Model-Based Optimization Algorithms for Empirical Performance Tuning

The CACHE Math/CS Institute
&
Derivative-free Optimization of Complex Systems

Prasanna Balaprakash
Argonne National Laboratory
Mathematics and Computer Science Division
9700 South Cass Avenue, Bldg. 240-1154, Argonne, IL 60439

Abstract

The increasing complexity, heterogeneity, and rapid evolution of modern computer architectures present obstacles for achieving high performance of scientific codes. Even after algorithmic improvements — seeking to improve scalability or minimize communication, for example — are made, performance can vary greatly from machine to machine. Empirical performance tuning addresses this issue by selecting code variants based on their measured performance on the target machine. A major bottleneck in empirical performance tuning is the computation time associated with testing a large number of possible code variants, which grows exponentially with the number of tuning parameters.

We formulate this tuning problem as a mathematical mixed-integer, nonlinear optimization problem over a decision space consisting of source code transformations, compiler options, and internal parameters. This is a challenging search space because integer variables (such as loop unroll factors) cannot be relaxed, continuous parameters (such as internal tolerances) can lead to incorrect output, and categorical variables (such as compiler type) do not admit a natural ordinal relationship. This difficulty is exacerbated by constraints on the search space, including algebraic expressions capturing the space of acceptable transformations, expensive black-box constraints (such as the number of cache misses or correctness of the output), and hidden constraints associated with segmentation faults or compilation errors.

In this poster we present our initial work on a model-based solver for these types of challenging optimization problems based on a trust-region framework and local surrogate interpolation or regression models over a relaxation of the decision space. We discuss the effect of different performance metrics on the noise of the objective and constraints, the importance of appropriately scaling the trust region, and steps made to ensure that the local minimum obtained is of sufficient quality. Our numerical results for tuning frequently occurring kernels on DOE leadership-class machines at NERSC and the ALCF illustrate the ability of optimization algorithms to quickly find high-performing code variants.

Joint work with Paul Hovland, Aswin Kannan, and Stefan Wild.