Declarative coordination of parallel declarative programs

Abstract

Declarative programming has been hailed as a promising approach to parallel programming since it hides the implementation details of parallelism away from the programmer. However, some control is lost when code is automatically parallelized, leaving the programmer with very little opportunities for fine tuning programs. We extended Linear Meld, a concurrent forward-chaining linear logic programming language, with coordination facts that allow the programmer to change computation scheduling and data layout. Our approach allows the programmer to write declarative parallel programs and then optionally use coordination facts to fine tune programs, while keeping the program fully declarative and amenable to correctness proofs. We show several application examples, experimental results and some advantages of our novel approach.

Categories and Subject Descriptors D.1.3 [*PROGRAMMING TECHNIQUES*]: Concurrent Programming—Parallel Programming; D.3.4 [*PROCESSORS*]: Interpreters; D.3.4 [*PROCESSORS*]: Run-time environments

General Terms Design, Languages, Performance

Keywords Parallel Programming, Linear Logic, Virtual Machine, Implementation

1. Introduction

Writing parallel programs in sequential languages is hard because manipulating shared state using multiple threads may result in data race conditions. Such issues are handled with low level constructs such as locks and condition variables, requiring a fair amount of effort to get right. Declarative programming has been hailed as a solution to this issue, since the problem of implementing the details of parallelism is moved away from the programmer to the compiler and runtime of the language. The programmer writes code without having to deal with parallel programming constructs and the compiler automatically parallelizes the program in order to take advantage of multiple threads of execution. The programming paradigm has been adopted with huge success in domain specific languages such as SQL and MapReduce [Dean and Ghemawat 2008]. Although general declarative languages have yet to be as successful, the future looks promising for this particular approach.

We argue, however, that declarative programming leaves little to no programmer control over how execution is scheduled or how data is laid out. This introduces some performance issues because even if the runtime system is able to reasonably parallelize the program using a general algorithm, there is a lack of specific information about the program that a compiler does not easily understand. In turn, such information would make execution better in terms of run time time, memory usage and scalability.

In this paper, we extend Linear Meld (LM), a declarative language by [Cruz et al. 2014a,b] with coordination facts that give programmer control over scheduling and data placement. LM is a linear logic programming language designed for programs that operate on graphs. The use of linear logic [Girard 1987] allows support for structured manipulation of mutable state. In LM, computation is divided so that each node of the graph computes independently and is allowed to communicate with other nodes. Both computation and communication happen through derivation of logical rules.

Given that LM uses mutable state, the order in which nodes are executed may be crucial in terms of performance. We extended the language with coordination facts that can be used as a regular fact in the program logic to improve program execution. The first kind of coordination facts are called *sensing facts* and are used to sense information about the system the program is running on. This allows the programmer to write logical rules that depend on the current state of the program and the system. The second kind of coordination facts are *action facts* that when deleted in rules are used to apply a scheduling operation during execution.

This allows the programmer to prioritize node computation or place nodes in different threads.

To the best of our knowledge, this is the first time that a declarative language allows control over execution without a significant change in the language itself and without using non-declarative operations. This is specially important since programs can proven to be correct even in the presence of coordination facts. In order to prove our thesis, we first give a brief overview of the base language, including a small example. Next, in Section 4 we introduce the available coordination facts, followed by Section 5, where we describe the implementation of the runtime required to implement the desired coordination mechanisms. Afterwards, in Section 6 we present several applications where we add programmer control to declarative programs in order to make them perform better.

2. Related Work

There has been an increasing interest in declarative and data-centric languages. MapReduce [Dean and Ghemawat 2008], for instance, is a popular data-centric programming model that is optimized for large clusters. The scheduling and data sharing model is very simple: in the *map phase*, data is transformed at each node and the result reduced to a final result in the *reduce phase*. In order to facilitate the writing of programs over large datasets, SQL-like languages such as PigLatin [Olston et al. 2008] have been developed. PigLatin automatically compiles declarative code into physical plans that are executed over Hadoop but provides very little control to the programmer.

A declarative approach that is regaining popularity is Datalog [Ullman 1990], also a forward-chaining logic programming language. Traditionally used in deductive databases, it is now being increasingly used in different fields such as distributed networking [Loo et al. 2006], sensor nets [Chu et al. 2007] and cloud computing [Alvaro et al. 2010]. LM itself is inspired in P2 [Loo et al. 2006] since it also models programs as a graph structure. Another programming system that uses the graph abstraction is Dryad [Isard et al. 2007], which combines computational vertices with communication channels (edges) to form a data-flow graph. The program is then scheduled to run on multiple computers/cores and data is partitioned automatically during runtime.

Many programming languages follow the so-called *coordination paradigm* [Papadopoulos and Arbab 1998], a form of distributed programming that divides execution in two parts: *computation*, where the actual computation is performed, and *coordination*, which deals with communication and cooperation between processing units. This paradigm attempts to clearly distinguish between these two parts by providing abstractions for coordination in an attempt to provide architecture and system-independent forms of communication.

Linda [Ahuja et al. 1986] is arguably the most famous coordination model. Linda implements a data-driven coordination model and features a *tuple space* that can be manipulated using the following coordination directives: `out(t)` to write a tuple `t` into the tuple space; `in(t)` to read a tuple using the template `t`; `rd(t)` to retrieve a copy of the tuple `t` from the tuple space; and `eval(p)` to add a process `p` in the tuple space and execute it in parallel. Linda is be implemented on top of many popular languages by simply creating a communication and storage mechanism for the tuple space and then adding the directives as a language library.

Another important coordination language is Delirium [Lucco and Sharp 1990]. Unlike Linda which is embbeded into another language, Delirium actually embeds operators written in other languages inside the Delirium language. The advantages of Delirium are improved abstraction and easier debugging because sequential operators are isolated from the coordination language.

Linda and Delirium are limited in the sense that the programmer can only coordinate the scheduling of processing units, while the placement of data is left to the implementation. LM also differs from those languages since it raises the abstraction level from the processing units to nodes of a graph. Furthermore, the LM language is both a coordination language and a computation language and there is no distinction between the two components.

3. Base Language

Linear Meld (LM) is a forward-chaining linear logic programming language [Cruz et al. 2014a] with roots on Datalog [Ramakrishnan and Ullman 1993]. Programs consist of a set of *rules* and a *database of facts*. Rules such as `a(X), b(Y) -o c(X, Y)` can be read as follows: if fact `a(X)` and fact `b(Y)` exist in the database then `c(X, Y)` is added to the database. The expression `a(X), b(Y)` is called the *body* of the rule and `c(X, Y)` is the *head* of the rule. A fact is a predicate and its associated tuple of values, representing the arguments. Since LM uses linear logic as its foundation, we distinguish between *linear* and *persistent facts*. Linear facts are deleted during the process of deriving rule, while persistent facts are not. Program execution starts by adding the *axioms* (the initial facts) of the program to the database. Next, the rules are recursively applied and the database is updated with new and deleted facts. When no more rules are applicable, the program terminates.

LM has been designed for writing programs that operate on graphs. LM partitions the database by using the first argument of each fact. The first argument is typed as a *node* and represents a node in the graph. For instance, the fact `f(@1, 2)` is stored in node `@1`, while fact `p(@2)` is stored in node `@2`. LM restricts the body of every rule to only refer to the same node so that nodes can derive rules independently. Although the body is restricted, the head of the rule may refer to any node as long as the node is refered somewhere in the body. This allows communication between nodes during rule derivation, since a node must send a fact to another node. Rule restrictions in turn make LM implicitly parallel because nodes are able to compute independently as long as they have new facts to be applied in rules. This makes LM non-deterministic since nodes can be picked to run in any order, affecting which rules are applied and which facts are deleted or derived.

In order to fully understand how the language works, we present a very simple algorithm: the single source shortest path program (SSSP). Later in the paper, we add coordination facts to improve the execution of the program.

The SSSP program in Fig. 1 starts (lines 1-3) with the declaration of the predicates. Predicates specify the kinds of facts used in the program. The first predicate, `edge`, is a persistent predicate that describes the relationship between the nodes of the graph, where the third argument represents the weight of the edge (the `route` modifier informs the compiler about the structure of the program graph). The predicates `shortest` and `relax` are specified as linear facts and thus are deleted when deriving new facts. The idea of the algorithm is to compute the shortest distance from node `@1` to all other nodes in the graph. Every node has a `shortest` fact that is improved with new `relax` facts. Lines 5-9 declare the axioms of the program: `edge` facts describe the graph; `shortest(A, +00, [])` is the initial shortest distance (infinity) for all nodes; and `relax(@1, 0, [@1])` starts the algorithm with the initial distance from `@1` to `@1`.

```
1   type route edge(node, node, int).
2   type linear shortest(node, int, list int).
3   type linear relax(node, int, list int).
4
5   !edge(@1, @2, 3). !edge(@1, @3, 1).
6   !edge(@3, @2, 1). !edge(@3, @4, 5).
7   !edge(@2, @4, 1).
8   shortest(A, +00, []).
9   relax(@1, 0, [@1]).
10
11  relax(A, D1, P1), shortest(A, D2, P2),
12  D1 < D2
13     -o shortest(A, D1, P1),
14        {B, W | !edge(A, B, W) |
15           relax(B, D1 + W, P1 ++ [B])}.
16
17  relax(A, D1, P1), shortest(A, D2, P2),
18  D1 >= D2
19     -o path(A, D2, P2).
```

Figure 1: Single Source Shortest Path program code.

The first rule of the program (lines 11-15) replaces the current path in `shortest` with a shorter one in `relax`. The rule deletes both `relax` and `shortest` facts and derives a new `shortest` fact. In lines 14-15, we have a *comprehension* where the program iterates over the edges of node `A` and derives a new `relax` fact at `B` with the new shorter distance plus the weight of the edge. For instance, in Fig. 2 (a) we apply rule 1 in node `@1` where two new `relax` facts are derived at node `@2` and `@3`. Fig. 2 (b) is the result after applying the same rule but at node `2`.

The second rule of the program (lines 17-19) deals with the case where the new distance in `relax` is not better than the current one, so it is thrown away and the current distance is kept.

There many opportunities for concurrency in the SSSP program. For instance, after applying rule 1 in Fig. 2 (a), it is possible to either apply rules in either node `@2` or node `@3`. In our implementation, we partition the graph into subgraphs that are processed by multiple threads of execution. Eventually, it is no longer to possible to apply rules and the final result present in Fig. 2 (c) is achieved.

4. Coordination

Since LM uses linear logic and supports deletion of facts, the order in which nodes are scheduled can impact the performance and even the results of the program.

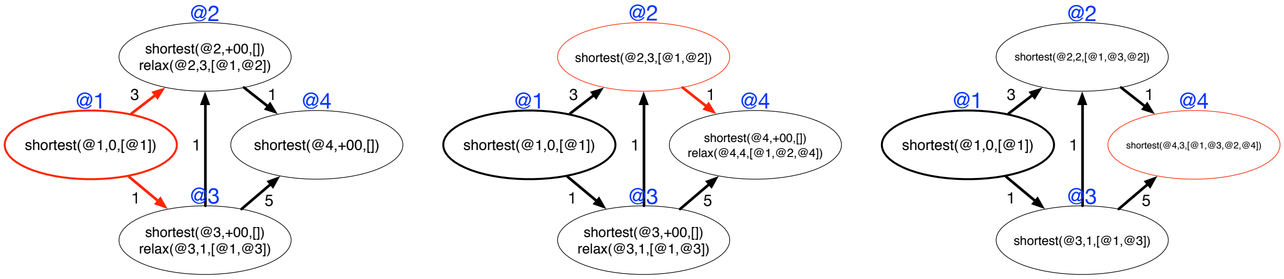
The SSSP program present in Fig. 1 is concise and declarative but its performance may depend in the order in which nodes are executed. If nodes with greater distances are prioritized over other nodes, the program will generate more `relax` facts since it will take longer to reach the shortest distances. From Fig. 2, it is clear that the best computational ordering is the following: `@1`, `@3`, `@2` and then `@4`, where only 4 `relax` facts are generated. If we had decided to process nodes in order `@1`, `@2`, `@4`, `@3`, `@4`, `@2`, then 6 `relax` facts would have been generated. Therefore, the optimal solution for SSSP is to schedule the node with the shortest distance, which is essentially the Dijkstra shortest path algorithm [Dijkstra 1959]. Note how it is possible to change the nature of the algorithm by simply changing the order of node computation, but still retain the declarative nature of the program.

We introduce the concept of *coordination facts* that allow the programmer to change how the run time schedules nodes and partitions the nodes among threads of execution. Coordination facts can be used in the body or head of the rules to allow the programmer to reason and change how execution is to be done. In this fashion, we keep the language declarative because we reason logically about the state of execution, without the need to introduce extra-logical operators into the language that would introduce problems when attempting to prove properties about programs.

There are two classes of coordination facts. The first class of coordination facts are called *sensing facts* and are used to sense information about the underlying runtime system, including node placement and node scheduling. The second kind of coordination facts are *action facts* that are deleted to apply a coordination operation in the runtime system.

4.1 Scheduling Facts

We can use action facts to change the order in which nodes are evaluated by adding *priorities*. Node priority works at the



(a)

(b)

(c)

Figure 2: Graphical representation of the SSSP program. Figure (a) represents the program after propagating initial distance at node `@1`, followed by Figure (b) where the first rule is applied in node `@3`. Figure (c) represents the state of the final program, where all the shortest paths have been computed.

thread level so that each thread can make a local decision about which node to execute next. Note that, by default, nodes are picked using a FIFO approach, because that tends to work better for most programs since older nodes tend to have more facts to be used.

We have two kinds of priorities: a *temporary priority* and a *default priority*. A temporary priority momentarily changes the default priority D of a node, so that once the node is done, the priority will default back to D. Initially, all nodes have a default priority of 0.

The following list presents the action facts available to manipulate the scheduling decisions of the system:

- `type linear set-priority(node, float)`: This sets the temporary priority of a node. If the node already has a priority, we only change the priority if the new one is higher. The programmer can decide if priorities are to be ordered in ascending or descending order.
- `type linear action add-priority(node, float)`: Increases, temporarily, the priority of the node.
- `type linear schedule-next(node)`: The work will fetch the highest priority node's priority P from its set of nodes and set the action's argument node's priority as $P + 1.0$. If using the priorities in ascending order, we pick the lowest priority and subtract 1.0.
- `type linear set-default-priority(node, float)`: Sets the default priority of the node.
- `type linear action stop-program(node)`: Immediately stops the execution of the whole program.

LM provides the sensing fact `priority(A, B, P)` in order to sense the priority `P` of node `B` from node `A`. Sensing facts are only used in the body of the rules in order to fetch information from the runtime system. Note that when sensing facts are deleted, they are re-derived automatically. Logically, `set-priority` and `set-default-priority` update the value of `priority` facts, but this done automatically by the runtime system.

4.1.1 Partitioning facts

We provide several coordination facts for dealing with node partitioning among the available threads executing the program. In terms of action facts, we have the following:

- `type linear set-cpu(node, int)`: Moves the node to a thread of execution.
- `type linear set-affinity(node, node)`: Places the first node in the same thread as the second node.
- `type linear set-moving(node)`: Allows the node to move freely between threads.
- `type linear set-static(node)`: Forces the node to stay in the same thread indefinitely.

For sensing facts, we have the following set of coordination facts:

- `type linear cpu-id(node, node, int)`: The third argument indicates the thread where the node of the second argument is currently running.
- `type linear moving(node, node)`: Fact available if the node in the second argument is allowed to move between threads.
- `type linear static(node, node)`: Fact available if the node in the second argument is not allowed to move between threads.

5. Implementation

The implementation of LM includes a compiler and a virtual machine that runs byte-code. The virtual machine uses 32 registers for operations and executes procedures to iterate over the database in order to match facts. The implementation of the original machine is described in [Cruz et al. 2014b]. Here, we review the most important details of that paper, while focusing on the changes for efficiently implementing coordination.

5.1 Compilation

The compiler translates each rule to a procedure and a list of facts that need to exist to satisfy the rule. This procedure is executed by the virtual machine whenever we have enough facts to satisfy the rule. The procedure loops over all possible combinations of the rule, retrieving facts from the database, performing join operations and then consuming and deriving facts. Optimizations such as join optimizations (to support efficient rule filtering) and fact updates are implemented by the compiler. The latter transforms fact derivations of the same predicate to a simple update operation, avoiding a deallocation operation followed by an allocation of the new fact.

5.1.1 Coordination directives

Coordination directives are compiled in two different ways, depending on whether they appear in the body or in the head of the rule. Coordination facts in the body are compiled into special instructions that inspect the state of the virtual machine. For example, the directive `node-priority` will inspect the target node, retrieve the current priority and assign the priority to a register. Coordination facts in the head of the rule are also implemented as special instructions of the virtual machine, but they perform some action, instead of returning something. While these are still facts from the point of view of the language, they are deleted immediatelly in order to apply the operation, instead of being derived like any other regular fact. This is more efficient since we do not need to create a new fact.

5.2 Execution

The virtual machine is implemented in C++11 and uses the threading system from the standard library to implement multithreading. Each thread is responsible for executing a subset of nodes. Nodes are either inactive, with no new facts, or active, where they will be placed in the corresponding thread's queue. Threads do useful work by processing the nodes in their queues. Whenever a thread becomes idle, it attempts to steal nodes from a random thread into its own queues. If sucessful, the thread becomes idle and waits for program termination while periodically attempting to steal nodes.

5.2.1 Nodes

A node is represented as a collection of facts (per predicate) and an indexing structure that keeps track of the available predicates and potential candidate rules. We need a separate indexing structure per node since rules run locally.

Facts need to be stored efficiently because the virtual machine instructions perform searches on the database by fixing arguments of a predicate to concrete values. Each predicate is stored using one of the following data structures:

- *Tries* are used exclusively to store persistent facts. Tries are trees where facts are indexed by common prefix arguments.
- *Doubly Linked Lists* are used to store linear facts. We use a double linked list because it is a very efficient way to add and remove facts.
- *Hash Tables* are used to improve lookup when linked lists are too long and when we need to do search filtered by a fixed argument. The virtual machine decides which arguments are best to be indexed and then uses a hash table indexed by the appropriate argument.

In order to facilitate the implementation of coordination, we added a state variable for each node. The state machine in Fig. 3 represents the valid state transitions of a node:

working : the node is being executed.

idle : the node is not active, has no new facts and is not in any queue for processing.

queue : the node is active with new facts and is waiting in some queue to be processed.

stealing : the node has just been stolen is about to move to another thread.

coordinating : the node is moving from one queue to another (i.e. changing priority).

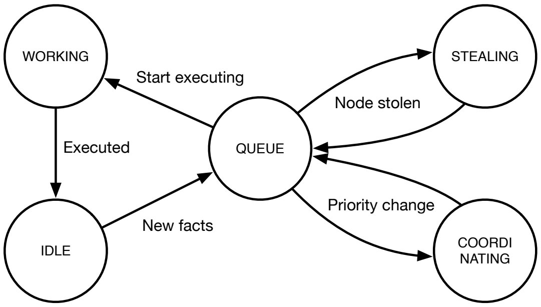


Figure 3: The node state machine. During the lifetime of a program, each node goes through different states as specified by the state machine.

Each node is protected by a main spinlock that allows threads to change node attributes: incoming facts, owner thread, node state and locality information. There is also an internal spinlock that is locked whenever the node enters into the **working** state. If a node sends facts to node is placed in another thread, the current thread first attempts to lock the internal lock in order to update the indexing structures of the target node, otherwise it adds the facts to the list of incoming facts that are later processed by the owner thread.

5.2.2 Threads

Threads pop active nodes from their queues and execute new candidate rules. In order to allow node priorities, we use 2 queues per thread: a doubly linked list known as the *standard queue* and a min/max heap known as the *priority queue*. The regular queue contains nodes without priorities and implements the operations `push_tail(node)` (push to the tail), `pop_head()` (remove node from the head), `remove(node)` (remove an arbitrary node) and `remove_half()` (removes first half). The priority queue contains nodes with priorities and is implemented as a binary heap array. It supports the operations `push(node,priority)` (add new node to the heap), `pop_min()` (remove the best node), `remove(node)` (remove an arbitrary node), `pop_half_min()` (remove the best half) and `move(node,new_priority)` (update priority). Operations such as `remove_half()` are supported in order to support node stealing, while operations `remove(node)` or `move(node,new_priority)` allows threads to change the priority of nodes.

The `next` and `prev` pointers of the regular queue are part of the node structure in order to save space. These pointers are also used as the index in the priority queue and current priority, respectively.

Both the regular and priority queue are actually implemented as 2 queues each. The *static queue* contains nodes that are local to a thread (and thus cannot be stolen) and the *dynamic queue* maybe accessed by other threads during node stealing.

5.2.3 Communication

Threads synchronize with each other using mutual exclusion. We use a spinlock in each queue to protect queue operations and one spinlock to protect the thread's state. Given threads T_1 and T_2, we enumerate the most important synchronization intensive places in the virtual machine:

New facts: When a node executes on T_1 and derives facts to a node in T_2, T_1 first buffers the facts and then sends them to the target node. Here, it checks if the node is currently **idle** and then synchronizes with T_2 to add the node to the T_2's queue.

Thread activation: If T_2 is inactive when adding facts to a node in T_2, T_1 also synchronizes with T_2 to change T_2's state to *active*.

Node stealing: T_1 synchronizes with T_2 when it attempts to steal nodes from T_2 by removing half of the nodes from T_2's queues.

Coordination: If T_1 needs to perform coordination operations to a node in T_2, it may need to sychronize with T_2 during priority updates in order to move the node in T_2's queues.

5.2.4 Termination

There is a global atomic counter, a global boolean flag and one boolean flag (active/idle) per thread for detecting termination. Once a thread goes idle, it decrements the global counter and changes its flag to idle. If the counter goes to zero, the global flag is set to idle. Since every thread will be busy-waiting and checking the global flag, they will detect the change and exit the program.

5.2.5 Node collection

Whenever a node has an empty database and no references from other nodes, it is deleted from the graph. The virtual machine detects such cases by keeping a count of fact references to every node. Once the count drops to zero and the database is empty, the node is immediatelly deleted in order to reduce memory usage.

6. Applications

To better understand how coordination facts are used, we next present some programs that take advantage of them. In our experimental setup, we we used a machine with four AMD Six-Core Opteron TM 8425 HE (2100 MHz) chips (24 cores) and 64 GB of DDR-2 667MHz (16x4GB) RAM, running GNU/Linux (kernel 3.15.10-201 64 bits). We compiled our virtual machine using GCC 4.8.3 (g++) with the flags `-O3 -std=c++11 -fno-rtti -march=x86-64` [1].

6.1 Single Source Shortest Path

The Single Source Shortest Path (SSSP) program was already presented in Section 3. We are going to update the code in Fig. 1 so that nodes with the shortest distances are selected to run before others nodes. The coordinated version of the same algorithm is presented in Fig. 4 and it uses one coordination fact, namely `set-priority` (line 15). Note that we also use a global program directive to order priorities in ascending order (line 5).

When using a single thread, the algorithm behaves like Dijkstra's shortest path algorithm [Dijkstra 1959]. However, when using multiple threads, each thread will pick the smallest distance from their subset of nodes. While this is not optimal since threads do not share a global view of priorities, many rule derivations are going to be avoided.

The most interesting property of the SSSP program presented in Fig. 4 is that it remains provably correct, although it applies rules using smarter ordering. The derivation of `set-priority` does not change the behavior of the logical rules and the code remains declarative.

Figure 5 shows experimental results for a variant of the SSSP program that computes the SSSP for many of the nodes of the graph. There are some situations where unecessary facts are propagated because although the shortest distance is selected, sub-optimal distances may be propagated because many SSSP distances are computed at the same time. However, we see a reduction of over 50% in the number of derived facts when using the coordinated version **C** over the regular version **R**. The results also show that the cordinated version using 16 threads is 1.3 times faster than the regular version using 16 threads. In turn, this makes the

[1] Implementation available in `http://github.com/.../meld`

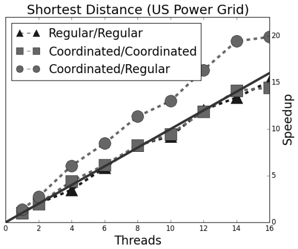
```
type route edge(node, node, int).
type linear shortest(node, int, list int).
type linear relax(node, int, list int).

priority @order asc.

shortest(A, +00, []).
relax(@1, 0, [@1]).

relax(A, D1, P1), shortest(A, D2, P2),
D1 < D2
   -o shortest(A, D1, P1),
      {B, W | !edge(A, B, W) |
         relax(B, D1 + W, P1 ++ [B]),
         set-priority(B, float(D1 + W))}.

relax(A, D1, P1), shortest(A, D2, P2),
D1 >= D2
   -o shortest(A, D2, P2).
```

Figure 4: Shortest Path Program.

coordinated version 20 times faster than the regular version using 1 thread. Note that the coordinated program is as scalable as the regular program.

# T	R	C
1	333K	206K
2	300K	210K
4	316K	208K
8	328K	211K
16	343K	212K

(a)



(b)

Figure 5: Experimental results for the SSSP program using US powergrid network. On the left we show the number of facts derived for the regular **R** and coordinated version **C** using a variable number of threads **# T**. On the right, we have the scalability of the regular and coordinated version. The speedup values are computed using the execution time for 1 thread.

6.2 MiniMax

The MiniMax algorithm is a decision rule algorithm for minimizing the possible loss for a worst case (maximum loss) scenario in a zero sum game for 2 (or more) players that play in turns [Edwards 1954].

The algorithm proceeds by building a game tree, where each tree node represents a game state and the children represent the possible game moves that can be made by either player 1 or player 2. The tree levels are built in a way that each player plays in turns and the first level represents the moves made by player 1. An evaluation function `f(State)` is used to compute the score of the board for each leaf of the tree. A node is a leaf when the game state can no longer be expanded. Finally, the algorithm recursively minimizes or maximizes the scores of each node. To select the best move for player 1, the algorithm picks the move maximized at the root node.

In LM, the program starts with a root node (with the initial game state) that is expanded with the available moves at each level. The graph of the program is dynamic since nodes are created and then deleted once they are no longer needed. The latter happens when the leaf scores are computed or when a node fully minimizes or maximizes the children scores. When the program ends, only the root node has facts in its database.

The LM code in Fig. 6 implements the initial expansion of the tree. The first three rules (lines 1-10) deal with the cases where no children nodes are created and the last three rules (12-29) deal with cases where new nodes may be created. The rule in lines 12-25 generate new nodes using the `exists` language construct, which creates a new node address `B`, where a new `play` fact is created with the new board. A `parent(B,A)` fact is derived to allow node `B` to communicate with its parent.

As noted before in Section 4.1, we use a FIFO ordering for nodes, meaning that when children are expanded, we first expand the children and then the children of the children. We note however, that this is not optimal for this program since the complete tree needs to be expanded before computing the scores at the leaves. The memory complexity of the program is $\mathcal{O}(n)$, where n is the number of nodes in the tree.

```
1   expand(A, Board, [], 0, P, Depth)
2     -o leaf(A, Board).
3
4   expand(A, Board, [], N, P, Depth),
5   N > 0, P = player1
6     -o maximize(A, N, -00, 0).
7
8   expand(A, Board, [], N, P, Depth),
9   N > 0, P = player2
10    -o minimize(A, N, +00, 0).
11
12  expand(A, Board, [0 | Xs], N, P, Depth),
13  Depth >= 5
14    -o exists B. (set-static(B),
15         set-default-priority(B, float(Depth + 1)),
16         play(B, Board ++ [P | Xs], next(P), Depth + 1),
17         expand(A, Board ++ [0], Xs, N + 1, P, Depth),
18         parent(B, A)).
19
20  expand(A, Board, [0 | Xs], N, P, Depth),
21  Depth < 5
22    -o exists B. (set-default-priority(B, float(Depth + 1)),
23         play(B, Board ++ [P | Xs], next(P), Depth + 1),
24         expand(A, Board ++ [0], Xs, N + 1, P, Depth),
25         parent(B, A)).
26
27  expand(A, Board, [C | Xs], N, P, Depth)
28  C <> 0
29    -o expand(A, Board ++ [C], Xs, N, P, Depth).
```

Figure 6: MiniMax: checking if the game has ended and expanding the tree

With coordination, we set the priority nodes to be the same as the depth so that the tree is expanded in a depth-first fashion (lines 15 and 22). The memory complexity of the program then becomes $\mathcal{O}(dt)$, where d is the depth of the tree and t is the number of threads. In Fig. 7, we have two threads (squares) and the first one has expanded the tree on the left side. When a second thread is idle, it steals the first unexpanded node and expands the tree from that position. Since threads prioritize deeper nodes, the scores of the first leaves are immediatelly computed and then sent to the parent node. At this point, the leaves are deleted and reused for other nodes in the tree, resulting in minimal memory usage.

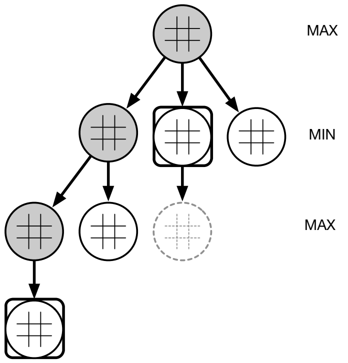
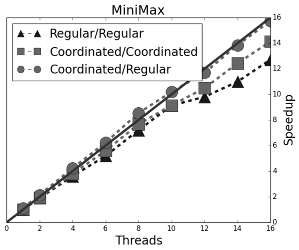


Figure 7: Expanding the MiniMax tree using coordination. By prioritizing deeper nodes, threads are forced to expand the tree using a depth-first approach, which is superior since there is no need to expand the whole tree before computing the node scores.

We also take advantage of memory locality by using `set-static` (line 14), so that nodes after a certain level are not stolen by other threads. While this is not critical for performance in shared memory systems where node stealing is fairly efficient, we expect that such coordination to be critical in distributed systems.

# T	R	C
1	11.80GB	0.50MB
2	12.19GB	1.45MB
4	13.82GB	2.35MB
8	14.87GB	4.36MB
16	13.79GB	8.1MB

(a)



(b)

Figure 8: Memory usage and scalability of the regular and coordinated versions of MiniMax.

In Fig. 8 we compare the memory usage and scalability of the coordinated MiniMax against the regular MiniMax. The coordinated version sees a huge reduction of memory usage, needing at most 8MB for 16 threads, while the regular version needs around 14GB. Note that as the number of threads goes up, memory usage also goes up. This is an artifact of our parallel memory allocator that allocates large chunks of memory beforehand. In terms of run time, our experimental results show a 14-fold speedup for the coordinated version against a 12-fold for the regular version when using 16 threads. When comparing the two versions directly, there is a 20% run time reduction when using 16 threads in the coordinated version.

6.3 Heat Transfer

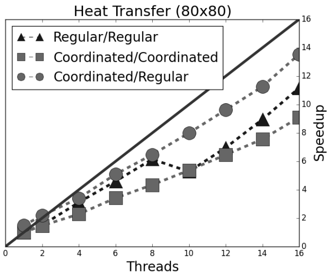
In the Heat Transfer algorithm (HT), we have a graph where heat values are exchanged between nodes. The goal of the program is to exchange heat to a point where new heat value $\delta = |H_i - H_{i-1}| \leq \epsilon$ for all nodes of the graph. The algorithm works asynchronously since heat values can be updated by using new partial information coming from neighbor nodes. This increases parallelism since nodes do not need to synchronize if we had to compute heat values using iterations.

Fig. 9 shows the HT logical rules that send new heat values to neighbor nodes. We added the first rule to increase the priority of the neighbor nodes if the current node has a large δ. The idea is to prioritize the computation of heat values of nodes (using `update`) that have a neighbor node that changed significantly. Multiple `add-priority` facts will increase the priority of a node so that nodes with multiple large deltas will have more priority. Note that the LM compiler will merge the three rules in Fig. 9 into a single one that handles the 3 branches.

```
new-heat(A, New, Old),
Delta = fabs(New - Old),
Delta > epsilon * 2.0
   -o {B, W | !edge(A, B, W) |
         new-neighbor-heat(B, A, New),
         update(B), add-priority(B, Delta)}.

new-heat(A, New, Old),
Delta = fabs(New - Old)
Delta < epsilon * 2.0,
Delta > epsilon
   -o {B, W | !edge(A, B, W) |
         new-neighbor-heat(B, A, New), update(B)}.

new-heat(A, New, Old)
Delta = fabs(New - Old)
Delta < epsilon
   -o {B, W | !edge(A, B, W) |
         new-neighbor-heat(B, A, New)}.
```

Figure 9: Coordination code for the Heat Transfer program. We added one rule that covers cases where Δ is at least 2ϵ in order to increase the priority of neighbors nodes. The code logic remains exactly the same as before, however bigger changes in heat values are now propagated faster.

Fig. 10 presents the scalability results for the uncoordinated and coordinated version. The dataset used is a square grid with an inner square with high heat nodes. With 1 thread, there's a 33% reduction in run time, while for 16 threads, there's a 18% reduction.



(a)

Figure 10: Experimental results for the HT program. The coordinated version is, on average, 30% faster than the regular version although it has a slightly worse scalability due to a reduction in work available.

6.4 Splash Belief Propagation

Randomized and approximation algorithms can obtain significant benefits from coordination directives because although the final program results will not be exact, they follow important statistical properties and can be computed faster. Examples of such programs are Loopy Belief Propagation. Loopy Belief Propagation [Murphy et al. 1999] (LBP) is an approximate inference algorithm used in graphical models with cycles. In its essence, LBP is a sum-product message passing algorithm where nodes exchange messages with their immediate neighbors and apply some computations to the messages received.

LBP is an algorithm that maps very well to the graph based model of LM. In its original form, the belief values of nodes are computed by synchronous iterations. LBP offers more concurrency when belief values are computed asynchronously leading to faster convergence. For this, every node keeps track of all messages sent/received and recomputes the belief using partial information from neighbor nodes. It is then possible to prioritize the computation of beliefs when a neighbor's belief changes significantly.

The asynchronous approach proves to be a nice improvement over the synchronous version. Still, it is possible to do even better. Gonzalez et al [Gonzalez et al. 2009] developed an optimal algorithm to compute this algorithm by first building a tree and then updating the beliefs of each node twice, first from the leaves to the root and then from the root to the leaves. The root of this tree is the node with the highest priority (based on belief) while the other nodes in the tree must have a non-zero priority. Note that the priorities are updated whenever a neighbor updates their belief. These *splash trees* are built iteratively until we reach convergence. In Fig. 11 we represent two threads creating two different splash trees. It is encouraged to create several splash trees concurrently as long as we have threads to create them.

The code for Splash Belief Propagation (SBP) in Fig. 12 presents the coordination code for LBP. Please note that we just appended the code in Fig. 12 to a working but unoptimized version of the algorithm, every other logical rule remains the same. We add new logical ruels that coordinate the creation and execution of the splash trees:

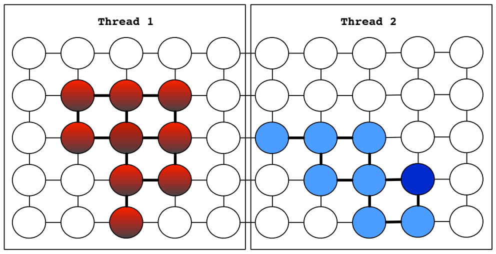


Figure 11: Creating splash trees for belief propagation. Each threads picks the highest priority node and creates a tree from that node. The belief values are updated in two phases: first from the leaves to the root and then from the root to the leaves.

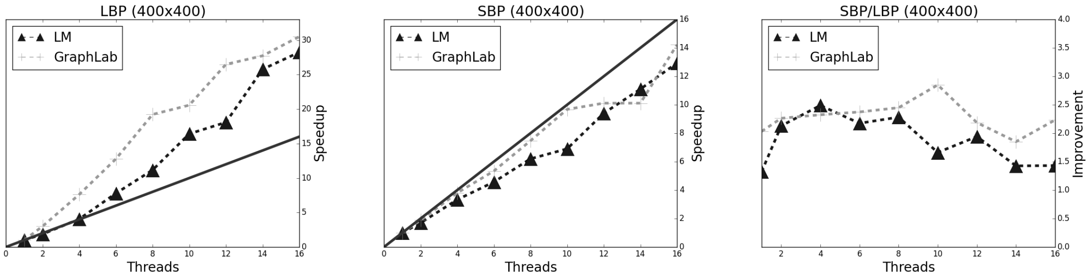
Tree building : Each node has a `waiting` fact that is used to start the tree building process. When the highest priority is picked, a token is created that will navigate through the tree. Note that in the rule located in lines 14-21, newly added nodes to the tree must have a positive priority (due to new belief updates) and must be residing in the same thread. We want the tree to be kept in the same thread in order to maximize parallelism with 1 tree per thread.

First phase : In the third rule (lines 11-12), when the number of nodes of the tree reaches a certain limit, a `first-phase` is generated in order to update the beliefs of all nodes in the tree, starting from the leaves and ending at the root As the nodes are updated, an `update` fact is derived to update the belief values (line 35).

Second phase : In the second phase, the computation of beliefs is performed from the root to the leaves and the belief values are updated a second time (line 42).

Note that the `set-static` and `set-cpu` action fact are used in line 2 in order to (1) force nodes to stay in the thread and (2) to partition nodes as a grid of threads. This setups a well defined area of nodes for threads to build splash trees on.

In this program, coordination assumes a far more important role than we have seen before. Coordination rules fully drive the behavior of the algorithm and although the final result of the algorithm is identical to the original algorithm (minus probabilitistic errors), SBP works in a very different way. A system that also implements SBP is GraphLab [Low et al. 2010], a C++ framework for writing machine algorithms. GraphLab provides the splash scheduler for these types of inferences as part of the framework. This particular scheduler is around 350 lines of complicated C++ code. With



(a)

(b)

(c)

Figure 13: Experimental results for LBP and SBP. Figure (a) shows the scalability of LBP for both LM and GraphLab and Figure (b) shows the scalability of SBP. Figure (c) presents the improvements seen in SBP against LBP, where SBP runs, on average, 1.5 to 2.5 times faster than LBP.

our coordination facts, it is possible to create fairly complicated coordination patterns with only 12 simple logical rules.

We measured the behavior of LBP and SBP for both LM and GraphLab. Fig. 13 shows that both systems have very similar behavior when using a variable number of threads. Note that GraphLab avoids message passing between nodes since new neighbor belief updates are changed immediatelly, resulting in slightly better scalability for GraphLab.

6.5 N Queens

The N-Queens puzzle is the problem of placing N chess queens on an NxN chessboard so that no pair of two queens attack each other [Hoffman et al. 1969]. The challenge of finding all the distinct solutions to this problem is a good benchmark in designing parallel algorithms.

The LM solution considers the chess board as a graph where nodes communicate with each other to build valid configurations. First, we start with the empty state in the first row of the chess board. Then, each square adds its own position to the state and sends the state down, to the next row. Once a square receives new configurations, it attempts to add its position to the configurations. If valid, that configuration is then sent, recursively, to the next row, until all rows are traversed. At the end of the program, the nodes at the bottom row will have all the valid configurations.

Since computation goes from the top row to the bottom row, not all placements of nodes on threads may be equally performant. This is specially true because the bottom rows tend to perform most work. The best placement is then to split the board vertically, so that, in optimal conditions, each thread gets the same number of columns. Our experiments show that, on shared memory, it does not matter much if we start with a bad placement since node stealing overcomes those issues by improving load balancing dynamically. But what if we want run the program on a distributed system, where moving nodes between workers can be expensive? We measured the performance of the program without node stealing in order to assess how important data placement would be for such systems. We used the default partition of the program as it is, which results in the board being split by rows and another configuration using the coordination action `set-cpu(A, vertical(X, Y))` that forces nodes to be moved to another thread using a vertical partitioning. Note that `X` and `Y` are the coordinates of the node in the chessboard.

Figure 14 shows the performance results of coordinating the placing of nodes on threads without work stealing. There is a noticiable performance improvement when using an optimal data placement strategy. Note however that in the plot we do not take into account the costs of moving nodes between workers. In a real system, the performance gap may not be as wide since in the worst case there may be 144 moves.

7. Coordination Overhead

In the paper [Cruz et al. 2014b], the LM virtual machine was measured and compared against implementations of similar programs but written in other languages. For instance, the LBP program was found to be around 2 times slower than GraphLab, while N Queens program was found to be around 10 to 15 slower than a sequential C implementation. Some of those programs were compared against a Python implementations and LM faired fairly well. In this section, we measure the impact of the coordination mechanisms we have implemented for the new virtual machine.

Fig. 15 shows the comparison between the original virtual machine and the new virtual machine with coordination mechanisms. All the programs benchmarked do not use any kind of coordination. As expected, there is no degration of performance by adding coordination mechanisms.

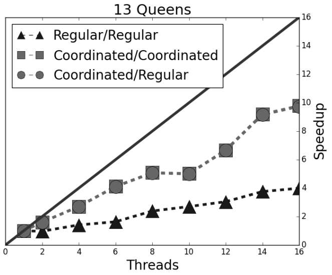
8. Conclusions

We think that our particular approach is applicable in distributed systems as we have seen. Explore functional programming and coordination?

```
!coord(A, X, Y), start(A)
   -o set-static(A), set-cpu(A, grid(X, Y)).

// TREE BUILDING
// continue tree
waiting(A), token(A, All, Next) -o token(A, All, Next).
// start tree
waiting(A), @priority(A, A, P), P > 0.0
   -o token(A, [A], [A]).
// end tree building
token(A, All, Next), length(All) > maxnodes
   -o first-phase(A, All, reverse(All)).
// expand tree
token(A, All, [A | Next])
   -o [collect => L | Side | !edge(A, L, Side),
         0 = count(All, L),
         0 = count(Next, L),
         @priority(A, L, P), P > 0.0,
         @cpu-id(A, L, Id1),
         @cpu-id(A, A, Id2), Id1 = Id2 |
         send-token(A, All, Next ++ L)].

send-token(A, All, [])
   -o first-phase(A, All, reverse(All)).
send-token(A, All, [B | Next])
   -o schedule-next(B),
      token(B, All ++ [B], [B | Next]).

// FIRST PHASE
first-phase(A, [A], [A]) -o second-phase(A, [], A).
first-phase(A, [A, B | Next], [A])
   -o update(A), schedule-next(B),
      second-phase(B, [B | Next], A).
first-phase(A, All, [A, B | Next])
   -o update(A), schedule-next(B),
      first-phase(B, All, [B | Next]).

// SECOND PHASE
second-phase(A, [], _)
   -o set-priority(A, 0.0), waiting(A), update(A).
second-phase(A, [A], Back)
   -o update(A), waiting(Back),
      waiting(A), set-priority(A, 0.0).
second-phase(A, [A, B | Next], Back)
   -o update(A), waiting(Back), schedule-next(B),
      second-phase(B, [B | Next], A).
```

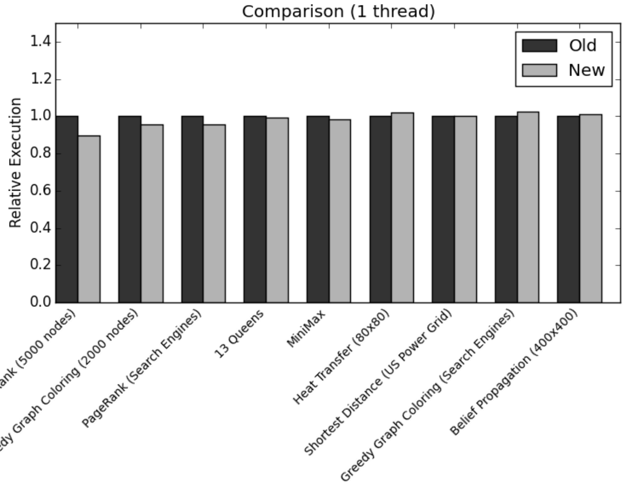
Figure 12: Coordination code for the Splash Belief Propagation program. LM needs 50 lines of rules to implement splash trees, while GraphLab needs a total of 350 lines of C++ to implement the same functionality. Note that when a linear fact is prefixed by @, it indicates that the fact is going to be re-derived.

9. Acknowledgements



(a)

Figure 14: Experimental results for the 13 Queens program. The coordinated version is on average 2 times faster than the version with bad data placement.



(a)

Figure 15: Measuring the overhead of coordination mechanisms. The **Old** bars represent the performance of the original virtual machine, while **New** is the relative performance of the new version that includes coordination mechanisms.

References

S. Ahuja, N. Carriero, and D. Gelernter. Linda and friends. *Computer*, 19(8):26–34, 1986.

P. Alvaro, T. Condie, N. Conway, K. Elmeleegy, J. M. Hellerstein, and R. Sears. Boom analytics: exploring data-centric, declarative programming for the cloud. In *European Conference on Computer Systems (EuroSys)*, pages 223–236, 2010.

D. Chu, L. Popa, A. Tavakoli, J. M. Hellerstein, P. Levis, S. Shenker, and I. Stoica. The design and implementation of a declarative sensor network system. In *International Conference on Embedded Networked Sensor Systems (SenSys)*, pages 175–188, 2007.

F. Cruz, R. Rocha, S. Goldstein, and F. Pfenning. A Linear Logic Programming Language for Concurrent Programming over Graph Structures. *Journal of Theory and Practice of Logic Programming, 30th International Conference on Logic Programming (ICLP 2014), Special Issue*, pages 493–507, July 2014a.

F. Cruz, R. Rocha, and S. C. Goldstein. Design and Implementation of a Multithreaded Virtual Machine for Executing Linear Logic Programs. In O. Danvy, editor, *Proceedings of the International Symposium on Principles and Practice of Declarative Programming (PPDP 2014)*, pages 43–53, Canterbury, UK, September 2014b. ACM Press.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51:107–113, 2008.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

W. Edwards. The theory of decision making. *Psychological Bulletin*, 41:380–417, 1954.

J.-Y. Girard. Linear logic. *Theoretical Computer Science*, 50(1): 1–102, 1987.

J. Gonzalez, Y. Low, and C. Guestrin. Residual splash for optimally parallelizing belief propagation. In *Artificial Intelligence and Statistics (AISTATS)*, 2009.

E. J. Hoffman, J. C. Loessi, and R. C. Moore. Construction for the solutions of the M queens problem. *Mathematics Magazine*, 42 (2):66–72, 1969.

M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *European Conference on Computer Systems (EuroSys)*, pages 59–72, 2007.

B. T. Loo, T. Condie, M. Garofalakis, D. E. Gay, and J. M. Hellerstein. Declarative networking: Language, execution and optimization. In *International Conference on Management of Data (SIGMOD)*, pages 97–108, 2006.

Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new framework for parallel machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 340–349, 2010.

S. Lucco and O. Sharp. Delirium: an embedding coordination language. In *Supercomputing '90., Proceedings of*, pages 515–524, Nov 1990.

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 467–475, 1999.

C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: A not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1099–1110, New York, NY, USA, 2008. ACM.

G. A. Papadopoulos and F. Arbab. Coordination models and languages. In *Advances in Computers*, pages 329–400, 1998.

R. Ramakrishnan and J. D. Ullman. A survey of research on deductive database systems. *Journal of Logic Programming*, 23: 125–149, 1993.

J. D. Ullman. *Principles of Database and Knowledge-Base Systems: Volume II: The New Technologies*. 1990.