ACCOLADES: A Scalable Workflow Framework for Large-Scale Simulation and Analyses of Automotive Engines

Shashi M. Aithal(✉) and Stefan M. Wild

Argonne National Laboratory, Argonne, IL 60439, USA
{aithal,wild}@anl.gov

Abstract. Analysis and optimization of simulation-generated data have myriads of scientific and industrial applications. Fuel consumption and emissions over the entire drive cycle of a large fleet of vehicles is an example of such an application and the focus of this study. Temporal variation of fuel consumption and emissions in an automotive engine are functions of over twenty variables. Determining relationships between fuel consumption or emissions and the dependent variables plays a crucial role in designing an automotive engine. This paper describes the development of ACCOLADES (Advanced Concurrent COmputing for LArge-scale Dynamic Engine Simulations), a scalable workflow framework that exploits the task parallelism inherent in such analyses by using large-scale computing. Excellent weak scaling is observed on 4,096 cores of both an Intel Sandy Bridge-based cluster and a Blue-Gene/Q supercomputer.

Keywords: Workflow management · Industrial simulations · Large-scale vehicle simulations · Task parallelism

1 Introduction

Discrete time series occur in many scientific and industrial applications [7,9,11,13]. Examples of these applications include solar radiation, temporal variations of the load requirements on a power-grid, temperature variation of power-generating equipment and variation in the price of a commodity, among others. An observed value of a time series is typically a function of several variables; developing an understanding of the effect of the variables on the observed value is a computationally intensive task. Furthermore, optimization of time-averaged or integrated values of these functions often requires analyses of large datasets.

Fuel consumption and emissions over the entire drive cycle of a large fleet of cars is an example of such a problem, and hence the focus of this study. Temporal variation of fuel consumption and emissions in an automotive engine

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under Contract DE-AC02-06CH11357.

J.M. Kunkel and T. Ludwig (Eds.): ISC High Performance 2015, LNCS 9137, pp. 87–95, 2015.
DOI: 10.1007/978-3-319-20119-1_7

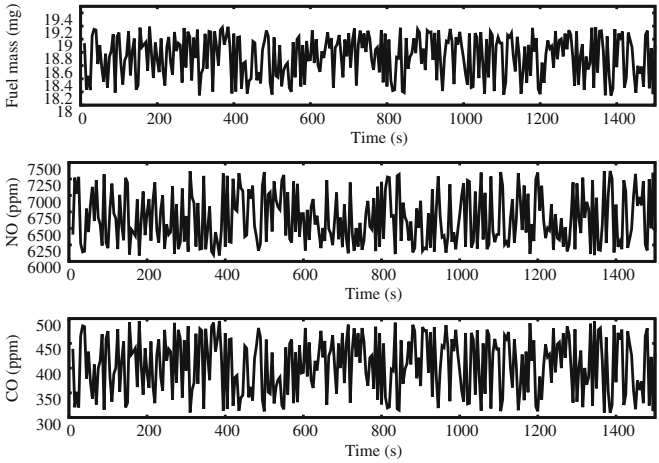

Fig. 1. Typical time series data for fuel mass (*top*), NO emissions (*middle*), and CO emissions (*bottom*) as generated by pMODES over a single drive cycle.

can be functions of over twenty independent variables, including engine speed (i.e., RPM), torque, type of fuel/additive, air-to-fuel ratio, ambient temperature, inlet pressure, humidity, ignition and valve timings, and driving conditions (e.g., city or highway). Deriving correlations between the observed values (such as fuel consumption and emissions) and the independent variables plays a crucial role during the design and development stages of an automotive engine. For instance, a study on the effect of the inflow air temperature on the fuel consumption and (NO, CO, soot, and unburned hydrocarbons) emissions might consist of four different drive cycles and five different temperatures (e.g., expressed as a percentage of the nominal temperature). Such a study would result in twenty different time series for fuel consumption and eighty time series for emissions (i.e., one for each type of emission). A typical drive cycle of an automotive engine has a duration of 25–30 min; for data sampled every second, one obtains approximately 1,500 data points per drive cycle. Dynamometer testing and measurements (called "dyno testing") are usually conducted to obtain data for engine performance and emissions. Numerical simulations can be used to complement dyno data or to estimate engine performance and emissions during the engine design process.

Figure 1 shows an instance of the temporal variation of fuel flow along with the computed temporal variation of nitric oxide (NO) and carbon monoxide (CO) emissions. For each sampled data point, which represents one (compression and expansion) engine cycle, engine state variables (e.g., temperature, pressure, andfuel-air mixture combination) are computed over 360 crank angle degrees (CAD) in intervals of roughly 0.5 CAD. These engine state variables are needed in order to compute the engine performance (e.g., torque and power) and engine-out emissions (e.g., NO and CO). Hence, each drive cycle requires the evaluation of over a million engine ($\approx$1,500 $\times$ 720) CAD.

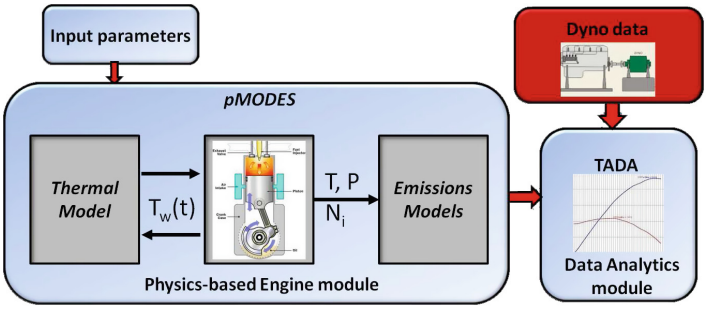

Fig. 2. Block diagram showing the structure of ACCOLADES.

The above example represents a simplified case wherein the effect of variation of a single parameter on the engine performance and emissions is studied. Typical fleet studies require the simultaneous variation of multiple design variables over specified ranges for a larger number of drive cycles, resulting in a large set of input configurations. For instance, if one were to consider the effect of variation of four design parameters (e.g., inlet pressure, inlet temperature, humidity, and engine RPM) with four different values for each of these design parameters over sixteen different drive cycles, one would need $(4^4 \times 16 = 4,096)$ different independent cases. Each of these 4,096 cases would require 1,500 engine-cycle evaluations. Conducting large parametric sweeps on the drive cycles of a fleet of cars with varying combinations of operating conditions places stringent demands on the required computational resources. Furthermore, analyses of the results of these large-scale simulations present significant challenges from a data-analytics standpoint. Transient multidimensional numerical simulation of a single engine cycle (360 CAD) running on 24 to 48 cores (approaching the strong scaling limit for physically meaningful grid sizes) can take several hours to days, depending on the complexity of the physical models used. Hence, conducting multi-cycle simulations for the scenario described above would require enormous computational resources, and thus precluding their use for initial design/development studies or analyses of large transients.

Physics-based reduced-order models, which capture the temporal variation, for example, of average engine temperature, pressure, and mixture composition, are ideally suited for such large-scale studies. Given the wide range of operating conditions (engine speed, load, equivalence ratio, etc.) the reduced-order models have to be robust and fast in order to compute emissions and performance at real-time speeds. Real-time analysis would require a typical data point in any given drive cycle to be computed in approximately 250–30 ms.

This paper describes the development of ACCOLADES (Advanced Concurrent COmputing for LArge-scale Dynamic Engine Simulations), a scalable workflow management framework that enables automotive design engineers to exploit the task parallelism inherent in the study of such systems using large-scale computing (e.g., GPGPUs, multicore architectures, or the cloud). As shown

in Fig. 2 and detailed in Sect. 2, ACCOLADES consists of two main components, pMODES (parallel Multi-fuel Otto Diesel Engine Simulator) and TADA (Toolkit for Advanced Data Analytics). pMODES is a fast, robust, physics-based reduced-order engine simulator that can concurrently compute the performance and emissions of the various parametric cases required for a vehicle fleet simulation. TADA is a data analytics toolbox used to post-process the results generated by pMODES or directly from dyno data.

Although large-scale system-level optimization has been performed for military vehicles [5,10], to the author's knowledge, this work is the first to implement physics-based engine models for large-scale analysis of a fleet of cars. As illustrated by our results in Sect. 3, ACCOLADES can be used in the design and conceptual analyses phase of new engine systems and can streamline workflow management in the analyses of large amounts of data obtained in dyno tests for various engine operating conditions.

2 Main Components of ACCOLADES

ACCOLADES consists of the reduced-order engine simulator p-MODES and the data analytics toolbox TADA.

2.1 pMODES

pMODES is used to compute the temporal variation of various engine parameters such as pressure, temperature and mixture composition for each CAD over an entire drive cycle. The energy equation shown in Eq. (1) describes the relationship between the engine crank-angle θ and pressure.

$$\frac{dP(\theta)}{d\theta} = \frac{\gamma - 1}{V(\theta)}(Q_{in} - Q_{loss}) - \gamma \frac{P(\theta)}{V(\theta)} \frac{dV}{d\theta} \tag{1}$$

Solution of this equation yields the temporal variation of cylinder pressure for a given set of operating conditions (such as load, combustion duration, fuel type, engine RPM, etc). The instantaneous values of temperature and composition of the burned and unburned gas zones can be obtained from the instantaneous value of computed pressure. Knowing the instantaneous temperature, pressure and composition of the burned zone enables the computation of emissions such as NO, CO, soot, and unburned hydrocarbons using simplified reduced chemistry models. Details of these models and the solution procedure are discussed in Ref. [2]. Instantaneous values of equilibrium concentrations of the combustion products are needed to compute various emissions. Computation of these equilibrium concentrations pose serious numerical challenges on account of the stiffness of the system of nonlinear equations describing the formation of combustion products. References [1,3] discuss the details of the computation procedure and steps taken to ensure a fast, robust solution. Following the solution procedure discussed above enables one to obtain temporal variation of emissions such as NO and CO for a given fuel input profile. Figure 1 shows the NO and CO emissions for a single-cylinder gasoline engine obtained using pMODES.

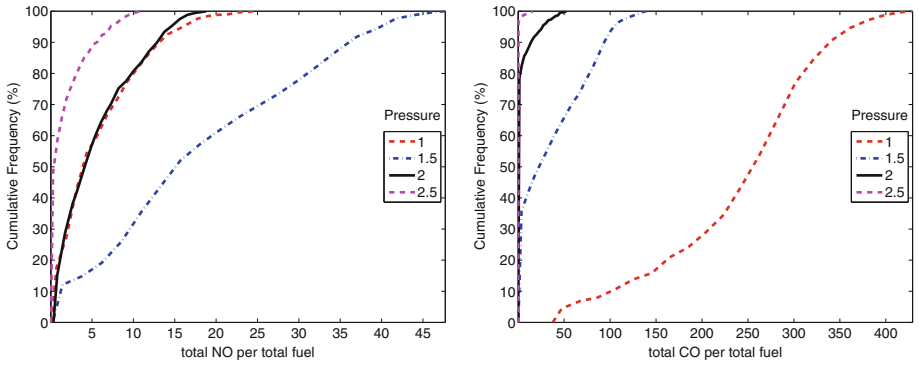

Fig. 3. Cumulative distributions of NO (*left*) and CO (*right*) emissions for different initial cylinder pressure conditions. Each curve shows the percentage of drive cycle runs for which the NO/CO is at or below the value given on the horizontal axis.

2.2 TADA

The Toolkit for Advanced Data Analytics (TADA) provides a framework for post-processing of experimental- and simulation-generated time series data. Here we overview some of the operations possible with TADA.

Whether from physical experiment or numerical simulation, TADA takes as input time series data $\{f_o(x;t;\theta_t(x)) : o = 1,\ldots,O; t = 1,\ldots,T\}$, where o indexes O different dependent variable outputs, t indexes T time periods $\tau_1 < \tau_2 < \cdots < \tau_T$, $x \in \mathbb{R}^n$ parameterizes the independent design and operational variables, and $\theta_t(x) \in \mathbb{R}^m$ denotes the state variables at time τ_t with input x.

Typical data analysis operations on these sets of time series data include

Filtering to extract basic statistics and identify input configurations of interest. For example, one can determine peak temperatures and pressures in order to characterize engine damage; peaks can be computed for each configuration, or all peaks above a threshold can be extracted.

Empirical distribution characterization to provide cross-configuration information. Such distributions can be used, for example, to determine fleet-wide fuel economy [12] or to characterize emissions as a function of ambient pressure as is done in Fig. 3 for the case study in Sect. 3.

Sensitivity analysis to analyze how operating conditions or other independent variables effect observables of interest. Sensitivity analysis can be used, for example to determine fleet-wide implications for performance and emissions of increased adoption of novel fuel types or additives.

Tradeoff visualization and analysis can be used to flag a configuration that is worse in all metrics of interest than some other configuration. Such analyses can also be used, for example, to identify vehicle configurations that sacrifice little in terms of performance while providing substantial gains in fuel economy.

We expect that the capabilities in TADA will be fully used as one "closes the loop" between the pMODES simulation and analysis for purposes of simulation-based design optimization [4] or optimal experimental design to determine configurations that should be tested on a dyno. In this view, TADA can be used to generate input configurations and/or in order to optimize a design objective of interest. Distributional information can be used to generate scenarios (e.g., ambient or operating conditions, drive cycle variations) for use in sample average approximation for optimization under uncertainty [8]. Similarly, tradeoff analysis forms the basis for simultaneously optimizing multiple conflicting objectives [6,14], such as performance and engine lifetime/reliability.

3 Results and Discussions

As an illustrative example, we discuss the simulation of a single-cylinder gasoline engine operating at 1,100 RPM, wherein the inlet gas temperature, air humidity, initial cylinder pressure, and exhaust gas recirculation (EGR) fraction are varied for realistic engine operating conditions. Each of the parameters have four values and for each configuration sixteen different drive cycles are considered, leading to 4,096 individual configurations (or parametric cases). The inlet gas temperature is varied from 28 to 31 °C in steps of 1 °C, the initial cylinder pressure is varied from 0.88 atm to 1.0 atm, the relative humidity is varied from 0 to 100 %, and the EGR fraction is varied from 0 to 3 %. These 4,096 parametric cases, each with 1,500 temporal data points in the drive cycle, were run on IBM Blue Gene/Q (BG/Q) and Sandy Bridge clusters at Argonne National Laboratory. The BG/Q supercomputer (called "Mira") is equipped with 786,432 cores, 768 TB of memory and has a peak performance of 10 petaflops. Each compute node has a PowerPC A2 1600 MHz processor containing 16 cores, each with 4 hardware threads, running at 1.6 GHz, and 16 GB of DDR3 memory. The Sandy Bridge cluster (called "Blues") is a 2.6 GHz, 4960 processors system with 16 cores per compute node and 4 GB memory per core. These systems were chosen to ensure portability of the code on different architectures (and compilers) and also to compare and contrast the relative performance of ACCOLADES on these machines. Each of the cases considered in the study was assigned to one MPI rank on the machine. Each MPI rank read its input data (i.e., operating conditions such as initial pressure, humidity, inlet air temperature, and EGR fraction) from a separate input file and wrote two different files: (a) the computed solution (e.g., emissions, maximum temperature, pressure, exhaust temperature and pressure, peak ion current, location of peak ion current) and (b) operating conditions to its own uniquely named file. This methodology was chosen to ensure no communication between different case configurations (thus ensuring task parallelism), and also to facilitate data analytics by TADA. A typical case's solution file was 306 kB while the file containing information about each of 1,500 data points was 351 kB. Since each of the parametric cases are independent of the others, increasing the number of cases directly proportional to the number of cores yields a weak scaling study. The study was run both with no optimization

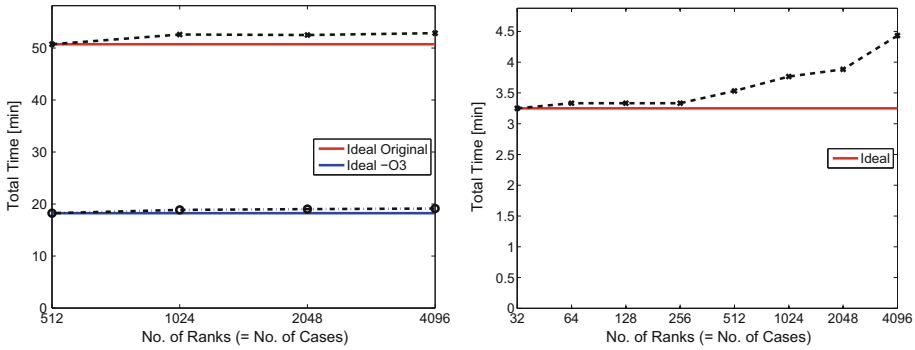

Fig. 4. Weak scaling results on the BlueGene/Q machine Mira (*left*) and the Sandy Bridge-based machine Blues (*right*).

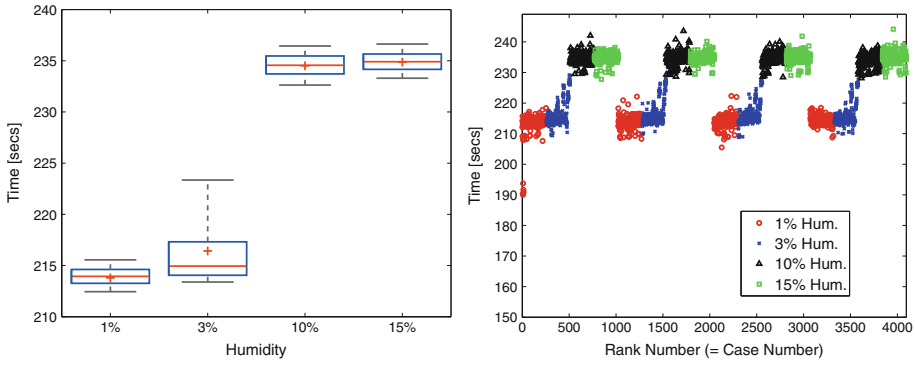

Fig. 5. Case timings on Sandy Bridge-based machine Blues: (*left*) global sensitivity analysis (outliers removed) demonstrating effect of humidity input on per case timing; (*right*) case timings as a function of rank number show a cyclic pattern associated with varying humidity input.

and with 'O3' optimization option on both machines. The Intel 13.1 compiler was used on Blues whereas the xl.legacy.ndebug (libraries with MPICH compiled with the XL compilers) was used on Mira.

Figure 4 shows the scaling for Blues and Mira (with and without optimization). Excellent weak scaling is seen on both machines. It was seen that the optimization level did not change the overall compute time on Blues (hence it is not shown in Fig. 4), whereas using -O3 level optimization on Mira reduced the computational time by a factor of nearly 2.7. We attribute the slight increase in overall computational time as the number of cores (and thus cases) increases primarily to imbalances in individual case solution times and to increased contention for the I/O operations. Furthermore, we see greater imbalances across cases for Blues than for Mira.

Deeper analysis of the timings associated with the 4,096 cases provides insight to the scaling behavior seen. For the Sandy Bridge-based system, Fig. 5 shows that the input value of humidity (for this study, selected from {1 %, 3 %, 10 %, 15 %}) has a significant effect on the timing of a run. This information can be used to perform application-informed load balancing in ACCOLADES, whereby the population of tasks is partitioned and scheduled based on their input values. For the study presented in Fig. 4, the cases were selected as ordered in Fig. 5 (right). As a result of this ordering, the results using fewer than 512 ranks benefit from the fact that they only involve low input humidity values, which result in lower time per case.

4 Conclusions

In this work, we discuss the development of a parallel design and data analysis tool, named ACCOLADES, for conducting large-scale parametric studies of a fleet of cars. A parallel, fast robust physics-based engine model used to compute performance and emissions of automotive engines was coupled to a data-analytics module to enable a wide range of operations in support of design- and decision-makers as well as vehicle experimentalists.

An illustrative example consisting of 4,096 parametric cases was run on a Sandy Bridge cluster and an IBM BG/Q supercomputer. It was shown that the emission and performance characteristics of a 25-min-long synthetic drive cycle can be obtained numerically in acceptable computing time ($\approx$ 4–20 min, depending on the machine). Excellent weak scaling was observed on both machines as expected in such inherently task parallel problems. Although no serious I/O bottlenecks were observed for the simulations considered in this work, we expect that additional care will need to be taken when performing I/O operations for massively parallel studies (e.g., involving a million cases) in order avoid overloading a parallel file system.

Acknowledgments. We gratefully acknowledge the computing resources provided by the Argonne Leadership Computing Facility and the Laboratory Computing Resource Center at Argonne National Laboratory.

References

1. Aithal, S.M.: Analysis of the current signature in a constant-volume combustion chamber. Combust. Sci. Technol. **185**, 336–349 (2013)
2. Aithal, S. M.: Development of an integrated design tool for real-time analyses of performance and emissions in engines powered by alternative fuels. In: Proceedings of the SAE 11th International Conference on Engines and Vehicles (2013). SAE Paper 2013–24-0134
3. Aithal, S.M.: Prediction of voltage signature in a homogeneous charge compression ignition (HCCI) engine fueled with propane and acetylene. Combust. Sci. Technol. **185**, 1184–1201 (2013)

4. Aithal, S.M., Wild, S.M.: Development of a fast, robust numerical tool for the design, optimization, and control of IC engines. In: Proceedings of the SAE 11th International Conference on Engines and Vehicles (2013). SAE Paper 2013–24-0141
5. Belludi, N., Receveur, J., Raymond, J.: High-performance grid computing for cummins vehicle mission simulation: architecture and applications. In: Proceedings of the SAE 2011 Commercial Vehicle Engineering Congress (2011). SAE Paper 2011–01-2268
6. Ehrgott, M.: Multicriteria Optimization, 2nd edn. Springer-Verlag, Heidelberg (2005)
7. Fu, T.-C.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**, 164–181 (2011)
8. Homem-de-Mello, T., Bayraksan, G.: Monte Carlo sampling-based methods for stochastic optimization. Surv. Oper. Res. Man. Sci. **19**, 56–85 (2014)
9. Kieckhafer, K., Walther, G., Axmann, J., Spengler, T.: Integrating agent-based simulation and system dynamics to support product strategy decisions in the automotive industry. In: Proceedings of the Winter Simulation Conference (2009), pp. 1433–1443
10. Lamb, D.A., Gorsich, D., Krayterman, D., Choi, K.K., Hardee, E., Du, L., Youn, B.D., Bettig, B., Ghiocel, D.: System level RBDO for military ground vehicles using high performance computing. In: Proceedings of the SAE 2008 World Congress and Exhibition. SAE Technical Paper 2008–01-0543 (2008)
11. Liu, Y., Wang, Z., Liang, J., Liu, X.: Synchronization and state estimation for discrete-time complex networks with distributed delays. IEEE Trans. Syst. Man Cybern. B **38**, 1314–1325 (2008)
12. Moawad, A., Balaprakash, P., Rousseau, A., Wild, S.M.: Novel large scale simulation process to support DOT's CAFE modeling system. In: Proceedings of the International Electric Vehicle Symposium and Exhibition, May 2015
13. Thornton, P.E., Running, S.W.: An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agric. For. Meteorol. **93**, 211–228 (1999)
14. Vijayagopal, R., Sharer, P., Wild, S.M., Rousseau, A., Chen, R., Bhide, S., Dongarkar, G., Zhang, M., Meier, R.: Using multi-objective optimization for HEV component sizing. In: Proceedings of the International Electric Vehicle Symposium and Exhibition, no. EVS28_0153, May 2015