Implementing high performance Synchronous Message Exchange University of Copenhagen

Truls Asheim <truls@asheim.dk>

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla laoreet lobortis erat, ac fringilla quam accumsan ut. Nam eget tortor neque. Nulla dictum dapibus venenatis. Vestibulum interdum sem vestibulum enim vulputate ullamcorper nec ut velit. Suspendisse rhoncus a nisi eget viverra. Curabitur porttitor hendrerit elit vel porta. Phasellus elementum nulla aliquet, rhoncus sapien sit amet, maximus libero. Morbi ipsum sapien, faucibus et tempor sed, commodo at enim. Curabitur maximus pretium felis, ac elementum nisi tincidunt et. Sed facilisis lacus purus, sit amet blandit elit ultricies sit amet. Maecenas et cursus libero, ac tincidunt arcu. In at tincidunt ante. Pellentesque lacinia dui nec nulla accumsan auctor faucibus elit suscipit. Pellentesque quis justo turpiso.

Praesent in dignissim lacus, in auctor felis. Phasellus iaculis finibus sapien, vitae vestibulum leo rutrum ac. Etiam in tellus eget sapien suscipit ornare eget sit amet lectus. Vestibulum sagittis consectetur varius. Aenean consectetur ante quis dui iaculis fringilla. Etiam enim sem, facilisis et interdum sit amet, congue nec metus. Fusce eu imperdiet enim, non sollicitudin sem. Vestibulum malesuada mattis justo, vel dictum massa tristique ac. Vivamus pretium turpis ut ligula porta, eget facilisis eros pulvinar. Aenean elit mi, semper vitae ornare vel, pretium sit amet ex. Cras commodo justo quis dolor gravida auctor. Vivamus ultrices arcu dolor, vitae porta ante viverra et. Quisque at aliquet massa, vel gravida felis. Donec eu imperdiet mi. Aliquam in tortor a libero iaculis luctus.

Contents

List of Figures

Chapter 1

Introduction

In this report, we describe the design and implementation of a highly efficient library for parallel execution of new, globally synchronous, message passing framework called Synchronous Message Exchange (SME). We will present a parallel, compiled framework, named C++SME, which can be used to implement and execute applications implementing the SME model.

The remainder of this chapter, will describe and define the SME model. The second chapter will describe the process behind designing C++SME. In the third chapter we will describe the implementation process and finally, we will show the benchmarks performed of our implementation.

1.1 Hardware Description Languages

A Hardware Description Language (HDL) is a programming language for describing hardware designs. A program written in such a language is usually

1.2 Background and Motivation

Field-programmable Gate Arrays (FPGA) provides several advantages over using GPGPUS for processing work including a significantly improved performance-per-watt ratio. Due to the low-level nature of current tools for programming FPGAs, their use are largely restricted to engineers with working knowledge in the field of hardware design. In order for software developers to take advantage of FPGAs, improved high-level hardware design utilities are required [1].

The creation of SME was motivated by attempts to use CSP for modeling hardware which, for simple cases, proved successful [4], but additional testing revealed that the CSP model introduced a significant overhead when simulating more complicated hardware designs. Modeling the clock-cycle driven global synchrony that exists in hardware proved to be particularly difficult with CSP and required the introduction of several additional processes and channels. This added complexity

reduced simulation performance and limited the usefulness of the CSP-based hardware design model [5].

These attempts were performed as a part of a master thesis which, despite acheiving their original goal found that the concurrency model of CSP was directly at odds with simulating hardware. Since CSP is based on the idea that any process can communicate at any time, the addition of a large number of additional channels and processes were needed in order to tame this behaviour. A central clock process were needed in order to propagate a clock signal to all processes, telling them when, and when not to, communicate. Furthermore, latch processes had to be inserted in order to dimulyr

SME is an attempt to provide a programming framework that, while leveraging and maintaining the properties of CSP that proved useful and f enforcing a hardware-like paradigm, that is accessible to software developers. By [5]

1.3 Synchronous Message Exchange

In this section, we elaborate on the description of SME from the previous section. We will describe the

We would like to start by defining the terms used henceforth in this report to avoid any ambiguities. Many of the terms used, have a specific meaning when used to describe computer hardware and this meaning will generally give a good idea of the meaning of a term in SME.

1.3.1 Definitions

Network A network is the highest level structure in the SME-model. It is simply a network of processes connected by buses

Cycle During a cycle, all processes has executed and all busses has propagated their values. A cycle goes through two distinct phases which we will refer to as the process execution phase and the value propagation phase. The process execution phase will activate all the execution units in the network to allow them to perform data calculation. The value propagation phase will transfer the output-values generated by processes in the current cycle, to serve as input-values in the next cycle.

1.3.2 Components

Compared to CSP, a much smaller and simpler set of components are used to model the process network. In this section, we describe those c

Process A process is an execution unit performing a unit of work. A process is defined by input and output busses used for communicating with other processes in the network and function which is called when the process is executed. The internal state of a process persistent between executions, but it's

execution cannot have any side-effects. Thus, the only way the execution of a process can alter the state of other processes is by bus communication.

Bus A bus enables communication between processes and should be considered analogous to buses found in actual hardware. A bus consists of a writing-slot and a reading-slot, both of which can hold a signed integer value. A bus in SME implements the CSP-equivalent of a one-to-all channel with a one message overwrite-buffer which means that only the final value written to the writing-slot will persist in the next cycle. The value of the reading-slot, on the other hand, can be read by all connected processes during a cycle. The value of the reading-slot is idempotent and is guaranteed to remain constant during the process execution phase of a cycle. During the value-propagation phase of a cycle the value of the writing-slot is copied to the reading-slot. From the point of view of the processes, the value-propagation phase is atomic, meaning that the values of all buses can be observed changing "at conce". If no value is written the writing-slot of a bus during a cycle, the value of its writing-slot will be `0` in the subsequent cycle.

1.3.3 Properties

The SME model has a number of special properties which must be maintained in order to ensure correct execution of the network. These properties also influences the design of our execution model

Property 1 (Implicit clock)**.** One defining feature of hardware is that all processing is driven by a clock beat. In order to enable fulfillment of this goal we introduce a simulated clock beat in our implementation of SME and thus the defining property of hardware is preserved in the SME model.

Property 2 (Global synchrony)**.** As a consequence of implementing the simulated clock best, all events and communications of the network occurs completely synchronous from the point of view of a process. FIXME

Property 3 (Shared Nothing)**.** A process is completely autonomous and can only change state through receiving a message on its incoming bus. A process is also self contained in the sense

1.4 Limitations

This report will not discuss details related to design of hardware

1.5 Related work

Discuss master thesis

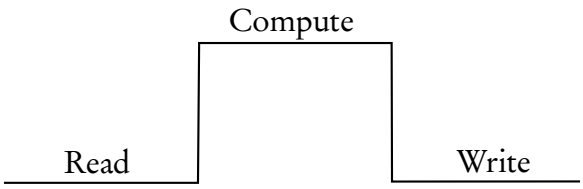

Figure 1.1: The execution cycle of a SME process visualized as a hardware clock-cycle. Before every cycle data from a input bus is read into the process and after a cycle, data is written back to the gate

Chapter 2

Analysis and Design

This chapter will describe, not only the final design, but also the considerations that went into the designs and ideas that was discarded during the process. n

2.1 Overall goal and success parameters

2.2 Paralellization model

A common way of parallelizing CSP-like networks is to use user-level threads to represent a process. In comparison with OS-level threads, user-level threads has a significantly lower overhead both with regards to context switching penalty and memory cost. Due to these limitations, implementing these kinds of message passing systems using only OS-level threads are generally not feasible. Therefore, user-level threads are used by other message passing systems such as the C++CSP2 library and the goroutines in the Go language. However, user-level threads are complicated to implement since we need to implement a scheduler which preempts running processes and schedules them on top of OS-level threads.

FiXme Note: CSP probably isn't preemtive

Comparing, once again, to CSP, the concurrency in CSP is inherently asynchronous while SME is entirely synchronous in nature. This means that a CSP library needs to implement a scheduler which decides when to give control to a process based on certain events, e.g. a process wishing to communicate or a process receiving a message from another process.

Due to the enforced synchrony of SME we don't have the same need to schedule processes "intelligently". We know that all processes need to run during a cycle and all busses have to propagate their values. By additionally taking into account he shared nothing property of SME, we can paralellize SME using a much simpler model compared to the aforementioned message passing network implementations:

The basic idea that we base our design on is conceptually similar to a classic producer-consumer setup. In our case, the work "produced" is the processes to be executed, and the consumers are the threads executing the processes. In this setup, a process is run simply by calling a function, whereas user-level threads are usually im-

plemented using the `setcontext` and `getcontext` library functions, which, while extremely fast, still causes a slightly larger overhead compared to a simple function call.

by simply letting a number of OS-threads run SME processes in a worker-consumer like manner. This approach also make it simple enforce the synchronicity properties of SME, since we know hen a process has finished executing. Unlike

In this project, we have explored two different variations of this basic idea. Both models are based on the idea described in the previous paragraph. Our overall goal in parallelizing execution of SME-networks is to minimize the amount of core idle time. We expect the hereafter presented model to achieve this goal under different circumstances .

FiXme Note: different networks

2.2.1 Pros and cons

Pros: Shared nothing,forced synchrony

Cons: Forced syncrhony

2.2.2 "Worker quwue" model

This approach is similar to a classic producer-consumer model where we have a number of workers which takes tasks off a circular queue and executes the processes. The main advantage of this model is that it allows processes with different execution times to "interleave" leading to a higher overall execution time. The primary problem of this model is that we need to make the queue thread-safe. The locking mechanism needed to do this isn't free, and could therefore come at a significant cost when we're executing a network consisting of small processes.

2.2.3 Static orchestration model

In this model, we assign separate queues to each thread of execution and distribute the processes amongst them. Due to the properties of SME, this distribution of processes only needs to happen once, before we start network execution. The main advantage of this model is that eliminates any shared state in our network, and therefore we don't need to consider the thread-safety of our queues. This reduces the fixed cost of executing a process significantly. This model, however, is more sensitive to uneven distributions in process workloads. For instance, if we end up assigning predominantly small processes to one core and large processes to another, the core executing the small processes would be left idle until the other core has finished executing it's part of the cycle.

Overall, we expect the latter model to have a significant advantage in executing networks with small processes while the queue-locking cost of the former model will perform better when executing networks with large or unevenly distributed workloads since the queue-locking costs will be amortized and allow its process-interleaving ability to shine through.

The optimal way of showing

2.2.4 Identifying optimal process scheduling

In order to determine the efficiency of various methods of process scheduling we need to identify the optimality condition for our process scheduling. An illustration of our threading model can be seen in figure 2.1. The green boxes represents processes while the red boxes indicates core idle time. Notice, how we by redistributing the processes across threads could reduce the idle-time of our cores.

2.3 Synchronizing cycles

The main problem of executing an SME network is to make sure that the cycles are synchronized, or that all the processes "meet up" at the end of an execution cycle. While the problem of executing the actual processes, due to the shared-nothing property of SME, is embarrassingly parallel, the need for synchronization makes it not quite so. Therefore, the time spent on synchronization will significantly impact the overall performance of our network.

1. In order to know when to stop executing, we need to count the number of cycles performed

2. In order to know when a cycle is complete, we need to keep track of the number of processes that has been executed.

2.3.1 Cost of synchronization

The cost of synchronization arises from two areas

There are two places in the network execution where this "accounting" could be preformed. We could either place a "guard" around each ... An alternative way of performing synchronization is to insert special p

"Naive" way of doing would be to let each execution thread count the number of processes that has been executed. The problem with this approach is that it adds a fixed computational cost to each process execution. This would become especially pronounced in model 1 which would require

Furthermore, we would need to keep the state of the network as a global shared state which would need to be protected by locks when accessed by a thread. Both of these factors would significantly limit the concurrent scalability of the parallel execution.

One of the central parts of managing an execution cycle is how we synchronize our threads before leaving each cycle phase . In order to maintain the previously described synchrony property, ... Furthermore, the network execution must be controlled so that we are able to stop the execution after a specified number of cycles has completed.

FiXme Note: elaborate

An alternative method, which allows the

2.4 Implementing the queues

An actual circular linked list where the last element points to the first would be the most natural representation of the conceptual circular queue that we just described. The usual advantage of using a linked-list structure is that it allows for O(1) addition of elements. The disadvantage if that element access is slower, even though we would never need to actually traverse the list in order to find a specific element, the cost associated with simply getting the next element of linked list is not insignificant when performed enough times

FiXme Note: crap

A straight-forward array is much better suited for the task z7

2.4.1 Locking primitices

Classic locking mechanisms such as semaphores and mutexes needs no introduction. We will, however, spend a little bit of time on explaining the new kid on the block – atomic operations. Atomic operations

2.4.2 Process orchestration

As we discussed in the previous section, the primary limiting factor for our multi-threaded network is an uneven and suboptimal distribution of processes across CPU-cores. If no attempt is made to optimize process distribution, the order of process execution will depend on the order of which processes are defined in the source code. Due to property 3 and property 2 of SME networks there is no scenario where it would be necessary or beneficial for a programmer to exercise ultimate control over the order of process execution. Therefore, maximizing CPU-core utilization would be an unreasonable burden to put on the programmer, especially since their optimization efforts would be specific to a certain number of CPU-cores.

The optimal method and timing of process orchestration depends on the dynamicity of the work performed by the network we are executing. A network where each process performs a fixed amount of work per iteration will only need to be orchestrated once, while a network where the workload of the processes are variable will need to be continuously evaluated at runtime in order to maintain our optimality condition. These various methods will be discussed for the remainder of this section.

FiXme Note: is predictability a better word?

2.4.3 Round-robin orchestration

Processes will be executed on the first available core as seen in figure 2.2

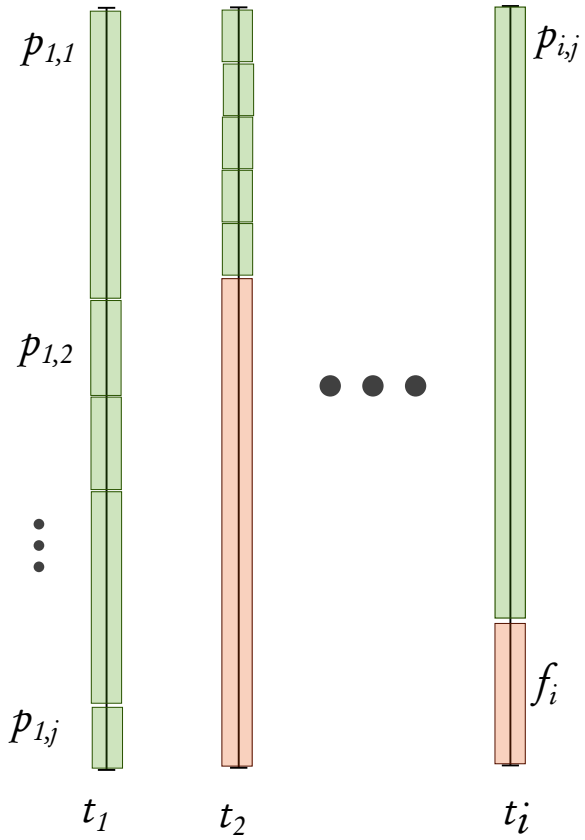

Figure 2.1: Example of suboptimal distribution of processes across processing threads. Green blocks represents processes while red blocks represents thread idle time. Threads are named $p_{i,j}$ where i is the number of the thread the process has been assigned to and f_i is the combined idle time for each thread. Threads are named t_i.

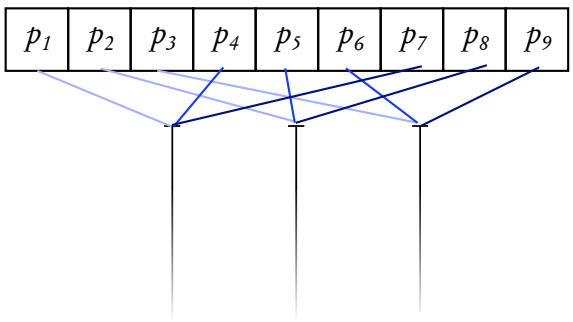

Figure 2.2: Illustration of round-robin process orchestration. Progressive iterations are shown as
increasing color intensity of arrows

Chapter 3

Implementation

We have chosen to implement the SME library in the C++ language. C++ combines the availability of high-level structures, such as classes, with the ability to, when needed, assert low-level control over the code generated. Furthermore, the C++11 revision of the language allows for easy access to features that were previously hard to use. Such feature include the <atomics> header which enables the use of atomic instructions and enforced memory ordering without the need for inline assembly and similar. Having access to atomics is a highly desire able feature for us since they can be used as high-performance synchronization primitives. Furthermore, classes in C++ are well suited for representing SME constructs and specifically, they provide a natural enclosure of the state maintained by a SME process.

The library is meant to be imported by applications wishing to take advantage of the SME-model.

The implementation was performed in several phases. The initial version of the C++SME code was purely single-threaded and was implemented to play around with the C++ API's and for defineing the API used to define SME networks.

Adding support for multi-threading required a lot of the code from the initial single-threaded implementation to be refactored and rewritten.

Since the networks that we benchmark are large enough that it would be tedious to write them by hand, features were also added mainly for the purpose of supporting benchmarking. ur initial approach to benchmarking used python script for generating the benchmark networks, however, this method quickly proved to be in infeasible since GCC's compilation time increases, seamingly, exponentially with the amount of objects defined in the code being compiled. We therefore had to enable the SME-library to support runtime definition of networks. Mind you, that networks are still statically defined in the sense that the orchestration of processes must be performed before the start of network execution. Networks that change at runtime is beyond the scope of SME since it simply isn't possible in hardware, which SME is intended to map FiXme Note: rewrite

Since

We want the API to be as seamless as possible, that is, it should get int he way of

the programmer as little as possible. Several phases of refinement led to the current API which reduces the amount of boilerplate code required significantly compared to the original version

3.1 Initial implementation

Our initial implementation was a sequential implementation of the SME execution environment. This implementation was done simply, as a proof of concept and to experiment with different API's for defining SME networks.

3.2 Queue implementation

How we performed process orchestration and, in particular, the workqueue mechanism got a lot of attention in the previous chapter. In this section, we will how we made the actual implementations of the work queues

The locking mechanisms used in sought

3.2.1 Locking mechanisms and atomics

Atomics, and particularlily lockless algorithms have recently been made availible for "casual" use by programmer following their inclusion in recent language revisions.

https://www.arangodb.com/2015/02/comparing-atomic-mutex-rwlocks/

3.3 Design goals

The library takes advantage of the fact that the initial process orchestration is only executed once and thus can be implemented with focus on code clarity rather than performance. This

3.4 Public API

Boilerplate code is a major annoyance

3.5 Testing

In order to check that our network works as intended and is able to execute a SME network without violating any of the properties of the SME model, we have implemented a special *validator network* which is design to reveal any inconsistencies arising from incohesion with the invariants of the SME model.

FiXme Note: is this a word?

Specifically, the validator network intends to check if all processes has been executed during a cycle and b) if all values

Chapter 4

Benchmarks and Discussion

We will present a number of benchmarks designed to compare and quantify the differences in performance of the parallelization models that we have implemented.

Since the execution time is only dependent on the total amount of work that a network performs and not how the processes in the network are connected, all of our benchmarks will use a ring-shaped (figure 4.1) network with the participating processes performing varying amounts of work.

We conjecture that the scalability of our implementation will depend strongly on the nature of the workload performed by the SME-networks benchmarked. We will therefore benchmark both light and heavy

As our previously presented hypotheses states, we expect our benchmarks to show that the effects of syncing becomes more pronounced as we decrease the amount of work performed by our processes, while it will become amortized as the amount of work performed by each process increases.

4.1 Testing methodology

All of the time measurements presented here were performed within the SME framework itself and measures only the actual execution time of the network. It therefore does not include the constant time required to generate the benchmarked networks. Two different hardware platforms has been used for performing the benchmarks: One AMD and one Intel platform.

Since the instruction set used by the two CPU's support incompatible optimizationns, code generated for one will not run unmodified on the other. Therefore, code executed on the AMD CPU were compiled with the GCC flags `-mtune=barcelona -march=barcelona`, while code executed on the Intel CPU were compiled with `-march=native` on a Core i7 machine. GCC 4.9 was used in both cases. Furthermore, due to incompatible versions of libstdc++ on the test machines, all benchmarks has been performed using statically linked binaries.

All of the benchmarks has been executed 5 times and the graphs are based on the averages of these. The benchmark results have been stable between runs with

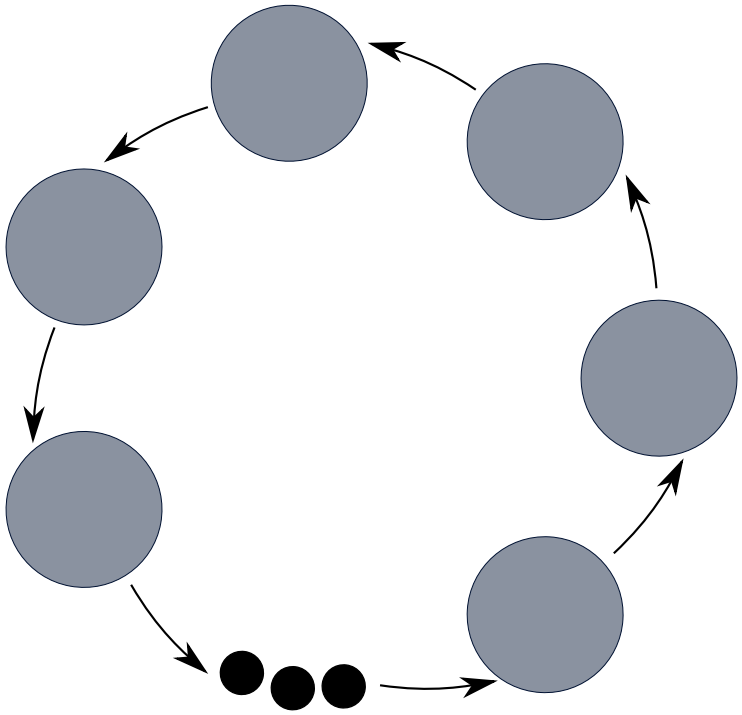

Figure 4.1: Illustration showing the layout of the network used for benchmarking. The blue circles represents processes and the arrows represents busses

error ranges between $3 \cdot 10^{-2}$ and $7 \cdot 10^{-3}$. Since these values are small enough to be insignificant, we haven't made any effort to plot them in our graphs, especially since they are too small to create visible error bars on included graphs.

We calculate our speedup using the formula

$$S = \frac{T_{\text{old}}}{T_{\text{new}}}$$

where S is the achieved speedup, T_{old} is the original (pre-improvement) speed and T_{new} is the new (post-improvement) speed [3].

4.2 Synchronization dominated

In this section, we present a benchmark, where the performance is predominantly determined by the efficiency of the synchronization mechanisms.

We perform this benchmark, by creating a ring which does nothing other than passing an integer value from process to process. Sine each process only takes a few clock cycles we expect that this benchmark will reveal the overhead

urthermore, since we actually e

The following source code used in the execution unit of the process

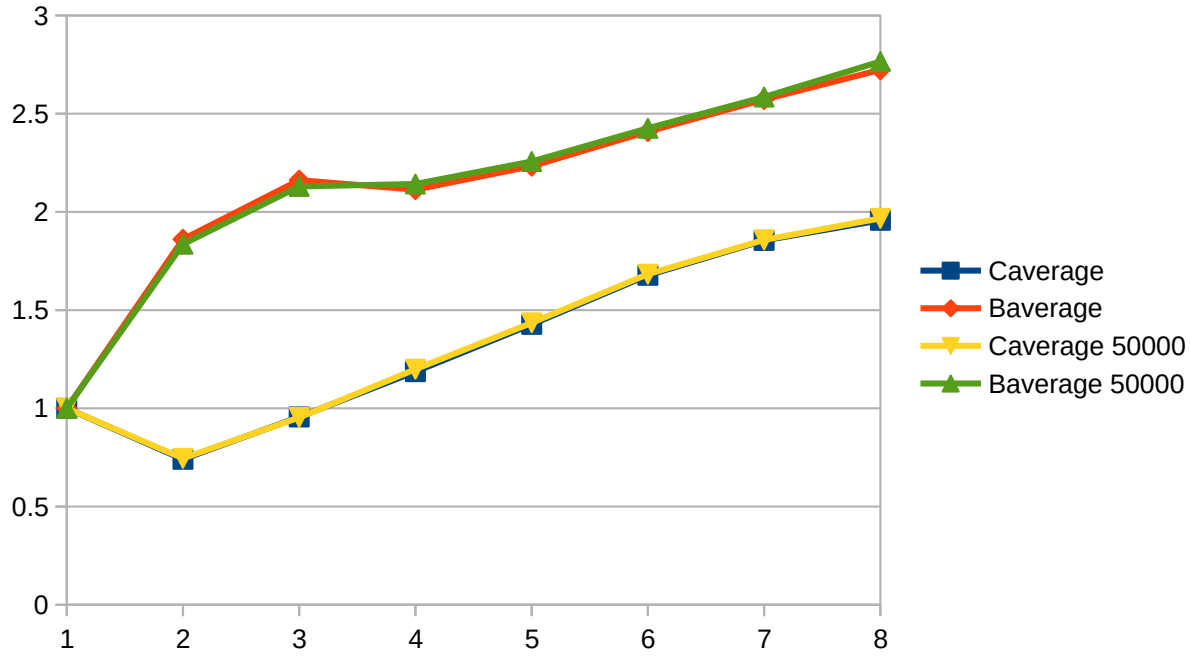

Figure 4.2: Benchmark graph

```
void step() {
  int val = in->read();
  out->write(++val);
}
```

4.2.1 Discussion

We can observe a number of things from the results that can bee seen in figure 4.2

One thing that is clearly visible from this benchmark is the overhead produced by the atomic increment that is required.. This model is doubly penalized when running the benchmark since we, addition to then time required by the atomic increment, also need to wait for all of the threads to sync up at the end of a cycle. What is slightly surprising, however, is the actual performance that this method shows. It performs significantly worse when going from one to two threads. The most likely explanation for this is that the CPU must somehow make an optimization to make atomic updates less costly when the operation only occurs from one thread.

We can see that our

Our model 2 performs, quite decently and produces almost 2x speedup when going from 1 to 2 threads.

Another interesting observation is that Hyper Threading seems to give a significant additional speedup. One hypothesis as to why this is the case, could be

that branch-prediction isn't very effective at predicting which functions we're going to call in our SME network. A branch mis-prediction causes the CPU-pipeline to be cleared, creating an optimal condition for Hyper Threading to make use of the empty pipeline-stages[2] While branch-predictors This hypothesis could be tested by running the program through a profiler in order to measure the number of mis-predictions occurring. At this time, these results are not available.

4.3 Cycle domintaed

In this benchmark, the processes in the network performs a significant amount work. We expect that this will, to some extent, amortize the synchronization overhead inherent in the SME model. Combined with the fact that the individual processes contain no shared state, we conjecture that this benchmark will scale significantly stronger than the previous synchronization dominated benchmark that we performed.

The unit of work being performed by every process in every cycle is simply to divide a `long double` number by 3 10000 times. Since the busses in our SME-implementation only supports transporting integer values nothing is being done value calculated, but as long as our workload isn't being optimized away at compilation time this is irrelevant.

FiXme Note: Yes... good question, Why exactly do we do that?

We use floating point numbers as opposed to integer

The following code is used as workload in our processes

```
private:
  long double n;
  int i;
protected:
void step() {
  n = 533.63556434;
  for (i = 0; i < 10000; i++) {
    n = n/3;
  }
  int val = in->read();
   out->write(++val);
}
```

Listing 1: Code used for generating work in the cycle-dominated benchmarks

4.4 Discussion

A significant problem with this benchmark is that the work that we perform is entirely cache-local to a CPU-core. This allows us to scale more strongly than when

benchmarking a problem which to a larger extent is limited by memory bandwidth and/or CPU-cache misses.

4.5 Future works

More benchmarks:

The results that we have shown, although reasonable, can not be easily explained by

4.5.1 One-shot process orchestration

In this model, we orchestrate the processes in our network as soon as possible after execution start and

4.5.2 Monte Carlo orchestration

In this approach, we simply randomize the order of the processes. The main advantage of this approach is that is computationally cheap compared to

4.5.3 Optimization-based orchestration

Another way to orchestrate the processes is to use a

4.5.4 Adaptive process orchestration

The benefits of using a oneshot orchestration approach diminishes when we execute process networks where the processes performs a variable amount of work per iteration. In these kinds of networks, CPU-core load distribution will gradually become uneven and suboptimal as the network execution progresses. In order to keep this from happening and maximize CPU-core utilization, we need to monitor process execution time and core idle time as the network execution progresses. This is what we refer to s adaptive orchestration. This approach, however introduces another trade-off that we need to consider. producing an

4.5.5 Adaptive Monte Carlo process orchestration

4.5.6 Adaptive Optimization-based process orchestration

Bibliography

[1] David F Bacon, Rodric Rabbah, and Sunil Shukla. "FPGA Programming for the Masses". In: *Communications of the ACM* 56.4 (2013), pp. 56–63 (cit. on p. 5).

[2] Agner Fog. *The microarchitecture of Intel, AMD and VIA CPUs/An optimization guide for assembly programmers and compiler makers*. 2014 (cit. on p. 20).

[3] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2012, pp. 46–47. ISBN: 978-0-12-383872-8 (cit. on p. 18).

[4] Martin REHR, Kenneth SKOVHEDE, and Brian VINTER. "BPU Simulator". In: *Sommunicateind Process Architectures 2013* (2013) (cit. on p. 5).

[5] Brian Vinter and Kenneth Skovhede. "Synchronous Message Exchange for Hardware Designs". In: *Communicating Process Architectures 2014* (2014) (cit. on p. 6).

Appendix A

Compiling and executing

This section will contain information about how to compile and execute cppsme