On-the-Fly Pipeline Parallelism

I-Ting Angelina Lee* Charles E. Leiserson* Tao B. Schardl† Jim Sukha† Zhunping Zhang‡

*MIT CSAIL
32 Vassar Street
Cambridge, MA 02139
{angelee, cel, neboat, jzz}@mit.edu

†Intel Corporation
25 Manchester Street, Suite 200
Merrimack, NH 03054
jim.sukha@intel.com

ABSTRACT

Pipeline parallelism organizes a parallel program as a linear sequence of $s$ stages. Each stage processes elements of a data stream, passing each processed data element to the next stage, and then taking on a new element before the subsequent stages have necessarily completed their processing. Pipeline parallelism is used especially in streaming applications that perform video, audio, and digital signal processing. Three out of 13 benchmarks in PARSEC, a popular software benchmark suite designed for shared-memory multiprocessors, can be expressed as pipeline parallelism.

Whereas most concurrency platforms that support pipeline parallelism use a “construct-and-run” approach, this paper investigates “on-the-fly” pipeline parallelism, where the structure of the pipeline emerges as the program executes rather than being specified a priori. On-the-fly pipeline parallelism allows the number of stages to vary from iteration to iteration and dependencies to be data dependent. We propose simple linguistics for specifying on-the-fly pipeline parallelism and describe a provably efficient scheduling algorithm, the PIPER algorithm, which integrates pipeline parallelism into a work-stealing scheduler, allowing pipeline and fork-join parallelism to be arbitrarily nested. The PIPER algorithm automatically throttles the parallelism, precluding “runaway” pipelines. Given a pipeline computation with $T_1$ work and $T_{\text{run}}$ span (critical-path length), PIPER executes the computation on $P$ processors in $T_P \leq T_1 + O(T_{\text{run}} + \log P)$ expected time. PIPER also limits stack space, ensuring that it does not grow unboundedly with running time.

We have incorporated on-the-fly pipeline parallelism into a Cilk-based work-stealing runtime system. Our prototype Cilk-P implementation exploits optimizations such as lazy enabling and dependency folding. We have ported the three PARSEC benchmarks that exhibit pipeline parallelism to run on Cilk-P. One of these, x264, cannot readily be executed by systems that support only construct-and-run pipeline parallelism. Benchmark results indicate that Cilk-P has low serial overhead and good scalability. On x264, for example, Cilk-P exhibits a speedup of 13.87 over its respective serial counterpart when running on 16 processors.

Categories and Subject Descriptors
D.3.3 [Language Constructs and Features]: Concurrent programming structures; D.3.4 [Programming Languages]: Processors—Run-time environments.

General Terms
Algorithms, Languages, Theory.

Keywords
Cilk, multithreading, parallel programming, pipeline parallelism, on-the-fly pipelining, scheduling, work stealing.

1. INTRODUCTION

Pipeline parallelism1 [6, 16, 17, 25, 27, 28, 31, 33, 35, 37] is a well-known parallel-programming pattern that can be used to parallelize a variety of applications, including streaming applications from the domains of video, audio, and digital signal processing. Many applications, including the ferret, dedup, and x264 benchmarks from the PARSEC benchmark suite [4, 5], exhibit parallelism in the form of a linear pipeline, where a linear sequence $S = \langle S_0, \ldots, S_{n-1} \rangle$ of abstract functions, called stages, are executed on an input stream $I = \langle a_0, a_1, \ldots, a_{n-1} \rangle$. Conceptually, a linear pipeline can be thought of as a loop over the elements of $I$, where each loop iteration $i$ processes an element $a_i$ of the input stream. The loop body encodes the sequence $S$ of stages through which each element is processed. Parallelism arises in linear pipelines because the execution of iterations can overlap in time, that is, iteration $i$ may start after the preceding iteration $i-1$ has started, but before $i-1$ has necessarily completed.

Most systems that provide pipeline parallelism employ a construct-and-run model, as exemplified by the pipeline model in Intel Threading Building Blocks (TBB) [27], where the pipeline stages and their dependencies are defined a priori before execution. Systems that support construct-and-run pipeline parallelism include the following: [1, 11, 17, 25–27, 29–32, 35, 37, 38].

We have extended the Cilk parallel-programming model [15, 20, 24] to augment its native fork-join parallelism with on-the-fly pipeline parallelism, where the linear pipeline is constructed dynamically as the program executes. The Cilk-P system provides a flexible linguistic model for pipelining that allows the structure of the pipeline to be determined dynamically as a function of data in the input stream. Cilk-P also admits a variable number of stages across iterations, allowing the pipeline to take on shapes other than

---

1Pipeline parallelism should not be confused with instruction pipelining in hardware [34] or software pipelining [22].
and indicates that between corresponding stages of adjacent iterations are optional.

cross edges, it is a serial stage. The edges between nodes denote dependencies of the stage edge. A

careful analysis of the model leads to a natural theoretical question, how much parallelism is inherent in

A natural theoretical question is, how much parallelism is inherent in the ferret pipeline (or in any pipeline)? How much speedup can one hope for? Since the computation is represented as a dag \( G = (V, E) \), one can use a simple work/span analysis \([12, \text{Ch. 27}]\) to answer this question. In this analytical model, we assume that each vertex \( v \in V \) executes in some time \( w(v) \). The work of the computation, denoted \( T_1 \), is essentially the serial execution time, that is, \( T_1 = \sum_{v \in V} w(v) \). The span of the computation, denoted \( T_{\text{span}} \), is the length of a longest weighted path through \( G \), which is essentially the time of an infinite-processor execution. The parallelism is the ratio \( \frac{T_1}{T_{\text{span}}} \), which is the maximum possible speedup attainable on any number of processors, using any scheduler.

Unlike in other applications, in the ferret pipeline, each node executes serially, that is, its work and span are the same. Let \( w(i,j) \) be the execution time of node \((i,j)\). Assume that the serial Stages 0 and 2 execute in unit time, that is, for all \( i \), we have \( w(i,0) = w(i,2) = 1 \), and that the parallel Stage 1 executes in time \( r \gg 1 \), that is, for all \( i \), we have \( w(i,1) = r \). Because the pipeline dag is grid-like, the span of this SPS pipeline can be realized by some staircase walk through the dag from node \((0,0)\) to node \((n−1,2)\). The work of this pipeline is therefore \( T_1 = r(n+2) \), and the span is

\[
T_{\text{span}} = \max_{0 \leq x < n} \left\{ \sum_{i=0}^{r} w(i,0) + w(x,1) + \sum_{i=r}^{n-1} w(i,2) \right\} = n + r. 
\]

Consequently, the parallelism of this dag is \( \frac{T_1}{T_{\text{span}}} = \frac{n(r+2)}{n+r} \), which for \( 1 \ll r \ll n \) is at least \( r/2+1 \). Thus, if Stage 1 contains much more work than the other two stages, the pipeline exhibits good parallelism.

Cilk-P guarantees to execute the ferret pipeline efficiently. In particular, on an ideal shared-memory computer with up to \( P \) processors, Cilk-P guarantees linear speedup. Generally, Cilk-P executes a pipeline with linear speedup as long as the parallelism of the pipeline exceeds the number of processors on which the computation is scheduled. Moreover, as Section 3 will describe, Cilk-P allows stages of the pipeline to be parallel using recursive pipelining or fork-join parallelism.

In practice, it is also important to limit the space used during an execution. Unbounded space can cause thrashing of the memory system, leading to slowdowns not predicted by simple execution models. In particular, a bind-to-element scheduler must avoid creating a runaway pipeline — a situation where the scheduler allows many new iterations to be started before finishing old ones. In Figure 1, a runaway pipeline might correspond to executing many nodes in Stage 0 (the top row) without finishing the other stages of the computation in the earlier iterations. Runaway pipelines can cause space utilization to grow unboundedly, since every started but incomplete iteration requires space to store local variables.

Cilk-P automatically throttles pipelines to avoid runaway pipelines. On a system with \( P \) workers, Cilk-P inhibits the start of iteration \( i+K \) until iteration \( i \) has completed, where \( K = \Theta(P) \) is the throttling limit. Throttling corresponds to putting "throttling edges" from the last node in each iteration \( i \) to the first node in iterations.
tion $i + K$. For the simple pipeline from Figure 1, throttling does not adversely affect asymptotic scalability if stages are uniform, but it can be a concern for more complex pipelines, as Section 11 will discuss. The Cilk-P scheduler guarantees efficient scheduling of pipelines as a function of the parallelism of the dag in which throttling edges are included in the calculation of span.

Contributions

Our prototype Cilk-P system adapts the Cilk-M [23] work-stealing scheduler to support on-the-fly pipeline parallelism using a bind-to-element approach. This paper makes the following contributions:

- We describe linguistics for Cilk-P that allow on-the-fly pipeline parallelism to be incorporated into the Cilk fork-join parallel programming model (Section 2).
- We illustrate how Cilk-P linguistics can be used to express the x264 benchmark as a pipeline program (Section 3).
- We characterize the execution dag of a Cilk-P pipeline program as an extension of a fork-join program (Section 4).
- We introduce the PIPER scheduling algorithm, a theoretically sound randomized work-stealing scheduler (Section 5).
- We prove that PIPER is asymptotically efficient, executing Cilk-P programs on $P$ processors uses $S_P \leq P(S_1 + fDK)$ stack space for pipeline iterations, where $S_1$ is the serial stack space, $f$ is the “frame size,” $D$ is the depth of nested pipelines, and $K$ is the throttling limit (Sections 6 and 7).
- We bound space usage, proving that $P_{\text{IPER}} \leq P + O(T_0 + \log P)$ expected time (Sections 6 and 7).
- We introduce the P pipe_while construct. This paper makes the following contributions:
  - We demonstrate that the pipe_while keyword. A pipe_while loop is similar to a serial while loop, except that loop iterations can execute in parallel in a pipelined fashion. The body of the pipe_while can be subdivided into stages, with stages named by user-specified integer values that strictly increase as the iteration executes. Each stage can contain nested fork-join and pipeline parallelism.

To support on-the-fly pipeline parallelism, Cilk-P provides a pipe_while keyword. A pipe_while loop is similar to a serial while loop, except that loop iterations can execute in parallel in a pipelined fashion. The body of the pipe_while can be subdivided into stages, with stages named by user-specified integer values that strictly increase as the iteration executes. Each stage can contain nested fork-join and pipeline parallelism.

The boundaries of stages are denoted in the body of a pipe_while using the special functions pipe_continue and pipe_wait. These functions accept an integer stage argument, which is the number of the next stage to execute and which must strictly increase during the execution of an iteration. Every iteration $i$ begins executing Stage 0, represented by node $(i,0)$. While executing a node $(i,j)$, if control flow encounters a pipe_wait($j$) or pipe_continue($j$) statement, where $j > j'$, then node $(i,j')$ ends, and control flow proceeds to node $(i,j)$. A pipe_continue($j$) statement indicates that node $(i,j)$ can start executing immediately, whereas a pipe_wait($j$) statement indicates that node $(i,j)$ cannot start until node $(i-1,j)$ completes. The pipe_wait($j$) in iteration $i$ creates a cross edge from node $(i-1,j)$ to node $(i,j)$ in the pipeline dag. Thus, by design choice, Cilk-P imposes the restriction that pipeline dependencies only go between adjacent iterations. As we shall see in Section 9, this design choice facilitates the “lazy enabling” and “dependency folding” runtime optimizations.

The pipe_continue and pipe_wait functions can be used without an explicit stage argument. Omitting the stage argument while executing stage $j$ corresponds to an implicit stage argument of $j+1$, i.e., control moves onto the next stage.

Cilk-P’s semantics for pipe_continue and pipe_wait statements allow for stage skipping, where execution in an iteration $i$ can jump stages from node $(i,j')$ to node $(i,j)$, even if $j > j' + 1$. If control flow in iteration $i+1$ enters node $(i+1,j')$ after a pipe_wait, where $j' < j'' < j$, then we implicitly create a null node $(i,j'')$ in the pipeline dag, which has no associated work and incurs no scheduling overhead, and insert stage edges from $(i,j')$ to $(i,j'')$ and from $(i,j'')$ to $(i,j)$, as well as a cross edge from $(i,j'')$ to $(i+1,j'')$.

3. ON-THE-FLY PIPELINING OF x264

To illustrate the use of Cilk-P’s pipe_while loop, this section describes how to parallelize the x264 video encoder [39].

We begin with a simplified description of x264. Given a stream $(f_0, f_1, \ldots)$ of video frames to encode, x264 partitions the frame into two dimensional arrays of “macroblocks” and encodes each macroblock. A macroblock in frame $f_i$ is encoded as a function of the encodings of similar macroblocks within $f_i$ and similar macroblocks in frames “near” $f_i$. A frame $f_j$ is near a frame $f_i$ if $i - b \leq j \leq i + b$ for some constant $b$. In addition, we define a macroblock $(x', y')$ to be near a macroblock $(x, y)$ if $x - w \leq x' \leq x + w$ and $y - w \leq y' \leq y + w$ for some constant $w$.

The type of a frame $f_i$ determines how a macroblock $(x, y)$ in $f_i$ is encoded. If $f_i$ is an I-frame, then macroblock $(x, y)$ can be encoded using only previous macroblocks within $f_i$ — macroblocks at positions $(x', y')$ where $y' < y$ or $y' = y$ and $x' < x$. If $f_i$ is a P-frame, then macroblock $(x, y)$’s encoding can also be based on nearby macroblocks in nearby preceding frames, up to the most recent preceding I-frame, if one exists within the nearby range. If $f_i$ is a B-frame, then macroblock $(x, y)$’s encoding can also be based on nearby macroblocks in nearby frames, likewise, up to the most recently preceding I-frame and up to the next succeeding I- or P-frame.

\footnote{To be precise, up to a particular type of I-frame called an IDR-frame.}
Based on these frame types, an x264 encoder must ensure that frames are processed in a valid order such that dependencies between encoded macroblocks are satisfied. A parallel x264 encoder can pipeline the encoding of I- and P-frames in the input stream, processing each set of intervening B-frames after encoding the latest I- or P-frame on which the B-frame may depend.

Figure 2 shows pseudocode for an x264 linear pipeline. Conceptually, the x264 pipeline begins with a serial stage (lines 8–17) that reads frames from the input stream and determines the type of each frame. This stage buffers all B-frames at the head of the input stream until it encounters an I- or P-frame. After this initial stage, x264 hybrid stages process this I- or P-frame row by row (lines 18–25), where \( s \) is the number of rows in the video frame. After all rows of this I- or P-frame have been processed, the PROCESS_BFRAMES stage processes all B-frames in parallel (lines 26–29), and then the END stage updates the output stream with the processed frames (line 31).

Two issues arise with this general pipelining strategy, both of which can be handled using on-the-fly pipeline parallelism. First, the encoding of a P-frame must wait for the encoding of rows in the previous frame to be completed, whereas the encoding of an I-frame need not. These conditional dependencies are implemented in lines 20–24 of Figure 2 by executing a pipe_continue or pipe_wait statement conditionally based on the frame’s type. In contrast, many construct-and-run pipeline mechanisms assume that the dependencies on a stage are fixed for the entirety of a pipeline’s execution, making such dynamic dependencies more difficult to handle. Second, the encoding of a macroblock in row \( x \) of P-frame \( f \) may depend on the encoding of a macroblock in a later row \( x + w \) in the preceding I- or P-frame \( f_{i-1} \). The code in Figure 2 handles such offset dependencies on line 17 by skipping \( w \) additional stages relative to the previous iteration. A similar stage-skipping trick is used on line 26 to ensure that the processing of a P-frame in iteration \( i \) depends only on the processing of the previous I- or P-frame, and not on the processing of preceding B-frames.

Figure 3 illustrates the pipeline dag corresponding to the execution of the code in Figure 2, assuming that \( w = 1 \). Skipping stages shifts the nodes of an iteration down, adding null nodes to the pipeline, which do not increase the work or span.

### 4. COMPUTATION-DAG MODEL

Although the pipeline-dag model provides intuition for programmers to understand the execution of a pipeline program, it is not as precise as we shall require. For example, a pipeline dag has no real way of representing nested fork-join or pipeline parallelism within a node. This section describes how to represent the execution of a pipeline program as a more refined “computation dag.”

Let us first review the notion of a computation dag for ordinary fork-join Cilk programs [7, 8] without pipeline parallelism. A fork-join computation dag \( G = (V, E) \) represents the execution of a Cilk program, where the vertices belonging to \( V \) are unit-cost instructions. Edges in \( E \) indicate ordering dependencies between instructions. The normal serial execution of one instruction after another creates a serial edge from the first instruction to the next.

A cilk_spawn of a function creates two dependency edges emanating from the instruction immediately before the cilk_spawn: the spawn edge goes to the first instruction of the spawned function, and the continue edge goes to the first instruction after the spawned function. A cilk_sync creates a return edge from the final instruction of each spawned function to the instruction immediately after the cilk_sync (as well as an ordinary serial edge from the instruction that executed immediately before the cilk_sync).

To model an arbitrary pipeline-program execution as a (pipeline) computation dag, we follow a three-step process. First, we translate the code executed in each pipe_while loop into ordinary Cilk code augmented with special functions to handle cross and throttling dependencies. Second, we model the execution of this augmented Cilk program as a fork-join computation dag. Third, we show how to augment the fork-join computation dag with cross and throttling edges using the special functions.

The first step of this process does not reflect how a Cilk-P com-

---

Figure 2: Example C++-like pseudocode for the x264 linear pipeline. This pseudocode uses Cilk-P’s dualistics to define hybrid pipeline stages on the fly, specifically with the pipe_wait on line 17, the input-data dependent pipe_continue or pipe_wait on lines 20–24, and the pipe_continue on line 26.

```c++
// Symbolic names for important stages
const uint64_t PROCESS_IPFRAME = 1;
const uint64_t PROCESS_BFRAMES = 1 << 4;
const uint64_t END = PROCESS_BFRAMES + 1;
int i = 0;
int w = mv_range / pixel_per_row;

pipe_while(frame_t *f = next_frame()) {
    vector<frame_t *> bframes;
    f->type = decide_frame_type(f);
    write_out_frames(frame, bframes);
}
```

---

Figure 3: The pipeline dag generated for x264. Each iteration processes either an I- or P-frame, each consisting of \( s \) rows. As the iteration index \( i \) increases, the number of initial stages skipped in the iteration also increases. This stage skipping produces cross edges into an iteration \( i \) from null nodes in iteration \( i-1 \). Null nodes are represented as the intersection between two edges.

---

In contrast, many construct-and-run pipeline mechanisms assume that the dependencies on a stage are fixed for the entirety of a pipeline’s execution, making such dynamic dependencies more difficult to handle. Second, the encoding of a macroblock in row \( x \) of P-frame \( f \) may depend on the encoding of a macroblock in a later row \( x + w \) in the preceding I- or P-frame \( f_{i-1} \). The code in Figure 2 handles such offset dependencies on line 17 by skipping \( w \) additional stages relative to the previous iteration. A similar stage-skipping trick is used on line 26 to ensure that the processing of a P-frame in iteration \( i \) depends only on the processing of the previous I- or P-frame, and not on the processing of preceding B-frames.

Figure 3 illustrates the pipeline dag corresponding to the execution of the code in Figure 2, assuming that \( w = 1 \). Skipping stages shifts the nodes of an iteration down, adding null nodes to the pipeline, which do not increase the work or span.
1 int fd_out = open_output_file();
2 bool done = false;
3 pipe_while (!done) {
4   chunk_t *chunk = get_next_chunk();
5   if (chunk == NULL)
6      done = true;
7  }
8  else {
9    pipe_wait();
10   bool isDuplicate = deduplicate(chunk);
11   pipe_continue();
12   compress(chunk);
13   write_to_file(fd_out, chunk);
14 }
15 }

Figure 4: Cilk-P pseudocode for the parallelization of the dedup compression program as an SSPS pipeline.

The second step models the execution of this Cilk-P program as a fork-join computation dag by augmenting the fork-join computation dag with cross and throttling edges based on the special functions pipe_wait, pipe_continue, and throttle special functions. The unbound variable \( K \) is the throttling limit.

As the first step in building the computation dag for an execution of this Cilk-P program, we transform the code executed from running the code in Figure 4 into the ordinary Cilk program shown in Figure 5. As shown in lines 3–32, a pipe while is “lifted” using a C++ lambda function [36, Sec.11.4] and converted into an ordinary while loop using the variable \( i \) to index iterations. The loop body executes Stage 0 and spawns off a C++ lambda function that executes the remainder of the iteration (line 12). As one can see from this transformation, Stage 0 of a pipe while loop is always a serial stage and the test condition of the pipe while loop is considered part of Stage 0. These constraints guarantee that the repeated tests of the pipe while loop-termination condition execute serially. Each stage ends with a cilk sync (lines 10, 16, 21, and 25.)

The last statement in the loop (line 29) is a call to a special function throttle, which implements the throttling dependency. The cilk sync immediately after the end of the while loop (line 31) waits for completion of the spawned iterations.

The second step models the execution of this Cilk program as a fork-join computation dag \( G = (V,E) \) as in [7,8].

The third step is to produce the final pipeline computation dag by augmenting the fork-join computation dag with cross and throttling edges based on the special functions pipe_wait, pipe_continue, and throttle. For example, when iteration \( i \) executes the pipe_wait call in line 22, it specifies the start of

5. THE PIPER SCHEDULER

Piper executes a pipeline program on a set of \( P \) workers using work-stealing. For the most part, Piper’s execution model can be viewed as modification of the scheduler described by Arora, Blumofe, and Plaxton [2] (henceforth referred to as the ABP model) for computation daggs arising from pipeline programs. Piper deviates from the ABP model in one significant way, however, in that it performs a "tail-swap" operation.

We describe the operation of Piper in terms of the pipeline computation dag \( G = (V,E) \). Each worker \( p \) in Piper maintains an assigned vertex corresponding to the instruction that \( p \) executes on the current time step. We say that a vertex \( x \) is ready if all its predecessors have been executed. Executing an assigned vertex \( v \) may enable a vertex \( y \) that is a direct successor of \( v \) in \( G \) by making \( y \) ready. Each worker maintains a deque of ready vertices. Normally, a worker pushes and pops vertices from the tail of its deque. A "thief," however, may try to steal a vertex from the head of another worker’s deque. It is convenient to define the extended deques \( \langle v_0, v_1, \ldots, v_r \rangle \) of a worker \( p \), where \( v_0 \in V \) is \( p \)’s assigned vertex and \( v_1, v_2, \ldots, v_r \in V \) are the vertices in \( p \)’s deque in order from tail to head.

On each time step, each Piper worker \( p \) follows a few sim-
ple rules for execution based on the type of \( p \)'s assigned vertex \( v \) and how many direct successors are enabled by the execution of \( v \), which is at most 2. (Although \( v \) may have multiple immediate successors in the next iteration due to the collapsing of null nodes, executing \( v \) can enable at most one such vertex, since the stages in the next iteration execute serially.) We assume that the rules are executed atomically.

If the assigned vertex \( v \) is not the last vertex of an iteration and its execution enables only one direct successor \( x \), then \( p \) simply changes its assigned vertex from \( v \) to \( x \). Executing a vertex \( v \) can enable two successors if \( v \) spawns a child \( x \) with continuation \( y \) or if \( v \) is the last vertex in a node in iteration \( i \), which enables both the first vertex \( x \) of the next node in \( i \) and the first vertex \( y \) in a node in iteration \( i + 1 \). In either case, \( p \) pushes \( y \) onto the tail of its deque and changes its assigned vertex from \( v \) to \( x \). If executing \( v \) enables no successors, then \( p \) tries to pop an element \( z \) from the tail of its deque, changing its assigned vertex from \( v \) to \( z \). If \( p \)'s deque is empty, \( p \) becomes a thief. As a thief, \( p \) randomly picks another worker to be its victim, tries to steal the vertex \( z \) at the head of the victim's deque, and sets the assigned vertex of \( p \) to \( z \) if successful. These cases are consistent with the normal ABP model.

Piper handles the end of an iteration differently, however, due to throttling edges. Suppose that a worker \( p \) has an assigned vertex \( v \) representing the last vertex in a given iteration in a given pipe_while loop, and suppose that the edge leaving \( v \) is a throttling edge to a vertex \( z \). When \( p \) executes \( v \), two cases are possible. In the first case, executing \( v \) does not enable \( z \), in which case no new vertices are enabled, and \( p \) acts accordingly. In the second case, however, executing \( v \) does enable \( z \), in which case \( p \) performs two actions. First, \( p \) changes its assigned vertex from \( v \) to \( z \). Second, if \( p \) has a nonempty deque, then \( p \) performs a tail swap: it exchanges its assigned vertex \( z \) with the vertex at the tail of its deque.

This tail-swap operation is designed to reduce Piper's space usage. Without the tail swap, in a normal ABP-style execution, when a worker \( p \) finishes an iteration \( i \) that enables a vertex via a throttling edge, \( p \) would conceptually choose to start a new iteration \( i + K \), even if iteration \( i + 1 \) were already suspended and on its deque. With the tail swap, \( p \) resumes iteration \( i + 1 \), leaving \( i + K \) available for stealing. The tail swap also enhances cache locality by encouraging \( p \) to execute consecutive iterations.

It may seem, at first glance, that a tail-swap operation might significantly reduce the parallelism, since the vertex \( z \) enabled by the throttling edge is pushed onto the bottom of the deque. Intuitively, if there were additional work above \( z \) in the deque, then a tail swap could significantly delay the start of iteration \( i + K \). Lemma 4 will show, however, that a tail-swap operation only occurs on deques with exactly 1 element. Thus, whenever a tail swap occurs, \( z \) is at the top of the deque and is immediately available to be stolen.

6. STRUCTURAL INVARIANTS

During the execution of a pipeline program by Piper, the worker deques satisfy two structural invariants, called the "contour" property and the "depth" property. This section states and proves these invariants.

Intuitively, we would like to describe the structure of the worker deques in terms of frames — activation records — of functions' local variables, since the deques implement a "cactus stack" [18, 23]. A pipe_while loop would correspond to a parent frame with a spawned child for each iteration. Although the actual Cilk-P implementation manages frames in this fashion, the control of a pipe_while really does follow the schema illustrated in Figure 5, where Stage 0 of an iteration executes in the same lambda function as the parent, rather than in the child lambda function which contains the rest of \( i \). Consequently, we introduce "contours" to represent this structure.

Consider a computation dag \( G = (V, E) \) that arises from the execution of a pipeline program. A contour is a path in \( G \) composed only of serial and continue edges. A contour must be a path, because there can be at most one serial or continue edge entering or leaving any vertex. We call the first vertex of a contour the root of the contour, which (except for the initial instruction of the entire computation) is the only vertex in the contour that has an incoming spawn edge. Consequently, contours can be organized into a tree hierarchy, where one contour is a parent of a second if the first contour contains a vertex that spawns the root of the second. Given a vertex \( v \in V \), let \( c(v) \) denote the contour to which \( v \) belongs.

The following two lemmas describe two important properties exhibited in the execution of a pipeline program.

**Lemma 1.** Only one vertex in a contour can belong to any extended deque at any time.

**Proof.** The vertices in a contour form a chain and are, therefore, enabled serially.

The structure of a pipe_while guarantees that the "top-level" vertices of each iteration correspond to a contour, and that all iterations of the pipe_while share a common parent in the contour tree. These properties lead to the following lemma.

**Lemma 2.** If an edge \( (x, y) \) is a cross edge, then \( c(x) \) and \( c(y) \) are siblings in the contour tree and correspond to adjacent iterations in a pipe_while loop. If an edge \( (x, y) \) is a throttling edge, then \( c(y) \) is the parent of \( c(x) \) in contour tree.

As Piper executes a pipeline program, the deques of workers are highly structured with respect to contours.

**Definition 3.** At any time during an execution of a pipeline program which produces a computation dag \( G = (V, E) \), consider the extended deque \( \langle v_0, v_1, \ldots, v_i \rangle \) of a worker \( p \). This deque satisfies the contour property if for all \( k = 0, 1, \ldots, r - 1 \), one of the following two conditions holds:

1. \( c(v_{k+1}) \) is the parent of \( c(v_k) \).
2. The root of \( c(v_k) \) is the start of some iteration \( i \), the root of \( c(v_{k+1}) \) is the start of the next iteration \( i + 1 \), and if \( k + 2 \leq r \), then \( c(v_{k+2}) \) is the common parent of both \( c(v_k) \) and \( c(v_{k+1}) \).

Contours allow us to prove an important property of the tail-swap operation.

**Lemma 4.** At any time during an execution of a pipeline program which produces a computation dag \( G = (V, E) \), suppose that worker \( p \) enables a vertex \( x \) via a throttling edge as a result of executing its assigned vertex \( v_0 \). If \( p \)'s deque satisfies the contour property (Definition 3), then either

1. \( p \)'s deque is empty and \( x \) becomes \( p \)'s new assigned vertex, or
2. \( p \)'s deque contains a single vertex \( v_1 \) which becomes \( p \)'s new assigned vertex and \( x \) is pushed onto \( p \)'s deque.

**Proof.** Because \( x \) is enabled by a throttling edge, \( v_0 \) must be the last node of some iteration \( i \), which by Lemma 2 means that \( c(x) \) is the parent of \( c(v_0) \). Because \( x \) is just being enabled, Lemma 1 implies that no other vertex in \( c(x) \) can belong to \( p \)'s deque. Suppose that \( p \)'s extended deque \( \langle v_0, v_1, \ldots, v_i \rangle \) contains \( r \geq 2 \) vertices. By Lemma 1, either \( v_1 \) or \( v_2 \) belongs to contour \( c(x) \), neither of which is possible, and hence \( r = 0 \) or \( r = 1 \). If \( r = 0 \), then \( x \) is \( p \)'s assigned vertex. If \( r = 1 \), then the root of \( c(v_1) \) is the start of iteration \( i + 1 \). Since \( x \) is enabled by a throttling edge, a tail swap occurs, making \( v_1 \) the assigned vertex of \( p \) and putting \( x \) onto \( p \)'s deque.
To analyze the time required for Piper to execute a computation dag $G = (V, E)$, define the enabling tree $G_T = (V, E_T)$ as the tree containing an edge $(x, y) \in E_T$ if $x$ is the last predecessor of $y$ to execute. The enabling depth $d(x)$ of $x \in V$ is the depth of $x$ in the enabling tree $G_T$.

**Definition 5.** At any time during an execution of a pipeline program which produces a computation dag $G = (V, E)$, consider the extended deque $(v_0, v_1, \ldots, v_r)$ of a worker $p$. The deque satisfies the depth property if the following conditions hold:

1. For $k = 1, 2, \ldots, r - 1$, we have $d(v_{k-1}) \geq d(v_k)$.
2. For $k = r$, we have $d(v_{r-1}) \geq d(v_r)$ or $v_r$ has an incoming throttling edge.
3. The inequalities are strict for $k > 1$.

**Theorem 6.** At all times during an execution of a pipeline program by Piper, all deques satisfy the contour and depth properties (Definitions 3 and 5).

**Proof.** The proof is similar to the inductive proof of Lemma 3 from [2]. Intuitively, we replace the “designated parents” discussed in [2] with contours, which exhibit similar parent-child relationships. Although most of the proof follows from this substitution, we address the two most salient differences. The other minor cases are either straightforward or similar to these two cases.

First, we consider the consequences of the tail-swap operation, which may occur if the assigned vertex $v_0$ is the end of an iteration and executing $v_0$ enables a vertex $z$ via a throttling edge. Lemma 4 describes the structure of a worker $p$'s extended deque in this case, and in particular, states that the deque contains at most 1 vertex. If $r = 0$, the deque is empty and the properties hold vacuously. Otherwise, $r = 1$ and the deque contains one element $v_1$, in which case the tail-swap operation assigns $v_1$ to $p$ and puts $z$ into $p$'s deque. The contour property holds, because $c(z)$ is the parent of $(v_1)$. The depth property holds, because $z$ is enabled by a throttling edge.

Second, we must show that the contour and depth properties hold when a worker $p$'s assigned vertex $v_0$ belongs to some iteration $i$ of a `pipeline`, and executing $v_0$ enables a vertex $y$ belonging to iteration $i + 1$ via a cross edge $(v_0, y)$. Assume that executing $v_0$ also enables $x$, where $x$ also belongs to iteration $i$. (The case where $x$ is not enabled is similar.) Since both $v_0$ and $x$ belong to the same iteration, $c(v_0) = c(x)$, and by Lemma 2, $c(x)$ is a sibling of $c(y)$ in the contour tree. Suppose that before $v_0$ executes, $p$'s extended deque is $(v_0, v_1, \ldots, v_r)$, and thus after $v_0$ executes, $p$'s extended deque is $(x, y, v_1, \ldots, v_r)$. For vertices $v_2, v_3, \ldots, v_r$, if they exist, the conditions of the contour property continue to hold by induction. Since $c(x)$ and $c(y)$ are adjacent siblings in the contour tree, we need only show that $c(v_1)$, if it exists, is their parent. But if $c(v_1)$ is not the parent of $c(v_0) = c(x)$, then by induction it must be that $c(x)$ and $c(v_1)$ are adjacent siblings. In this case $c(v_1) = c(y)$, which is impossible by Lemma 1. The depth property holds because $d(x) = d(y) = d(v_0) + 1 \geq d(v_1) + 1 > d(v_1)$.

**7. TIME ANALYSIS OF PIPER**

This section bounds the completion time for Piper, showing that Piper executes pipeline program asymptotically efficiently. Specifically, suppose that a pipeline program produces a computation dag $G = (V, E)$ with work $T_1$ and span $T_C$, when executed by Piper on $P$ processors. We show that for any $\epsilon > 0$, the running time is $T_P \leq T_1/P + O(T_C + \lg P + \lg(1/\epsilon))$ with probability at least $1 - \epsilon$, which implies that the expected running time is $E[T_P] \leq T_1/P + O(T_C + \lg P)$. This bound is comparable to the work-stealing bound for fork-join dags originally proved in [8].

We adapt the potential-function argument of Arora, Blumofe, and Plaxton [2]. Piper executes computation dags in a style similar to their work-stealing scheduler, except for tail swapping. Although Arora et al. ignore the issue of memory contention, we handle it using the “recycling game” analysis from [8], which contributes the additive $O(\lg P)$ term to the bounds.

As in [2], the crux of the proof is to bound the number of steal attempts performed during the execution of a computation dag $G$ in terms of its span $T_C$. We measure progress through the computation by defining a potential function for a vertex in the computation dag based on its depth in the enabling tree. Consider a particular execution of a computation dag $G = (V, E)$ by Piper. For that execution, we define the weight of a vertex $v$ as $w(v) = T_C - d(v)$, and we define the potential of vertex $v$ at a given time as

$$
\Phi(v) = \begin{cases} 
\frac{3^{2w(v)} - 1}{3^{2w(v)}} & \text{if } v \text{ is assigned,} \\
0 & \text{otherwise.}
\end{cases}
$$

We define the potential of a worker $p$'s extended deque $(v_0, v_1, \ldots, v_r)$ as $\Phi(p) = \sum_{v \in p} \Phi(v)$.

Given this potential function, the proof of the time bound follows the same overall structure as the proof in [2]. We sketch the proof.

First, we prove two properties of worker deques involving the potential function.

**Lemma 7.** At any time during an execution of a pipeline program which produces a computation dag $G = (V, E)$, the extended deque $(v_0, v_1, \ldots, v_r)$ of every worker $p$ satisfies the following:

1. $\Phi(v_r) + \Phi(v_{r-1}) \geq 3\Phi(v_0)$.
2. Let $\Phi'$ denote the potential after $p$ executes $v_0$. Then we have $\Phi(p) - \Phi'(p) = 2(\Phi(v_0) + \Phi(v_1)) / 3$, if $p$ performs a tail swap, and $\Phi(p) - \Phi'(p) \geq 5\Phi(v_0)/9$ otherwise.

**Proof.** Property 1 follows from the depth property of Theorem 6. Property 2 follows from Lemma 4, if $p$ performs a tail swap, and the analysis in [2] otherwise.

As in [2], we analyze the behavior of workers randomly stealing from each other using a balls-and-weighted-bins analog. We want to analyze the case where the top 2 elements are stolen out of any deque, however, not just the top element. To address this case, we modify Lemma 7 of [2] to consider the probability that 2 out of 2P balls land in the same bin.

**Lemma 8.** Consider P bins, where for $p = 1, 2, \ldots, P$, bin $p$ has weight $W_p$. Suppose that 2P balls are thrown independently and uniformly at random into the P bins. For bin $p$, define the random variable $X_p$ as

$$
X_p = \begin{cases} 
W_p & \text{if at least 2 balls land in bin } p \\
0 & \text{otherwise.}
\end{cases}
$$

Let $W = \sum_{p=1}^{P} W_p$ and $X = \sum_{p=1}^{P} X_p$. For any $\beta$ in the range $0 < \beta < 1$, we have Pr $\{X \geq \beta W\} > 1 - 3/(1 - \beta)^2$.

**Proof.** For each bin $p$, consider the random variable $W_p - X_p$. It takes on the value $W_p$ when 0 or 1 ball lands in bin $p$, and otherwise it is 0. Thus, we have

$$
E[W_p - X_p] = W_p \left((1 - 1/P)^2 + 2P(1 - 1/P)^{2P-1}(1/P)\right) = W_p \left(1 - 1/P\right)^{2P} (3P - 1) / (P - 1).
$$

Since $(1 - 1/P)^P$ approaches $1/e$ and $(3P - 1)/(P - 1)$ approaches 3, we have $\lim_{P \to \infty} E[W_p - X_p] = 3W_p/e^2$. In fact, one can show that $E[W_p - X_p]$ is monotonically increasing, approaching the limit
from below, and thus $E|W - X| \leq 3W/e^2$. By Markov’s inequality, we have that $Pr\{W - X > (1 - \beta)W\} < E|W - X|/(1 - \beta)W$, from which we conclude that $Pr\{X < \beta W\} \leq 3/(1 - \beta)e^2$.

To use Lemma 8 to analyze Piper, we divide the time steps of the execution of $G$ into a sequence of rounds, where each round (except the first, which starts at time 0) starts at the time step after the previous round ends and continues until the first time step such that at least $2P$ steal attempts — and hence less than $3P$ steal attempts — occur within the round. The following lemma shows that a constant fraction of the total potential in all deques is lost in each round, thereby demonstrating progress.

**Lemma 9.** Consider a pipeline program executed by Piper on $P$ processors. Suppose that a round starts at time step $t$ and finishes at time step $t'$. Let $\Phi$ denote the potential at time $t$, let $\Phi'$ denote the potential at time $t'$, let $\Phi = \sum_{p=1}^{P} \phi(p)$, and let $\Phi' = \sum_{p=1}^{P} \phi(p)$. Then we have $Pr\{\Phi - \Phi' \geq \Phi/4\} > 1 - 6/e^2$.

**Proof.** We first show that stealing twice from a worker $p$’s deque contributes a potential drop of at least $\phi(p)/2$. The proof follows a similar case analysis to that in the proof of Lemma 8 in [2] with two main differences. First, we use the two properties of $\phi$ in Lemma 7. Second, we must consider the case unique to Piper, where $p$ performs a tail swap after executing its assigned vertex $v_0$. In this case, $p$’s deque contains a single ready vertex $v_1$ and $p$ may perform a tail swap if executing $v_0$ enables a vertex via an outgoing throttling edge. If so, however, then by Lemma 7, the potential drops by at least $2(\phi(v_0) + \phi(v_1))/3 > \phi(p)/2$, since $\phi(p) = \phi(v_0) + \phi(v_1)$.

Now, suppose that we assign each worker $p$ a weight of $W_p = \phi(p)/2$. These weights $W_p$ sum to $W = \Phi/2$. If we think of steal attempts as ball tosses, then the random variable $X$ from Lemma 8 bounds from below the potential decrease due to actions on $p$’s deque. Specifically, if at least 2 steal attempts target $p$’s deque in a round (which corresponds conceptually to at least 2 balls landing in bin $p$), then the potential drops by at least $W_p$. Moreover, $X$ is a lower bound on the potential decrease within the round, i.e., $X \leq \Phi - \Phi'$. By Lemma 8, we have $Pr\{X \geq W/2\} > 1 - 6/e^2$. Substituting for $X$ and $W$, we conclude that $Pr\{\Phi - \Phi' \geq \Phi/4\} > 1 - 6/e^2$.

We are now ready to prove the completion-time bound.

**Theorem 10.** Consider an execution of a pipeline program by Piper on $P$ processors which produces a computation dag with work $T_1$ and span $T_m$. For any $\epsilon > 0$, the running time is $T_p \leq T_1/P + O(T_m + \log P + \log(1/\epsilon))$ with probability at least $1 - \epsilon$.

**Proof.** On every time step, consider each worker as placing a token in a bucket depending on its action. If a worker $p$ executes an assigned vertex, $p$ places a token in the work bucket. Otherwise, $p$ is a thief and places a token in the steal bucket. There are exactly $T_1$ tokens in the work bucket at the end of the computation. The interesting part is bounding the size of the steal bucket.

Divide the execution of $G$ into rounds. Recall that each round contains at least $2P$ and less than $3P$ steal attempts. Call a round successful if after that round finishes, the potential drops by at least $1/4$ fraction. From Lemma 9, a round is successful with probability at least $1 - 6/e^2 \geq 1/6$. Since the potential starts at $\Phi_0 = 2^{2P-1}$, ends at 0, and is always an integer, the number of successful rounds is at most $(2P - 1) \log_{1/6}(3) < 8T_m$. Consequently, the expected number of rounds needed to obtain $8T_m$ successful rounds is at most $48T_m$, and the expected number of tokens in the steal bucket is therefore at most $3P \cdot 48T_m = 144PT_m$.

For the high-probability bound, suppose that the execution takes $n = 48T_m + m$ rounds. Because each round succeeds with probability at least $p = 1/6$, the expected number of successes is at least $np = 8T_m + m/6$. We now compute the probability that the number $X$ of successes is less than $8T_m$. As in [2], we use the Chernoff bound $Pr\{X < np - a\} < e^{-a^2/2np}$, with $a = m/6$. Choosing $m = 48T_m + 21\ln(1/\epsilon)$, we have

$$Pr\{X < 8T_m\} < e^{-(m/2)/(2np)} = e^{-m/21} \leq \epsilon.$$ 

Hence, the probability that the execution takes $n = 96T_m + 21\ln(1/\epsilon)$ rounds or more is less than $\epsilon$, and the number of tokens in the steal bucket is at most $288T_m + 63\ln(1/\epsilon)$.

The additional $\log P$ term comes from the “recycling game” analysis described in [8], which bounds any delay that might be incurred when multiple processors try to access the same deque in the same time step in randomized work-stealing.

**8. SPACE ANALYSIS OF PIPER**

This section derives bounds on the stack space required by Piper by extending the bounds in [8] for fully strict fork-join parallelism to include pipeline parallelism. We show that Piper on $P$ processors uses $S_P \leq P(S_1 + fDK)$ stack space for pipeline iterations, where $S_1$ is the serial stack space, $f$ is the “frame size,” $D$ is the depth of nested linear pipelines, and $K$ is the throttling limit.

To model Piper’s usage of stack space, we partition the vertices of the computation dag $G$ of the pipeline program into a tree of contours, as described in Section 6. Each contour in this partition is rooted at the start of a spawned subcomputation. The control for each pipe_while loop, which corresponds to a while loop as in line 5 of Figure 5, belongs to some contour in the contour tree with its iterations as children. Define the pipe nesting depth $D$ of $G$ as the maximum number of pipe_while contours on any path from leaf to root in the contour tree.

We assume that every contour $c$ of $G$ has an associated frame size representing the stack space consumed by $c$ while it or any of its descendant contours are executing. The space used by Piper on any time step is the sum of frame sizes of all contours $c$ which are either (1) associated with a vertex in some worker’s extended deque, or (2) suspended, meaning that the earliest unexecuted vertex in the contour is not ready. Let $S_P$ denote the maximum over all time steps of the stack space used by Piper during a $P$-worker execution of $G$. Thus, $S_1$ is the stack space used by Piper for a serial execution. We now generalize the space bound $S_P \leq P S_1$ from [8], which deals only with fork-join parallelism, to pipeline programs.

**Theorem 11.** Consider a pipeline program with pipe nesting depth $D$ executed on $P$ processors by Piper with throttling limit $K$. The execution requires $S_P \leq P(S_1 + fDK)$ stack space, where $f$ is the maximum frame size of any contour of any pipe_while iteration and $S_1$ is the serial stack space.

**Proof.** We show that except for suspended contours that are pipe_while iterations, Piper still satisfies the “busy-leaves property” [8]. More precisely, at any point during the execution, in the tree of active and suspended contours, each leaf contour either (1) is currently executing on some worker, or (2) is a suspended pipe_while iteration with a sibling iteration that is currently executing on some worker. In fact, one can show that for any pipe_while loop, the contour for the leftmost (smallest) iteration that has not completed is either active or has an active descendant in the contour tree. The bound of $PS_1$ covers the space used by all contours that fall into Case (1).
To bound the space used by contours from Case (2), observe that any pipe\_while loop uses at most fK space for iteration contours, since the throttling edge from the leftmost active iteration precludes having more than K active or suspended iterations in any one pipe\_while loop. Thus, each worker p has at most fDK iteration contours for any pipe\_while loop that is an ancestor of the contour p's assigned vertex. Summing the space used over all workers gives PfDK additional stack-space usage.

9. CILK-P RUNTIME DESIGN

This section describes the Cilk-P implementation of the PIPER scheduler. We first introduce the data structures Cilk-P uses to implement a pipe\_while loop. Then we describe the two main optimizations that the Cilk-P runtime exploits: lazy enabling and dependency folding.

Data structures
Like the Cilk-M runtime [23] on which it is based, Cilk-P organizes runtime data into frames. Cilk-P executes a pipe\_while loop in its own function, whose frame, called a control frame, handles the spawning and throttling of iterations. Furthermore, each iteration of a pipe\_while loop executes as an independent child function, with its own iteration frame. This frame structure is similar to that of an ordinary while loop in Cilk-M, where each iteration spawns a function to execute the loop body. Cross and throttling edges, however, may cause the iteration and control frames to suspend.

Cilk-P's runtime employs a simple mechanism to track progress of an iteration i. The frame of iteration i maintains a stage counter, which stores the stage number of the current node in i, and a status field, which indicates whether i is suspended due to an unsatisfied cross edge. Because executed nodes in an iteration i have strictly increasing stage numbers, checking whether a cross edge into iteration i is satisfied amounts to comparing the stage counters of iterations i and i−1. Any iteration frame that is not suspended corresponds to either a currently executing or a completed iteration.

Cilk-P implements throttling using a join counter in the control frame. Normally in Cilk-M, a frame's join counter simply stores the number of active child frames. Cilk-P also uses the join counter to limit the number of active iteration frames in a pipe\_while loop to the throttling limit K. Starting an iteration increments the join counter, while returning from an iteration decrements it. If a worker tries to start a new iteration when the control frame's join counter is K, the control frame suspends until a child iteration returns.

Using these data structures, one could implement PIPER directly, by pushing and popping the appropriate frames onto deques as specified by PIPER's execution model. In particular, the normal THE protocol [15] could be used for pushing and popping frames from a deque, and frame locks could be used to update fields in the frames atomically. Although this approach directly matches the model analyzed in Sections 7 and 8, it incurs unnecessary overhead for every node in an iteration. Cilk-P implements lazy enabling and dependency folding to reduce this overhead.

Lazy enabling
In the PIPER algorithm, when a worker p finishes executing a node in iteration i, it may enable an instruction in iteration i+1, in which case p pushes this instruction onto its deque. To implement this behavior, intuitively, p must check right — read the stage counter and status of iteration i+1 — whenever it finishes executing a node. The work to check right at the end of every node could amount to substantial overhead in a pipeline with fine-grained stages.

Lazy enabling allows p's execution of an iteration i to defer the check-right operation, as well as avoid any operations on its deque involving iteration i+1. Conceptually, when p enables work in iteration i+1, this work is kept on p's deque implicitly. When a thief p' tries to steal iteration i's frame from p's deque, p' first checks right on behalf of p to see whether any work from iteration i+1 is implicitly on the deque. If so, p' resumes iteration i+1 as if it had found it on p's deque. In a similar vein, the Cilk-P runtime system also uses lazy enabling to optimize the check-parent operation — the enabling of a control frame suspended due to throttling.

Lazy enabling requires p to behave differently when p completes an iteration. When p finishes iteration i, it first checks right, and if that fails (i.e., iteration i+1 need not be resumed), it checks its parent. It turns out that these checks find work only if p's deque is empty. Therefore, p can avoid performing these checks at the end of an iteration if its deque is not empty.

Lazy enabling is an application of the work-first principle [15]: minimize the scheduling overheads borne by the work of a computation, and amortize them against the span. Requiring a worker to check right every time it completes a node adds overhead proportional to the work of the pipe\_while in the worst case. With lazy enabling, the overhead can be amortized against the span of the computation. For programs with sufficient parallelism, the work dominates the span, and the overhead becomes negligible.

Dependency folding
In dependency folding, the frame for iteration i stores a cached value of the stage counter of iteration i−1, hoping to avoid the checking of already satisfied cross edges. In a straightforward implementation of PIPER, before a worker p executes each node in iteration i with an incoming cross edge, it reads the stage counter of iteration i−1 to see if the cross edge is satisfied. Reading the stage counter of iteration i−1, however, can be expensive. Besides the work involved, the access may contend with whatever worker p' is executing iteration i−1, because p' may be constantly updating the stage counter of iteration i−1.

Dependency folding mitigates this overhead by exploiting the fact that an iteration's stage counter must strictly increase. By caching the most recently read stage-counter value from iteration i−1, worker p can sometimes avoid reading this stage counter before each node with an incoming cross edge. For instance, if p' finishes executing a node (i−1, j), then all cross edges from nodes (i−1, 0) through (i−1, j) are necessarily satisfied. Thus, if p reads j from iteration i−1's stage counter, p need not reread the stage counter of i−1 until it tries to execute a node with an incoming cross edge (i, j') where j' > j. This optimization is particularly useful for fine-grained stages that execute quickly.

10. EVALUATION

This section presents empirical studies of the Cilk-P prototype system. We investigated the performance and scalability of Cilk-P using the three PARSEC [4, 5] benchmarks that we ported, namely ferret, dedup, and x264. The results show that Cilk-P's implementation of pipeline parallelism has negligible overhead compared to its serial counterpart. We compared the Cilk-P implementations to TBB and Pthreaded implementations of these benchmarks. We found that the Cilk-P and TBB implementations perform comparably, as do the Cilk-P and Pthreaded implementations for ferret and x264. The Pthreaded version of dedup outperforms both Cilk-P and TBB, because the bind-to-element approaches of Cilk-P and TBB produce less parallelism than the Pthreaded bind-to-stage approach. Moreover, the Pthreaded approach benefits more from “oversubscription.” We study the effectiveness of dependency folding on a
The experiments were conducted using native, the largest input data set that comes with the PARSEC benchmark suite. The leftmost column shows the number of cores used \((P)\). Subsequent columns show the running time \((T_p)\), speedup over serial running time \((T_{s,p})\), and scalability \((T_1/T_p)\) for each system. The throttling limit was \(K = 10P\).

### Table 1: Performance comparison of the three dedup implementations.

<table>
<thead>
<tr>
<th>(P)</th>
<th>Processing Time ((T_p))</th>
<th>Speedup ((T_s/T_p))</th>
<th>Scalability ((T_1/T_p))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.94</td>
<td>2.01</td>
<td>1.97</td>
</tr>
<tr>
<td>4</td>
<td>3.92</td>
<td>4.06</td>
<td>3.93</td>
</tr>
<tr>
<td>8</td>
<td>7.76</td>
<td>7.96</td>
<td>7.75</td>
</tr>
<tr>
<td>12</td>
<td>11.46</td>
<td>11.70</td>
<td>11.36</td>
</tr>
<tr>
<td>16</td>
<td>14.98</td>
<td>14.93</td>
<td>14.75</td>
</tr>
</tbody>
</table>

**Figure 6:** Performance comparison of the three dedup implementations.

The experiments were conducted using native, the largest input data set and come with the PARSEC benchmark suite. The column headers are the same as in Figure 6. The throttling limit was \(K = 4P\).

### Table 2: Performance evaluation on PARSEC benchmarks.

<table>
<thead>
<tr>
<th>(P)</th>
<th>Encoding Time ((T_e))</th>
<th>Speedup ((T_s/T_e))</th>
<th>Scalability ((T_1/T_e))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.02</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>2.27</td>
<td>2.09</td>
<td>2.24</td>
</tr>
<tr>
<td>4</td>
<td>4.63</td>
<td>4.14</td>
<td>4.55</td>
</tr>
<tr>
<td>8</td>
<td>8.53</td>
<td>8.27</td>
<td>8.40</td>
</tr>
<tr>
<td>12</td>
<td>11.84</td>
<td>11.44</td>
<td>11.66</td>
</tr>
<tr>
<td>16</td>
<td>13.87</td>
<td>13.63</td>
<td>13.66</td>
</tr>
</tbody>
</table>

**Figure 7:** Performance comparison of the three dedup implementations.

The experiments were conducted using native, the largest input data set that comes with the PARSEC benchmark suite. The column headers are the same as in Figure 6. The throttling limit was \(K = 4P\).

**Figure 8:** Performance comparison between the Cilk-P implementation and the Pthreaded implementation of x264 (encoding only). The experiments were conducted using native, the largest input data set that comes with the PARSEC benchmark suite. The column headers are the same as in Figure 6. The throttling limit was \(K = 4P\).
Figure 9: Performance evaluation using the pipe-fib benchmark. We tested the system with two different programs, the ordinary pipe-fib, and the pipe-fib-256, which is coarsened. Each program is tested with and without the dependency folding optimization. For each program for a given setting, we show the running time of its serial counterpart (T₀), running time executing on a single worker (T₁), on 16 workers (T₁₆), its serial overhead, scalability, and speedup obtained running on 16 workers.

Figure 10: Sketch of the pathological unthrottled linear pipeline dag, which can be used to prove Theorem 13. Small circles represent nodes with unit work, medium circles represent nodes with T₁/3 work, and large circles represent nodes with T₂/3 work. The number of iterations per cluster is T₁₁/3 + 1, and the total number of iterations is (T₁₂/3 + T₁₁/3)/2.

like the system to coarsen automatically, which is what dependency folding effectively does.

Further investigation revealed that the time spent checking for cross edges increases noticeably when the number of workers increases from 1 to 2. It turns out that when iterations are run in parallel, each check for a cross-edge dependency necessarily incurs a true-sharing conflict between the two adjacent active iterations, an overhead that occurs only during parallel execution. Dependency folding eliminated much of this overhead for pipe-fib, as shown in the third row of Figure 9, leading to scalability that exceeds the coarsened version without the optimization, although a slight price is still paid in speedup. Employing both optimizations, as shown in the last row of the table, produces the best numbers for both speedup and scalability.

11. CONCLUSION

What impact does throttling have on theoretical performance? PIPER relies on throttling to achieve its provable space bound and avoid runaway pipelines. Ideally, the user should not worry about throttling, and the system should perform well automatically, and indeed, PIPER’s throttling of a pipeline computation is encapsulated in Cilk-P’s runtime system. But what price is paid?

We can pose this question theoretically in terms of a pipeline computation G’s unthrottled dag: the dag G = (V, E) with the same vertices and edges as G, except without throttling edges. How does adding throttling edges to an unthrottled dag affect span and parallelism?

The following two theorems provide two partial answers to this question. First, for uniform pipelines, where the cost of a node (i, j) is identical across all iterations i — all stages have the same cost — throttling does not affect the asymptotic performance of PIPER executing G.

THEOREM 12. Consider a uniform unthrottled linear pipeline G = (V, E) having n iterations and s stages. Suppose that PIPER throttles the execution of G on P processors using a window size of K = aP, for some constant a > 1. Then PIPER executes G in time T_P ≤ (1 + c/a)T₁/P + cT₁₀ for some sufficiently large constant c, where T₁ is the total work in G and T₁₀ is the span of G.

Second, we consider nonuniform pipelines, where the cost of a node (i, j) may vary across iterations. It turns out that nonuniform pipelines can pose performance problems, not only for PIPER, but for any scheduler that throttles the computation. Figure 10 illustrates a pathological nonuniform pipeline for any scheduler that uses throttling. In this dag, T₁ work is distributed among (T₁₁/3 + T₁₂/3)/2 iterations such that any T₁₁/3 + 1 consecutive iterations consist of 1 heavy iteration with T₂/3 work and T₁₁/3 light iterations of T₁₁/3 work each. Intuitively, achieving a speedup of 3 on this dag requires having at least 1 heavy iteration and Θ(T₁₁/3)

3Figure 9 shows the results from a single run, but these data are representative of other runs with different input sizes.
light iterations active simultaneously, which is impossible for any scheduler that uses a throttling limit of \( K = o(T_1^{1/3}) \). The following theorem formalizes this intuition.

**Theorem 13.** Let \( \hat{G} = (V, \hat{E}) \) denote the nonuniform unthrottled linear pipeline shown in Figure 10, with work \( T_1 \) and span \( T_{\infty} \leq 2T_1^{2/3} \). Let \( S_1 \) denote the optimal stack-space usage when \( G \) is executed on 1 processor. Any P-processor execution of \( G \) that achieves \( T_p \leq T_1/\rho \), where \( \rho \) satisfies \( 3 \leq \rho \leq O(T_1/T_{\infty}) \), uses space \( S_p \geq S_1 + (\rho - 3)T_1^{1/3}/2 - 1 \).

Intuitively, these two theorems present two extremes of the effect of throttling on pipeline dags. One interesting avenue for research is to determine what are the minimum restrictions on the structure of an unthrottled linear pipeline \( G \) that would allow a scheduler to achieve parallel speedup on \( P \) processors using a throttling limit of only \( \Theta(P) \).

12. **ACKNOWLEDGMENTS**

Thanks to Loren Merritt of x264 LLC and Hank Hoffman of University of Chicago (formerly of MIT CSAIL) for answering questions about x264. Thanks to Yungang Bao of Institute of Computing Technology, Chinese Academy of Sciences (formerly of Princeton) for answering questions about the PARSEC benchmark suite. Thanks to Will Hasenplag of MIT CSAIL and Nasro Min-Allah of COMSATS Institute of Information Technology in Pakistan for helpful discussions. We especially thank the reviewers for their thoughtful comments.

13. **REFERENCES**


