Parallel Computing: Theory and Practice

COLLABORATORS			
	TITLE : Parallel Computing: Theory and Practice		
ACTION	*NAME*	*DATE*	*SIGNATURE*
WRITTEN BY		February 9, 2016	

REVISION HISTORY			
NUMBER	DATE	DESCRIPTION	NAME

Contents

List of Figures

Preface

The goal of this book is to cover the fundamental concepts of parallel computing, including models of computation, parallel algorithms, and techniques for implementing and evaluating parallel algorithms.

Our primany focus will be hardware-shared memory parallelism.

For implementations, we use a C++ library, called **PASL**, which we have been developing over the past 5 years. PASL stands for Parallel Algorithm Scheduling Library. It also sounds a bit like the French phrase "pas seul" (pronounced "pa-sole"), meaning "not alone".

The library provides several scheduling algorithms for executing parallel programs on modern multicores and provides a range of utilities for inspecting the empirical behavior of parallel programs. We expect that the instructions in this book to allow the reader to write performant parallel programs at a relatively high level (essentially at the same level of C++ code) without having to worry too much about lower level details such as machine specific optimizations, which might otherwise be necessary.

All code that discussed in this book can be found at the Github repository linked by the following URL:

https://github.com/deepsea-inria/pasl/tree/edu

This code-base includes the examples presented in the book, see file `minicourse/examples.hpp`.

Some of the material in this book is based on the course, 15-210, co-taught with Guy Blelloch at CMU. The interested reader can find more details on this material in this book.

Starting point for this book was this book on PASL.

v1.0 2016-01

Author: Umut A. Acar

1 Administrative Matters

Course combines theory and practice. We will try ask the following two questions.

1. Does it work in practice?
2. Does it work in theory?

As you know this is a graduate class. This means that I don't care about your grade. But if you are sitting here, you might as well work for an A.

Grading will be based on some assignments, one midterm exam, and one project. We shall make time so that you can devote a good chunk of your semester to the project. You will also be receive a participation score, which amounts to 20\% of the grade.

For each lecture, I will make some notes and we shall make them available for you to comment, perhaps via a github repository.

2 Preliminaries

2.1 Processors, Processes, and Threads

We assume a machine model that consists of a shared memory by a number of processors, usually written as P. The processors have access to a shared memory, which is readable and writable by all processors.

We assume that an operating system or a that allows us to create ***processes***. The kernel schedules processes on the available processors in a way that is mostly out of our control with one exception: the kernel allows us to create any number of processes and ***pin*** them on the available processors as long as no more than one process is pinned on a processor.

We define a ***thread*** to be a piece of sequential computation whose boundaries, i.e., its start and end points, are defined on a case by case basis, usually based on the programming model. In reality, there different notions of threads. For example, a system-level thread is created by a call to the kernel and scheduled by the kernel much like a process. A user-level thread is created by the application program and is scheduled by the application—user level threads are invisible to the kernel. Common property of all threads is that they perform a sequential computation. In this class, we will usually talk about user-level threads. In the literature, you will encounter many different terms for a user-level thread, such as "fiber", "sparc", "strand", etc.

For our purposes in this book an ***application***, a piece of runnable software, can only create threads but no processes. We will assume that we can assign to an application any number of processes to be used for execution. If an application is run all by itself (without any other application running at the same time) and if all of its processes are pinned, then we refer to such an execution as occurring in the ***dedicated mode***.

Note
For now, we leave the details of the memory consistency model unspecified.

2.2 C++ Background

This book is entirely based on C++ and a library for writing parallel programs in C++. We use recent features of C++ such as closures or lambda expressions and templates. A deep understanding of these topics is not necessary to follow the course notes, because we explain them at a high level as we go, but such prior knowledge might be helpful; some pointers are provided below.

2.2.1 Template metaprogramming

Templates are C++'s way of providing for parametric polymorphism, which allows using the same code at multiple types. For example, in modern functional languages such as the ML family or Haskell, you can write a function $\lambda\ x.x$ as an identity function that returns its argument for any type of x. You don't have to write the function at every type that you plan to apply. Since functional languages such as ML and Haskell rely on type inference and have powerful type systems, they can infer from your code the most general type (within the constraints of the type system). For example, the function $\lambda\ x.x$ can be given the type $\forall\alpha.\alpha \rightarrow \alpha$. This type says that the function works for any type α and given an argument of type α, it returns a value of type α.

C++ provides for polymorphism with ***templates***. In its most basic form, a template is a class declaration or a function declaration, which is explicitly stated to be polymorphic, by making explicit the type variable. Since C++ does not in general perform type inference (in a rigorous sense of the word), it requires some help from the programmer.

For example, the following code below defines an array class that is parametric in the type of its elements. The declaration `template <class T>` says that the declaration of `class array`, which follows is parameterized by the identifier `T`. The definition of `class array` then uses `T` as a type variable. For example, the array defines a pointer to element sequences of type `T`, and the `sub` function returns an element of type `T` etc.

```
template <class T>
class array {
  public:
   array (int size) {a = new T[size];}
   T sub (int i) { a[i];}

  private:
    *T a;
}
```

Note that the only part of the syntax `template <class T>` that is changeable is the identifier `T`. In other words, you should think of the syntax `template <class T>` as a binding form that allows you to pick an identifier (in this case `T`). You might ask why the type identifier/variable `T` is a `class`. This is a good question. The authors find it most helpful to not think much about such questions, especially in the context of the C++ language.

Once defined a template class can be initialized with different type variables by using the `< >` syntax. For examples, we can define different arrays such as the following.

```
array<int> myFavoriteNumbers(7);
array<char*> myFavoriteNames(7);
```

Again, since C++ does not perform type inference for class instances, the C++ compiler expects the programmer to eliminate explicitly parametricity by specifying the argument type.

It is also possible to define polymorphic or generic functions. For example, the following declarations defines a generic identity function.

```
template <class T>
T identity(T x) { return x;}
```

Once defined, this function can be used without explicitly specializing it at various types. In contrast to templated classes, C++ does provide some type inference for calls to templated functions. So generic functions can be specialized implicitly, as shown in the examples below.

```
i = identity (3)
s = identity ("template programming can be ugly")
```

This brief summary of templates should suffice for the purposes of the material covered in this book. Templates are covered in significant detail by many books, blogs, and discussions boards. We refer the interested reader to those sources for further information.

2.2.2 Lambda expressions

The C++11 reference provides good documentation on lambda expressions.

3 Introduction

This class is motivated by recent advances in architecture that put sequential hardware on the path to extinction. Due to fundamental architectural limitations, sequential performance of processors have not been increasing since 2005. In fact performance has been decreasing by some measures because hardware manufacturer's have been simplifying processorns by simplifying the handling of instruction level parallelism.

Moore's law, however, continues to be holding, at least for the time being, enabling hardware manufacturers to squeeze an increasing number of transistors into the same chip area. The result, not surprisingly, has been increased paralellelism, more processors that is. In particular, manufacturers have been producing multicore chips where each chip consists of a number of processors that fit snugly into a small area, making communication between them fast and efficient. You can read more about the history of modern multicore computers.

This simple change in hardware has led to dramatic changes in computing. Parallel computing, once a niche domain for computational scientists, is now an everyday reality. Essentially all computing media ranging from mobile phones to laptops and computers operate on parallel computers.

This change was anticipated by computers scientists. There was in fact much work on parallel algorithms in 80's and 90's. The models of computation assumed then turned out to be unrealistic. This makes it somewhat of a challenge to use the algorithms from that era. Some of the ideas, however, transcends those earlier models can still be used today to design and implement parallel algorithms on modern architectures.

The goal of this class is to give an introduction to the theory and the practice of parallol computing. Specifically, we will cover the following topics.

1. Multithreaded computation
2. Work and span
3. Offline scheduling
4. Structured or implicit parallel computation
    a. Fork-join, async-finish, nested parallelism.
    b. Futures.
    c. Parallelism versus concurrency
5. Parallel algorithms for sequences
6. Online scheduling: work-stealing algorithms and its analysis
7. Parallel algorithms for trees
8. Parallel algorithms for graphs
9. Concurrency

4 Chapter: Multithreading, Parallelism, and Concurrency

The term ***multithreading*** refers to computing with multiple threads of control. Once created, a thread performs a computation by executing a sequence of instructions, as specified by the program, until it terminates. A multithreaded computation starts by executing a ***main thread***, which is the thread at which the execution starts. A thread can create or ***spawn*** another thread and ***synchronize*** with other threads by using a variety of synchronization constructs such as locks, mutex's, synchronization variables, and semaphores.

4.1 DAG Representation

A multithreaded computation can be represented by a dag, a Directed Acyclic Graph, or written also more simply a ***dag***, of vertices. The figure below show an example multithreaded computation and its dag. Each vertex represents the execution of an ***instruction***, such as an addition, a multiplication, a memory operation, a (thread) spawn operation, or a synchronization operation. A vertex representing a spawn operation has outdegree two. A synchronization operation waits for an operation belonging to a thread to complete, and thus a vertex representing a synchronization operation has indegree two. Recall that a dag represents a partial order. Thus the dag of the computation represents the partial ordering of the dependencies between the instructions in the computation.

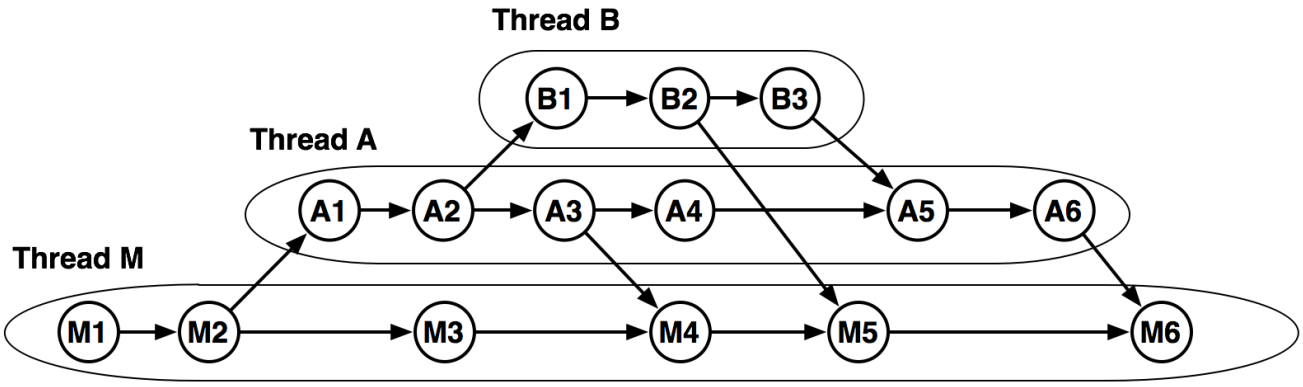


Figure 1: A multithreaded computation.

Throughout this book, we make two assumptions about the structure of the dag:

1. Each vertex has outdegree at most two.
2. The dag has exactly one ***root vertex*** with indegree zero and one ***final vertex*** vertex with outdegree zero. The root is the first instruction of the ***root thread***.

The outdegree assumption naturally follows by the fact that each vertex represents an instruction, which can create at most one thread.

4.2 Cost Model: Work and Span

For analyzing the efficiency and performance of multithreaded programs, we use several cost measures, the most important ones include work and span. We define the ***work*** of a computation as the number of vertices in the dag and the ***span*** as the length of the longest path in the dag. In the example dag above, work is 15 and span is 9.

4.3 Execution and Scheduling

The execution of a multithreaded computation executes the vertices in the dag of the computation in some partial order that is consistent with the partial order specified by the dag, that is, if vertices u, v are ordered then the execution orders them in the same way.

Multithreaded programs are executed by using a ***scheduler*** that assigns vertices of the dag to processes.

Definition: Execution Schedule

Given a dag G, an (execution) schedule for G is a function from processes and (time) ***steps*** to instructions of the dag such that

1. if a vertex u is ordered before another v in G, then v is not executed at a time step before u, and
2. each vertex in G is executed exactly once.

The ***length*** of a schedule is the number of steps in the schedule.

The first condition ensures that a schedule observes the dependencies in the dag. Specifically, for each arc (u, v) in the dag, the vertex u is executed before vertex v.

For any step in the execution, we call a vertex ***ready*** if all the ancestors of the vertex in the dag are executed prior to that step. Similarly, we say that a thread is ready if it contains a ready vertex. Note that a thread can contain only one ready vertex at any time.

Example 4.1 Schedule

An example schedule with 3 processes. The length of this schedule is 10

Time Step	Process 1	Process 2	Process 3
1	M1		
2	M2		
3	M3	A1	
4		A2	
5	B1	A3	
6	B2	A4	
7	B2		M4
8		A5	M5
9	A6		
10		M6	

Fact: Scheduling Invariant

Consider a computation dag G and consider an execution using any scheduling algorithm. At any time during the execution, color the vertices that are executed as blue and the others as red.

1. The blue vertices induce a blue sub-dag of G that is connected and that has the same root as G.
2. The red vertices incude a red sub-dag of G that is connected.
3. All the vertices of G are in the blue or the red sub-dag. In other words, the blue and red vertices partitions the dag into two sub-dags.

4.4 Scheduling Lower Bounds

Theorem: Lower bounds

Consider any multithreaded computation with work W and span S and P processes. The following lower bounds hold.

1. Every execution schedule has length at least $\frac{W}{P}$.
2. Every execution schedule has length at least S.

The first lower bound follows by the simple observation that a schedule can only execute P instructions at a time. Since all vertices must be executed, a schedule has length at least $\frac{W}{P}$. The second lower bound follows by the observation that the schedule cannot execute a vertex before its ancestors and thus the length of the schedule is at least as long as any path in the dag, which can be as large as the span S.

4.5 Offline Scheduling

Having established a lower bound, we now move on to establish an upper bound for the ***offline scheduling problem***, where we are given a dag and wish to find an execution schedule that minimizes the run time. It is known that the related decision problem in NP-complete but that 2-approximation is relatively easy. We shall consider two distinct schedulers: ***level-by-level scheduler*** and ***greedy scheduler***.

A ***level-by-level schedule*** is a schedule that executes the instructions in a given dag level order, where the ***level*** of a vertex is the longest distance from the root of the dag to the vertex. More specifically, the vertices in level 0 are executed first, followed by the vertices in level 1 and so on.

Theorem:[Offline level-by-level schedule]

For a dag with work W and span S, the length of a level-by-level schedule is $W/P+S$.

Proof

Let W_i denote the work of the instructions at level i. These instructions can be executed in $\lceil \frac{W_i}{P} \rceil$ steps. Thus the total time is

$$\sum_{i=1}^{S} \lceil \frac{W_i}{P} \rceil \leq \sum_{i=1}^{S} \lfloor \frac{W_i}{P} \rfloor + 1 \leq \lfloor \frac{W}{P} \rfloor + S$$

This theorem called https://www.google.com/search?q=Brent%27s+theorem&gws_rd=ssl [Brent's theorem] was proved by Brent in 1974. It shows that the lower bound can be approximated within a factor of 2.

Brent's theorem has later been generalized to all greedy schedules. A ***greedy schedule*** is a schedule that never leaves a process idle unless there are no ready vertices. In other words, greedy schedules keep processes as busy as possibly by greedily assigning ready vertices.

Theorem: Offline Greedy Schedule

Consider a dag G with work W and span S. Any greedy P-process schedule has length at most $\frac{W}{P} + S \cdot \frac{P-1}{P}$.

Proof

Consider any greedy schedule with length T.

For each step $1 \leq i \leq T$, and for each process that is scheduled at that step, collect a token. The token goes to the ***work bucket*** if the process executes a vertex in that step, otherwise the process is idle and the token goes to an ***idle bucket***.

Since each token in the work bucket corresponds to an executed vertex, there are exactly W tokens in that bucket.

We will now bound the tokens in the idle bucket by $S \cdot (P-1)$. Observe that at any step in the execution schedule, there is a ready vertex to be executed (because otherwise the execution is complete). This means that at each step, at most $P-1$ processes can contribute to the idle bucket. Furthermore at each step where there is at least one idle process, we know that the number of ready vertices is less than the number of available processes. Note now that at that step, all the ready vertices have no incoming edges in the red sub-dag consisting of the vertices that are not yet executed, and all the vertices that have no incoming edges in the red sub-dag are ready. Thus executing all the ready vertices at the step reduces the length of all the paths that originate at these vertices and end at the final vertex by one. This means that the span of the red sub-dag is reduced by one because all paths with length equal to span must originate in a ready vertex. Since the red-subdag is initially equal to the dag G, its span is S, and thus there are at most S steps at which a process is idle. As a result the total number of tokens in the idle bucket is at most $S \cdot (P-1)$.

Since we collect P tokens in each step, the bound thus follows.

Exercise

Show that the bounds for Brent's level-by-level scheduler and for any greedy scheduler is within a factor 2 of optimal.

4.6 Online Scheduling

In offline scheduling, we are given a dag and are interested in finding a schedule with minimal length. When executing multi-threaded program, however, we don't have full knowledge of the dag. Instead, the dag unfolds as we run the program. Furthermore, we are interested in not minimizing the length of the schedule but also the work and time it takes to compute the schedule. These two additional conditions define the ***online scheduling problem.***

There are many different algorithms for online scheduling but these algorithms all operate similarly. We can outline a typical scheduling algorithm as follows. The algorithm maintain a ***work pool*** of work, consisting of ready threads, and execute them. Execution starts with the root thread in the pool. It ends when the final vertex is executed. In order to minimize the cost of computing the schedule, the algorithm executes a thread until there is a need for synchronization with other threads.

To obtain work, a process removes a thread from the pool and executes its ready vertex. We refer to the thread executed by a process as the ***assigned thread***. When executed, the ready vertex can make the next vertex of the thread ready, which then also gets executed an so on until one of the following ***synchronization*** actions occur.

1. ***Die:*** The process executes last vertex of the thread, causing the thread to die. The process then obtains other work.
2. ***Block:*** The assigned vertex executes but the next vertex does not become ready. This blocks the thread and thus the process obtains other work.
3. ***Enable:*** The assigned vertex makes ready the continuation of the vertex and unblocks another previously blocked thread by making a vertex from that thread ready. In this case, the process inserts both (any) one thread into the work pool and continues to execute the other.
4. ***Spawn:*** The assigned vertex spaws another thread. As in the previous case, the process inserts one thread into the work pool and continues to execute the other.

These actions are not mutually exclusive. For example, a thread may spawn/enable a thread and die. In this case, the process performs the corresponding steps for each action.

Exercise: Scheduling Invariant

Convince yourself that the scheduling invariant holds in online scheduling.

For a given schedule generated by an online scheduling algorithm, we can define a tree of vertices, which tell us far a vertex, the vertex that enabled it.

Definition: Enabling Tree

Consider the execution of a dag. If the execution of a vertex u enables another vertex v, then we call the edge (u,v) an ***enabling edge*** and we call u the ***enabling parent*** of v. For simplicity, we simply use the term ***parent*** instead of enabling parent.

Note that any vertex other than the root vertex has one enabling parent. Thus the subgraph induced by the enabling edges is a rooted tree that we call the ***enabling tree***.

4.7 Writing Multithreaded Programs: Pthreads

Multithreaded programs can be written using a variety of language abstractions interfaces. One of the most widely used interfaces is the ***POSIX Threads* or *Pthreads*** interface, which specifies a programming interface for a standardized C language in the IEEE POSIX 1003.1c standard. Pthreads provide a rich interface that enable the programmer to create multiple threads of control that can synchronize by using the nearly the whole range of the synchronization facilities mentioned above.

Hello world with Pthreads An example Pthread program is shown below. The main thread (executing function `main`) creates 8 child threads and terminates. Each child in turn runs the function `helloWorld` and immediately terminates. Since the main thread does not wait for the children to terminate, it may terminate before the children does, depending on how threads are scheduled on the available processors.

```
#include <iostream>
#include <cstdlib>
#include <pthread.h>

using namespace std;

#define NTHREADS 8

void *helloWorld(void *threadid)
{
   long tid;
   tid = (long)threadid;
   cout << "Hello world! It is me, 00" << tid << endl;
   pthread_exit(NULL);
}

int main ()
{
   pthread_t threads[NTHREADS];
   int rc;
   int i;
   for( i=0; i < NTHREADS; i++ ){
      cout << "main: creating thread 00" << i << endl;
      error = pthread_create(&threads[i], NULL, helloWorld, (void *) i);
      if (error) {
         cout << "Error: unable to create thread," << error << endl;
         exit(-1);
      }
   }
   pthread_exit(NULL);
}
```

When executed this program may print the following.

```
main: creating thread 000
main: creating thread 001
```

```
main: creating thread 002
main: creating thread 003
main: creating thread 004
main: creating thread 005
main: creating thread 006
main: creating thread 007
Hello world! It is me, 000
Hello world! It is me, 001
Hello world! It is me, 002
Hello world! It is me, 003
Hello world! It is me, 004
Hello world! It is me, 005
Hello world! It is me, 006
Hello world! It is me, 007
```

But that would be unlikely, a more likely output would look like this:

```
main: creating thread 000
main: creating thread 001
main: creating thread 002
main: creating thread 003
main: creating thread 004
main: creating thread 005
main: creating thread 006
main: creating thread 007
Hello world! It is me, 000
Hello world! It is me, 001
Hello world! It is me, 006
Hello world! It is me, 003
Hello world! It is me, 002
Hello world! It is me, 005
Hello world! It is me, 004
Hello world! It is me, 007
```

And may even look like this

```
main: creating thread 000
main: creating thread 001
main: creating thread 002
main: creating thread 003
Hello world! It is me, 000
Hello world! It is me, 001
Hello world! It is me, 003
Hello world! It is me, 002
main: creating thread 004
main: creating thread 005
main: creating thread 006
main: creating thread 007
Hello world! It is me, 006
Hello world! It is me, 005
Hello world! It is me, 004
Hello world! It is me, 007
```

4.8 Writing Multithreaded Programs: Structured or Implicit Multithreading

Interface such as Pthreads enable the programmer to create a wide variety of multithreaded computations that can be structured in many different ways. Large classes of interesting multithreaded computations, however, can be expcessed using a more structured approach, where threads are restricted in the way that they synchronize with other threads. One such interesting class of computations is fork-join computations where a thread can spawn or "fork" another thread or "join" with another existing

thread. Joining a thread is the only mechanism through which threads synchronize. The figure below illustrates a fork-join computation. The main threads forks thread A, which then spaws thread B. Thread B then joins thread A, which then joins Thread M.

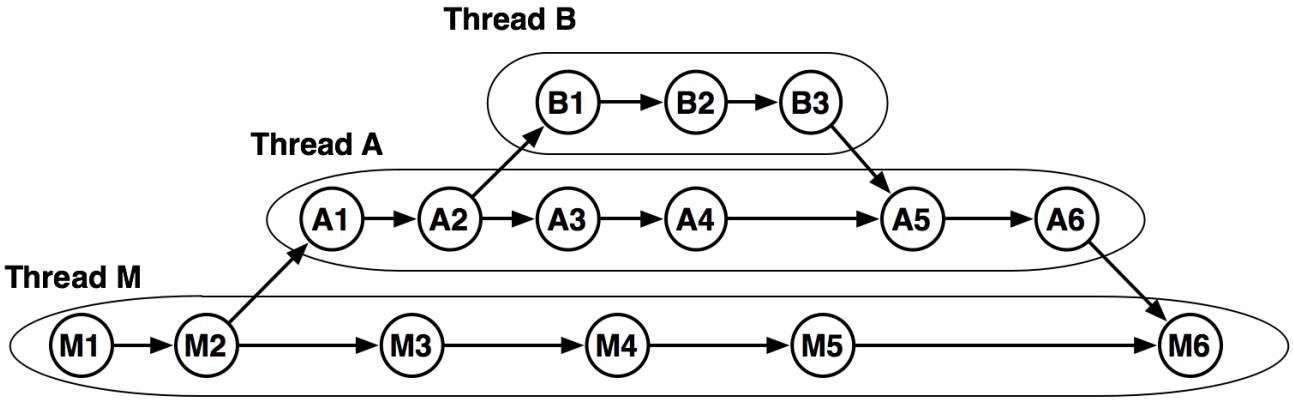


Figure 2: A multithreaded fork-join computation.

In addition to fork-join, there are other interfaces for structured multithreading such as async-finish, and futures. These interfaces are adopted in many programming languages: the Cilk language is primarily based on fork-join but also has some limited support for async-finish; X10 language is primarily based on async-finish but also supports futures; the Haskell language Haskell language provides support for fork-join and futures as well as others; Parallel ML language as implemented by the Manticore project is primarily based on fork-join parallelism. Such languages are sometimes called ***implicitly parallel***.

The class computations that can be expressed as fork-join and async-finish programs are sometimes called ***nested parallel***. The term "nested" refers to the fact that a parallel computation can be nested within another parallel computation. This is as opposed to ***flat parallelism*** where a parallel computation can only perform sequential computations in parallel. Flat parallelism used to be common technique in the past but becoming increasingly less prominent.

4.9 Parallelism versus concurrency

Structured multithreading offers important benefits both in terms of efficiency and expressiveness. Using programming constructs such as fork-join and futures, it is usually possible to write parallel programs such that the program accepts a "sequential semantics" but executes in parallel. The sequential semantics enables the programmer to treat the program as a serial program for the purposes of correctness. A run-time system then creates threads as necessary to execute the program in parallel. This approach offers is some ways the best of both worlds: the programmer can reason about correctness sequentially but the program executes in parallel. The benefit of structured multithreading in terms of efficiency stems from the fact that threads are restricted in the way that they communicate. This makes it possible to implement an efficient run-time system.

More precisely, consider some sequential language such as the untyped (pure) lambda calculus and its sequential dynamic semantics specified as a strict, small step transition relation. We can extend this language with the structured multithreading by enriching the syntax language with "fork-join" and "futures" constructs. We can now extend the dynamic semantics of the language in two ways: 1) trivially ignore these constructs and execute serially as usual, and 2) execute in parallel by creating parallel threads. We can then show that these two semantics are in fact identical, i.e., that they produce the same value for the same expressions. In other words, we can extend a rich programming language with fork-join and futures and still give the language a sequential semantics. This shows that structured multithreading is nothing but an efficiency and performance concern; it can be ignored from the perspective of correctness.

We use the term ***parallelism*** to refer to the idea of computing in parallel by using such structured multithreading constructs. As we shall see, we can write parallel algorithms for many interesting problems. While parallel algorithms or applications constitute a large class, they don't cover all applications. Specifically applications that can be expressed by using richer forms of multithreading such as the one offered by Pthreads do not always accept a sequential semantics. In such ***concurrent*** applications, threads can communicate and coordinate in complex ways to accomplish the intended result. A classic concurrency example is the "producer-consumer problem", where a consumer and a producer thread coordinate by using a fixed size buffer of items. The

producer fills the buffer with items and the consumer removes items from the buffer and they coordinate to make sure that the buffer is never filled more than it can take. We can use operating-system level processes instead of threads to implement similar concurrent applications.

In summary, parallelism is a property of the hardware or the software platform where the computation takes place, whereas concurrency is a property of the application. Pure parallelism can be ignored for the purposes of correctness; concurrency cannot be ignored for understanding the behavior of the program.

Parallelism and concurrency are orthogonal dimensions in the space of all applications. Some applications are concurrent, some are not. Many concurrent applications can benefit from parallelism. For example, a browser, which is a concurrent application itself as it may use a parallel algorithm to perform certain tasks. On the other hand, there is often no need to add concurrency to a parallel application, because this unnecessarily complicates software. It can, however, lead to improvements in efficiency.

The following quote from Dijkstra suggest pursuing the approach of making parallelism just a matter of execution (not one of semantics), which is the goal of the much of the work on the development of programming languages today. Note that in this particular quote, Dijkstra does not mention that parallel algorithm design requires thinking carefully about parallelism, which is one aspect where parallel and serial computations differ.

> From the past terms such as "sequential programming" and "parallel programming" are still with us, and we should try to get rid of them, for they are a great source of confusion. They date from the period that it was the purpose of our programs to instruct our machines, now it is the purpose of the machines to execute our programs. Whether the machine does so sequentially, one thing at a time, or with considerable amount of concurrency, is a matter of implementation, and should *not* be regarded as a property of the programming language.
>
> — Edsger W. Dijkstra *Selected Writings on Computing: A Personal Perspective (EWD 508)*

5 Chapter: Fork-join parallelism

Fork-join parallelism, a fundamental model in parallel computing, dates back to 1963 and has since been widely used in parallel computing. In fork join parallelism, computations create opportunities for parallelism by branching at certain points that are specified by annotations in the program text.

Each branching point ***forks*** the control flow of the computation into two or more logical threads. When control reaches the branching point, the branches start running. When all branches complete, the control ***joins*** back to unify the flows from the branches. Results computed by the branches are typically read from memory and merged at the join point. Parallel regions can fork and join recursively in the same manner that divide and conquer programs split and join recursively. In this sense, fork join is the divide and conquer of parallel computing.

As we will see, it is often possible to extend an existing language with support for fork-join parallelism by providing libraries or compiler extensions that support a few simple primitives. Such extensions to a language make it easy to derive a sequential program from a parallel program by syntactically substituting the parallelism annotations with corresponding serial annotations. This in turn enables reasoning about the semantics or the meaning of parallel programs by essentially "ignoring" parallelism.

PASL is a C++ library that enables writing implicitly parallel programs. In PASL, fork join is expressed by application of the `fork2()` function. The function expects two arguments: one for each of the two branches. Each branch is specified by one C++ lambda expression.

Example 5.1 Fork join

In the sample code below, the first branch writes the value 1 into the cell `b1` and the second 2 into `b2`; at the join point, the sum of the contents of `b1` and `b2` is written into the cell `j`.

```
long b1 = 0;
long b2 = 0;
long j  = 0;

fork2([&] {
  // first branch
  b1 = 1;
}, [&] {
  // second branch
```

```
  b2 = 2;
});
// join point
j = b1 + b2;

std::cout << "b1 = " << b1 << "; b2 = " << b2 << "; ";
std::cout << "j = " << j << ";" << std::endl;
```

Output:

```
b1 = 1; b2 = 2; j = 3;
```

When this code runs, the two branches of the fork join are both run to completion. The branches may or may not run in parallel (i.e., on different cores). In general, the choice of whether or not any two such branches are run in parallel is chosen by the PASL runtime system. The join point is scheduled to run by the PASL runtime only after both branches complete. Before both branches complete, the join point is effectively blocked. Later, we will explain in some more detail the scheduling algorithms that the PASL uses to handle such load balancing and synchronization duties.

In fork-join programs, a thread is a sequence of instructions that do not contain calls to `fork2()`. A thread is essentially a piece of sequential computation. The two branches passed to `fork2()` in the example above correspond, for example, to two independent threads. Moreover, the statement following the join point (i.e., the continuation) is also a thread.

Note
If the syntax in the code above is unfamiliar, it might be a good idea to review the notes on lambda expressions in C++11. In a nutshell, the two branches of `fork2()` are provided as lambda-expressions where all free variables are passed by reference.

Note
Fork join of arbitrary arity is readily derived by repeated application of binary fork join. As such, binary fork join is universal because it is powerful enough to generalize to fork join of arbitrary arity.

All writes performed by the branches of the binary fork join are guaranteed by the PASL runtime to commit all of the changes that they make to memory before the join statement runs. In terms of our code snippet, all writes performed by two branches of `fork2` are committed to memory before the join point is scheduled. The PASL runtime guarantees this property by using a local barrier. Such barriers are efficient, because they involve just a single dynamic synchronization point between at most two processors.

Example 5.2 Writes and the join statement

In the example below, both writes into `b1` and `b2` are guaranteed to be performed before the print statement.

```
long b1 = 0;
long b2 = 0;

fork2([&] {
  b1 = 1;
}, [&] {
  b2 = 2;
});

std::cout << "b1 = " << b1 << "; b2 = " << b2 << std::endl;
```

Output:

```
b1 = 1; b2 = 2
```

PASL provides no guarantee on the visibility of writes between any two parallel branches. In the code just above, for example, writes performed by the first branch (e.g., the write to `b1`) may or may not be visible to the second, and vice versa.

5.1 Parallel Fibonacci

Now, we have all the tools we need to describe our first parallel code: the recursive Fibonacci function. Although useless as a program because of efficiency issues, this example is the "hello world" program of parallel computing.

Recall that the n^{th} Fibonnacci number is defined by the recurrence relation

$$F(n) \quad = \quad F(n-1)+F(n-2)$$

with base cases

$$F(0)=0, F(1)=1$$

Let us start by considering a sequential algorithm. Following the definition of Fibonacci numbers, we can write the code for (inefficiently) computing the n^{th} Fibonnacci number as follows. This function for computing the Fibonacci numbers is inefficient because the algorithm takes exponential time, whereas there exist dynamic programming solutions that take linear time.

```
long fib_seq (long n) {
  long result;
  if (n < 2) {
    result = n;
  } else {
    long a, b;
    a = fib_seq(n-1);
    b = fib_seq(n-2);
    result = a + b;
  }
  return result;
}
```

To write a parallel version, we remark that the two recursive calls are completely ***independent***: they do not depend on each other (neither uses a piece of data generated or written by another). We can therefore perform the recursive calls in parallel. In general, any two independent functions can be run in parallel. To indicate that two functions can be run in parallel, we use `fork2()`.

```
long fib_par(long n) {
  long result;
  if (n < 2) {
    result = n;
  } else {
    long a, b;
    fork2([&] {
      a = fib_par(n-1);
    }, [&] {
      b = fib_par(n-2);
    });
    result = a + b;
  }
  return result;
}
```

5.2 Incrementing an array, in parallel

Suppose that we wish to map an array to another by incrementing each element by one. We can write the code for a function `map_incr` that performs this computation serially.

```
void map_incr(const long* source, long* dest, long n) {
  for (long i = 0; i < n; i++)
    dest[i] = source[i] + 1;
}
```

Example 5.3 Example: Using `map_incr`.

The code below illustrates an example use of `map_incr`.

```
const long n = 4;
long xs[n] = { 1, 2, 3, 4 };
long ys[n];
map_incr(xs, ys, n);
for (long i = 0; i < n; i++)
  std::cout << ys[i] << " ";
std::cout << std::endl;
```

Output:

```
2 3 4 5
```

This is not a good parallel algorithm but it is not difficult to give a parallel algorithm for incrementing an array. The code for such an algorithm is given below.

```
void map_incr_rec(const long* source, long* dest, long lo, long hi) {
  long n = hi - lo;
  if (n == 0) {
    // do nothing
  } else if (n == 1) {
    dest[lo] = source[lo] + 1;
  } else {
    long mid = (lo + hi) / 2;
    fork2([&] {
      map_incr_rec(source, dest, lo, mid);
    }, [&] {
      map_incr_rec(source, dest, mid, hi);
    });
  }
}
```

It is easy to see that this algorithm has $O(n)$ work and $O(\log n)$ span.

5.3 The sequential elision

In the Fibonacci example, we started with a sequential algorithm and derived a parallel algorithm by annotating independent functions. It is also possible to go the other way and derive a sequential algorithm from a parallel one. As you have probably guessed this direction is easier, because all we have to do is remove the calls to the `fork2` function. The sequential elision of our parallel Fibonacci code can be written by replacing the call to `fork2()` with a statement that performs the two calls (arguments of `fork2()`) sequentially as follows.

```
long fib_par(long n) {
  long result;
  if (n < 2) {
    result = n;
  } else {
    long a, b;
    ([&] {
      a = fib_par(n-1);
    })();
    ([&] {
      b = fib_par(n-2);
    })();
    result = a + b;
  }
  return result;
}
```

Note
Although this code is slightly different than the sequential version that we wrote, it is not too far away, because the only the difference is the creation and application of the lambda-expressions. An optimizing compiler for C++ can easily "inline" such computations. Indeed, After an optimizing compiler applies certain optimizations, the performance of this code the same as the performance of `fib_seq`.

The sequential elision is often useful for debugging and for optimization. It is useful for debugging because it is usually easier to find bugs in sequential runs of parallel code than in parallel runs of the same code. It is useful in optimization because the sequentialized code helps us to isolate the purely algorithmic overheads that are introduced by parallelism. By isolating these costs, we can more effectively pinpoint inefficiencies in our code.

5.4 Executing fork-join algorithms

We defined fork-join programs as a subclass case of multithreaded programs. Let's see more precisely how we can "map" a fork-join program to a multithreaded program. An our running example, let's use the `map_incr_rec`, whose code is reproduced below.

```
void map_incr_rec(const long* source, long* dest, long lo, long hi) {
  long n = hi - lo;
  if (n == 0) {
    // do nothing
  } else if (n == 1) {
    dest[lo] = source[lo] + 1;
  } else {
    long mid = (lo + hi) / 2;
    fork2([&] {
      map_incr_rec(source, dest, lo, mid);
    }, [&] {
      map_incr_rec(source, dest, mid, hi);
    });
  }
}
```

The key question is this: what is a thread? Clearly, when we write a multithreaded program, we don't manipulate explicitly threads. The basic idea is to partition a computation, which is a run of a parallel algorithm on a specified input, into pieces of serial computations, and define these pieces as threads.

We call a piece of serial computation a ***thread***, if it executes without performing parallel operations (`fork2`) except perhaps as its last action. The term thread is short for ***user-level thread*** (as opposed to a operating-system thread). When partitioning the computation into threads, it is important for threads to be maximal; technically a thread can be as small as a single instruction.

Definition: Thread

A **thread** is a maximal computation consisting of a sequence of instructions that do not contain calls to `fork2()` except perhaps at the very end.

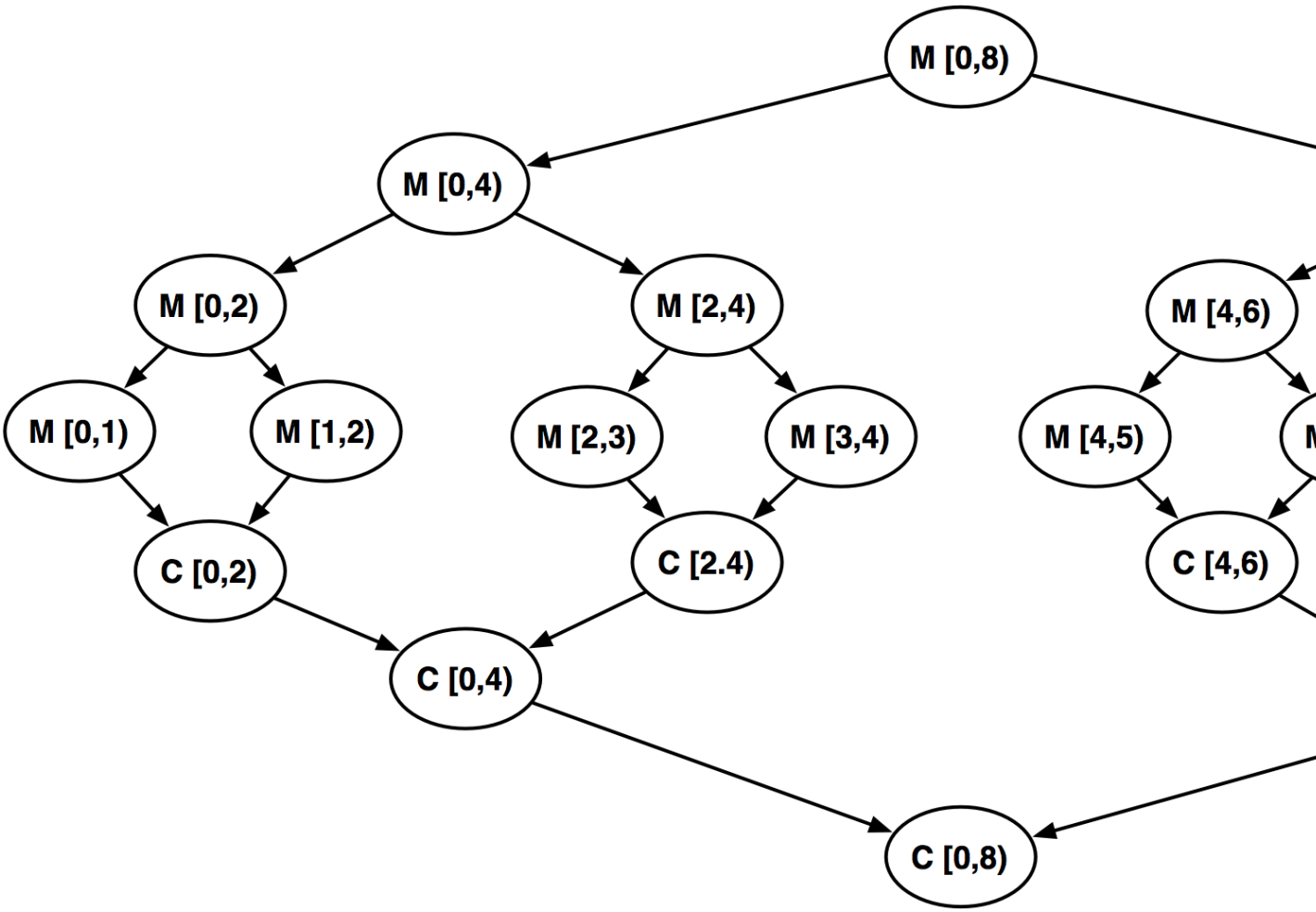


Figure 3: DAG for parallel increment on an array of 8: Each vertex corresponds a call to `map_inc_rec` excluding the `fork2` or the continuation of `fork2`, which is empty, an is annotated with the interval of the input array that it operates on (its argument).

The Figure above illustrates the DAG for an execution of `map_incr_rec`. We partition each invocation of this function into two threads labeled by "M" and "C" respectively. The threads labeled by $M[i, j]$ corresponds to the part of the invocation of `map_incr_rec` with arguments `lo` and `hi` set to i and j respectively; this first part corresponds to the part of execution up and including the `fork2` or all of the function if this is a base case. The second corresponds to the "continuation" of the `fork2`, which is in this case includes no computation.

Based on this dag, we can create another dag, wherea each thread is replaced by the sequence of instructions that it represents. This would give us a picture similar to the dag we drew before for general multithreaded programs. Such a dag representation, where we represent each instruction by a vertex, gives us a direct way to calculate the work and span of the computation. If we want to calculate work and span ond the dag of threads, we can label each vertex with a weight that corresponds to the number of instruction in that thread.

Having talked about DAG's we are now ready to talk about how to map parallel computations to actual hardware so as to minimize their run-time, i.e., scheduling.

```
 But before we move on to scheduling let
us observe a few properties of implicitly parallel computations.
```

1. The computation DAG of a parallel algorithm applied to an input unfolds dynamically as the algorithm executes. For example, when we run `map_inc_rec` with an input with n elements, the DAG initially contains just the root vertex (thread) corresponding to the first call to `map_inc_rec` but it grows as the execution proceeds.

2. An execution of a parallel algorithm can generate a massive number of threads. For example, our 'map_inc_rec' function generates approximately $4n$ threads for an input with n elements.

3. The work/span of each thread can vary from a small amount to a very large amount depending on the algorithm. In our example, each thread performs either a conditional, sometimes an addition and a fork operation or performs no actual computation (continuation threads).

Suppose now we are given a computation DAG and we wish to execute the DAG by mapping each thread to one of the P processor that is available on the hardware. When a thread is mapped to a processor, it will be executed requiring time proportional to the work (weight) of the thread. Once the processor completes the thread, we can map another thread to the processor, so that the processor does not idle unnecessarily. As we map threads to processors, it is important that we observe the dependencies between threads, i.e., we don't execute a thread before its parents.

, and map them to available processors while observing the dependencies between them. The task of mapping the threads to available processors is called ***thread scheduling*** or simply ***scheduling***.

When scheduling a parallel computation, it is important that we don't alter the intended meaning of the computation. Specifically, if a thread depends another thread, because for example, it reads a piece of data generated by the latter, it cannot be executed before the thread that it depends on. We can conservatively approximate such dependencies by observing the `fork2` expressions and by organizing dependencies consistently with them. More specifically, we can represent a computations as a graph where each vertex represents a thread and each edge represents a dependency. Vertices and edges are created by execution of `fork2`. Each `fork2` creates two threads (vertices) corresponding to the two branches and inserts an edge between each branch and the thread that performs the `fork2` branches; in addition, each `fork2` creates a join or continuation thread (vertex) that depends on the two branches. Since such a graph cannot contain cycles, it is a ***Directed Acyclic Graph (DAG)***.

Definition: Scheduling

The (thread) scheduling problem requires assigning to each thread in a given DAG a processor number and a time step such that

1. each thread is assigned to a unique processor for as many consecutive steps as its weight,
2. no thread is executed before its descendants in the DAG, and
3. no processor is assigned more at most one thread at a time.

The goal of scheduling to minimize critical resources such as time. Computing the shortest schedule for a DAG turns out to be highly nontrivial. In fact, the related decision problem in NP-complete. It is possible, however, to give a good approximation algorithm for the offline version of the problem to generate a 2-factor approximation. Such an appraximation yields a schedule for a given DAG within a factor-two of the shortest schedule. In the online version of the problem, where the DAG unfolds as the computation executes, we don't know the DAG a priori and we have to account for the costs for scheduling such as migrating threads between schedulers and finding work. To execute parallel programs, we need an solution to this online version of the problem.

An online scheduler or a simply a ***scheduler*** is an algorithm that performs scheduling by mapping threads to available processors. For example, if only one processor is available, a scheduler can map all threads to that one processor. If two processors are available, then the scheduler can divide the threads between the two processors as evenly as possible in an attempt to keep the two processors as busy as possible by ***load balancing***.

Example 5.4 An example 2-processor schedule

The following is a valid schedule for the DAG shown in this Figure assuming that each thread takes unit time.

Time Step	Processor 1	Processor 2
1	M [0,8)	
2	M [0,4)	M [4,8)
3	M [0,2)	M [4,6)
4	M [0,1)	M [4,5)
5	M [1,2)	M [5,6)
6	C [0,2)	C [4,6)

Time Step	Processor 1	Processor 2
7	M [2,4)	M [6,8)
8	M [2,3)	M [6,7)
9	M [3,4)	M [7,8)
10	C [2,4)	C [6,8)
11	C [0,4)	C [4,8)
12	C [0,8)	_

We say that a scheduler is ***greedy*** if, whenever there is a processor available and a thread ready to be executed, then the scheduler assigns the thread to the processor and starts running the thread immediately. Greedy schedulers have a nice property that is summarized by the following theorem.

Theorem: Greedy Scheduling Principle

If a computation is run on P processors using a perfect greedy scheduler that incurs no costs in creating, locating, and moving threads, then the total time (clock cycles) for running the computation T_P is bounded by

$$T_P < \frac{W}{p} + S.$$

Here W is the work of the computation, and S is the span of the computation (both measured in units of clock cycles).

This simple statement is powerful. To see this, note that the time to execute the computation is at least $\frac{W}{P}$ because we have a total of W work. As such, the best possible execution strategy is to divide it evenly among the processors. Furthermore, execution time cannot be less than S since S represents the longest chain of sequential dependencies. Thus we have: $T_P \geq \max\left(\frac{W}{P}, S\right)$.

This means that a greedy schudeler yields a schedule that is within a factor two of optimal: $\frac{W}{P} + S$ is never more than twice $\max(\frac{W}{P}, S)$. Furthermore, when $\frac{W}{P} \gg S$, the difference between the greedy scheduler and the optimal scheduler is very small. In fact, we can rewrite equation above in terms of the average parallelism $\mathbb{P} = W/S$ as follows:

$$\begin{aligned} T_p &< \frac{W}{P} + S \\ &= \frac{W}{P} + \frac{W}{\mathbb{P}} \\ &= \frac{W}{P}\left(1 + \frac{p}{\mathbb{P}}\right) \end{aligned}$$

Therefore as long as $\mathbb{P} \gg P$ (the parallelism is much greater than the number of processors), then we obtain near perfect speedup. (Speedup is W/T_p and perfect speedup would be p).

The quantity $\mathbb{P}$, sometimes called ***average parallelism***, is usually quite large, because it usually grows polynomially with the input size.

Example 5.5 Scheduler with a global thread queue.

We can give a simple greedy scheduler by using a queue of threads. At the start of the execution, the scheduler places the root of the DAG into the queue and then repeats the following step until the queue becomes empty: for each idle processor, take the vertex at the front of the queue and assign it to the processor, let each processor run for one step, if at the end of the step, there is a vertex in the DAG whose parents have all completed their execution, then insert that vertex at the tail of the queue.

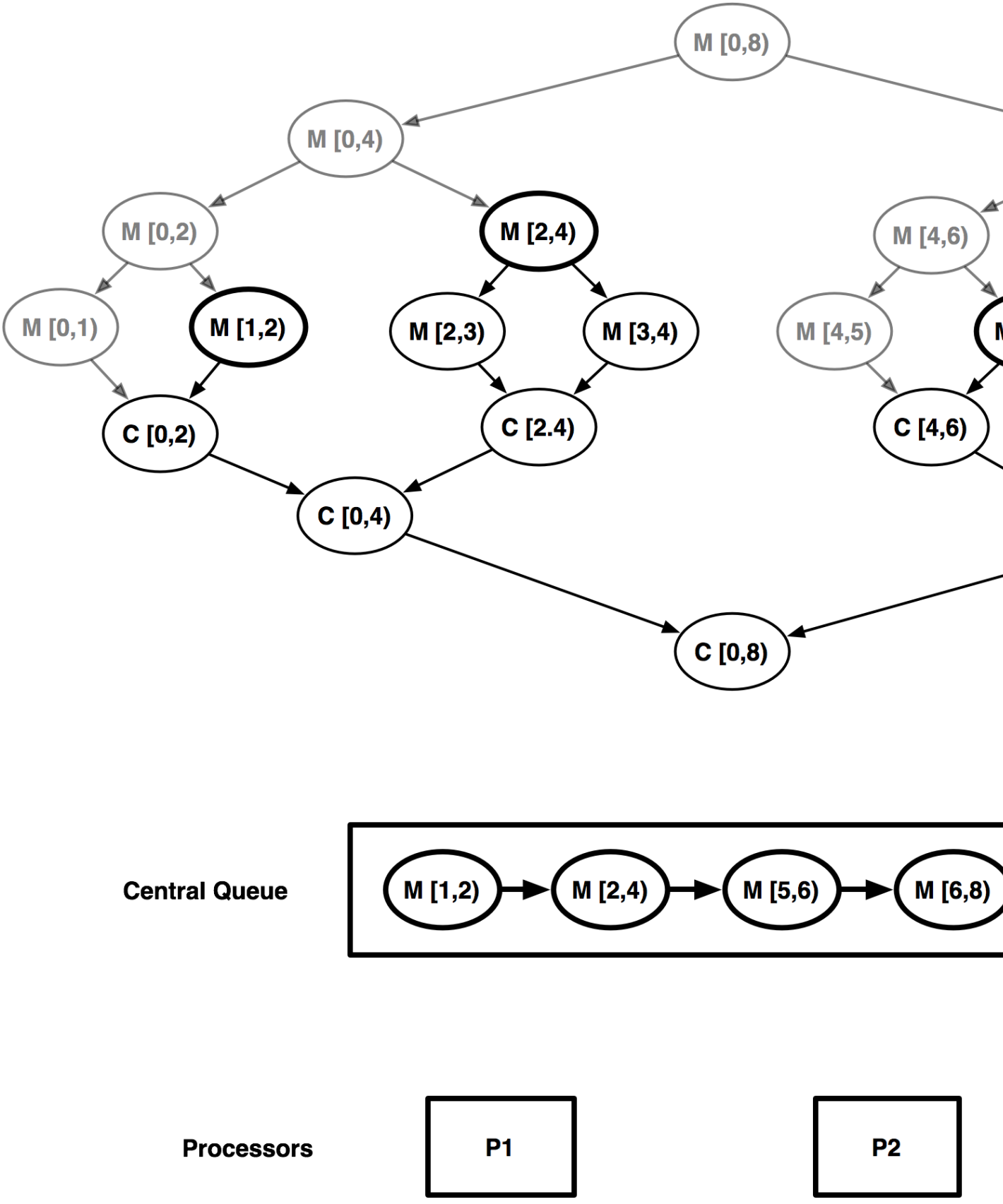


Figure 4: Centralized scheduler illustrated: the state of the queue and the DAG after step 4. Completed vertices are drawn in grey (shaded).

The centralized scheduler with the global thread queue is a greedy scheduler that generates a greedy schedule under the assumption that the queue operations take zero time and that the DAG is given. This algorithm, however, does not work well for online scheduling the operations on the queue take time. In fact, since the thread queue is global, the algorithm can only insert and remove one thread at a time. For this reason, centralized schedulers do not scale beyond a handful of processors.

Definition: Scheduling friction.

No matter how efficient a scheduler is there is real cost to creating threads, inserting and deleting them from queues, and to performing load balancing. We refer to these costs cumulatively as ***scheduling friction***, or simply as ***friction***.

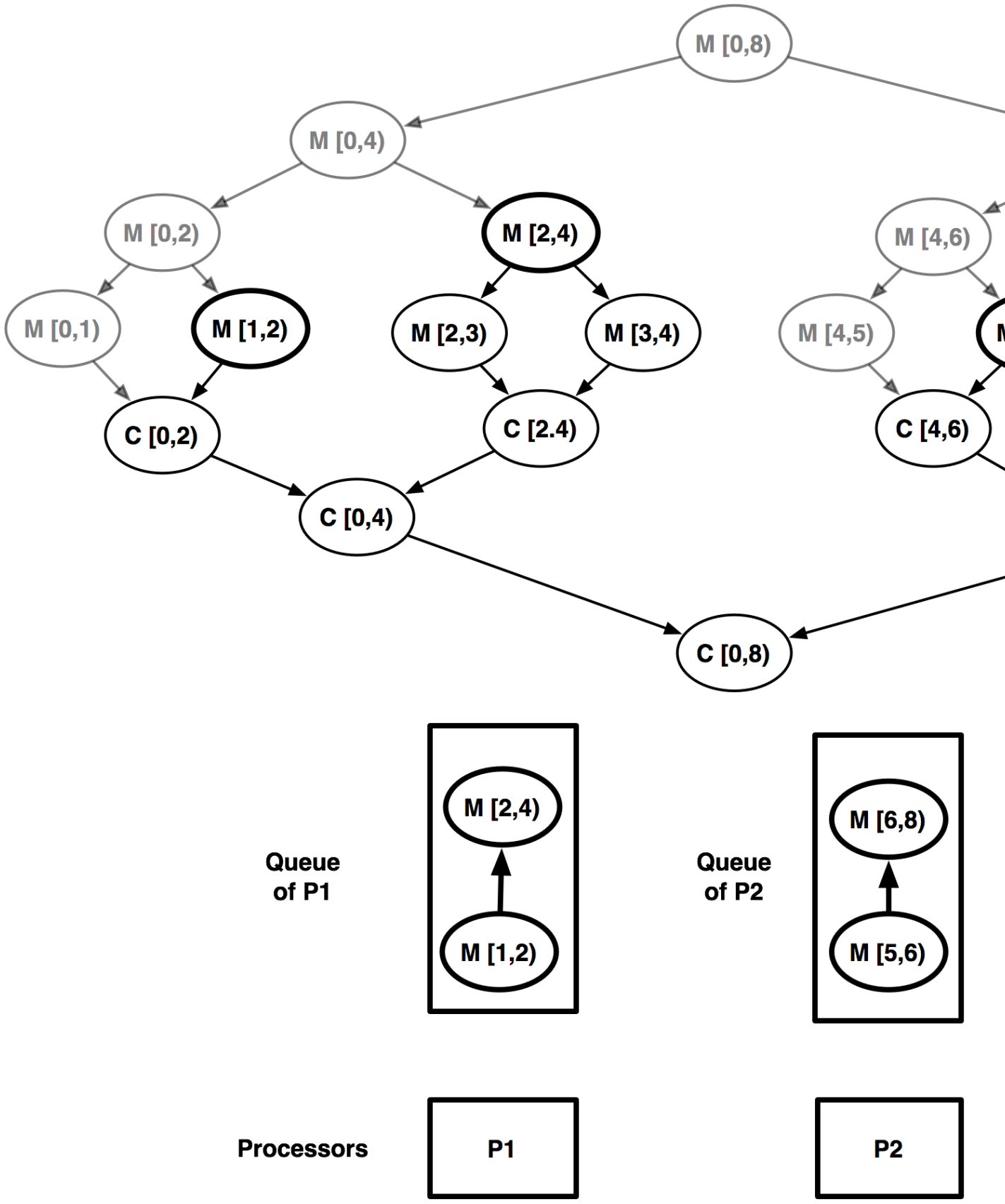


Figure 5: Distributed scheduler illustrated: the state of the queue and the DAG after step 4. Completed vertices are drawn in grey (shaded).

There has been much research on the problem of reducing friction in scheduling. This research shows that distrubuted scheduling algorithms can work quite well. In a distributed algorithm, each processor has its own queue and primarily operates on its own queue. A load-balancing technique is then used to balance the load among the existing processors by redistributing threads, usually on a needs basis. This strategy ensures that processors can operate in parallel to obtain work from their queues.

A specific kind of distributed scheduling technique that can leads to schedules that are close to optimal is ***work stealing*** schedulers. In a work-stealing scheduler, processors work on their own queues as long as their is work in them, and if not, go "steal" work from other processors by removing the thread at the tail end of the queue. It has been proven that randomized work-stealing algorithm, where idle processors randomly select processors to steal from, deliver close to optimal schedules in expectation (in fact with high probability) and furthermore incur minimal friction. Randomized schedulers can also be implemented efficiently in practice. PASL uses an scheduling algorithm that is based on work stealing.

6 Critical Sections and Mutual Exclusion

In a multithreaded program, a ***critical section*** is a part of the program that may not be executed by more than one thread at the same time. Critical sections typically contain code that alters shared objects, such as shared (e.g., global) variables. This means that the a critical section requires ***mutual exclusion***: only one thread can be inside the critical section at any time.

Since only one thread can be inside a critical section at a time, threads must coordinate to make sure that they don't enter the critical section at the same time. If threads do not coordinate and multiple threads enter the critical section at the same time, we say that a ***race condition*** occurs, because the outcome of the program depends on the relative timing of the threads, and thus can vary from one execution to another. Race conditions are sometimes benign but usually not so, because they can lead to incorrect behavior. Spectacular examples of race conditions' effects include the "Northeast Blackout" of 2003, which affected 45 million people in the US and 10 million people in Canada.

It can be extremely difficult to find a race condition, because of the non-determinacy of execution. A race condition may lead to an incorrect behavior only a tiny fraction of the time, making it extremely difficult to observe and reproduce it. For example, the software fault that lead to the Northeast blackout took software engineers "weeks of poring through millions of lines of code and data to find it" according to one of the companies involved.

The problem of designing algorithms or protocols for ensuring mutual exclusion is called the ***mutual exclusion problem*** or the ***critical section*** problem. There are many ways of solving instances of the mutual exclusion problem. But broadly, we can distinguish two categories: spin-locks and blocking-locks. The idea in ***spin locks*** is to busy wait until the critical section is clear of other threads. Solutions based on ***blocking locks*** is similar except that instead of waiting, threads simply block. When the critical section is clear, a blocked thread receives a signal that allows it to proceed. The term ***mutex***, short for "mutual exclusion" is sometimes used to refer to a lock.

Mutual exclusions problems have been studied extensively in the context of several areas of computer science. For example, in operating systems research, processes, which like threads are independent threads of control, belonging usually but not always to different programs, can share certain systems' resources. To enable such sharing safely and efficiently, researchers have proposed various forms of locks such as ***semaphores***, which accepts both a busy-waiting and blocking semantics. Another class of locks, called ***condition variables*** enable blocking synchronization by conditioning an the value of a variable.

6.1 Parallelism and Mutual Exclusion

In parallel programming, mutual exclusion problems do not have to arise. For example, if we program in a purely functional language extended with structured multithreading primitives such as fork-join and futures, programs remain purely functional and mutual-exclusion problems, and hence race conditions, do not arise. If we program in an imperative language, however, where memory is always a shared resource, even when it is not intended to be so, threads can easily share memory objects, even unintentionally, leading to race conditions. .Writing to the same location in parallel.

In the code below, both branches of `fork2` are writing into `b`. What should then the output of this program be?

```
long b = 0;

fork2([&] {
  b = 1;
}, [&] {
```

```
  b = 2;
});

std::cout << "b = " << std::endl;
```

At the time of the print, the contents of `b` is determined by the last write. Thus depending on which of the two branches perform the write, we can see both possibilities:

Output:

```
b = 1
```

Output:

```
b = 2
```

Example 6.1 Fibonacci

Consider the following alternative implementation of the Fibonacci function. By "inlining" the plus operation in both branches, the programmer got rid of the addition operation after the `fork2`.

```
long fib_par_racy(long n) {
  long result = 0;
  if (n < 2) {
    result = n;
  } else {
    fork2([&] {
      result += fib_par_racy(n-1);
    },[&] {
      result += fib_par_racy(n-2);
    });
  }
  return result;
}
```

This code is not correct because it has a race condition.

As in the example shows, separate threads are updating the value result but it might look like this is not a race condition because the update consists of an addition operation, which reads the value and then writes to i. The race condition might be easier to see if we expand out the applications of the += operator.

```
long fib_par_racy(long n) {
  long result = 0;
  if (n < 2) {
    result = n;
  } else {
    fork2([&] {
      long a1 = fib_par_racy(n-1);
      long a2 = result;
      result = a1 + a2;
    },[&] {
      long b1 = fib_par_racy(n-2);
      long b2 = result;
      result = b1 + b2;
    });
  }
  return result;
}
```

When written in this way, it is clear that these two parallel threads are not independent: they both read `result` and write to `result`. Thus the outcome depends on the order in which these reads and writes are performed, as shown in the next example.

Example 6.2 Execution trace of a race condition

The following table takes us through one possible execution trace of the call `fib_par_racy(2)`. The number to the left of each instruction describes the time at which the instruction is executed. Note that since this is a parallel program, multiple instructions can be executed at the same time. The particular execution that we have in this example gives us a bogus result: the result is 0, not 1 as it should be.

Time step	Thread 1	Thread 2
1	a1 = fib_par_racy(1)	b2 = fib_par_racy(0)
2	a2 = result	b3 = result
3	result = a1 + a2	_
4	_	result = b1 + b2

The reason we get a bogus result is that both threads read the initial value of result at the same time and thus do not see each others write. In this example, the second thread "wins the race" and writes into `result`. The value 1 written by the first thread is effectively lost by being overwritten by the second thread.

6.2 Synchronization Hardware

Since mutual exclusion is a common problem in computer science, many hardware systems provide specific synchronization operations that can help solve instances of the problem. These operations may allow, for example, testing the contents of a (machine) word then modifying it, perhaps by swapping it with another word. Such operations are sometimes called atomic ***read-modify-write*** or ***RMW***, for short, operations.

A handful of different RMW operations have been proposed. They include operations such as ***load-link/store-conditional***, ***fetch-and-add***, and ***compare-and-swap***. They typically take the memory location `x`, and a value `v` and replace the value of stored at `x` with `f(x,v)`. For example, the fetch-and-add operation takes the location `x` and the increment-amount, and atomically increments the value at that location by the specified amount, i.e., `f(x,v) =*x + v`.

The compare-and-swap operation takes the location `x` and takes a pair of values `(a,b)` as the second argument, and stores `b` into `x` if the value in `x` is `a`, i.e., `f(x,(a,b)) =if *x =a then b else a`; the operation returns a Boolean indicating whether the operation successfully stored a new value in `x`. The operation "compare-and-swap" is a reasonably powerful synchronization operation: it can be used by arbitrarily many threads to agree (reach consensus) on a value. This instruction therefore is frequently provided by modern parallel architectures such as Intel's X86.

In C++, the `atomic` class can be used to perform synchronization. Objects of this type are guarantee to be free of race conditions; and in fact, in C++, they are the only objects that are guaranteed to be free from race conditions. The contents of an `atomic` object can be accessed by `load` operations, updated by `store` operation, and also updated by `compare_exchange_weak` and `compare_exchange_strong` operations, the latter of which implement the compare-and-swap operation.

Example 6.3 Accessing the contents of atomic memory cells

Access to the contents of any given cell is achieved by the `load()` and `store()` methods.

```
std::atomic<bool> flag;

flag.store(false);
std::cout << flag.load() << std::endl;
flag.store(true);
std::cout << flag.load() << std::endl;
```

Output:

```
0
1
```

The key operation that help with race conditions is the compare-and-exchange operation.

Definition: compare and swap

When executed with a 'target` atomic object and an `expected` cell and a new value `new' this operation performs the following steps, atomically:

1. Read the contents of `target`.
2. If the contents equals the contents of `expected`, then writes `new` into the `target` and returns `true`.
3. Otherwise, returns `false`.

Example 6.4 Reading and writing atomic objects

```
std::atomic<bool> flag;

flag.store(false);
bool expected = false;
bool was_successful = flag.compare_exchange_strong(expected, true);
std::cout << "was_successful = " << was_successful << "; flag = " << flag.load() << std:: ↩
    endl;
bool expected2 = false;
bool was_successful2 = flag.compare_exchange_strong(expected2, true);
std::cout << "was_successful2 = " << was_successful2 << "; flag = " <<
flag.load() << std::endl;
```

Output:

```
was_successful = 1; flag = 1
was_successful2 = 0; flag = 1
```

As another example use of the `atomic` class, recall our Fibonacci example with the race condition. In that example, race condition arises because of concurrent writes to the `result` variable. We can eliminate this kind of race condition by using different memory locations, or by using an atomic class and using a `compare_exchange_strong` operation.

Example 6.5 Fibonacci

The following implementation of Fibonacci is not safe because the variable `result` is shared and updated by multiple threads.

```
long fib_par_racy(long n) {
  long result = 0;
  if (n < 2) {
    result = n;
  } else {
    fork2([&] {
      result += fib_par_racy(n-1);
    },[&] {
      result += fib_par_racy(n-2);
    });
  }
  return result;
}
```

We can solve this problem by declaring `result` to be an atomic type and using a standard busy-waiting protocol based on compare-and-swap.

```
long fib_par_atomic(long n) {
  atomic<long> result = 0;
  if (n < 2) {
    result.store(n);
  } else {
    fork2([&] {
      long r = fib_par_racy(n-1);
```

```
      // Atomically update result.
      while (true) {
        long exp = result.load();
        bool flag = result.compare_exchange_strong(exp,exp+r)
        if (flag) {break;}
      }
    },[&] {
      long r = fib_par_racy(n-2);
      // Atomically update result.
      while (true) {
        long exp = result.load();
        bool flag = result.compare_exchange_strong(exp,exp+r)
        if (flag) {break;}
      }
    });
  }
  return result;
}
```

The idea behind the solution is to load the current value of `result` and atomically update `result` only if it has not been modified (by another thread) since it was loaded. This guarantees that the `result` is always updated (read and modified) correctly without missing an update from another thread.

The example above illustrates a typical use of the compare-and-swap operation. In this particular example, we can probably prove our code is correct. But this is not always as easy due to a problem called the "ABA problem."

6.3 ABA problem

While reasonably powerful, compare-and-swap suffers from the so-called ***ABA*** problem. To see this consider the following scenario where a shared variable `result` is update by multiple threads in parallel: a thread, say T, reads the `result` and stores its current value, say `2`, in `current`. In the mean time some other thread also reads `result` and performs some operations on it, setting it back to `2` after it is done. Now, thread T takes its turn again and attempts to store a new value into `result` by using `2` as the old value and being successful in doing so, because the value stored in `result` appears to have not changed. The trouble is that the value has actually changed and has been changed back to the value that it used to be. Thus, compare-and-swap was not able to detect this change because it only relies on a simple shallow notion of equality. If for example, the value stored in `result` was a pointer, the fact that the pointer remains the same does not mean that values accessible from the pointer has not been modified; if for example, the pointer led to a tree structure, an update deep in the tree could leave the pointer unchanged, even though the tree has changed.

This problem is called the ***ABA*** problem, because it involves cycling the atomic memory between the three values A, B, and again A). The ABA problem is an important limitation of compare-and-swap: the operation itself is not atomic but is able to behave as if it is atomic if it can be ensured that the equality test of the subject memory cell suffices for correctness.

In the example below, ABA problem may happen (if the counter is incremented and decremented again in between a load and a store) but it is impossible to observe because it is harmless. If however, the compare-and-swap was on a memory object with references, the ABA problem could have had observable effects.

The ***ABA*** problem can be exploited to give seemingly correct implementations that are in fact incorrect. To reduce the changes of bugs due to the ABA problem, memory objects subject to compare-and-swap are usually tagged with an additional field that counts the number of updates. This solves the basic problem but only up to a point because the counter itself can also wrap around. The load-link/store-conditional operation solves this problem by performing the write only if the memory location has not been updated since the last read (load) but its practical implementations are hard to come by.

7 Chapter: Experimenting with PASL

We are now going to study the practical performance of our parallel algorithms written with PASL on multicore computers.

To be concrete with our instructions, we assume that our username is `pasl` and that our home directory is `/home/pasl/`. You need to replace these settings with your own where appropriate.

7.1 Obtain source files

Let's start by downloading the PASL sources. The PASL sources that we are going to use are part of a branch that we created specifically for this course. You can access the sources either via the tarball linked by the github webpage or, if you have `git`, via the command below.

```
$ cd /home/pasl
$ git clone -b edu https://github.com/deepsea-inria/pasl.git
```

7.2 Software Setup

You can skip this section if you are using a computer already setup by us or you have installed an image file containing our software. To skip this part and use installed binaries, see the heading "Starting with installed binaries", below.

7.2.1 Check for software dependencies

Currently, the software associated with this course supports Linux only. Any machine that is configured with a recent version of Linux and has access to at least two processors should be fine for the purposes of this course. Before we can get started, however, the following packages need to be installed on your system.

Software dependency	Version	Nature of dependency
`gcc`	>= 4.9.0	required to build PASL binaries
`php`	>= 5.3.10	required by PASL makefiles to build PASL binaries
`ocaml`	>= 4.0.0	required to build the benchmarking tools (i.e., pbench and pview)
`R`	>= 2.4.1	required by benchmarking tools to generate reports in bar plot and scatter plot form
`latex`	recent	optional; required by benchmarking tools to generate reports in tabular form
`git`	recent	optional; can be used to access PASL source files
`tcmalloc`	>= 2.2	optional; may be useful to improve performance of PASL binaries
`hwloc`	recent	optional; might be useful to improve performance on large systems with NUMA (see below)

The rest of this section explains what are the optional software dependencies and how to configure PASL to use them. We are going to assume that all of these software dependencies have been installed in the folder `/home/pasl/Installs/`.

7.2.2 Use a custom parallel heap allocator

At the time of writing this document, the system-default implementations of `malloc` and `free` that are provided by Linux distributions do not scale well with even moderately large amounts of concurrent allocations. Fortunately, for this reason, organizations, such as Google and Facebook, have implemented their own scalable allocators that serve as drop-in replacements for `malloc` and `free`. We have observed the best results from Google's allocator, namely, tcmalloc. Using tcmalloc for your own experiements is easy. Just add to the `/home/pasl/pasl/minicourse` folder a file named `settings.sh` with the following contents.

Example 7.1 Configuration to select tcmalloc

We assume that the package that contains `tcmalloc`, namely `gperftools`, has been installed already in the folder `/home/pasl/Installs/gperftools-install/`. The following lines need to be in the `settings.sh` file in the `/home/pasl/pasl/minicourse` folder.

```
USE_ALLOCATOR=tcmalloc
TCMALLOC_PATH=/home/pasl/Installs/gperftools-install/lib/
```

Also, the environment linkder needs to be instructed where to find `tcmalloc`.

```
export LD_PRELOAD=/home/pasl/Installs/gperftools-install/lib/libtcmalloc.so
```

This assignment can be issued either at the command line or in the environment loader script, e.g., `~/.bashrc`.

Warning
Changes to the `settings.sh` file take effect only after recompiling the binaries.

7.2.3 Use `hwloc`

If your system has a non-uniform memory architecture (i.e., NUMA), then you may improve performance of PASL applications by using optional support for `hwloc`, which is a library that reports detailed information about the host system, such as NUMA layout. Currently, PASL leverages `hwloc` to configure the NUMA allocation policy for the program. The particular policy that works best for our applications is round-robin NUMA page allocation. Do not worry if that term is unfamiliar: all it does is disable NUMA support, anyway!

Example 7.2 How to know whether my machine has NUMA

Run the following command.

```
$ dmesg | grep -i numa
```

If the output that you see is something like the following, then your machine has NUMA. Otherwise, it probably does not.

```
[    0.000000] NUMA: Initialized distance table, cnt=8
[    0.000000] NUMA: Node 4 [0,80000000) + [100000000,280000000) -> [0,280000000)
```

We are going to assume that `hwloc` has been installed already and is located at `/home/pasl/Installs/hwloc-install/`. To configure PASL to use `hwloc`, add the following lines to the `settings.sh` file in the `/home/pasl/pasl/minicourse` folder.

Example 7.3 Configuration to use `hwloc`

```
USE_HWLOC=1
HWLOC_PATH=/home/pasl/Installs/hwloc-install/
```

7.3 Starting with installed binaries

At this point, you have either installed all the necessary software to work with PASL or these are installed for you. In either case, make sure that your `PATH` variable makes the software visible. For setting up your `PATH` variable on andrew.cmu domain, see below.

7.3.1 Specific set up for the andrew.cmu domain

We have installed much of the needed software on andrew.cmu.edu. So you need to go through a relatively minimal set up.

First set up your `PATH` variable to refer to the right directories. Using cshell

```
setenv PATH  /opt/rh/devtoolset-3/root/usr/bin:/usr/lib64/qt-3.3/bin:/usr/lib64/ccache:/usr ↩
    /local/bin:/bin:/usr/bin:./
```

The part added to the default PATH on andrew is

```
/opt/rh/devtoolset-3/root/usr/bin
```

It is important that this is at the beginning of the `PATH` variable. To make interaction easier, we also added the relative path `./` to the `PATH` variable.

7.3.2 Fetch the benchmarking tools (pbench)

We are going to use two command-line tools to help us to run experiments and to analyze the data. These tools are part of a library that we developed, which is named pbench. The pbench sources are available via github.

```
$ cd /home/pasl
$ git clone https://github.com/deepsea-inria/pbench.git
```

The tarball of the sources can be downloaded from the github page.

7.3.3 Build the tools

The following command builds the tools, namely `prun` and `pplot`. The former handles the collection of data and the latter the human-readable output (e.g., plots, tables, etc.).

```
$ make -C /home/pasl/pbench/
```

Make sure that the build succeeded by checking the pbench directory for the files `prun` and `pplot`. If these files do not appear, then the build failed.

7.3.4 Create aliases

We recommend creating the following aliases.

```
$ alias prun '/home/pasl/pbench/prun'
$ alias pplot '/home/pasl/pbench/pplot'
```

It will be convenient for you to make these aliases persistent, so that next time you log in, the aliases will be set. Add the commands above to your shell configuration file.

7.3.5 Visualizer Tool

When we are tuning our parallel algorithms, it can be helpful to visualize their processor utilization over time, just in case there are patterns that help to assign blame to certain regions of code. Later, we are going to use the utilization visualizer that comes packaged along with PASL. To build the tool, run the following make command.

```
$ make -C /home/pasl/pasl/tools/pview pview
```

Let us create an alias for the tool.

```
$ alias pview '/home/pasl/pasl/tools/pview/pview'
```

We recommend that you make this alias persistent by putting it into your shell configuration file (as you did above for the pbench tools).

7.4 Using the Makefile

PASL comes equipped with a `Makefile` that can generate several different kinds of executables. These different kinds of executables and how they can be generated is described below for a benchmark program `pgm`.

- ***baseline***: build the baseline with command `make pgm.baseline`
- ***elision***: build the sequential elision with command `make pgm.elision`
- ***optimized***: build the optimized binary with command `make pgm.opt`
- ***log***: build the log binary with command `make pgm.log`
- ***debug***: build the debug binary with the command `make pgm.dbg`

To speed up the build process, add to the `make` command the option `-j` (e.g., `make -j pgm.opt`). This option enables `make` to parallelize the build process. Note that, if the build fails, the error messages that are printed to the terminal may be somewhat garbled. As such, it is better to use `-j` only if after the debugging process is complete.

7.5 Task 1: Run the baseline Fibonacci

We are going to start our experimentation with three different instances of the same program, namely `bench`. This program serves as a "driver" for the benchmarks that we have implemented. These implementations are good parallel codes that we expect to deliver good performance. We first build the baseline version.

```
$ cd /home/pasl/pasl/minicourse
$ make bench.baseline
```

Warning
The command-line examples that we show here assume that you have `.` in your `$PATH`. If not, you may need to prefix command-line calls to binaries with `./` (e.g., `./bench.baseline`).

The file extension `.baseline` means that every benchmark in the binary uses the sequential-baseline version of the specified algorithm.

We can now run the baseline for one of our benchmarks, say Fibonacci by using the `-bench` argument to specify the benchmark and the `-n` argument to specify the input value for the Fibonacci function.

```
$ bench.baseline -bench fib -n 39
```

On our machine, the output of this run is the following.

```
exectime 0.556
utilization 1.0000
result 63245986
```

The three lines above provide useful information about the run.

- The `exectime` indicates the wall-clock time in seconds that is taken by the benchmark. In general, this time measures only the time taken by the benchmark under consideration. It does not include the time taken to generate the input data, for example.
- The `utilization` relates to the utilization of the processors available to the program. In the present case, for a single-processor run, the utilization is by definition 100%. We will return to this measure soon.
- The `result` field reports a value computed by the benchmark. In this case, the value is the 39^{th} Fibonacci number.

7.6 Task 2: Run the sequential elision of Fibonacci

The `.elision` extension means that parallel algorithms (not sequential baseline algorithms) are compiled. However, all instances of `fork2()` are erased as described in an earlier chapter.

```
$ make bench.elision
$ bench.elision -bench fib -n 39
```

The run time of the sequential elision in this case is similar to the run time of the sequential baseline because the two are similar codes. However, for most other algorithms, the baseline will typically be at least a little faster.

```
exectime 0.553
utilization 1.0000
result 63245986
```

7.7 Task 3: Run parallel Fibonacci

The `.opt` extension means that the program is compiled with full support for parallel execution. Unless specified otherwise, however, the parallel binary uses just one processor.

```
$ make bench.opt
$ bench.opt -bench fib -n 39
```

The output of this program is similar to the output of the previous two programs.

```
exectime 0.553
utilization 1.0000
result 63245986
```

Because our machine has 40 processors, we can run the same application using all available processors. Before running this command, please adjust the `-proc` option to match the number of cores that your machine has. Note that you can use any number of cores up to the number you have available. You can use `nproc` or `lscpu` to determine the number of cores your machine has.

```
$ bench.opt -bench fib -n 39 -proc 40
```

We see from the output of the 40-processor run that our program ran faster than the sequential runs. Moreover, the `utilization` field tells us that approximately 86% of the total time spent by the 40 processors was spent performing useful work, not idling.

```
exectime 0.019
utilization 0.8659
result 63245986
```

Warning
PASL allows the user to select the number of processors by the `-proc` key. The maximum value for this key is the number of processors that are available on the machine. PASL raises an error if the programmer asks for more processors than are available.

7.8 Measuring performance with "speedup"

We may ask at this point: What is the improvement that we just observed from the parallel run of our program? One common way to answer this question is to measure the "speedup".

Definition: P-processor speedup

The speedup on P processors is the ratio T_B/T_P, where the term T_B represents the run time of the sequential baseline program and the term T_P the time measured for the P-processor run.

The importance of selecting a good baseline

Note that speedup is defined with respect to a baseline program. How exactly should this baseline program be chosen? One option is to take the sequential elision as a baseline. The speedup curve with such a baseline can be helpful in determining the scalability of a parallel algorithm but it can also be misleading, especially if speedups are taken as a indicator of good performance, which they are not because they are only relative to a specific baseline. For speedups to be a valid indication of good performance, they must be calculated against an optimized implementation of the best serial algorithm (for the same problem.)

The speedup at a given number of processors is a good starting point on the way to evaluating the scalability of the implementation of a parallel algorithm. The next step typically involves considering speedups taken from varying numbers of processors available to the program. The data collected from such a speedup experiment yields a ***speedup curve***, which is a curve that plots the trend of the speedup as the number of processors increases. The shape of the speedup curve provides valuable clues for performance and possibly for tuning: a flattening curve suggests lack of parallelism; a curve that arcs up and then downward suggests that processors may be wasting time by accessing a shared resource in an inefficient manner (e.g., false sharing); a speedup curve with a constant slope indicates at least some scaling.

Example 7.4 Speedup for our run of Fibonacci on 40 processors

The speedup T_B/T_{40} equals $0.556/0.019 = 29.26$x. Although not linear (i.e., 40x), this speedup is decent considering factors such as: the capabilities of our machine; the overheads relating to parallelism; and the small size of the problem compared to the computing power that our machine offers.

7.8.1 Generate a speedup plot

Let us see what a speedup curve can tell us about our parallel Fibonacci program. We need to first get some data. The following command performs a sequence of runs of the Fibonacci program for varying numbers of processors. You can now run the command yourself.

```
$ prun speedup -baseline "bench.baseline" -parallel "bench.opt -proc 1,10,20,30,40" -bench ↩
    fib -n 39
```

Here is another example on a 24-core machine.

```
$ prun speedup -baseline "bench.baseline" -parallel "bench.opt -proc 1,4,8,16,24" -bench ↩
    fib -n 39
```

Run the following command to generate the speedup plot.

```
$ pplot speedup
```

If successful, the command generates a file named `plots.pdf`. The output should look something like the plot in speedup plot below.

```
Starting to generate 1 charts.
Produced file plots.pdf.
```

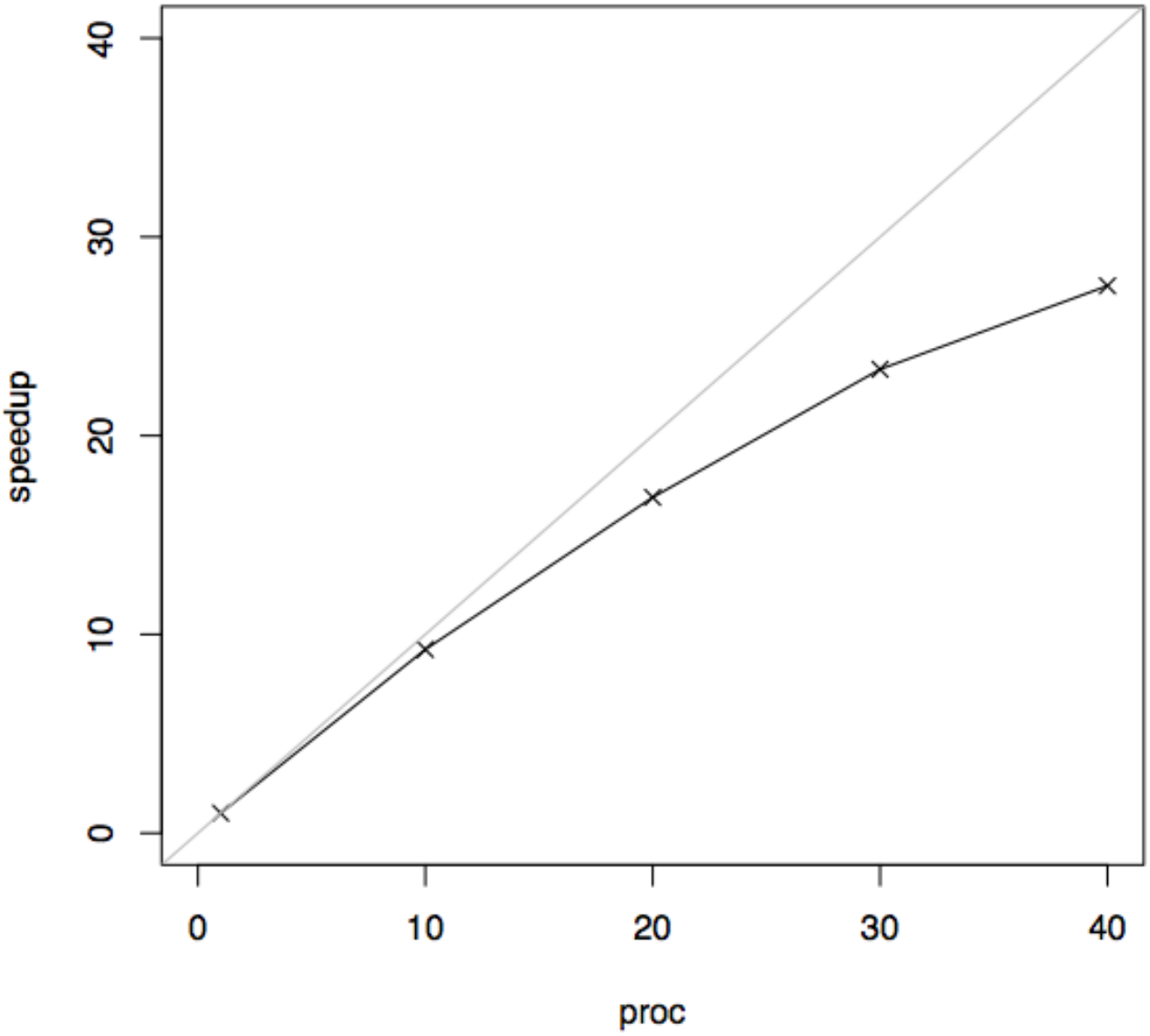


Figure 6: Speedup curve for the computation of the 39th Fibonacci number.

The plot shows that our Fibonacci application scales well, up to about twenty processors. As expected, at twenty processors, the curve dips downward somewhat. We know that the problem size is the primary factor leading to this dip. How much does the problem size matter? The speedup plot in the Figure below shows clearly the trend. As our problem size grows, so does the speedup improve, until at the calculation of the 45^{th} Fibonacci number, the speedup curve is close to being linear.

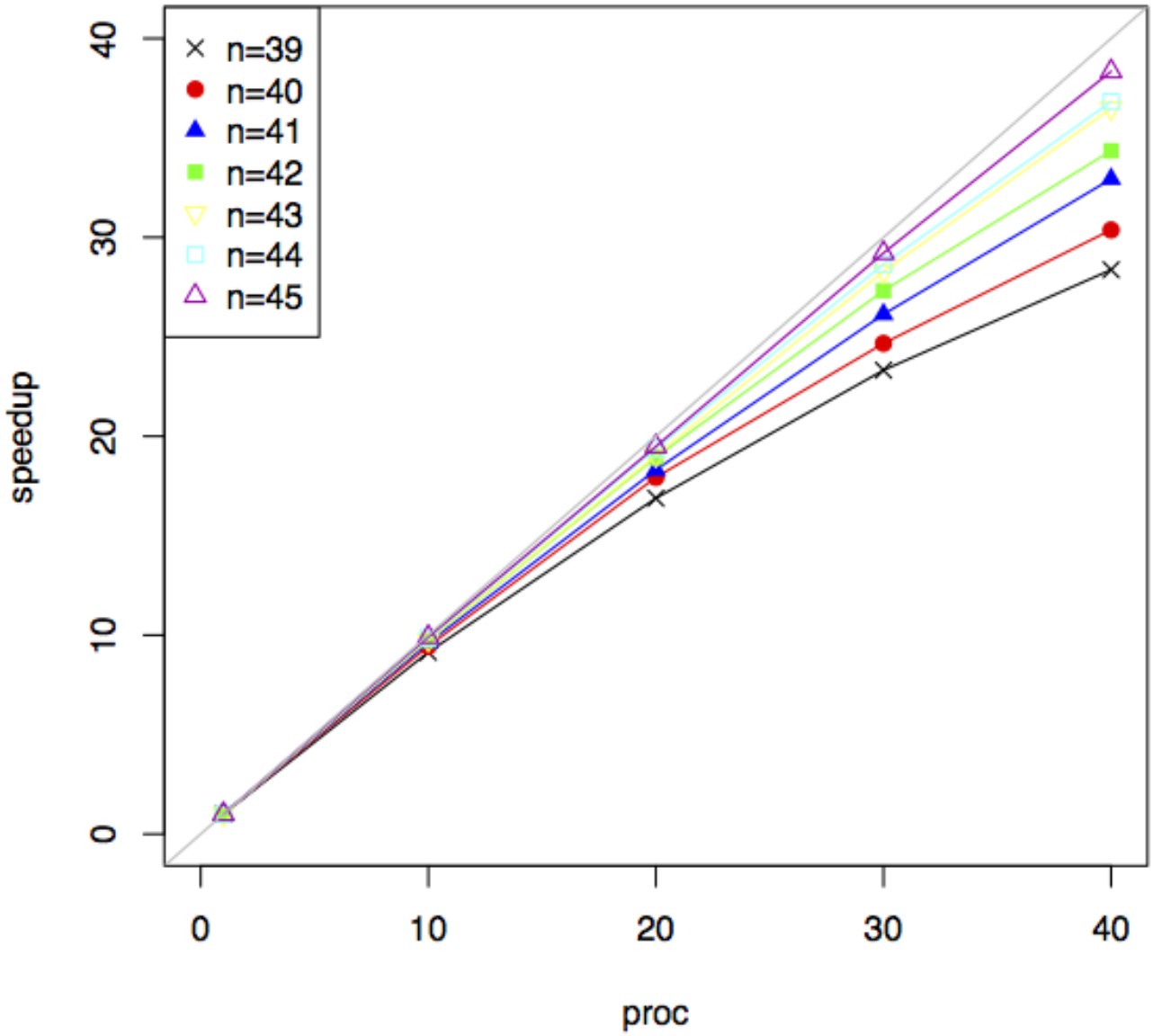


Figure 7: Speedup plot showing speedup curves at different problem sizes.

Note

The `prun` and `pplot` tools have many more features than those demonstrated here. For details, see the documentation provided with the tools in the file named `README.md`.

Noise in experiments

The run time that a given parallel program takes to solve the same problem can vary noticeably because of certain effects that are not under our control, such as OS scheduling, cache effects, paging, etc. We can consider such noise in our experiments random noise. Noise can be a problem for us because noise can lead us to make incorrect conclusions when, say, comparing the performance of two algorithms that perform roughly the same. To deal with randomness, we can perform multiple runs for each data point that we want to measure and consider the mean over these runs. The `prun` tool enables taking multiple runs via the `-runs` argument. Moreover, the `pplot` tool by default shows mean values for any given set of runs and optionally shows error bars. The documentation for these tools gives more detail on how to use the statistics-related features.

7.8.2 Superlinear speedup

Suppose that, on our 40-processor machine, the speedup that we observe is larger than 40x. It might sound improbable or even impossible. But it can happen. Ordinary circumstances should preclude such a **superlinear speedup**, because, after all, we have only forty processors helping to speed up the computation. Superlinear speedups often indicate that the sequential baseline program is suboptimal. This situation is easy to check: just compare its run time with that of the sequential elision. If the sequential elision is faster, then the baseline is suboptimal. Other factors can cause superlinear speedup: sometimes parallel programs running on multiple processors with private caches benefit from the larger cache capacity. These issues are, however, outside the scope of this course. As a rule of thumb, superlinear speedups should be regarded with suspicion and the cause should be investigated.

7.9 Visualize processor utilization

The 29x speedup that we just calculated for our Fibonacci benchmark was a little dissapointing, and the 86% processor utilization of the run left 14% utilization for improvement. We should be suspicious that, although seemingly large, the problem size that we chose, that is, $n = 39$, was probably a little too small to yield enough work to keep all the processors well fed. To put this hunch to the test, let us examine the utilization of the processors in our system. We need to first build a binary that collects and outputs logging data.

```
$ make bench.log
```

We run the program with the new binary in the same fashion as before.

```
$ bench.log -bench fib -proc 40 -n 39
```

The output looks something like the following.

```
exectime 0.019
launch_duration 0.019
utilization     0.8639
thread_send     205
thread_exec     4258
thread_alloc    2838
utilization 0.8639
result 63245986
```

We need to explain what the new fields mean.

- The `thread_send` field tells us that 233 threads were exchaged between processors for the purpose of load balancing;
- the `thread_exec` field that 5179 threads were executed by the scheduler;
- the `thread_alloc` field that 3452 threads were freshly allocated.

Each of these fields can be useful for tracking down inefficiencies. The number of freshly allocated threads can be a strong indicator because in C++ thread allocation costs can sometimes add up to a significant cost. In the present case, however, none of the new values shown above are highly suspicious, considering that there are all at most in the thousands.

Since we have not yet found the problem, let us look at the visualization of the processor utilization using our `pview` tool. To get the necessary logging data, we need to run our program again, this time passing the argument `--pview`.

```
$ bench.log -bench fib -n 39 -proc 40 --pview
```

When the run completes, a binary log file named `LOG_BIN` should be generated in the current directory. Every time we run with `--pview` this binary file is overwritten. To see the visualization of the log data, we call the visualizer tool from the same directory.

```
$ pview
```

The output we see on our 40-processor machine is shown in the Figure below. The window shows one bar per processor. Time goes from left to right. Idle time is represented by red and time spent busy with work by grey. You can zoom in any part of the plot by clicking on the region with the mouse. To reset to the original plot, press the space bar. From the visualization, we can see that most of the time, particularly in the middle, all of the processors keep busy. However, there is a lot of idle time in the beginning and end of the run. This pattern suggests that there just is not enough parallelism in the early and late stages of our Fibonacci computation.

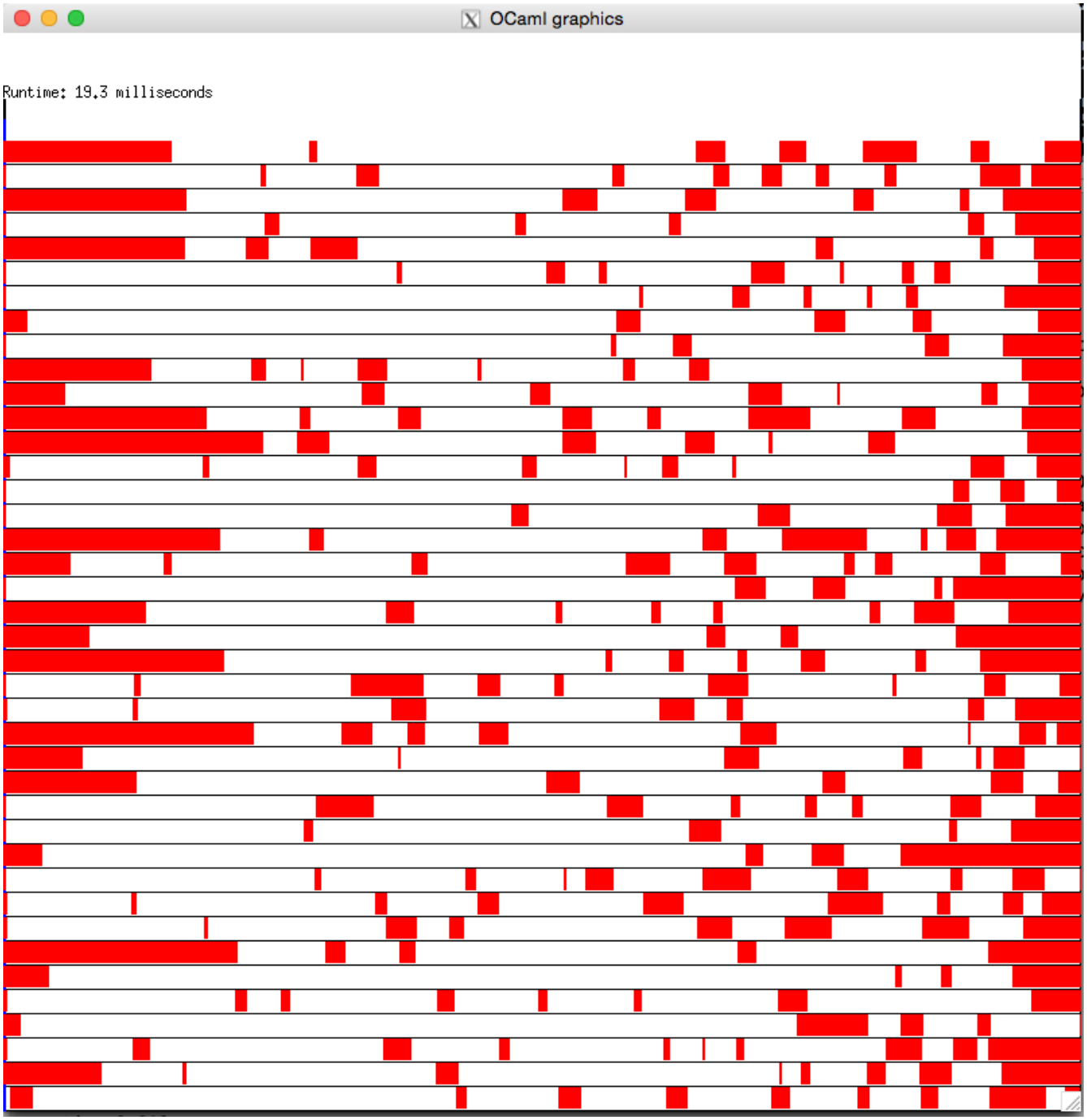


Figure 8: Utilization plot for computation of 39th Fibonacci number.

7.10 Strong versus weak scaling

We are pretty sure that or Fibonacci program is not scaling as well is it could. But poor scaling on one particular input for n does not necessarily mean there is a problem with the scalability our parallel Fibonacci program in general. What is important is to know more precisely what it is that we want our Fibonacci program to achieve. To this end, let us consider a distinction that is important in high-performance computing: the distinction between strong and weak scaling. So far, we have been studying the strong-scaling profile of the computation of the 39^{th} Fibonacci number. In general, strong scaling concerns how the run time varies with the number of processors for a fixed problem size. Sometimes strong scaling is either too ambitious, owing to hardware limitations, or not necessary, because the programmer is happy to live with a looser notion of scaling, namely weak scaling. In weak scaling, the programmer considers a fixed-size problem per processor. We are going to consider something similar to weak scaling. In the Figure below, we have a plot showing how processor utilization varies with the input size. The situation dramatically improves from 12% idle time for the 39^{th} Fibonacci number down to 5% idle time for the 41^{st} and finally to 1% for the 45^{th}. At just 1% idle time, the utilization is excellent.

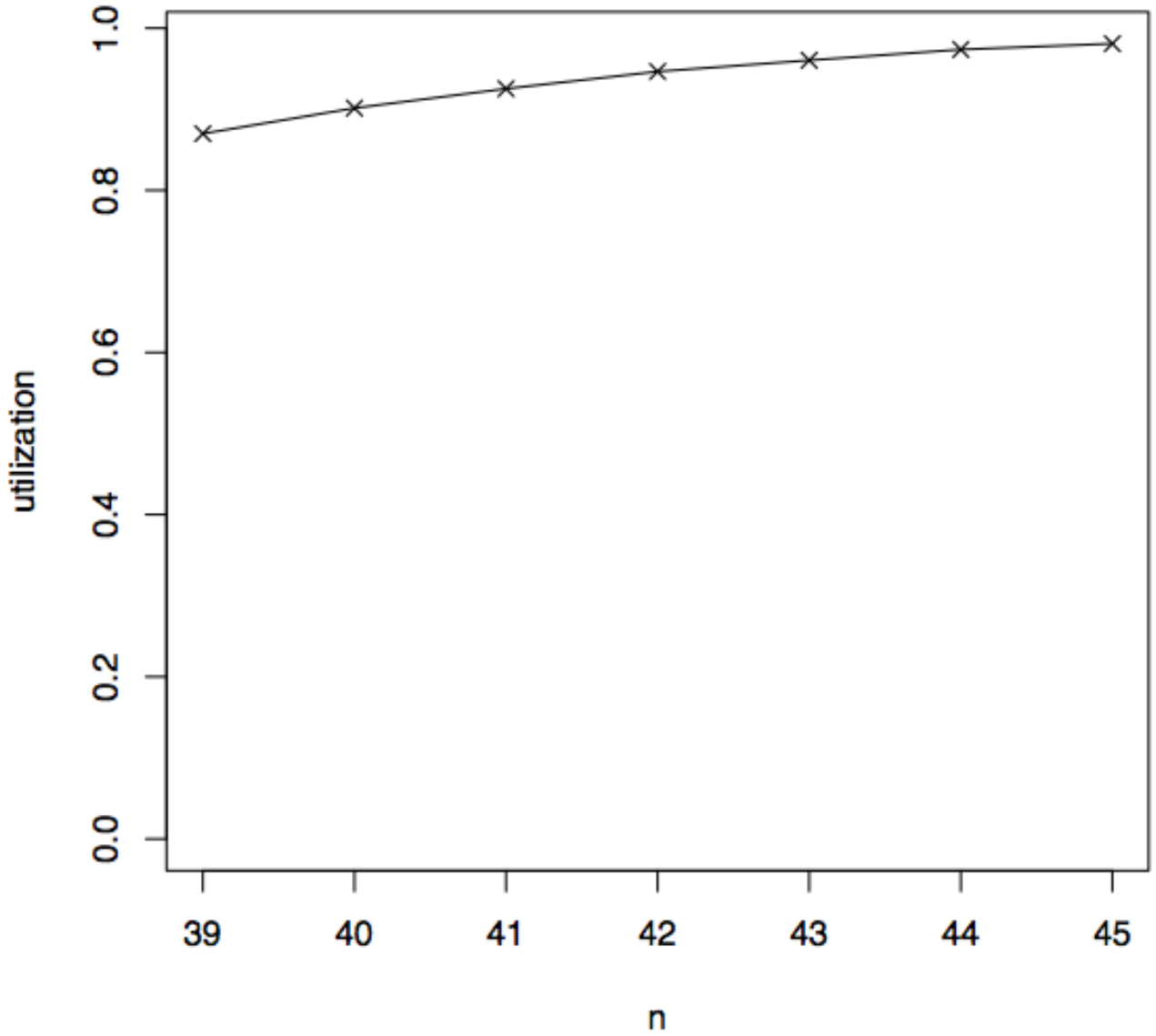


Figure 9: How processor utilization of Fibonacci computation varies with input size.

The scenario that we just observed is typical of multicore systems. For computations that perform relatively little work, such

as the computation of the 39^{th} Fibonacci number, properties that are specific to the hardware, OS, and PASL load-balancing algorithm can noticeably limit processor utilization. For computations that perform lots of highly parallel work, such limitations are barely noticeable, because processors spend most of their time performing useful work. Let us return to the largest Fibonacci instance that we considered, namely the computation of the 45^{th} Fibonacci number, and consider its utilization plot.

```
$ bench.log -bench fib -n 45 -proc 40 --pview
$ pview
```

The utilization plot is shown in the Figure below. Compared the to utilization plot we saw in the Figure above for n=39, the red regions are much less prominent overall and the idle regions at the beginning and end are barely noticeable.

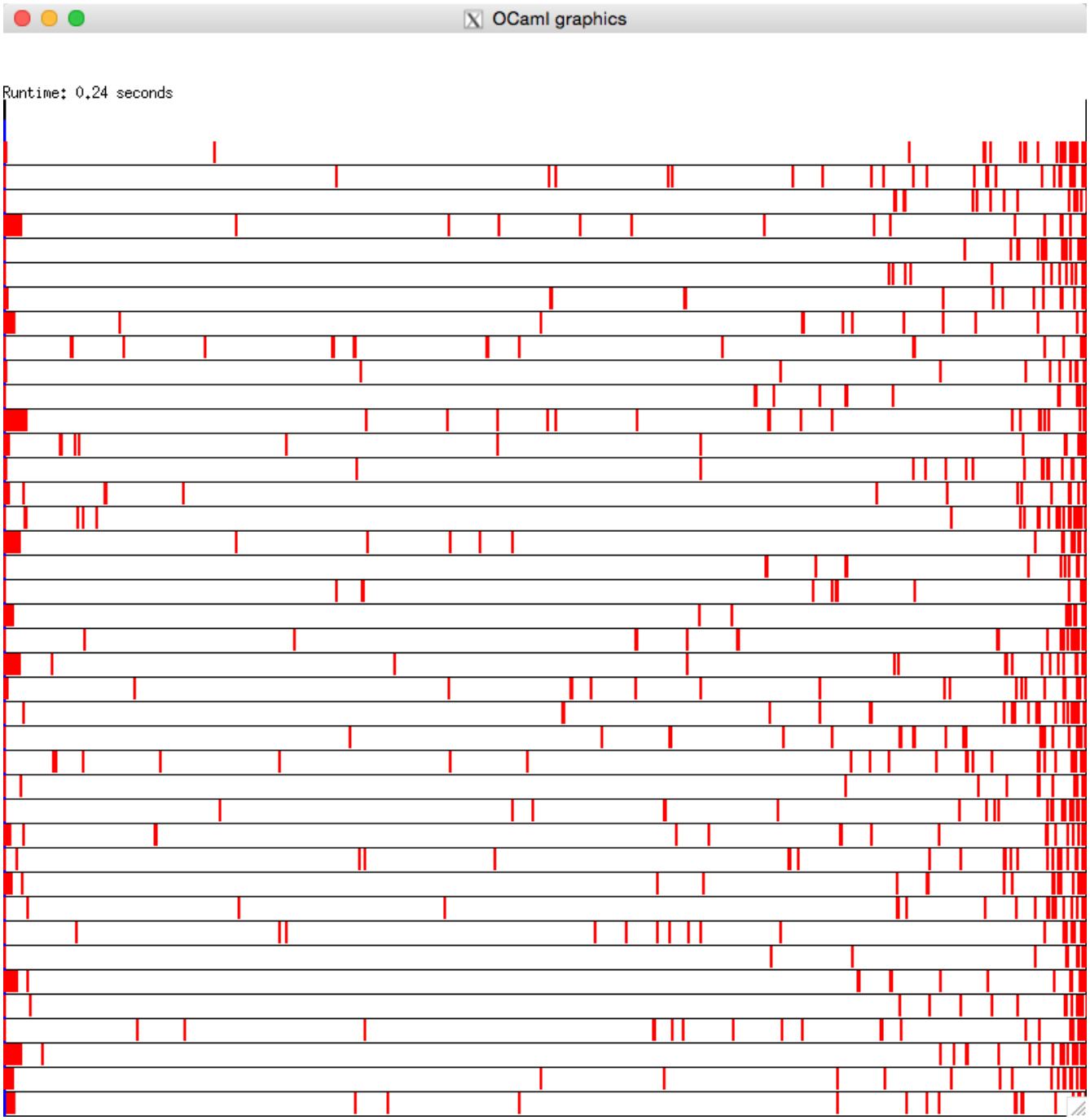


Figure 10: Utilization plot for computation of 45th Fibonacci number.

7.11 Chapter Summary

We have seen in this lab how to build, run, and evaluate our parallel programs. Concepts that we have seen, such as speedup curves, are going to be useful for evaluating the scalability of our future solutions. Strong scaling is the gold standard for a parallel implementation. But as we have seen, weak scaling is a more realistic target in most cases.

8 Chapter: Work efficiency

In many cases, a parallel algorithm which solves a given problem performs more work than the fastest sequential algorithm that solves the same problem. This extra work deserves careful consideration for several reasons. First, since it performs additional work with respect to the serial algorithm, a parallel algorithm will generally require more resources such as time and energy. By using more processors, it may be possible to reduce the time penalty, but only by using more hardware resources.

For example, if an algorithm performs $O(\log n)$-factor more work than the serial algorithm, then, assuming that the constant factor hidden by the asymptotic notation is 1, when $n = 2^{20}$, it will perform 20-times more actual work, consuming 20 times more energy consumption than the serial algorithm. Assuming perfect scaling, we can reduce the time penalty by using more processors. For example, with 40 processors, the algorithm may require half the time of the serial algorithm.

Sometimes, a parallel algorithm has the same asymptotic complexity of the best serial algorithm for the problem but it has larger constant factors. This is generally true because scheduling friction, especially the cost of creating threads, can be significant. In addition to friction, parallel algorithms can incur more communication overhead than serial algorithms because data and processors may be placed far away in hardware. For example, it is not unusual for a parallel algorithm to incur a $10-100\times$ overhead over a similar serial algorithm because of scheduling friction and communication.

These considerations motivate considering "work efficiency" of parallel algorithm. Work efficiency is a measure of the extra work performed by the parallel algorithm with respect to the serial algorithm. We define two types of work efficiency: ***asymptotic work efficiency*** and ***observed work efficiency***. The former relates to the asymptotic performance of a parallel algorithm relative to the fastest sequential algorithm. The latter relates to running time of a parallel algorithm relative to that of the fastest sequential algorithm.

Definition: asymptotic work efficiency

An algorithm is ***asymptotically work efficient*** if the work of the algorithm is the same as the work of the best known serial algorithm.

Example 8.1 Asymptotic work efficiency

- A parallel algorithm that comparison-sorts n keys in span $O(\log^3 n)$ and work $O(n \log n)$ is asymptotically work efficient because the work cost is as fast as the best known sequential comparison-based sorting algorithm. However, a parallel sorting algorithm that takes $O(n \log^2 n)$ work is not asymptotically work efficient.
- The parallel array increment algorithm that we consider in an earlier Chapter is asymptotically work efficient, because it performs linear work, which is optimal (any sequential algorithm must perform at least linear work).

To assess the practical effectiveness of a parallel algorithm, we define observed work efficiency, parameterized a value r.

Definition: observed work efficiency

A parallel algorithm that runs in time T_1 on a single processor has **observed work efficient factor of** r if $r = \frac{T_1}{T_{seq}}$, where T_{seq} is the time taken by the fastest known sequential algorithm.

Example 8.2 Observed work efficiency

- A parallel algorithm that runs $10\times$ slower on a single processor that the fastest sequential algorithm has an observed work efficiency factor of 10. We consider such algorithms unacceptable, as they are too slow and wasteful.
- A parallel algorithm that runs $1.1\times - 1.2\times$. slower on a single processor that the fastest sequential algorithm has observed work efficiency factor of 1.2. We consider such algorithms to be acceptable.

Example 8.3 Observed work efficiency of parallel increment

To obtain this measure, we first run the baseline version of our parallel-increment algorithm.

```
$ bench.baseline -bench map_incr -n 100000000
exectime 0.884
utilization 1.0000
result 2
```

We then run the parallel algorithm, which is the same exact code as `map_incr_rec`. We build this code by using the special `optfp` "force parallel" file extension. This special file extension forces parallelism to be exposed all the way down to the base cases. Later, we will see how to use this special binary mode for other purposes.

```
$ make bench.optfp
$ bench.optfp -bench map_incr -n 100000000
exectime 45.967
utilization 1.0000
result 2
```

Our algorithm has an observed work efficiency factor of $60\times$. Such poor observed work efficiency suggests that the parallel algorithm would require more than an order of magnitude more energy and that it would not run faster than the serial algorithm even when using less than 60 processors.

In practice, observed work efficiency is a major concern. First, the whole effort of parallel computing is wasted if parallel algorithms consistently require more work than the best sequential algorithms. In other words, in parallel computing, both asymptotic complexity and constant factors matter.

Based on these discussions, we define a ***good parallel algorithm*** as follows.

Definition: good parallel algorithm

We say that a parallel algorithm is ***good*** if it has the following three characteristics:

1. it is asymptotically work efficient;
2. it is observably work efficient;
3. it has low span.

8.1 Improving work efficiency with granularity control

It is common for a parallel algorithm to be asymptotically and/or observably work inefficient but it is often possible to improve work efficiency by observing that work efficiency increases with parallelism and can thus be controlled by limiting it.

For example, a parallel algorithm that performs linear work and has logarithmic span leads to average parallelism in the orders of thousands with the small input size of one million. For such a small problem size, we usually would not need to employ thousands of processors. It would be sufficient to limit the parallelism so as to feed tens of processors and as a result reduce impact of excess parallelism on work efficiency.

In many parallel algorithms such as the algorithms based on divide-and-conquer, there is a simple way to achieve this goal: switch from parallel to sequential algorithm when the problem size falls below a certain threshold. This technique is sometimes called ***coarsening*** or ***granularity control***.

But which code should we switch to: one idea is to simply switch to the sequential elision, which we always have available in PASL. If, however, the parallel algorithm is asymptotically work inefficient, this would be ineffective. In such cases, we can specify a separate sequential algorithm for small instances.

Optimizing the practical efficiency of a parallel algorithm by controlling its parallelism is sometimes called ***optimization***, sometimes it is called ***performance engineering***, and sometimes ***performance tuning*** or simply ***tuning***. In the rest of this document, we use the term "tuning."

Example 8.4 Tuning the parallel array-increment function

We can apply coarsening to `map_inc_rec` code by switching to the sequential algorithm when the input falls below an established threshold.

```
long threshold = Some Number;

void map_incr_rec(const long* source, long* dest, long lo, long hi) {
  long n = hi - lo;
  if (n <= threshold) {
    for (long i = lo; i < hi; i++)
      dest[i] = source[i] + 1;
  } else {
    long mid = (lo + hi) / 2;
    fork2([&] {
      map_incr_rec(source, dest, lo, mid);
    }, [&] {
      map_incr_rec(source, dest, mid, hi);
    });
  }
}
```

Note

Even in sequential algorithms, it is not uncommon to revert to a different algorithm for small instances of the problem. For example, it is well known that insertion sort is faster than other sorting algorithms for very small inputs containing 30 keys or less. Many optimize sorting algorithm therefore revert to insertion sort when the input size falls within that range.

Example 8.5 Observed work efficiency of tuned array increment

As can be seen below, after some tuning, `map_incr_rec` program becomes highly work efficient. In fact, there is barely a difference between the serial and the parallel runs. The tuning is actually done automatically here by using an automatic-granularity-control technique described in the section.

```
$ bench.baseline -bench map_incr -n 100000000
exectime 0.884
utilization 1.0000
result 2
$ bench.opt -bench map_incr -n 100000000
exectime 0.895
utilization 1.0000
result 2
```

In this case, we have $r = \frac{T_1}{T_{seq}} = \frac{0.895}{0.884} = 1.012$ observed work efficiency. Our parallel program on a single processor is one percent slower than the sequential baseline. Such work efficiency is excellent.

8.2 Determining the threshold

The basic idea behind coarsening or granularity control is to revert to a fast serial algorithm when the input size falls below a certain threshold. To determine the optimal threshold, we can simply perform a search by running the code with different threshold settings.

While this approach can help find the right threshold on the particular machine that we performed the search, there is no guarantee that the same threshold would work on another machine. In fact, there are examples in the literature that show that such optimizations are not ***portable***, i.e., a piece of code optimized for a particular architecture may behave poorly on another.

In the general case, determining the right threshold is even more difficult. To see the difficulty consider a generic (polymorphic), higher-order function such as `map` that takes a sequence and a function and applies the function to the sequence. The problem is that the threshold depends both on the type of the elements of the sequence and the function to be mapped over the sequence. For example, if each element itself is a sequence (the sequence is nested), the threshold can be relatively small. If, however, the elements are integers, then the threshold will likely need to be relatively large. This makes it difficult to determine the threshold because it depends on arguments that are unknown at compile time. Essentially the same argument applies to the function being mapped over the sequence: if the function is expensive, then the threshold can be relatively small, but otherwise it will need to be relatively large.

As we describe in this chapter, it is sometimes possible to determine the threshold completely automatically.

9 Chapter: Automatic granularity control

There has been significant research into determining the right threshold for a particular algorithm. This problem, known as the ***granularity-control problem***, turns out to be a rather difficult one, especially if we wish to find a technique that can ensure close-to-optimal performance across different architectures. In this section, we present a technique for automatically controlling granularity by using asymptotic cost functions.

9.1 Complexity functions

Our automatic granularity-control technique requires assistance from the application programmer: for each parallel region of the program, the programmer must annotate the region with a **complexity function**, which is simply a C++ function that returns the asymptotic work cost associated with running the associated region of code.

Example 9.1 Complexity function of `map_incr_rec`

An application of our `map_incr_rec` function on a given range $[lo, hi)$ of an array has work cost $cn = c(hi - lo)$ for some constant c. As such, the following lambda expression is one valid complexity function for our parallel `map_incr_rec` program.

```
auto map_incr_rec_complexity_fct = [&] (long lo, long hi) {
  return hi - lo;
};
```

In general, the value returned by the complexity function need only be precise with respect to the asymptotic complexity class of the associated computation. The following lambda expression is another valid complexity function for function `map_incr_rec`. The complexity function above is preferable, however, because it is simpler.

```
const long k = 123;
auto map_incr_rec_complexity_fct2 = [&] (long lo, long hi) {
  return k * (hi - lo);
};
```

More generally, suppose that we know that a given algorithm has work cost of $W = n + \log n$. Although it would be fine to assign to this algorithm exactly W, we could just as well assign to the algorithm the cost n, because the second term is dominated by the first. In other words, when expressing work costs, we only need to be precise up to the asymptotic complexity class of the work.

9.2 Controlled statements

In PASL, a ***controlled statement***, or `cstmt`, is an annotation in the program text that activates automatic granularity control for a specified region of code. In particular, a controlled statement behaves as a C++ statement that has the special ability to choose on the fly whether or not the computation rooted at the body of the statement spawns parallel threads. To support such automatic granularity control PASL uses a prediction algorithm to map the asymptotic work cost (as returned by the complexity function)

to actual processor cycles. When the predicted processor cycles of a particular instance of the controlled statement falls below a threshold (determined automatically for the specific machine), then that instance is sequentialized, by turning off the ability to spawn parallel threads for the execution of that instance. If the predicted processor cycle count is higher than the threshold, then the statement instance is executed in parallel.

In other words, the reader can think of a controlled statement as a statement that executes in parallel when the benefits of parallel execution far outweigh its cost and that executes sequentially in a way similar to the sequential elision of the body of the controlled statement would if the cost of parallelism exceeds its benefits. We note that while the sequential exection is similar to a sequential elision, it is not exactly the same, because every call to `fork2` must check whether it should create parallel threads or run sequentially. Thus the execution may differ from the sequential elision in terms of performance but not in terms of behavior or semantics.

Example 9.2 Array-increment function with automatic granularity control

The code below uses a controlled statement to automatically select, at run time, the threshold size for our parallel array-increment function.

```
controller_type map_incr_rec_contr("map_incr_rec");

void map_incr_rec(const long* source, long* dest, long lo, long hi) {
  long n = hi - lo;
  cstmt(map_incr_rec_contr, [&] { return n; }, [&] {
    if (n == 0) {
      // do nothing
    } else if (n == 1) {
      dest[lo] = source[lo] + 1;
    } else {
      long mid = (lo + hi) / 2;
      fork2([&] {
        map_incr_rec(source, dest, lo, mid);
      }, [&] {
        map_incr_rec(source, dest, mid, hi);
      });
    }
  });
}
```

The controlled statement takes three arguments, whose requirements are specified below, and returns nothing (i.e., `void`). The effectiveness of the granularity controller may be compromised if any of the requirements are not met.

- The first argument is a reference to the controller object. The controller object is used by the controlled statement to collect profiling data from the program as the program runs. Every controller object is initialized with a string label (in the code above `"map_incr_rec"`). The label must be unique to the particular controller. Moreover, the controller must be declared as a global variable.
- The second argument is the complexity function. The type of the return value should be `long`.
- The third argument is the body of the controlled statement. The return type of the controlled statement should be `void`.

When the controlled statement chooses sequential evaluation for its body the effect is similar to the effect where in the code above the input size falls below the threshold size: the body and the recursion tree rooted there is sequentialized. When the controlled statement chooses parallel evaluation, the calls to `fork2()` create parallel threads.

9.2.1 Granularity control with alternative sequential bodies

It is not unusual for a divide-and-conquer algorithm to switch to a different algorithm at the leaves of its recursion tree. For example, sorting algorithms, such as quicksort, may switch to insertion sort at small problem sizes. In the same way, it is not unusual for parallel algorithms to switch to different sequential algorithms for handling small problem sizes. Such switching can be beneficial especially when the parallel algorithm is not asymptotically work efficient.

To provide such algorithmic switching, PASL provides an alternative form of controlled statement that accepts a fourth argument: the ***alternative sequential body***. This alternative form of controlled statement behaves essentially the same way as the original described above, with the exception that when PASL run time decides to sequentialize a particular instance of the controlled statement, it falls through to the provided alternative sequential body instead of the "sequential elision."

Example 9.3 Array-increment function with automatic granularity control and sequential body

```
controller_type map_incr_rec_contr("map_incr_rec");

void map_incr_rec(const long* source, long* dest, long lo, long hi) {
  long n = hi - lo;
  cstmt(map_incr_rec_contr, [&] { return n; }, [&] {
    if (n == 0) {
      // do nothing
    } else if (n == 1) {
      dest[lo] = source[lo] + 1;
    } else {
      long mid = (lo + hi) / 2;
      fork2([&] {
        map_incr_rec(source, dest, lo, mid);
      }, [&] {
        map_incr_rec(source, dest, mid, hi);
      });
    }
  }, [&] {
    for (long i = lo; i < hi; i++)
      dest[i] = source[i] + 1;
  });
}
```

Even though the parallel and sequential array-increment algorithms are algorithmically identical, except for the calls to `fork2()`, there is still an advantage to using the alternative sequential body: the sequential code does not pay for the parallelism overheads due to `fork2()`. Even when eliding `fork2()`, the run-time-system has to perform a conditional branch to check whether or not the context of the `fork2()` call is parallel or sequential. Because the cost of these conditional branches adds up, the version with the sequential body is going to be more work efficient. Another reason for why a sequential body may be more efficient is that it can be written more simply, as for example using a for-loop instead of recursion, which will be faster in practice.

Recommended style for programming with controlled statements
In general, we recommend that the code of the parallel body be written so as to be completely self contained, at least in the sense that the parallel body code contains the logic that is necessary to handle recursion all the way down to the base cases. The code for `map_incr_rec` honors this style by the fact that the parallel body handles the cases where `n` is zero or one (base cases) or is greater than one (recursive case). Put differently, it should be the case that, if the parallelism-specific annotations (including the alternative sequential body) are erased, the resulting program is a correct program.
We recommend this style because such parallel codes can be debugged, verified, and tuned, in isolation, without relying on alternative sequential codes.

9.3 Controlled parallel-for loops

Let us add one more component to our granularity-control toolkit: the ***parallel-for*** from. By using this loop construct, we can avoid having to explicitly express recursion-trees over and over again. For example, the following function performs the same computation as the example function we defined in the first lecture. Only, this function is much more compact and readable. Moreover, this code takes advantage of our automatic granularity control, also by replacing the parallel-for with a serial-for.

```
loop_controller_type map_incr_contr("map_incr");

void map_incr(const long* source, long* dest, long n) {
```

```
  parallel_for(map_incr_contr, (long)0, n, [&] (long i) {
    dest[i] = source[i] + 1;
  });
}
```

Underneath, the parallel-for loop uses a divide-and-conquer routine whose structure is similar to the structure of the divide-and-conquer routine of our `map_incr_rec`. Because the parallel-for loop generates the log-height recursion tree, the `map_incr` routine just above has the same span as the `map_incr` routine that we defined earlier: $\log n$, where n is the size of the input array.

Notice that the code above specifies no complexity function. The reason is that this particular instance of the parallel-for loop implicitly defines a complexity function. The implicit complexity function reports a linear-time cost for any given range of the iteration space of the loop. In other words, the implicit complexity function assumes that per iteration the body of the loop performs a constant amount of work. Of course, this assumption does not hold in general. If we want to specify explicitly the complexity function, we can use the form shown in the example below. The complexity function is passed to the parallel-for loop as the fourth argument. The complexity function takes as argument the range `[lo, hi)`. In this case, the complexity is linear in the number of iterations. The function simply returns the number of iterations as the complexity.

```
loop_controller_type map_incr_contr("map_incr");

void map_incr(const long* source, long* dest, long n) {
  auto linear_complexity_fct = [] (long lo, long hi) {
    return hi-lo;
  };
  parallel_for(map_incr_contr, linear_complexity_fct, (long)0, n, [&] (long i) {
    dest[i] = source[i] + 1;
  });
}
```

The following code snippet shows a more interesting case for the complexity function. In this case, we are performing a multiplication of a dense matrix by a dense vector. The outer loop iterates over the rows of the matrix. The complexity function in this case gives to each of these row-wise iterations a cost in proportion to the number of scalars in each column.

```
loop_controller_type dmdvmult_contr("dmdvmult");

// mtx: nxn dense matrix, vec: length n dense vector
// dest: length n dense vector
void dmdvmult(double* mtx, double* vec, double* dest, long n) {
  auto compl_fct = [&] (long lo, long hi) {
    return (hi-lo)*n;
  };
  parallel_for(dmdvmult_contr, compl_fct, (long)0, n, [&] (long i) {
    ddotprod(mtx, v, dest, i);
  });
  return dest;
}
```

Example 9.4 Speedup for matrix multiply

Matrix multiplication has been widely used as an example for parallel computing since the early days of the field. There are good reasons for this. First, matrix multiplication is a key operation that can be used to solve many interesting problems. Second, it is an expansive computation that is nearly cubic in the size of the input---it can thus can become very expensive even with modest inputs.

Fortunately, matrix multiplication can be parallelized relatively easily as shown above. The figure below shows the speedup for a sample run of this code. Observe that the speedup is rather good, achieving nearly excellent utilization.

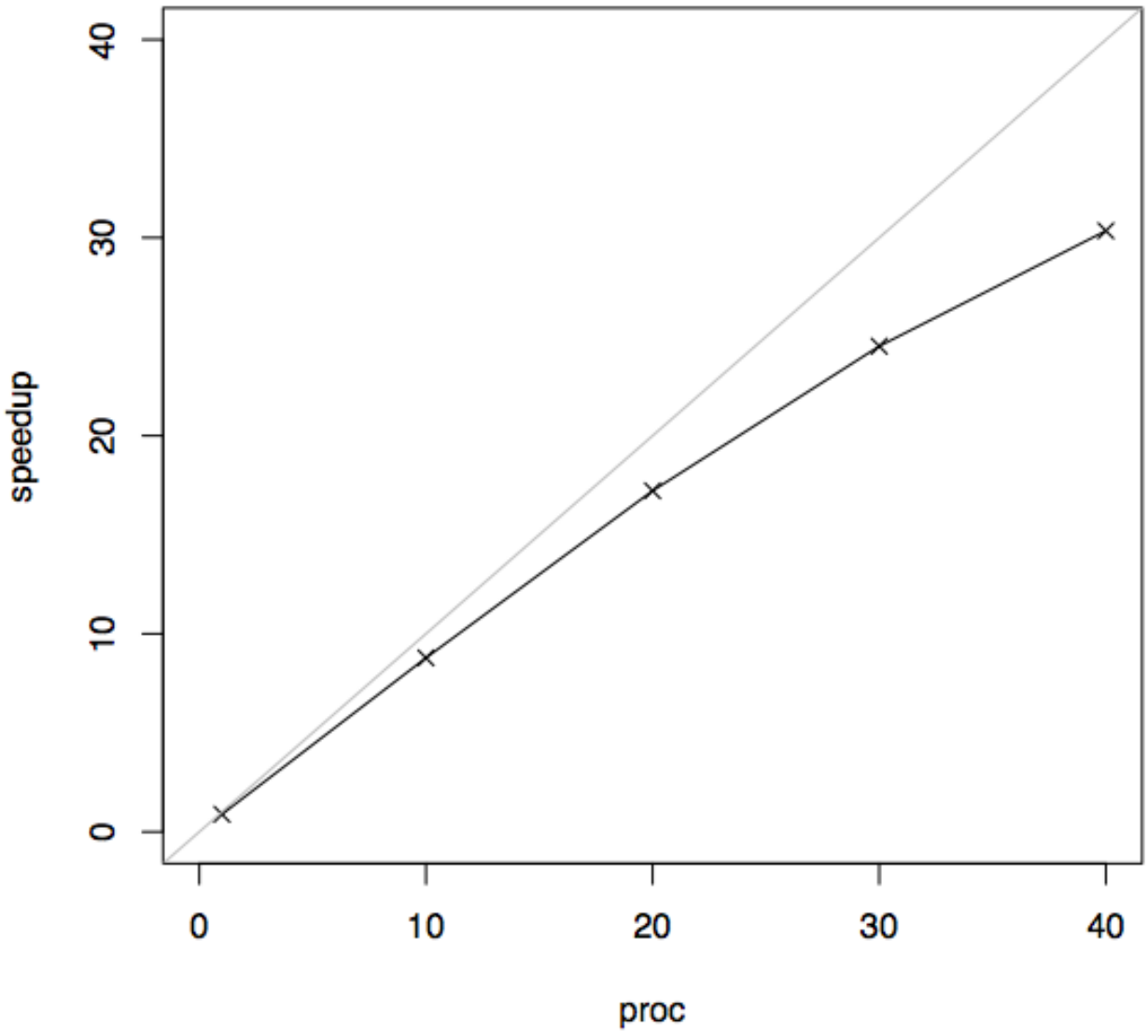


Figure 11: Speedup plot for matrix multiplication for 25000×25000 matrices.

While parallel matrix multiplication delivers excellent speedups, this is not common for many other algorithms on modern multicore machines where many computations can quickly become limited by the availability of bandwidth. Matrix multiplication does not suffer as much from the memory-bandwidth limitations because it performs significant work per memory operation: it touches $\Theta(n^2)$ memory cells, while performing nearly $\Theta(n^3)$ work.

10 Simple Parallel Arrays

Arrays are a fundamental data structure in sequential and parallel computing. When computing sequentially, arrays can sometimes be replaced by linked lists, especially because linked lists are more flexible. Unfortunately, linked lists are deadly for parallelism, because they require serial traversals to find elements; this makes arrays all the more important in parallel computing.

Unfortunately, it is difficult to find a good treatment of parallel arrays in C++: the various array implementations provided by C++ have been designed primarily for sequential computing. Each one has various pitfalls for parallel use.

Example 10.1 C++ arrays

By default, C++ arrays that are created by the `new[]` operator are initialized sequentially. Therefore, the work and span cost of the call `new[n]` is `n`. But we can initialize an array in logarithmic span in the number of items.

Example 10.2 STL vectors

The "vector" data structure that is provided by the Standard Template Library (STL) has similar issues. The STL vector implements a dynamically resizable array that provides push, pop, and indexing operations. The push and pop operations take amortized constant time and the indexing operation constant time. As with C++ arrays, initialization of vectors can require linear work and span. The STL vector also provides the method `resize(n)` which changes the size of the array to be `n`. The resize operation takes, in the worst case, linear work and span in proportion to the new size, `n`. In other words, the resize function uses a sequential algorithm to fill the cells in the vector. The `resize` operation is therefore not parallel for the same reason as for the default C++ arrays.

Such sequential computations that exist behind the wall of abstraction of a language or library can harm parallelism by introducing implicit sequential dependencies. Finding the source of such sequential bottlenecks can be time consuming, because they are hidden behind the abstraction boundary of the native array abstraction that is provided by the programming language.

We can avoid such pitfalls by carefully designing our own array data structure. Because array implementations are quite subtle, we consider our own implementation of parallel arrays, which makes explicit the cost of array operation, allowing us to control them quite carefully. Specifically, we carefully control initialization and disallow implicit copy operations on arrays, because copy operations can harm observable work efficiency (their asymptotic work cost is linear).

10.1 Interface and cost model

The key components of our array data structure, `sparray`, are shown by the code snippet below. An `sparray` can store 64-bit words only; in particular, they are monomorphic and fixed to values of type `long`. Generalizing `sparray` to store values of arbitrary types can be achieved by use of C++ templates. We stick to monomorphic arrays here to simplify the presentation.

```
using value_type = long;

class sparray {
public:

  // n: size to give to the array; by default 0
  sparray(unsigned long n = 0);

  // constructor from list
  sparray(std::initializer_list<value_type> xs);

  // indexing operator
  value_type& operator[](unsigned long i);

  // size of the array
  unsigned long size() const;

};
```

The class `sparray` provides two constructors. The first one takes in the size of the array (set to 0 by default) and allocates an unitialized array of the specified size (`nullptr` if size is 0). The second constructor takes in a list specified by curly braces and allocates an array with the same size. Since the argument to this constructor must be specified explicitly in the program, its size is constant by definition.

The cost model guaranteed by our implementation of parallel array is as follows:

- **Constructors/Allocation:** The work and span of simply allocating an array on the heap, without initialization, is constant. The second constructor performs initialization, based on constant-size lists, and thus also has constant work and span.

- **Array indexing:** Each array-indexing operation, that is the operation which accesses an individual cell, requires constant work and constant span.
- **Size operation:** The work and the span of accessing the size of the array is constant.
- **Destructors/Deallocation:** Not shown, the class includes a destructor that frees the array. Combined with the "move assignment operator" that C++ allows us to define, destructors can be used to deallocate arrays when they are out of scope. The destructor takes constant time because the contents of the array are just bits that do not need to be destructed individually.
- **Move assignment operator:** Not shown, the class includes a move-assignment operator that gets fired when an array is assigned to a variable. This operator moves the contents of the right-hand side of the assigned array into that of the left-hand side. This operation takes constant time.
- **Copy constructor:** The copy constructor of `sparray` is disabled. This prohibits copying an array unintentionally, for example, by passing the array by value to a function.

Note

The constructors of our array class do not perform initializations that involve non-constant work. If desired, the programmer can write an initializer that performs linear work and logarithmic span (if the values used for initialization have non-constant time cost, these bounds may need to be scaled accordingly).

Example 10.3 Simple use of arrays

This program below shows a basic use `sparray`'s. The first line allocates and initializes the contents of the array to be three numbers. The second uses the familiar indexing operator to access the item at the second position in the array. The third line extracts the size of the array. The fourth line assigns to the second cell the value 5. The fifth prints the contents of the cell.

```
sparray xs = { 1, 2, 3 };
std::cout << "xs[1] = " << xs[1] << std::endl;
std::cout << "xs.size() = " << xs.size() << std::endl;
xs[2] = 5;
std::cout << "xs[2] = " << xs[2] << std::endl;
```

Output:

```
xs[1] = 2
xs.size() = 3
xs[2] = 5
```

10.2 Allocation and deallocation

Arrays can be allocated by specifying the size of the array.

Example 10.4 Allocation and deallocation

```
sparray zero_length = sparray();
sparray another_zero_length = sparray(0);
sparray yet_another_zero_length;
sparray length_five = sparray(5);    // contents uninitialized
std::cout << "|zero_length| = " << zero_length.size() << std::endl;
std::cout << "|another_zero_length| = " << another_zero_length.size() << std::endl;
std::cout << "|yet_another_zero_length| = " << yet_another_zero_length.size() << std::endl;
std::cout << "|length_five| = " << length_five.size() << std::endl;
```

Output:

```
|zero_length| = 0
|another_zero_length| = 0
|yet_another_zero_length| = 0
|length_five| = 5
```

Just after creation, the array contents consist of uninitialized bits. We use this convention because the programmer needs flexibility to decide the parallelization strategy to initialize the contents. Internally, the `sparray` class consists of a size field and a pointer to the first item in the array. The contents of the array are heap allocated (automatically) by constructor of the `sparray` class. Deallocation occurs when the array's destructor is called. The destructor can be called by the programmer or by run-time system (of C++) if an object storing the array is destructed. Since C++ destructs (stack allocated) variables that go out of scope when a function returns, we can combine the stack discipline with heap-allocated arrays to manage the deallocation of arrays mostly automatically. We give several examples of this automatic deallocation scheme below.

Example 10.5 Automatic deallocation of arrays upon return

In the function below, the `sparray` object that is allocated on the frame of `foo` is deallocated just before `foo` returns, because the variable `xs` containing it goes out of scope.

```
void foo() {
  sparray xs = sparray(10);
  // array deallocated just before foo() returns
}
```

Example 10.6 Dangling pointers in arrays

Care must be taken when managing arrays, because nothing prevents the programmer from returning a dangling pointer.

```
value_type* foo() {
  sparray xs = sparray(10);
  ...
  // array deallocated just before foo() returns
  return &xs[0]
}

...

std::cout << "contents of deallocated memory: " << *foo() << std::endl;
```

Output:

```
contents of deallocated memory: .... (undefined)
```

It is safe to take a pointer to a cell in the array, when the array itself is still in scope. For example, in the code below, the contents of the array are used strictly when the array is in scope.

```
void foo() {
  sparray xs = sparray(10);
  xs[0] = 34;
  bar(&xs[0]);
  ...
  // array deallocated just before foo() returns
}

void bar(value_type* p) {
  std::cout << "xs[0] = " << *p << std::endl;
}
```

Output:

```
xs[0] = 34
```

We are going to see that we can rely on cleaner conventions for passing to functions references on arrays.

10.3 Passing to and returning from functions

If you are familiar with C++ container libraries, such as STL, this aspect of our array implementation, namely the calling conventions, may be unfamiliar: our arrays cannot be passed by value. We forbid passing by value because passing by value implies creating a fresh copy for each array being passed to or returned by a function. Of course, sometimes we really need to copy an array. In this case, we choose to copy the array explicitly, so that it is obvious where in our code we are paying a linear-time cost for copying out the contents. We will return to the issue of copying later.

Example 10.7 Incorrect use of copy constructor

What then happens if the program tries to pass an array to a function? The program will be rejected by the compiler. The code below does **not** compile, because we have disabled the copy constructor of our `sparray` class.

```
value_type foo(sparray xs) {
  return xs[0];
}

void bar() {
  sparray xs = { 1, 2 };
  foo(xs);
}
```

Example 10.8 Correctly passing an array by reference

The following code does compile, because in this case we pass the array `xs` to `foo` by reference.

```
value_type foo(const sparray& xs) {
  return xs[0];
}

void bar() {
  sparray xs = { 1, 2 };
  foo(xs);
}
```

Returning an array is straightforward: we take advantage of a feature of modern C++11 which automatically detects when it is safe to move a structure by a constant-time pointer swap. Code of the following form is perfectly legal, even though we disabled the copy constructor of `sparray`, because the compiler is able to transfer ownership of the array to the caller of the function. Moreover, the transfer is guaranteed to be constant work — not linear like a copy would take. The return is fast, because internally all that happens is that a couple words are being exchanged. Such "move on return" is achieved by the "move-assignment operator" of `sparray` class.

Example 10.9 Create and initialize an array (sequentially)

```
sparray fill_seq(long n, value_type x) {
  sparray tmp = sparray(n);
  for (long i = 0; i < n; i++)
    tmp[i] = x;
  return tmp;
}

void bar() {
  sparray xs = fill_seq(4, 1234);
  std::cout << "xs = " << xs << std::endl;
}
```

Output after calling `bar()`:

```
xs = { 1234, 1234, 1234, 1234 }
```

Although it is perfectly fine to assign to an array variable the contents of a given array, what happens may be surprising to those who know the usual conventions of C++11 container libraries. Consider the following program.

Example 10.10 Move constructor

```
sparray xs = fill_seq(4, 1234);
sparray ys = fill_seq(3, 333);
ys = std::move(xs);
std::cout << "xs = " << xs << std::endl;
std::cout << "ys = " << ys << std::endl;
```

The assignment from `xs` to `ys` simultaneously destroys the contents of `ys` (by calling its destructor, which nulls it out), namely the array `{ 333, 333, 333 }`, moves the contents of `xs` to `ys`, and empties out the contents of `xs`. This behavior is defined as part of the move operator of `sparray`. The result is the following.

```
xs = { }
ys = { 1234, 1234, 1234, 1234 }
```

The reason we use this semantics for assignment is that the assignment takes constant time. Later, we are going to see that we can efficiently copy items out of an array. But for reasons we already discussed, the copy operation is going to be explicit.

Exercise: duplicating items in parallel

The aim of this exercise is to combine our knowledge of parallelism and arrays. To this end, the exercise is to implement two functions. The first, namely `duplicate`, is to return a new array in which each item appearing in the given array `xs` appears twice.

```
sparray duplicate(const sparray& xs) {
  // fill in
}
```

For example:

```
sparray xs = { 1, 2, 3 };
std::cout << "xs = " << duplicate(xs) << std::endl;
```

Expected output:

```
xs = { 1, 1, 2, 2, 3, 3 }
```

The second function is a generalization of the first: the value returned by `ktimes` should be an array in which each item `x` that is in the given array `xs` is replaced by `k` duplicate items.

```
sparray ktimes(const sparray& xs, long k) {
  // fill in
}
```

For example:

```
sparray xs = { 5, 7 };
std::cout << "xs = " << ktimes(xs, 3) << std::endl;
```

Expected output:

```
xs = { 5, 5, 5, 7, 7, 7 }
```

Notice that the `k` parameter of `ktimes` is not bounded. Your solution to this problem should be highly parallel not only in the number of items in the input array, `xs`, but also in the duplication-degree parameter, `k`.

1. What is the work and span complexity of your solution?
2. Does your solution expose ample parallelism? How much, precisely?
3. What is the speedup do you observe in practice on various input sizes?

10.4 Tabulation

A **tabulation** is a parallel operation which creates a new array of a given size and initializes the contents according to a given "generator function". The call `tabulate(g, n)` allocates an array of length `n` and assigns to each valid index in the array `i` the value returned by `g(i)`.

```
template <class Generator>
sparray tabulate(Generator g, long n);
```

Tabulations can be used to generate sequences according to a specified formula.

Example 10.11 Sequences of even numbers

```
sparray evens = tabulate([&] (long i) { return 2*i; }, 5);
std::cout << "evens = " << evens << std::endl;
```

Output:

```
evens = { 0, 2, 4, 6, 8 }
```

Copying an array can be expressed as a tabulation.

Example 10.12 Parallel array copy function using tabulation

```
sparray mycopy(const sparray& xs) {
  return tabulate([&] (long i) { return xs[i]; }, xs.size());
}
```

Exercise

Solve the `duplicate` and `ktimes` problems that were given in homework, this time using tabulations.

Solutions appear below.

Example 10.13 Solution to `duplicate` and `ktimes` exercises

```
sparray ktimes(const sparray& xs, long k) {
  long m = xs.size() * k;
  return tabulate([&] (long i) { return xs[i/k]; }, m);
}

sparray duplicate(const sparray& xs) {
  return ktimes(xs, 2);
}
```

The implementation of tabulate is a straightforward application of the parallel-for loop.

```
loop_controller_type tabulate_contr("tabulate");

template <class Generator>
sparray tabulate(Generator g, long n) {
  sparray tmp = sparray(n);
  parallel_for(tabulate_contr, (long)0, n, [&] (long i) {
    tmp[i] = g(i);
  });
  return tmp;
}
```

Note that the work and span of the generator function depends on the generator function passed as an argument to the tabulation. Let us first analyze for the simple case, where the generator function takes constant work (and hence, constant span). In this case, it should be clear that a tabulation should take work linear in the size of the array. The reason is that the only work performed by the body of the loop is performed by the constant-time generator function. Since the loop itself performs as many iterations as positions in the array, the work cost is indeed linear in the size of the array. The span cost of the tabulation is the sum of two quantitles: the span taken by the loop and the maximum value of the spans taken by the applications of the generator function. Recall that we saw before that the span cost of a parallel-for loop with n iterations is $\log n$. The maximum of the spans of the generator applications is a constant. Therefore, we can conclude that, in this case, the span cost is logarithmic in the size of the array.

The story is only a little more complicated when we generalize to consider non-constant time generator functions. Let $W(\mathrm{g}(i))$ denote the work performed by an application of the generator function to a specified value i. Similarly, let $S(\mathrm{g}(i))$ denote the span. Then the tabulation takes work

$$\sum_{i=0}^{n} W(\mathrm{g}(i))$$

and span

$$\log n + \max_{i=0}^{n} S(\mathrm{g}(i))$$

10.5 Higher-order granularity controllers

We just observed that each application of our tabulate operation can have different work and span cost depending on the selection of the generator function. Pause for a moment and consider how this observation could impact our granularity-control scheme. Consider, in particular, the way that the tabulate function uses its granularity-controller object, `tabulate_contr`. This one controller object is shared by every call site of `tabulate()`.

The problem is that all of the profiling data that the granularity controller collects at run time is lumped together, even though each generator function that is passed to the tabulate function can have completely different performance characteristics. The threshold that is best for one generator function is not necessarily good for another generator function. For this reason, there must be one distinct granularity-control object for each generator function that is passed to `tabulate`. For this reason, we refine our solution from the one above to the one below, which relies on C++ template programming to effectively key accesses to the granularity controller object by the type of the generator function.

```
template <class Generator>
class tabulate_controller {
public:
  static loop_controller_type contr;
};

template <class Generator>
loop_controller_type
  tabulate_controller<Generator>::contr("tabulate"+std::string(typeid(Generator).name()));

template <class Generator>
sparray tabulate(Generator g, long n) {
  sparray tmp = sparray(n);
  parallel_for(tabulate_controller<Generator>::contr, (long)0, n, [&] (long i) {
    tmp[i] = g(i);
  });
  return tmp;
}
```

To be more precise, the use of the template parameter in the class `tabulate_controller` ensures that each generator function in the program that is passed to `tabulate()` gets its own unique instance of the controller object. The rules of the template system regarding static class members that appear in templated classes ensure this behavior. Although it is not essential for our purposes to have a precise understanding of the template system, it is useful to know that the template system provides us with the exact mechanism that we need to properly separate granularity controllers of distinct instances of higher-order functions, such as tabulation.

Note: Above, we still assume constant-work generator functions.

10.6 Reduction

A **reduction** is an operation which combines a given set of values according to a specified **identity element** and a specified **associative combining operator**. Let S denote a set. Recall from algebra that an associative combining operator is any binary operator $\oplus$ such that, for any three items $x, y, z \in S$, the following holds.

$$x \oplus (y \oplus z) = (x \oplus y) \oplus z$$

An element $\mathbf{I} \in S$ is an identity element if for any $x \in S$ the following holds.

$$(x \oplus \mathbf{I}) = (\mathbf{I} \oplus x) = x$$

This algebraic structure consisting of $(S, \oplus, \mathbf{I})$ is called a **monoid** and is particularly worth knowing because this structure is a common pattern in parallel computing.

Example 10.14 Addition monoid

- S = the set of all 64-bit unsigned integers; $\oplus$ = addition modulo 2^{64}; $\mathbf{I} = 0$

Example 10.15 Multiplication monoid

- S = the set of all 64-bit unsigned integers; $\oplus$ = multiplication modulo 2^{64}; $\mathbf{I} = 1$

Example 10.16 Max monoid

- S = the set of all 64-bit unsigned integers; $\oplus$ = max function; $\mathbf{I} = 0$

The identity element is important because we are working with sequences: having a base element is essential for dealing with empty sequences. For example, what should the sum of the empty sequence? More interestingly, what should be the maximum (or minimum) element of an empty sequence? The identity element specifies this behavior.

What about the associativity of $\oplus$? Why does associativity matter? Suppose we are given the sequence $[a_0, a_1, \ldots, a_n]$. The serial reduction of this sequence always computes the expression $(a_0 \oplus a_1 \oplus \ldots \oplus a_n)$. However, when the reduction is performed in parallel, the expression computed by the reduction could be $((a_0 \oplus a_1 \oplus a_2 \oplus a_3) \oplus (a_4 \oplus a_5) \oplus \ldots \oplus (a_{n-1} \oplus a_n))$ or $((a_0 \oplus a_1 \oplus a_2) \oplus (a_3 \oplus a_4 \oplus a_5) \oplus \ldots \oplus (a_{n-1} \oplus a_n))$. In general, the exact placement of the parentheses in the parallel computation depends on the way that the parallel algorithm decomposes the problem. Associativity gives the parallel algorithm the flexibility to choose an efficient order of evaluation and still get the same result in the end. The flexibility to choose the decomposition of the problem is exploited by efficient parallel algorithms, for reasons that should be clear by now. In summary, associativity is a key building block to the solution of many problems in parallel algorithms.

Now that we have monoids for describing a generic method for combining two items, we can consider a generic method for combining many items in parallel. Once we have this ability, we will see that we can solve the remaining problems from last homework by simply plugging the appropriate monoids into our generic operator, `reduce`. The interface of this operator in our framework is specified below. The first parameter corresponds to $\oplus$, the second to the identity element, and the third to the sequence to be processed.

```
template <class Assoc_binop>
value_type reduce(Assoc_binop b, value_type id, const sparray& xs);
```

We can solve our first problem by plugging integer plus as $\oplus$ and 0 as $\mathbf{I}$.

Example 10.17 Summing elements of array

```
auto plus_fct = [&] (value_type x, value_type y) {
  return x+y;
};

sparray xs = { 1, 2, 3 };
std::cout << "sum_xs = " << reduce(plus_fct, 0, xs) << std::endl;
```

Output:

```
reduce(plus_fct, 0, xs) = 6
```

We can solve our second problem in a similar fashion. Note that in this case, since we know that the input sequence is nonempty, we can pass the first item of the sequence as the identity element. What could we do if we instead wanted a solution that can deal with zero-length sequences? What identity element might make sense in that case? Why?

Example 10.18 Taking max of elements of array

Let us start by solving a special case: the one where the input sequence is nonempty.

```
auto max_fct = [&] (value_type x, value_type y) {
  return std::max(x, y);
};

sparray xs = { -3, 1, 634, 2, 3 };
std::cout << "reduce(max_fct, xs[0], xs) = " << reduce(max_fct, xs[0], xs) << std::endl;
```

Output:

```
reduce(max_fct, xs[0], xs) = 634
```

Observe that in order to seed the reduction we selected the provisional maximum value to be the item at the first position of the input sequence. Now let us handle the general case by seeding with the smallest possible value of type `long`.

```
long max(const sparray& xs) {
  return reduce(max_fct, LONG_MIN, xs);
}
```

The value of `LONG_MIN` is defined by `<limits.h>`.

Like the tabulate function, reduce is a higher-order function. Just like any other higher-order function, the work and span costs have to account for the cost of the client-supplied function, which is in this case, the associative combining operator.

10.7 Scan

A **scan** is an iterated reduction that is typically expressed in one of two forms: inclusive and exclusive. The inclusive form maps a given sequence $[x_0, x_1, x_2, \ldots, x_{n-1}]$ to $[x_0, x_0 \oplus x_1, x_0 \oplus x_1 \oplus x_2, \ldots, x_0 \oplus x_1 \oplus \ldots \oplus x_{n-1}]$.

```
template <class Assoc_binop>
sparray scan_incl(Assoc_binop b, value_type id, const sparray& xs);
```

Example 10.19 Inclusive scan

```
  scan_incl(b, 0, sparray({ 2, 1, 8, 3 }))
= { reduce(b, id, { 2 }),       reduce(b, id, { 2, 1 }),
    reduce(b, id, { 2, 1, 8 }), reduce(b, id, { 2, 1, 8, 3}) }
= { 0+2, 0+2+1, 0+2+1+8, 0+2+1+8+3 }
= { 2, 3, 11, 14 }
```

The exclusive form maps a given sequence $[x_0, x_1, x_2, \ldots, x_{n-1}]$ to $[\mathbf{I}, x_0, x_0 \oplus x_1, x_0 \oplus x_1 \oplus x_2, \ldots, x_0 \oplus x_1 \oplus \ldots \oplus x_{n-2}]$. For convenience, we extend the result of the exclusive form with the total $x_0 \oplus \ldots \oplus x_{n-1}$.

```
class scan_excl_result {
public:
  sparray partials;
  value_type total;
};

template <class Assoc_binop>
scan_excl_result scan_excl(Assoc_binop b, value_type id, const sparray& xs);
```

Example 10.20 Exclusive scan

The example below represents the logical behavior of `scan`, but actually says nothing about the way scan is implemented.

```
  scan_excl(b, 0, { 2, 1, 8, 3 }).partials
= { reduce(b, 0, { }),       reduce(b, 0, { 2 }),
    reduce(b, 0, { 2, 1 }), reduce(b, 0, { 2, 1, 8}) }
= { 0, 0+2, 0+2+1, 0+2+1+8 }
= { 0, 2, 3, 11 }

  scan_excl(b, 0, { 2, 1, 8, 3 }).total
= reduce(b, 0, { 2, 1, 8, 3})
= { 0+2+1+8+3 }
= 14
```

Scan has applications in many parallel algorithms. To name just a few, scan has been used to implement radix sort, search for regular expressions, dynamically allocate processors, evaluate polynomials, etc. Suffice to say, scan is important and worth knowing about because scan is a key component of so many efficient parallel algorithms. In this course, we are going to study a few more applications not in this list.

The expansions shown above suggest the following sequential algorithm.

```
template <class Assoc_binop>
scan_excl_result scan_excl_seq(Assoc_binop b, value_type id, const sparray& xs) {
  long n = xs.size();
  sparray r = array(n);
  value_type x = id;
  for (long i = 0; i < n; i++) {
    r[i] = x;
    x = b(x, xs[i]);
  }
  return make_scan_result(r, x);
}
```

If we just blindly follow the specification above, we might be tempted to try the solution below.

```
loop_controller_type scan_contr("scan");

template <class Assoc_binop>
sparray scan_excl(Assoc_binop b, value_type id, const sparray& xs) {
  long n = xs.size();
  sparray result = array(n);
  result[0] = id;
  parallel_for(scan_contr, 1l, n, [&] (long i) {
    result[i] = reduce(b, id, slice(xs, 0, i-1));
  });
  return result;
}
```

Question

Although it is highly parallel, this solution has a major problem. What is it?

Consider that our sequential algorithm takes linear time in the size of the input array. As such, finding a work-efficient parallel solution means finding a solution that also takes linear work in the size of the input array. The problem is that our parallel algorithm takes quadratic work: it is not even asymptotically work efficient! Even worse, the algorithm performs a lot of redundant work.

Can we do better? Yes, in fact, there exist solutions that take, in the size of the input, both linear time and logarithmic span, assuming that the given associative operator takes constant time. It might be worth pausing for a moment to consider this fact, because the specification of scan may at first look like it would resist a solution that is both highly parallel and work efficient.

10.8 Derived operations

The remaining operations that we are going to consider are useful for writing more succinct code and for expressing special cases where certain optimizations are possible. All of the the operations that are presented in this section are derived forms of tabulate, reduce, and scan.

10.8.1 Map

The `map(f, xs)` operation applies `f` to each item in `xs` returning the array of results. It is straightforward to implement as a kind of tabulation, as we have at our disposal efficient indexing.

```
template <class Func>
sparray map(Func f, sparray xs) {
  return tabulate([&] (long i) { return f(xs[i]); }, xs.size());
}
```

The array-increment operation that we defined on the first day of lecture is simple to express via `map`.

Example 10.21 Incrementing via `map`

```
sparray map_incr(sparray xs) {
  return map([&] (value_type x) { return x+1; }, xs);
}
```

The work and span costs of `map` are similar to those of tabulate. Granularity control is handled similarly as well. However, that the granularity controller object corresponding to `map` is instantiated properly is not obvious. It turns out that, for no extra effort, the behavior that we want is indeed preserved: each distinct function that is passed to `map` is assigned a distinct granularity controller. Although it is outside the scope of this course, the reason that this scheme works in our current design owes to specifics of the C++ template system.

10.8.2 Fill

The call `fill(v, n)` creates an array that is initialized with a specified number of items of the same value. Although just another special case for tabulation, this function is worth having around because internally the `fill` operation can take advantage of special hardware optimizations, such as SIMD instructions, that increase parallelism.

```
sparray fill(value_type v, long n);
```

Example 10.22 Creating an array of all 3s

```
sparray threes = fill(3, 5);
std::cout << "threes = " << threes << std::endl;
```

Output:

```
threes = { 3, 3, 3, 3, 3 }
```

10.8.3 Copy

Just like `fill`, the `copy` operation can take advantage of special hardware optimizations that accellerate memory traffic. For the same reason, the `copy` operation is a good choice when a full copy is needed.

```
sparray copy(const sparray& xs);
```

Example 10.23 Copying an array

```
sparray xs = { 3, 2, 1 };
sparray ys = copy(xs);
std::cout << "xs = " << xs << std::endl;
std::cout << "ys = " << ys << std::endl;
```

Output:

```
xs = { 3, 2, 1 }
ys = { 3, 2, 1 }
```

10.8.4 Slice

We now consider a slight generalization on the copy operator: with the `slice` operation we can copy out a range of positions from a given array rather than the entire array.

```
sparray slice(const sparray& xs, long lo, long hi);
```

The `slice` operation takes a source array and a range to copy out and returns a fresh array that contains copies of the items in the given range.

Example 10.24 Slicing an array

```
sparray xs = { 1, 2, 3, 4, 5 };
std::cout << "slice(xs, 1, 3) = " << slice(xs, 1, 3) << std::endl;
std::cout << "slice(xs, 0, 4) = " << slice(xs, 0, 4) << std::endl;
```

Output:

```
{ 2, 3 }
{ 1, 2, 3, 4 }
```

10.8.5 Concat

In contrast to `slice`, the `concat` operation lets us "copy in" to a fresh array.

```
sparray concat(const sparray& xs, const sparray& ys);
```

Example 10.25 Concatenating two arrays

```
sparray xs = { 1, 2, 3 };
sparray ys = { 4, 5 };
std::cout << "concat(xs, ys) = " << concat(xs, ys) << std::endl;
```

Output:

```
{ 1, 2, 3, 4, 5 }
```

10.8.6 Prefix sums

The prefix sums problem is a special case of the scan problem. We have defined two solutions for two variants of the problem: one for the exclusive prefix sums and one for the inclusive case.

```
sparray prefix_sums_incl(const sparray& xs);
scan_excl_result prefix_sums_excl(const sparray& xs);
```

Example 10.26 Inclusive and exclusive prefix sums

```
sparray xs = { 2, 1, 8, 3 };
sparray incl = prefix_sums_incl(xs);
scan_excl_result excl = prefix_sums_excl(xs);
std::cout << "incl = " << incl << std::endl;
std::cout << "excl.partials = " << excl.partials << "; excl.total = " << excl.total << std ↩
    ::endl;
```

Output:

```
incl = { 2, 3, 11, 14 }
excl.partials = { 0, 2, 3, 11 }; excl.total = 147
```

10.8.7 Filter

The last data-parallel operation that we are going to consider is the operation that copies out items from a given array based on a given predicate function.

```
template <class Predicate>
sparray filter(Predicate pred, const sparray& xs);
```

For our purposes, a predicate function is any function that takes a value of type `long` (i.e., `value_type`) and returns a value of type `bool`.

Example 10.27 Extracting even numbers

The following function copies out the even numbers it receives in the array of its argument.

```
bool is_even(value_type x) {
  return (x%2) == 0;
}

sparray extract_evens(const sparray& xs) {
  return filter([&] (value_type x) { return is_even(x); }, xs);
}

sparray xs = { 3, 5, 8, 12, 2, 13, 0 };
std::cout << "extract_evens(xs) = " << extract_evens(xs) << std::endl;
```

Output:

```
extract_evens(xs) = { 8, 12, 2, 0 }
```

Example 10.28 Solution to the sequential-filter problem

The particular instance of the filter problem that we are considering is a little tricky because we are working with fixed-size arrays. In particular, what requires care is the method that we use to copy the selected items out of the input array to the output array. We need to first run a pass over the input array, applying the predicate function to the items, to determine which items are to be written to the output array. Furthermore, we need to track how many items are to be written so that we know how much space to allocate for the output array.

```
template <class Predicate>
sparray filter(Predicate pred, const sparray& xs) {
  long n = xs.size();
  long m = 0;
  sparray flags = array(n);
  for (long i = 0; i < n; i++)
    if (pred(xs[i])) {
      flags[i] = true;
      m++;
    }
  sparray result = array(m);
  long k = 0;
  for (long i = 0; i < n; i++)
    if (flags[i])
      result[k++] = xs[i];
  return result;
}
```

Question

In the sequential solution above, it appears that there are two particular obstacles to parallelization. What are they?

Hint: the obstacles relate to the use of variables `m` and `k`.

Question

Under one particular assumption regarding the predicate, this sequential solution takes linear time in the size of the input, using two passes. What is the assumption?

10.8.8 Parallel-filter problem

The starting point for our solution is the following code.

```
template <class Predicate>
sparray filter(Predicate p, const sparray& xs) {
  sparray flags = map(p, xs);
  return pack(flags, xs);
}
```

The challenge of this exercise is to solve the following problem: given two arrays of the same size, the first consisting of boolean valued fields and the second containing the values, return the array that contains (in the same relative order as the items from the input) the values selected by the flags. Your solution should take linear work and logarithmic span in the size of the input.

```
sparray pack(const sparray& flags, const sparray& xs);
```

Example 10.29 The allocation problem

```
sparray flags = { true, false, false, true, false, true, true };
sparray xs    = {   34,    13,     5,    1,    41,   11,   10 };
std::cout << "pack(flags, xs) = " << pack(flags, xs) << std::endl;
```

Output:

```
pack(flags, xs) = { 34, 1, 11, 10 }
```

Tip

You can use scans to implement `pack`.

Note

Even though our arrays can store only 64-bit values of type `long`, we can nevertheless store values of type `bool`, as we have done just above with the `flags` array. The compiler automatically promotes boolean values to long values without causing us any problems, at least with respect to the correctness of our solutions. However, if we want to be more space efficient, we need to use arrays that are capable of packing values of type `bool` more efficiently, e.g., into single- or eight-bit fields. It should easy to convince yourself that achieving such specialized arrays is not difficult, especially given that the template system makes it easy to write polymorphic containers.

10.9 Summary of operations

10.9.1 Tabulate

```
template <class Generator>
sparray tabulate(Generator g, long n);
```

The call `tabulate(g, n)` returns the length-`n` array where the `i` th element is given by `g(i)`.

Let $W(\mathrm{g}(i))$ denote the work performed by an application of the generator function to a specified value i. Similarly, let $S(\mathrm{g}(i))$ denote the span. Then the tabulation takes work

$$\sum_{i=0}^{n} W(\mathrm{g}(i))$$

and span

$$\log \mathrm{n} + \max_{i=0}^{n} S(\mathrm{g}(i))$$

10.9.2 Reduce

```
template <class Assoc_binop>
value_type reduce(Assoc_binop b, value_type id, const sparray& xs);
```

The call `reduce(b, id, xs)` is logically equal to `id` if `xs.size() ==0`, `xs[0]` if `xs.size() ==1`, and

```
b(reduce(b, id, slice(xs, 0, n/2)),
  reduce(b, id, slice(xs, n/2, n)))
```

otherwise where `n ==xs.size()`.

The work and span cost are $O(n)$ and $O(\log n)$ respectively, where n denotes the size of the input sequence `xs`. This cost assumes that the work and span of `b` are constant. If it's not the case, then refer directly to the implementation of `reduce`.

10.9.3 Scan

```
template <class Assoc_binop>
sparray scan_incl(Assoc_binop b, value_type id, const sparray& xs);
```

For an associative function `b` and corresponding identity `id`, the return result of the call `scan_incl(b, id, xs)` is logically equivalent to

```
tabulate([&] (long i) { return reduce(b, id, slice(xs, 0, i+1)); }, xs.size())
```

```
class scan_excl_result {
public:
  sparray partials;
  value_type total;
};

template <class Assoc_binop>
scan_excl_result scan_excl(Assoc_binop b, value_type id, const sparray& xs);
```

For an associative function `b` and corresponding identity `id`, the call `scan_excl(b, id, xs)` returns the object `res`, such that `res.partials` is logically equivalent to

```
tabulate([&] (long i) { return reduce(b, id, slice(xs, 0, i)); }, xs.size())
```

and `res.total` is logically equivalent to

```
reduce(b, id, xs)
```

The work and span cost are $O(n)$ and $O(\log n)$ respectively, where n denotes the size of the input sequence `xs`. This cost assumes that the work and span of `b` are constant. If it's not the case, then refer directly to the implementation of `scan_incl` and `scan_excl`.

10.9.4 Map

```
template <class Func>
sparray map(Func f, sparray xs) {
  return tabulate([&] (long i) { return f(xs[i]); }, xs.size());
}
```

Let $W(\texttt{f}(x))$ denote the work performed by an application of the function `f` to a specified value x. Similarly, let $S(\texttt{f}(x))$ denote the span. Then the map takes work

$$\sum_{x \in xs} W(\texttt{f}(x))$$

and span

$$\log \texttt{xs.size}() + \max_{x \in xs} S(\texttt{f}(x))$$

10.9.5 Fill

```
sparray fill(value_type v, long n);
```

Returns a length-`n` array with all cells initialized to `v`.

Work and span are linear and logarithmic, respectively.

10.9.6 Copy

```
sparray copy(const sparray& xs);
```

Returns a fresh copy of `xs`.

Work and span are linear and logarithmic, respectively.

10.9.7 Slice

```
sparray slice(const sparray& xs, long lo, long hi);
```

The call `slice(xs, lo, hi)` returns the array `{xs[lo], xs[lo+1], ..., xs[hi-1]}`.

Work and span are linear and logarithmic, respectively.

10.9.8 Concat

```
sparray concat(const sparray& xs, const sparray& ys);
```

Concatenate the two sequences.

Work and span are linear and logarithmic, respectively.

10.9.9 Prefix sums

```
sparray prefix_sums_incl(const sparray& xs);
```

The result of the call `prefix_sums_incl(xs)` is logically equivalent to the result of the call

```
scan_incl([&] (value_type x, value_type y) { return x+y; }, 0, xs)
```

```
scan_excl_result prefix_sums_excl(const sparray& xs);
```

The result of the call `prefix_sums_excl(xs)` is logically equivalent to the result of the call

```
scan_excl([&] (value_type x, value_type y) { return x+y; }, 0, xs)
```

Work and span are linear and logarithmic, respectively.

10.9.10 Filter

```
template <class Predicate>
sparray filter(Predicate p, const sparray& xs);
```

The call `filter(p, xs)` returns the subsequence of `xs` which contains each `xs[i]` for which `p(xs[i])` returns `true`.

Warning
Depending on the implementation, the predicate function `p` may be called multiple times by the `filter` function.

Work and span are linear and logarithmic, respectively.

11 Chapter: Parallel Sorting

In this chapter, we are going to study parallel implementations of quicksort and mergesort.

11.1 Quicksort

The quicksort algorithm for sorting an array (sequence) of elements is known to be a very efficient sequential sorting algorithm. A natural question thus is whether quicksort is similarly effective as a parallel algorithm?

Let us first convince ourselves, at least informally, that quicksort is actually a good parallel algorithm. But first, what do we mean by "parallel quicksort." Chances are that you think of quicksort as an algorithm that, given an array, starts by reorganizing the elements in the array around a randomly selected pivot by using an in-place *partitioning* algorithm, and then sorts the two parts of the array to the left and the right of the array recursively.

While this implementation of the quicksort algorithm is not immediately parallel, it can be parallelized. Note that the recursive calls are naturally independent. So we really ought to focus on the partitioning algorithm. There is a rather simple way to do such a partition in parallel by performing three `filter` calls on the input array, one for picking the elements less that the pivot, one for picking the elements equal to the pivot, and another one for picking the elements greater than the pivot. This algorithm can be described as follows.

Algorithm: parallel quicksort

1. Pick from the input sequence a pivot item.
2. Based on the pivot item, create a three-way partition of the input sequence:
    a. the sequence of items that are less than the pivot item,
    b. those that are equal to the pivot item, and
    c. those that are greater than the pivot item.
3. Recursively sort the "less-than" and "greater-than" parts
4. Concatenate the sorted arrays.

Now that we have a parallel algorithm, we can check whether it is a good algorithm or not. Recall that a good parallel algorithm is one that has the following three characteristics

1. It is asymptotically work efficient
2. It is observably work efficient
3. It is highly parallel, i.e., has low span.

Let us first convince ourselves that Quicksort is a highly parallel algorithm. Observe that

1. the dividing process is highly parallel because no dependencies exist among the intermediate steps involved in creating the three-way partition,
2. two recursive calls are parallel, and
3. concatenations are themselves highly parallel.

11.1.1 Asymptotic Work Efficiency and Parallelism

Let us now turn our attention to asymptotic and observed work efficiency. Recall first that quicksort can exhibit a quadratic-work worst-case behavior on certain inputs if we select the pivot deterministically. To avoid this, we can pick a random element as a pivot by using a random-number generator, but then we need a parallel random number generator. Here, we are going to side-step this issue by assuming that the input is randomly permuted in advance. Under this assumption, we can simply pick the pivot to be the first item of the sequence. With this assumption, our algorithm performs asymptotically the same work as sequential quicksort implementations that perform $\Theta(n \log n)$ in expectation.

For the analysis, let's assume a version of quicksort that compares the pivot p to each key in the input once (instead of 3 times. The figure below illustrates the structure of an execution of quicksort by using a tree. Each node corresponds to a call to the quicksort function and is labeled with the key at that call. Note that the tree is a binary search tree.

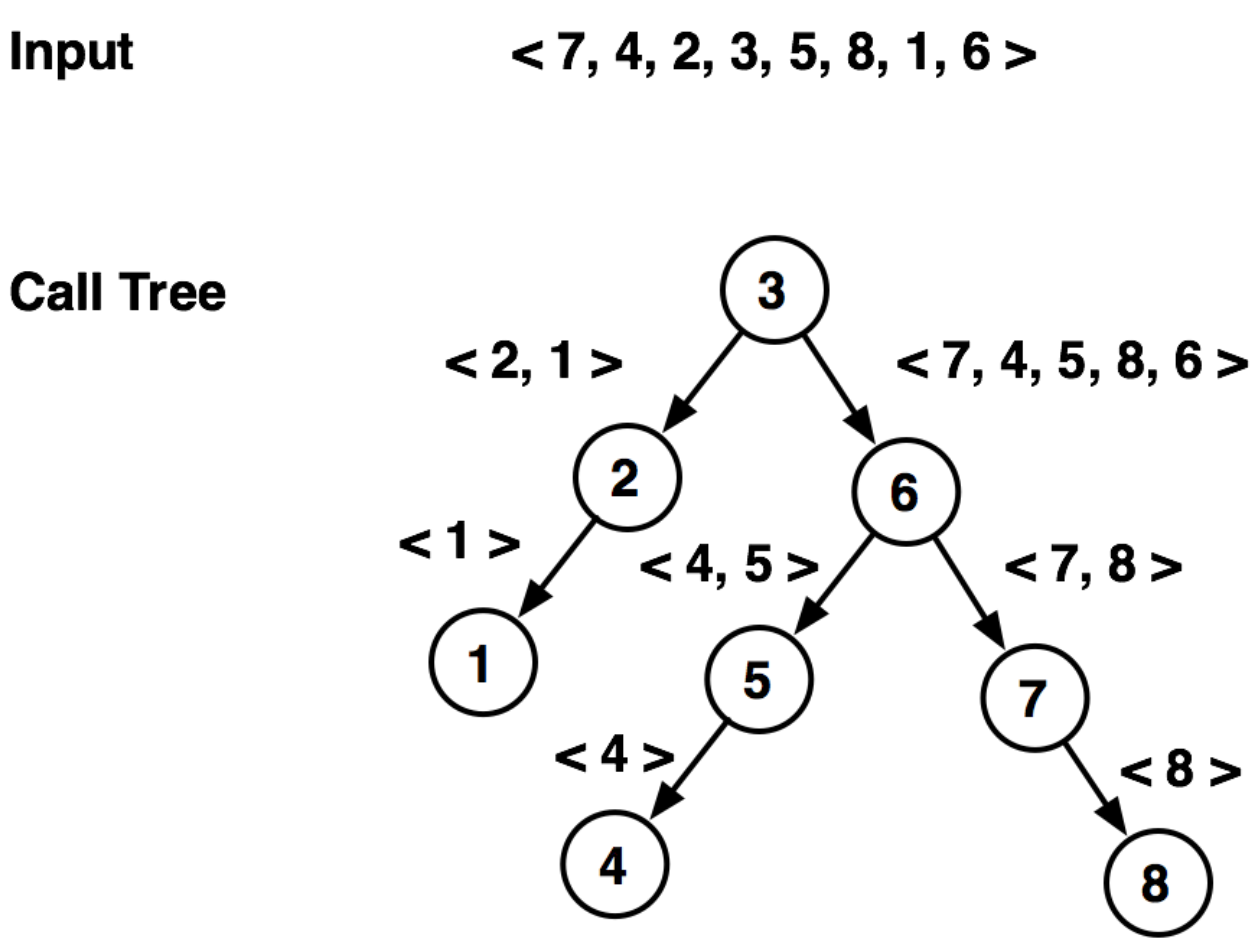


Figure 12: Quicksort call tree.

Let's observe some properties of quicksort.

1. In quicksort, a comparison always involves a pivot and another key. Since, the pivot is never sent to a recursive call, a key is selected as a pivot exactly once, and is not involved in further comparisons (after is becomes a pivot). Before a key is selected as a pivot, it may be compared to other pivots, once per pivot, and thus two keys are never compared more than once.
2. When the algorithm selects a key y as a pivot and if y is between two other keys x, z such that $x < y < z$, it sends the two keys latexmath:[y, z to two separate subtrees. The two keys x and z separated in this way are never compared again.
3. We can sum up the two observations: a key is compared with all its ancestors in the call tree and all its descendants in the call tree, and with no other keys.

Since a pair of key are never compared more than once, the total number of comparisons performed by quicksort can be expressed as the sum over all pairs of keys.

Let X_n be the random variable denoting the total number of comparisons performed in an execution of quicksort with a randomly permuted input of size latexmath:[n.

We want to bound the expectation of X_n, $E[X_n]$.

For the analysis, let's consider the final sorted order of the keys T, which corresponds to the output. Consider two positions $i, j \in \{1, \ldots, n\}$ in the sequence T. We define following random variable:

$$A_{ij} = \begin{cases} 1 & \text{if } T_i \text{ and } T_j \text{are compared} \\ 0 & \text{otherwise} \end{cases}$$

We can write X_n by summing over all A_{ij}'s:

$$X_n \leq \sum_{i=1}^{n} \sum_{j=i+1}^{n} A_{ij}$$

By linearity of expectation, we have

$$E[X_n] \leq \sum_{i=1}^{n} \sum_{j=i+1}^{n} E[A_{ij}]$$

Furthermore, since each A_{ij} is an indicator random variable, $E[A_{ij}] = P(A_{ij}) = 1$. Our task therefore comes down to computing the probability that T_i and T_j are compared, i.e., $P(A_{ij} = 1)$, and working out the sum.

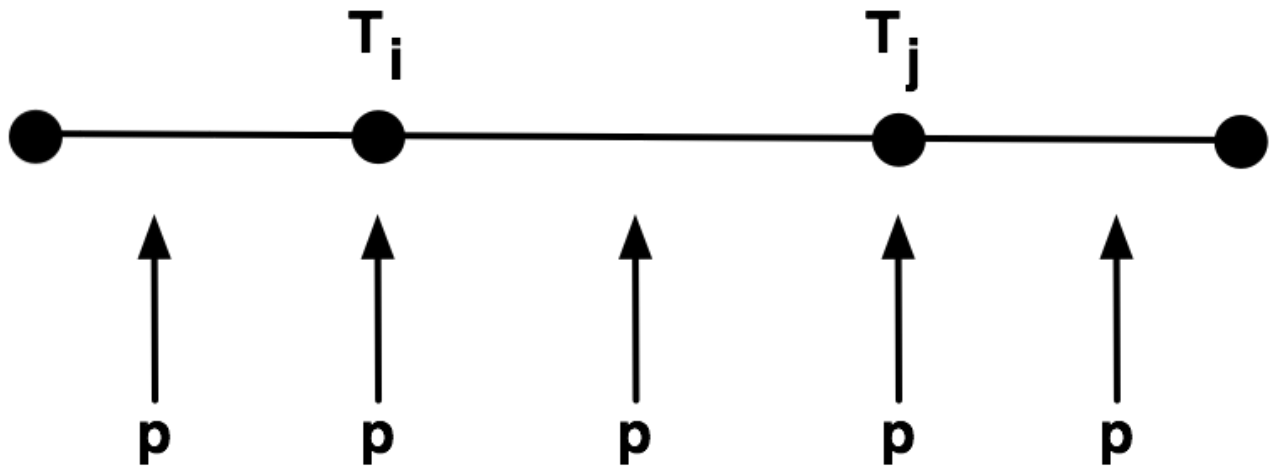


Figure 13: Relationship between the pivot and other keys.

To compute this probability, note that each call takes the pivot p and splits the sequence into two parts, one with keys larger than p and the other with keys smaller than p. For any one call to quicksort there are three possibilities as illustrated in the figure above.

1. The pivot is (equal to) either T_i or T_j, in which case T_i and T_j are compared and $A_{ij} = 1$. Since there are $j-i+1$ keys in the interval $T_i \ldots T_j$ and since each one is equally likely to be the first in the randomly permuted input, the $P(A_{ij} = 1) = \frac{2}{j-i+1}$.
2. The pivot is a key between T_i and T_j, and T_i and T_j will never be compared; thus $A_{ij} = 0$.
3. The pivot is less than T_i or greater than T_j. Then latexmath:[T_i and T_j are sent to the same recursive call. Whether T_i and T_j are compared will be determined in some later call.

$$\begin{aligned} E[X_n] &\leq \textstyle\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[A_{ij}] \\ &= \textstyle\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \\ &= \textstyle\sum_{i=1}^{n-1} n \sum_{k=2}^{n-i+1} \frac{2}{k} \\ &\leq 2 \textstyle\sum_{i=1}^{n-1} H_n \\ &= 2nH_n \in O(n \log n) \end{aligned}$$

The last step follows by the fact that $H_n = \ln n + O(1)$.

Having completed work, let's analyze the span of quicksort. Recall that each call to quicksort partitions the input sequence of length n into three subsequences L, E, and R, consisting of the elements less than, equal to, and greater than the pivot.

Let's first bound the size of the left sequence L.

$$\begin{aligned} E[|L|] &= \textstyle\sum_{i=1}^{n-1} \frac{i-1}{n} \\ &\leq \frac{n}{2}. \end{aligned}$$

By symmetry, $E[|R|] \leq \frac{n}{2}$. This reasoning applies at any level of the quicksort call tree. In other words, the size of the input decreases by $1/2$ in expectation at each level in the call tree.

Since pivot choice at each call is independent of the other calls. The expected size of the input at level i is $E[Y_i] = \frac{n}{2^i}$.

Let's calculate the expected size of the input at depth $i = 5\lg n$. By basic arithmetic, we obtain

$$E[Y_{5\lg n}] = n\frac{1}{2^{5\lg n}} = nn^{-5\lg 2} = n^{-4}.$$

Since Y_i's are always non-negative, we can use Markov's inequality, to turn expectations into probabilities as follows.

$$P(Y_{5\lg n} \geq 1) \leq \frac{E[Y_{5\lg n}]}{1} = n^{-4}.$$

In other words, the probability that a given path in the quicksort call tree has a depth that exceeds $5\lg n$ is tiny.

From this bound, we can calculute a bound on the depth of the whole tree. Note first that the tree has exactly $n+1$ leaves because it has n internal nodes. Thus we have at most $n+1$ to consider. The probability that a given path expeeds a depth of $5\lg n$ is n^{-4}. Thus, by union bound the probability that any one of the paths exceed the depeth of $5\lg n$ is $(n+1)\cdot n^{-4} \leq n^{-2}$. Thus the probability that the depth of the tree is greater than $5\lg n$ is n^{-2}.

By using Total Expectation Theorem or the Law of total expectation, we can now calculate expected span by dividing the sample space into mutually exclusive and exhaustive space as follows.

$$\begin{aligned} E[S] &= E[S \mid \text{Depth is no greater than } 5\lg n] \cdot P(\text{Depth is no greater than } 5\lg n) + E[S \mid \text{Depth is greater than } 5\lg n] \cdot P(\text{Depth is greate} \\ &\leq \lg^2 n \cdot (1 - 1/n^2) + n^2 \cdot 1/n^2 \\ &= O(lg^2 n) \end{aligned}$$

In thes bound, we used the fact that each call to quicksort has a span of $O(\lg n)$ because this is the span of `filter`.

Here is an alternative analysis.

Let $M_n = \max\{|L|, |R|\}$, which is the size of larger subsequence. The span of quicksort is determined by the sizes of these larger subsequences. For ease of analysis, we will assume that $|E| = 0$, as more equal elements will only decrease the span. As the partition step uses `filter` we have the following recurrence for span:

$$S(n) = S(M_n) + O(\log n)$$

To develop some intuition for the span analysis, let's consider the probability that we split the input sequence more or less evenly. If we select a pivot that is greater than $T_{n/4}$ and less than $T_{3n/4}$ then M_n is at most $3n/4$. Since all keys are equally likely to be selected as a pivot this probability is $\frac{3n/4 - n/4}{n} = 1/2$. The figure below illustrates this.

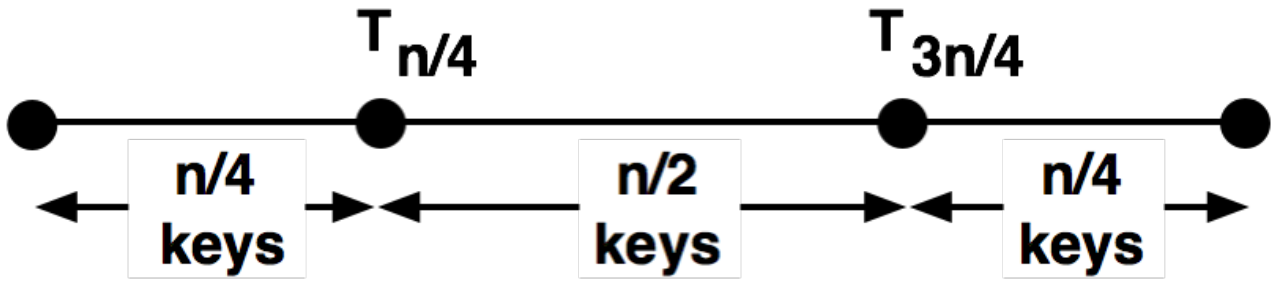


Figure 14: Quicksort span intuition.

This observations implies that at each level of the call tree (every time a new pivot is selected), the size of the input to both calls decrease by a constant fraction (of $4/3$). At every two levels, the probability that the input size decreases by latexmath:[4/3 is the probability that it decreases at either step, which is at least $3/4$, etc. Thus at a small costant number of steps, the probability

that we observe a $4/3$ factor decrease in the size of the input approches 1 quickly. This suggest that at some after $c \log n$ levels quicksort should complete. We now make this intuition more precise.

For the analysis, we use the conditioning technique for computing expectations as suggested by the total expectation theorem. Let X be a random variable and let A_i be disjoint events that form a a partition of the sample space such that $P(A_i) > 0$. The Total Expectation Theorem or the Law of total expectation states that

$$E[X] = \sum_{i=1}^{n} P(A_i) \cdot E[X|A_i].$$

Note first that $P(X_n \leq 3n/4) = 1/2$, since half of the randomly chosen pivots results in the larger partition to be at most $3n/4$ elements: any pivot in the range $T_{n/4}$ to $T_{3n/4}$ will do, where T is the sorted input sequence.

By conditioning S_n on the random variable M_n, we write,

$$E[S_n] = \sum_{m=n/2}^{n} P(M_n = m) \cdot E[S_n|(M_n = m)].$$

We can re-write this

$$E[S_n] = \sum_{m=n/2}^{n} P(M_n = m) \cdot E[S_m]$$

The rest is algebra

$$\begin{aligned} E[S_n] &= \textstyle\sum_{m=n/2}^{n} P(M_n = m) \cdot E[S_m] \\ &\leq P(M_n \leq \tfrac{3n}{4}) \cdot E[S_{\frac{3n}{4}}] + P(M_n > \tfrac{3n}{4}) \cdot E[S_n] + c \cdot \log n \\ &\leq \tfrac{1}{2} E[S_{\frac{3n}{4}}] + \tfrac{1}{2} E[S_n] \\ &\implies E[S_n] \leq E[S_{\frac{3n}{4}}] + 2c \log n. \end{aligned}$$

This is a recursion in $E[S(\cdot)]$ and solves easily to $E[S(n)] = O(\log^2 n)$.

11.1.2 Observable Work Efficency and Scalability

For an implementation to be observably work efficient, we know that we must control granularity by switching to a fast sequential sorting algorithm when the input is small. This is easy to achieve using our granularity control technique by using `seqsort()`, a fast sequential algorithm provided in the code base; `seqsort()` is really a call to STL's sort function. Of course, we have to assess observable work efficiency experimentally after specifying the implementation.

The code for quicksort is shown below. Note that we use our array class `sparray` to store the input and output. To partition the input, we use our parallel `filter` function from the previous lecture to parallelize the partitioning phase. Similarly, we use our parallel concatenation function to constructed the sorted output.

```
controller_type quicksort_contr("quicksort");

sparray quicksort(const sparray& xs) {
  long n = xs.size();
  sparray result = { };
  cstmt(quicksort_contr, [&] { return n * std::log2(n); }, [&] {
    if (n == 0) {
      result = { };
    } else if (n == 1) {
      result = { xs[0] };
    } else {
      value_type p = xs[0];
      sparray less = filter([&] (value_type x) { return x < p; }, xs);
      sparray equal = filter([&] (value_type x) { return x == p; }, xs);
      sparray greater = filter([&] (value_type x) { return x > p; }, xs);
      sparray left = { };
      sparray right = { };
      fork2([&] {
        left = quicksort(less);
```

```
      }, [&] {
        right = quicksort(greater);
      });
      result = concat(left, equal, right);
    }
  }, [&] {
    result = seqsort(xs);
  });
  return result;
}
```

By using randomized-analysis techniques, it is possible to analyze the work and span of this algorithm. The techniques needed to do so are beyond the scope of this book. The interested reader can find more details in another book.

Fact

The randomized quicksort algorithm above has expected work of $O(n \log n)$ and expected span of $O(\log^2 n)$, where n is the number of items in the input sequence.

One consequence of the work and span bounds that we have stated above is that our quicksort algorithm is highly parallel: its average parallelism is $\frac{O(n \log n)}{O(\log^2 n)} = \frac{n}{\log n}$. When the input is large, there should be ample parallelism to keep many processors well fed with work. For instance, when $n = 100$ million items, the average parallelism is $\frac{10^8}{\log 10^8} \approx \frac{10^8}{40} \approx 3.7$ million. Since 3.7 million is much larger than the number of processors in our machine, that is, forty, we have a ample parallelism.

Unfortunately, the code that we wrote leaves much to be desired in terms of observable work efficiency. Consider the following benchmarking runs that we performed on our 40-processor machine.

```
$ prun speedup -baseline "bench.baseline" -parallel "bench.opt -proc 1,10,20,30,40" -bench ↩
    quicksort -n 100000000
```

The first two runs show that, on a single processor, our parallel algorithm is roughly 6x slower than the sequential algorithm that we are using as baseline! In other words, our quicksort appears to have "6-observed work efficiency". That means we need at least six processors working on the problem to see even a small improvement compared to a good sequential baseline.

```
[1/6]
bench.baseline -bench quicksort -n 100000000
exectime 12.518
[2/6]
bench.opt -bench quicksort -n 100000000 -proc 1
exectime 78.960
```

The rest of the results confirm that it takes about ten processors to see a little improvement and forty processors to see approximately a 2.5x speedup. This plot shows the speedup plot for this program. Clearly, it does not look good.

```
[3/6]
bench.opt -bench quicksort -n 100000000 -proc 10
exectime 9.807
[4/6]
bench.opt -bench quicksort -n 100000000 -proc 20
exectime 6.546
[5/6]
bench.opt -bench quicksort -n 100000000 -proc 30
exectime 5.531
[6/6]
bench.opt -bench quicksort -n 100000000 -proc 40
exectime 4.761
```

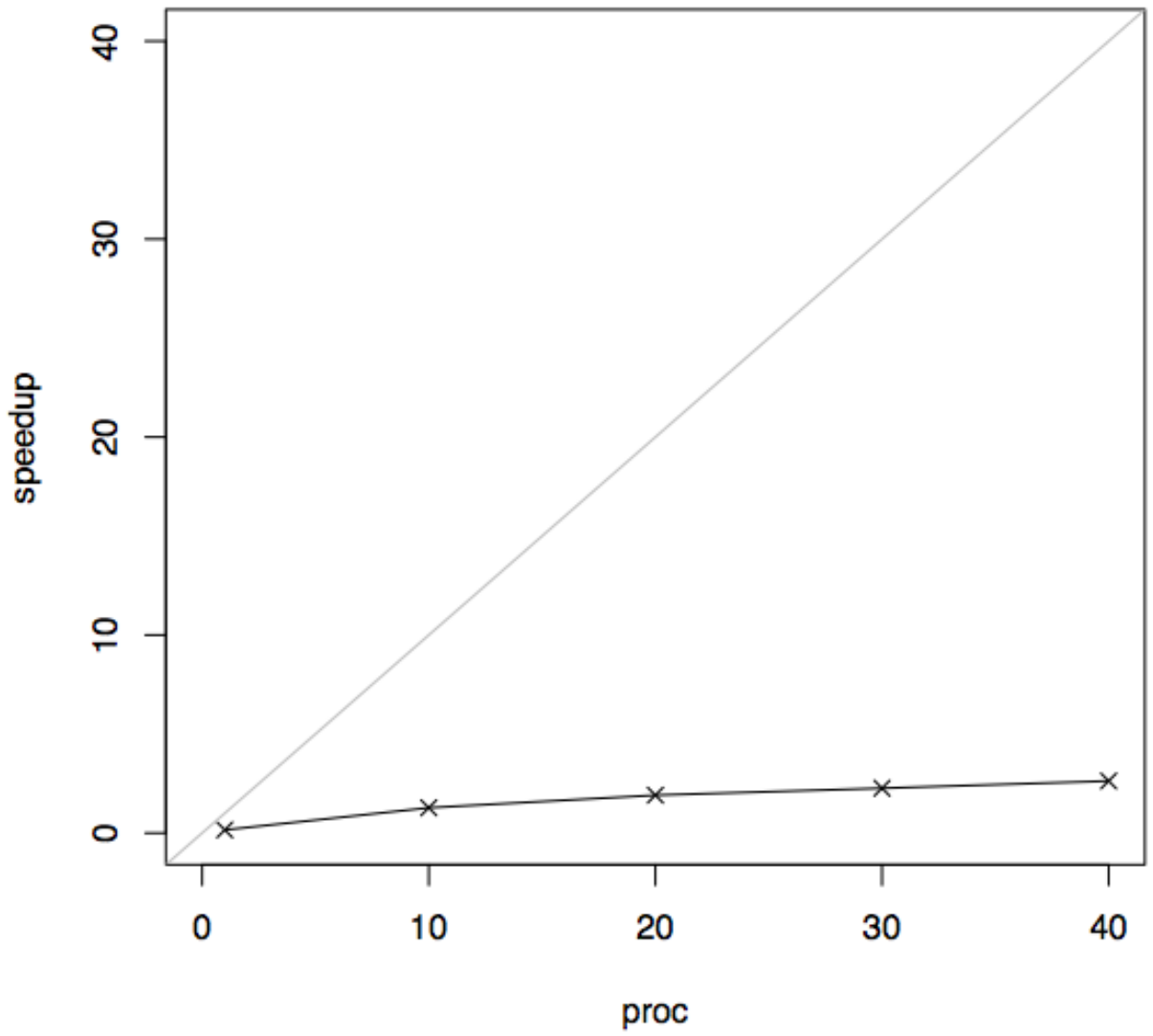


Figure 15: Speedup plot for quicksort with 100000000 elements.

Our analysis suggests that we have a good parallel algorithm for quicksort, yet our observations suggest that, at least on our test machine, our implementation is rather slow relative to our baseline program. In particular, we noticed that our parallel quicksort started out being 6x slower than the baseline algorithm. What could be to blame?

In fact, there are many implementation details that could be to blame. The problem we face is that identifying those causes experimentally could take a lot of time and effort. Fortunately, our quicksort code contains a few clues that will guide us in a good direction.

It should be clear that our quicksort is copying a lot of data and, moreover, that much of the copying could be avoided. The copying operations that could be avoided, in particular, are the array copies that are performed by each of the the three calls to filter and the one call to concat. Each of these operations has to touch each item in the input array.

Let us now consider a (mostly) in-place version of quicksort. This code is mostly in place because the algorithm copies out the input array in the beginning, but otherwise sorts in place on the result array. The code for this algorithm appears just below.

```
sparray in_place_quicksort(const sparray& xs) {
  sparray result = copy(xs);
  long n = xs.size();
```

```
  if (n == 0) {
    return result;
  }
  in_place_quicksort_rec(&result[0], n);
  return result;
}
```

```
controller_type in_place_quicksort_contr("in_place_quicksort");

void in_place_quicksort_rec(value_type* A, long n) {
  if (n < 2) {
    return;
  }
  cstmt(in_place_quicksort_contr, [&] { return nlogn(n); }, [&] {
    value_type p = A[0];
    value_type* L = A;   // below L are less than pivot
    value_type* M = A;   // between L and M are equal to pivot
    value_type* R = A+n-1; // above R are greater than pivot
    while (true) {
      while (! (p < *M)) {
        if (*M < p) std::swap(*M,*(L++));
        if (M >= R) break;
        M++;
      }
      while (p < *R) R--;
      if (M >= R) break;
      std::swap(*M,*R--);
      if (*M < p) std::swap(*M,*(L++));
      M++;
    }
    fork2([&] {
      in_place_quicksort_rec(A, L-A);
    }, [&] {
      in_place_quicksort_rec(M, A+n-M); // Exclude all elts that equal pivot
    });
  }, [&] {
    std::sort(A, A+n);
  });
}
```

We have good reason to believe that this code is, at least, going to be more work efficient than our original solution. First, it avoids the allocation and copying of intermediate arrays. And, second, it performs the partitioning phase in a single pass. There is a catch, however: in order to work mostly in place, our second quicksort code sacrified on parallelism. In specific, observe that the partitioning phase is now sequential. The span of this second quicksort is therefore linear in the size of the input and its average parallelism is therefore logarithmic in the size of the input.

Exercise

Verify that the span of our second quicksort has linear span and that the average parallelism is logarithmic.

So, we expect that the second quicksort is more work efficient but should scale poorly. To test the first hypothesis, let us run the second quicksort on a single processor.

```
$ bench.opt -bench in_place_quicksort -n 100000000 -proc 1
```

Indeed, the running time of this code is essentially same as what we observed for our baseline program.

```
exectime 12.500
total_idle_time 0.000
utilization 1.0000
result 1048575
```

Now, let us see how well the second quicksort scales by performing another speedup experiment.

```
$ prun speedup -baseline "bench.baseline" -parallel "bench.opt -proc 1,20,30,40" -bench  ↩
    quicksort,in_place_quicksort -n 100000000
```

```
[1/10]
bench.baseline -bench quicksort -n 100000000
exectime 12.031
[2/10]
bench.opt -bench quicksort -n 100000000 -proc 1
exectime 68.998
[3/10]
bench.opt -bench quicksort -n 100000000 -proc 20
exectime 5.968
[4/10]
bench.opt -bench quicksort -n 100000000 -proc 30
exectime 5.115
[5/10]
bench.opt -bench quicksort -n 100000000 -proc 40
exectime 4.871
[6/10]
bench.baseline -bench in_place_quicksort -n 100000000
exectime 12.028
[7/10]
bench.opt -bench in_place_quicksort -n 100000000 -proc 1
exectime 12.578
[8/10]
bench.opt -bench in_place_quicksort -n 100000000 -proc 20
exectime 1.731
[9/10]
bench.opt -bench in_place_quicksort -n 100000000 -proc 30
exectime 1.697
[10/10]
bench.opt -bench in_place_quicksort -n 100000000 -proc 40
exectime 1.661
Benchmark successful.
Results written to results.txt.
```

```
$ pplot speedup -series bench
```

The plot below shows one speedup curve for each of our two quicksort implementations. The in-place quicksort is always faster. However, the in-place quicksort starts slowing down a lot at 20 cores and stops after 30 cores. So, we have one solution that is observably not work efficient and one that is, and another that is the opposite. The question now is whether we can find a happy middle ground.

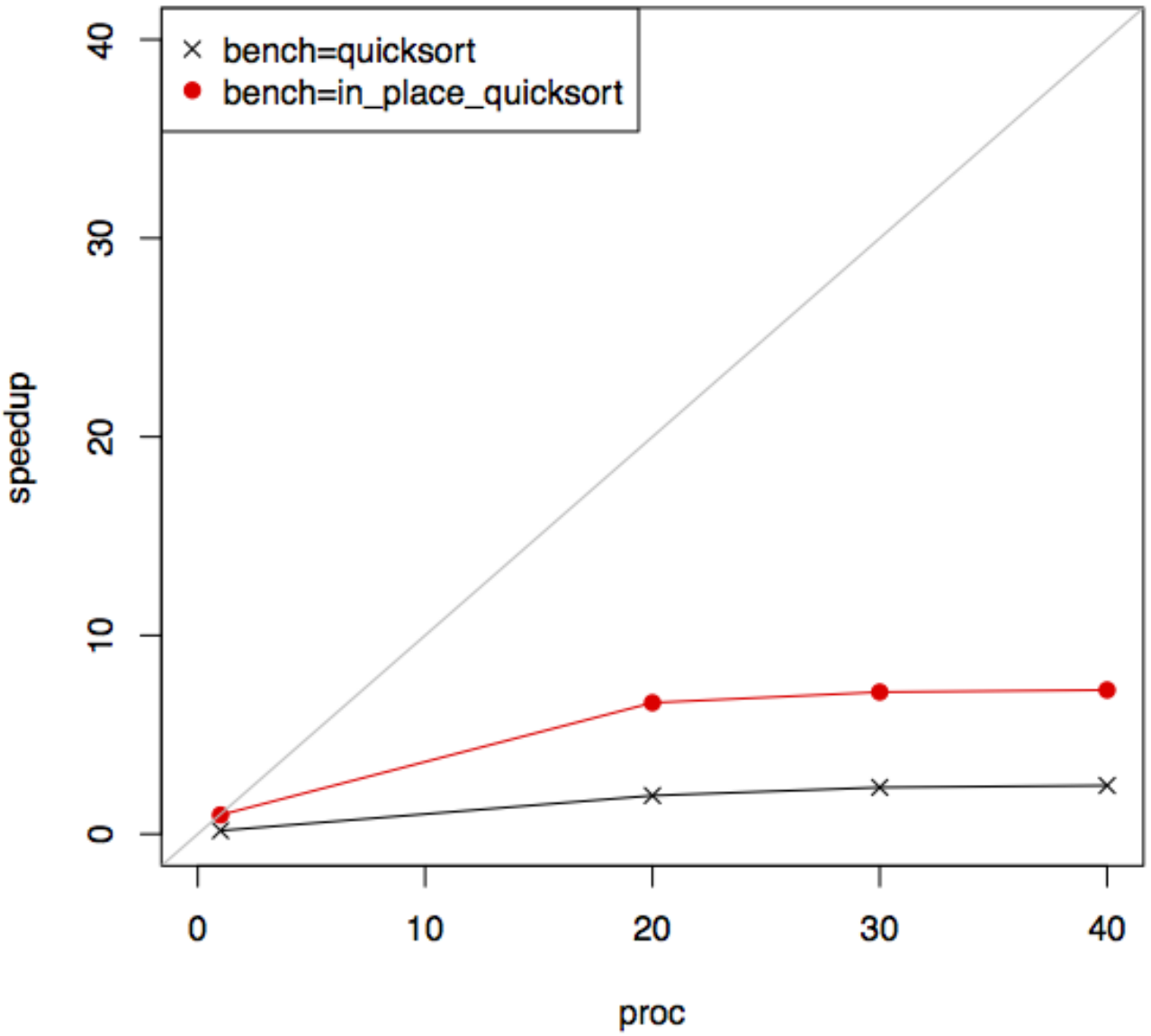


Figure 16: Speedup plot showing our quicksort and the in-place quicksort side by side. As before, we used 100000000 elements.

Question

What can we do to write a better quicksort?

Tip

Eliminate unecessary copying and array allocations.

Tip

Eliminate redundant work by building the partition in one pass instead of three.

Tip
Find a solution that has the same span as our first quicksort code.

We encourage students to look for improvements to quicksort independently. For now, we are going to consider parallel mergesort. This time, we are going to focus more on achieving better speedups.

11.2 Mergesort

As a divide-and-conquer algorithm, the mergesort algorithm, is a good candidate for parallelization, because the two recursive calls for sorting the two halves of the input can be independent. The final merge operation, however, is typically performed sequentially. It turns out to be not too difficult to parallelize the merge operation to obtain good work and span bounds for parallel mergesort. The resulting algorithm turns out to be a good parallel algorithm, delivering asymptotic, and observably work efficiency, as well as low span.

Mergesort algorithm

1. Divide the (unsorted) items in the input array into two equally sized subrange.
2. Recursively and in parallel sort each subrange.
3. Merge the sorted subranges.

This process requires a "merge" routine which merges the contents of two specified subranges of a given array. The merge routine assumes that the two given subarrays are in ascending order. The result is the combined contents of the items of the subranges, in ascending order.

The precise signature of the merge routine appears below and its description follows. In mergesort, every pair of ranges that are merged are adjacent in memory. This observation enables us to write the following function. The function merges two ranges of source array `xs`: `[lo, mid)` and `[mid, hi)`. A temporary array `tmp` is used as scratch space by the merge operation. The function writes the result from the temporary array back into the original range of the source array: `[lo, hi)`.

```
void merge(sparray& xs, sparray& tmp, long lo, long mid, long hi);
```

Example 11.1 Use of merge function

```
sparray xs = {
  // first range: [0, 4)
  5, 10, 13, 14,
  // second range: [4, 9)
  1, 8,  10, 100, 101 };

merge(xs, sparray(xs.size()), (long)0, 4, 9);

std::cout << "xs = " << xs << std::endl;
```

Output:

```
xs = { 1, 5, 8, 10, 10, 13, 14, 100, 101 }
```

To see why sequential merging does not work, let us implement the merge function by using one provided by STL: `std::merge()`. This merge implementation performs linear work and span in the number of items being merged (i.e., `hi -lo`). In our code, we use this STL implementation underneath the `merge()` interface that we described just above.

Now, we can assess our parallel mergesort with a sequential merge, as implemented by the code below. The code uses the traditional divide-and-conquer approach that we have seen several times already.

Question

Is the implementation asymptotically work efficient?

The code is asymptotically work efficient, because nothing significant has changed between this parallel code and the serial code: just erase the parallel annotations and we have a textbook sequential mergesort!

```
sparray mergesort(const sparray& xs) {
  long n = xs.size();
  sparray result = copy(xs);
  mergesort_rec(result, sparray(n), (long)0, n);
  return result;
}

controller_type mergesort_contr("mergesort");

void mergesort_rec(sparray& xs, sparray& tmp, long lo, long hi) {
  long n = hi - lo;
  cstmt(mergesort_contr, [&] { return n * std::log2(n); }, [&] {
    if (n == 0) {
      // nothing to do
    } else if (n == 1) {
      tmp[lo] = xs[lo];
    } else {
      long mid = (lo + hi) / 2;
      fork2([&] {
        mergesort_rec(xs, tmp, lo, mid);
      }, [&] {
        mergesort_rec(xs, tmp, mid, hi);
      });
      merge(xs, tmp, lo, mid, hi);
    }
  }, [&] {
    if (hi-lo < 2)
      return;
    std::sort(&xs[lo], &xs[hi-1]+1);
  });
}
```

Question

How well does our "parallel" mergesort scale to multiple processors, i.e., does it have a low span?

Unfortunately, this implementation has a large span: it is linear, owing to the sequential merge operations after each pair of parallel calls. More precisely, we can write the work and span of this implementation as follows:

$$W(n) = \begin{cases} 1 & \text{if } n \leq 1 \\ W(n/2) + W(n/2) + n \end{cases}$$

$$S(n) = \begin{cases} 1 & \text{if } n \leq 1 \\ \max(W(n/2), W(n/2)) + n \end{cases}$$

EQUATION 11.1: Analyzing work and span of mergesort

It is not difficult to show that these recursive equation solve to $W(n) = \Theta(n \log n)$ and $S(n) = \Theta(n)$.

With these work and span costs, the average parallelism of our solution is $\frac{cn\log n}{2cn} = \frac{\log n}{2}$. Consider the implication: if $n = 2^{30}$, then the average parallelism is $\frac{\log 2^{30}}{2} = 15$. That is terrible, because it means that the greatest speedup we can ever hope to achieve is 15x!

The analysis above suggests that, with sequential merging, our parallel mergesort does not expose ample parallelism. Let us put that prediction to the test. The following experiment considers this algorithm on our 40-processor test machine. We are going to sort a random sequence of 100 million items. The baseline sorting algorithm is the same sequential sorting algorithm that we used for our quicksort experiments: `std::sort()`.

```
$ prun speedup -baseline "bench.baseline" -parallel "bench.opt -proc 1,10,20,30,40" -bench ↩
    mergesort_seqmerge -n 100000000
```

The first two runs suggest that our mergesort has better observable work efficiency than our quicksort. The single-processor run of parallel mergesort is roughly 50% slower than that of the sequential baseline algorithm. Compare that to the 6x-slower running time for single-processor parallel quicksort! We have a good start.

```
[1/6]
bench.baseline -bench mergesort_seqmerge -n 100000000
exectime 12.483
[2/6]
bench.opt -bench mergesort_seqmerge -n 100000000 -proc 1
exectime 19.407
```

The parallel runs are encouraging: we get 5x speedup with 40 processors.

```
[3/6]
bench.opt -bench mergesort_seqmerge -n 100000000 -proc 10
exectime 3.627
[4/6]
bench.opt -bench mergesort_seqmerge -n 100000000 -proc 20
exectime 2.840
[5/6]
bench.opt -bench mergesort_seqmerge -n 100000000 -proc 30
exectime 2.587
[6/6]
bench.opt -bench mergesort_seqmerge -n 100000000 -proc 40
exectime 2.436
```

But we can do better by using a parallel merge instead of a sequential one: the speedup plot in Figure 13 shows three speedup curves, one for each of three mergesort algorithms. The `mergesort()` algorithm is the same mergesort routine that we have seen here, except that we have replaced the sequential merge step by our own parallel merge algorithm. The `cilksort()` algorithm is the carefully optimized algorithm taken from the Cilk benchmark suite. What this plot shows is, first, that the parallel merge significantly improves performance, by at least a factor of two. The second thing we can see is that the optimized Cilk algorithm is just a little faster than the one we presented here. That's pretty good, considering the simplicity of the code that we had to write.

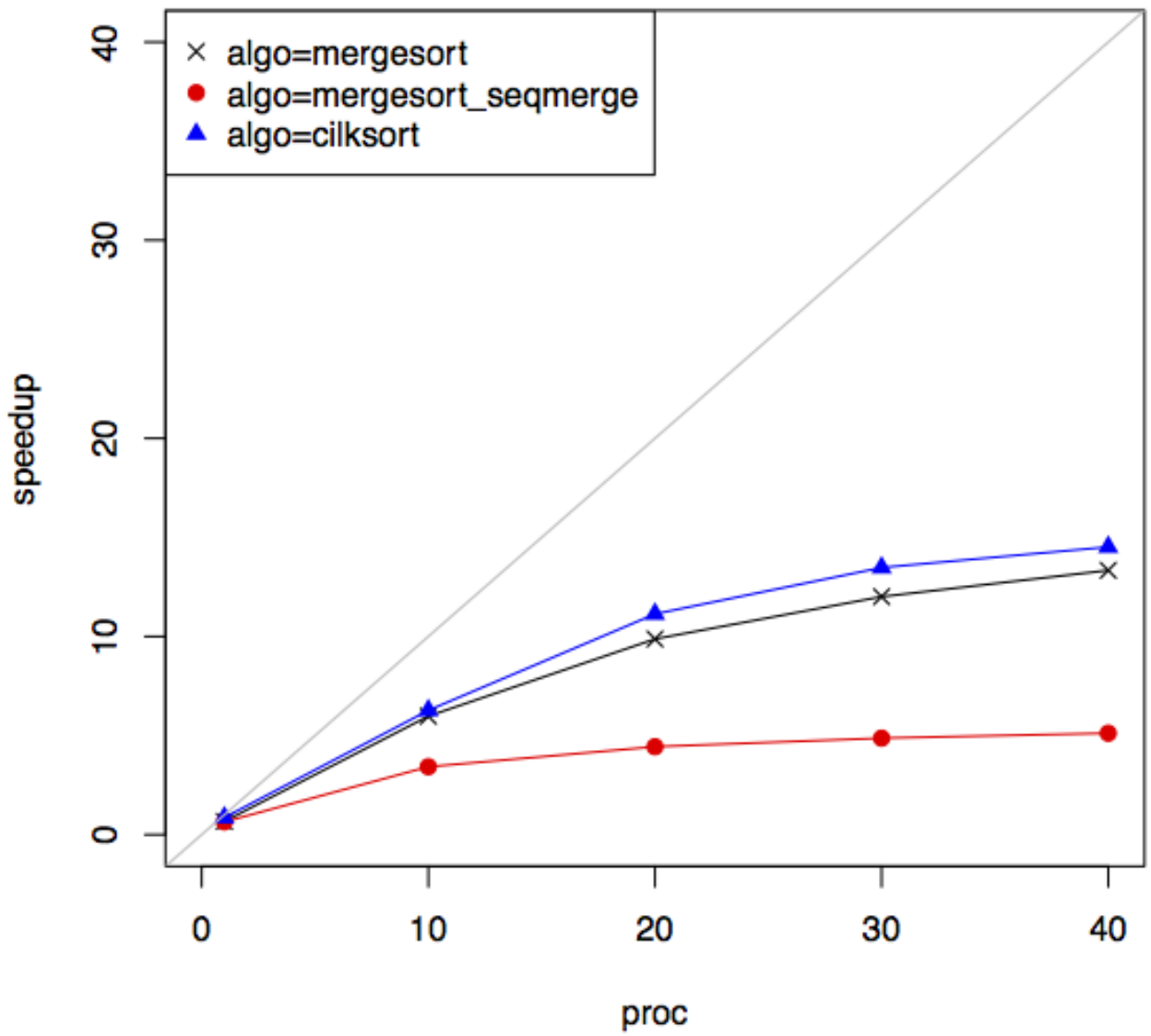


Figure 17: Speedup plot for three different implementations of mergesort using 100 million items.

It turns out that we can do better by simply changing some of the variables in our experiment. The plot shown in Figure 14 shows the speedup plot that we get when we change two variables: the input size and the sizes of the items. In particular, we are selecting a larger number of items, namely 250 million instead of 100 million, in order to increase the amount of parallelism. And, we are selecting a smaller type for the items, namely 32 bits instead of 64 bits per item. The speedups in this new plot get closer to linear, topping out at approximately 20x.

Practically speaking, the mergesort algorithm is memory bound because the amount of memory used by mergesort and the amount of work performed by mergesort are both approximately roughly linear. It is an unfortunate reality of current multicore machines that the main limiting factor for memory-bound algorithms is amount of parallelism that can be achieved by the memory bus. The memory bus in our test machine simply lacks the parallelism needed to match the parallelism of the cores. The effect is clear after just a little experimentation with mergesort. You can see this effect yourself, if you are interested to change in the source code the type aliased by `value_type`. For a sufficiently large input array, you should observe a significant performance improvement by changing just the representation of `value_type` from 64 to 32 bits, owing to the fact that with 32-bit items is a greater amount of computation relative to the number of memory transfers.

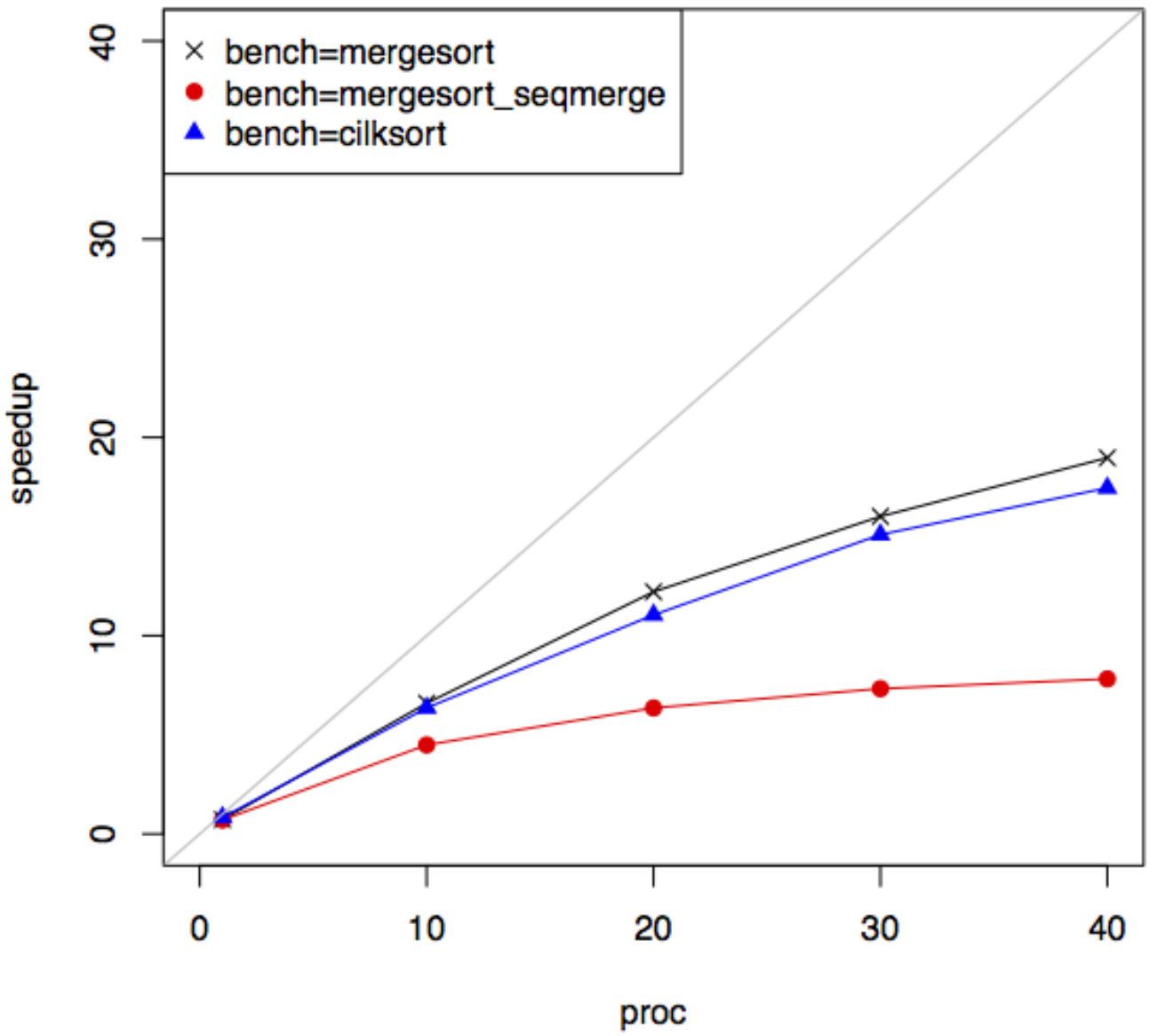


Figure 18: Speedup plot for three different implementations of mergesort using 250 million items.

Question: Stable Mergesort

An important property of the sequential merge-sort algorithm is that it is stable: it can be written in such a way that it preserves the relative order of equal elements in the input. Is the parallel merge-sort algorithm that you designed stable? If not, then can you find a way to make it stable?

12 Chapter: Graph processing

In just the past few years, a great deal of interest has grown for frameworks that can process very large graphs. Interest comes from a diverse collection of fields. To name a few: physicists use graph frameworks to simulate emergent properties from large networks of particles; companies such as Google mine the web for the purpose of web search; social scientists test theories regarding the origins of social trends.

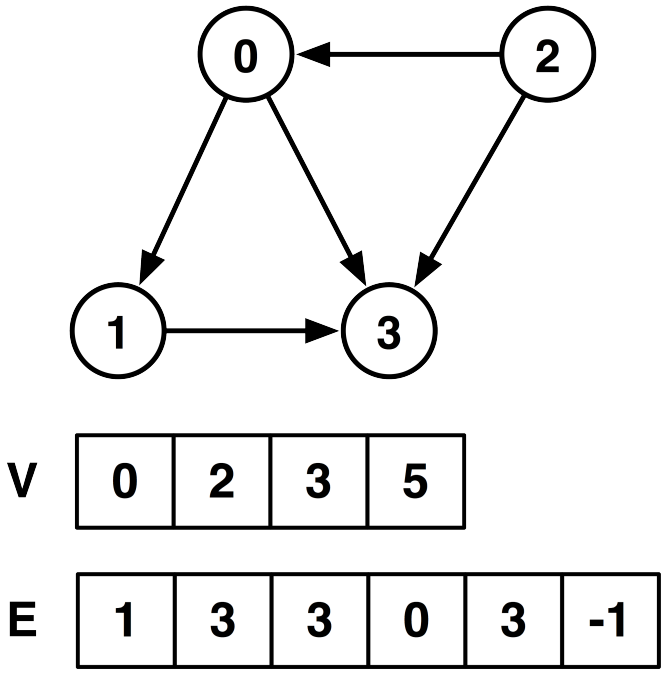
In response, many graph-processing frameworks have been implemented both in academia and in the industry. Such frameworks offer to client programs a particular application programming interface. The purpose of the interface is to give the client programmer a high-level view of the basic operations of graph processing. Internally, at a lower level of abstraction, the framework provides key algorithms to perform basic functions, such as one or more functions that "drive" the traversal of a given graph.

The exact interface and the underlying algorithms vary from one graph-processing framework to another. One commonality among the frameworks is that it is crucial to harness parallelism, because interesting graphs are often huge, making it practically infeasible to perform sequentially interesting computations.

12.1 Graph representation

We will use an adjacency lists representation based on ***compressed arrays*** to represent directed graphs. In this representation, a graph is a collection of unordered neighbor lists. Each vertex in the graph $G = (V, E)$ is assigned an identifier $v \in \{0, \ldots, n-1\}$, where $n = |V|$ is the number of vertices in the graph. For each vertex v in the graph, the adjacency list stores one neighbor list, `out_edges_of[v]`, that describes the set of out-neighbors of `v`. The adjacency lists of the vertices are stored is a single array, which the `out_edges_of` array references as shows below.

A graph (top) and its compressed-array representation (bottom).



The compressed-array representation supports efficiently several operations that are key for parallel graph search. For example, we can determine the out-neighbors af a given vertex with constant work. Similarly, we can determine the out-degree of a given vertex with constant work.

Exercise

Give a constant-work algorithm for computing the out-degree of a vertex.

Another important concern in representing graphs is space use. Space use is a major concern because graphs can have tens of billions of edges or more. The Facebook social network graph (including just the network and no metadata) uses 100 billion edges, for example, and as such could fit snugly into a machine with 2TB of memory. Such a large graph is a greater than the capacity of the RAM available on current personal computers. But it is not that far off, and there are many other interesting graphs that easily fit into just a few gigabytes. Our adjacency list consumes a total of $n+m$ vertex-id cells in memory, where $n = |V|$ and $m = |E|$. For simplicity, we always use 64 bits to represent vertex identifiers but note that a practical library would

support 32 bit representations as well. Although the format that we use is reasonably space efficient for storing large graphs, we should point out that there are other representations that may offer more compact graphs. In fact, graph-compression techniques are an active area of research at present.

We implemented the adjacency-list representation based on compressed arrays with a class called `adjlist`.

```
using vtxid_type = value_type;
using neighbor_list = const value_type*;

class adjlist {
public:
  long get_nb_vertices() const;
  long get_nb_edges() const;
  long get_out_degree_of(vtxid_type v) const;
  neighbor_list get_out_edges_of(vtxid_type v) const;
};
```

Example 12.1 Graph creation

Sometimes it is useful for testing and debugging purposes to create a graph from a handwritten example. For this purpose, we define a type to express an edge. The type is a pair type where the first component of the pair represents the source and the second the destination vertex, respectively.

```
using edge_type = std::pair<vtxid_type, vtxid_type>;
```

In order to create an edge, we use the following function, which takes a source and a destination vertex and returns the corresponding edge.

```
edge_type mk_edge(vtxid_type source, vtxid_type dest) {
  return std::make_pair(source, dest);
}
```

Now, specifying a (small) graph in textual format is as easy as specifying an edge list. Moreover, getting a textual representation of the graph is as easy as printing the graph by `cout`.

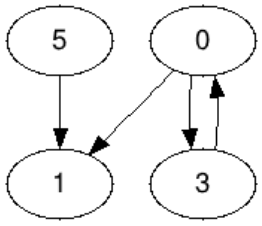
```
adjlist graph = { mk_edge(0, 1), mk_edge(0, 3), mk_edge(5, 1), mk_edge(3, 0) };
std::cout << graph << std::endl;
```

Output:

```
digraph {
0 -> 1;
0 -> 3;
3 -> 0;
5 -> 1;
}
```

Note

The output above is an instance of the "dot" format. This format is used by a well-known graph-visualization tool called graphviz. The diagram below shows the visualization of our example graph that is output by the graphviz tool. You can easily generate such visualizations for your graphs by using online tools, such is Click this one.



Example 12.2 Adjacency-list interface

```
adjlist graph = { mk_edge(0, 1), mk_edge(0, 3), mk_edge(5, 1), mk_edge(3, 0),
                  mk_edge(3, 5), mk_edge(3, 2), mk_edge(5, 3) };
std::cout << "nb_vertices = " << graph.get_nb_vertices() << std::endl;
std::cout << "nb_edges = " << graph.get_nb_edges() << std::endl;
std::cout << "neighbors of vertex 3:" << std::endl;
neighbor_list neighbors_of_3 = graph.get_out_edges_of(3);
for (long i = 0; i < graph.get_out_degree_of(3); i++)
  std::cout << " " << neighbors_of_3[i];
std::cout << std::endl;
```

Output:

```
nb_vertices = 6
nb_edges = 7
neighbors of vertex 3:
 0 5 2
```

Next, we are going to study a version of breadth-first search that is useful for searching in large in-memory graphs in parallel. After seeing the basic pattern of BFS, we are going to generalize a little to consider general-purpose graph-traversal techniques that are useful for implementing a large class of parallel graph algorithms.

12.2 Breadth-first search

The breadth-first algorithm is a particular graph-search algorithm that can be applied to solve a variety of problems such as finding all the vertices reachable from a given vertex, finding if an undirected graph is connected, finding (in an unweighted graph) the shortest path from a given vertex to all other vertices, determining if a graph is bipartite, bounding the diameter of an undirected graph, partitioning graphs, and as a subroutine for finding the maximum flow in a flow network (using Ford-Fulkerson's algorithm). As with the other graph searches, BFS can be applied to both directed and undirected graphs.

The idea of ***breadth first search***, or ***BFS*** for short, is to start at a *source* vertex s and explore the graph outward in all directions level by level, first visiting all vertices that are the (out-)neighbors of s (i.e. have distance 1 from s), then vertices that have distance two from s, then distance three, etc. More precisely, suppose that we are given a graph G and a source s. We define the ***level*** of a vertex v as the shortest distance from s to v, that is the number of edges on the shortest path connecting s to v.

12.2.1 Sequential BFS

Many variations of BFS have been proposed over the years. The one that may be most widely known is the classic sequential BFS that uses a FIFO queue to buffer vertices that are waiting to be visited. The FIFO-based approach is a poor approach for parallelization because accesses to the FIFO queue are by definition serialized.

12.2.2 Parallel BFS

Our goal is to design and implement a parallel algorithm for BFS that is observably work efficient and has plenty of parallelism. To this end, let's consider an algorithm that traverses the graph in level order. We can implement such a level-order traversal by maintaining a ***frontier*** as a set of vertices that have not yet been visited but will be visited next, and visiting the vertices in the frontier (which represents all the vertices is a level) all together in parallel. The pseudo-code for this is shown below.

Set-based pseudocode for parallel BFS

```
frontier = { source }
visited = {}
while frontier not empty
  start level
    next = {}
    foreach vertex v in current frontier
      visit v
```

```
    visited = visited set-union frontier

    foreach v in frontier
      next = next set-union (neighbors of v)
    frontier = next set-difference visited
  end level
```

Exercise

Convince yourself that this algorithm does indeed perform a BFS by performing a level-by-level traversal.

Assuming that we have a parallel set data structure, let us parallelize this algorithm. First, note that we can visit all the vertices in the frontier in parallel. That is, we can parallelize the first `foreach` loop. Second, we can also compute the next set (frontier) in parallel by performing a reduce with the set-union operation, and then by taking a set-difference operation.

Our goal is to implement an observably work-efficient version of this algorithm on a hardware-shared memory parallel machine such as a modern multicore computer. The key challenge in doing so will be the elimination of the set operations performed for computing the next frontier. Apart from maintaining a visited set to prevent a vertex from being visited more than once, the serial algorithm does not have to perform these operations.

We will use atomic read-modify-write operations to achieve observable work efficiency and competitiveness with the serial BFS. Specifically. we shall use the compare-and-swap operation. This will also allow us to represent the frontier set with arrays. To achieve observable work efficiency, we will change the notion of the frontier slightly. Instead of holding the vertices that we are will visit next, the frontier will hold the vertices we just visited. At each level, we will visit the neighbors of the vertices in the frontier, but only if they have not yet been visited. This guard is necessary, because two vertices in the frontier can both have a vertex as their neighbor. In fact, without such a guard, the algorithm can may fail to terminate if the graph has a cycle. After we visited all the neighbors of the vertices in the frontier at this level, we assign the frontier to be the vertices visited. The pseudocode for this algorithm is shown below.

Pseudocode for parallel BFS

```
visit source
frontier = { source }
while frontier not empty
  start level
    foreach vertex v in current frontier
      foreach neighbor u of vertex
        if u is not visited
          visit u
    frontier = vertices visited at this level
   end level
```

Let's turn our attention to parallelism. From the pseudocode, we see that there are at least two clear opportunities for parallelism. The first is the foreach loop that processes the frontier and the second the foreach loop that processes the neighbors of the vertex that is currently being visited. These two loops should expose a lot of parallelism, at least for certain classes of graphs. The outer loop exposes a lot of parallelism when the frontier gets to be large. The inner loop exposes a lot of parallelism when the traversal reaches a vertex that has a high out degree.

Parallelizing the two foreach loops requires some extra care, because parallelizing in a naive fashion would enable a race condition. To see why, we need to consider how an implementation of BFS keeps track of which vertices have been visited already and which have not. Suppose that we use an array of booleans `visited[v]` of size n that is keyed by the vertex identifier. If `visited[v] ==true`, then vertex `v` has been visited already and has not otherwise. Suppose now that two processors, namely A and B, concurrently attempt to gain access to the same vertex `v` (via two different neighbors of `v`). If A and B both read `visited[v]` at the same time, then both consider that they have gained access to `v`. Both processors then mark `v` as visited and then proceed to visit the neighbors of `v`. As such, `v` will be visited twice and subsequently have its outgoing neighbors processed twice.

Consider now the implication: owing to the race condition, such an implementation of parallel BFS cannot in general guarantee that each reachable vertex is visited once and only once. Of course, the race does not necessarily happen every time a vertex is visited. But even if it happens only rarely, there is a real chance that a huge, in fact unbounded, amount of redundant work is

performed by the BFS: when the same vertex is visited twice, all of its neighbors are processed twice. The amount of redundant work that is performed is high when the outdegree of the vertex is high. In other words, a racy parallel BFS is not even an asymptotically work-efficient BFS due to the unbounded amount of redundant work that it could perform in any given round.

Exercise

Clearly, the race conditions on the visited array that we described above can cause BFS to visit any given vertex twice.

- Could such race conditions cause the BFS to visit some vertex that is not reachable? Why or why not?
- Could such race conditions cause the BFS to not visit some vertex that is reachable? Why or why not?
- Could such race conditions trigger infinite loops? Why or why not?

The issues relating to the race conditon leads us to consider lightweight atomic memory. We can use lightweight atomic memory, as described in this chapter to both eliminate race conditions and avoid having to sacrifice a lot of performance. The basic idea is to guard each cell in our "visited" array by an atomic type.

Example 12.3 Accessing the contents of atomic memory cells

Access to the contents of any given cell is achieved by the `load()` and `store()` methods.

```
const long n = 3;
std::atomic<bool> visited[n];
long v = 2;
visited[v].store(false);
std::cout << visited[v].load() << std::endl;
visited[v].store(true);
std::cout << visited[v].load() << std::endl;
```

Output:

```
0
1
```

The key operation that enables us to eliminate the race condition is the **compare and exchange** operation. This operation performs the following steps, atomically:

1. Read the contents of the target cell in the visited array.
2. If the contents is false (i.e., equals the contents of `orig`), then write `true` into the cell and return `true`.
3. Otherwise, just return `false`.

```
const long n = 3;
std::atomic<bool> visited[n];
long v = 2;
visited[v].store(false);
bool orig = false;
bool was_successful = visited[v].compare_exchange_strong(orig, true);
std::cout << "was_successful = " << was_successful << "; visited[v] = " << visited[v].load ↩
    () << std::endl;
bool orig2 = false;
bool was_successful2 = visited[v].compare_exchange_strong(orig2, true);
std::cout << "was_successful2 = " << was_successful2 << "; visited[v] = " << visited[v]. ↩
    load() << std::endl;
```

Output:

```
was_successful = 1; visited[v] = 1
was_successful2 = 0; visited[v] = 1
```

12.3 Implementing parallel BFS

So far, we have seen pseudocode that describes at a high level the idea behind the parallel BFS. We have seen that special care is required to eliminate problematic race conditions. Let's now put these ideas together to complete and implementation. The following function signature is the signature for our parallel BFS implementation. The function takes as parameters a graph and the identifier of a source vertex and returns an array of boolean flags.

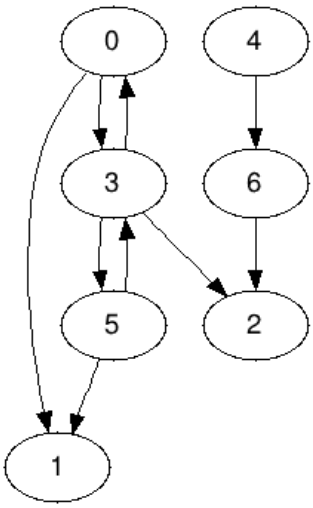
```
sparray bfs(const adjlist& graph, vtxid_type source);
```

The flags array is a length $|V|$ array that specifies the set of vertices in the graph which are reachable from the source vertex: a vertex with identifier `v` is reachable from the given source vertex if and only if there is a `true` value in the v^{th} position of the flags array that is returned by `bfs`.

Example 12.4 Parallel BFS

```
adjlist graph = { mk_edge(0, 1), mk_edge(0, 3), mk_edge(5, 1), mk_edge(3, 0),
                  mk_edge(3, 5), mk_edge(3, 2), mk_edge(5, 3),
                  mk_edge(4, 6), mk_edge(6, 2) };
std::cout << graph << std::endl;
sparray reachable_from_0 = bfs(graph, 0);
std::cout << "reachable from 0: " << reachable_from_0 << std::endl;
sparray reachable_from_4 = bfs(graph, 4);
std::cout << "reachable from 4: " << reachable_from_4 << std::endl;
```

The following diagram shows the structure represented by `graph`.



Output:

```
digraph {
0 -> 1;
0 -> 3;
3 -> 0;
3 -> 5;
3 -> 2;
4 -> 6;
5 -> 1;
5 -> 3;
6 -> 2;
}
reachable from 0: { 1, 1, 1, 1, 0, 1, 0 }
reachable from 4: { 0, 0, 1, 0, 1, 0, 1 }
```

To complete our implementation, let's assume that we have a function called `edge_map` with the following signature the "edge map" operation. This operation takes as parameters a graph, an array of atomic flag values, and a frontier and returns a new frontier.

```
sparray edge_map(const adjlist& graph, std::atomic<bool>* visited, const sparray&  ↩
    in_frontier);
```

The main loop of BFS is shown below. The algorithm uses the edge-map function to advance level by level through the graph. The traversal stops when the frontier is empty.

```
loop_controller_type bfs_init_contr("bfs_init");

sparray bfs(const adjlist& graph, vtxid_type source) {
  long n = graph.get_nb_vertices();
  std::atomic<bool>* visited = my_malloc<std::atomic<bool>>(n);
  parallel_for(bfs_init_contr, 0l, n, [&] (long i) {
    visited[i].store(false);
  });
  visited[source].store(true);
  sparray cur_frontier = { source };
  while (cur_frontier.size() > 0)
    cur_frontier = edge_map(graph, visited, cur_frontier);
  sparray result = tabulate([&] (value_type i) { return visited[i].load(); }, n);
  free(visited);
  return result;
}
```

One minor technical complication relates to the result value: our algorithm performs extra work to copy out the values from the visited array. Although it could be avoided, we choose to copy out the values because it is more convenient for us to program with ordinary `sparray`'s. Here is an example describing the behavior of the `edge_map` function.

Example 12.5 A run of `edge_map`

```
adjlist graph = // same graph as shown in the previous example
const long n = graph.get_nb_vertices();
std::atomic<bool> visited[n];
for (long i = 0; i < n; i++)
  visited[i] = false;
visited[0].store(true);
visited[1].store(true);
visited[3].store(true);
sparray in_frontier = { 3 };
sparray out_frontier = edge_map(graph, visited, in_frontier);
std::cout << out_frontier << std::endl;
sparray out_frontier2 = edge_map(graph, visited, out_frontier);
std::cout << out_frontier2 << std::endl;
```

Output:

```
{ 5, 2 }
{  }
```

From the perspective of BFS, the edge-map function is the function that advances one level ahead in the level-by-level traversal of the graph. More concretely, this function takes as argument the frontier at level i in the BFS traversal and returns the frontier at level $i+1$.

To implement `edge_map`, we shall use the following sentinel value to represent empty cells in sparse arrays of vertex identifiers.

```
const vtxid_type not_a_vertexid = -1l;
```

Example 12.6 Sparse-array representation of a set of vertex identifiers

The following array represents a set of three valid vertex identifiers, with two positions in the array being empty.

```
{ 3, not_a_vertexid, 0, 1, not_a_vertexid }
```

Let us define two helper functions. The first one takes a sparse array of vertex identifiers and copies out the valid vertex identifiers.

```
sparray just_vertexids(const sparray& vs) {
  return filter([&] (vtxid_type v) { return v != not_a_vertexid; }, vs);
}
```

The other function takes a graph and an array of vertex identifiers and returns the array of the degrees of the vertex identifiers.

```
sparray get_out_degrees_of(const adjlist& graph, const sparray& vs) {
  return map([&] (vtxid_type v) { return graph.get_out_degree_of(v); }, vs);
}
```

At a high level, our solution is the following. First, we construct a sparse-array representation of the set of vertex ids that are to be returned by the edge map. Second, we construct a compact representation of the sparse the array and return the compacted result array. Aside from this sparse-array detail, the algorithm implemented here corresponds exactly to the high level description that we presented in the previous section. That is, the outer loop processes the vertex ids from the frontier of the previous level and the inner loop processes the neighbors of each vertex in the frontier of the previous level, adding newly visited neighbors to the result set.

```
loop_controller_type process_out_edges_contr("process_out_edges");
loop_controller_type edge_map_contr("edge_map");

sparray edge_map(const adjlist& graph, std::atomic<bool>* visited, const sparray&  ↩
    in_frontier) {
  // temporarily removed.
}
```

The complexity function used by the outer loop in the edge map is interesting because the complexity function treats the vertices in the frontier as weighted items. In particular, each vertex is weighted by its out degree in the graph. The reason that we use such weighting is because the amount of work involved in processing that vertex is proportional to its out degree. We cannot treat the out degree as a constant, unfortunately, because the out degree of any given vertex is unbounded, in general. As such, it should be clear why we need to account for the out degrees explicitly in the complexity function of the outer loop.

Question

What changes you need to make to BFS to have BFS annotate each vertex v by the length of the shortest path between v and the source vertex?

12.3.1 Performance analysis

Our parallel BFS is asymptotically work efficient: the BFS takes work $O(n+m)$. To establish this bound, we need to assume that the compare-and-exchange operation takes constant time. After that, confirming the bound is only a matter of inspecting the code line by line. On the other hand, the span is more interesting.

Question

What is the span of our parallel BFS?

Tip

In order to answer this question, we need to know first about the graph **diameter**. The diameter of a graph is the length of the shortest path between the two most distant vertices. It should be clear that the number of iterations performed by the while loop of the BFS is at most the same as the diameter.

Exercise

By using sentinel values, it might be possible to implement BFS to eliminate the compaction used by `edge_map`. Describe and implement such an algorithm. Does in perform better?

13 Chapter: Work Stealing in Dedicated Environments

I AM TRYING TO SIMPLIFY AND STREAMLINE THE PROOF SO THAT WE CAN DEAL WITH EDGE WEIGHTS.

We will present a randomized work stealing algorithm and analyze its complexity.

We consider multithreaded computations represented as dags as described in Chapter Multithreading. To streamline the analysis, we assume without loss of generality that the root vertex has a single child.

13.1 Offline and online scheduling

INSERT LOWER BOUNDS

INSERT UPPER BOUND FOR BRENT

INSERT UPPER BOUND GREEDY

13.1.1 Online scheduling

Offline scheduling problem discussed above shows that any greedy scheduler has within factor two of the optimal, based on the lower bounds established above. We now turn our attention to the problem of constructing such an execution schedule online. In this ***online scheduling problem***, we are given a P-processor kernel schedule and a computation dag, and we are interested in constructing an execution schedule with minimal length. As we shall describe next, the non-blocking work-stealing algorithm achieves this goal, also when including the cost of scheduling itself.

13.2 Work-Stealing Algorithm

In work stealing, each process maintains a ***deque***, doubly ended queue, of vertices. Each process tries to work on its local deque as much as possible. Execution starts with the root vertex in the deque of one of the processes. It ends when the final vertex is executed.

A work stealing scheduler operates as described by our generic scheduling algorithm but instead of operating on threads, it operates on vertices of the dag.

To obtain work, a process pops the vertex at the bottom of its deque and executes it. We refer to the vertex executed by a process as the ***assigned vertex***. When executed, the ready vertex can make the other vertices ready, which are then pushed onto the bottom end of the deque in an arbitrary order. If a process finds its deque empty, then the process becomes a ***thief***. A thief picks a ***victim*** process at random and attempts to steals a thread from another it by popping a thread off the top of the victim's deque.

Such a ***steal attempt*** can fail if

1. the victim's deque is empty, or
2. contention between processors occurs and the vertex targeted by the thief is executed by the process that own the dequer or stolen by another thief.

The thief performs steal attempts until it successfully steals a thread, at which point, the thief goes back to work and the stolen thread becomes its assigned thread.

The pseudo-code for the algorithm is shown below. The algorithm operates is round. In each round, a process executes the assigned vertex if any, pushes the newly enabled vertices to its deque, and obtains a new assigned vertex from its deque. If the

round starts with no assigned vertex then the process becomes a thief performs a steal attempt. For simplicity, we refer to a steal attempt as a ***throw***. Noe that a steal attempt start and completes in the same round.

For the analysis of the algorithm, we shall assume that each instruction and each deque operations executes in a single step to execute. As a result, each iteration of the loop, a round, completes in constant steps.

```
// Assign root to process zero.
assignedVertex = NULL
if (self == ProcessZero) {
  assignedVertex = rootVertex
}


// Run scheduling loop.
while (computationDone == false) {

  // Execute assigned vertex.
  if (assignedVertex <> NULL) {
    (nChildren, child1, child2) = execute (assignedVertex)

    if (nChildren == 1) {
      self.pushBottom child1
    }
    else {
      self.pushBottom child1
      self.pushBottom child2
    }
    assignedVertex = self.popBottom ()
  }
  else {
    // Make steal attempt.
    victim = randomProcess ()
    assignedVertex = victim.popTop ()
  }
}
```

13.2.1 Deque Specification

The deque supports three methods:

1. `pushBottom`, which pushed a vertex at the bottom end of the deque.
2. `popBottom`, which returns the vertex at the bottom end of the deque if any, or returns `NULL` otherwise.
3. `popTop`, returns the vertex at the top end of the deque, if any, or returns `NULL` if the deque is empty.

These operations take place atomically. We can think of them start by first obtaining a lock for the deque and then performing the desired operation.

For the analysis, we shall assume that all these operations take constant time and in fact complete in one step. This assumption is somewhat unrealistic, because it is not known whether `popTop` can be implemented in constant time. But a relaxed version of `popTop`, which allows `popTop` to return `NULL` if another concurrent operation removes the top vertex in the deque, accepts a constant-time implementation. This relaxed version suffices for our purposes.

13.2.2 Work sequence of a process

Consider the execution of the work stealing algorithm and let q be any process. We define the ***work sequence*** of q as the sequence of vertices defined by the assigned vertex of q followed by the vertices in its deque ordered from bottom to top. If a vertex is in the work sequence of a process, then we say that it ***belongs*** to that process.

For example, if a process has just completed the execution of its assigned vertex but has not pushed the children enabled onto its deque, then the work sequence consist of the vertices in the deque in the bottom to top order.

13.2.3 Enabling Tree and Weights

Let's recall first the notion of an enabling three. If execution of u enables v, then we call the edge (u,v) an ***enabling edge*** and call u the ***parent*** of v. Every vertex except for the root has a parent. Therefore the subgraph of the dag consisting of the enabling edges form a rooted tree, called the ***enabling tree***. Note each execution can have a different enabling tree.

If $d(u)$ is the depth of a node u in the enabling tree, then we define the weight of u, written $w(u) = S - d(u)$. The root has weight S. Intuitively, the weight is equal to the distance of a vertex to the completion.

13.2.4 Structural Lemma

Lemma[Structural Lemma]

Consider any time in an execution of the work-stealing algorithm after the execution of the root vertex. Let $v_0, v_1, \ldots, v_k$ denote the work sequence of a process. Let $u_0, u_1, \ldots, u_k$ be the sequence consisting of the parents of the vertices in the working sequence in the same order. Then u_i is an ancestor of u_{i-1} in the enabling tree. Moreover, we may have $u_0 = u_1$ but for the ancestor relationship is proper.

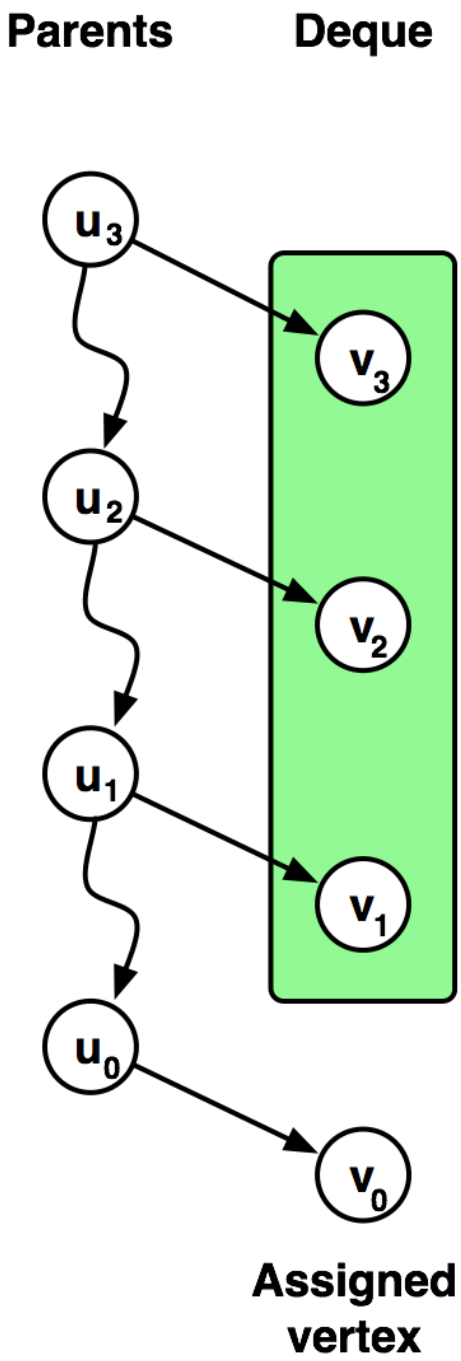
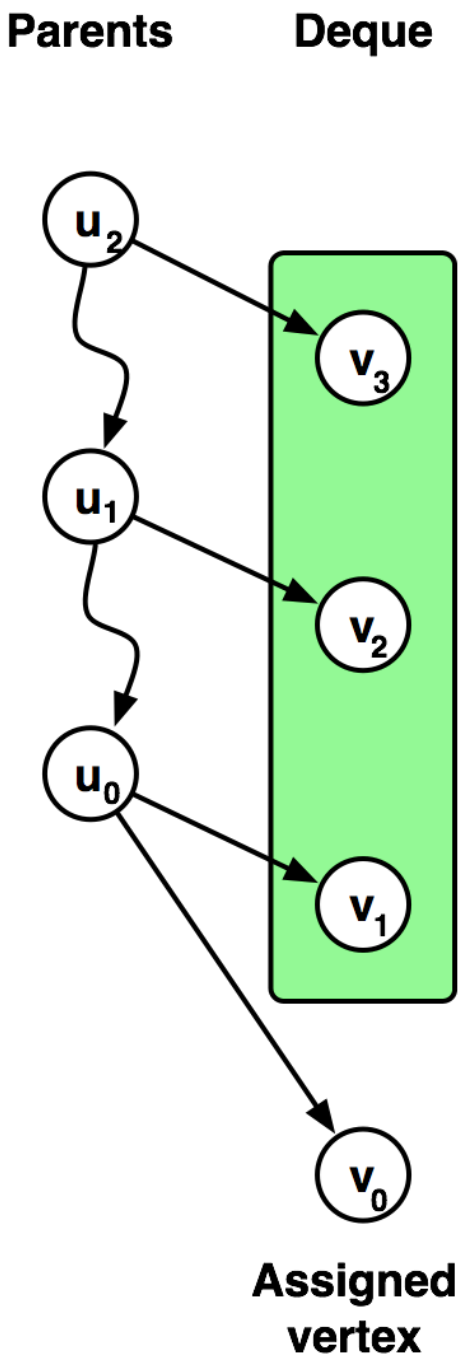


Figure 19: Structural lemma illustrated.



Before we prove the lemma, we state an important corollary.

Corrollary

Let $v_0, v_1, v_2 \dots v_k$ defined as in the structural lemma above. Then, we have $w(v_0) \leq w(v_1) < w(v_2) \dots w(k-1) < w(v_k)$ defined as in the structural lemma above.

The proof is by induction on the number of rounds.

At the initialization and before the beginning of the first round, all deques are empty, root is assigned to a process but has not been executed. The root vertex is then executed and enables a single vertex, this vertex is pushed onto the deque and popped again becoming the assigned vertex at the beginning of the second round. At any of these points in time after the execution of the root, process zero's work sequence consist of the child of the root, v. The parent of v is root and the lemma holds trivially.

For the inductive case, assume that the lemma holds up to beginning of some later round. We will show that it holds at any point during the round and also after the completion of the round.

Consider any process and its deque. We have two cases to consider.

Case 1: There is an assigned node, v_0, which is executed.

By the definition of work sequences, we know that $v_1, \ldots, v_k$ are the vertices in the deque. Let $u_1, \ldots, u_k$ be their parents. By induction, we know that u_i is an ancestor of u_{i-1} in the enabling tree and the ancestor relationship is proper except for $i = 1$, where it is possible that $u_0 = u_1$. Immediately after the execution of the assigned node, the work sequence of the process consists of all the vertices in the deque and the lemma follows easily.

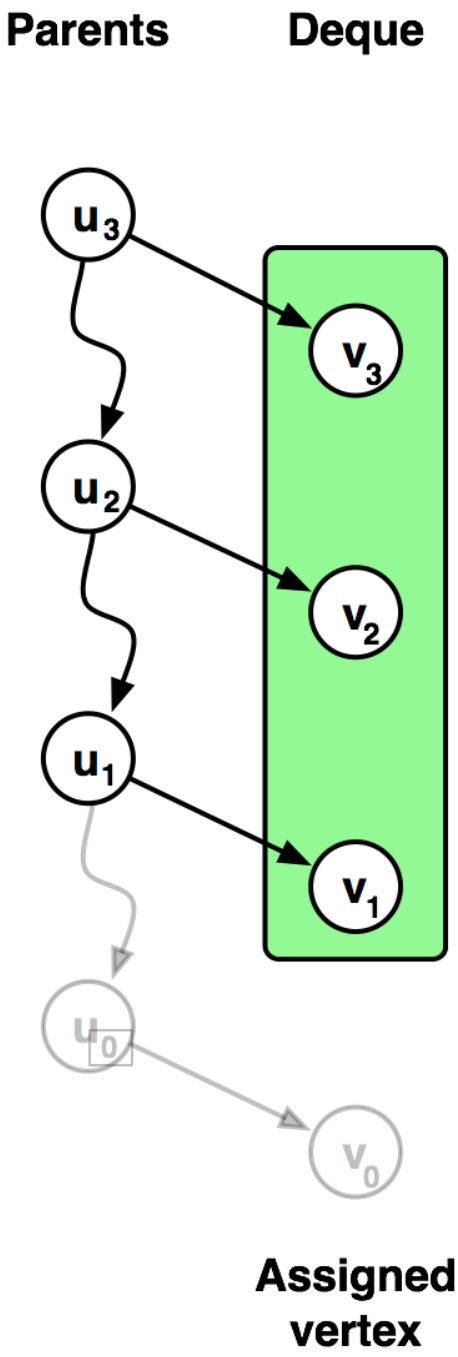


Figure 20: Structural lemma illustrated after the assigned vertex is executed.

After the execution of the assigned vertex v_0, we have several sub-cases to consider.

Case 1.1: execution of v_0 enables no children.

Since the deque remains the same, the lemma holds trivially.

Case 1.2: execution of v_0 enables one child x, which is pushed to the bottom of the deque. In this case, v_0 becomes the parent of x. The lemma holds.

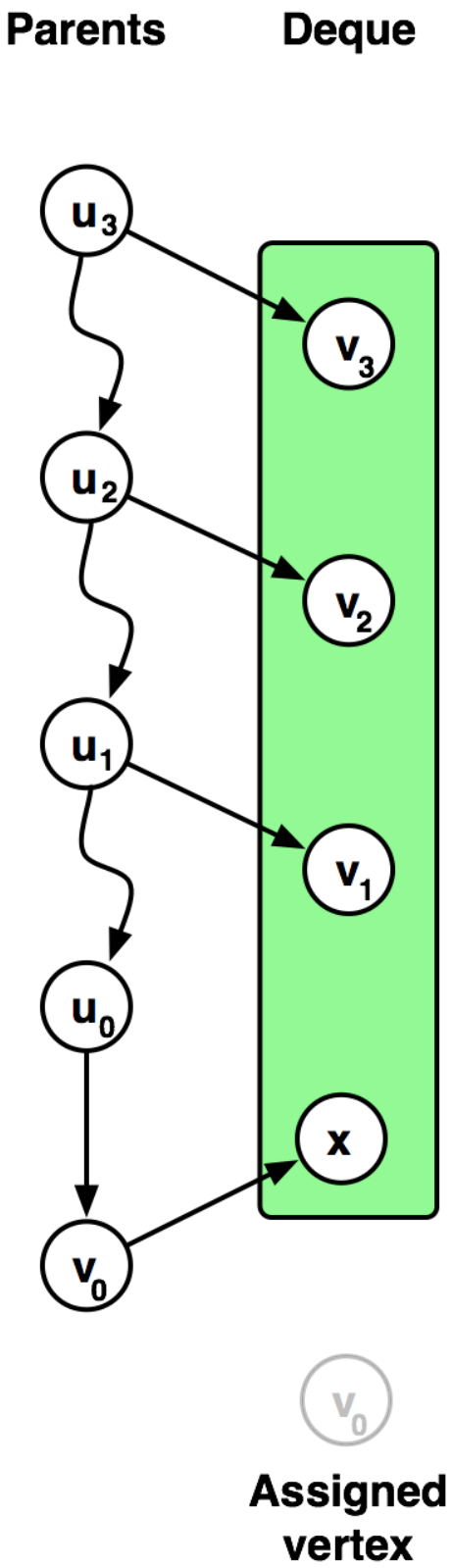


Figure 21: Structural lemma illustrated after the assigned vertex enables one child.

Case 1.2: execution of v_0 enables two children x, y, which are pushed to the bottom of the deque in an arbitrary order.

In this case, v_0 becomes the parent of x and y. The lemma holds.

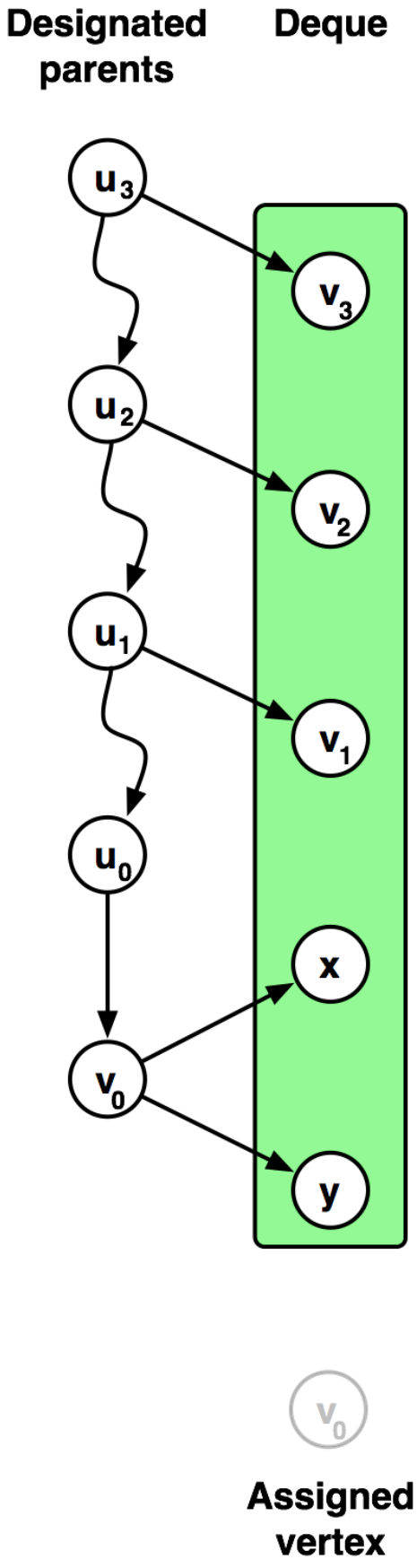


Figure 22: Structural lemma illustrated after the assigned vertex enables two children.

After the execution of the assigned vertex completes and the children are pushed, the process pops the vertex at the bottom of the deque. There are two cases to consider.

1. If the deque is empty, then the process finds no vertex in its deque and there is no assigned vertex at the end of the round, thus the lemma holds trivially.
2. If the deque is not empty, then the vertex at the bottom of the deque becomes the new assigned vertex. The lemma holds trivially because making the bottom vertex the assigned vertex has no impact on the work sequence of the process and thus the correctness of the lemma.

Case 2: A successful steal takes place and removes the top vertex in the deque. In this case, the victim process loses its top vertex, which becomes the assigned vertex of the thief. The lemma holds.

13.3 Analysis

13.3.1 Balls and Bins Game

One crucial fact behind the analysis is a probabilistic lemma.

Lemma[Balls and Bins]

Suppose that P balls are thrown uniformly and randomly into P bins, where bin $1 \leq i \leq P$ has weight W_i, and $W = \sum_{i=1}^{P} W_i$. For each bin, define the random variable $X_i = \begin{cases} W_i & \text{if a ball lands in bin } i \\ 0 & \text{otherwise} \end{cases}$ If $X = \sum_{i=1}^{P} X_i$, then for any $\beta, 0 < \beta < 1$, we have $P[X \geq \beta W] > 1 - \frac{1}{(1-\beta)e}$.

Example 13.1 An application of the lemma

Let's calculate the probability for $\beta = 1/2$. By the lemma, we know that if P balls are thrown into P bins, then the probability that the total weight of the bins that have a ball in them is at least half the total weight is $P[X \geq \frac{W}{2}]1 - \frac{1}{0.5e}$. Since $e > 2.71$, this quantity can be canculated as at least 0.25. We thus conclude that we using the Ball and Bins lemma, we can "collect" at least half of the weight with probability at least 0.25

The Ball and Bins lemma proves something relatively intuitive. If you throw as many ball as there are bins, changes are non-trivial that you will have a ball in at least a constant fraction of the bins, because chances of all balls landing in a small number of bins is law.

Proof

The proof of the lemma is a relatively straightforward application of Markov's inequality. Consider the random variable $W_i - X_i$. This random variable takes on the value 0 if a ball lands in bin i and W_i otherwise. Thus,

$E[W_i - X_i] = W_i \cdot (1 - 1/P)^P \leq W_i/e$.

It follows that $E[W - X] \leq W/e$.

By Markov's inequality we have

$P[W - X > (1-\beta)W] \leq \frac{E[W-X]}{(1-\beta)W}$, and thus $P[X < \beta W] < \frac{1}{(1-\beta)e}$.

13.3.2 Bound in terms of Throws

Lemma[Work-Throw Bound]

Consider any multithreaded computation with work W. The P-processor execution time is $O(W/P + S/P)$ steps where S is the number of throws.

IMPORTANT The proof assumes that each instructions including deque operations takes one step by assuming that each round contributes to the work or to the steal bucket. If this assumption is not valid, then we might need to change the notion of rounds. This is what ABP does.

Proof

We shall consider the execution in terms of rounds and collect a token from each processor in each round. Since rounds are a constant number of steps, we will thus collect as many tokes as total number of steps divided by some constant.

Consider the execution in terms of rounds

1. If a vertex is executed in that round, then the processor places a token into the work bucket.
2. If a steal attempt takes place in that round, then a throw occurs in this round, and the processor places a token into the throw bucket.

There are exactly S tokens in the throw bucket and exactly W tokens is the work bucket. Thus the total number of tokens is at most $W+S$. Since in each round a process either executes a vertex or performs a throw, P-process execution time, $T_P = \frac{W+S}{P}$.

13.3.3 Bounding the Number of Throws

Our analysis will use a potential-function based method. We shall divide the computation into phases each of which decreases the potential by a constant factor.

Consider some round i and let R_i denote the set of ready vertices in the beginning of that round. A vertex is R_i is either assigned to a process or is in a deque. For each vertex $v \in R_i$, we define its ***potential*** as

1. $\phi_i(v) = 3^{2w(v)} - 1$, if v is assigned, or
2. $3^{2w(v)}$, otherwise.

We define the potential at round i, written Φ_i as $\Phi_i = \sum_{v \in R_i} \phi_i(v)$.

Definition: Beginning and termination potential

At the beginning of the computation, the only ready vertex is the root, which has a weight of S, because it is also the root of the enabling tree. Therefore the potential in the beginning of the computation is 3^{2S-1}.

At the end of the computation, there are no ready notes and thus the potential is zero.

Definition: Basic definitions for potentials

1. The ***potential of process*** q is $\Phi_i(q) = \sum_{v \in R_i(q)} \phi_i(v)$.
2. We write A_i for the set of processes whose deques are empty at the beginning of round i. We write $\Phi_i(A_i)$ for the total potential of the processes A_i, $\Phi_i(A_i) = \sum_{q \in A_i} \Phi_i(q)$.
3. We write D_i for the set of other processes. We write $\Phi_i(D_i)$ for the total potential of the processes D_i, $\Phi_i(D_i) = \sum_{q \in D_i} \Phi_i(q)$.
4. We can write the total potential in round i as follows $\Phi_i = \Phi_i(A_i) + \Phi_i(D_i)$.

Lemma: Vertex Assignment

Consider any round i and let v be a vertex that is ready but not assigned in the beginning of that round. Suppose that the scheduler assigns v to a process in that round. In this case, the potential decreases by $2/3\phi_i(v)$.

Proof

This is a simple consequence of the definition of the potential function: $\phi_i(v) - \phi_{i+1}(v) = 3^{2w(v)} - 3^{2w(v)-1} = 2/3\phi_i(v)$.

Lemma: Monotonicity of Total Potential

Total potential does not increase from one round to the next, i.e., $\Phi_{i+1} \leq \Phi_i$.

Proof

Let's now consider how the potential changes during each round. There are two cases to consider based on scheduler actions.

Case 1: Suppose that the scheduler assign a vertex v to a process. By the Vertex Assignment Lemma, we know that the potential decreases by $2/3\phi_i(v)$. Since $\phi_i(v)$ is positive, the potential decreases.

Note that this calculation holds regardless of where in the deque the vertex v is. Specifically, it could have been the bottom or the top vertex.

Case 2: suppose that the scheduler executes a vertex v and pushes its children onto the deque. There are two sub-cases to consider.

Case 2.1: suppose that the scheduler pushes onto the deque the only child x of a vertex v. Assuming that the child stays in the deque until the beginning of the next round, the potential decreases by

$$\phi_i(v) - \phi_{i+1}(x) = 3^{2w(v)-1} - 3^{2w(v)-2} = 3^{2w(v)-1}(1-1/3) = 2/3 \cdot \phi_i(v).$$

Case 2.2: suppose that the scheduler pushes onto the deque two children x, y of a vertex v. Assuming that the children remain in the deque until the beginning of the next round, then potential decreases by

$$\begin{aligned} \phi_i(v) - \phi_{i+1}(x) - \phi_{i+1}(y) &= 3^{2w(v)} - 3^{2w(v)-2} - 3^{2w(v)-2} \\ &= 3^{2w(v)-1}(1-1/3-1/3) \\ &= 1/3 \cdot \phi_i(v). \end{aligned}$$

Since $\phi_i(v)$ is positive, the potential decreases in both cases.

In each round each process performs none, one, or both of these actions. Thus the potential never increases.

Remark: in the second case, it is safe to assume that the children remain in the deque until the next round, because assignment of a vertex only decreases the potential further.

We have thus ustablished that the potential decreases but this by itself does not suffice. We also need to show that it decreases by some significant amount. This is our next step in the analysis.

We are now going to show that after P throws the total potential decreases with constant probability.

Lemma[Top-Heavy Deques]

Consider any round i and any process $q \in D_i$. The topmost vertex in q's deque contributes at least $3/4$ of the potential of process q.

This lemma follows directly from the structural lemma. The case in which the topmost vertex v contributes the least of the

process is when the assigned vertex and v have the same parent. In this case, both vertices have the same depth and thus we have

$$\Phi_i(q) = \phi_i(v) + \phi_i(x) = 3^{2w(v)} + 3^{2w(x)-1} = 3^{2w(v)} + 3^{2w(v)-1} = (4/3)\phi_i(v).$$

Lemma: P Throws

Consider any round i and any later round j such that at least P throws occur at rounds from i (inclusive) to j (exclusive). Then, we have

$$Pr[\Phi_i - \Phi_j \geq \frac{1}{4}\Phi_i(D_i)] > \frac{1}{4}.$$

Proof:

First we use the Top-Heavy Deques Lemma to establish the following. If a throw targets a process with a nonempty deque as its victim, then the potential decreases by at least of a half of the potential of the victim.

Consider any process q in D_i and let v denote the vertex at the top of its deque at round i. By Top-Heavy Deques Lemma, we have $\phi_i(v) \geq \frac{3}{4}\Phi_i(q)$.

Consider any throw that occurs at round $k \geq i$.

1. Suppose that this throw is sucessful with `popTop` returning a vertex. The two subcases both follow by the Vertex Assignment Lemma.
   a. If the returned vertex is v, then after round k, vertex v has been assigned and possibly executed. Thus, the potential has decreased by at least $\frac{2}{3}\phi_i(u)$.
   b. If the returned vertex is not v, then v is already assigned and possibly executed. Thus, the potential has decreased by at least $\frac{2}{3}\phi_i(v)$.
2. Suppose that the throw is not succesful, and `popTop` returns `NULL`. In this case, we know that q's deque was empty during `popTop`, or some other `popTop` or `popBottom` operation returned v. In all cases, vertex u has been assigned or possibly executed by the end of round k and thus, potential decrases by $\frac{2}{3}\phi_i(u)$.

Thus, we conclude that if a thief targets a process $q \in D_i$ as victim at round $k \geq i$, then the potential decreases by at least

$$\frac{2}{3}\phi_i(u) \geq \frac{2}{3}\frac{3}{4}\Phi_i(q) = \frac{1}{2}\Phi_i(q).$$

Second, we use Ball and Bins Lemma to establish the total decrease in potential.

Consider now all P processes and P throws that occur at or after round i. For each process q in D_i, assign a weight of $\frac{1}{2}\Phi_i(q)$ and for each process in A_i, we assign a weight of 0. The total weight is thus $\frac{1}{2}\Phi_i(D_i)$. Using the Balls-and-Bins Lemma with $\beta = 1/2$, we conclude that the potential decreases by at least $\beta W = \frac{1}{4}\Phi_i(D_i)$ with probability greater that $1 - \frac{1}{(1-\beta)e} > \frac{1}{4}$.

Important
For this lemma to hold, it is crucial that a steal attempt does not fail unless the deque is empty or the vertex being targeted at the time is popped from the deque is some other way.

Theorem: Dedicated Environments

Consider any multithreaded computation with work W and span S and execute it with non-blocking work stealing with P processes in a dedicated environment. The exection time is

1. $O(W/P+S)$ in expectation, and
2. $O(W/P+S+\lg(1/\varepsilon))$ with probability at least $1-\varepsilon$ for any $\varepsilon > 0$.

Proof

The Throw-Bound Lemma bounds the execution time in terms of throws. We shall prove bounds on the number of throws.

We break execution into ***phases*** of $\Theta(P)$ throws. We show that with constant probability, a phase causes the potential to drop by a constant factor, and since we that the the potential starts at $\Phi_0 = 3^{2S-1}$ and ends at zero, we can bound the number of phases.

The first phase begins at round $t_1 = 1$ and ends at the first round t_i', where at least P thorws occun during the interval $[t_1, t_1']$. The second phase begins at round $t_2 = t_1' + 1$ and so on.

Consider a phase $[i, j)$, where the next round begins at round j. We show that $\Pr[\Phi_j \leq \frac{3}{4}\Phi_i] < \frac{1}{4}$.

Recall that the potential can be partitioned as $\Phi_i = \Phi_i(A_i) + \Phi_i(D_i)$.

1. Since the phase contains at least P throws, by P Throws Lemma, we know that $P[\Phi_i - \Phi_j \geq \frac{1}{4}\Phi_i(D_i)] > \frac{1}{4}$.
2. We need to show that the potential also drops by a constant fraction of $\Phi_i(A_i)$. Consider some process q in A_i. If q does not have an assigned node, then $\Phi_i(q) = 0$. If q has an assigned node u, then $\Phi_i(q) = \phi_i(u)$. In this case, process q executes node u at round i and the potential drops by at least $\frac{5}{9}\phi_i(u)$. Summing over all processes in A_i, we have $\Phi_i - \Phi_j \geq \frac{5}{9}\Phi_i(A_i)$.

Thus we conclude that $P[\Phi_i - \Phi_j \geq \frac{1}{4}\Phi_i] > \frac{1}{4}$. In other words, we have established that in any phase, potential drops by a quarter with some probability $\frac{1}{4}$.

Define a phase to be ***successful*** if it causes the potential do trop by at least a quarter fraction. We just established that phase is successful with probability at least 0.25. Since the start potential $\Phi_0 = 3^{2S-1}$ and ends at zero and is always an integer, the number of successful phases it as most $(2S-1)\log_{4/3} 3 < 8S$.

Thus, the expceted number of phases is $O(S)$ and since each contains $O(P)$ throws, the expected number of throws is $O(SP)$.

We now establish the high-probability bound.

Suppose that the execution takes $n = 32S + m$ phases. Each phase succeds with probability at least $p = \frac{1}{4}$, so the expceted number of successes is at least $np = 8S + m/4$. Let X the number of successes. By Chernoff bound

$$P[X < np - a] < e^{-a^2/2np},$$

with $a = m/4$. Thus if we choose $m = 32S + 16\ln 1/\varepsilon$, then we have

$$P[X < 8S] < e^{-(m/4)^2/(16S+m/2)} \leq e^{-(m/4)^2/(m/2+m/2)} = e^{-m/16} \leq e^{-16\ln 1/\varepsilon/16} = \varepsilon.$$

Thus the probabilty that the executian takes $64S + 18\ln 1/epsilon)$ phases or more is less than ε.

We conclude that the number of throws is $O(S + \lg 1/\varepsilon P)$ with probability at least $1-\varepsilon$.