An Investigation Into The Performance Of A Plastic Parallel Programming System

Mark Jenkins

School Of Informatics
University Of Edinburgh
2017

Abstract

This is the first part of an MInf project which will span two years. In this report, we combine the ideas of contention aware scheduling, plastic programming, and skeleton programming, to create a library where incarnations of its skeletons will explicitly work together to share and optimize the use of system resources. We then implement multiple experiments to investigate if we can obtain any performance improvements over traditional parallel programming, finding

Insert results summary here. We found X, Y, Z..

Assess it's usefulness,

We have shown X, Y, Z

how it could be taken further, and possibly used to implement a useful tool for programmers.

Planned for part 2 of Minf project, and other possible future work.

Contents

Chapter 1

Introduction

Throughout the history of computing, computer programmers and hardware engineers have exploited parallelism, with software and architectural innovation complementing technological improvements to provide increased performance. Architects have exploited bit and instruction parallelism, and now programmers are increasingly exploiting task/data parallelism in their applications.

It is recognised that writing correct and efficient parallel programs is hard, [3] as the introduction of parallelism introduces a whole host of new problems, (e.g. unreproducible bugs due to race conditions, and difficult debugging as high level instructions need ot be decomposed into atomic assembly code to understand what is going on). The main purpose of multi-threading is improved performance, however parallel programs can often be slower then their serial counterparts, be it due to dependencies or hardware contention. Overhead must be balanced such that we don't introduce so much more work organizing threads and computations that our performance gains disappear.

Utilizing hardware efficiently is a major research challenge, especially considering that, today, a single program or library needs to deal with multiple different incarnations of the task it is trying to run. These arise from different circumstances, ranging from the hardware it is running on to the task size. Creating a "one size fits all" solution has become increasingly difficult, especially for high performance parallel applications. As such, there are solutions to help mitigate this problem [1], which utilize this idea of plastic programming, that is, changing the specifics of an implementation depending on the circumstances. However, such solutions are only "plastic" at compile time, and during runtime if circumstances change they cannot adapt. In particular, the most common situation a program may encounter would be the sharing of the computer's resources with other programs. This is again exacerbated for high performance parallel programs, as they typically attempt to use all the resources they can

get their hands on.

Even when we have an ideal parallel program, often resources are shared between multiple programs and users, leading to resource contention. This is the case for most use cases, from the serious computing resources in data centres, to the low powered hardware of mobile phones. Even in our own computing labs, we have messages to the effect of "Do not leave applications running on this machine" or "please nice your programs", which are often unseen or disregarded.

It is known that in such a situation with two programs, with careful selection of program parameters, we can obtain a better average runtime for both programs [2].

The aim of this project is to simplify the challenges of parallel programming and to provide improved performance by utilizing three key ideas:

- Co-Scheduling
- Plastic Programming
- Skeleton Programming

and to investigate the performance ramifications. Combining these ideas results in these problems becoming particularly tricky, with many different challenges involved in incorporating them.

The layout of the report is as follows; In chapter 2 we will provide an exploration of the ideas behind the report, and give a quick overview of how they are combined. In chapter 3, we will detail the thought process behind the design of our system, and then in chapter 4 we will then go into the specifics of the implementation of the system. Chapter 5 presents a carefully selected series of experiments to assess this system, and in chapter 6 we discuss the results of these experiments and their ramifications. Chapter 7 provides the roadmap of future work for the second part of this two year MInf project and beyond. We then end with providing our final thoughts on the topic in chapter 8.

Chapter 2

Background

In this chapter, we will detail the current approaches to parallel programming. We will then explore the three key ideas requisite to this project, such that we can discuss how they are combined and the implications.

The main new idea in this project is that of dynamic contention aware scheduling and optimization. It has been shown to be an important factor in multiprogramming systems with performance implications [2]. Plasticity is a technique to respond to this challenge, and take it further. This results in complex code, making it hard to ensure correctness. So we use skeletons to abstract this complexity away from the programmer. It also has the nice side effect of dividing the challenge into a pattern-by-pattern basis.

In this project we will produce such skeletons, and investigate the performance implications of these ideas, as it is not known whether they will have a significant effect.

2.1 Current Solutions

Current solutions for parallel programming include:

- Pthreads (POSIX Threads)
- MPI (Message Passing Interface)
- OpenMP (Open Multi-Processing)

These are the more conventional methods of parallel programming.

Pthreads provides the highest level of fine grain control, leaving most of the work to the programmer. As such, some parallel processing techniques may only be possible with Pthreads. It is implemented in the POSIX library, and is well integrated with the standard GNU compiler collection (**gcc**). Pthreads only makes sense with shared memory architectures.

MPI is a communication library, and as such mostly details how processes can communicate. It is the dominant model used in the high performance computing industry today [4]. It can not only be used for a parallel program running on a single machine, but can also be used to implement a parallel program running on a distributed memory system.

OpenMP is comprised of compiler directives, runtime library routines, and environment variables. It is mostly used for loop parallelization, and is much higher level. It is also not limited to C. Like Pthreads, OpenMP is intended for shared memory architecture as it is thread bound.

Each of these methods have their own way of dealing with the complications introduced with parallel programming, which range from race conditions to limited scalability. These new problems can certainly be overwhelming to a traditionally sequential application programmer, so much so that there are entire books dedicated to the use of each of these particular parallel programming methods.

2.2 What Is Contention Aware Scheduling?

It is known that in multiprogramming systems, with many programs running simultaneously, the choice of program to socket mapping significantly affects the performance of the system [2]. Just considering two programs running on the same socket, we can see from the graph in figure 2.1 that certain programs perform differently with others, with some strange cases where the programs actually display better performance when running in contention with another. This problem is called co-scheduling.

So with this evidence, we can see that if we take into account these factors in our scheduler, we may obtain better overall performance. The outcome of the LIRA paper concludes that throughput gains of 3-7% can be seen. Socket/resource aware scheduling in this manner is called a contention aware scheduler. Adding in the plastic programming idea could make this particularly powerful, because we know and control the specifics of the implementations, and not only can we control what program runs where, we can also adjust the implementation the program is using.

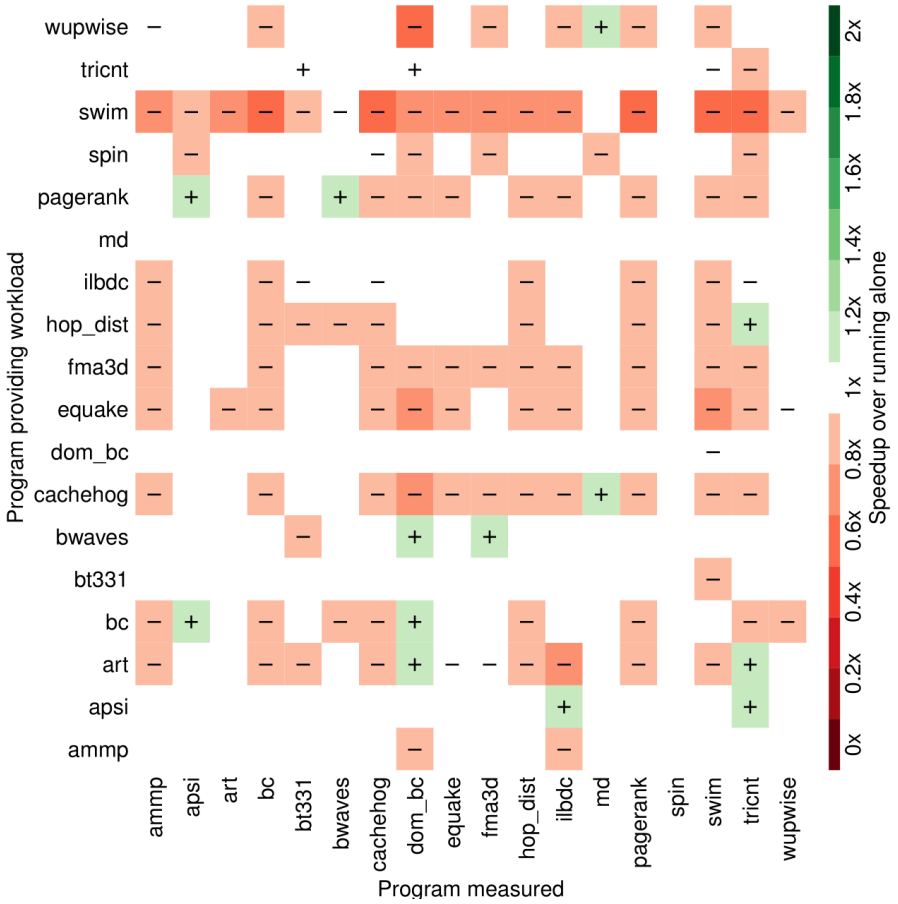

Figure 2.1: Taken from the LIRA paper (CITE LIRA): Pair-wise speedup of programs, comparing sharing a socket to using separate sockets. Boxes annotated with a − indicate cases where performance decreased, and + where performance increased.

2.3 What Is Plastic Programming?

When programming an algorithm, there are often many choices about the specific implementation which can greatly affect performance, and the best choice depends on the circumstances of the problem. We tend to have more choices with parallel programs, but this is the case even for sequential programs. As an example, for a sorting problem with a large input size, radix sort would perform best, whereas for a small input size, insertion sort would be better. So naturally, in the interests of performance, we can conceive of a better overall implementation by combining the two approaches, so while the task size is large we would use radix sort, and then once it is reduced we would use insertion sort.

An example of such a situation can be seen in figure 2.2, which shows three algorithms with different runtime curves, which depend upon the array size. Combining these algorithms would provide an improved algorithm, with D1 and D2 showing the optimal decision points where a plastic programming system should switch algorithms.

Such compositions are commonplace, such as the sorting example discussed in the PetaBricks paper introduction [1]. Compared to PetaBricks, our implementation of plastic programming will be a little different. With the PetaBricks system, the programmer must specify multiple implementations for the compiler to switch between. In our system, we provide the various implementations, meaning no extra effort is required from the programmer.

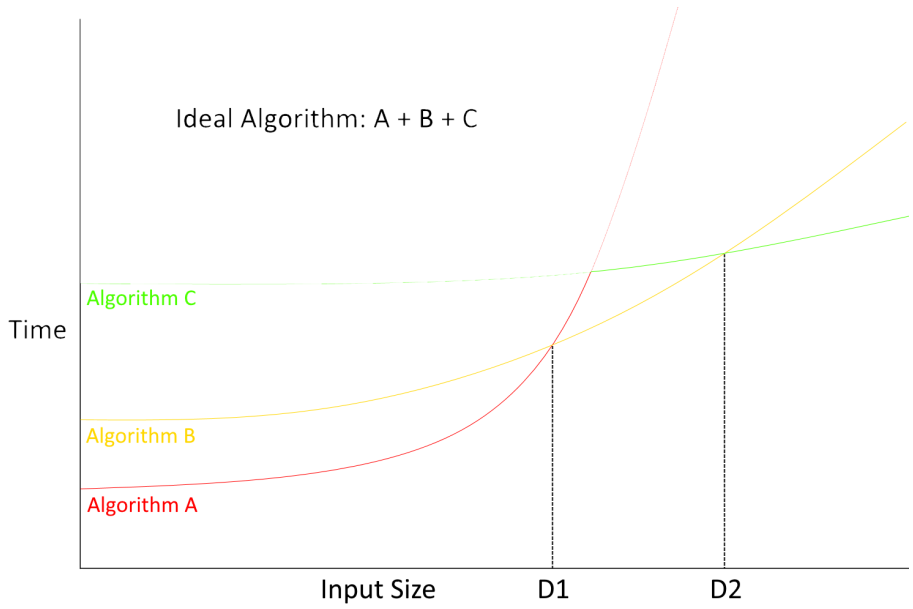

Figure 2.2: A graph showing three algorithms with different runtime curves, which depend upon the array size. Combining these algorithms would provide an improved algorithm, with D1 and D2 showing the optimal decision points where a plastic programming system should switch algorithms

2.4 What Is Skeleton Programming?

Skeleton programming is a high-level programming model. Skeletons will allow us to abstract away all the complexity involved in parallel programming, plastic programming, and co-scheduling. The essence of skeleton programming is that the skeleton provides the core structure of an algorithm, the user provides some code (In our case, a function), which then produces a correct program for the task at hand. The skeleton handles the hard-work of providing and optimizing the code (In our case, dealing with parallelism, plasticity, and co-scheduling).

The consequences of this are twofold:

- Errors are reduced substantially, as parallel programming is not easy, even without plasticity and contention aware scheduling.
- We can assess the program's complexity, since we know the algorithmic details of the skeleton.

Typically, multiple skeletons are combined to produce a more complex program, for example, a common combination is map and reduce (specific skeletons will be discussed in section 3.1.1.) The ability to combine skeletons makes them a powerful tool, allowing programmers to easily create clean complex programs.

Previous work to make parallel programming simple utilizing skeletons includes:

- SkePU
- The Mnster Skeleton Library Muesli

These methods simplify the problem, with both utilizing skeleton programming to do so. Since each SkePU skeleton has multiple implementations, and SkePU automatically selects the expected fastest implementation variant, it is the closest to the system we wish to implement. In contrast to our system, this selection is not dynamic, i.e. it does not change during runtime.

2.5 Summary

In this chapter, we described the three fundamental concepts behind this project, contention aware scheduling, plastic programming, and skeleton programming. With just contention aware scheduling, we can achieve greater performance [2]. In this project we will take this idea further, adding plastic programming, so we

can adapt the implementation of programs at runtime. This lets us change the implementations to ones which we know work well together. Combining these two ideas results in considerable complexity, so to make it simple for a programmer to use and easier for us to program, we use skeletons to abstract this complexity. The details of this complexity and how these ideas are combined are described in chapter 4.

The intent of this project is to explore the extent to which this approach can outperform a contention aware scheduled system which doesn't utilize plasticity.

Chapter 3

Design

In this chapter, we will discuss the design of the system created to investigate this problem, and detail how such a system could be extended for the future.

3.1 System description

The ideas described in the background section are combined to produce a skeleton programming library with plasticity and contention aware scheduling. To keep it simple, we build the system incrementally, starting with a single parallel pattern, later adding plasticity and contention aware scheduling.

3.1.1 Skeleton Foundation

As discussed in the background section, one of the key ideas behind the project is that of skeleton programming, using predefined patterns to aid the programmer. Possibly the most common skeleton is map, which is a skeleton that takes a function, an array of data, and applies the function to each member of the array. The map operation itself is inherently parallelizable, since each 'task' (processing one element of the input array) is independent, so we can simply assign different chunks of the input array to different threads. This is called an "embarrassingly parallel" problem.

Other such problems include map-array, reduce, and scan. These are all implemented in SkePU with a corresponding skeleton. In our case, we will focus on the map-array pattern, with further patterns left for possible future work. Map-array is similar to the map pattern in that it applies a given user function

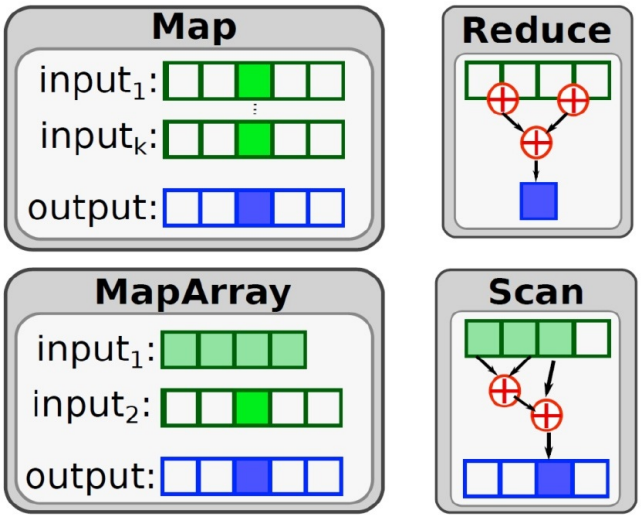

Figure 3.1: Taken from SkePU documentation: Examples of skeletons

to each element in a list, however map-array also allows the function to access a user provided array. A diagram to better explain this is given in figure 3.1.

Map-array was chosen as the map pattern is likely the most well known pattern and certainly one of the most useful, and map-array provides further functionality on top of this. It also provides a good basis for developing further patterns, and allows for complex testing, which will be covered in the evaluation section of this report.

3.1.2 Adding Plasticity

To implement plasticity, we add the ability to vary three key aspects of the implementation of a single instance of map-array:

- Thread count - The number of threads we split the tasks between
- Thread pinnings - The particular CPU core each thread runs on
- Schedule - How to divide tasks between threads

The thread count and pinnings are self explanatory. The schedule however requires some explanation.

The most basic method to divide the tasks is to give an equal amount to each thread. This is fine if the complexity of the tasks is uniform, but if it is skewed,

the amount of computation to be done by each thread is imbalanced. This is illustrated in figure 3.2. This is because we have idle cores during computation, which is a wasted resource in a multi-threaded execution. However, if we allocate the tasks differently, we can obtain better performance, as illustrated in figure 3.3.

So load balancing a workload is critical to performance in such a multi-threaded application. However, optimizing the task distribution in this manner is non trivial, and it depends upon the computation to be done as well as the number of threads and other resources available at runtime.

A solution to this problem is to provide many different task distributions, and let the user pick or the machine select which distribution to use. OpenMP documentation calls these schedules, and some examples of these are:

- Static - An equal number of tasks allocated to all threads
- Dynamic individual - Each thread retrieves one task at a time, and once completed, it goes back for more
- Dynamic chunks - Each thread retrieves N tasks at a time, and once completed, it goes back for more
- Tapered - Each thread starts by retrieving N tasks at a time, and as the computation continues, it retrieves fewer and fewer

Thread count and schedule were chosen as they seem the most critical to performance, and thread pinnings was added as this is was investigated in the LIRA paper [2] as a factor contention aware scheduling could exploit.

Once we have added plasticity, we can experiment with the specifics of an implementation, and see how they affect the performance of the system. This would be the use case of utilizing our library with no other program running, (So no contention aware scheduling), and we can explore how we can adapt the program using plasticity at runtime in this case. We may be able to improve performance even under these conditions, depending upon the configuration of the machine (e.g., is there more CPU cores available) and the problem (e.g. do we have many small tasks or few large tasks.)

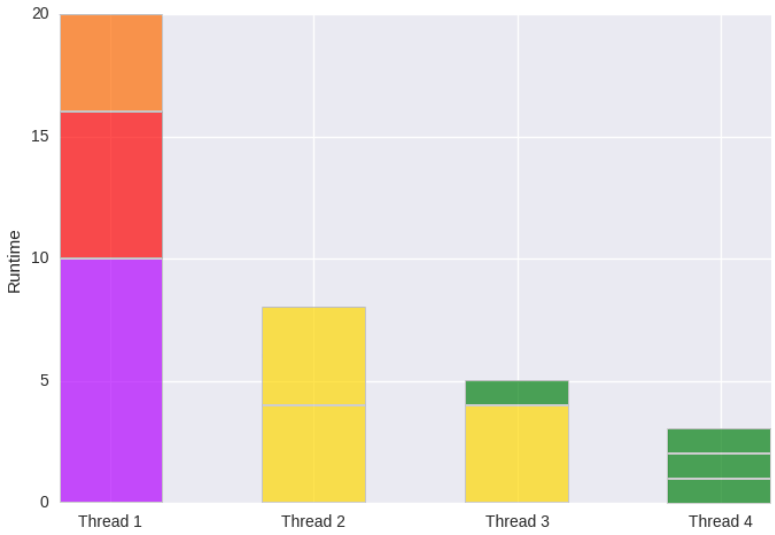

Figure 3.2: A worst case scenario of a static schedule assigning each thread an equal number of tasks. Suppose we have a set of twelve independent tasks with the following set of execution times: {10, 6, 4, 4, 2, 2, 2, 2, 1, 1, 1, 1}. With four threads, a simple division of tasks would be three tasks each distributed in order. This figure illustrates this distribution of work. Note that the total execution time is 20 time units.

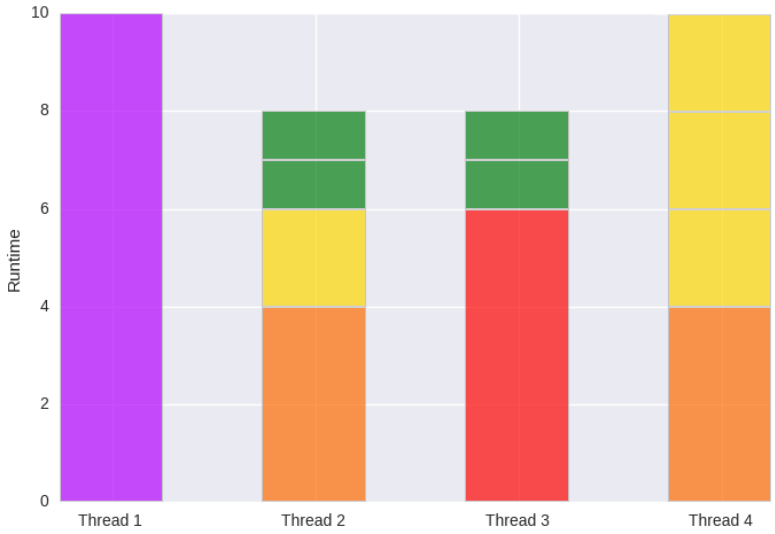

Figure 3.3: An optimized version of figure 3.2. Here we have the same set of tasks, but we have a better distribution of work, so the total execution time is 10 time units.

3.1.3 Contention Aware Scheduling

To add contention aware scheduling, we need multiple applications using our library to be able to collaborate, and adapt their behaviour accordingly. To do this, we use a separate controller application, with which all instances of our program can communicate. This provides a single known point of contact, and a designated thread for computing program parameters with respect to all aspects of the system.

Once our programs can communicate, and we can control each aspect of them, we can implement contention aware scheduling. In this phase of the project, we simply program a set of predefined actions for the controller to take, in order to manually control what each implementation does, as we are only investigating if this approach seems promising. We leave implementing some algorithm for automatic parameter tuning for future work.

As an example of the communication model is given in figure 3.4.

Now that we have contention aware scheduling, we can experiment with multiple programs running on a system at once. An example of how contention aware scheduling can be enhanced with plastic programming is given in figure 3.5.

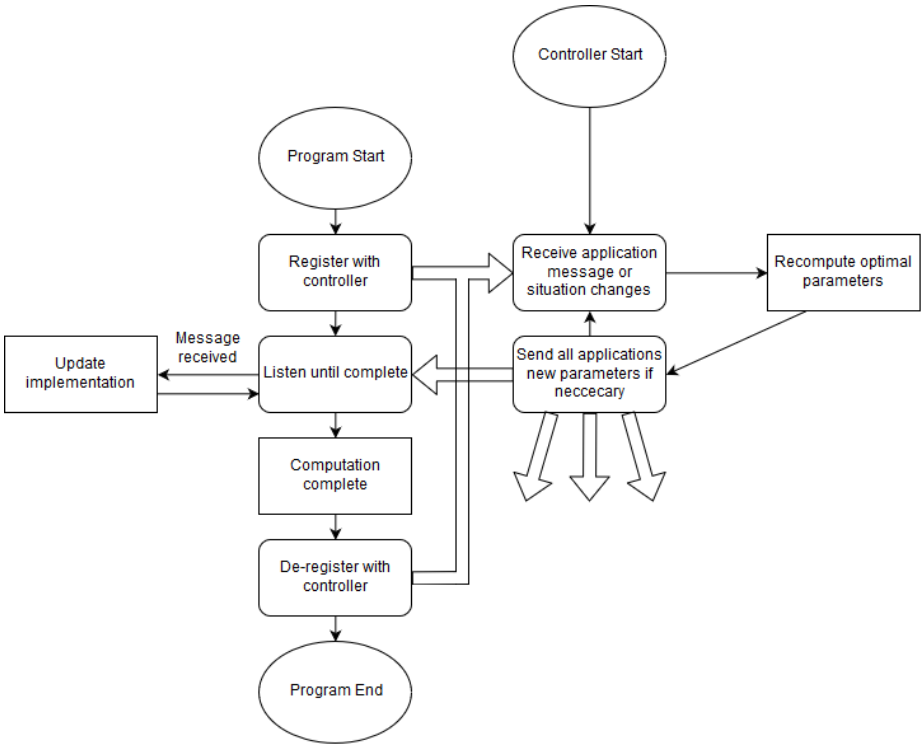

Figure 3.4: Communication model for communications between applications and the controller. Thin lines represent program flow, thick lines represent inter-process messages.

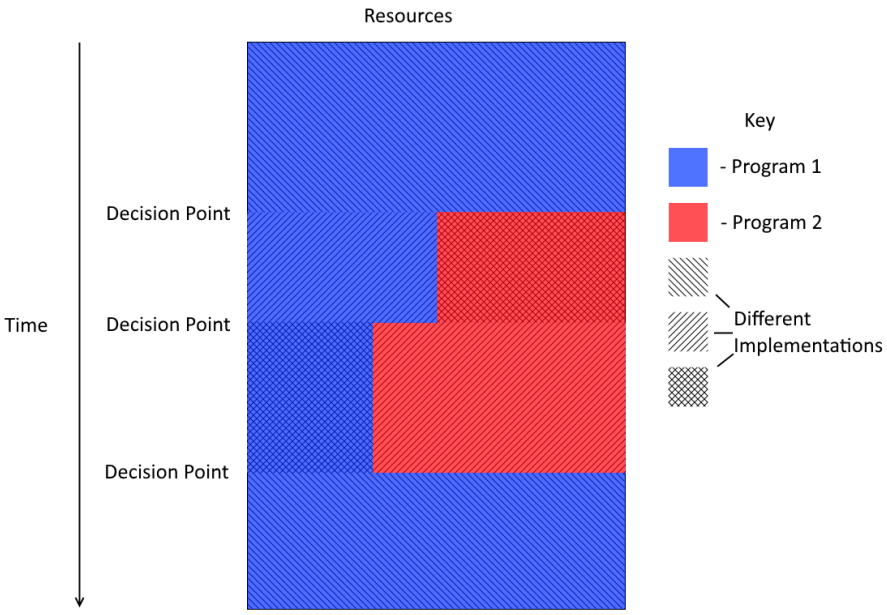

Figure 3.5: An example of how contention aware scheduling can be enhanced with plastic programming. We have two programs, represented by different colours. As time progresses, we see that when program 2 launches, we have a decision point. Here, the system decides how many resources to give each program and what implementation they should use, according to what is optimal (In this project, we do this manually.)

Moving further in time, we have another decision point. This one is triggered by some change in the machine, which means the optimal configuration has changed. So the system updates the configuration of each program, and continues.

Program 2 then terminates, triggering another decision point, and the system again updates the configuration.

3.1.4 Evaluation

To properly evaluate the outcome of this project, we need some way of testing the libraries performance. To this end, we implemented a synthetic program which evaluates the library with an artificial workload, collecting and recording metrics detailing the libraries performance with different parameters.

We also need points of comparison in terms of performance. So, in addition to our synthetic program, we also implement an equivalent sequential and an OpenMP version, in which we can vary similar parameters and produce comparable statistics.

The detailed experimental program will be discussed in 5.

Chapter 4

Implementation

In this chapter, we present specific implementation details of the project, and cover the problems that occurred and how they were solved.

Since the system was developed incrementally, we will cover the implementation in a similar fashion, starting with the implementation of a basic skeleton function, then adding plasticity, and contention aware scheduling. Finally, we finish with the applications developed to evaluate the project.

C++ was chosen as a basis, since it's fast, and provides language constructs such as templates and overloading. Another, personal, reason is that I wished to learn something new, as I had not used C++ before.

The parallel backend that the system is based upon is Pthreads due to its wide availability, and the level of fine control. It allows us to tune all parameters of the program and implement functions which are not possible with other solutions, e.g. detailed metric analysis.

4.1 Skeleton Foundation

The first skeleton we will implement is the map-array skeleton, as described in section 3.1.1. It is reasonably straight forward to create a sequential skeleton, using c++ templates to create a templated function which takes a function amongst other things as it's arguments. This will be our skeleton. The interface of our skeleton is given in figure 4.1, and a usage example in figure 4.2.

```
1
2   template <typename in1, typename in2, typename out>
3   void map_array(deque<in1>& input1,
4                  deque<in2>& input2,
5                  out (*user_function) (in1, deque<in2>),
6                  deque<out>& output,
7                  string output_filename = "",
8                  parameters params = parameters())
```

Figure 4.1: Interface of our map_array skeleton. The first four variables are the two input arrays, the function to apply, and the output array respectively. The output_filename variable is the filename to record the metrics output in, and params sets up the initial parameters we will use. These last two are optional.

```
1
2   int user_function(int in1, deque<int> in2) {
3      return in1 + in2[in1];
4   }
5
6   int main() {
7      // Inputs.
8      deque<int> input1(ARRAY_SIZE);
9      deque<int> input2(ARRAY_SIZE_2);
10
11     /*
12      * Put data in inputs.
13      */
14
15     // Output.
16     deque<int> output(ARRAY_SIZE);
17
18     // Start mapArray.
19     map_array(input1, input2, user_function, output);
20
21     /*
22      * Record output.
23      */
24  }
```

Figure 4.2: A usage example of map_array, here we apply our user_function to each element of input1. The size of our two input arrays need not match, but the size of the input1 and output arrays must.

Parallelizing this presents two problems, namely:

- How to divide tasks amongst threads (In map array, one task is one application of the user function to an element of the input array)
- How many threads should be used

Currently in our implementation, both of these parameters must be specified beforehand and will not change at runtime. Both are aspects of the skeleton which we would like to be variable, to provide different implementations for the plasticity portion of this project. To this end, we need to implement multiple different schedules, and the ability to use a variable number of worker threads.

To implement multiple schedules, we use a bag of tasks object. A bag of tasks is a collection of independent, usually similar tasks which are to be executed. It is a model which is usually combined with some form of parallelism.

Each thread will be given the location of the same bag, and will retrieve a specified number of tasks at a time from the bag. Since each schedule is essentially consists of retrieving a different number tasks, we can use this to implement the basic schedules (Static, Dynamic chunks, Dynamic Individual). More complex schedules can easily be added in the future, as they can reuse the get tasks method, an simply adjust how many tasks each thread retrieves at one time.

The bag also provides the main source of inter-thread communication using shared memory, and contains various semaphores for controlling the threads. This will be expanded upon in sections 4.2 and 4.3.

Because we use this bag of tasks object, it gives us the basis for further extensions, such as using multiple bags, adding task stealing, or adding tasks to the bag during computation.

Providing a variable amount of worker threads is simple now that we have the bag of tasks, all we need to do is to adjust our initial calculations when calculating how many tasks each thread should receive according to the schedule. These worker threads will be spawned by the original thread, the "main" thread. It will block until all computation is complete, it will then join with each of the worker threads.

(*** I think that the organization of these sections, particularly discussing the bag of tasks, can be organized better ***)

4.2 Adding Plasticity

We already have a sort of compile time plasticity in our system, in that we can choose some parameters of the skeleton before compilation. We can choose the number of worker threads and the schedule used. The other main aspect we would like to control is what CPU core each thread executes on. This is called processor affinity or thread pinning. This was added to the skeleton by adding control variables to the bag of tasks, which control the CPU affinity of each thread. Each thread then simply sets it's affinity to the intended CPU.

To add runtime plasticity, we need to be able to change the implementation of the skeleton on the fly. The most straightforward method of achieving this is to stop all computation, terminate worker threads, update the parameters, and resume computation. Previously, (as described at the end of section 4.1,) the main thread would block until the computation is complete, and would then join with each of the worker threads. We now change this so that, instead of blocking, it monitors the computation and can instruct the threads to terminate at any point (using the afore mentioned bag of tasks described in section 4.1 for communication). This allows it to prematurely terminate the computation, and then restart it with the new desired parameters. When this is done and what the new parameters should be is another matter. A future system would calculate these things dependent upon the current state of the system, but for our preliminary investigation, we manually produce them in a synthetic environment, and communicate them to the main thread using the controller application, as described in section 4.3.

(*** Above paragraph may need to be split up, as it's pretty huge as is ***)

An obvious optimization of this system would be to modify the implementation without the need to terminate threads. This graceful switching of the implementation is left for future work, as this optimization is very complex, and would likely result in a marginal speedup proportional to the number of times we wish to change the implementation. We don't plan on switching implementations excessively, and any delay added could be overcome in testing by increasing the input size, and this project is only investigating if this approach to parallel programming is promising.

4.3 Contention Aware Scheduling

As detailed in section 3.1.3 of the design chapter, we will use a separate controller application to co-ordinate and control all programs using our library. This simplifies the implementation significantly, as it gives us a single point of contact, and a single place to calculate an optimal configuration for each program. To

achieve this, we need our separate controller application, and two main additions to our system:

- Inter-process communication
- Separate communication/control thread for each program

We use the zeromq library to provide inter-process communication for it's speed, although it could be replaced with another method. It uses tcp sockets, and serves to simplify inter-process communication. Currently I have made simple use of zeromq for now, and may optimize it in the future.

In order to communicate with the controller while a program is using our library, we utilize our main thread, which will manage the worker threads, switch implementations when instructed, and will clean up after we have finished our computation. The main thread communicates with the controller, registering with it when we start and de-registering once complete. During computation, it listens to messages from the controller using non-blocking communication. This allows us to also check if we are finished processing. For the initial registration and de-registration, we use blocking communication, as these messages are vital to the system. This system is illustrated in figure 3.4, and figure 4.3 shows the communication model.

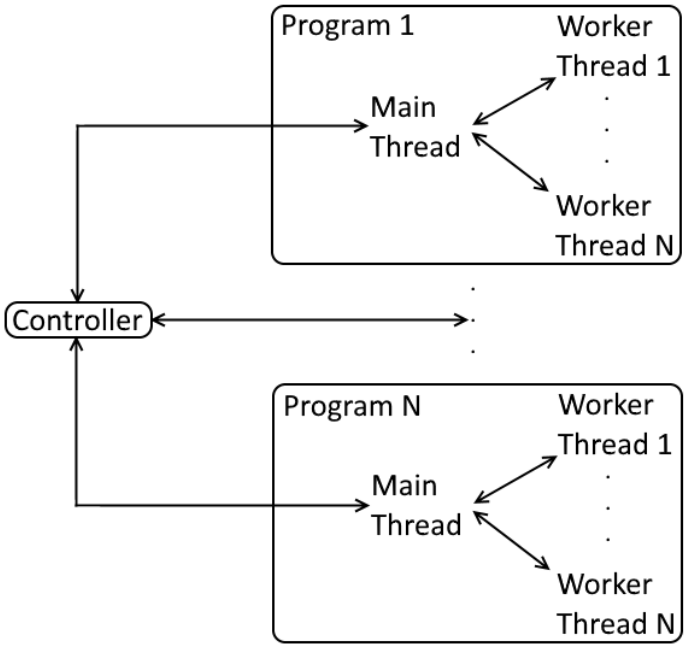

Figure 4.3: Overall communication model of the system, with an arbitrary number of programs, with an arbitrary number of threads. Two way communication occurs between the controller and each main thread, and then between each main thread and it's worker threads.

4.4 Evaluation

To evaluate our library, we need programs to run and test the library under differing conditions, and we need programs to represent competing approaches to parallel programming.

The main map-array test application is a tool which can read experiment parameters from a configuration file, and then run each experiment in sequence, recording various metrics for later analysis. This provides a convenient framework for carrying out experiments, and easily enables us to queue up a set of experiments. This is important because such experiments can take hours.

For meaningful analysis, we need to synthesize a workload. One which we can scale, so we can test different sized tasks and varying task distributions. To do this we use the Collatz function to generate a CPU intensive workload. A constant starting number is used, and the sequence is repeated multiple times, to scale the workload.

The comparison programs consist of a purely sequential implementation, and an OpenMP implementation. These were chosen because the former represents an implementation a traditional sequential programmer would use, and OpenMP because it is a popular method of parallelizing code, with a focus on performance and a simple interface. They are set up such that they can use the same synthetic workload we use for testing our library.

The final program used for evaluation is a python program which was used to create graphs from the generated data.

Chapter 5

Experimental Methodology And Program

The primary goal of this project is to investigate if we can achieve better performance using our approach to parallel programming compared to other methods. Accordingly, we will focus our experiments on evaluating the performance of the system. To this end, we have created a carefully designed set of experiments to asses the system under different conditions, conditions which may be found in future applications. We also evaluate other approaches to parallel programming in order to directly compare their performance with ours. In this chapter we describe our evaluation methodology and the particular experiments we will run. The outcome of each experiment will be discussed in the results chapter in the same order they are presented here.

5.1 Evaluation Methodology

All experiments were run on an Ubuntu virtual machine, with four cores at 4GHz, 4096MB of memory, and the rest of the system at idle. All programs were compiled at optimization level 3, and every result averaged over 5 repeats.

Multiple different statistics can be collected for each thread:

1. Total runtime
2. Time spent doing work
3. Time spent in overhead

4. Time blocked by main thread
5. Number of tasks completed

In these experiments however, we will focus on total runtime, as we are mostly concerned with the overall performance. The other statistics were mostly used during development.

5.2 Programme of Experiments

In this section, we detail each experiment, with the thought process behind it, how it will be executed, and the expected results if appropriate. The actual results will be presented in chapter 6, where we will compare them with our expectations and discuss their implications.

In each experiment we will vary a different set of parameters. These parameters can be partitioned into three separate classes; Input Parameters, Resources Granted, and Skeleton Parameters. More parameters can be added, (e.g. different resources granted,) but this is left for future work, as currently we only want a preliminary investigation into these ideas. The parameters we can change are:

Input Parameters:

- Number of tasks - The number of tasks to be carried out
- Task grain - The complexity of the tasks, possible values are small, medium, and large. (*** Add med takes 10x as much time as small etc ***)
- Task grain distribution - If we have range of task grains, this describes how they distributed throughout the input array. Possible values are uniform (evenly distributed throughout,) and biased (The more complex tasks are gathered together at the start of the array. This is the worst case scenario for the static schedule.)

Resources Granted:

- Number of CPU cores - The amount of CPU cores provided in the machine.

Skeleton Parameters:

- Number of threads used - The number of threads our program uses for multi-threading.

- Thread pinning - The configuration of threads pinned to CPU cores. Possible values are (*** INSERT POSSIBLE VALUES HERE ***)
- Schedule - How the tasks are distributed amongst threads. See section 3.1.2 for the possible schedules.

(*** Add in biased = 1st 1/4 large, rest small ***)

Our first experiment will test how much overhead our metrics collection adds. Experiment 2 then establishes the baseline performance of our library, comparing it to see if it is on a par with an OpenMP solution. Next in experiment 3, we investigate the overhead that has been introduced by contention aware scheduling and plasticity. Experiment 4 then shows the importance of choosing the correct schedule, and how our system could deal with a suboptimal schedule using plasticity. This is again compared to an OpenMP solution.

Thus far, experiments 1-4 have all been conducted with a single application utilizing our library. In experiment 5, we simulate a situation with multiple applications using our library. We investigate the performance of the system with many applications competing for resources, with multiple strategies for how to share them.

5.2.1 Experiment 1 - Metrics Collection Overhead

The aim of our first experiment is to investigate if our method for collecting metrics about the system adds any significant overhead that we must take into account in further experiments. To this end, we compare the runtimes of our system both with and without metrics collection, to see if there is any significant difference. The system is run with no controller application or runtime plasticity, since we only wish to measure the overhead added by our metrics collection.

The experiment parameters are designed such that it is a worse case scenario, and as a result would produce the most significant difference in runtimes. This worst case scenario would be many small tasks, with multiple threads, since metrics functions are called before and after tasks are completed, and are managed on a per-thread basis.

Since our metrics collection consists of some function calls and updating tallys, I don't expect the performance impact to be significant.

Number of tasks	Large (100,000)
Task Grain	Small (100,000)
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Dynamic Chunks (Chunk Size = 1000)

Table 5.1: Experiment 1 Parameters

5.2.2 Experiment 2 - Absolute Performance

The purpose of this experiment is to gauge the absolute performance of our system compared to other approaches. This should test if our approach has significantly more overhead than other approaches, and also if the underlying implementation is correct such that performance is on a par with competing approaches. To provide these points of comparison, we will test a purely sequential approach, and another using OpenMP. We will also vary the schedule, again to verify that our implementation is on a par with a current parallel programming method (OpenMP) with different schedules.

Note here that the Static schedule and the Dynamic Chunks schedule with a chunk size of 1 represent opposing ends of an overhead spectrum. With a static schedule, we have the least amount of overhead possible, (good for when the task variance is uniform), and with the Dynamic Chunks schedule and a chunk size of 1, we have the most amount of overhead possible, but since we fetch tasks one at a time, we will have the best task complexity balance across our threads (good for skewed/high variance.)

As for the expected outcome, I would predict that we would be close to a purely sequential implementation in single threaded circumstances, if we use a static schedule, as the only additional overhead would be in managing the one thread. For other schedules, we may see a significant performance drop, since we would be using a sub-optimal schedule. But, OpenMP would likely see a similar performance drop with a sub-optimal schedule.

The main question we want to answer is how we compare to a current parallel programming approach, OpenMP. Do we scale with thread/CPU core count in a similar fashion? I would expect that we may be behind OpenMP in terms of performance, as OpenMP is huge, well established project, and as such may be much more optimized than our implementation. Whilst the outcome of this experiment will no doubt be useful, we are mostly concerned with if we can improve upon our own performance by adding plasticity and contention aware scheduling, so we need only compare against ourselves in the final experiment.

Number of tasks	10,000
Task Grain	Medium
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	1, 2, 3, 4
Thread pinning	Uniform
Schedule	Static, Dynamic Chunks (Chunk Size = 1,000), Dynamic Chunks (Chunk Size = 1)

Table 5.2: Experiment 2 Parameters

5.2.3 Experiment 3 - Plasticity And Contention Aware Scheduling Framework Overhead

This experiment is designed to investigate whether the two main additions to our approach incur any significant overhead. That is, plasticity and contention aware scheduling. To do this, we will simulate the overhead of communicating with the controller (contention aware scheduling) and switching strategies (plasticity), without actually changing any details of the implementation, so that there will be no performance difference. We will also be running the controller application, which will tie up (a small amount of) system resources.

We will then run the same experiment with no controller or plasticity, as in experiment 1. The runtimes of both cases will be compared to see the total amount of delay introduced.

Whilst I expect that we will incur some overhead, I don't expect it to be significant, as it will be of constant time and not scale with regards to input size.

Number of tasks	100,000 10,000 1,000
Task Grain	Medium
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Dynamic Chunks

Table 5.3: Experiment 3 Parameters

5.2.4 Experiment 4 - Schedule Choice Importance

This experiment is designed to highlight the importance of the choice of schedule, and the benefits plasticity can bring even with just a single application. To do this, we compare the runtimes of different schedules with different task grain distributions. With a uniform task distribution, the static schedule should perform best, and the dynamic individual schedule should be the worst. With a biased task distribution, we should see the static schedule perform the worst, and the dynamic individual or dynamic chunks perform the best.

For the plasticity portion of this experiment, we will also add cases where we start with a sub-optimal schedule, and then, after 10 seconds, change to a better schedule. This will be done with a uniform schedule, starting with the dynamic individual schedule, and switching to the static schedule.

Number of tasks	50,000
Task Grain	Small, Large
Task Grain Distribution	Uniform, Biased
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Static, Dynamic Chunks (Chunk Size = 1,000), Dynamic Chunks (Chunk Size = 1)

Table 5.4: Experiment 4 Parameters

5.2.5 Experiment 5 - Absolute Multiprogramming Performance

The purpose of this experiment is to investigate the absolute performance of our library in a situation with multiple applications.

(*** Expand ***)

Number of tasks	50,000
Task Grain	Small
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	2, 4, 8
Thread pinning	Uniform, Random
Schedule	Static

Table 5.5: Experiment 5 Parameters

Chapter 6

Results

In this chapter of the report, we discuss the results of the experiments and their ramifications. The structure of this section mirrors that of the experiments section, so that the first results discussed will be the first experiment detailed in chapter 5.

6.1 Experiment Results

6.1.1 Experiment 1 - Metrics Collection Overhead

Runtime without metrics: 67862.0 Runtime with metrics: 68047.0

(*** Should I make this a graph? ***)

185ms of delay, 0.003% of total runtime.

Number of tasks	Large (100,000)
Task Grain	Small (100,000)
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Dynamic Chunks (Chunk Size = 1000)

Table 6.1: Experiment 1 Parameters

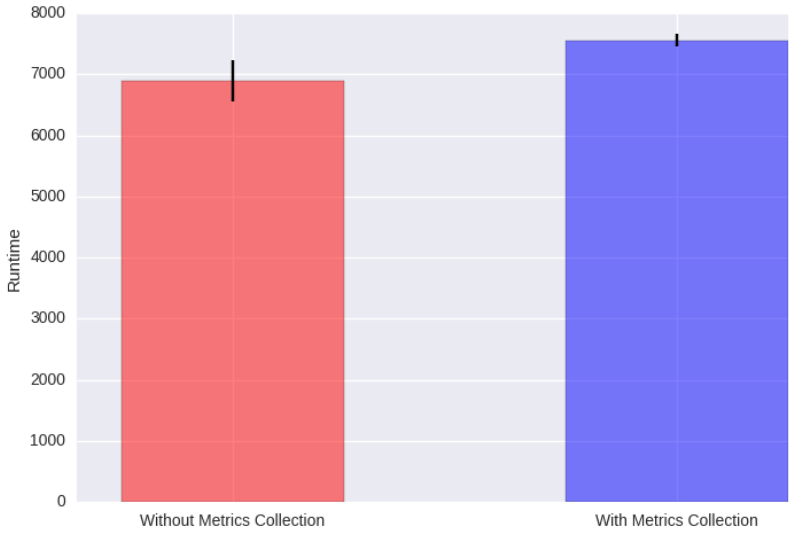

Figure 6.1: Experiment 1 results

6.1.2 Experiment 2 - Absolute Performance

In this experiment, the total run times of three implementations are measured. One sequential implementation, a standard modern parallel implementation utilizing OMP, and our plastic implementation. Our plastic implementation is running with no plasticity for the moment, and with no messaging functionality at all. This is so it is comparable to a standard parallel implementation.

The OMP and our implementation are using a dynamic chunks schedule, with a chunk size of 500. All programs were compiled at optimization level 3.

Fig. 6.2 shows us that with a single thread, our performance is similar to a sequential implementation, and as we increase the thread count, our performance scales accordingly. Overall, this shows that our baseline implementation performs on a par with current parallel implementations, providing a good baseline performance.

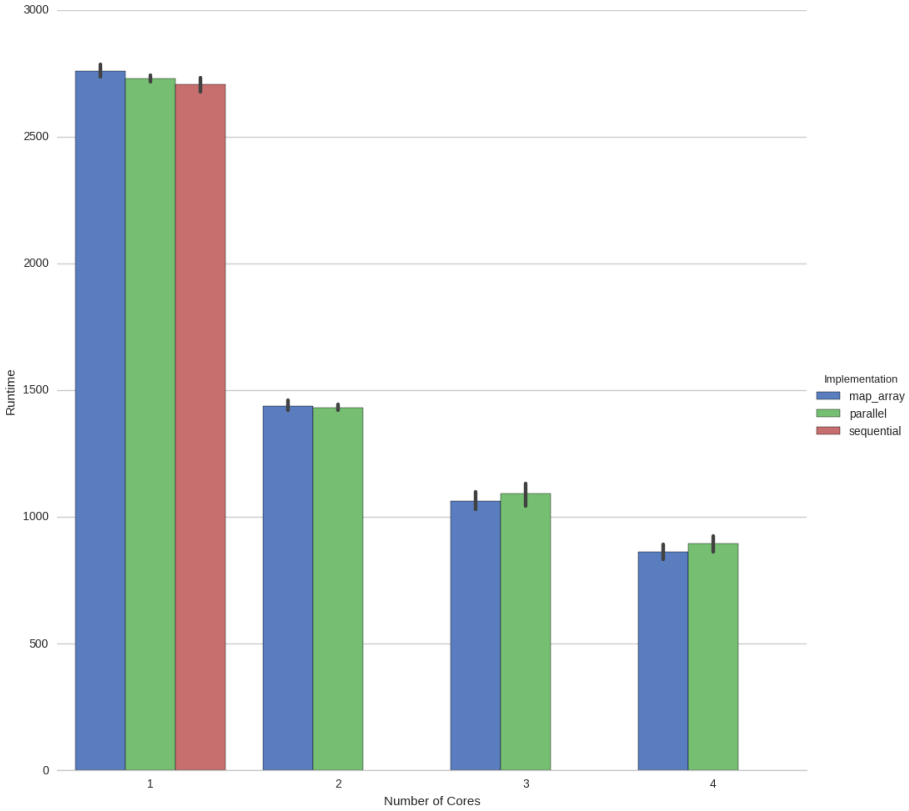

Figure 6.2: Experiment 2 results

Number of tasks	10,000
Task Grain	Medium
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	1, 2, 3, 4
Thread pinning	Uniform
Schedule	Static, Dynamic Chunks (Chunk Size = 1,000), Dynamic Chunks (Chunk Size = 1)

Table 6.2: Experiment 2 Parameters

6.1.3 Experiment 3 - Plasticity And Contention Aware Scheduling Framework Overhead

Results:

Runtime without plasticity etc: 24153.0 Runtime with plasticity etc: 24684.0

Standard deviation!

(*** Should I make this a graph? Yes - different input sizes***)

531ms of delay, 0.02% of total runtime.

Number of tasks	100,000 10,000 1,000
Task Grain	Medium
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Dynamic Chunks

Table 6.3: Experiment 3 Parameters

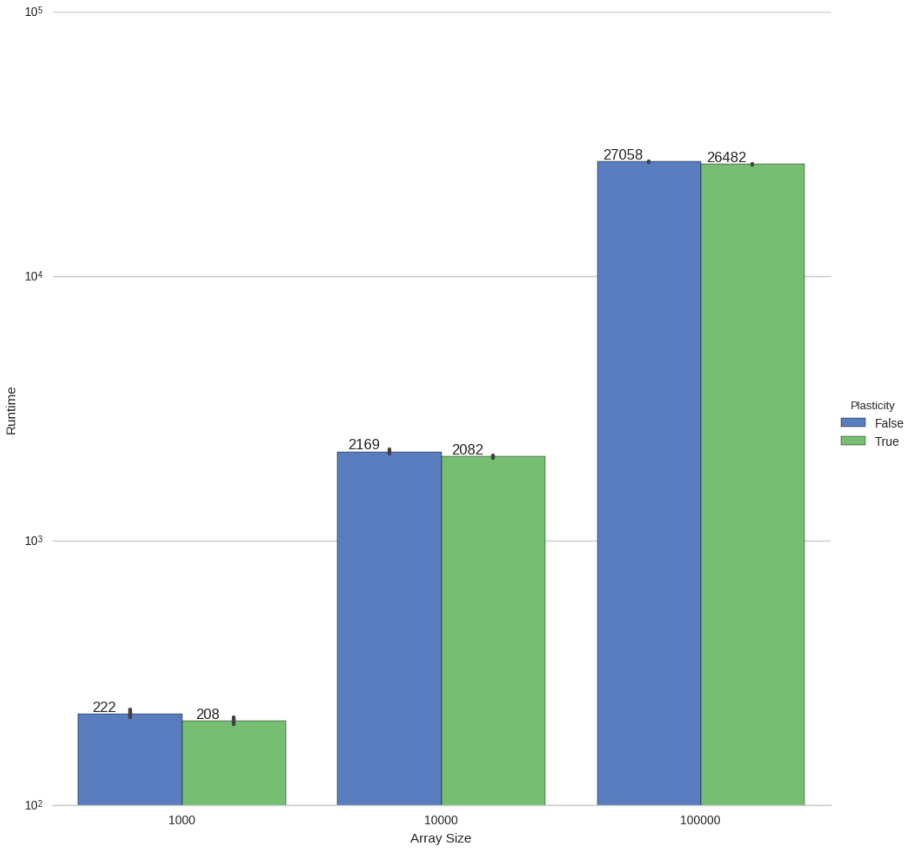

Figure 6.3: Experiment 3 results

6.1.4 Experiment 4 - Schedule Choice Importance

Number of tasks	50,000
Task Grain	Small, Large
Task Grain Distribution	Uniform, Biased
Number of CPU cores	4
Number of threads used	4
Thread pinning	Uniform
Schedule	Static, Dynamic Chunks (Chunk Size = 1,000), Dynamic Chunks (Chunk Size = 1)

Table 6.4: Experiment 4 Parameters

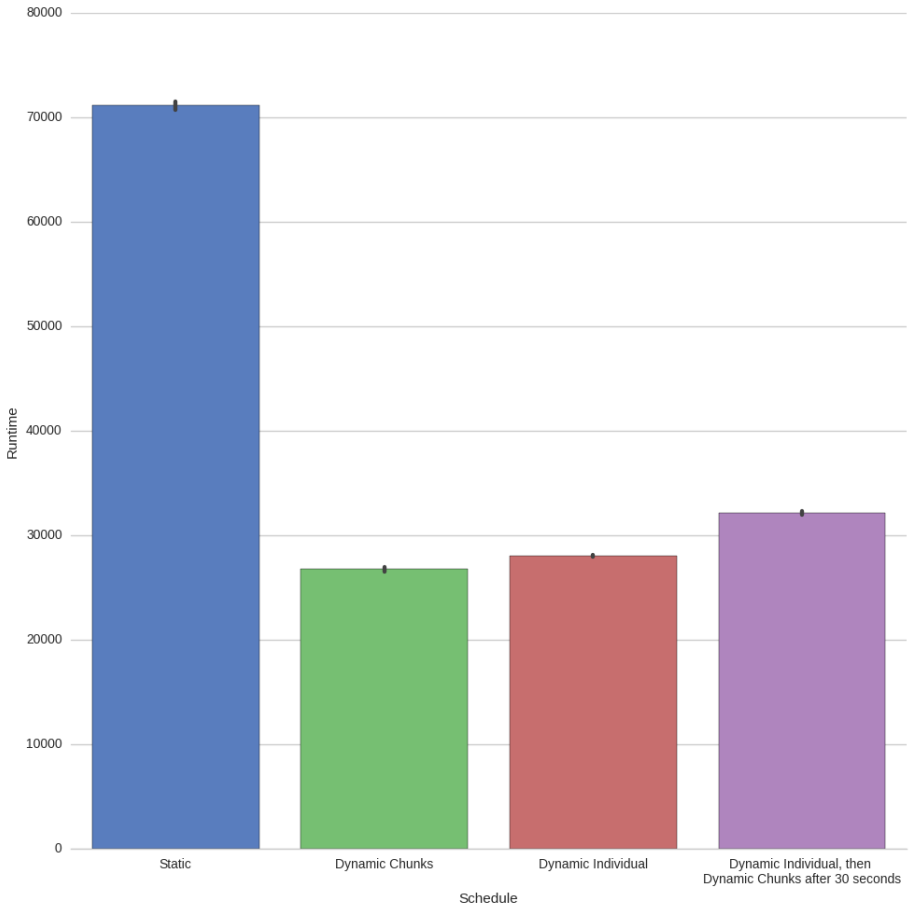

Figure 6.4: Experiment 4 results

6.1.5 Experiment 5 - Absolute Multiprogramming Performance

Number of tasks	50,000
Task Grain	Small
Task Grain Distribution	Uniform
Number of CPU cores	4
Number of threads used	2, 4, 8
Thread pinning	Uniform, Random
Schedule	Static

Table 6.5: Experiment 5 Parameters

6.2 Discussion

Discuss the findings of the results, (Mention weird runtimes with many small tasks!)

Chapter 7

Future Work

7.1

Overview of possible future work here, including work for next year and after that

Currently optimizing for performance, can optimize for other things e.g. energy efficiency for mobile applications etc.

Utilize GPUs, more exotic hardware.

Test on XXXII and other hardware.

further patterns left for possible future work. Producer consumer?

task stealing?

distributed system?

implementing some algorithm for automatic parameter tuning for future work.

Make work interruptable, e.g. if we assign static schedule we should still be able to change to other schedules

more experiments with varying memory access patterns

Vary Type of tasks (CPU Bottleneck/Memory Bottleneck)

(*** schedule choice importance - Should the same be done for thread pinnings? And we have already kind of seen importance of number of threads, but should

we have another experiment where num threads ¿ num CPU cores?)

7.2 Designing For Future Applications

A real-world version of our library would include multiple common patterns of parallel programming, and may even utilize multiple backends allowing for different features (e.g. Standard Pthreads, OpenCL/CUDA for multi-GPU computation). It's feasible that the system could assess both the tasks presented and the environment (e.g. the particulars of the machine), and automatically allocate the resources of the machine so we perform in the most efficient manner.

This system would be useful in any performance orientated application, even when the machine will only be running a single instance, as we can still optimize the implementation to the environment on that machine. It would, however, come into it's own when we have multiple instances running simultaneously on a machine, a common situation with modern multiprogramming machines.

Could be used on multiple nodes in a distributed system

Chapter 8

Conclusion

8.1

Provide overall conclusion, and discuss future work (next year)

Chapter 9

Bibliography

[1] Ansel, J., Chan, C., Wong, Y. L., Olszewski, M., Zhao, Q., Edelman, A., and Amarasinghe, S. Petabricks. *Proceedings of the 2009 ACM SIGPLAN conference on Programming language design and implementation - PLDI '09* (2009).

[2] Collins, A., Harris, T., Cole, M., and Fensch, C. Lira. *Proceedings of the 5th International Workshop on Runtime and Operating Systems for Supercomputers - ROSS '15* (2015).

[3] Meade, A., Buckley, J., and Collins, J. J. Challenges of evolving sequential to parallel code. *Proceedings of the 12th international workshop and the 7th annual ERCIM workshop on Principles on software evolution and software evolution - IWPSE-EVOL '11* (2011).

[4] Sur, S., Koop, M. J., and Panda, D. K. Mpi and communication—high-performance and scalable mpi over infiniband with reduced memory usage. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06* (2006).