**Process Cooperativity as a Feedback Metric**

**in Concurrent Message-Passing Languages**

APPROVED BY

SUPERVISING COMMITTEE:

Dr. Matthew Fluet, Supervisor

Dr. James Heliotis, Reader

Dr. Rajendra K. Raj, Observer

# Process Cooperativity as a Feedback Metric in Concurrent Message-Passing Languages

**by**

**Alexander Dean, B.S.**

**THESIS**

Presented to the Faculty of the Golisano College of Computer and Information Sciences

Rochester Institute of Technology

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Science**

**Rochester Institute of Technology**

August 2014

**Abstract**

**Process Cooperativity as a Feedback Metric**

**in Concurrent Message-Passing Languages**

Alexander Dean, M.S.

Rochester Institute of Technology, 2014

Supervisor: Dr. Matthew Fluet

Runtime systems for concurrent languages have begun to utilize feedback mechanisms to influence their scheduling behavior as the application proceeds. These feedback mechanisms rely on metrics by which to grade any alterations made to the schedule of the multi-threaded application. As the application's phase shifts, the feedback mechanism is tasked with modifying the scheduler to reduce its overhead and increase the application's efficiency.

Cooperativity is a novel possible metric by which to grade a system. In biochemistry the term cooperativity is defined as the increase or decrease in the rate of interaction between a reactant and a protein as the reactant concentration increases. This definition translates well as an information theoretic definition as: the increase or decrease in the rate of interaction between a process and a communication method as the number of processes increase.

This work proposes several unique takes on feedback mechanisms and scheduling algorithms which take advantage of cooperative behavior. It further compares these algorithms to other common mechanisms via a custom extensible runtime system developed to support swappable scheduling mechanisms. A minimalistic language with interesting characteristics, which lend themselves to easier statistical metric accumulation and simulated application implementation, is also introduced.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Runtime systems can be broken up into multiple distinct parts: the garbage collector, dynamic type-checker, resource allocator, and much more. One sub-system of a language's run-time is the task-scheduler. The scheduler is responsible for order of task evaluation and the distribution of these tasks across the available processing units.

Tasks are typically spawned when there is a chance for parallelism, either explicitly through `spawn` or `fork` commands or implicitly through calls to parallel built-in functions like `pmap`. In either case it is assumed that the job of a task is to perform some action concurrent to the parent task because it would be quicker if given the chance to be parallel.

It is up to the scheduler of these tasks to try and optimize for where there is opportunity for parallelism. However, it's not as simple as evenly distributing the tasks over the set of processing units. Sometimes, these tasks need particular resources which other tasks are currently using, or perhaps some tasks are waiting for user input and don't have anything to do. Still worse, some tasks may be trying to work together to complete an objective, and rely on dynamic dependencies that change over time.

Tasks however, in functional language verbiage, are typically called *processes* due to the inherent isolation this term brings and the language paradigm calls for. Instead, message passing is a common alternative to, and sometimes abstraction of, shared memory. Message passing is akin to emailing a colleague a question. You operate asynchronously, and your colleague can check her mailbox and then respond at her leisure. Meanwhile you are free to operate on an assumption until proven wrong, wait until she gets back to you, or even ask someone else.

While message passing is a good method for inter-process communication, it is also a nice mechanism for catching when two processes are working together. For example, consider a purely functional `pmap`, where all workers are given subsections of the list. Each worker thread will have no need to access another's subsection and thus no messages will need to be passed. However, what in the event the function being mapped on a particular subsection uses several processes? Each may access a shared resource via message

passing. We would see a close coupling in this case. This highlights the granularity of process coupling, in that the `pmap` workers exhibit course-grained coupling, which allows the scheduler greater flexibility to run them in parallel. The opposite is true for the processes which show close coupling, like the mapped function.

There exists a large number of mechanisms that scheduling systems can use in an attempt to improve work-load across all processing units. Some of these mechanisms use what's called a feedback system. Namely, they observe the running behaviour of the application as a whole, (i.e. collect *metrics*), and modify themselves to improve operation. We define the granularity of process coupling as **Process Cooperativity**.

Process Cooperativity is an interesting metric by which to grade a system. In biochemistry the term cooperativity is defined as an increase or decrease in the rate of interaction between a reactant and a protein as the reactant concentration increases. We can translate this into an information theoretic definition:

**Definition 1.** *The* **degree of cooperativity** *of a system is the increase or decrease in the rate of interaction between processes and an inter-process communication method as the concentration of processes fluctuate.*

Thus, when a process attempts to pass a message to another we know it's trying to cooperate on some level. When this frequency of interaction is high, it may indicate a tight coupling of processes or fine-grained parallelism. If it is low, this could indicate course-grained parallelism. In either event, a scheduler able to recognize these clusters of cooperative and non-cooperative processes should have an edge over those that don't.

Chapter 2 will look first at the background of classical scheduling systems as well as the recent feedback-enabled approaches. Then, we will also examine the types of message passing implementations and how these effect scheduling decisions, now that we are looking at process cooperativity. Chapter 3 introduces our work on a language and compiler, built to easily simulate system cooperativity and visualize the effects of scheduling mechanisms on these systems. We also discuss a few example mechanics which take advantage of cooperativity. Some example applications which demonstrate different degrees of cooperativity and phase changes are also explained. In Chapter 4 we run our cooperativity-enabled schedulers along with a few common non-feedback-enabled schedulers on the example applications and discuss the results. Finally, in Chapter 5 we give some concluding remarks and avenues of future research we believe would be fruitful.

# Chapter 2

# Background

## 2.1 A Note on Control Theory

Since the formalization of feedback driven systems and the advent of Cybernetics, multiple fields have attempted to mold these principles to their own models; and run-time schedulers are no exception. This is due, in part, from process scheduling in parallel systems being fundamentally an NP-Complete problem [1].

Note that the base case of runtime scheduling is called the Multiprocessor Scheduling Problem which is used in job-batch scheduling, and states:

**Definition 2. MULTIPROCESSOR SCHEDULING PROBLEM**
*Given a set of jobs* $\mathscr{J} = (J_1, J_2, \ldots, J_n)$*, a directed acyclic graph (lattice)* $L = (\mathscr{J}, C)$ *(indicating job dependence, and thus precedence constraints), an integer* $P$ *(the number of processors) and an integer* $D$ *(the deadline), is there a function* $S$ *(the schedule) mapping* $\mathscr{J} \to \{1, 2, \ldots D\}$ *such that:*

1. *For all* $t \leq D, |\{J_i : S(J_i) = t\}| \leq P$.
   *(*i.e. *The quantity of processors is greater than the quantity of jobs per time slice.)*
2. *If* $(J_i, J_j) \in C$*, then* $S(J_i) < S(J_j)$.
   *(*i.e. *Jobs cannot be scheduled before their dependencies.)*

In runtime scheduling however, the deadline, $D$, is incremented for each timeslice we pass. As such, it is possible for $|\mathscr{J}|$ and thus $C$ to fluctuate causing a need to re-find $S$. The continuous nature of this problem complicates the scheduling problem substantially. Instead, focus has been more fruitful when pursuing the optimization of various measurements using some particular objective function [2] to tune for particular edge cases. As such, scheduling based on such feedback metrics is not a new practice [3].

There is a big distinction though, which can be made between the effects of control theory in classical cybernetic applications versus that of run-time systems. This is primarily in the adaptation of the controller in the generic feedback loop (figure 2.1).

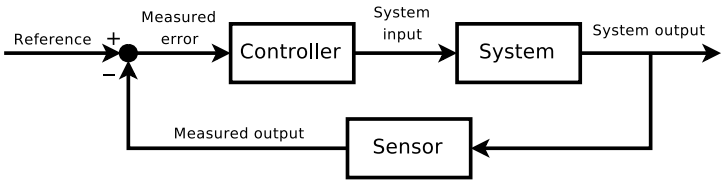



Figure 2.1: A classical feedback loop representation.

The classic feedback loop starts by making reading the current state of the system and applying some operation to it (via the controller). The operation in some way affects the system which can be observed by measuring particular metrics. These measurements can be fed back to the controller along with details about the system's output. If the controller wishes to modify the system further via the same or an opposite operation, it may do so. The canonical example is that of an automobile's cruise control. The controller can correct the speed of the vehicle by applying or releasing the throttle based on readings of the current speed.

In typical physical feedback loops there are two scenarios which need to be avoided: resonance and rapid compensation. Resonance in physical systems is when a spike in the amplitude of a system's oscillation can cause it to fail at a particular frequency. It can be seen that most controller models will attempt to damp the adjustments to reduce oscillation which could cause resonance or sharp spikes in behavior based on its output. This is due to the limitations of the physical space in which they are having to work. But frequent or extreme damping or can stress physical systems to the point of failure as well.

However, in run-time scheduling systems we would very much like to do the opposite. We would prefer tight oscillations or consistent behavior of our runtime so as to achieve minimal overhead from our modifications. We can also compensate, to reach our reference signal, as quickly as we need to as there are no physical restrictions for our modifications. As such these feedback systems are closely coupled with the design of the scheduling algorithm, rather than being an interchangeable sensor, and controller modules. As such we make an effort to trace the feedback optimizations during our evaluation and explanation of the scheduler designs.

Another distinction must be made as far as the level of foresight the scheduling systems have, at least, within this paper. There is a spectrum of clairvoyance in classical job-scheduling, in that on one end, job-schedulers have full foresight over the jobs which will enter the queue and their order (*i.e.* the full $\mathscr{J}$ set will always be known). These

schedulers have the opportunity to optimize for future events (by constructing a valid lattice $L$ based on the current time $t \leq D$), which is a luxury the scheduling systems that this paper discusses do not have.

However, as it is a spectrum, there is a single point of knowledge this subrange of schedulers can assume. Namely, that the first job will always be the last, and all other jobs will spawn from it. Thus there will always be only a single process in the queue at the beginning. This is true as the runtime will always require an initial primary process (*e.g.* the 'main' function), and once that function is completed, the system is terminated (despite the cases of unjoined children). Apart from this, all other insights will need to be gleaned from the evaluation of this initial process.

## 2.2 Message-Passing

In concurrent systems, there are a number of methods for inter-process communication. Arguably though, one of the more popular abstractions is the idea of message passing. This is especially true in functional languages as the language assumes shared-nothing by default. Also, just as compilers can optimize using language constraints, so can a run-time using the language implementation. We will therefore examine possible message passing designs and how their implementation might effect our schedulers.

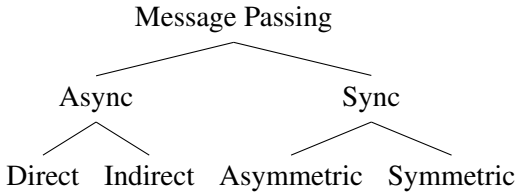



Figure 2.2: A High-Level Message-Passing Taxonomy

Message passing in general can be broken down into two types based on the language's implementation; asynchronous or synchronous. In asynchronous message passing, a process can send the message directly to another process like in our example of emailing a coworker; these implementations are aptly named mailbox message passing. Another implementation sends the message indirectly via a rendezvous point, like a socket.

To send a message in either case requires pushing/copying the message into a shared or global memory space for another process to access (possibly) at a later time. This

push/copy can be done in a lock free manner with some lower level atomic data structures such as a double-ended queue. But in either a locked or lock-free manner, the process performing the send still forces a block somewhere along in the operation, to push the message into some shared storage. As asynchronous operations require the additional capabilities to both store and resend a message at a later time.

In terms of scheduling, a language with asynchronous message passing does not hint much in regard to whether progress is being made. If a consumer process requires a value before continuing and therefore is repeatedly trying to receive from the channel, the schedule for the system would be better served by coming back to that process at a later time rather than repeatedly looping. However, asynchronous does benefit from process placement so as to take advantage of possible gains in cache affinity [4]. For example, the effects of the cache on direct message passing (*e.g.* a process mailbox) can be substantial as two processes on different cores share a location to store and thus check for content. This shared location if accessed from two processes will have to be updated in possibly multiple locations and validated for consistency at the cache level. However, if the two processes are local to the same cache there will be time saved in context switching and cache-line validation. In indirect message passing the task is even worse as more than two processes may need access to the same space.

In synchronous message passing, a process must meet another at a provided rendezvous point but can either be symmetrical or asymmetrical. Note that the rendezvous point is not a requirement in the sense that direct synchronous messaging isn't possible. Instead we think of a rendezvous point in synchronous communication to be time bound rather than location bound (*i.e.* two processes are blocked until communication occurs, the implementation of this passing is irrelevant to this classification).

Asymmetrical message passing is synonymous with Milner's Calculus of Communicating Systems [5] or standard $\Pi$-Calculus [6], in that you have a sender, and then a receiver which will both block on their respective functions around an anonymous channel until the pass has been completed. This differs from symmetrical message passing in that the only operation on the channel is a blocking function which swaps values with the process at the other end.

It's worth noting that asynchronous message-passing can be simulated using synchronous channels with a secondary buffer process. But by simulating it in this fashion we, as the scheduler, elevate the problem of cache locality to a problem of process locality. The same methods suggested to alleviate some of the lost efficiency due to cache locality [7,

8] are the same techniques which could be simulated for process locality; namely process batching and process affinity.

Note also, it is possible to simulate symmetrical message passing on asymmetrical message channels, but in terms of scheduling of synchronizing processes, order is now a factor that needs to be considered. On top of this, directionality can also be a factor which complicates the channel implementation. Namely, the internal queuing of senders or receivers may not percolate hints up to the scheduler regarding their queue position.

For the alternative, symmetrical message passing or swap channels, the order is directly handled by the scheduling of the system (*i.e.* the order at which the channels evaluate the $swap$ command can be directly governed). And it is for this purpose along with simplifying our core language we have chosen to base our semantics on symmetric synchronous message-passing.

## 2.3 Classic Runtime Scheduling

Operating Systems research have long been a leading front for scheduling topics. However, most of the early concern in scheduling was devoted to job scheduling over a group/shared system. As such, their concerns were largely devoted to fairness and job priority. Choosing processes based on job-length are also not a possibility due to a lack of *a priori* knowledge. Yet there are a few topics of scheduling in general, which lend themselves to run-times too, such as how to distribute a set of processes across processing units.

There are several mechanisms, first, to choose a process from the set of processes. We could use a First-Come, First-Serve method, which means ordering the processes in a queue and running them as they come. However, if a particular process is computationally intensive, processes involved with user-interaction for example would have to wait. This results in an obvious lag or hang in the system as the interactivity of the system stalls to finish computation.

To solve this problem a scheduler can $preempt$ a process after a certain amount of time has passed. This time slice is also called a time interval or a $quantum$ and has quite a literature involved with its selection. Too short, doesn't allow a process enough time to progress and the runtime system starts to spend more time context switching than computation. Too long and the preemptive-scheduler effectively becomes non-preemptive as all computation-bound processes hog the CPU from the interactive ones.

cite

After preempting a process, the scheduler has a choice as to where the process is placed. The common choice is to place it at the end of the queue, his behaviour is called Round-Robin. It is a common choice, not necessarily because of its simplicity, but due to its fairness. Each process in the queue is guaranteed an equal amount of time on the CPU and starvation of processes can therefore never happen. However, this isn't always the case as it's based on how process spawning is implemented. For example, if the newly spawned process was placed at the front of the queue or preempts the currently running process, a fork-bomb like process could hog the CPU and effectively shut out all other processes. Spawning to the end of the queue is the only effective way to avoid these scenarios in Round-Robin.

This, however, has all been using the assumption of a single process queue. While it is possible to implement a single global queue for all $P$ processors, we will eventually get into an issue of contention where all the processors are attempting to take or add a process to the queue while another one is. However, in the event of multiple process queues there needs to be a mechanism in place for dispersing the processes across them all in an even or fair way.

There are two mechanisms for this, *work-sharing* and *work-stealing*. In work-sharing, the processor with more than enough work to do, will offload any new processes onto another (either randomly or by some heuristic). In work-stealing, it's the scheduler with the empty or small process queue that contacts another scheduler (either randomly or by some heuristic) so as to steal one or more. In the case of work-stealing the victim processor can be working on a process while another processor steals from it. This means that the cost of performing the process transfer is potentially hidden by the parallelism gained. However, in the case of work-sharing there is always an additional cost involved on top of the time of execution as the overloaded processor must wait to work until after the transfer is complete.

This is why most schedulers which support multiple processing units utilize some work-stealing implementation. Of the implementations, there are two which we would like to highlight as they are provided by the ErLam toolkit: Shared-Queues and Interrupting-Steal. The Shared-Queues work-stealing scheduler allows other processes to directly access an end of their local process queue. This means, while a processor is potentially popping from one end of the queue, another could be stealing from the other end (assuming a lock-free doubly-ended queue like structure).

The alternative, Interrupting-Steal, has gone by several names like Work-Requesting, and Thief Processes. It's mechanism is to send a fake or dummy process to one or more other schedulers so when they run them they steal a process and send it back to its parent process. This reduces the overhead involved in synchronizing on the victim's process queue, but will instead stall it during the steal.

## 2.4 Feedback-Enabled Scheduling

Operating systems have also had motivation for designing intelligent feedback-enabled schedulers. As systems move away from perfect knowledge about the jobs it will be running, scheduling has needed to make guesses about the length of time jobs will need to run. A well-known example to this effect is called the Multi-Level Feedback Queue (MLFQ) scheduler, first described by Corbató *et al.* [9, 10].

The scheduler maintains $N$ separate process queues, for $N$ priority levels. All new processes would be spawned to the highest priority and would be subsequently demoted if they ended up running their whole designated quantum. However, a process may inadvertently game the system by running just up to the quantum before yielding. To fix this, after some time, $S$, the MLFQ is reset and all processes are boosted to the highest priority. This helps with adapting to new system behaviour which may arise as well as coping with process starvation.

The goal of the MLFQ model is two-fold: to prefer interactive processes and to subsequently reduce the strain of computation bound processes on the overall system. This allows the system to prune the short-running processes out quickly and also maintain an adequate level of interactivity. A MLFQ implementer would also be able to heuristically set the quantum, $N$, and $S$ based on the needs of the system as it's running, so as to introduce a second layer of feedback. For example, one could observe how much of a particular time period each priority queue is using. If a lower priority queue is being starved, it could trigger a reset [11].

The MLFQ idea in general is highly maliable and can be adapted to a number of situations. As such it transferred well into the level of runtime systems quite well. Concurrent ML (CML), uses this idea of a MLFQ to improve application interactivity.

CML is an extension to SML which adds the *spawn* function, and channel operations, among other things (such as asynchronous events) [12]. CML's scheduler defines

a MLFQ where $N = 2$ and uses a single promotion algorithm instead of a reset. However, there is a key difference: CML uses process tagging to mark whether a process has communicated in the past.

As all newly spawned processes are appended to the primary queue, CML tells the difference between these newcomers and the short-running processes by tagging any process which makes a communication, or demoting it if not. A promotion can only happen if a previously marked process gets a demotion. However, the demotion process of a marked process is just a mark removal. Thus, the primary queue is essentially two queues in one.

CML's dual-queue system has the effect of reacting to new processes by testing them for longevity. It then makes an assumption about their behaviour immediately, but a process can change the scheduler's first impression of them through consistent behaviour to the contrary. A marked communication-bound process will, if it continues to use it's entire quantum, eventually be demoted. A computation-bound process can eventually be promoted and marked as a communication-bound if it continues to communicate. Thus the system eventually adapts it's behaviour to the new phase of the process.

However, recently an alternative mechanism has been utilized to adapt to system behaviour, that of process batching. The occam-$\pi$ language, and specifically the Kent Retargetable occam Compiler (KRoC), allows processes which frequently communicate to be batched and processed together [13]. This has two side effects: cache-affinity, and informed work-stealing.

The goal of the KRoC scheduler is primarily to take advantage of cache-locality when scheduling processes. It does so by reducing the chances for cache-misses by grouping processes which have a higher likelihood to communicate. The rational being, if two processes communicate, the data which is being shared will be in cache unless too many context-switches forces it out, thus place them close together in the queue. As a side effect of this, instead of stealing single processes, the KRoC schedulers will steal batches from each-other. This results in a quicker equilibrium in work-load saturation than stealing single processes.

Process migration between batches is done in two ways: 1. A channel synchronizes and causes the process to be de-scheduled from one scheduler and sent to the one which unblocks it. 2. A batch is split when more than one process in a batch is active, by popping the head of the batch into a new one. We explain this de-scheduling method in greater detail in Section 3.2.2, as we've implemented this mechanic for testing purposes. However,

the mechanism absorbs a blocking process into the channel it's blocked on until another process unblocks it. At that time, the scheduler which unblocked it, now becomes its owner. Occam-$\pi$ uses this mechanism as a method to build up batches of cooperating processes.

Ritson *et al.* mention however, that without a method to break up the batches, the system will eventually become one large batch. Therefore, whenever a new process joins a batch, the batch is allowed to split if there are more than one currently active processes within it (*e.g.* non-blocked or waiting processes). Thus, if a parent spawns a large number of processes (*i.e.* passed the batch size limit), the parent can start a new batch, while the batch of children can be stolen.

While KRoC's primary goal was cache-affinity, and CML's was optimizing interactivity, their feedback systems enabled a closer to optimal schedule than would have otherwise been possible with a classical scheduler focused on work-saturation. We now discuss another feedback metric, *Process Cooperativity*, which KRoC's scheduler, and our algorithms presented later, were able to benefit from.

### 2.4.1 Cooperativity as a Metric

Process Cooperativity stands out as a critical feedback metric in process-oriented programming. In fact the KRoC scheduler showed this through their performance gains. They showed that when a scheduler can recognize when two or more processes form a subcomponent, treating them that way improves cache utilization, reduces context switching time, and makes for smarter work-stealing. From this, we can take that recognizing cooperativity gives a good mechanism for determining a potential for fine-grained parallelism.

However, we would like to revisit the concern Ritson *et al.* expressed regarding when a component becomes too large. They introduced the mechanism of splitting batches based on an arbitrary max size of a batch, without regard to the substructure of the component expressed by the processes cooperation.

To illustrate this problem, figure 2.3 visualizes two possible components which may occur naturally. On the left we get a ring like structure, where the dependency of one process is the one to it's right. We can abstractly envision a data-flow like application which acts like a token ring network. On the right we have a cluster of processes, all communicating with a random other on the same channel. We can envision this as an abstraction over a single shared resource.

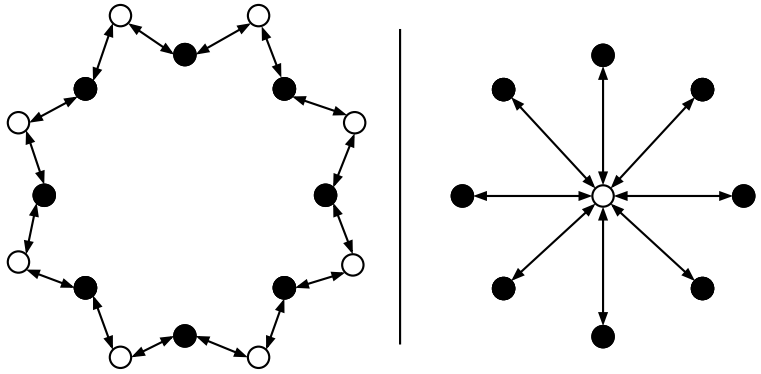


Figure 2.3: Two subcomponents formed by process cooperativity. Black dots represent processes, and the white dots represent a channel.

Both the ring and cluster subcomponents would gradually become grouped into a KRoC batch. This is optimal for the ring component, no matter how large the ring may be. There is nothing to be gained from splitting it into multiple batches, and by doing so, we may actually hinder it. However, the cluster component may improve if given the chance for more parallelism, this would depend entirely on the longevity of the processes. KRoC attempts to account for this by recognizing if there are multiple active processes in a batch, and splitting an arbitrary one into a new batch (which ever happens to be at the head of the process queue).

While this may not be avoidable based on observing structure alone, we may now run into an issue. Suppose all processes are active, but run for a length of time under their designated time quantum. At every preemption, when the size of the batch forces a split, we will create a singleton batch which must be reabsorbed after a single run. Ignoring the overhead, fairness properties also start to percolate. Namely, the processes within the batch after being trimmed will get preferential treatment to the processes in the singleton queue. Especially in the case of a single processor. Inevitably, the worst-case scenario for KRoC is below that of the work-stealing scheduler.

From this we can take that the longevity of a process can effect it's cooperativity. In fact, looking at definition 1, we can apply it to a single process too. Now, a *process' degree of cooperativity* can be defined as its *frequency* of interaction with a set of channels. Thus, a cooperativity-conscious scheduler should also want to consider both the longevity of a process, and which channels it communicates with. This would give a much more complete picture of cooperativity.

# Chapter 3

# Methodology

## 3.1 Overview

To examine the effects of cooperativity-conscious schedulers we needed to have a method for comparing several scheduler implementations without needing to modify the underlying implementation of processes, channels, or application source code. It would be also beneficial if our solution were able to visualize these differences similar to Haskell's ThreadScope [14].

Our solution, $ErLam$, is a compiler for an experimental version of Lambda Calculus with Swap Channels and a runtime system which allows for swappable scheduler mechanisms and an optional logging system which can be fed into a custom report generator. We break up our solution description into three parts; Section 3.2 will discuss our language syntax and semantics. It will also demonstrate our Runtime Scheduler API by breaking down the CML Interactivity scheduler. Section 3.3 will go more into depth about our testing environment which involves our logging system, the report generator, and the set of example applications we used to represent different cooperativity levels. Finally, Section 3.4 will go over our example schedulers we wrote which demonstrate cooperative-conscious behavior. These will be the schedulers we provide our results against.

## 3.2 ErLam

The ErLam toolkit is itself broken down into three parts, the language and its semantics, the Runtime System, and the Scheduler API. We will first lay out the language and it's basic semantics, as the finer-details are reliant on the exact selected scheduling solution as well as the chosen swap-channel implementation. We will then examine the possible channel implementations and how they effect the given semantics. Next, we will discuss the Scheduler API using an example scheduler implementation. We conclude this chapter with a summary of each of the classic schedulers that are included in the ErLam toolkit.

```
<Expression> ::= <Variable>
               |  <Integer>
               |  'newchan'
               |  '(' <Expression> ')'
               |  <Expression> <Expression>
               |  'if' <Expression> <Expression> <Expression>
               |  'swap' <Expression> <Expression>
               |  'spawn' <Expression>
               |  'fun' <Variable> '.' <Expression>
```

Figure 3.1: The ErLam language grammar, without syntax sugar or types.

### 3.2.1 The ErLam Language

The ErLam Language is based on Lambda Calculus, with first-class single variable functions, but deviates somewhat in that it provides other first-class entities. It deviates from Church representation to provide Integers, this is purely for ease of use. It also provides a symmetric synchronous Channel type for interprocess communication. As a note, this language can also be classified as a Simply-Typed Lambda Calculus.

ErLam also makes a number of ease-of-use decisions like providing a default branch operator and has some useful syntactic sugar such as SML style $let$ expressions and multi-variable function definitions. There is also a set of built in functions for numeric operations, type checking, and standard functional behaviors (*e.g.* combinators, *etc.*) which are ignored in this document.

Figure 3.1 expresses ErLam in its simplified BN-Form. The semantics for the language is fairly straight forward, but it's operational semantics are layed out in appendix A. All expressions reduce to one of the terminal types: Integer, Channel, or Function. To spawn for instance, if any terminal is passed other than a function, it returns a $0$ (*e.g.* false). When the function is passed, it is applied with $nil$ to initialize the internal expression.

ErLam extends this grammar only a little to add SML style $let$ expressions and multiple variable functions which are curried from left to right (see figure 3.2 for syntactic transformation). We will be using this syntactic sugar throughout this document to make our source easier to review.

Also, note the possible steps *swap* can take: either returning a block or another expression and a set of functions. On a semantic level, either event is transparent and results in blocking behaviour until a successful swap. However, in the former case, the channel has blocked and the only course of action for the scheduler is to get another expression to work

$$\textbf{let}\ x\ =\ e_1\ \textbf{in}\ e_2 \Rightarrow ((\textbf{fun}\ x.e_2)\ e_1)$$

$$\textbf{fun}\ x,y,z.e \Rightarrow \textbf{fun}\ x.(\textbf{fun}\ y.(\textbf{fun}\ z.e))$$

Figure 3.2: Syntactic sugar parse transformations.

on. In the later case, we have an expression to work on, but we also may have unblocked other processes by doing so, so we need to reschedule them. Note that in this case the function set may be null and the expression returned may be another attempt at swapping (*i.e.* $e = (\textbf{swap}\ c\ v)$). This would let the scheduler choose whether to retry immediately or reschedule it for a later time and work on something in the mean time. Thus, there are several possible channel implementations we could provide while still adhering to the above semantics.

### 3.2.2 Channel Implementations

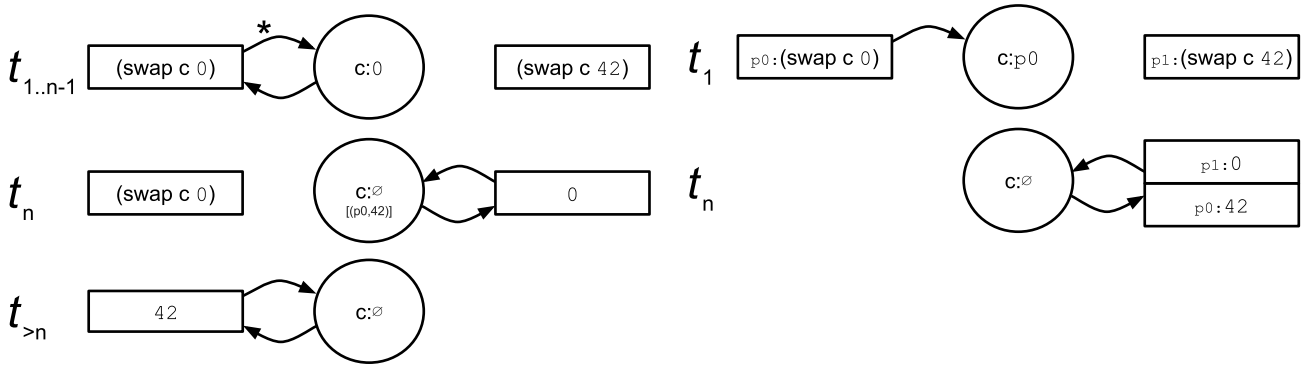
ErLam provides a selection of channel implementations to allow for interchangeable scheduler comparisons with different synchronization methods. We chose two channel implementations the *Blocking* Swap, and the *Absorbing* Swap as they highlight key differences for the runtime. We will now look at an example application and its execution using both methods for comparison.

Figure 3.3 gives an example ErLam application. It first creates a new channel for processes to communicate on. It then creates a null-function to spawn, who's sole purpose is to swap on the channel the number $42$ and quit. Finally, it swaps on the channel the number $0$ and returns the result of the whole evaluation, which in this case will be the value passed from the other end of the swap, $42$.

```
let c = newchan in
let f = (fun _.(swap c 42)) in
let _ = (spawn f) in
in (swap c 0)
```

Figure 3.3: A simple ErLam application which swaps on a channel before returning.

As ErLam is innately concurrent, we do not know which process will ask to swap first. It may even be possible that $0$ asks to swap several times before $42$ even tries. In fact, the *Blocking* channel allows this behaviour of multiple swap attempts. We can see an illustration of this in figure 3.4(a). The first row shows the arbitrary time-slice $t_1$ where




(a) Process Blocking Swap

(b) Process Absorption Swap

Figure 3.4: Channel operation over time. Note arbitrary time-slice $t_1$ is when the first swap operation is evaluated.

the process swapping $0$, $p0$, first contacts the channel with its value. The process may be scheduled again, repeatedly, to check the channel up to some arbitrary time-slice $t_{n-1}$. At $t_n$ however, the process swapping $42$ requests a swap and immediately gets a value back and logs that the process which it swapped with can get its value when it returns. Thus on the third line, for any arbitrary time-slice in the future $t_{>n}$, the process $p0$ can ask for a swap and get the value $p1$ stored.

Note the illustration makes no explicit mention of the scheduler or its functionality. It may be the case that the two processes are on different processing units and are in different process queues. Or it may be the case that they both exist in the same queue and upon a block, the scheduler chooses the next one, which will immediately unblock the channel.

The *Blocking* channel effectively simulates a common spin-lock over a shared piece of memory. These channels represent a worst-case, albeit common, application implementation for concurrent software. Even so, they do allow for some hints to the scheduler, which can be taken advantage of. Alternatives on the spin-lock could be added to the ErLam toolkit though, such as a push-notifying semaphore, however we provide a simpler and functionally more common alternative: the process absorption channel.

In the *Absorption* channel (figure 3.4(b)), the first process to get to the channel will get absorbed by it. The scheduler which evaluated the swap will be out a process, and but the scheduler which unblocks the channel by arriving second will get back two processes (the one performing the swap, and the absorbed one). In terms of scheduling efficiency

```
-callback layout( erlang:cpu_topology(), scheduler_opts() ) ->
                scheduler_layout().

-callback init( scheduler_opts() ) ->
                {ok, scheduler_state()} |
                {error, log_msg()}      |
                {warn, log_msg(), scheduler_state()}.

-callback cleanup( scheduler_state() ) ->
                ok | {error, log_msg()} | {warn, log_msg()}.

-callback tick( scheduler_status(), scheduler_state() ) ->
                {ok, scheduler_status(), scheduler_state()} |
                {return, term()} | {stop, scheduler_state()}.

-callback spawn_process( erlam_process(), scheduler_state() ) ->
                {ok, scheduler_state()} | {error, log_msg()}.
```

Figure 3.5: The ErLam Scheduler API

this type of message passing channel has provided enormous improvements for run-times which do not wish to introduce channel inspection into their scheduler.

### 3.2.3 The Scheduler API

ErLam was written in Erlang, and as such, can take advantage of Erlang's callback behaviour specifications. An *erlam_scheduler* behaviour was defined which requires a minimum of 5 callback functions (figure 3.5).

Upon instantiation the runtime system will call the *layout/2* function with the NUMA layout of the system that the application is running on, along with any parameters the user specified at runtime. The result of this function is to be the scheduler layout.

For example, let's assume we are running our application on a Intel Core i7 which has 4 logical cores which support hyper-threading. The *layout/2* function will be given the following structure:

```
[{processor,[{core,[{thread,{logical,0}},{thread,{logical,1}}]},
             {core,[{thread,{logical,2}},{thread,{logical,3}}]},
             {core,[{thread,{logical,4}},{thread,{logical,5}}]},
             {core,[{thread,{logical,6}},{thread,{logical,7}}]}]}].
```

This indicates to the scheduler implementation that it, at max, can spawn $8$ instances of itself which would be bound to each logical processing unit (LPU). Although we could of course have a scheduler which acts differently based on the architecture. However, the schedulers we have limited ourselves to are either single or fully multi-core (*i.e.* uses all available LPUs).

To spin up an instance of the scheduler on the particular core, the *init/1* function is called which should return the scheduler's state. As Erlang is a functional language, we use this state object as a means to maintain some global state for each scheduler process by threading it through all subsequent callback calls. Upon shutdown, the opposite function *cleanup/1* is called.

The last two functions are the most interesting as they pertain to the core of what each new scheduler provides, namely how to evaluate the world in a given time-slice (*tick/2*) and how a new process should be handled (*spawn_process/2*). An explanation of these callbacks is best done through example.

### 3.2.4 Example Usage: The CML Scheduler

CML's scheduler utilizes a dual-queue structure rather than a simple unary-process-queue. The scheduler attempts to differentiate between *communication* and *computation*-bound processes so as to reduce the effects of highly computationally intensive processes from choking the system. The scheduling system thus improves on application interactivity by demoting *computation*-bound processes to the secondary queue (which isn't accessed until another process is demoted).

Spawning a process in the CML scheduler (figure 3.6) does not go onto the primary queue, instead we enqueue the current process and start evaluating the new process. This is a fairly simplistic example, but it shows how one would go about updating the state between ticks. Note also, that the *spawn_process/2* call happens on the same scheduler instance which evaluated the $spawn$. While this is not of consequence for this scheduler, a multi-core scheduler could be confident in appending a new process to its local queue without interfering with another LPU's scheduler.

In the original CML scheduler, it defined a quantum which it would let the current process run for, it would preempt it if it attempted to run for longer. The ErLam runtime avoids the use of time based quantum as logging and other factors directly effect the usefulness of this. Instead it uses a 'tick', which emulates one step forward in the execution of the

```
spawn_process( Process, State ) ->
    enqueueAndSwitchCurThread( Process, State ).

enqueueAndSwitchCurThread( Process, #state{curThread=T}=State ) ->
    case T of
        nil ->
            setCurThread( Process, State );
        _   ->
            % New process takes over
            {ok, NewState} = enqueue1( T, State ),
            setCurThread( Process, NewState )
    end.
```

Figure 3.6: CML Process Spawning.

```
tick( _Status, #state{ curReduct=0 }=State ) ->
    {ok, NState} =  pick_next( State ),
    reduce( NState );
tick( _Status, State ) -> reduce( State ).

pick_next( State ) ->
    {ok, NewState} = preempt( State ),      % Place cur thread onto queue
    {ok, Top, Next} = dequeue1( NewState ), % Pop next off
    setCurThread( Top, Next ).              % Set as cur and return state
```

Figure 3.7: CML Process evaluation.

application. Thus to simulate a quantum we instead keep track of the number of reductions performed on the current process and decrement the counter until we reach $0$.

The *tick/2* function (figure 3.7) performs one of two things based on what the state of the system is. If the current reduction count is $0$, then we can pick a new process from the queue, otherwise we can perform a reduction.

Note for our scheduler simulation we ignore the first parameter to the *tick/2* function for either case. The first parameter was the status of the scheduler returned from the previous tick (*e.g.* running, waiting, *etc.*). This would be useful if the CML scheduler utilized work-stealing to get work to do from other LPUs when in *waiting* mode.

### 3.2.5 Provided Schedulers

Along with the Single-Threaded Dual-Queue CML scheduler *(STDQ)*, ErLam comes with several basic scheduling mechanics. We utilize these as bases cases on which to compare the behaviour of all subsequent feedback-enabled schedulers.

- **The Single-Threaded Round-Robin Scheduler** *(STRR)*
  This scheduler uses a single FIFO queue which all processes are spawned to. There is no rearrangement of order, and the single-thread scheduler will just round-robin the queue performing a set number of reductions per process before enqueuing and poping the next one.
- **The Multi-Threaded Round-Robin Global-Queue Scheduler** *(MTRRGQ)*
  A multi-core version of the previous scheduler. This uses a single global process queue which all schedulers share and attempt to work from.
- **The Multi-Threaded Round-Robin Work-Stealing Scheduler** *(MTRRWS)*
  An improvement on the previous scheduler. Instead of a global process queue, each scheduler maintains their own. A waiting scheduler will randomly sleep-and-steal until it finds a process to work on from another scheduler. The provided implementation gives two example stealing mechanisms:
  - **Shared-Queue** *(MTRRWS-SQ)*
    Stealing a process involved performing an atomic dequeue from the bottom (rather than the top) of another scheduler's process queue. This will only block the other scheduler from performing a dequeue for a very short window of time, but involves accessing "remote" memory.
  - **Interrupting-Steal** *(MTRRWS-IS)*
    Simulates sending a thief-process over to another scheduler. When the victim scheduler preempts or yields their current process and selects the next one from the queue, they will instead get a thief process which will syphon a process away to spawn on the thief's home scheduler. This blocks the process for a longer period of time, but does not involve accessing remote memory.

ErLam also comes with three cooperativity-conscious schedulers: the Longevity-Based Batching Scheduler (section 3.4.1), the Channel Pinning Scheduler (section 3.4.2), and the Bipartite Graph Aided Shuffling Scheduler (section 3.4.3). The first two build on

the same shared queue module as provided by $MTRRWS^*$, while the third utilizes it's own implementation.

For any compiled ErLam script, the runtime installs a command line option for selecting the scheduler used (among several other options). We are able to specify that we wish to run $pfib$, for example, with $MTRRGQ$ with the following command:

```
./pfib -s erlam_sched_global_queue
```

Any new schedulers can be added to the ErLam toolkit without needing to recompile the scripts as they are dynamically fetched and loaded at runtime.

## 3.3 Simulation & Visualization

The second primary goal of the ErLam toolkit was the ability to visualize how a scheduler proceeded to evaluate an ErLam application. We therefore needed a way to log all events over time, including unique per-scheduler events, such as the size of both the primary and secondary queues in the CML scheduler. It would also be advantageous to be as finely grained as possible and leave it up to the visualization mechanism to dial the accuracy.

We also needed a sample set of application simulations to run our set of schedulers against. These simulations needed to be minimal to reduce extraneous data but still demonstrate various levels of cooperativity and phase changes. We would like to also have the ability to compose test cases together to better create realistic work-sets for the schedulers to react to.

### 3.3.1 Runtime Log Reports

Logging in Erlang is a fairly simple matter. We utilize a simplistic data logging module based on syslog. The output of running an application could look like this:

```
timestamp,lpu,event,value
...
983847.935268,3,sched_state,running
983847.935333,0,queue_length,59
983847.935677,24,channel_blocked,6102
983847.935683,6,yield,""
983847.936003,4,queue_length,50
```

```
983847.936430,3,tick,""
983847.936439,3,reduction,""
...
```

The time-stamp given is a concatenation of the second and microsecond that the event happened in. The lpu is the scheduler which caused the event, unless it's a channel based event, such as a *channel_blocked* event, in which case it's the channel ID.

Our logging API is fairly simplistic as we only need to capture two types of metrics from our events: quantity and frequency. With frequency, we want to know the amount of events which happened in a time range, but with quantity we would like things like length of the scheduler's process queue over time or the amount of time spent in the running or waiting state.

Note time is not consistent per LPU, it may be the case that another OS application is getting time instead of one of the ErLam schedulers. This could result in one or more of the LPUs getting far less *"tick"* events. Worse yet, there may be a large gap of time missing from one scheduler to the next. For our purposes though we would like to compare the state of the scheduler while it is executing and would be fine with averaging over the largest gap. These from experimentation have not been found to be very frequent or large on an otherwise unoccupied processor.

Anecdotal! I need to prove this

To explain this averaging technique we'll now discuss the report generation method. The ErLam toolkit comes with a secondary R script which can be given a generated log file for processing. This script dynamically loads chart creation scripts based on the types of events it sees in the log file. The toolkit comes with five charting scripts which should work for all schedulers: Channel Usage (Communication Density) over time, Channel State (blocked vs. unblocked) over time, Process Queue Length per LPU over time, Reductions (Computation Density) over time, and Scheduler State (running vs waiting) over time.

Communication Density for example (see figure 3.8, creates a heatmap based on the frequency of *yield* events which occur whenever a process attempts a $swap$. Each cell of the heatmap is a color intensity based on the number of *yield* events seen in a given time-slice for a given LPU. This time-slice is where the averages come into play. R heatmaps have a max number of colors of 9, so any range we select must be modulo 9. However, the constant multiplicand is based on the mean amount of time $N$ ticks take place across each LPU. We can obviously tune the accuracy of these averages on a per-LPU basis by modifying $N$. Anecdotally, this turned out to be advantageous on several occasions when

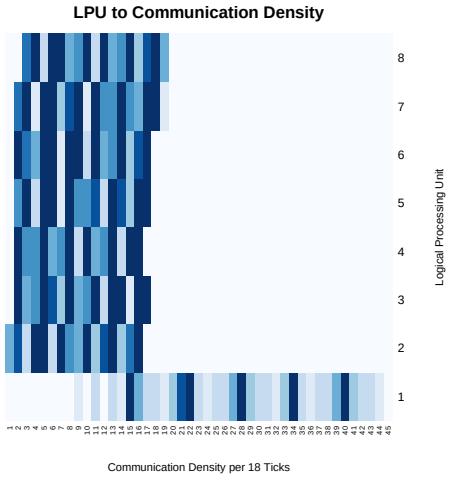



Figure 3.8: Example of Communication Density graph for the Work-Stealing scheduler on a Core i7 running the $PRing$ application.

debugging scheduler implementations. As decreasing the number of ticks to average over, increased the number of samples and thus accuracy.

### 3.3.2 Cooperativity Testing

As part of the thought experiment, we needed to implement a decent set of test cases which would give us a good coverage of the range of cooperative behaviour in common applications.

On one hand we have an axis depicting the amount of parallelism possible in an application. A system which is completely parallel, would be one where all processes spawned have no dependence on any of the others. For our toolkit, we called this behaviour $ChugMachine_N$ (figure 3.9(d)) ; where $N$ depicts the number of parallel processes. On the other side of the axis, we would have a system which had absolutely no parallelism possible. We called this behaviour $PRing_N$ (figure 3.9(b)), as it would spawn $N$ processes in a ring formation and pass a token in one direction. Each process has a channel to its left and right and would synchronize to the right until it receives a token to continue.

$PRing_N$ also gives an example of full-system cooperation, except we would instead like some degree of parallelism possible. To experiment with that, we would have to throttle the degree of cooperativity. This behaviour is called $ClusterComm_{(N,M)}$ (fig-

ure 3.9(c)) as it spawns $N$ processes and $M$ channels which can be synchronized with by any process. Note for this system to work with swap channels we limit $M$ to be at most $\lfloor N/2 \rfloor$ for all tests.

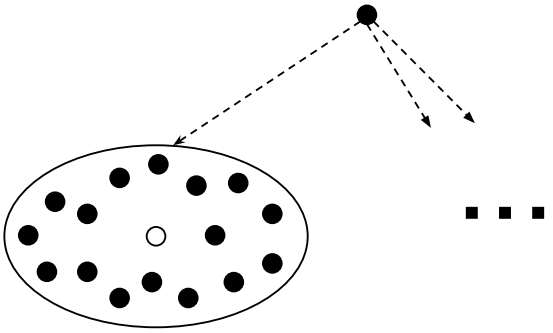
$ClusterComm_{(N,M)}$ is also an example of full-system cooperation, we also want to have a possible case for partial-system cooperation. We begin this range of experiments with a behaviour which acts like a bunch of $ClusterComm_{(N,1)}$ running in parallel. We call this special case behaviour $PTree_W$ (figure 3.9(a)); where $W$ is the number of work groups to run in parallel. This is the cleanest case of partial-system cooperation. We would expect to see obvious clustering of processes by work-group affiliation if the scheduler was cooperativity-conscious.

However, to expand on the concept of partial-system cooperativity, we would also like to experiment with lop-sided behaviours where a work-group exists along with other processes which may not be affiliated with one another. An application like this would be the combination of $ClusterComm_{(N,M)}$, $ChugMachine_N$, and/or $PRing_N$ running in parallel. For this reason, we made our behaviours composable.

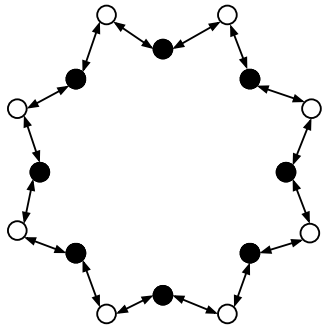
We are missing two important behaviour simulations: application phase changes, and hanging processes (typical of I/O bound processes). In the case of the latter, a simple built-in command $hang$ was provided which would simulate hanging for a random amount of time before allowing the process to proceed with evaluation. If a scheduler attempted to reduce the process before the $hang$ time was completed, it would be immediately pre-empted. The behaviour which implements this is called $UserInput_{(T,C)}$ (figure 3.9(e)); where $T$ is the max time in seconds the process would hang before continuing, and $C$ is the number of times it would simulate "waiting for user input". This simple behaviour would also compose with the others.

For the former missing behaviour, phase changes, we decided to make a variation of $PTree_W$ called $JumpShip_{(W,P)}$ (figure 3.9(f)) which would act like $PTree_W$ but would "change phase" $P$ times before completion. The act of "changing phase" would be the successive relocation of all the processes from one work-group to another, effectively having all processes from work-group $X$ "jump-ship" to $X + 1$ **mod** $N$.

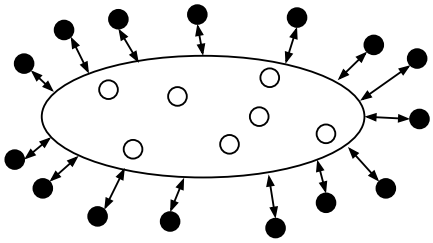
We would have liked to possibly experiment with variations on the "jump-ship" behaviour so as to inject phase changes into $PRing_N$ (by perhaps reversing direction) or $ClusterComm_{(N,M)}$ (by switching to $ChugMachine_N$ for a brief period before returning to $ClusterComm_{(N,M)}$). Yet time constraints have limited us to the aforementioned.



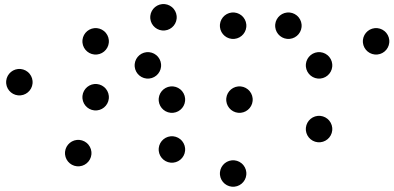
(a) Graphical representation of $PTree$, $N$ Parallel work groups.



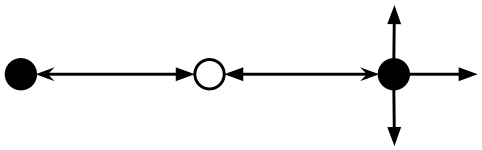
(b) Graphical representation of $PRing$, full system predictable cooperation.



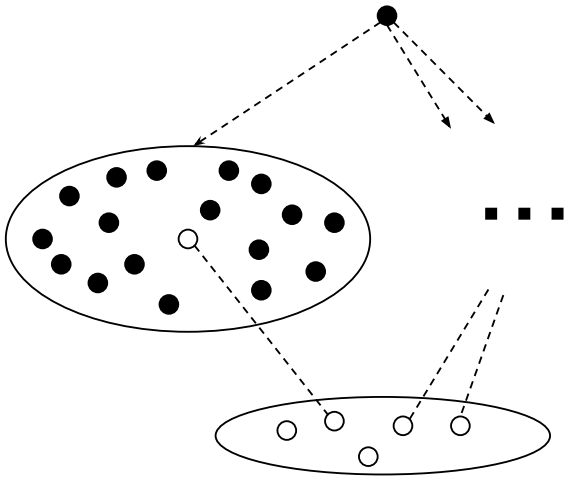
(c) Graphical representation of $ClusterComm$, $N$ processes to $M$ channels for unpredictable full system cooperation.



(d) Graphical representation of $ChugMachine$, $N$ worker processes without cooperation.



(e) Graphical representation of $UserInput$, single randomly hanging process.



(f) Graphical representation of $JumpShip$, $N$ Parallel phase shifting work groups.

Figure 3.9: Simulated behaviour examples.

## 3.4 Cooperativity Mechanics

As mentioned previously, ErLam provides a number of feedback-enabled cooperativity-conscious schedulers for comparison purposes. Through our exploration of cooperative behaviour we noticed that cooperativity has, on some level, been captured in previous scheduling systems before. Through techniques keeping in mind cache locality, channel efficiency, and work-stealing efficiency, we see a common theme of keeping processes which communicate together, in close proximity.

Ultimately cooperativity gives us a mechanism for recognizing when processes may be moving along the spectrum of application parallelism. We would therefore like to look at systems which recognise particular behaviours which utilize this.

### 3.4.1 Longevity-Based Batching

As we've mentioned before, batching processes together has been a common mechanism in capturing some of the lost efficiency of cache locality. To take advantage of the cache a message passing language is required to be considerate of their channel implementation to allow for it. For example, occam-$\pi$ makes sure their channel fits into a single 32 bit cache line (technically two, as they have asymmetrical channels and split it based on send and receive).

However, ErLam elevates the concerns of cache locality into an issue of process locality by relying only on symmetrical message passing. Note that a channel is loaded into memory at the location (*i.e.* LPU) of whichever process first blocks it and then again at whichever process unblocks it. Thus to reduce the number of times the channel needs to be loaded, we can still use the batching mechanism. Namely, if two processes are communicating frequently, the channel never needs to be migrated if they exist in the same batch and thus on the same LPU. This mechanism would be applicable to any other language which uses message passing too.

However, this batching mechanism is great when the application is going through a phase of frequent communication and thus fine-grained parallelism. But in the event the phase changes or the system is course-grained, then a batch actually harms the ability of the scheduler to parallelize to the fullest. This issue however is one we've seen before, the struggle of computation and communication bound processes.

The CML Interactivity based scheduler breaks up the processes into groups of long running (computation-bound) and short running (communication-bound) processes. We

can take advantage of this in our case by kicking processes which are too long-running out of a batch (into a singleton batch for example). To merge communicating processes back together in the same batch would be as easy as turning on Process Absorption for our channels. A process would then be batched with processes that have all begun communicating on a particular set of channels.

The mechanism we just described has been implemented as the Longevity-Based Batching Scheduler. A process belongs to a batch as long as it does not exceed it's quantum (reduction count). If it is preempted, it is thusly kicked out from the batch it's apart of (unless it's a singleton). There are of course alternatives to this (*e.g.* allow for $N$ quantum to pass before kicking a process from a batch), but we, for testing purposes, can just extend the size of the quantum for the same effect.

To gain entry back into a batch, all a process needs to do is perform a swap with Process Absorption turned on. Note as Process Absorption must be toggled on for this to work, we can experiment with max-parallelism possible in the application. Our assumption prior to experimentation was that, the longevity-batching scheduler would drop-down into a common work-stealing scheduler in the worst case (*i.e.* all processes are singleton batches). We talk more on these assumptions and their validity in Chapter 4.

### 3.4.2 Channel Pinning

An alternative approach to batching, which moves the channels to the processes, is to set an affinity to a core based on which channels you cooperate on. We call this mechanism Channel Pinning, as we bind a channel upon creation to a particular LPU and force processes to that location to perform a swap. There are a large number of interesting mechanics for this behaviour. But we will look at three: 1. How to spread the channel pinnings? 2. How should processes react when attempting a swap on a remotely bound channel? 3. And based on the decisions made in the previous, how should a scheduler steal/spawn a process?

When choosing a channel spread algorithm there are two things which need to be compared, cost of creating a channel and channel usage of the running application. In stark contrast to the previous scheduler, channel pinning would either need to be an expensive heuristic or a programmer-aided decision based on the application itself. For example, if our channel pinning algorithm was a sane even spread across all processors we would have ignored the possibility that a subset of the channels could be used more frequently.

This could therefore cause more harm than good. If we chose to do an expensive check across all processors to compute the saturation each time we create a channel we would be harming the application which uses a map-reduce style, and uses a lot of temporary one-use channels.

On top of this complication, this also ignores the possibility of phase changes. It may be the case that during a start up phase, the frequency of particular channel usage may offset the saturation to such a degree that any further checks will suggest alternate processors despite possibly being inaccurate. However, these issues are beyond the scope of this discussion and we will instead focus on scheduling around channel pinning. As such we only provide the following channel pinning implementations: 1. *same*, which pins the channel to the same processor it is created on (*i.e.* the processor which evaluates $newchan$). 2. *even*, which pins the channel in a round-robin fashion starting at LPU $0$.

Note *same* pinning, would not be ideal for an application which creates all channels in one process and then hands them out to it's children like in $JumpShip_{(W,P)}$, or $ClusterComm_{(N,M)}$. However, *even* pinning will be perfect for them. We expect experimentation with composed application will have interesting effects in this scheduler.

We have a couple of possible mechanisms for how a process should be handled when it comes time to communicate. In the event it is wanting to communicate on a channel that is not local, we propose the following mechanism: let them anyway but if they block, spawn them to the LPU which owns the channel.

The rational for this is, in the event of a block, the local LPU won't gain anything from having it in its queue (if process absorption is turned on it would loose it anyway). However, if the process completes a swap, both the remote LPU and the local can continue in parallel.

Due to this selective spawn feature, we have the opportunity to look at a selective steal opportunity. Namely, when stealing we can attempt to grab from a random scheduler, one or more processes which have communicated with a randomly selected channel which the thief owns. This is akin to the children's card game "Go Fish" where in our case a scheduler may ask if another "has any channel 3's".

However, with this mechanism it may be the case that there is never any work for a particular scheduler. In this case, we have added the post condition that if a process has never communicated, or if the scheduler is not the owner of a channel, then they act as wild

cards and can match anything. This is both a simplistic algorithm, but it also makes sure the scheduler falls back to a standard work-stealing algorithm by default.

We now note that the primary interest in comparing this scheduler is to look at this work-stealing mechanism. We would like to see how this type of mechanism fairs against both a best and worst case scenario. We first thought the worst case scenario would be a $ChugMachine_N$ due to the reliance on wild-cards, however after more consideration a single $ClusterComm_{N,1}$ may end up having a worse behaviour due to the constant re-spawning of processes back to the owner of the primary channel. A best case in this scenario would be a $PTree_W$ when $W$ is greater than or equal to the number of processing units available.

### 3.4.3 Bipartite Graph Aided Shuffling

Both of the previous schedulers ignore the effects ordering can have on execution behaviour. We hypothesize that order of process execution could have a drastic consequence on highly cooperative processes by pairing channels such that if a process where to block, the unblock would happen as soon as possible (*i.e.* the scheduler would not choose a process which had no probability of unblocking it).

Granted the extreme of this type of scheduler would be excruciatingly unfair, and as such we still maintain a round-robin like behaviour, except now we rely on a sorted process queue. Sorting the process queue would allow us the chance to increase the likelihood of immediately unblocking a blocked channel, while still maintaining execution fairness. Sorting the process queue would also have the aided side effect of making work-stealing potentially more efficient when stealing more than one from a victim queue. Namely, it would be much more likely that the group of processes stolen are cooperating.

Due to this, we hope to show an interesting case where process absorption may not be the preferred channel implementation. The blocking channel, one which will immediately return, will allow for the process to stay sorted and allow for the next process following it to unblock it for it's next turn. Thus relying on work-stealing alone to migrate processes in a potentially smarter and more-grouped way.

There are three metrics which could effect the order of processes, and which would subsequently trigger a resorting of the process queue. A process yielding, returning, or spawned could all mean that a particular process or set of processes need to be relocated.

However, frequent sorting may cause the scheduler's fairness to suffer. We therefore set a variable, $\Gamma$ which defines how many "events" can happen before causing a resort.

To provide a mechanism for sorting, we structure our process queue as a bipartite graph, where one side is our process queue, and the other is the set of channels. We generate a "pseudo-priority" based on recency of the communications over the number of channels it's communicated with. Namely, we sort the process queue by $\Delta(P_i) = \Sigma E_i/|C_i|$ where $E_i$ is the edge set of timestamps, and $C_i$ is the set of channels $P_i$ communicated with. The default priority function, $\Delta$, is intended to give our processes which communicate frequently, a head start, and all other processes can be pushed to the end (and more likely stolen).

If our process queue is long, we maintain preferential treatment for the set of interactive processes. If our process queue is short, then the effects of sorting are non-existent. As such, we should only be concerned with sorting a process queue after a particular size. This is another chance for heuristic analysis. Due to this though, we would expect to see poorer behaviour on most simple simulated applications, but would steadily gain in advantage when subjected to composed applications.

# Chapter 4

# Results and Discussion

## 4.1 Evaluation

Evaluation of each scheduler will consist of complete testing:
- Run $PTree_N$ for $1 \leq N \leq P + 1$ where $P$ is number of processors.
- Run $PRing_N$ for $N \in \{1, P, B - 1, B, 2 * B\}$ where $B$ is max batch size where applicable.
- Run $ClusterComm_{(N,M)}$ for $N \in \{2, 4, ..., P, B, 2 * B\}$ and $1 \leq M \leq \lfloor N/2 \rfloor$
- Run $ChugMachine_N$ for $1 \leq N \leq P + 2$
- Run $JumpShip_{(W,H)}$ for $2 \leq W \leq P + 1$ and $1 \leq H \leq 3$

Composing simulations may lead to interesting results though. I would like to test, at least:
- $UserInput$ and $ChugMachine$ to examine Interactivity.
- $PTree$ and $ChugMachine$ or $ClusterComm$ to have both structured subcomponents, and otherwise.

### 4.1.1 Classical Schedulers

There are 5 schedulers, STRR, STDQ, MTRRGQ, MTRRWS-SQ, MTRRWS-IS which need to be evaluated. However, I wont know what's interesting until after all tests have been run.

### 4.1.2 Cooperativity Feedback Schedulers

There are 3 schedulers, Longevity-Batcher, Channel-Pinner, Graph Shuffling. I will primarily be looking at:
- How does Longevity Batcher degrade? Does it turn into MTRRWS?
- How quickly and thoroughly can Channel-Pinner saturate the cores with work?
- How does Graph Shuffling handle large and small process queues?

## 4.2 Comparisons

Including the above, I will want to compare:
- MTRRWS-SQ and MTRRWS-IS
- STRR and MTRRGQ
- Long-Batcher and Channel-Pinner in terms of work-stealing
- All feedback schedulers in terms of: execution time (tick count), saturation, evenness of work-load.

## 4.3 A Comment on Swap Channels

Namely, we're using an odd channel implementation that most languages do not choose. How did that work out for us?
- In implementing example applications,
- In terms of applicability of findings to other message-passing implementations,
- etc.

# Chapter 5

# Conclusion and Future Work

# Appendices

# Appendix A

# ErLam Operational Semantics

$$\text{Variable}\ \frac{E(x) \Rightarrow v}{S, C, E : x \rightarrow S, C, E : v} \qquad \text{Integer}\ \frac{}{S, C, E : n \rightarrow S, C, E : n}$$

$$\text{Fun}\ \frac{}{S, C, E : \textbf{fun}\ x\textbf{.}e \rightarrow S, C, E : \textbf{fun}\ x\textbf{.}e} \qquad \text{Unwrap}\ \frac{}{S, C, E : (e) \rightarrow S, C, E : e}$$

$$\text{NewChan}\ \frac{|C| + 1 = n \qquad C \downarrow n \Rightarrow chan_n}{S, C, E : \textbf{newchan} \rightarrow S, C; \{chan_n\}, E : chan_n}$$

$$\text{App(1)}\ \frac{S, C, E : e_1 \rightarrow S', C', E' : e_1'}{S, C, E : e_1 e_2 \rightarrow S', C', E' : e_1' e_2}$$

$$\text{App(2)}\ \frac{S, C, E : e_2 \rightarrow S', C', E' : e_2'}{S, C, E : \textbf{fun}\ x\textbf{.}e_1 e_2 \rightarrow S', C', E' : \textbf{fun}\ x\textbf{.}e_1 e_2'}$$

$$\text{App(3)}\ \frac{}{S, C, E : \textbf{fun}\ x\textbf{.}e_1 v \rightarrow S, C, E; (x, v) : e_1}$$

$$\text{If(1)}\ \frac{S, C, E : e_1 \rightarrow S', C', E' : e_1'}{S, C, E : \textbf{if}\ e_1\ e_2\ e_3 \rightarrow S', C', E' : \textbf{if}\ e_1'\ e_2\ e_3}$$

$$\text{If(2)}\ \frac{v \geq 1}{S, C, E : \textbf{if}\ v\ e_2\ e_3 \rightarrow S, C, E : e_2} \qquad \text{If(3)}\ \frac{v \leq 0}{S, C, E : \textbf{if}\ v\ e_2\ e_3 \rightarrow S, C, E : e_3}$$

$$\text{Swap(1)}\ \frac{S, C, E : e_1 \rightarrow S', C', E' : e_1'}{S, C, E : \textbf{swap}\ e_1 e_2 \rightarrow S', C', E' : \textbf{swap}\ e_1' e_2}$$

$$\text{Swap(2)}\ \frac{S, C, E : e_2 \rightarrow S', C', E' : e_2'}{S, C, E : \textbf{swap}\ c e_2 \rightarrow S', C', E' : \textbf{swap}\ c e_2'}$$

$$\text{Swap(3)}\ \frac{C(c, v) \Rightarrow \emptyset \qquad S \downarrow (S', e)}{S, C, E : \textbf{swap}\ c v \rightarrow S', C, E : e}$$

$$\text{Swap(4)}\ \frac{C(c, v) \Rightarrow (e, F) \qquad \{S \uparrow f \Rightarrow S' : \forall f \in F\}}{S, C, E : \textbf{swap}\ c v \rightarrow S', C, E : e}$$

$$\text{Spawn(1)}\ \frac{S, C, E : e \rightarrow S', C', E' : e'}{S, C, E : \textbf{spawn} e \rightarrow S', C', E' : \textbf{spawn} e'}$$

$$\text{Spawn(2)}\ \frac{S \uparrow f \Rightarrow S'}{S, C, E : \textbf{spawn}\ f \rightarrow S', C, E : 1} \qquad \text{Spawn(3)}\ \frac{}{S, C, E : \textbf{spawn}\ v \rightarrow S, C, E : 0}$$

# Bibliography


[1] John L Bruno et al. *Computer and job-shop scheduling theory*. Wiley, 1976.

[2] Michael R Garey, Ronald L Graham, and DS Johnson. "Performance guarantees for scheduling algorithms." In: *Operations Research* 26.1 (1978), pp. 3–21.

[3] Richard D Dietz et al. "The use of feedback in scheduling parallel computations." In: *Parallel Algorithms/Architecture Synthesis, 1997. Proceedings., Second Aizu International Symposium*. IEEE. 1997, pp. 124–132.

[4] Kurt Debattista, Kevin Vella, and Joseph Cordina. "Cache-affinity scheduling for fine grain multithreading." In: *Communicating Process Architectures* 2002 (2002), pp. 135–146.

[5] Robin Milner. *A calculus of communicating systems*. Springer-Verlag New York, Inc., 1982.

[6] Catuscia Palamidessi. "Comparing the expressive power of the synchronous and the asynchronous $\Pi$-calculus." In: *Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM. 1997, pp. 256–265.

[7] Evangelos P Markatos and Thomas J LeBlanc. "Load balancing vs. locality management in shared-memory multiprocessors." In: (1991).

[8] Evangelos P Markatos and Thomas J LeBlanc. "Memory-Conscious Scheduling in Shared-Memory Multiprocessors." In: *Computer Science Dept., University of Rochester* (1991).

[9] Fernando J Corbató, Marjorie Merwin-Daggett, and Robert C Daley. "An experimental time-sharing system." In: *Proceedings of the May 1-3, 1962, spring joint computer conference*. ACM. 1962, pp. 335–344.

[10] Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau. *Operating Systems: Three Easy Pieces*. 0.80. Arpaci-Dusseau Books, 2014.

[11] Kenneth Hoganson. "Reducing MLFQ scheduling starvation with feedback and exponential averaging." In: *Journal of Computing Sciences in Colleges* 25.2 (2009), pp. 196–202.

[12] John H Reppy. "Concurrent ML: Design, application and semantics." In: *Functional Programming, Concurrency, Simulation and Automated Reasoning*. Springer. 1993, pp. 165–198.

[13] Carl G Ritson, Adam T Sampson, and Frederick RM Barnes. "Multicore scheduling for lightweight communicating processes." In: *Science of Computer Programming* 77.6 (2012), pp. 727–740.

[14] Don Jones Jr, Simon Marlow, and Satnam Singh. “Parallel performance tuning for Haskell.” In: *Proceedings of the 2nd ACM SIGPLAN symposium on Haskell*. ACM. 2009, pp. 81–92.

[15] David R White et al. “Automated heap sizing in the poly/ML runtime.” In: *Trends in Functional Programming* (2012).

[16] Kunal Agrawal et al. “Adaptive work-stealing with parallelism feedback.” In: *ACM Transactions on Computer Systems (TOCS)* 26.3 (2008), p. 7.

[17] Yuxiong He, Wen-Jing Hsu, and Charles E Leiserson. “Provably efficient online nonclairvoyant adaptive scheduling.” In: *Parallel and Distributed Systems, IEEE Transactions on* 19.9 (2008), pp. 1263–1279.

[18] Thomas L. Casavant and Jon G. Kuhl. “A taxonomy of scheduling in general-purpose distributed computing systems.” In: *Software Engineering, IEEE Transactions on* 14.2 (1988), pp. 141–154.

[19] Hagai Abeliovich. “An empirical extremum principle for the hill coefficient in ligand-protein interactions showing negative cooperativity.” In: *Biophysical journal* 89.1 (2005), pp. 76–79.