Green: A Framework for Supporting Energy-Conscious Programming using Controlled Approximation

Woongki Baek *
Computer Systems Laboratory
Stanford University
Stanford, CA 94305
wkbaek@stanford.edu

Trishul M. Chilimbi
Microsoft Research
One Microsoft Way
Redmond, WA 98052
trishulc@microsoft.com

Abstract

Energy-efficient computing is important in several systems ranging from embedded devices to large scale data centers. Several application domains offer the opportunity to tradeoff quality of service/solution (QoS) for improvements in performance and reduction in energy consumption. Programmers sometimes take advantage of such opportunities, albeit in an ad-hoc manner and often without providing any QoS guarantees.

We propose a system called Green that provides a simple and flexible framework that allows programmers to take advantage of such approximation opportunities in a systematic manner while providing statistical QoS guarantees. Green enables programmers to approximate expensive functions and loops and operates in two phases. In the calibration phase, it builds a model of the QoS loss produced by the approximation. This model is used in the operational phase to make approximation decisions based on the QoS constraints specified by the programmer. The operational phase also includes an adaptation function that occasionally monitors the runtime behavior and changes the approximation decisions and QoS model to provide strong statistical QoS guarantees.

To evaluate the effectiveness of Green, we implemented our system and language extensions using the Phoenix compiler framework. Our experiments using benchmarks from domains such as graphics, machine learning, signal processing, and finance, and an in-production, real-world web search engine, indicate that Green can produce significant improvements in performance and energy consumption with small and controlled QoS degradation.

Categories and Subject Descriptors D.1.m [*Programming Techniques*]: Miscellaneous; D.3.3 [*Programming Languages*]: Language Constructs and Features

General Terms Performance, Measurement, Languages, Designs

Keywords Energy-Conscious Programming, Controlled Approximation

* A part of this work was performed while the author was an intern at Microsoft Research, Redmond.

PLDI'10, June 5–10, 2010, Toronto, Ontario, Canada.

1. Introduction

Companies such as Amazon, Google, Microsoft, and Yahoo are building several large data centers containing tens of thousands of machines to provide the processing capability necessary to support web services such as search, email, online shopping, etc. [3]. Not surprisingly, power is a large component of the monthly operational costs of running these facilities and companies have attempted to address this by locating them in places where power is cheap [13]. In addition, energy is often a key design constraint in the mobile and embedded devices space given the current limitations of battery technology.

There are several application domains where it is acceptable to provide an approximate answer when the cost and resources required to provide a precise answer are unavailable or not justified. For example, real-time ray tracing is infeasible on current PCs so games employ a variety of techniques to produce realistic looking lighting and shadows while still rendering at 60 frames per second. Content such as images, music, and movies are compressed and encoded to various degrees that provide a tradeoff between size requirements and fidelity. Such approximations typically result in the program performing a smaller amount of processing and consequently consuming less energy while still producing acceptable output.

Programmers often take advantage of such Quality of Service (QoS) tradeoffs by making use of domain-specific characteristics and employing a variety of heuristics. Unfortunately, these techniques are often used in an ad-hoc manner and programmers rarely quantify the impact of these approximations on the application. Even in cases where they attempt to quantify the impact of the heuristics used, these are hard to maintain and keep up-to-date as programs evolve and add new features and functionality.

To address these issues, this paper proposes Green shown in Figure 1, which is a framework for supporting energy-conscious programming using loop and function approximation. Green provides a simple, yet flexible framework and programming support for approximating expensive loops and functions in programs. It allows programmers to specify a maximal QoS loss that will be tolerated and provides statistical guarantees that the application will meet this QoS target. Programmers must provide (possibly multiple) approximate versions of the function for function approximation. Unless directed otherwise, Green uses the function return value as the QoS measure and computes loss in QoS by comparing against the value returned by the precise function version given the same input. Loops are approximated by running fewer loop iterations. In this case, a programmer must provide a function that computes QoS to enable Green to calculate the loss in QoS that arises from early loop termination.

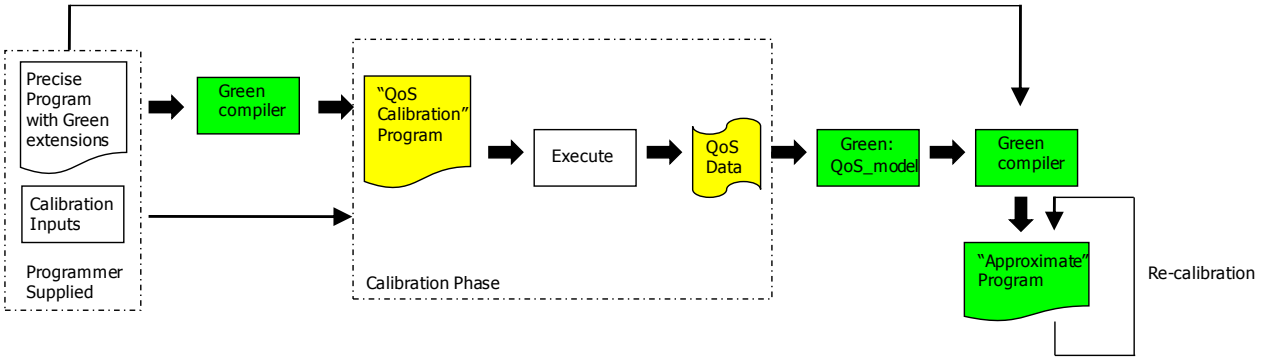


Figure 1. Overview of the Green system.

Green uses this information about loops and functions that are candidates for approximation to construct a "calibration" program version. Green runs this application version with a programmer-provided set of calibration inputs to construct a QoS model that quantifies the loss in QoS that results from using the approximate version of the function or loop and the corresponding improvement in application performance and energy consumption. Green then generates an "approximate" version of the program that uses this QoS model in conjunction with the programmer supplied QoS target to determine when to use the approximate version of the function or terminate the loop early while still meeting the QoS requirement. Since the QoS degradation on actual program inputs may differ from that observed during calibration, Green also samples the QoS loss observed at runtime and updates the approximation decision logic to meet the specified QoS target. In this way, Green attempts to provide statistical guarantees that the specified QoS will be met. Such statistical guarantees are becoming important as cloud-based companies provide web services with Service Level Agreements (SLAs) that typically take the form: "the service will provide a response within 300ms for 99.9% of its requests for a peak client load of 500 requests per second" [8].

Our experimental results indicate that Green can significantly improve performance and reduce energy consumption of several applications with only a small degradation in QoS. In particular, we improved the performance and reduced the energy consumption of `Bing Search`, a back-end implementation of a commercial, in-production web-search engine by 21.0% and 14.0% respectively with 0.27% of QoS degradation (three queries out of a thousand returned a search result that included at least one different document or the same documents in a different rank order). We also empirically demonstrate that Green can generate a robust QoS model with a relatively small training data-set and that runtime re-calibration can enable applications to meet their QoS targets even if they use imperfect QoS models. The QoS models produced by Green during the calibration phase have also proved useful in providing developers with a better understanding of their application.

This paper makes the following main contributions.

- Describes the design and implementation of the Green system, which provides simple, yet flexible support for energy-conscious programming using function and loop approximations.
- Experimental evaluation of the Green system that shows significant improvements in performance and energy consumption with little QoS degradation.
- Experimental evidence that indicates that Green's QoS modeling is robust and that its adaptation supports meeting target QoS requirements.

The rest of the paper is organized as follows. Section 2 describes the design of the Green system. Section 3 discusses our implementation of Green. Experimental evaluation of Green is described in Section 4. Section 5 provides a brief overview of related work, and Section 6 concludes the paper.

2. Green Design

Figure 1 provides a high-level overview of the Green system that we introduce and discuss in more detail in this section.

2.1 Controlled Program Approximation

Expensive functions and loops present attractive targets for program approximation because they are modular and time-consuming portions of programs. Green considers function approximations that use an alternative, programmer-supplied approximate version of the function and loop approximations that terminate the loop earlier than the base (precise) version. But we would like to quantify the impact these approximations have on the QoS of the program. To do this, the programmer must supply code to compute a QoS metric for the application and some training inputs. These training inputs are used during a calibration phase as shown in Figure 1. During these training runs, Green monitors and records the impact function or loop approximation has on the program's QoS and its performance and energy consumption. This data is used to build a QoS model that is subsequently used by Green to decide when to approximate and when to use the precise version in order to guarantee user-specified QoS Service Level Agreements (SLAs). Figure 2 illustrates at a high-level how Green approximates loops or functions while still attempting to meet required QoS SLAs. The QoS model constructed in the calibration phase is used by the Green synthesized QoS_Approx() to decide whether approximation is appropriate in the current situation as determined by the function input or loop iteration count. Since the QoS degradation on the actual program inputs may differ from that observed during the calibration phase, Green provides a mechanism to occasionally measure the program's QoS and update the QoS approximation decisions at runtime.

2.2 Green Mechanisms

As described, Green requires a QoS_Compute() function for computing the program's QoS (for function approximation, the original "precise" function serves this purpose), and then synthesizes a QoS_Approx() function for determining whether to perform an approximation, and a QoS_ReCalibrate() function for revisiting approximation decisions at runtime. In addition, it requires a set of training inputs to construct the QoS model used to synthesize the QoS_Approx() function and a QoS_SLA value that must be met.

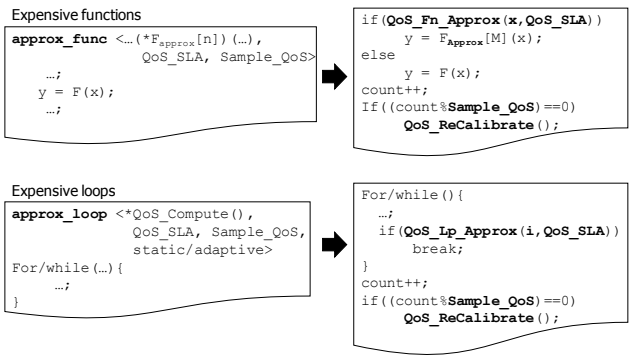


Figure 2. High-level overview of Green's approximation.

We provide a high-level description of these mechanisms here and leave a detailed discussion to the next section.

2.2.1 QoS Calibration and Modeling

Green's calibration phase collects the data required to build the QoS model. It requires QoS_Compute(), which is application dependent and can range from trivial as in the case of using the approximated function's return value to a more involved computation involving pixel values rendered on the screen. Green uses this function along with the set of training inputs to construct a QoS model that relates function inputs in the case of function approximation and loop iteration count in the case of loop approximation to loss in QoS and performance and energy consumption improvements. This QoS model is used in conjunction with a provided target QoS SLA to make approximation decisions.

2.2.2 QoS Approximation

QoS_Approx() comes in two main flavors for loop approximations. In the static variety, the approximation is solely determined by the QoS model constructed in the calibration phase. Here the loop iteration count threshold is determined by the QoS model and the user-specified QoS SLA. Once the loop iteration count exceeds this threshold, the approximation breaks out of the loop. The adaptive variety is based on the law of diminishing returns. Here the approximation uses the QoS model in conjunction with the QoS SLA to determine appropriate intervals at which to measure change in QoS and the amount of QoS improvement needed to continue iterating the loop. For function approximation, the QoS model is used in conjunction with the QoS SLA and the function input to determine if the function should be approximated and which approximate version of the function to use.

2.2.3 QoS Re-Calibration

The program's behavior may occasionally differ from that observed on its training inputs and QoS_ReCalibrate() provides a mechanism to detect and correct for this effect. In the case of loop approximation, when used with static approximation re-calibration can update the QoS model and increase the loop iteration count threshold to compensate for higher than expected QoS degradation or decrease this threshold to improve performance and energy consumption when QoS degradation is lower than expected. Similarly, when used with adaptive approximation re-calibration can appropriately change the interval used to measure change in QoS and/or QoS improvement required to continue. For function approximation, re-calibration allows Green to switch the approximate version of the function used to one that is more or less precise as determined by the observed QoS loss.

2.2.4 Discussion

To tie these concepts together, we illustrate an end-to-end example of applying loop approximation to the main loop of a simple program that estimates Pi in Figure 3. The underlined terms correspond to functions or variables that are supplied by the programmer. The rest of the code is generated by the Green system.

Since, as shown in Section 4, Green's re-calibration mechanism is quite effective, one might underestimate the importance of the calibration phase and attempt to solely rely on the re-calibration mechanism. However, the calibration phase is still important and necessary because it provides (1) faster convergence to a good state, (2) reliable operation even when users choose to avoid or minimize re-calibration to lower overhead, and (3) programmer insight into the application's QoS tradeoff through the QoS model constructed by Green during the calibration phase. In fact, Green users have provided consistent feedback that the QoS model constructed has provided extremely valuable and often unexpected information about their application behavior.

Green's reliance on a runtime re-calibration mechanism to ensure that the desired QoS is met permits approximating multiple expensive loops and functions within the same application. In contrast, a static approach to estimating QoS would have to account for non-linear effects arising from combining multiple approximations. Our current implementation builds a local QoS model for each approximated program unit, uses these to construct a global QoS model for the application, and coordinates re-calibration across these.

3. Green Implementation

This section describes our implementation of the Green system that provides a simple and flexible framework for constructing a wide variety of approximation policies (see Figure 1 for overview). Our goal is to provide a minimal and simple interface that satisfies the requirements of the majority of programmers while providing the hooks that allow expert programmers to craft and implement custom, complex policies. To achieve this, Green comes with a couple of simple, default policies that meet the needs of many applications and enables them to benefit from using controlled QoS approximation with minimal effort. At the same time, it allows programmers to override these default policies and supply their own by writing custom versions of QoS_Approx() and QoS_ReCalibrate() while still benefiting from Green's calibration and modeling capability. We discuss Green's interface and default policies and provide an instance of a customized policy.

3.1 Green Programming Support

3.1.1 Loop Approximation

Green supports loop approximation with a new language keyword approx_loop as shown in Figure 2. The programmer uses approx_loop just before the target loop and supplies a pointer to a user-defined QoS_Compute() function, the value of the desired QoS_SLA and indicates whether they want to use static or adaptive approximation. In addition, if the programmer wants to avail of runtime re-calibration she must provide the sampling rate (sample_QoS) to perform re-calibration. If a programmer wishes to construct a custom policy, they must also supply pointers to custom QoS_Approx() and/or QoS_ReCalibrate() routines.

3.1.2 Function Approximation

For function approximation, Green introduces a new keyword approx_function as shown in Figure 2. The programmer uses approx_function before the target function implementation and supplies a function pointer array that contains pointers to user-defined

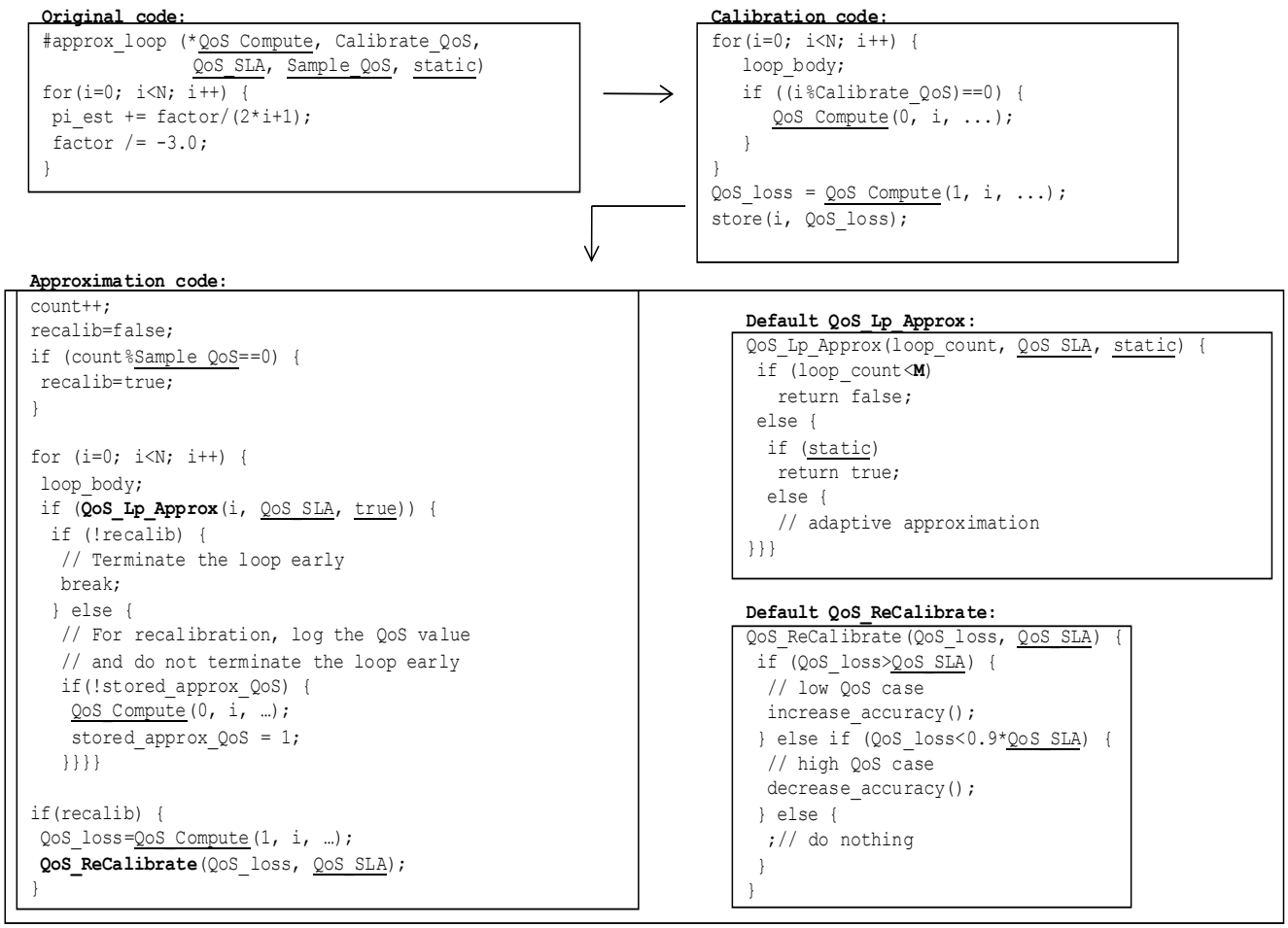


Figure 3. An end-to-end example of applying loop approximation to the Pi estimation program.

approximate versions of that function in increasing order of precision, along with the value of the desired QoS_SLA and a sampling rate (sample_QoS) if re-calibration is required. If the function return value does not provide the desired QoS metric, the programmer must also supply a pointer to a custom QoS_Compute() function. If the function takes multiple arguments, Green requires the parameter positions of the arguments that should be used while building the QoS model[1]. As with approx_loop, pointers to custom QoS_Approx() and/or QoS_ReCalibrate() routines are optional and only needed for custom policies.

3.2 Green System Implementation

Figure 1 and Figure 3 provide an overview of the Green system implementation. The Green compiler first generates a "calibration" version of the program that is run with user-provided calibration inputs to generate QoS data needed to build the QoS model. Then the compiler uses this constructed QoS model to generate an "approximate" version of the program that can be run in place of the original. It synthesizes code to implement QoS_Approx() and QoS_ReCalibrate().

[1] Our current implementation constructs models based on a single input parameter. However, this can be extended to multiple parameters.

```
QoS_Compute(return_QoS, loop_count)
{
 if(!return_QoS) {
  Store_top_N_docs(loop_count);
  return -1;
 } else {
  if(Same_docs(Get_top_N_docs(loop_count),
               Current_top_N_docs())
   return 0;
  else
   return 1;
 }}
```

Figure 4. QoS_Compute for `Bing Search`.

3.2.1 Loop Approximation

The programmer-supplied QoS_Compute() function is used in the calibration phase to tabulate the loss in QoS resulting from early loop termination at loop iteration counts specified by Calibrate_QoS. QoS_Compute() function has the following interface: QoS_Compute (return_QoS, loop_count, calibrate, Calibrate_QoS, ...) and the search application's version is shown

```
QoS_Lp_Approx(loop_count, QoS_SLA, static) {
 if (loop_count<M)
   return false;
 else {
  if (static)
   return true;
  else {
   if(loop_count%Period==0) {
    QoS_improve=QoS_Compute(1, loop_count, ...);
    QoS_Compute(0, loop_count, ...);
    if (QoS_improve>Target_delta)
     return false;
    else
     return true;
   } else {
    return false;
   }}}}
```

Figure 5. Code generated for loop approximation.

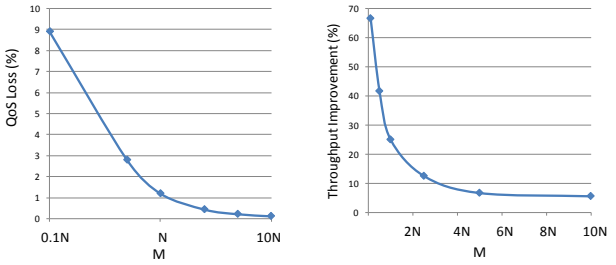


Figure 6. Calibration data for `Bing Search`.

in Figure 4. Note that when QoS_Compute() is called with return_QoS unset it stores the QoS computed at that point and only returns QoS_loss when this flag is set. Then, it compares the current QoS against the stored QoS to return the QoS loss. When it is called with the calibrate flag set at the end of the loop in calibration mode, it computes and stores the % QoS loss when the loop terminates early at loop iteration counts specified by Calibrate_QoS.

Figure 6 shows the calibration data generated for `Bing Search`, that quantifies the impact of limiting the documents searched to M rather than all matching documents on QoS and throughput. This calibration data is then used by Green's QoS modeling routine that is implemented as a MATLAB program. The MATLAB program is used for interpolation and curve fitting to construct a function from these measurements. It automatically selects the appropriate approximation level based on the QoS value desired by the programmer and this empirically constructed model. The programmer only supplies the desired QoS Value. This model supports the following interface for loops:

$$M = QoS\_Model\_Loop(QoS\_SLA, static) \quad (1)$$

$$< M, Period, Target\_Delta >= QoS\_Model\_Loop(QoS\_SLA, adaptive) \quad (2)$$

For static approximation the QoS model supplies the loop iteration count that is used by QoS_Approx() for early loop termination. In the adaptive approximation case, the QoS model additionally determines the period and target QoS improvement required to continue iterating the loop.

The synthesized QoS_Approx() code shown in Figure 5 uses the parameters generated by the QoS model to perform approximation.

```
QoS_Fn_Approx(x, QoS_SLA){
 if (x < 0.5)
  return false;
 else if (x < 0.8)
  M = 0;
 else if (x < 1.1)
  M = 1;
 else
  return false;
 return true;
}
```

Figure 7. Code generated for function approximation.

When the approximation is being re-calibrated, the synthesized approximation code stores the QoS value that would have been generated with early loop termination and continues running the loop as many times as the original (precise) program version would have in order to compute the QoS loss (see Figure 3).

The QoS_ReCalibrate() code generated by Green compares this QoS loss against the target QoS SLA and either decreases/increases the approximation by either increasing/decreasing the value M of the early termination loop iteration count (static approximation) or decreasing/increasing the value of Target_Delta (adaptive approximation).

3.2.2 Function Approximation

By default, Green uses the function return value to compute QoS unless the programmer defines a separate QoS_Compute() function. For calibration, Green generates code that computes and stores the loss in precision that results from using the family of approximation functions at each call site of the function selected for approximation but our current implementation does not differentiate between call sites and uses the same QoS_Approx() function for all sites. Figures 8(a) and 8(b) shows the calibration data generated for the `exp` and `log` functions in the `blackscholes` application over the input argument range observed on the training inputs.

Green's modeling routine uses this data and supports the following interface for functions[2]:

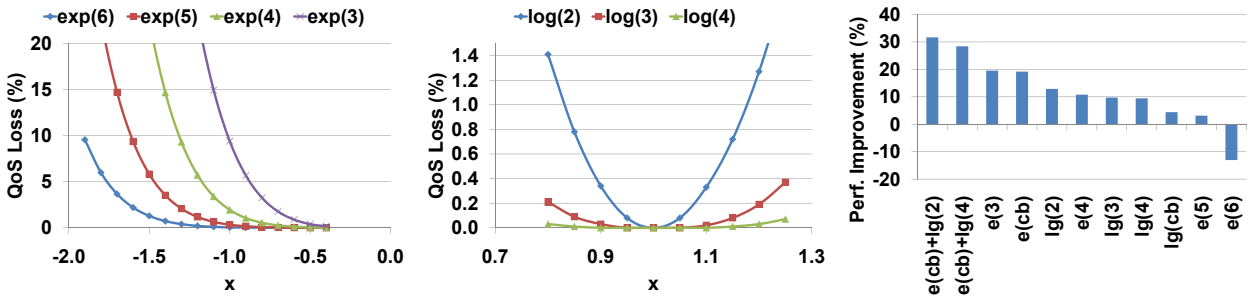
$$< (M_i, lb_i, ub_i) >= QoS\_Model\_Func(QoS\_SLA)$$

where it returns the best approximate function version and corresponding input argument range where the QoS loss for that function satisfies the specified target QoS SLA. In the event that none of the approximate function versions meet the QoS requirement, it returns an empty set and the precise function is used.

The synthesized QoS_Approx() function for the `exp` function in `blackscholes` is shown in Figure 7, where it uses `exp`(3) for $0.5 \leq$ abs(x) < 0.8, exp(4) for $0.8 \leq$ abs(x) < 1.1 and the precise `exp` function for `abs`(x) ≥ 1.1 and `abs`(x) < 0.5[3]. Note that `exp`(5) and `exp`(6) (see Figure 8(a)) were discarded because they did not provide a competitive QoS_loss to performance improvement ratio. The QoS_ReCalibrate() function replaces the current approximate function version with a more precise one, to address low QoS, and uses a more approximate version to address higher than necessary QoS.

[2] Our current QoS modeling scheme only works for functions that take numerical data as input and would need to be extended to handle functions that take structured data as input.

[3] The approximate versions of the `exp` (and `log`) functions correspond to different Taylor series expansions with the number indicating the highest degree polynomial used.



(a) QoS (exp functions) (b) QoS (log functions) (c) Performance

Figure 8. Calibration data for `blackscholes`.

3.3 Custom Approximation

Green allows programmers to override its default synthesis of QoS_Approx() and QoS_ReCalibrate() to implement custom approximation policies. Figure 9 shows an example of a custom QoS_ReCalibrate() that we used for the Search application. For `Bing Search`, QoS_Compute() returns a QoS loss of 1 (100%) if the top N documents returned by the precise and approximate version do not exactly match and a QoS loss of 0 otherwise. To perform re-calibration and compare against a target QoS SLA that is of the form, "the application returns identical results for 99% of the queries", we need to measure the QoS loss across multiple queries as implemented in the code shown. It is certainly possible to relax this stringent QoS requirement and allow document reordering within the top N documents returned. But we used this strict QoS requirement for the search experiments to avoid addressing how such reorderings may affect the perceived quality of the search results returned. The calibration and QoS_Approx() uses the default code synthesized by Green. `Bing Search` was the only application of those evaluated in this paper that needed a custom recalibration routine mainly because its QoS metric is computed over an aggregate set of queries and not an individual query.

3.4 Green Support for Multiple Approximations

So far, we have discussed Green support for individual loop or function approximation and this section discusses Green's mechanisms for combining multiple approximations. Green requires the application developer to provide an additional QoS_Compute() function for the application and an application QoS SLA. In many cases, this is identical to the QoS_Compute() function already supplied for loop approximation.

3.4.1 QoS Approximation Modeling

Green performs the calibration for each function or loop approximation in isolation as shown in Figures 6, 8(a), and 8(b), and then performs an exhaustive search space exploration that attempts to combine these and still meet the specified application QoS SLA. Figure 8(c) illustrates this process for the `exp` and `log` functions in the `blackscholes` application, where exp(cb) represents the combination of exp(3) and exp(4) that was selected for approximating `exp` as shown in Figure 7. This search process can result in individual function/loop approximation decisions being changed so that the overall application QoS SLA is still satisfied. In the case of `blackscholes`, the local approximation decision to use `log`(2) was changed to use `log`(4), so that the combined approximation of using exp(cb) with `log`(4) was able to satisfy the application QoS. Note that unlike `exp` the calibration process did not find a combination of `log` functions viable for this application (strictly worse than the individual `log` approximations) and hence there are no combined `log`(cb) bars shown in Figure 8(c).

```
QoS_ReCalibrate(QoS_loss, QoS_SLA) {
 // n_m: number of monitored queries
 // n_l: number of low QoS queries in monitored queries
 if (n_m==0) {
  // Set Sample_QoS to 1 to trigger QoS_ReCalibrate
  // for the next 100 consecutive queries
  Saved_Sample_QoS=Sample_QoS;
  Sample_QoS=1;
 }
 n_m++;
 if (QoS_loss !=0)
  n_l++;
 if (n_m==100) {
  QoS_loss=n_l/n_m;
  if(QoS_loss>QoS_SLA) {
   // low QoS case
   increase_accuracy();
  } else if (QoS_loss < 0.9*QoS_SLA) {
   // high QoS case
   decrease_accuracy();
  } else {
   ; // no change
  }
  Sample_QoS=Saved_Sample_QoS;
 }
}
```

Figure 9. Customized QoS_ReCalibrate for `Bing Search`.

3.4.2 Global Recalibration

When an application that has multiple approximate functions and/or loops requires recalibration to meet its QoS SLA, Green initiates a global recalibration process. This entails selecting a subset of functions and loops for recalibration and then performing local recalibration of these. Our current implementation of global recalibration initially assumes that the individual approximations are independent of each other and the approximations are additive, but subsequently detects and applies corrections to the cases where this assumption is not valid. In the case where the measured application QoS falls below the specified SLA, recalibration candidates are ranked by their QoS_loss/performance gain sensitivity as per the QoS model. This enables recalibration to be first applied to candidates where a large QoS change produces a small performance change. To handle the case where the approximations interact and produce non-linear effects, Green uses an exponential

backoff scheme similar to that used for TCP/IP packet retransmission [19]. In this scheme, individual approximations are recalibrated as follows. Loop approximations are recalibrated using a random increase in the number of iterations within an acceptable range and approximated functions are replaced by higher precision ones over a random portion of their input argument range, until the non-linear effects disappear and the application QoS SLA is satisfied or the approximation is disabled and the precise loop/function is used. This scheme appears to perform well and converge fast on artificial test examples we constructed to validate its efficacy but we have been unable to force such non-linear behavior in any of our benchmark applications.

4. Green Evaluation

We performed two types of experiments. First, we show that Green can produce significant improvements in performance and reduction in energy consumption with little QoS degradation. Next, we show that the QoS models Green constructs are robust and in conjunction with runtime re-calibration provide strong QoS guarantees. For evaluation, we use five applications including `Bing Search`, a back-end implementation of a commercial web-search engine, and four other benchmarks including `252.eon` from SPEC CPU2000 [24], Cluster GA (`CGA`) [14], Discrete Fourier Transformation (`DFT`) [7], and `blackscholes` from the PARSEC benchmark suite [4]. The choice of `blackscholes` was driven by conversations with computational finance practitioners who pointed out that microseconds matter and controlled approximation is appropriate for their domain as they employ sophisticated risk analysis models.

4.1 Environment

We use two different machines for our experiments. A desktop machine is used for experiments with `252.eon`, `CGA`, `DFT`, and `blackscholes`. The desktop machine runs an Intel Core 2 Duo (3 GHz) processor with 4 GB (dual channel DDR2 667 MHz) main memory. A server-class machine is used for experiments with `Bing Search`. The server machine has two Intel 64-bit Xeon Quad Core (2.33 GHz) with 8 GB main memory.

For each application, we compare the approximated versions generated by the Green compiler implemented using the Phoenix compiler framework [20] against their corresponding precise (base) versions. For evaluation, we measure three key parameters for each version: performance, energy consumption, and QoS loss. For the other benchmarks, the wall-clock time between start and end of each run is used for performance evaluation. For `Bing Search`, we first run a set of warmup queries and use the measured throughput (i.e., queries per second (QPS)) while serving the test queries as the performance metric. To measure the energy consumption, we use an instrumentation device that measures the entire system energy consumption by periodically sampling the current and voltage values from the main power cable. The sampling period of the device is 1 second. Since the execution time of the applications we study are significantly longer, this sampling period is acceptable. Finally, we compute the QoS loss of each approximate version by comparing against results from the base versions. The QoS metric used for each application will be discussed later. We also measured the overhead of Green by having each call to QoS_Approx() eventually return false and found the performance to be indistinguishable from the base versions of the applications without the Green code, when the recalibration sampling rate was set to 1%.

4.2 Applications

In this section, we provide a high-level description and discuss Green approximation opportunities for each application. In addition, we discuss the evaluation metrics and input data-sets used.

4.2.1 Bing Search

Description: `Bing Search` is a back-end implementation of a commercial web-search engine that accepts a stream of user queries, searches its index for all documents that match the query, and ranks these documents before returning the top N documents that match the query in rank order. Web crawling and index updates are disabled. There are a number of places in this and subsequent sections where additional information about the `Bing Search` application may have been appropriate but where protecting Microsoft's business interests require us to reduce some level of detail. For this reason, performance metrics are normalized to the base version and the absolute number of documents processed are not disclosed.
Opportunities for Approximation: The base version of `Bing Search` processes all the matching candidate documents. Instead, we can limit the maximum number of documents (M) that each query must process to improve performance and reduce energy consumption while still attempting to provide a high QoS.
Evaluation Metrics: We use QPS as the performance metric since throughput is key for server applications. We use *Joules per Query* as the energy consumption metric. Finally, for our QoS loss metric, we use the percentage of queries that either return a different set of top N documents or return the same set of top N documents but in a different rank order, as compared to the base version.
Input data-sets: The `Bing Search` experiments are performed with a production index file and production query logs obtained from our data center. Each performance run uses two sets of queries: (1) warm-up queries: 200K queries to warm up the system and (2) test queries: 550K queries to measure the performance of the system.

4.2.2 Graphics: 252.eon

Description: `252.eon` is a probabilistic ray tracer that sends N^2 rays to rasterize a 3D polygonal model [24]. Among the three implemented algorithms in `252.eon`, we only used the Kajiya algorithm [12].
Opportunities for Approximation: The main loop in `252.eon` iterates N^2 iterations and sends a ray at each iteration to refine the rasterization. As the loop iteration count goes higher, QoS improvement per iteration can become more marginal. In this case, the main loop can be early terminated while still attempting to meet QoS requirements.
Evaluation Metrics: We measure the execution time and energy consumption to rasterize an input 3D model. To quantify the QoS loss of approximate versions, we compute the average normalized difference of pixel values between the precise and approximate versions.
Input data-sets: We generated 100 input data-sets by randomly changing the camera view using a reference input 3D model of `252.eon`.

4.2.3 Machine Learning: Cluster GA

Description: `Cluster GA (CGA)` solves the problem of scheduling a parallel program using a genetic algorithm [14]. `CGA` takes a task graph as an input where the execution time of each task, dependencies among tasks, and communication costs between processors are encoded using node weights, directed edges, and edge weights, respectively. `CGA` refines the QoS until it reaches the maximum generation (G). The output of `CGA` is the execution time of a parallel program scheduled by `CGA`.
Opportunities for Approximation: Depending on the size and characteristics of a problem, `CGA` can converge to a near-optimal solution even before reaching G. In addition, similar to `252.eon`, QoS improvement per iteration can become more marginal at higher iteration counts (i.e., generation). By terminating the main

loop earlier, we can achieve significant improvement in performance and reduction in energy consumption with little QoS degradation.

Evaluation Metrics: We use the same metrics for performance and energy consumption as for `252.eon`. For a QoS metric, we compute the normalized difference in the execution time of a parallel program scheduled by the base and approximate versions.

Input data-sets: We use 30 randomly generated task graphs described in [15]. To ensure various characteristics in the constructed task graphs, the number of nodes varies from 50 to 500 and communication to computation ratio (CCR) varies from 0.1 to 10 in randomly generating task graphs.

4.2.4 Signal Processing: Discrete Fourier Transform

Description: `Discrete Fourier Transform (DFT)` is one of the most widely used signal processing applications [7] that transforms signals in time domain to signals in frequency domain.

Opportunities for Approximation: In the core of `DFT`, `sin` and `cos` functions are heavily used. Since the precise version implemented in standard libraries can be expensive especially when the underlying architecture does not support complex FP operations, the approximated version of `sin` and `cos` functions can be effectively used if it provides sufficient QoS. We implement several approximated versions of `sin` and `cos` functions [9] and apply them to our `DFT` application.

Evaluation Metrics: We use the same metrics for performance and energy consumption as `252.eon`. As a QoS metric, we compute the normalized difference in each output sample of `DFT` between the precise and approximated versions.

Input data-sets: We randomly generate 100 different input data-sets. Each input sample has a random real value from 0 to 1.

4.2.5 Finance: blackscholes

Description: `blackscholes` computes the price of a portfolio of European options using a partial differential equation. The implemented method is a very popular technique for pricing options.

Opportunities for Approximation: The core computation makes heavy use of the exponentiation `exp` and logarithm `log` functions. We provided a series of approximate versions of these functions that use the corresponding Taylor series expansions with varying number of polynomial terms[4].

Evaluation Metrics: We use the same metrics for performance and energy consumption as `252.eon`. As a QoS metric, we compute the difference in option prices produced by the precise and approximate versions of the programs.

Input data-sets: We use the large simulation data set provided by the Parsec benchmark suite for training and report results using the native execution data set. The training data set prices 64 thousand options and the native data set prices 10 million options.

4.3 Experimental Results with `Bing Search`

Figures 10 and 11 demonstrate the tradeoff between the QoS loss and the improvement in performance[5] and reduction in energy consumption. The base version is the current implementation of `Bing Search`, while M-* versions are approximated. Specifically, M-*N statically terminates the main loop after processing *N matching documents for each query. M-PRO-0.5N samples QoS improvement after processing every 0.5N documents and adaptively terminates the main loop when there is no QoS improvement in the current period. As can be seen, some approximated versions significantly improve the performance and reduce the energy consumption (i.e., Joules per query) with very little QoS loss. For example, M-N improves throughput by 24.3% and reduces energy consumption by 17% with 0.94% of QoS loss. Another interesting point is that M-PRO-0.5N that uses the adaptive approximation leads to slightly better performance and less energy consumption while providing better QoS compared to M-2N version that uses the static

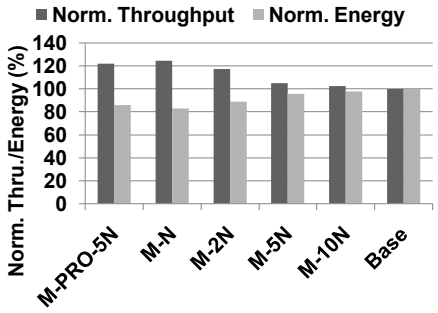


Figure 10. Performance and energy consumption of various versions of `Bing Search`.

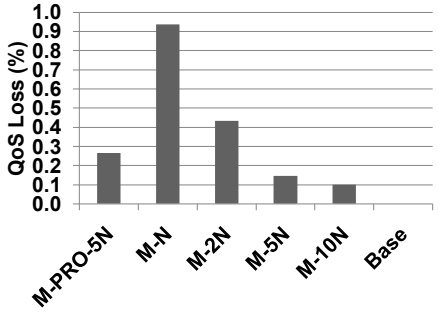


Figure 11. QoS loss of various versions of `Bing Search`.

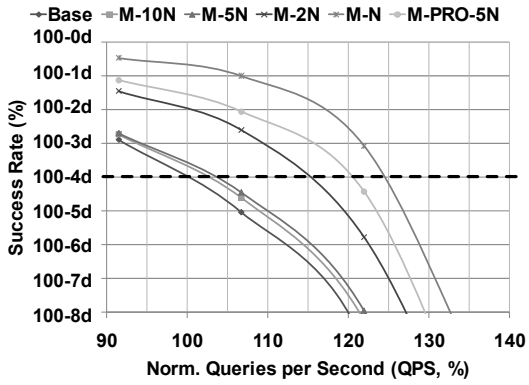


Figure 12. Varying throughput and resulting success rate of various versions of `Bing Search`.

[4] The `blackscholes` Parsec benchmark includes a loop that repeatedly computes the value of the option portfolio to increase the work performed by the application. All iterations of this loop except for the first can be skipped without affecting the results. We did not approximate this loop and measured similar improvements when this loop was executed once and when it executed the number of times specified in the benchmark.

[5] Figure 12 illustrates how the performance of each version of `Bing Search` is determined by the *cutoff QPS* metric. Cutoff QPS is defined as QPS where the success rate of queries goes as low as $(100 - 4d)\%$ shown as the dotted line. Any lower and the Search QoS SLA is considered to be violated.

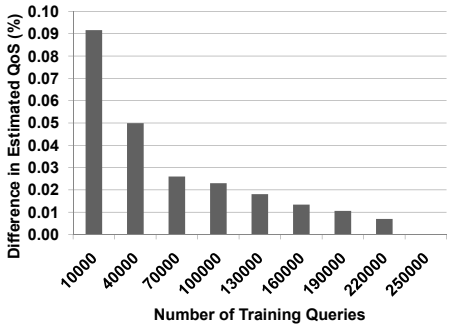


Figure 13. Sensitivity of Green's QoS model of `Bing Search` to the size of training data-sets.

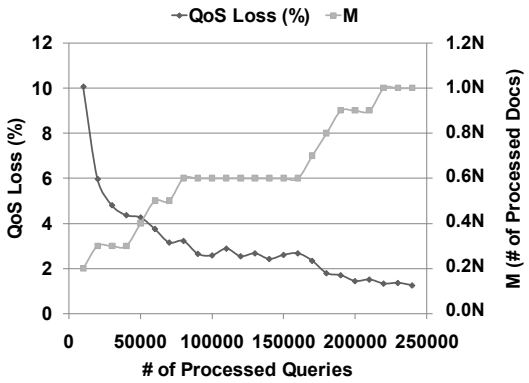


Figure 14. The effectiveness of Green's re-calibration mechanism for `Bing Search`.

approximation. This showcases the potential of adaptive techniques to optimize `Bing Search`.

To study the sensitivity of Green's QoS model to the training data-set size, we randomly permuted the warm-up and test queries and injected different number of queries ranging from 10K to 250K queries to build the QoS model. Figure 13 demonstrates the difference in estimated QoS loss (when M=N) with the varying size of training data-sets. QoS models generated with much fewer number of queries are very close to the one generated with 250K. For example, the QoS model generated using only 10K queries differs by only 0.1% compared to the one generated using 250K inputs. This provides empirical evidence that Green can construct a robust QoS model for `Bing Search` without requiring huge training data-sets.

To evaluate the effectiveness of Green's re-calibration mechanism, we performed an experiment simulating an imperfect QoS model. Say a user indicates his/her desired QoS target as 2%, but the constructed QoS model incorrectly supplies M = 0.1N (which typically results in a 10% QoS loss). Thus, without re-calibration `Bing Search` will perform poorly and not meet the target QoS. Figure 14 demonstrates how can Green provide robust QoS guarantees even when supplied with an inaccurate QoS model. After processing every 10K queries, Green monitors the next 100 consecutive queries (i.e., Sample_QoS=1%) by running the precise version while also computing the QoS loss, if it had used the approximated version for those 100 queries. Since the current QoS model is not accurate enough, the monitored results will keep reporting low QoS. Then, Green's re-calibration mechanism keeps increasing the

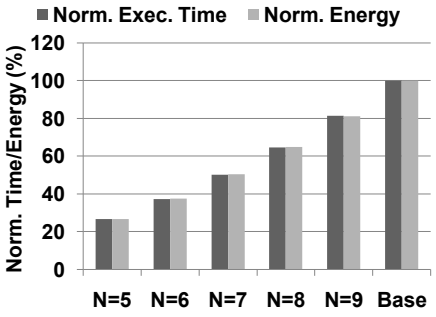


Figure 15. Performance and energy consumption of various versions of `252.eon`.

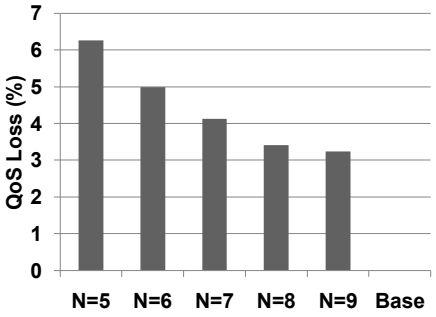


Figure 16. QoS loss of various versions of `252.eon`.

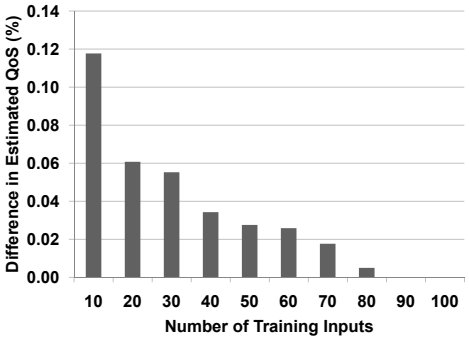


Figure 17. Sensitivity of Green's QoS model of `252.eon` to the size of training data-sets.

accuracy level (i.e., by increasing the M value by 0.1N) until it satisfies the user-defined QoS target. In Figure 14, Green meets the QoS target after processing 180K queries. The user could use a period smaller than 10K to make Green adapt faster but there is a tradeoff between quick adaptation and degradation in performance and energy consumption caused by more frequent monitoring. We performed identical re-calibration experiments for the other applications with similar results.

4.4 Experimental Results with Benchmarks

Figures 15 and 16 show the results for `252.eon` using 100 randomly generated input data-sets. Similar to `Bing Search`, approximated versions of `252.eon` significantly improve the performance and energy consumption with relatively low QoS loss. Figure 17 demonstrates the sensitivity of Green's QoS model for `252.eon` to the size of training data-sets. We varied the training data size

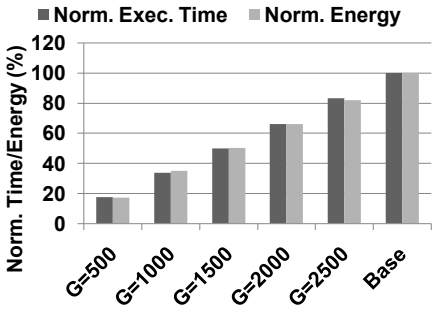


Figure 18. Performance and energy consumption of various versions of `CGA`.

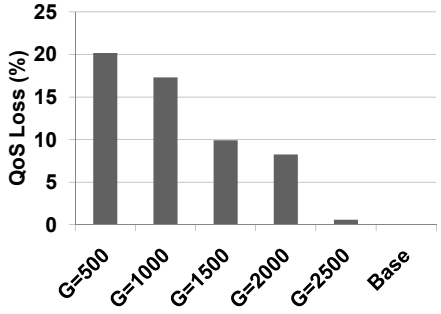


Figure 19. QoS loss of various versions of `CGA`.

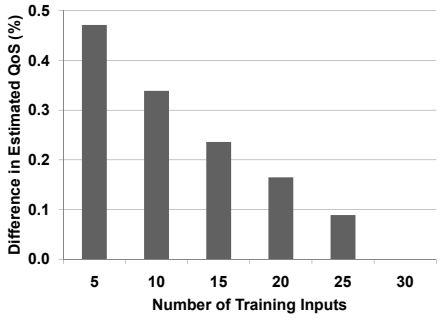


Figure 20. Sensitivity of Green's QoS model of `CGA` to the size of training data-sets.

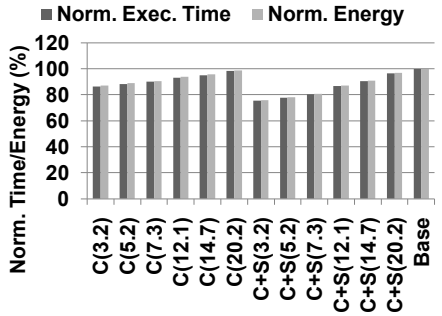


Figure 21. Performance and energy consumption of various versions of `DFT`.

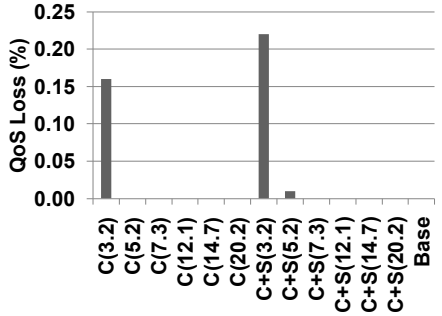


Figure 22. QoS loss of various versions of `DFT`.

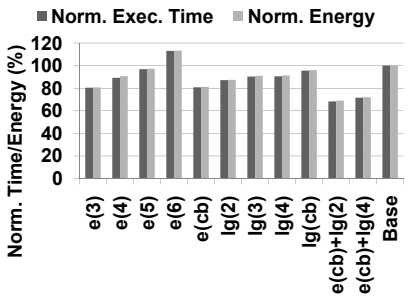


Figure 23. Performance and energy consumption of various versions of `blacksholes`.

from 10 to 100, and compared the estimated QoS loss difference (when N=9) of the generated QoS model to the one generated using 100 inputs. Figure 17 provides empirical evidence that Green's QoS model for `252.eon` can be constructed robustly with relatively small training data-sets. For example, the QoS model generated using 10 inputs differs by only 0.12% compared to the one generated using 100 inputs.

Figures 18 and 19 demonstrate the Green model of `CGA` using 30 randomly generated input data-sets. Up to G=1500, QoS loss is reasonable ($<$10%), while significantly improving performance and energy consumption by 50.1% and 49.8%, respectively. In Figure 20, We also present the sensitivity of Green's QoS model of `CGA` to the training data-set size. We varied the number of training inputs from 5 to 30 and compared the estimated QoS loss (G=2500) from the generated QoS model to the one generated using 30 inputs. While the difference in the estimated QoS loss is higher than other applications due to the discrete nature of the outcome of a parallel task scheduling problem, the difference is still low ($<$0.5% even when 5 inputs are used).

Figures 21 and 22 show the tradeoff between QoS loss and improvement in performance and energy consumption using various approximated versions of `DFT` generated by Green. More specifically, each `DFT` version uses `sin` and `cos` functions that provide different accuracy ranging from 3.2 digits to 23.1 (base) digits [9]. Up to the accuracy of 7.3 digits, no accuracy loss is observed due to the combined sin and cos approximation while improving performance and energy consumption of `DFT` by around 20.0%. Even using the accuracy of 3.2 digits, the observed QoS loss is very low (0.22%), while improving performance and energy consumption of `DFT` by 26.3% and 26.8%, respectively. This clearly indicates the potential of function-level approximation using Green. While not shown, the QoS model constructed for `DFT` is also similarly robust and can be accurately generated with very few inputs.

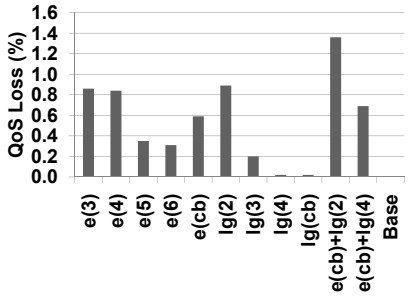


Figure 24. QoS loss of various versions of `blackscholes`.

Figures 23 and 24 show the tradeoff between QoS loss and improvement in performance and energy consumption using various approximated versions of `blackscholes` generated by Green. The version finally selected by Green uses a combination of approximate `exp` functions, (`exp(3)` and `exp(4)`) over different portions of the input argument range, with the `log(4)` function to provide performance and energy improvements of 28.5% and 28% respectively with a QoS loss of less than 0.8%. This result demonstrates that Green is able to successfully combine multiple approximations. The final approximation choice of `exp(cb)+log(4)` was automatically refined from the local choices of `exp(cb)` and `log(2)` to reduce the overall application QoS loss to less than 1%. The QoS model constructed for `blackscholes` is also very robust and the training data set used (64 thousand options) accurately predicts QoS loss on the test data to within 0.1%.

5. Related Work

There are several application domains such as machine learning and multimedia data processing where applications exhibit *soft computing* properties [5]. The common soft computing properties are *user-defined, relaxed correctness*, *redundancy in computation*, and *adaptivity to errors* [2, 16]. Researchers have studied improving the performance, energy consumption, and fault tolerance of applications and systems by exploiting these soft computing properties. However, to the best of our knowledge, we believe Green is the first system that provides a simple, yet flexible framework and programming support for controlled approximations that attempts to meet specified QoS requirements.

Green is most similar to Rinard's previous work [21, 22] in the sense of proposing a probabilistic QoS model (i.e., *distortion model* in [21]) and exploiting the tradeoff between the performance improvement and QoS loss. However, our proposal significantly differs in several aspects. Their focus was on surviving errors and faults with no mechanism for guaranteeing a desired goal would be met. They used a very coarse granularity approach of dropping tasks that only seems appropriate for their domain, which was parallel programs. On the other hand, Green provides support for finer-grain approximation at the loop and function level. In addition, Green provides a re-calibration mechanism that can effectively provide strong QoS guarantees even in the presence of some degree of inaccuracy in the constructed QoS model. Finally, we experimentally demonstrate that controlled approximation can be effective for a wider range of application domains (i.e., not only scientific applications) including an in-production, real-world web-search engine.

In parallel with this work, Hoffmann et al. proposed *SpeedPress*, a framework designed for exploiting performance/accuracy tradeoff using loop approximation [11]. In contrast to SpeedPress, Green also supports function approximation to target a wider range of applications (i.e., not only iterative algorithms) and provides a unifying framework for applying both loop and function approximations. SpeedPress only supports loop approximation but it pursues this goal much more aggressively than Green. It not only skips iterations at the end of a loop but also removes intermediate loop iterations. Green is a programmer assisted framework based on the principle that carefully optimizing a small number of application loops and functions in a controlled manner can provide significant benefits. On the other hand, SpeedPress is positioned as a compiler optimization framework that attempts to approximate as many program loops as possible. While that is a worthwhile goal, we do not believe it is safely achievable given the current state of program analysis technology and could result in unexpected program crashes, especially when intermediate loop iterations are skipped.

Sorber et al. proposed *Eon*, a language and runtime system for low-power embedded sensor systems [23]. Using Eon, control flows of a program may be annotated with abstract energy states. Then, Eon's runtime dynamically adapts the execution by adjusting the frequency of periodic tasks or providing high or low service levels by consulting the annotated energy state and the predicted energy availability. While Green is similar to Eon in the sense of exploiting the tradeoff between energy consumption and QoS loss and providing a dynamic runtime adaptation mechanism, our proposal significantly differs in three aspects. First, Green builds upon a probabilistic QoS model that enables a fast and robust convergence to a desired QoS target. Second, Green directly extends C/C++ languages to allow programmers to implement fine-grained and modular approximations at loop- and function-levels and fully exploit existing code optimizations implemented in C/C++ compilers. In contrast, Eon is based on a high-level coordination language [10] that can compose with conventional languages such as C/C++. Finally, we experimentally demonstrate Green's controlled approximation mechanism is highly effective and robust for a wider range of application domains (i.e., not only embedded sensor applications) including a commercial web-search engine.

In [18], Liu et al. discussed the *imprecise computation* technique with which each time-critical task can produce a usable, approximate result even in the presence of a transient failure or overload in real-time systems. In addition, they described a conceptual architectural framework that allows the integration of an imprecise mechanism with a traditional fault-tolerance mechanism. Our work differs in the following ways. First, Green supports approximation at a finer granularity than tasks. Second, Green provides a calibration mechanism that supports modeling sophisticated, application-specific error functions. In contrast, Liu et al. assumed a rather simple, predefined error function such as linear or convex. Finally, we implemented and evaluated a real (i.e., not conceptual) system to quantify the effectiveness of Green's controlled approximation mechanism.

Several researches focused on floating-point approximation techniques [1, 25]. Alvarez et al. proposed *fuzzy memoization* for floating-point (FP) multimedia applications [1]. Unlike the classical instruction memoization, fuzzy memoization associates similar inputs to the same output. Tong et al. proposed a *bitwidth reduction technique* that learns the fewest required bits in the FP representation for a set of signal processing applications to reduce the power dissipation in FP units without sacrificing any QoS [25]. While effective in improving performance and energy consumption of FP applications, applications with infrequent use of FP operations will not benefit from these schemes significantly. In addition, they do not provide any runtime re-calibration support for statistical QoS guarantees. In contrast, Green is more general, targets a wider range of applications (i.e., not only FP applications) and attempts to meet specified QoS requirements.

Several researches studied the impact of the soft computing properties on the error tolerance of systems in the presence of defects or faults in chips [6, 17]. Breuer et al. demonstrated that many

VLSI implementations of multimedia-related algorithms are error-tolerant due to the relaxed correctness. Based on this, they proposed design techniques to implement more error-resilient multimedia chips [6]. Li and Yeung investigated the fault tolerance of soft computations by performing fault-injection experiments [17]. They demonstrated that soft computations are much more resilient to faults than conventional workloads due to the relaxed program correctness. Our work differs as it focuses on performance and energy optimizations that meet specified QoS requirements instead of fault tolerance.

6. Conclusions

In this paper, we propose the Green system that supports energy-conscious programming using controlled approximation for expensive loops and functions. Green generates a calibration version of the program that it executes to construct a QoS model that quantifies the impact of the approximation. Next, it uses this QoS model to synthesize an approximate version of the program that attempts to meet a user-specified QoS target. Green also provides a runtime re-calibration mechanism to adjust the approximation decision logic to meet the QoS target.

To evaluate the effectiveness of Green, we built a prototype implementation using the Phoenix compiler framework and applied it to four programs including a real-world search application. The experimental results demonstrate that the Green version of these applications perform significantly better and consume less energy with only a small loss in QoS. In particular, we improved the performance and energy consumption of `Bing Search` by 21.0% and 14.0% respectively with 0.27% of QoS degradation. We also showed that the QoS models constructed for these applications are robust. In conjunction with Green's runtime re-calibration mechanism, this enables approximated applications to meet user-specified QoS targets.

Acknowledgements

We would like to thank Preet Bawa, William Casperson, Engin Ipek, Utkarsh Jain, Benjamin Lee, Onur Mutlu, Xuehai Qian, Gaurav Sareen, Neil Sharman, and Kushagra Vaid, who made contributions to this paper in the form of productive discussions and help with the evaluation infrastructure. Woongki Baek was supported by a Samsung Scholarship and an STMicroelectronics Stanford Graduate Fellowship.

References

[1] C. Alvarez and J. Corbal. Fuzzy memoization for floating-point multimedia applications. *IEEE Trans. Comput.*, 54(7):922–927, 2005.

[2] W. Baek, J. Chung, C. Cao Minh, C. Kozyrakis, and K. Olukotun. Towards soft optimization techniques for parallel cognitive applications. In *19th ACM Symposium on Parallelism in Algorithms and Architectures*. June 2007.

[3] L. A. Barroso. Warehouse-scale computers. In *USENIX Annual Technical Conference*, 2007.

[4] C. Bienia, S. Kumar, J. P. Singh, and K. Li. The parsec benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, October 2008.

[5] P. P. Bonissone. Soft computing: the convergence of emerging reasoning technologies. *Soft Computing—A Fusion of Foundations, Methodologies and Applications*, 1(1):6–18, 1997.

[6] M. A. Breuer, S. K. Gupta, and T. Mak. Defect and error tolerance in the presence of massive numbers of defects. *IEEE Design and Test of Computers*, 21(3):216–227, 2004.

[7] E. O. Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[8] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41(6):205–220, 2007.

[9] J. G. Ganssle. A Guide to Approximation. `http://www.ganssle.com/approx/approx.pdf`.

[10] D. Gelernter and N. Carriero. Coordination languages and their significance. *Commun. ACM*, 35(2):97–107, 1992.

[11] H. Hoffmann, S. Misailovic, S. Sidiroglou, A. Agarwal, and M. Rinard. Using code perforation to improve performance, reduce energy consumption, and respond to failures. Technical Report MIT-CSAIL-TR-2209-037, EECS, MIT, August 2009.

[12] J. T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150, 1986.

[13] R. Katz. Research directions in internet-scale computing. In *3rd International Week on Management of Networks and Services*, 2007.

[14] V. Kianzad and S. S. Bhattacharyya. Multiprocessor clustering for embedded systems. In *Euro-Par '01: Proceedings of the 7th International Euro-Par Conference Manchester on Parallel Processing*, pages 697–701, London, UK, 2001. Springer-Verlag.

[15] Y.-K. Kwok and I. Ahmad. Benchmarking the task graph scheduling algorithms. *Parallel Processing Symposium, 1998. IPPS/SPDP 1998. Proceedings of the Symposium on Parallel and Distributed Processing 1998*, pages 531–537, Mar-3 Apr 1998.

[16] X. Li and D. Yeung. Exploiting soft computing for increased fault tolerance. In *In Proceedings of Workshop on Architectural Support for Gigascale Integration*, 2006.

[17] X. Li and D. Yeung. Application-level correctness and its impact on fault tolerance. In *HPCA '07: Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture*, pages 181–192, Washington, DC, USA, 2007. IEEE Computer Society.

[18] J. Liu, W.-K. Shih, K.-J. Lin, R. Bettati, and J.-Y. Chung. Imprecise computations. *Proceedings of the IEEE*, 82(1):83–94, Jan 1994.

[19] R. M. Metcalfe and D. R. Boggs. Ethernet: distributed packet switching for local computer networks. *Commun. ACM*, 19(7):395–404, 1976.

[20] Phoenix Academic Program. `http://research.microsoft.com/Phoenix/`.

[21] M. Rinard. Probabilistic accuracy bounds for fault-tolerant computations that discard tasks. In *ICS '06: Proceedings of the 20th annual international conference on Supercomputing*, pages 324–334, New York, NY, USA, 2006. ACM.

[22] M. C. Rinard. Using early phase termination to eliminate load imbalances at barrier synchronization points. In *OOPSLA '07: Proceedings of the 22nd annual ACM SIGPLAN conference on Object-oriented programming systems and applications*, pages 369–386, New York, NY, USA, 2007. ACM.

[23] J. Sorber, A. Kostadinov, M. Garber, M. Brennan, M. D. Corner, and E. D. Berger. Eon: a language and runtime system for perpetual systems. In *SenSys '07: Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 161–174, New York, NY, USA, 2007. ACM.

[24] Standard Performance Evaluation Corporation, *SPEC CPU Benchmarks*. `http://www.specbench.org/`, 1995–2000.

[25] J. Tong, D. Nagle, and R. Rutenbar. Reducing power by optimizing the necessary precision/range of floating-point arithmetic. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 8(3):273–286, Jun 2000.