Work Scheduling on Heterogeneous Resources

Gabriele Keller Edward Pierzchalski

Supervisor

February 10, 2015

1 Introduction

There are two notable trends in computer hardware. First, CPUs are gaining cores per chip while having reduced gains in clock speed. Second, GPUs are developing into generic computing units. As more computing problems are found to be parallelizable, there is an incentive to try and move work that would have been done on the CPU to the GPU, particularly operations on vectors and streaming data.

1.1 Why Scheduling is Painful

There are several obstacles to effectively partitioning, let alone scheduling, a workload over different work processors. The major issues both revolve around memory, namely copy speed and access patterns.

Speed: Inside of a single GPU, data transfer rates can reach upwards of 100 GB/s, and similarly on a single CPU die up to 20GB/s. However, all of these devices are typically connected over a PCIe bus, which caps out at around 8GB/s. This places strong constraints on how frequently data can be shuttled between components. In the current Accelerate architecture, producer/consumer fusion reduces the number of intermediate structures required to perform work, which helps mitigate this issue. However, this fusion can impede attempts to partition work.

Access: Parallelizable computations may have varying degrees of contiguity in their memory accesses. When computing a simple `map` operation, adjacent output locations depend on adjacent input locations, and so we can partition the work by index almost arbitrarily. Similarly, a tree-traversal monoidal fold can, at any given depth, be split over processors. However, any output index of a backpermute can be affected by the value at any index of the input. Although we can parallelise the work, we cannot partition the memory access. (Expand on examples?)

Correctly partitioning and allocating work in the light of these issues will be a key component of this thesis.

2 Background

Phrase the problem as follows: you have some source code in some language, describing a computation you want performed. It would be nice if, given this source specification, we could partition and schedule the work on to any processing unit available. We will see that not all source descriptions are up to this task.

2.1 Describing the Problem

The most popular frameworks for GPU programming are CUDA and OpenCL, which are both low-level imperative languages. In theory, the problem of partitioning work over the CPU and GPU can be solved at the level of these languages. In practice, even for GPU-focused tasks such as a monoidal scan or fold, efficient implementations tend to require intimate knowledge of the memory and process architecture of the GPU. This is an added cognitive strain on users of these frameworks, and makes composing GPU tasks difficult: a map followed by a scan is substantially different to just a map, or just a scan.

These languages are also impractical for use as a single source language over multiple types of processors - for instance, while CUDA and OpenCL can target CPU architectures, the resulting binaries are substantially less performant than code originally written with a CPU architecture in mind.

The full semantics of the CUDA and OpenCL languages allow for arbitrary write-effects from any kernel to any array, and include operations such as global barriers. This makes reasoning about kernels difficult for the same reason that hidden state impedes compositional reasoning in many imperative languages. The effect is compounded when kernels are expected to run on separate devices; the result is that these low-level languages make for a poor source description.

2.2 Combinators

The frameworks described previously are distinguished by being *imperative:* they describe, step by step, the actions performed by the GPU. Alternatively, we can describe our program using *combinators*, which abstract the combination of computations. For example, `map` is a combinator that takes a function `f : a -> b`, along with some list `as : List a`, and produces a new list, `map f as : List b`, which contains the result of applying `f` to every element in `as`. We don't particularly care how `map` performs this: we just care that the result is what we expect.

This separation between *what* we want to happen, and *how* we want it to be done, makes combinators quite useful as a source language for computations to be performed, assuming the combinators have 'sufficiently sane' semantics. (yeah this needs a better definition) Once the user describes their program, we can interpret the description into whatever implementation we need to in order to satisfy the semantics of the given program.

2.3 Describing Semantics

The design of a language framework for a specialised purpose (for instance, automatically parallelised vector programming) results in a Domain Specific Language, or DSL. Many DSLs are implemented *inside* another host language, commonly called 'embedding'. There are essentially two flavours of Embedded DSLs (EDSLs):

2.3.1 Shallow Embedding

This is almost synonymous with constructing a library. Terms in the DSL correspond to library functions that, when evaluated, directly perform whatever computations that they represent.

2.3.2 Deep Embedding

Alternatively, we can use the host languages' data definitions to construct a direct representation of the combinators. Say we define a new version of `map`, call it `map'`. Instead of a function, define it as a data constructor: an object containing a function and a list, i.e `map' : (a -> b) -> List a -> DSL b`, where `DSL` is a type representing our DSL embedding. The representation data type is typically called the Abstract Syntax Tree, or AST.

This representation has issues: we can't nest it, for instance, since `map'` takes a `List` yet produces a `DSL`. However, if we change `map'` to take a `DSL`, we will have completely removed lists from our list-description language! The solution is to introduce a way to 'lift' lists into our DSL, which we can do by including a data constructor `lift : List a -> DSL a`. We can then give `map'` the type `(a -> b) -> DSL a -> DSL b`.

In order to do anything with these data structures, we need to interpret or evaluate them. For instance, we can define a recursive evaluation function `eval`:

```
eval : DSL a -> List a
eval dsl = dsl match
  lift list          -> list
  map' f anotherDsl -> map f (eval anotherDsl)
```

2.3.3 Benefits of Deep Embeddings

This may seem like unnecessarily complicated overhead, however it gives us the flexibility to do other things with our description of list computations. One of the most important is that once a user has constructed a term in the DSL, we can manipulate it in the host language. For array DSLs, most of these manipulations are for optimisation.

As an example, consider the evaluation of `map f (map g as)`. Semantically, the final result is the application of the composition of `f` and `g` on the elements of `as`. However, since `map` is a simple function, the two calls to `map` will produce two intermediate lists. On a CPU, this is a relatively small (but not ignorable!) overhead; yet if we were to naively send off the work to the GPU as two separate mapping steps (presumably over arrays instead of lists), we would suddenly run into memory bottlenecks.

Compare this with the equivalent deep emedding, `map' f (map' g (lift as))`. Since this is just a nested data structure, we can 'pull out' the functions and compose them, producing a new data structure `map' (f . g) (lift as)`. When we finally evaluate this new DSL value, it will be converted into a call to `map` that only makes a single list, avoiding memory bottlenecks.

Accelerate is an EDSL embedded in Haskell, a high-level general purpose functional programming language. We will discuss the inner workings of the Accelerate framework in a later section.

A large body of work has been produced on generation of parallel code for the GPU and for the CPU [1], however less has been done on the problem of doing so automatically for a general-purpose program. Previous work has investigated dynamic scheduling of a task composed of kernels, however the kernels had to have both CPU and GPU versions provided by the user of the library. (cite) Ideally, we should be able to separate the declaration of our problem from the generation of parallel code to run on whatever hardware happens to be available.

2.4 Accelerate

Accelerate models general GPU programming using a deeply-embedded AST, which is then transformed through a variety intermediate representations and functions over them. These transformations perform a wide variety of changes aimed at converting

Take up two pages describing the internals of Accelerate. Note: get a hand on R/T's code so we can see exactly how they go from fissioning the SimpleAcc AST to actually running code on separate processing units.

Aside: An Annoying Operation

There is one family of array operations in the Accelerate DSL that is particularly non-performant for heterogeneous scheduling. The `permute` and `backpermute` operations use a function from one set of array indices to another in order to permute an array. For instance, `backpermute f old` will produce an array `res` such that `res ! i = old ! (f i)`.

Since the range of `f` is unrestricted, the value at any index of the output array may depend on the value at any index of the input array. This complicates the partitioning of the work: although we can split the task of constructing the resulting array, we can't partition the input array between processors.

In many GPU use-cases, the strongest constraint on throughput is memory transfer speed, which makes copying the entire input array to each processor infeasible. The result is that these permutations are *de facto* synchronisation points: every operation before them, even when split, must all have their results returned and combined before the permutation.

2.5 Scheduling

3 Previous Work

Obviously going to mention R/T's stuff on implementation, but how much should we talk about scheduling theory? How relevant is it?

4 What We're Going To Do

There is much potential work within the scope of what has been discussed so far. For the purposes of this project, the high-level goal is to have Accelerate able to fission a program written in its DSL (already done by Trevor and Ryan - but not automatically?) into 'jobs' that can be scheduled onto a heterogeneous set of workers (CPUs, GPUs) with a demonstrable performance boost, while also addressing some of the issues expressed in previous work. (Need to clarify what I'm adding that isn't just 'making the scheduler better') In addition, the following are sub-goals that are amenable to investigation along the way:

Small Array Computations

One minor issue not addressed by existing approaches to heterogeneous resource scheduling is the cost of operations on small arrays. Loading, running, and using the result of a GPU kernel has significant overhead. Below a certain array size, the processing time of operations on an array is dominated by the high latencies of both loading the GPU kernel and data transfer between the CPU and GPU.

In these cases, it would be more effective to perform the operation on the CPU. The existing Accelerate framework does not differentiate workloads by array size, and so does not take advantage of this. Once we adapt

Learning to Share

Even purely parallel problems can benefit from also being split onto the CPU. (cite Ryan/Trevor, words about how CPUs can significantly assist a GPU. Note: find out why Ryan/Trevor didn't add the benchmark for the CPU+GPU performance on the Black Scholes stuff.)

What You're Good At

Recomputation can be beneficial for otherwise memory-bound programs... as long as they're running on the GPU. Otherwise, you want to minimize recomputation.

Extensions

Currently,

5 Work Schedule

As this is partially a development thesis, we must form a development plan. The below schedule details the hopes and expectations for when parts of the development phase will be completed. Included are milestones to be used as checks on whether these expectations have been met. The final report is expected to be written in tandem with the development stages: each week

and each milestone has an implicit requirement, 'have something written down about this'. (Need to clarify what parts are research and what parts are development. Also need to clarify what granularity is required other than 'mess with Accelerate to learn about it, then mess with Accelerate to add something to it'.)

Week 1 - 4: Familiarise myself with Accelerate.

Week 1: User-facing DSL.

Milestone: a simple Accelerate program, say a Newtonian gravity simulation.

Week 2 - 3: Compilation phases.

Milestone: implement a simple phase?

Week 3 - 4: Accelerate backends (mainly LLVM)

Week 5 - 9: Write scheduler.

Week 5 - 6: Have fissioning working on the AST.

Milestone: produce index-partitioned ASTs at kernel generation step.

Milestone: produce ASTs with optimisations for small array sizes.

Week 6 - 7: Emit to both CPU and GPU-facing backends.

Milestone: have above kernels produced and split between CPU and GPU targets, at some static fraction.

Milestone: implement small-array CPU optimisation as part of scheduling.

Week 7 - 9: Implement dynamic scheduling.

Milestone: a runtime utility to determine what backends and processors are available, along with their processing speeds and bus transfer limits (both between and within processors).

Milestone: have task fissioning and task allocation take into account memory dependencies and relative power of processors.

Week 10 - 12: Benchmarks.

Milestone: Compare performance on common problems (Black Scholes, n-body) with and without scheduler.

Milestone: Compare vs. previous scheduler implementation from make formal: Ryan and Trevor, hopefully seeing improvement.

Milestone: Show improvement due to separate stages of scheduler optimisation: small-array CPU scheduling vs. task partitioning.

Milestone: Show improvement due to dynamically informed allocation vs. static strategy.

Week 13 - Due: Buffer time for the inevitable. Clean-up for final report. Extensions, if applicable.

References

[1] Janghaeng Lee, Mehrzad Samadi, Yongjun Park, and Scott Mahlke. Transparent CPU-GPU collaboration for data-parallel kernels on heterogeneous systems. In *Proceedings of the 22nd international conference on Parallel architectures and compilation techniques*, pages 245–256. IEEE Press. URL `http://dl.acm.org/citation.cfm?id=2523756`.