on Intel OpenCL and OpenMP.

The SIMD efficiency is defined as

$$\varepsilon_{\text{SIMD}} = \frac{\bar{t}_{\text{unvec}}}{W \bar{t}_{\text{shallow}}} , \quad (18)$$

where $\bar{t}_{\text{unvec}}$ is the mean run time per condition of the unvectorized OpenCL code, $\bar{t}_{\text{shallow}}$ the same for the shallow-vectorized OpenCL code, and W is the vector width reported in number of double operations (see Table 1). This measure compares the actual speedup due to shallow vectorization with the ideal speedup based on the nominal vector width of the machine. Figure 8c shows the SIMD efficiency of source-term evaluation in `pyJac` on a single core of the `avx2` machine; the larger models (isopentanol and USC-Mech II) have higher SIMD efficiencies of 0.76–0.78, and the smaller models (H_2/CO , GRI-Mech 3.0) have lower SIMD efficiencies of 0.6–0.66. This again demonstrates that the SIMD vectorization becomes more efficient with increasing amounts of work to perform (i.e., with increasing model size). Interestingly, Fig. 8d shows the SIMD efficiency on the `sse4.2` machine as greater than one. This is likely caused by a combination of using an OpenCL vector width greater than the native CPU vector width (i.e., eight versus two) and improved data locality for the vectorized-data ordering as discussed in Section 2.1, and results in a modest 7–17% improvement over the nominal vector width.

3.5.2. Jacobian evaluation

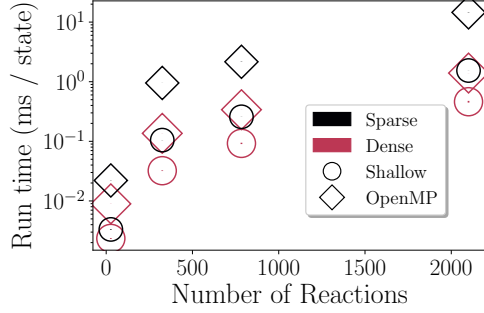
Figure 9 shows the performance of the sparse and dense Jacobian evaluations in `pyJac` on the CPU platforms. In Fig. 9a, the mean run time per condition is presented for the shallow-vectorized Intel OpenCL and OpenMP codes on the `avx2` CPU. The sparse Jacobian evaluates slower on both Intel OpenCL and OpenMP due to indirect lookup indexing requirements, as

discussed in Section 3.4. Interestingly, indirect lookup less-negatively impacts the shallow-vectorized OpenCL code: the sparse OpenMP code is $2.47\text{--}10.42\times$ slower than the dense OpenMP evaluation, while the sparse shallow-vectorized OpenCL code is just $1.41\text{--}3.34\times$ slower than its dense counterpart. As a result, the shallow-vectorized sparse OpenCL code performs as fast or faster than the dense OpenMP code in all cases on the `avx2` machine (Fig. 9a). Figure 9b shows the speedup of the shallow-vectorized OpenCL, sparse and dense Jacobian evaluations over the same on OpenMP; the dense OpenCL code is $3.03\text{--}4.23\times$ faster than the corresponding dense OpenMP code. This speedup increases to $6.63\text{--}9.44\times$ for the sparse Jacobian.

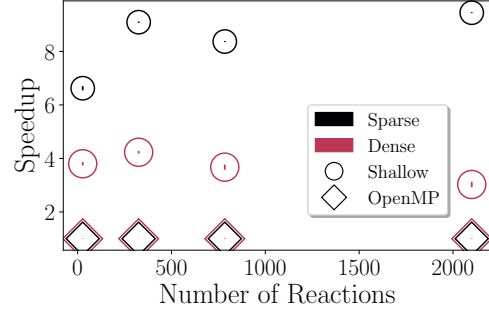
On the `sse4.2` machine, Fig. 9c shows similar results: the sparse OpenMP code is the slowest in all cases, and the shallow-vectorized OpenCL code is nearly as fast as the dense OpenMP code. Once again, indirect lookup less-negatively impacts the sparse OpenCL code, which is only $1.76\text{--}3.33\times$ slower than its dense counterpart, while the sparse OpenMP code is significantly ($2.25\text{--}8.72\times$) slower than the dense version. In Fig. 9d, the speedup of the sparse and dense shallow-vectorized OpenCL codes are compared with their OpenMP versions; the dense OpenCL code is $1.92\text{--}2.47\times$ faster while the sparse shallow-vectorization achieves a speedup of $3.14\text{--}5.03\times$.

Figure 10a compares the strong parallel scaling efficiency of the sparse shallow-vectorized OpenCL with the sparse OpenMP code. Although the plot is challenging to read since most of the data are clustered together, it shows that the shallow-vectorized OpenCL code scales similarly to the OpenMP code, in contrast to the parallel scaling efficiency of source-term evaluation (Fig. 8a). The H_2/CO model scales the worst for both codes, ranging from $0.94\text{--}0.54$ and $0.99\text{--}0.82$ efficiency for OpenCL and OpenMP respectively on 2–16 cores. As the model size increases, the efficiency of the OpenCL code improves dramatically, reaching $0.997\text{--}0.84$ for the isopentanol model.

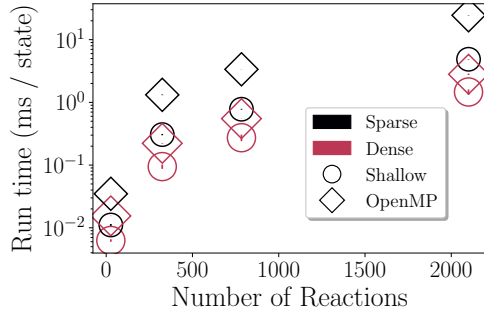
Figure 10b shows scaling for the OpenMP and shallow-vectorized dense Jacobian OpenCL codes. In this case, the isopentanol model scales the worst for both cases. The sheer size of the dense isopentanol Jacobian limited the total number of states for the dense isopentanol Jacobian evaluation to 50,000—storing the dense matrix for a single thermochemical state takes over 1 MB of data, so 50,000 states requires over 50 GB of memory); this greatly drops the computation cost for this case, and adversely affects the scaling efficiency as discussed in Section 3.5.1. Excluding the isopentanol



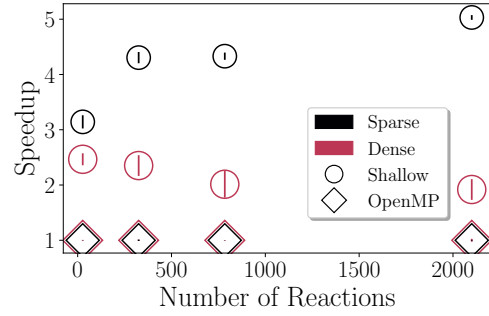
(a) The mean run time per condition of sparse and dense Jacobian evaluations for Intel OpenCL/OpenMP on the **avx2** machine.



(b) The speedup of the sparse and dense Jacobian evaluations for Intel OpenCL over the sparse/dense OpenMP baseline on the **avx2** machine.

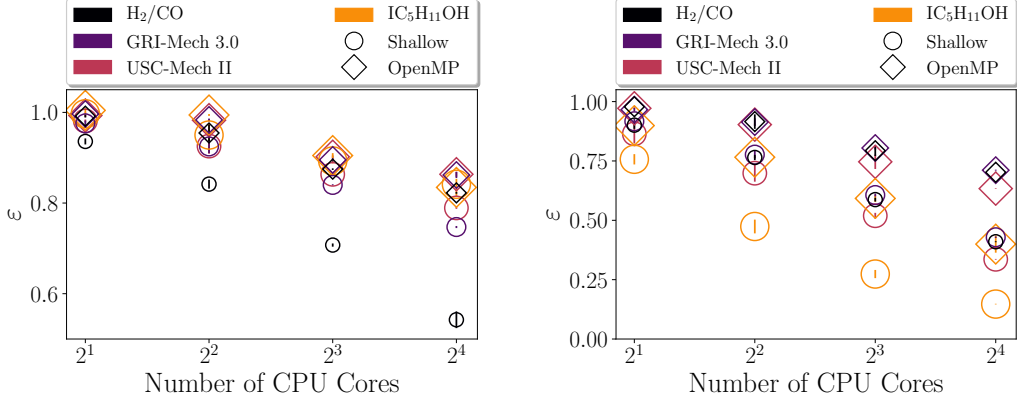


(c) The mean run time per condition of sparse and dense Jacobian evaluations for Intel OpenCL/OpenMP on the **sse4.2** machine.



(d) The speedup of the sparse and dense Jacobian evaluations for Intel OpenCL over the sparse/dense OpenMP baseline on the **sse4.2** machine.

Figure 9: Performance of sparse and dense Jacobian evaluations on the CPU platforms in pyJac.

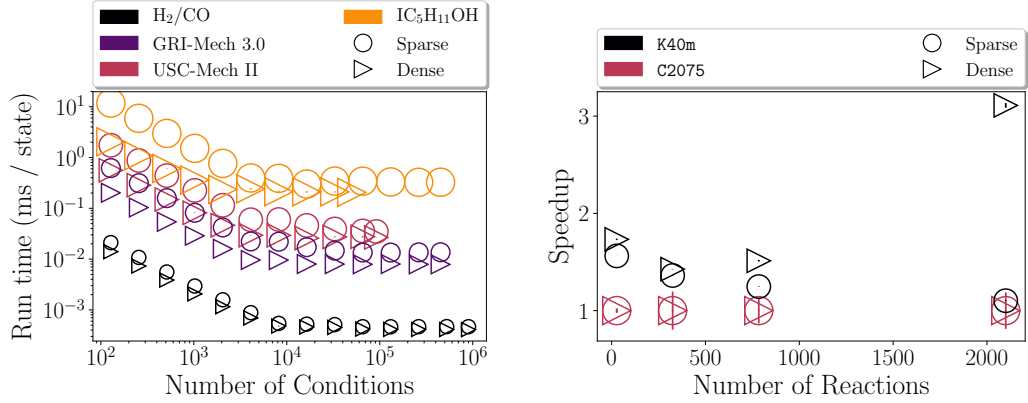


(a) Sparse Jacobian evaluation parallel scaling efficiency for Intel OpenCL/OpenMP on the **avx2** machine. (b) Dense Jacobian evaluation parallel scaling efficiency for Intel OpenCL/OpenMP on the **avx2** machine.

Figure 10: Strong parallel scaling efficiencies of sparse and dense Jacobian evaluations for the shallow-vectorized Intel OpenCL and OpenMP codes on the **avx2** CPU.

model, the dense shallow-vectorized OpenCL code scales slightly better than for source-term evaluation: 0.70–0.78 and 0.52–0.61 for four and eight cores, respectively (compared with 0.66–0.72 and 0.44–0.48 for shallow-vectorized OpenCL source-term evaluations). This results from the higher computational cost of Jacobian evaluation, and hence more available work per CPU core. As in Section 3.5.1, the shallow-vectorized dense Jacobian code running on 1 and 4 cores of the **avx2** machine performs roughly as fast as the OpenMP code on 4 and 16 cores, respectively. The parallel scaling efficiency of OpenMP on 16 cores (excluding isopentanol) is 0.63–0.71, similar to the shallow vectorization’s efficiency of 0.70–0.78 for 4 cores.

Figure 11a plots the mean run time per condition for the sparse and dense Jacobian evaluations on the **K40m** GPU. As in the species evaluation case, the mean run time per condition drops steadily, for both cases becoming roughly constant after just 3×10^3 states (except for H₂/CO, which levels off near 10^4 conditions). Sparse Jacobian evaluation is significantly slower for all models before the GPU becomes saturated (due to the indirect indexing lookup), but the performance gap between sparse and dense evaluations narrows past the saturation point. This is likely due to the ability to fit many more sparse Jacobian matrices in the **K40m**’s memory, as well as improved data locality/caching due to the smaller size of the sparse Jacobian. Figure 11b presents



(a) Mean run time per condition of sparse and dense Jacobian evaluations on the K40m GPU.

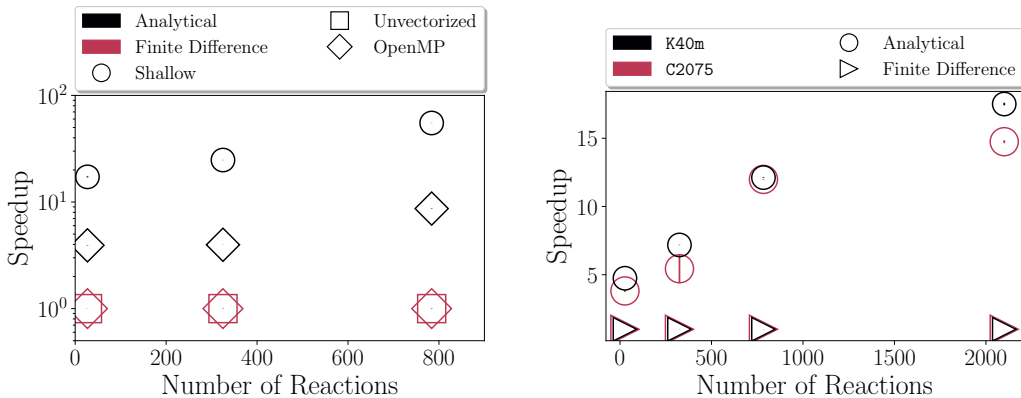
(b) Speedup of the K40m over C2075 GPUs for sparse and dense Jacobian evaluations, normalized per-Jacobian type (i.e., sparse versus dense).

Figure 11: The performance of sparse/dense Jacobian evaluation on the K40m and C2075 GPUs.

the speedup of the K40m over the C2075 GPU for sparse/dense Jacobian evaluation; sparse evaluation on the K40m is $1.10\text{--}1.59\times$ faster than on the C2075, while dense evaluation is $1.36\text{--}3.0\times$ faster. The speedup on the K40m decreases with increasing model size for the sparse formulation, but increases for larger models when dense; this likely results from the larger available memory of the K40m, and hence fewer data-transfer operations to/from the GPU.

Figure 12 compares the performance of the sparse analytical Jacobian with a sparse first-order finite-difference Jacobian on both the avx2 CPU and C2075/K40m GPUs. Figure 12a shows large speedups for both OpenMP and shallow-vectorized OpenCL; the analytical OpenMP Jacobian is $3.92\text{--}8.67\times$ faster, while the analytical OpenCL Jacobian achieves speedups of $17.22\text{--}55.11\times$. We excluded the isopentanol case here, since a single run of the sparse finite-difference Jacobian using either OpenCL or OpenMP took over 12 hours of run time. In addition the current finite-difference formulation breaks Intel’s auto-vectorizer, hence we compared OpenCL against the unvectorized OpenCL code (we did not prioritize fixing this issue, since we implemented the finite-difference Jacobian for comparison purposes only). Although we do not display the dense finite-difference Jacobian speedup

in Fig. 12a, the dense OpenCL and OpenMP analytical Jacobian codes outperform the finite-difference variants by even larger margins: $24.44\text{--}245.63\times$ for OpenCL and $9.68\text{--}112.73\times$ for OpenMP (these data do include the isopentanol model, though limited to 50,000 conditions as discussed previously). Figure 12b compares the sparse analytical and finite-difference Jacobians on the GPUs. The analytical Jacobian on the K40m and C2075 shows speedups of $3.81\text{--}17.60\times$ that increase with chemical model size; the K40m has a larger speedup than the C2075 for the isopentanol model ($17.60\times$ vs. $14.75\times$) due to its larger available memory. Although not pictured, the dense analytical Jacobian on the K40m GPU has larger speedups compared with the dense finite-difference Jacobian: $4.04\text{--}45.13\times$. The K40m GPU again shows significantly larger speedups over the C2075 for the isopentanol model ($45.13\times$ vs. $23.85\times$), further underscoring the effect of more available memory on the K40m.



(a) Speedup of the sparse analytical Jacobian versus finite-difference Jacobian evaluation on the `avx2` CPU, normalized per-language.

(b) Speedup of the analytical versus finite-difference Jacobian evaluation on both the K40m and C2075 GPUs, normalized per-GPU.

Figure 12: The performance of a sparse, first-order forward finite-difference Jacobian compared to the analytical Jacobian on the `avx2` CPU and both GPUs.

Figure 13 compares the performance of evaluating the dense analytical Jacobian of `pyJac-v2` with that of the previous version, `pyJac-v1` [54], on the `sse4.2` CPU and C2075 GPU. (We selected dense Jacobian evaluation for this comparison since it was the only type implemented in the previous version of `pyJac`.) On the `sse4.2` CPU, the `pyJac-v2` evaluates faster than `pyJac-v1`

for OpenMP for the larger chemical models; the static OpenMP code generated by `pyJac-v1` (see Section 2.3) is $1.79 \times$ faster for the H_2/CO model, and only $1.09 \times$ slower for the GRI-Mech 3.0 model. In contrast, the loop-based OpenMP code of `pyJac-v2` is $3.37\text{--}10.19 \times$ faster than `pyJac-v1` for the USC-Mech II and isopentanol models. The shallow-vectorized OpenCL `pyJac-v2` code is faster than `pyJac-v1` in all cases, achieving speedups of $1.37\text{--}19.56 \times$ that increase with model size. Figure 13b compares the performance of the `pyJac-v2` with `pyJac-v1` for evaluating dense analytical Jacobians on the C2075 GPU. As with the CPU, the static code of `pyJac-v1` is slightly faster for the H_2/CO model, but `pyJac-v2` outperforms `pyJac-v1` by $1.25\text{--}2.84 \times$ for the other models. Performance likely drops for the isopentanol model due to the limited number of conditions in dense evaluation for `pyJac-v2`, as noted earlier.

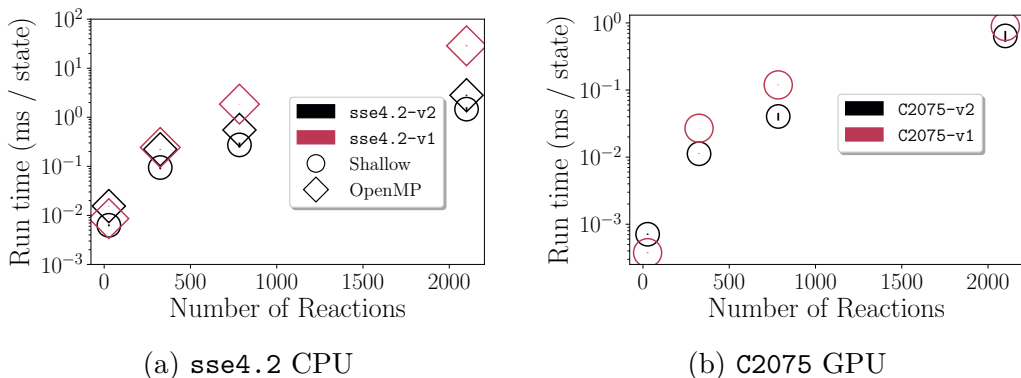


Figure 13: Performance comparison of dense Jacobian evaluation using `pyJac-v2` and `pyJac-v1` [54].

4. Practical notes on OpenCL use

While OpenCL provides a simple interface to enable cross-platform execution, and significant speedups were achieved via shallow-vectorized OpenCL code in this work, there are some serious potential pitfalls in its use. The closed-source OpenCL runtimes tested in this work (Intel and Nvidia) contain several bugs that result in compilation failures, simply incorrect vectorized machine code, or even segmentation faults. Further, these runtimes (in our experience) tend to be less responsive to fixing said bugs, with relatively infrequent new releases (or in Nvidia’s case, lack of changelogs/public records

of bug-fixes). On the other hand, the open-source OpenCL runtime used in this work, POCL, has far fewer implementation bugs, and when issues arise the community is very responsive to bug reports and user outreach; however, POCL fails to achieve vectorization as noted in Section 3.5.1.

To demonstrate the type of issue discussed, we created a minimum working example [77] that shows a failure of Nvidia’s OpenCL runtime corresponding to the GPU driver version 375.26 on a Tesla K40m GPU; simply changing to another runtime (e.g., Intel) produces the correct result—with no code changes or recompilation. This provides a particularly vexing problem for the programmer, since little can often be done to resolve the issue; thankfully, in this case we were able to upgrade the driver version to resolve the problem. Further, as noted throughout this work, certain code-generation patterns can break OpenCL execution/vectorization, e.g., the failure of POCL to achieve vectorization or Intel OpenCL’s failure to vectorize the finite-difference Jacobian. Indeed the vectorization process attempted by most OpenCL runtimes (and thus, reasons for incorrect/unvectorized code output) is obscured from the user, making reasoning about errors or performance trends quite difficult. Thankfully, `loo.py` allows relatively easy switching between output languages; the most significant code change requires building of the wrapper that initializes/transfers memory and calls the source-term/Jacobian kernel. These issues make it critical to provide adequate implementation details (e.g., runtime version, platform, etc., as given in Section 3.1) for codes utilizing OpenCL in order to enable reproducible results.

5. Conclusions

In this work, we developed automatically generated OpenCL codes for SIMD and SIMT-vectorized evaluations of thermochemical source-term and sparse/dense chemical kinetic Jacobian matrices. The main contributions of this work are:

- Deriving and verifying a new Jacobian formulation that greatly increases sparsity;
- Enabling vectorized execution on the CPU, GPU, and other accelerators; and
- Achieving significant speedups over a strictly parallel Jacobian and source-term evaluation on SIMD-enabled processors (e.g., CPUs).

These efforts are made publicly available (see [Appendix A](#)) via the open-source, high-performance, chemical kinetics code `pyJac`. The new molar-based formulation resulted in highly sparse chemical kinetic Jacobians, and allows selection of either the constant-volume or constant-pressure approximation. In addition, sparsity can be increased further by eliminating components associated with the bath gas, as discussed in [Section 3.4](#); this approximation is not a key feature of this work, and future efforts to incorporate more sophisticated Jacobian approximations would be a worthwhile endeavor.

We also demonstrated source-term and Jacobian evaluation for a range of chemical kinetic models [\[65–68\]](#) and multiple CPUs/GPUs ([Tables 1 and 2](#)). In addition to parallel OpenMP evaluation on the CPU, this work enabled the shallow-vectorized evaluation of the chemical-kinetic source terms and analytical Jacobian on both the CPU and GPU via OpenCL. Deep vectorization is possible on the Portable OpenCL (POCL) platform [\[59\]](#), but it yields no performance benefit as POCL did not achieve vectorizations for any execution pattern studied. Deep vectorization deserves further study with other platforms (e.g., CUDA).

We demonstrated significant speedups in shallow SIMD-vectorized execution over a parallel OpenMP code for evaluating the chemical-kinetic source terms and sparse/dense Jacobian: up to $4.09\times$, $9.44\times$, and $4.23\times$, respectively, on an `avx2`-capable CPU. Sparse Jacobians evaluate more slowly than dense Jacobians on all CPU/GPU platforms due to indirect lookup requirements in array indexing, but this adversely affects the shallow-vectorized OpenCL code less than OpenMP. Further, analytical Jacobians evaluate significantly faster than a first-order finite-difference-based approach on all platforms. Finally, we compared the performance of evaluating dense, analytical Jacobians in this new version of `pyJac` with the previous version [\[54\]](#). The OpenMP version evaluates moderately slower on the CPU for the smallest chemical model (e.g., $1.79\times$ on the `avx2` CPU), but significantly faster for the larger models—up to $10.19\times$. The shallow-vectorized OpenCL code runs faster than the previous version over all chemical models, reaching speedups of $19.56\times$.

The OpenMP code-generation is currently only capable of parallel execution, but extending this platform to shallow/deep-vectorizations (via `loopy` and compiler `#pragmas`) is a key priority going forwards since OpenMP is a standard library on most machines. In addition, CUDA [\[51\]](#) has been significantly more reliable in previous works [\[18, 24\]](#), while Intel’s open-source OpenCL alternative, ISPC [\[78\]](#), has been relatively stable and easy to

work with during preliminary efforts with the unit-testing discussed in Section 2.3. The current deep-vectorization formulation would be executable for both CUDA and ISPC targets, as these languages implement double-precision atomic operations, further recommending their use. It is also possible, particularly for the Intel OpenCL/POCL runtimes, that better performance/stability might be achieved using so-called “explicit” vectorization, i.e., through use of built-in vector types such as the `double8`. Specifically, this change could enable vectorization on the POCL runtime and might also enable deep vectorization on the Intel OpenCL runtime. Using OpenCL Image/CUDA Texture memory to accelerate the indirect lookup for sparse matrix evaluation should be investigated as well. Finally, future extensions of this work will include extending to additional target languages (e.g., vectorized OpenMP, CUDA, ISPC) to improve ease of use and reliability, improving the existing OpenCL targets (e.g., to enable meaningful deep-vectorized evaluation), and implementing reaction sorting [23] to improve SIMD efficiency (Appendix C).

One key component missing in this work is vectorized sparse/dense linear-algebra subroutines to maximize the performance of LU-factorization and matrix-vector multiplication (commonly used in implicit-integration techniques). Third-party/open-source options exist for some target languages, e.g., cuBLAS [79], clBLAS/clSPARSE [80] or SuperLU [20], but these do not necessarily cover all targets/required linear-algebra operations and, in the case of the CUDA/OpenCL libraries, are often optimized operation on one large matrix instead of many (relatively) smaller matrices. The extent to which these existing programs may be used needs to be assessed and missing operations should be implemented in `loo.py` to ensure easy switching between target languages and vectorization types.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under grants ACI-1534688 (Curtis and Sung) and ACI-1535065 (Niemeyer).

Appendix A. Availability of material

The results for this article were obtained using `pyJac v2.0.0b0` [81]. The most recent version of `pyJac` can be found at its GitHub repository: <https://github.com/SLACKHA/pyJac>. All figures, and the data and plotting scripts necessary to reproduce them, are available openly under the CC-BY license [82].

Appendix B. Jacobian error statistics per test platform

This section gives more detail on the results presented in Section 3.3, breaking down the reported error statistics per test platform/language. The error of the Intel OpenCL runtime is presented in Table B.1, the Portable OpenCL (POCL) runtime in Table B.2, OpenMP in Table B.3, and the Nvidia OpenCL runtime in Table B.4. POCL and OpenMP tend to have the smallest error norms, while Nvidia tends to have the largest; in particular the stringent filtered error norm $E_{C=10^{20}}$ is two orders of magnitude larger for the Nvidia runtime than the other test platforms with the H_2/CO and USC-Mech II models.

Model	$E_{\mathcal{L}}$	$E_{C=10^{20}}$	$E_{C=10^{15}}$
H_2/CO	1.455×10^{-14}	8.084×10^{-1}	1.907×10^{-5}
GRI-Mech 3.0	1.567×10^{-14}	1.469×10^{-7}	1.316×10^{-7}
USC-Mech II	9.632×10^{-15}	5.567×10^{-3}	1.704×10^{-7}
$\text{iC}_5\text{H}_{11}\text{OH}$	1.227×10^{-10}	1.363×10^{-3}	2.864×10^{-5}

Table B.1: Summary of Jacobian matrix verification results for the Intel OpenCL runtime. The reported error statistics are the maximum filtered relative error E_C and LAPACK error $E_{\mathcal{L}}$ over all vectorization patterns (Table 3), CONP/CONV, and sparse/dense Jacobians. The threshold for the filtered relative error is the same as reported in Section 3.3.

Model	$E_{\mathcal{L}}$	$E_{C=10^{20}}$	$E_{C=10^{15}}$
H ₂ /CO	1.456×10^{-14}	1.230×10^{-1}	3.951×10^{-6}
GRI-Mech 3.0	1.014×10^{-14}	1.890×10^{-7}	1.877×10^{-7}
USC-Mech II	9.632×10^{-15}	8.998×10^{-4}	1.201×10^{-8}
iC ₅ H ₁₁ OH	9.133×10^{-15}	1.723×10^{-5}	5.108×10^{-7}

Table B.2: Summary of Jacobian matrix verification results for the Portable OpenCL (POCL) runtime. The reported error statistics are the maximum filtered relative error E_C and LAPACK error $E_{\mathcal{L}}$ over all vectorization patterns (Table 3), CONP/CONV, and sparse/dense Jacobians. The threshold for the filtered relative error is the same as reported in Section 3.3.

Model	$E_{\mathcal{L}}$	$E_{C=10^{20}}$	$E_{C=10^{15}}$
H ₂ /CO	5.962×10^{-15}	3.614×10^{-2}	1.657×10^{-6}
GRI-Mech 3.0	1.297×10^{-15}	1.321×10^{-7}	1.316×10^{-7}
USC-Mech II	9.630×10^{-15}	4.185×10^{-4}	6.746×10^{-9}
iC ₅ H ₁₁ OH	6.131×10^{-15}	1.721×10^{-5}	5.108×10^{-7}

Table B.3: Summary of Jacobian matrix verification results for OpenMP execution. The reported error statistics are the maximum filtered relative error E_C and LAPACK error $E_{\mathcal{L}}$ over all vectorization patterns (Table 3), CONP/CONV, and sparse/dense Jacobians. The threshold for the filtered relative error is the same as reported in Section 3.3.

Model	$E_{\mathcal{L}}$	$E_{C=10^{20}}$	$E_{C=10^{15}}$
H ₂ /CO	1.862×10^{-14}	1.741×10^0	4.508×10^{-5}
GRI-Mech 3.0	1.489×10^{-14}	3.842×10^{-7}	3.687×10^{-7}
USC-Mech II	1.174×10^{-14}	1.119×10^{-2}	1.983×10^{-7}
iC ₅ H ₁₁ OH	8.602×10^{-15}	1.748×10^{-5}	5.109×10^{-7}

Table B.4: Summary of Jacobian matrix verification results for Nvidia OpenCL execution. The reported error statistics are the maximum filtered relative error E_C and LAPACK error $E_{\mathcal{L}}$ over all vectorization patterns (Table 3), CONP/CONV, and sparse/dense Jacobians. The threshold for the filtered relative error is the same as reported in Section 3.3.

Appendix C. SIMD efficiency Scaling Example

This simple example demonstrates how the SIMD efficiency of shallow-vectorized OpenCL source-term evaluation depends on the size of the chemical model in question, i.e., the amount of computational work per source-term evaluation. The base chemical model for this example was the isopentanol model [68] used throughout this article, and the same thermochemical state database described in Section 3.1 was used for source-term evaluation. As in Section 3.5 all reported results are based on 10 individual runs and in this example all cases were run on the `avx2` machine using the Intel OpenCL runtime.

Algorithm 1 A greedy selection algorithm to remove reactions from a base chemical model M , while preserving the number of active species.

Input: Base chemical model M with reactions R and species S

```
function DETERMINE SPECIES COUNT(active)
  for Species  $S_k$  in model  $M$  do
    species_rxn_count [ $k$ ]  $\leftarrow$  0
    for Reaction  $R_j$  in model  $M$  do
      if active [ $j$ ] and  $(|\nu'_{k,j}| + |\nu''_{k,j}|) > 0$  then
        species_rxn_count [ $k$ ]  $\leftarrow$  species_rxn_count [ $k$ ] + 1
  return species_rxn_count

procedure MODEL GENERATION( $M$ )
  active [ $j$ ]  $\leftarrow$  True for all reactions  $R_j$  in  $M$ 
  species_rxn_count  $\leftarrow$  DETERMINE SPECIES COUNT(active)
  while min (species_rxn_count)  $\geq 1$  do
    species_rxn_count  $\leftarrow$  DETERMINE SPECIES COUNT(active)
    for Reaction  $R_j$  in model  $M$  do
      rxn_count [ $j$ ]  $\leftarrow$  min $_{S_k \in R_j}$  (species_rxn_count [ $k$ ])
    remove_at  $\leftarrow$  argmax (rxn_count)
    active [remove_at]  $\leftarrow$  False
```

First, the reactions in the isopentanol model were converted to simple reversible Arrhenius reactions by either simply dropping third-body efficiency calculations (third-body enhanced reactions), using the high-pressure-limit coefficients (falloff/chemically-activated and P-Log reactions) or fitting Arrhenius parameters to the calculated rate constant at a fixed pressure (Chebyshev reactions). This conversion made the cost of reaction rate evaluation

roughly equivalent between all reactions in model, separating the effect of chemical model size from computational intensities of different reaction types on the SIMD efficiency. Next, a greedy reaction removal algorithm (Algorithm 1) generated a number of models ranging from 2100–186 reactions, in increments of 200 reactions (except the final increment from 200 to 186 reactions).

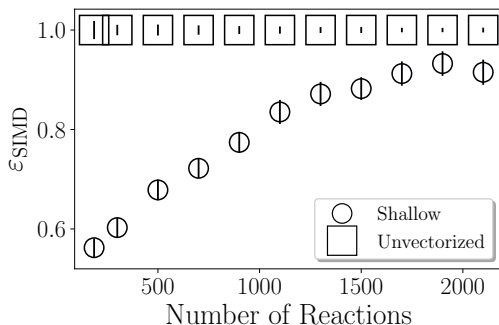


Figure C.1: The effect on SIMD efficiency of varying chemical model size.

To discern the effect of varying chemical model size on the SIMD efficiency, shallow-vectorized and unvectorized source-term evaluation performance tests were run. As demonstrated in Fig. C.1 the SIMD efficiency strongly depends on the generated model size and thus the amount of computational work per thermochemical state. In addition, the range of SIMD efficiency in this example (0.56–0.91) is larger than the range of SIMD efficiencies calculated for real chemical models, as seen in Fig. 8c. For smaller models—e.g., H_2/CO which had a SIMD efficiency of 0.6 on the `avx2` CPU—this suggests that the presence of more computationally intensive fall-off/chemically activated reactions in model can increase the SIMD efficiency. However, the base isopentanol model achieved a SIMD efficiency of only 0.78 on the `avx2` machine in Section 3.5.1, suggesting that more work could be done to optimize the source-term evaluations. In particular, it is likely that a reaction sorting method, such as suggested by Sewerin and Rigopoulos [23], would be particularly beneficial to reduce the number of vector gather/scatter/masking operations incurred during source-term evaluation.

- [1] T. Lu, C. K. Law, Toward accommodating realistic fuel chemistry in large-scale computations, *Prog. Energy Combust. Sci.* 35 (2) (2009) 192 – 215. [doi:10.1016/j.pecs.2008.10.002](https://doi.org/10.1016/j.pecs.2008.10.002).
- [2] C. Westbrook, C. Naik, O. Herbinet, W. Pitz, M. Mehl, S. Sarathy, H. Curran, Detailed chemical kinetic reaction mechanisms for soy and rapeseed biodiesel fuels, *Combust. Flame* 158 (4) (2011) 742 – 755, Special Issue on Kinetics. [doi:10.1016/j.combustflame.2010.10.020](https://doi.org/10.1016/j.combustflame.2010.10.020).
- [3] K. Spafford, J. Meredith, J. Vetter, J. Chen, R. Grout, R. Sankaran, Accelerating S3D: A GPGPU case study, in: *Euro-Par 2009 Parallel Process. Workshops*, LNCS 6043, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 122–131. [doi:10.1007/978-3-642-14122-5_16](https://doi.org/10.1007/978-3-642-14122-5_16).
- [4] T. Lu, C. K. Law, Linear time reduction of large kinetic mechanisms with directed relation graph: *n*-heptane and iso-octane, *Combust. Flame* 144 (1-2) (2006) 24–36. [doi:10.1016/j.combustflame.2005.02.015](https://doi.org/10.1016/j.combustflame.2005.02.015).
- [5] P. Pepiot-Desjardins, H. Pitsch, An efficient error-propagation-based reduction method for large chemical kinetic mechanisms, *Combust. Flame* 154 (1-2) (2008) 67–81. [doi:10.1016/j.combustflame.2007.10.020](https://doi.org/10.1016/j.combustflame.2007.10.020).
- [6] V. Hiremath, Z. Ren, S. B. Pope, A greedy algorithm for species selection in dimension reduction of combustion chemistry, *Combust. Theor. Model.* 14 (5) (2010) 619–652. [doi:10.1080/13647830.2010.499964](https://doi.org/10.1080/13647830.2010.499964).
- [7] K. E. Niemeyer, C.-J. Sung, M. P. Raju, Skeletal mechanism generation for surrogate fuels using directed relation graph with error propagation and sensitivity analysis, *Combust. Flame* 157 (9) (2010) 1760–1770. [doi:10.1016/j.combustflame.2009.12.022](https://doi.org/10.1016/j.combustflame.2009.12.022).
- [8] N. J. Curtis, K. E. Niemeyer, C.-J. Sung, An automated target species selection method for dynamic adaptive chemistry simulations, *Combust. Flame* 162 (4) (2015) 1358–1374. [doi:10.1016/j.combustflame.2014.11.004](https://doi.org/10.1016/j.combustflame.2014.11.004).
- [9] T. Lu, C. K. Law, Diffusion coefficient reduction through species bundling, *Combust. Flame* 148 (3) (2007) 117–126. [doi:10.1016/j.combustflame.2006.10.004](https://doi.org/10.1016/j.combustflame.2006.10.004).

- [10] S. S. Ahmed, F. Mauß, G. Moréac, T. Zeuch, A comprehensive and compact *n*-heptane oxidation model derived using chemical lumping, *Phys. Chem. Chem. Phys.* 9 (9) (2007) 1107–1126. doi:[10.1039/b614712g](https://doi.org/10.1039/b614712g).
- [11] P. Pepiot-Desjardins, H. Pitsch, An automatic chemical lumping method for the reduction of large chemical kinetic mechanisms, *Combust. Theor. Model.* 12 (6) (2008) 1089–1108. doi:[10.1080/13647830802245177](https://doi.org/10.1080/13647830802245177).
- [12] U. Maas, S. B. Pope, Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space, *Combust. Flame* 88 (3-4) (1992) 239–264. doi:[10.1016/0010-2180\(92\)90034-M](https://doi.org/10.1016/0010-2180(92)90034-M).
- [13] S.-H. Lam, D. A. Goussis, The CSP method for simplifying kinetics, *Int. J. Chem. Kinet.* 26 (4) (1994) 461–486. doi:[10.1002/kin.550260408](https://doi.org/10.1002/kin.550260408).
- [14] T. Lu, Y. Ju, C. K. Law, Complex CSP for chemistry reduction and analysis, *Combust. Flame* 126 (1–2) (2001) 1445–1455. doi:[10.1016/S0010-2180\(01\)00252-8](https://doi.org/10.1016/S0010-2180(01)00252-8).
- [15] X. Gou, W. Sun, Z. Chen, Y. Ju, A dynamic multi-timescale method for combustion modeling with detailed and reduced chemical kinetic mechanisms, *Combust. Flame* 157 (6) (2010) 1111–1121. doi:[10.1016/j.combustflame.2010.02.020](https://doi.org/10.1016/j.combustflame.2010.02.020).
- [16] T. Turányi, A. S. Tomlin, *Analysis of kinetic reaction mechanisms*, Springer, 2016.
- [17] D. A. Schwer, J. E. Tolsma, W. H. Green, P. I. Barton, On upgrading the numerics in combustion chemistry codes, *Combust. Flame* 128 (3) (2002) 270–291. doi:[10.1016/S0010-2180\(01\)00352-2](https://doi.org/10.1016/S0010-2180(01)00352-2).
- [18] K. E. Niemeyer, N. J. Curtis, C. J. Sung, pyJac: analytical Jacobian generator for chemical kinetics, *Comput. Phys. Comm.* 215 (2017) 188–203. doi:[10.1016/j.cpc.2017.02.004](https://doi.org/10.1016/j.cpc.2017.02.004).
- [19] Y. Gao, Y. Liu, Z. Ren, T. Lu, A dynamic adaptive method for hybrid integration of stiff chemistry, *Combust. Flame* 162 (2) (2015) 287–295. doi:[10.1016/j.combustflame.2014.07.023](https://doi.org/10.1016/j.combustflame.2014.07.023).
- [20] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, J. W. H. Liu, A supernodal approach to sparse partial pivoting, *SIAM J. Matrix Anal. Appl.* 20 (3) (1999) 720–755. doi:[10.1137/S0895479895291765](https://doi.org/10.1137/S0895479895291765).

- [21] Y. Shi, W. H. Green, H.-W. Wong, O. O. Oluwole, Accelerating multi-dimensional combustion simulations using GPU and hybrid explicit/implicit ODE integration, *Combust. Flame* 159 (7) (2012) 2388–2397. [doi:10.1016/j.combustflame.2012.02.016](https://doi.org/10.1016/j.combustflame.2012.02.016).
- [22] K. E. Niemeyer, C.-J. Sung, Accelerating moderately stiff chemical kinetics in reactive-flow simulations using GPUs, *J. Comput. Phys.* 256 (2014) 854–871. [doi:10.1016/j.jcp.2013.09.025](https://doi.org/10.1016/j.jcp.2013.09.025).
- [23] F. Sewerin, S. Rigopoulos, A methodology for the integration of stiff chemical kinetics on GPUs, *Combust. Flame* 162 (4) (2015) 1375–1394. [doi:10.1016/j.combustflame.2014.11.003](https://doi.org/10.1016/j.combustflame.2014.11.003).
- [24] N. J. Curtis, K. E. Niemeyer, C.-J. Sung, An investigation of GPU-based stiff chemical kinetics integration methods, *Combust. Flame* 179 (2017) 312–324. [doi:10.1016/j.combustflame.2017.02.005](https://doi.org/10.1016/j.combustflame.2017.02.005).
- [25] C. P. Stone, A. T. Alferman, K. E. Niemeyer, Accelerating finite-rate chemical kinetics with coprocessors: comparing vectorization methods on GPUs, MICs, and CPUs, *Comput. Phys. Comm.* 226 (2018) 18–29. [doi:10.1016/j.cpc.2018.01.015](https://doi.org/10.1016/j.cpc.2018.01.015).
- [26] H. N. Khan, D. A. Hounshell, E. R. Fuchs, Science and research policy at the end of Moore’s law, *Nat. Electron.* 1 (1) (2018) 14–21. [doi:10.1038/s41928-017-0005-9](https://doi.org/10.1038/s41928-017-0005-9).
- [27] J. E. Stone, D. Gohara, G. Shi, OpenCL: A parallel programming standard for heterogeneous computing systems, *IEEE Des. Test* 12 (3) (2010) 66–73. [doi:10.1109/MCSE.2010.69](https://doi.org/10.1109/MCSE.2010.69).
- [28] E. Lindholm, J. Nickolls, S. Oberman, J. Montrym, NVIDIA Tesla: A unified graphics and computing architecture, *IEEE micro* 28 (2) (2008) 39–55. [doi:10.1109/MM.2008.31](https://doi.org/10.1109/MM.2008.31).
- [29] C. Safta, H. N. Najm, O. M. Knio, TChem - a software toolkit for the analysis of complex kinetic models, Tech. Rep. SAND2011-3282, Sandia National Laboratories (May 2011).
- [30] N. J. Curtis, K. E. Niemeyer, Fileset for testing thread-safety of TChem, figshare (Jan. 2017). [doi:10.6084/m9.figshare.4563982.v1](https://doi.org/10.6084/m9.figshare.4563982.v1).

- [31] M. R. Youssefi, Development of analytic tools for computational flame diagnostics, Master’s thesis, University of Connecticut, http://digitalcommons.uconn.edu/gs_theses/145/ (Aug. 2011).
- [32] F. Bisetti, Integration of large chemical kinetic mechanisms via exponential methods with Krylov approximations to Jacobian matrix functions, *Combust. Theor. Model.* 16 (3) (2012) 387–418. doi:10.1080/13647830.2011.631032.
- [33] F. Perini, E. Galligani, R. D. Reitz, An analytical Jacobian approach to sparse reaction kinetics for computationally efficient combustion modeling with large reaction mechanisms, *Energy Fuels* 26 (8) (2012) 4804–4822. doi:10.1021/ef300747n.
- [34] M. A. Hansen, J. C. Sutherland, On the consistency of state vectors and Jacobian matrices, *Combust. Flame* 193 (2018) 257–271. doi:j.combustflame.2018.03.017.
- [35] T. Dijkmans, C. M. Schietekat, K. M. Van Geem, G. B. Marin, GPU based simulation of reactive mixtures with detailed chemistry in combination with tabulation and an analytical Jacobian, *Comput. Chem. Eng.* 71 (2014) 521–531. doi:10.1016/j.compchemeng.2014.09.016.
- [36] M. Bauer, S. Treichler, A. Aiken, Singe: Leveraging warp specialization for high performance on GPUs, *SIGPLAN Not.* 49 (8) (2014) 119–130. doi:10.1145/2692916.2555258.
- [37] T. F. LU, C. S. YOO, J. H. CHEN, C. K. LAW, Three-dimensional direct numerical simulation of a turbulent lifted hydrogen jet flame in heated coflow: a chemical explosive mode analysis, *J. Fluid Mech.* 652 (2010) 45–64. doi:10.1017/S002211201000039X.
- [38] Y. Shi, W. H. Green, H.-W. Wong, O. O. Oluwole, Redesigning combustion modeling algorithms for the graphics processing unit (GPU): Chemical kinetic rate evaluation and ordinary differential equation integration, *Combust. Flame* 158 (5) (2011) 836–847. doi:10.1016/j.combustflame.2011.01.024.
- [39] K. E. Niemeyer, C.-J. Sung, C. G. Fotache, J. C. Lee, Turbulence-chemistry closure method using graphics processing units: a preliminary

- test, in: Fall 2011 Technical Meeting of the Eastern States Section of the Combust. Institute. [doi:10.6084/m9.figshare.3384964](https://doi.org/10.6084/m9.figshare.3384964).
- [40] H. P. Le, J.-L. Cambier, L. K. Cole, GPU-based flow simulation with detailed chemical kinetics, *Comput. Phys. Comm.* 184 (3) (2013) 596–606. [doi:10.1016/j.cpc.2012.10.013](https://doi.org/10.1016/j.cpc.2012.10.013).
 - [41] C. P. Stone, R. L. Davis, Techniques for solving stiff chemical kinetics on graphical processing units, *J. Propul. Power* 29 (4) (2013) 764–773. [doi:10.2514/1.B34874](https://doi.org/10.2514/1.B34874).
 - [42] P. N. Brown, G. D. Byrne, A. C. Hindmarsh, VODE: a variable-coefficient ODE solver, *SIAM J. Sci. Stat. Comput.* 10 (5) (1989) 1038–1051. [doi:10.1137/0910062](https://doi.org/10.1137/0910062).
 - [43] N. Yonkee, J. C. Sutherland, PoKiTT: Exposing task and data parallelism on heterogeneous architectures for detailed chemical kinetics, transport, and thermodynamics calculations, *SIAM Journal on Scientific Computing* 38 (5) (2016) S264–S281. [doi:10.1137/15M1026237](https://doi.org/10.1137/15M1026237).
 - [44] G. Wanner, E. Hairer, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, 2nd Edition, Springer-Verlag, Berlin, 1996. [doi:10.1007/978-3-642-05221-7](https://doi.org/10.1007/978-3-642-05221-7).
 - [45] M. Hochbruck, C. Lubich, H. Selhofer, Exponential integrators for large systems of differential equations, *SIAM J. Sci. Comput.* 19 (5) (1998) 1552–1574. [doi:10.1137/S1064827595295337](https://doi.org/10.1137/S1064827595295337).
 - [46] M. Hochbruck, A. Ostermann, J. Schweitzer, Exponential Rosenbrock-type methods, *SIAM J. Numer. Anal.* 47 (1) (2009) 786–803. [doi:10.1137/080717717](https://doi.org/10.1137/080717717).
 - [47] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, C. S. Woodward, Sundials: Suite of nonlinear and differential/algebraic equation solvers, *ACM Trans. Math. Softw.* 31 (3) (2005) 363–396. [doi:10.1145/1089014.1089020](https://doi.org/10.1145/1089014.1089020).
 - [48] J. C. Linford, J. Michalakes, M. Vachharajani, A. Sandu, Automatic generation of multicore chemical kernels, *IEEE Trans. Parallel Distrib. Syst.* 22 (1) (2011) 119–131. [doi:10.1109/TPDS.2010.106](https://doi.org/10.1109/TPDS.2010.106).

- [49] A. Kroshko, R. J. Spiteri, Efficient SIMD solution of multiple systems of stiff IVPs, *J. Comput. Sci* 4 (5) (2013) 377–385. doi:[10.1016/j.jocs.2012.08.017](https://doi.org/10.1016/j.jocs.2012.08.017).
- [50] J. Gray, P. Shenoy, Rules of thumb in data engineering, in: *Data Engineering, 2000. Proceedings. 16th International Conference on*, IEEE, 2000, pp. 3–10.
- [51] NVIDIA, CUDA C programming guide, version 9.0, https://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf (Jan 2018).
- [52] S. R. Turns, *An Introduction to Combustion: Concepts and Applications*, 3rd Edition, McGraw-Hill, New York, 2012.
- [53] A. Klöckner, in: *Proc. ARRAY '14: ACM SIGPLAN Workshop Libr., Lang., Compil. Array Progr., Assoc. Comput. Mach., Edinburgh, Scotland.*, 2014. doi:[10.1145/2627373.2627387](https://doi.org/10.1145/2627373.2627387).
- [54] N. J. Curtis, K. E. Niemeyer, *pyJac v1.0.6* (Feb. 2018). doi:[10.5281/zenodo.1182789](https://doi.org/10.5281/zenodo.1182789).
- [55] A. Klöckner, N. Pinto, Y. Lee, B. Catanzaro, P. Ivanov, A. Fasih, PyCUDA and PyOpenCL: A scripting-based approach to GPU runtime code generation, *Parallel Comput.* 38 (3) (2012) 157–174. doi:[10.1016/j.parco.2011.09.001](https://doi.org/10.1016/j.parco.2011.09.001).
- [56] D. G. Goodwin, H. K. Moffat, R. L. Speth, *Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes*, <http://www.cantera.org>, version 2.3.0 (2017). doi:[10.5281/zenodo.170284](https://doi.org/10.5281/zenodo.170284).
- [57] R. J. Hogan, Fast reverse-mode automatic differentiation using expression templates in C++, *ACM Trans. Math. Software* 40 (4) (2014) 26. doi:[10.1145/2560359](https://doi.org/10.1145/2560359).
- [58] R. J. Hogan, *Adept v1.1*, Available at <https://github.com/rjhogan/Adept> (Jun. 2015).
- [59] P. Jääskeläinen, C. S. de La Lama, E. Schnetter, K. Raiskila, J. Takala, H. Berg, pocl: A performance-portable OpenCL implementation, *Int. J. Parallel Program.* 43 (5) (2015) 752–785. doi:[10.1007/s10766-014-0320-y](https://doi.org/10.1007/s10766-014-0320-y).

- [60] L. Dagum, R. Menon, OpenMP: an industry standard API for shared-memory programming, *Computational Sci. & Engineering*, IEEE 5 (1) (1998) 46–55. doi:10.1109/99.660313.
- [61] G. Travis CI, Travis CI - Test and Deploy Your Code with Confidence, <https://about.travis-ci.com/> (2018).
- [62] Intel® Corporation, OpenCL™ drivers and runtimes for Intel® architecture, https://software.intel.com/en-us/articles/opencl-drivers#latest_CPU_runtime (2018).
- [63] C. Lattner, V. Adve, LLVM: A compilation framework for lifelong program analysis & transformation, in: *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization*, CGO '04, IEEE Computer Society, Washington, DC, USA, 2004, pp. 75–. URL <http://dl.acm.org/citation.cfm?id=977395.977673>
- [64] MichaelE1000, Bug report on NVIDIA forums, [NVIDIA Devtalk Forums](#), accessed 03-06-18.
- [65] M. P. Burke, M. Chaos, Y. Ju, F. L. Dryer, S. J. Klippenstein, Comprehensive H₂/O₂ kinetic model for high-pressure combustion, *Int. J. Chem. Kinet.* 44 (7) (2011) 444–474. doi:10.1002/kin.20603.
- [66] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. Gardiner, V. V. Lissianski, Z. Qin, GRI-Mech 3.0, http://www.me.berkeley.edu/gri_mech/.
- [67] H. Wang, X. You, A. V. Joshi, S. G. Davis, A. Laskin, F. Egolfopoulos, C. K. Law, USC Mech Version II. High-temperature combustion reaction model of H₂/CO/C₁–C₄ compounds, http://ignis.usc.edu/USC_Mech_II.htm (May 2007).
- [68] S. M. Sarathy, S. Park, B. W. Weber, W. Wang, P. S. Veloo, A. C. Davis, C. Togbe, C. K. Westbrook, O. Park, G. Dayma, Z. Luo, M. A. Oehlschlaeger, F. N. Egolfopoulos, T. Lu, W. J. Pitz, C.-J. Sung, P. Dagaut, A comprehensive experimental and modeling study of iso-pentanol combustion, *Combust. Flame* 160 (12) (2013) 2712–2728. doi:10.1016/j.combustflame.2013.06.022.

- [69] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, LAPACK Users' Guide, 3rd Edition, SIAM, Philadelphia, PA, 1999.
- [70] M. J. McNenly, R. A. Whitesides, D. L. Flowers, Faster solvers for large kinetic mechanisms using adaptive preconditioners, Proceedings of the Combustion Institute 35 (1) (2015) 581–587. doi:[10.1016/j.proci.2014.05.113](https://doi.org/10.1016/j.proci.2014.05.113).
- [71] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. V. der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition, SIAM, Philadelphia, PA, 1994, note: the section on [sparse matrices formats](#).
- [72] M. Babej, P. Jääskeläinen, Debugging auto vectorizer, Private Communication, archived on [POCL mailing list](#) (Feb. 2018).
- [73] NVIDIA, Achieved occupancy, [Achieved Occupancy](#) (Mar 2018).
- [74] Intel® Corporation, Using vector data types, <https://software.intel.com/en-us/node/540561>, accessed on 02/19/18.
- [75] Intel® Corporation, Vectorizer knobs, <https://software.intel.com/en-us/node/540560>, accessed on 02/19/18.
- [76] G. G. Howes, Parallel performance and optimization, http://homepage.physics.uiowa.edu/~ghowes/teach/ihpc10/lec/ihpc10Lec_PerformanceHPC10.pdf, slides from Iowa High Performance Computing Summer School, University of Iowa, 08/2010 - Accessed on 02/19/18.
- [77] N. J. Curtis, A minimum working example showing the failure of simple OpenCL code on the NVIDIA Linux x64 Tesla 375.26 Driver, <https://figshare.com/s/03aa9064aa6fe3508d3d> (jun 2018). doi:[10.6084/m9.figshare.6533915](https://doi.org/10.6084/m9.figshare.6533915).
- [78] M. Pharr, W. R. Mark, ispc: A SPMD compiler for high-performance CPU programming, in: Innovat. Parallel Comput. (InPar), 2012, 2012, pp. 1–13. doi:[10.1109/InPar.2012.6339601](https://doi.org/10.1109/InPar.2012.6339601).

- [79] NVIDIA Corporation, Dense linear algebra on gpus, <https://developer.nvidia.com/cublas>, accessed: 03-12-18.
- [80] clMathLibraries, clmathlibraries, <https://github.com/clMathLibraries>, accessed: 03-12-18.
- [81] N. J. Curtis, K. E. Niemeyer, pyJac v2.0.0-beta.0 (Jun. 2018). doi: [10.5281/zenodo.1289979](https://doi.org/10.5281/zenodo.1289979).
- [82] N. J. Curtis, K. E. Niemeyer, C.-J. Sung, Data, plotting scripts, and figures for “using simd and simt vectorization to evaluate sparse chemical kinetic jacobian matrices and thermochemical source terms”. doi: [10.6084/m9.figshare.6534146](https://doi.org/10.6084/m9.figshare.6534146).