A Guide to Parallel Programming

Design Patterns for Decomposition, Coordination and Scalable Sharing

PRELIMINARY

April 20, 2010

Preface

This book describes patterns for parallel programming, with code examples that use the Parallel Extensions of .NET Framework 4.0 and the Parallel Patterns Library (PPL) that are shipped with Visual Studio 2010. You can use the patterns described in this book to improve your application's performance on multicore computers. Adopting the patterns in your code will make your application run faster today but will also help prepare for future hardware environments, which are expected to have an increasingly distributed computing architecture.

Who This Book Is For

The book is intended for programmers who write managed code in .NET or unmanaged code on the Windows platform. This includes programmers who write in C++, C#, Visual Basic and F#. No prior knowledge of parallel programming techniques is assumed. However, readers need to be familiar with features of C# such as delegate methods, lambda expressions, generic types and Language Integrated Query (LINQ) expressions. Readers should also have at least basic familiarity with the concepts of processes and threads of execution.

Note: Although the material in this book can be used with languages other than C#/C++, it focuses on C# with some callouts for the equivalent C++. Particular attention is given to relevant features in .NET Framework 4.0.

The examples in this book are written in C# and use the features of the .NET 4 Framework, including the Task Parallel Library (TPL) and Parallel LINQ (PLINQ). However, you can use the concepts presented here with other frameworks and libraries. For example, F# now has language support for concurrency and unmanaged code can be written to use the Parallel Patterns Library (PPL).

Complete code solutions are posted on CodePlex. See http://parallelpatterns.codeplex.com. There is both a C# (or PLINQ) version and a C++ version for every example.

Why This Book Is Pertinent Now

The advanced parallel programming features that are delivered with Visual Studio 2010 make it easier than ever to get started with parallel programming.

The Task Parallel Library (TPL) is available for .NET programmers who want to write parallel programs. It simplifies the process of adding parallelism and concurrency to applications. The TPL scales the degree of parallelism dynamically to most efficiently use all the processors that are available. In addition, the TPL assists in the partitioning of the work and the scheduling of tasks in the .NET thread pool. The library provides cancellation support, state management, and other services.

Parallel LINQ (PLINQ) is a parallel implementation of LINQ to Objects. PLINQ implements the full set of LINQ standard query operators as extension methods for the **System.Linq** namespace and has additional operators for parallel operations. PLINQ is a declarative, high-level interface for parallel loops and aggregation operations.

The Parallel Patterns Library (PPL) is for C++ programmers. It provides task parallelism, parallel algorithms, and parallel containers and objects.

Visual Studio 2010 includes tools for debugging parallel applications. The Parallel Stacks Window shows call stack information for all the threads in your application. It lets you navigate between threads and stack frames on those threads. The Parallel Tasks window resembles the Threads window, except that it shows information about each task handle object instead of each thread. The Concurrency Visualizer views enable you to see how your application interacts with the hardware, the operating system and other processes on the computer. You can use the Concurrency Visualizer to locate performance bottlenecks, CPU underutilization, thread contention, cross-core thread migration, synchronization delays, areas of overlapped I/O, and other information.

What You Need to Use the Code

The code that is used as examples in this book is at http://parallelpatterns.codeplex.com/. These are the system requirements:

- Microsoft Windows Vista, SP1, Windows 7, or Microsoft Windows Server 2008 (32-bit or 64-bit)
- Microsoft Visual Studio® 2010 (Ultimate edition is required for the Concurrency Visualizer, which allows you to analyze the performance of your application). This includes .NET 4 Framework, which is required to run the samples.

How to Use This Book

This book presents parallel programming techniques in terms of particular patterns. Figure 1 shows the different patterns and their relationships to each other. The numbers refer to the chapters in this book where the pattern is described.

Parallel programming patterns

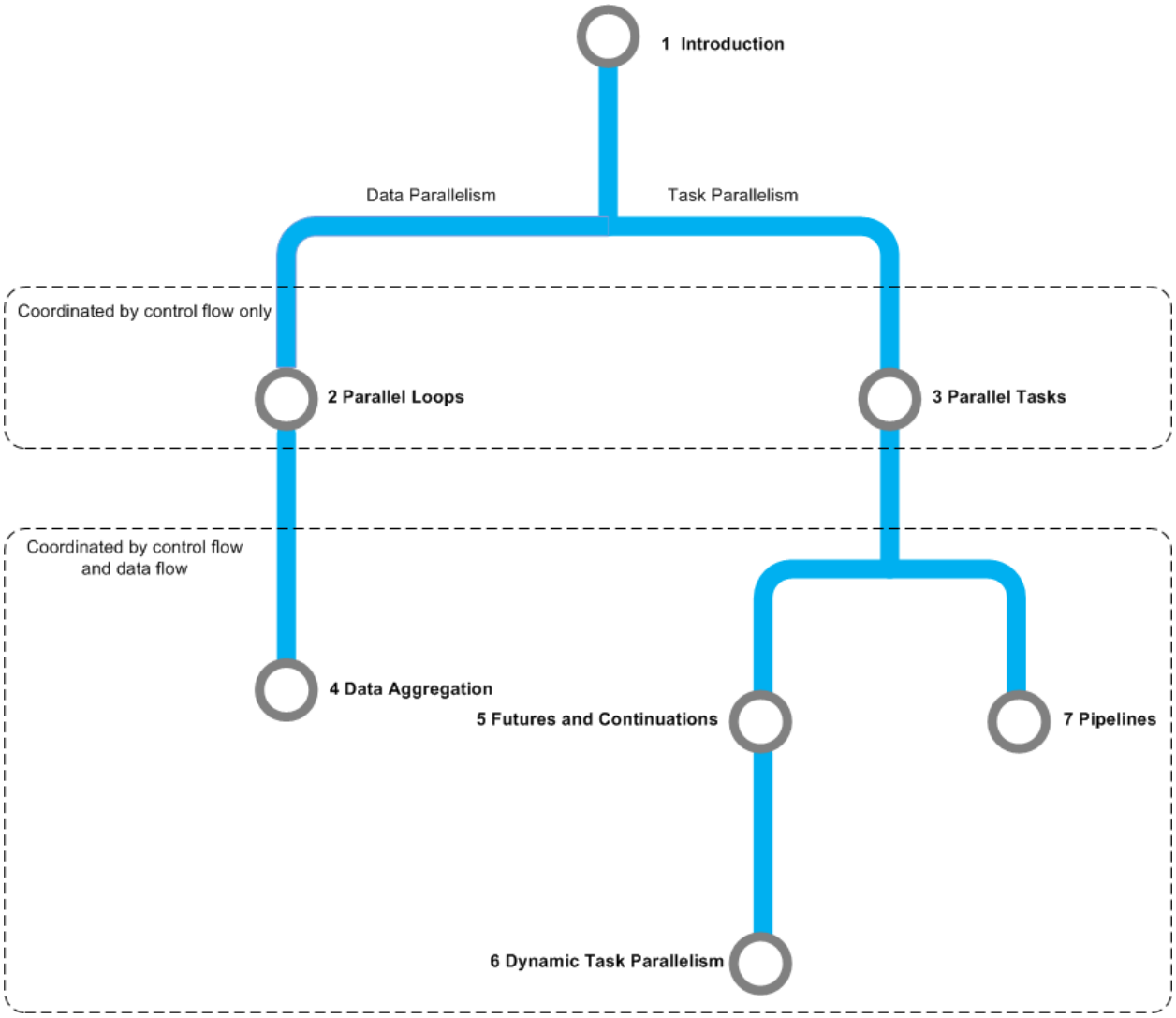


Figure 1

After the introduction, the book has two branches. One of them discusses data parallelism and the other discusses task parallelism. Both parallel loops and parallel tasks use only the program's control flow as the means of coordinating and ordering tasks. The other patterns use both control flow and data flow for coordination.

Introduction

Chapter 1 introduces the common problems faced by developers who want to use parallelism to make their applications run faster. It explains basic concepts and prepares you for the remaining chapters.

Parallelism with Control Dependencies Only

Chapters 2 and 3 deal with cases where asynchronous operations are ordered only by control flow constraints:

- **Chapter 2: Parallel Loops**. Use parallel loops when you want to perform the same calculation on each member of a collection or for a range of indices.
- **Chapter 3: Parallel Tasks**. Use parallel tasks when you have several distinct asynchronous operations to perform. This chapter explains why tasks and threads serve two distinct purposes.

Parallelism with Control and Data Dependencies

Chapters 4 and 5 show patterns for concurrent operations that are constrained by both control flow and data flow:

- **Chapter 4: Data Aggregation using Map/Reduce**. Patterns for data aggregation are used when the body of a parallel loop includes data dependencies, for example when calculating a sum or searching a collection for a maximum value.
- **Chapter 5: Futures and Continuations**. This pattern arises when operations produce some outputs that are needed as inputs to other operations. The order of operations is constrained by a directed graph of data dependencies. Some operations performed in parallel and some serially, depending on when inputs become available.

Dynamic Task Parallelism, and Pipelines

Chapters 6 and 7 discuss some more advanced scenarios.

- **Chapter 6: Dynamic Task Parallelism**. In some cases operations are dynamically added to the backlog of work as the computation proceeds. The pattern applies to several domains, including graph algorithms and sorting.
- **Chapter 7: Pipelines**. Use pipelines to feed successive outputs of one component to the input queue of another component, in the style of an assembly line. Parallelism results when the pipeline fills, and all components are simultaneously active.

Supporting Material

In addition to the patterns, there are several appendices.

- **Appendix A: Supporting Patterns**. This appendix gives tips for adapting some of the common object-oriented patterns such as facades, decorators and repositories to multicore architectures. For example, it shows how "lazy" data access can promote concurrency in a repository. Event-based coordination ("agents") is another familiar pattern that is also often adapted for parallel scenarios.
- **Appendix B: Debugging and Profiling Parallel Applications**. This appendix gives you an overview of how to debug and profile parallel applications in VS 2010.
- **Appendix C: Technology Roadmap**. A technology roadmap helps you place the various Microsoft technologies and frameworks for parallel programming in context.

- **Appendix D: QuickStart Examples**. This appendix contains very brief code examples for common tasks such as "How-to parallelize a loop over a collection." It is meant as a short-cut reference.

A glossary contains the definition of terms used in this book.

You can read the chapters in this book sequentially but everyone should read chapters 1, 2, and 3 for an introduction and overview of the basic principles. Although attention has been paid to presenting the material in a logical order, each chapter, from chapter 4 on, is independent.

Callouts show things you should watch out for. These are sometimes called "anti-patterns."

> **Anti-pattern:** Don't apply the patterns in this book blindly to your applications. (Hammer/nail problem)

What Is Not Covered

This book focuses more on CPU-bound workloads than on I/O-bound workloads. The goal is to make computationally intensive applications run faster by making better use of the computer's available cores. As a result, the book does not focus as much on the issue of I/O latency. Nonetheless, some attention is paid to balanced workloads that have are both CPU intensive and have large amounts of I/O (see Chapter 7, data pipelines). There is also an important example for user interfaces in Chapter 5 (Futures and Continuations) that illustrates concurrency for tasks with I/O.

The book describes parallelism within a single multicore node with shared memory rather than the cluster, High Performance Computing (HPC) Server approach that uses networked nodes with distributed memory. However, cluster programmers who want to take advantage of parallelism within a node may find the examples in this book helpful, since each node of a cluster may have multiple processing units.

1 Introduction

The CPU meter shows the problem. One core is running at 100%, but all of the others are idle. Your application is CPU bound, but you are using only a fraction of the computing power of your multicore system. What next?

The answer, in a nutshell, is *parallel programming*. Where you once would have written the kind of sequential code that is familiar to all programmers, you now find that this no longer meets your performance goals. To use your system's CPU resources efficiently you need to split your application into pieces that can run at the same time.

Parallel programming lets your application use more than one core at the same time. The goal is to improve the application's speed.

This is easier said than done. Parallel programming has a reputation for being the domain of experts and a minefield of subtle, hard-to-reproduce software defects. Everyone seems to have a favorite story about a parallel program that did not behave as expected or of an insidious and mysterious bug.

Anti-pattern: Some bugs appear only in programs that use more than one thread of execution. A favorite example of such a bug is known as a *torn write*. The problem starts when a thread updates a multiword structure in memory. Such updates take more than one machine instruction to complete and are therefore not guaranteed to be atomic. It is possible that a thread running simultaneously on another core could read the contents of a partially written structure, with disastrous results. This is not as esoteric as it sounds. Many programmers are unaware that common data types such as **double** and **System.Guid** are usually implemented as structures that span multiple words of memory. Without careful synchronization, simply setting and reading the value of a shared variable that contains a double-precision floating point number or a GUID can corrupt the memory of a parallel application. (Unfortunately, the cure, which is to add a lock, has side effects. It can drastically reduce the program's performance and, if not done carefully, can introduce its own bugs such as deadlock.)

Many of the trickiest bugs in parallel programs are performance related. Your program produces the expected results, but you find that it runs slowly for no apparent reason. A favorite example of this kind of problem is known as *false sharing*. Imagine that you write into adjacent locations of an array from two threads. Thread 1 reads and writes array locations indexed by even numbers. Thread 2 reads and writes only odd-numbered locations. Your program will run correctly and requires no special synchronization, but the performance is much worse than you expect. The reason is that your computer's memory cache operates on chunks of data that can include more than one array element. Even though you are writing to separate locations of memory, the memory cache doesn't see it this way. The memory cache is refreshed with every write, and this heavyweight operation cripples performance. Eliminating the interleaved writes by simply dividing the work differently (for example, by writing to separate regions of the array) clears up the mysterious problem.

Writing parallel programs has the reputation of being hard.

These stories should inspire a healthy respect for the difficulty of the problems you face in writing your own parallel programs. Fortunately, help has arrived. Microsoft's Visual Studio 2010 has introduced a new programming model for parallelism that significantly simplifies the job. Behind the scenes are supporting libraries with sophisticated algorithms that dynamically distribute computation on multicore architectures.

Proven design patterns are another source of help. This guide introduces you to the most important and frequently used patterns of parallel programming and gives executable code samples for them using the Task Parallel Library (TPL) and Parallel LINQ (PLINQ), with C++ examples that use the Parallel Patterns Library (PPL) provided in the online supplement. When thinking about where to begin, a good place to start is to review the patterns in this book. See if your problem has any attributes that match the seven basic patterns presented in the chapters. If it does, then delve more deeply into the relevant pattern or patterns and study the sample code.

The code examples for this guide are online at http://parallelpatterns.codeplex.com.

The Importance of Potential Parallelism

The patterns in this book are ways to express *potential parallelism*. This means that your program is written so that it runs faster when parallel hardware is available and the same as an equivalent sequential program when it's not. If you correctly structure your code then the run-time environment can automatically adapt to the workload on a particular computer. For this reason, the patterns in this book themselves only express potential parallelism. They do not guarantee it. This concept is a central organizing principle behind the parallel programming model of Visual Studio 2010. It deserves some explanation.

Declaring the potential parallelism of your program allows the execution environment to run it on all available cores.

Some parallel applications can be written for specific hardware. For example, creators of programs for a console gaming platform have detailed knowledge about the hardware resources that will be available at run time. They know the number of processors and the details of the memory architecture in advance. The game can be written to exploit the exact level of parallelism provided by the platform. Complete knowledge of the hardware environment is also a characteristic of some embedded applications such as industrial control. The lifecycle of such programs matches the lifecycle of the specific hardware they were designed to use.

In contrast, when you write programs that run on general-purpose computing platforms, such as desktop workstations and servers, there is less predictability about the hardware features. You may not always know how many cores will be available. You may not know whether memory access times will be uniform across all cores or whether the memory system of the computer optimizes memory access by grouping cores and memory regions into nodes.

It's generally a bad idea to hard code the degree of parallelism in your application. You can't always predict how many cores will be available at run time.

In addition to uncertainty about the hardware capabilities of the computer that will run your application, you also may be unable to predict what other software could be running at the same time as your application.

Even if you initially know about your application's environment, it can change over time. Planning for the future adds yet another layer of uncertainty, especially if the application will outlast the current generation of hardware. In the past, programmers assumed that their applications would automatically run faster on later generations of hardware. You could rely on this because processor clock speeds kept increasing. With multicore processors, clock speeds may not increase with newer hardware as much as they have in the past. Instead, the trend in processor design is toward more cores and more distribution of memory systems. If you want your application to benefit from hardware advances in the new multicore world, you need to adapt your programming model.

Hardware trends predict more cores and more distributed memory architectures instead of faster clock speeds.

Finally, you must plan for all of these contingencies in a way that does not penalize users who might not have access to the latest hardware. In other words, you want your parallel application to run as fast on a single-core computer as an application that was written using only sequential code. In other words, you want *scalable performance* from one to many cores.

A well-written parallel program runs at approximately the same speed as a sequential program when there is only one core available.

Allowing your application to adapt to varying hardware capabilities, both now and in the future, is the motivation for potential parallelism.

An example of potential parallelism is the parallel loop pattern described in chapter 2. If you have a **for** loop that performs a million independent iterations, it makes sense to divide those iterations among the available cores and do the work in parallel. It's easy to see that how you divide the work should depend on the number of cores. If you do this the speed of the loop will be approximately proportional to the number of cores for many common scenarios.

Decomposition, Coordination and Scalable Sharing

The patterns in this book contain some common themes. You will see that the process of designing and implementing a parallel application involves three aspects: *decomposing* the work into discrete units called tasks, ways of *coordinating* these tasks as they run in parallel, and scalable techniques for *sharing* the data needed to perform the tasks.

Understanding Tasks

Tasks are small sequential operations that work together to perform a larger operation. When you think about how to structure a parallel program, it's important to identify tasks a relatively fine level of granularity. Decomposing a problem into tasks requires a good understanding of the algorithmic and

structural aspects of your application. A guiding rule is that tasks should be as independent as possible. Also, the granularity of tasks must be carefully chosen—too fine and the overhead of managing tasks will dominate; too coarse and opportunities for parallelism may be lost if there aren't enough tasks to occupy the available hardware.

Tasks are small, sequential units of work. They should be as independent as possible.

Anti-pattern: Tasks are not threads. The distinction between tasks and threads is covered in chapter 3, Parallel Tasks.

Coordinating Tasks

It is often possible that more than one task can run at the same time. Tasks that are independent of one another can run in parallel. Some tasks can begin only after other tasks complete. The order of execution and the degree of parallelism are constrained by the application's underlying algorithms. Constraints can arise from control flow (the steps of the algorithm) or data flow (the availability of inputs and outputs). Various mechanisms for coordinating tasks are possible. How tasks are coordinated depends on which parallel pattern you use. For example, the pipeline pattern described in chapter 7 is distinguished by its use of concurrent queues to coordinate tasks.

Regardless of the mechanism you choose for coordinating tasks, in order to have a successful design you must thoroughly analyze the dependencies between tasks.

Scalable Sharing of Data

Tasks often need to share data. There are a number of techniques that allow data to be shared that don't degrade performance or make your program prone to error. These techniques include the use of immutable, read-only data, carefully isolating mutable state, limiting your program's reliance on shared variables and introducing new steps in your algorithm that merge mutable state at appropriate checkpoints. In other words, scalable sharing may involve changes to the algorithm. It's not just a matter of adding locks.

Scalable sharing may involve changes to your algorithm. Adding synchronization (locks) usually effect the scalability of your application.

Anti-pattern: Conventional object-oriented designs can have complex and highly interconnected in-memory graphs of object references. As a result, traditional object-oriented programming styles can be very difficult to adapt to scalable parallel execution. Your first impulse might be to consider all fields of a large, interconnected object graph as mutable shared state and to wrap access to these fields in serializing locks whenever there is the possibility that they may be shared by multiple tasks. However, this is *not* a scalable approach to sharing. Locks use *memory fences* that negatively affect the performance of all cores. (Refer to the glossary for more information about memory fences.) As the number of cores gets larger, the cost of each lock increases. As more and more tasks are added that share the same data, the overhead associated with locks can dominate the computation.

In addition to performance problems, programs that rely on complex synchronization are prone to deadlock and other software defects. Most of the horror stories about parallel programming are actually horror stories about the overuse of shared mutable state.

Of course, synchronizing (that is, locking) elements in an object graph plays a legitimate if limited role in scalable parallel programs. This book uses synchronization sparingly. You should too. Locks are the **goto** statements of parallel programming: error prone, occasionally necessary and usually best left to compilers and libraries.

Locks are the **goto** statements of parallel programming; use them sparingly.

Design Approaches

Techniques for decomposition, coordination and scalable sharing are interrelated. There's a circular dependency. You need to consider all of these aspects together when choosing your approach for a particular application.

After reading the preceding description you might complain that it all seems vague. How *specifically* do you divide your problem into tasks? Exactly what kinds of coordination techniques should be used?

Use patterns.

Questions like these are best answered by the patterns described in this book. Patterns are a true short cut to understanding. As you begin to see the design motivations behind the patterns in this book you will also develop your intuition about how the patterns can be applied to your own applications. The following section gives more detail on how you can use the parallel programming patterns described in this book.

Selecting the Right Pattern

To select the relevant pattern, use the following table.

Application characteristic	Relevant pattern
Do you have sequential loops where there's no communication among the steps of each iteration?	The Parallel Loop pattern (chapter 2). Parallel loops apply an independent operation to multiple inputs simultaneously.
Do you have specific units of works with well-defined control dependencies?	The Parallel Task pattern (chapter 3) Parallel tasks allow you to establish parallel control flow in the style of fork and join.
Do you need to summarize data by applying some kind of combination operator? Do you have loops with steps that are not fully independent?	The Parallel Aggregation pattern (chapter 4) Parallel aggregation introduces special steps in the algorithm for merging partial results. This pattern includes map/reduce as one of its variations.
Does the ordering of steps in your algorithm depend	The Futures and Continuations pattern (chapter 5).

on data flow constraints?	Futures and continuations make the data flow dependencies between tasks explicit.
Does your algorithm divide the problem domain dynamically during the run? Do you operate on recursive data structures such as graphs?	The Recursive Task Parallelism pattern (chapter 6) This pattern takes a divide-and-conquer approach and spawns new tasks on demand.
Does your application perform a sequence of operations repetitively? Does the input data have streaming characteristics?	The Pipeline pattern (chapter 7) Pipelines consist of components that are connected by queues, in the style of produces and consumers. All of the components run in parallel.

One way to familiarize yourself with the possibilities is to read the first page or two of each chapter. This will give you an overview of approaches that have been proven to work in a wide variety of applications. Then go back and more deeply explore patterns that may apply in your situation.

A Word about Terminology

You will often hear the words *parallelism* and *concurrency* used as synonyms. This book makes a distinction between the two terms. The goals of concurrency and parallelism are distinct.

Concurrency is a concept related to multitasking. It refers to the existence of multiple threads of execution that may each get a slice of time to execute before being preempted by another thread, which also gets a slice of time. Concurrency can exist even on a computer with a single core, for example, as a way to make a user interface more responsive. With concurrency not all components need to be simultaneously active.

With *parallelism* concurrent threads execute *at the same time* using multiple cores. Parallel programming focuses on improving the performance of CPU-bound applications when multiple cores are available.

The Limits of Parallelism

A theoretical result called Amdahl's law says that the amount of performance improvement parallelism provides is limited by the amount of sequential processing in your application. Amdahl's Law may at first seem counterintuitive.

Amdahl's law says that no matter how many cores you have, the maximum speedup you can ever achieve is (1 / percent of time spent in sequential processing). This is shown in Figure 1.

Amdahl's Law for an Application with 25% Sequential Processing

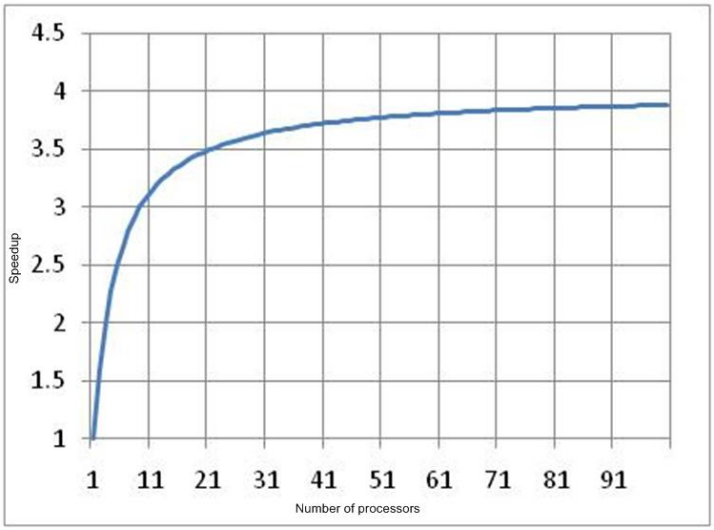


Figure 1

For example, with 11 processors, the application runs slightly more than three times faster than it would if it were entirely sequential.

Even with fewer cores, you can see that the expected speedup is not perfectly linear. This is shown in Figure 2.

Per-core Performance Improvement for a 25% Sequential Application

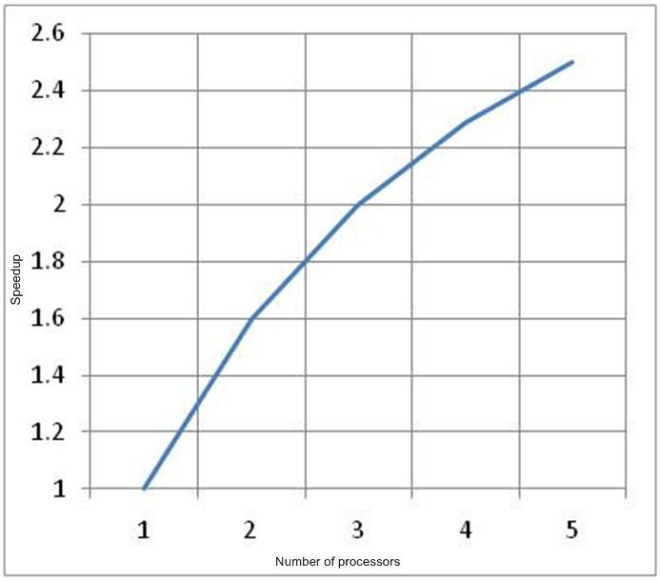


Figure 2

In practice, the speedup you can achieve is somewhat worse than Amdah's Law would predict. As the number of cores increases the overhead incurred by accessing shared memory increases.

This means that to see the most improvement when parallelizing your application you need to focus on the areas which take the most time. Sometimes this is obvious but often it requires careful

measurement and time spent with profiling tools to understand which parts of your program will benefit from parallelization and what the likely improvement in performance will be.

A Few Tips

Always try for the simplest approach. Here are some basic precepts:

- Whenever possible, use a library that does the parallel work for you.
- Make use of your application server's inherent parallelism; for example, in the Web server or database.
- Use an API to encapsulate parallelism, such as Microsoft's Parallel Extensions for .NET (TPL and PLINQ) or Microsoft's Parallel Patterns Library (PPL) for unmanaged code. These libraries were written by experts and avoid many of the common problems that arise in parallel programming.
- Use patterns, such as the ones described in this book.
- Don't share data among concurrent tasks unless absolutely necessary. If you do share data then use one of the containers provided by the API you are using, for example a shared queue.
- Use low-level primitives, such as threads and locks, only as a last resort. Raise the level of abstraction from threads to tasks in your applications.
- Often, restructuring your algorithm (for example, to eliminate the need for shared data) is better than making low-level improvements to code that was originally designed to run serially.

Obviously this introduction only gives a broad outline as to how to approach parallelizing your application.

2 Parallel Loops

Use the parallel loop pattern when the same independent operation needs to be performed for each element of a collection or for a fixed number of iterations. Iterations of a loop are independent if they don't write to memory locations or files that are read by other iterations. The syntax of a parallel loop is very similar to the **for** and **foreach** loops you already know, but the parallel loop performs much faster on multicore computers.

The parallel loop pattern independently applies an operation to multiple data elements. This kind of computation is sometimes called *data parallelism*.

Parallel loops express *potential parallelism*. This means that the degree of parallelism doesn't need to be specified by your code. Instead, the run-time environment executes iterations of your loop as concurrently as possible on the available hardware resources. The loop works correctly no matter how many CPU cores are available. If there is only one core then the performance is about the same as a sequential iteration. If there are many cores then performance improves proportionately.

The Problem

The parallel loop pattern allows an application that uses **for** or **foreach** loops to run faster on multicore computers. A parallel loop does the same work as a sequential loop, only in less time when more than one core is available. You can use both **parallel for** and **parallel foreach** loops, depending on whether you want to iterate over a range of integer indices or over values from a collection.

It's easy to make **for** or **foreach** loops with independent iterations run faster on multicore computers by using their parallel counterparts.

Parallel For Loops

Here is an example of a C# sequential **for** loop.

```
for (int i = 0; i < N; i++)
{
   // ... do some work ...
}
```

This example assumes that iterations of the loop body are independent of one another.

The computation can take advantage of multiple cores by simply replacing the **for** keyword with a call to the **Parallel.For** method.

```
Parallel.For(0, N, i =>
{
   // ... do some work ...
});
```

Parallel.For uses multiple cores to operate over an index range.

Parallel.For is one of the parallel extensions provided as part of the .NET 4 Framework. It is a normal static method with three arguments.

> **C++ Note:** Parallel loops in C++ can use the **Concurrency::parallel_for** function that is provided by the Parallel Patterns Library (PPL) in Visual Studio 2010.

The first two arguments specify the iteration limits. The first argument is the lowest index of the loop. The second argument is the exclusive upper bound. This is the largest index plus one. The third argument is a delegate method that will be invoked once per iteration. The delegate method takes the iteration's index as its argument and executes the loop body.

> **Note:** The code example includes a lambda expression in the form i => ... as the third parameter to the **Parallel.For** method. Lambda expressions denote unnamed or anonymous delegate methods. If you are unfamiliar with the syntax for lambda expressions, check the reference section at the end of this chapter for more information.

> **Anti-pattern**: If you need to use a step size other than one, transform the loop index in the body of the loop. Be aware that a step size other than one can indicate a data dependency, such the calculation of a sum. Analyze such a computation carefully before you convert it into a parallel loop.

Parallel ForEach

You can also execute a **foreach** loop in parallel. Here is a simple example of a sequential C# **foreach** loop.

```
MyObject[] myEnumerable = ...

foreach (var obj in myEnumerable)
{
   // ... do some work ...
}
```

This example assumes that iterations of the loop body don't write to memory locations or files that are read by other iterations. The loop body only makes updates to fields of the particular instance of the **MyObject** class that is passed to it with each iteration. It doesn't read fields that are updated by other iterations.

Take advantage of multiple cores by replacing the **foreach** keyword with a call to the **Parallel.ForEach** method.

```
MyObject[] myEnumerable = ...

Parallel.ForEach(myEnumerable, obj =>
{
```

```
   // ... do some work ...
});
```

Parallel.ForEach runs the loop body for each element in a collection.

Parallel.ForEach is one of the parallel extensions provided as part of the .NET 4 Framework. It is a normal static method with two arguments.

> **C++ Note:** Parallel loops in C++ can use the **Concurrency::parallel_for_each** function that is provided by the Parallel Patterns Library (PPL) in Visual Studio 2010.

The first argument contains a collection that provides the **IEnumerable<T>** interface. The second argument is a delegate method that will be invoked for each element of the input collection.

Unlike the sequential case, the order of execution is not guaranteed with either the **Parallel.For** or the **Parallel.ForEach** method.

Benefits

If you compare the **Parallel.For** loop with the original **for** loop, or the **Parallel.ForEach** loop with the original **foreach** loop, you'll notice that the sequential and parallel loop bodies are the same. The performance, however, is not. If you have more than one core on your computer, as most recently manufactured computers do, then the parallel loop will use them. By default, the degree of parallelism (that is, how many iterations run concurrently in hardware) depends on the number of cores on your computer. The more cores you use, the faster your loop executes. How much faster depends on the kind of work your loop does. See the Design Notes section of this chapter for more information.

If you use a parallel loop, you are planning for a future where the number of cores on a typical computer will increase. Your code will run unchanged, without recompilation, and automatically take advantage of the larger number of cores.

Implementations of the parallel loop pattern ensure that exceptions that are thrown during the execution of a loop body are not lost. For both the **Parallel.For** and **Parallel.ForEach** methods, exceptions that occur are collected into an **AggregateException** object and rethrown in the context of the calling thread. This ensures that all exceptions are propagated back to you. See the Variations section of this chapter to learn more about exception handling for parallel loops.

Robust exception handling is an important aspect of the parallel loop pattern.

So far you've seen only the most basic parallel loops. There are many variations. There are twelve overloaded methods for **Parallel.For** and twenty overloaded methods for **Parallel.ForEach**. In addition, the Language Integrated Query (LINQ) feature of .the NET framework includes a parallel version named PLINQ (Parallel LINQ). With over 100 extension methods, there are many options and variations on how to express PLINQ queries. See the Variations section of this chapter to learn about some of the most important cases.

The parallel loop pattern is one of the easiest parallel patterns to apply to existing code. Just look for the top-level loops in your application or the loops which consume most of the processor cycles and check

that the loop body is independent. If it is, replace **for** or **foreach** with **Parallel.For** or **Parallel.ForEach**. It's important to do this in a systematic manner, review and profile your code to establish which parts of your program will benefit from parallelization. See appendix ### for further discussion of profiling.

Anti-pattern: If the loop body is not independent, for example, when you use an iteration to calculate a sum, then you may need to apply the variation on a parallel loop that's described in chapter 4, Parallel Aggregation. If your loop has a negative step size (that is, it iterates from high to low values), then the ordering of the loop is probably significant and the iterations may not be independent.

An Example

Here's an example of when to use a parallel loop. Fabrikam Shipping extends credit to its commercial accounts. It wants to use customer credit trends to identify accounts that might pose a credit risk. Each customer account includes a history of past balance-due amounts. Fabrikam has noticed that customers who don't pay their bills often have histories of steadily increasing balances over a period of several months before they default.

To identify at-risk accounts, Fabrikam uses statistical trend analysis to calculate a projected credit balance for each account. If the analysis predicts that a customer account will exceed its credit limit within three months, the account is flagged for manual review by one of Fabrikam's credit analysts.

The account data includes balance histories for each customer. In the application, a top-level loop iterates over customers in the account repository. The body of the loop fits a trend line to the balance history, extrapolates the projected balance, compares it to the credit limit, and assigns the warning flag if necessary.

An important aspect of this application is that each customer's credit status can be calculated independently. The credit status of one customer doesn't depend on the credit status of any other customer. Because the operations are independent, making the credit analysis application run faster is simply a matter of replacing a sequential **foreach** loop with a parallel loop.

Anti-pattern: You will probably not get performance improvements if you use a parallel loop for very small loop bodies with only a limited number of data elements to process. In this case, the overhead required by the parallel loop itself will dominate the calculation. Simply changing every sequential **for** loop to **Parallel.For** will not necessarily produce good results.

Instead, the best place to replace a sequential loop with a parallel loop is at the top level of your application, as in the credit review sample. In other words, it's better to introduce parallel steps at a relatively coarse level of granularity if you have the choice. The only reason that large steps would be inappropriate is if there are too few of them, or if the steps are of such uneven size that processing the last step leaves many cores idle for a long time.

It's possible to nest parallel loops. In this situation, the run-time environment coordinates the use of processor resources. Another option to consider is to combine the nested loops into a single loop.

Nesting or unrolling loops is appropriate in cases where the amount of work in each top-level iteration is either very large or of uneven size.

Small loop bodies, even as small as a single multiplication, are not necessarily ineligible for parallel loops. It also depends on the number of iterations. If you have very small loop bodies but a large number of data elements, then a parallel loop may still provide a worthwhile improvement in performance. See Special Handling for Small Loop Bodies in the Variations section of this chapter for more information.

When you're thinking about how and where to add parallel loops, keep in mind the first rule of parallel programming: if in doubt, profile. Performance testing is the only way to guarantee that you are achieving the expected speedup.

If in doubt, profile.

The complete source code for this example can be found online at http://parallelpatterns.codeplex.com in the Chapter2\CreditReview project.

Sequential Prediction

Here is the sequential version of the credit analysis operation.

```
static void UpdatePredictionsSequential(
                                    AccountRepository accounts)
  {
    foreach (Account account in accounts.AllAccounts)
    {
      Trend trend = SampleUtilities.Fit(account.Balance);
      double prediction = trend.Predict(
                    account.Balance.Length + NumberOfMonths);
      account.SeqPrediction = prediction;
      account.SeqWarning = prediction < account.Overdraft;
  }
}
```

The **UpdatePredictionsSequential** method processes each account from the application's account repository. The **Fit** method is a utility function that uses the least squares method to create a trend line from an array of numbers. The prediction is a three-month projection based on the trend. If a prediction is more negative than the overdraft limit (credit balances are negative numbers in the accounting system), the account is flagged for review.

Prediction Using Parallel.ForEach

The parallel version of the credit scoring analysis is very similar to the sequential version.

```
static void UpdatePredictionsParallel(AccountRepository accounts)
{
  Parallel.ForEach(accounts.AllAccounts, account =>
  {
```

```
    Trend trend = SampleUtilities.Fit(account.Balance);
    double prediction = trend.Predict(
                        account.Balance.Length + NumberOfMonths);
    account.ParPrediction = prediction;
    account.ParWarning = prediction < account.Overdraft;
  });
}
```

The **UpdatePredictionsParallel** method is identical to the **UpdatePredictionsSequential** method, except that the **Parallel.ForEach** method replaces the **foreach** operator.

Prediction with PLINQ

You can also use PLINQ to express a parallel loop. Here is an example.

```
static void UpdatePredictionsPlinq(AccountRepository accounts)
{
  accounts.AllAccounts
  .AsParallel()
  .ForAll(account =>
  {
    Trend trend = SampleUtilities.Fit(account.Balance);
    double prediction = trend.Predict(
                        account.Balance.Length + NumberOfMonths);
    account.PlinqPrediction = prediction;
    account.PlinqWarning = prediction < account.Overdraft;
   });
}
```

Using PLINQ is almost exactly like using LINQ-to-Objects and LINQ-to-XML. PLINQ provides a **ParallelEnumerable** class that defines extension methods for various types in a manner very similar to LINQ's **Enumerable** class. One of the methods of **ParallelEnumerable** is the **AsParallel** extension method.

The **AsParallel** extension method allows you to convert a sequential collection of type **IEnumerable<T>** into an **IParallelEnumerable<T>** object. Applying **AsParallel** to the **accounts.GetAccountsList** collection returns an object of type **IParallelEnumerable<AccountRecord>**.

PLINQ's **ParallelEnumerable** class has more than a hundred extension methods that provide parallel queries for **IParallelEnumerable<T>** objects. In addition to parallel implementations of LINQ methods such as **Select** and **Where**, PLINQ provides a **ForAll** extension method that invokes a delegate method in parallel for each element.

In the PLINQ prediction example, the argument to **ForAll** is a lambda expression that performs the credit analysis for a specified account. The body is the same as in the sequential version.

Performance Comparison

Running the credit review example shows that on a dual-core computer, the **Parallel.ForEach** and PLINQ versions run slightly less than twice as fast as the sequential version.

Variations

The credit analysis example is a typical use case for parallel loops, but there can be variations. This section introduces some of the most important variations of the parallel loop pattern. These discussions are condensed summaries of topics presented in a paper by Stephen Toub [Toub09]. See the references section of this chapter for more information.

Breaking Out of Loops Early

Breaking out of loops is a familiar pattern in sequential iteration. Here is a basic example.

```
for (int i = 0; i < N; i++)
{
  // ... do some work ...
  if (/* condition is true */)
    break;
}
```

The situation is more complicated with parallel loops because more than one iteration may be active at the same time, and partitioning strategies may result in iterations that are not necessarily executed in order. Partitioning refers to the process of assigning the work to be done to available cores.

To address these scenarios, the **Parallel.For** method has an overload that provides a **ParallelLoopState** object as a second argument to the loop body. You can ask the loop to break by calling the **Break** method of the **ParallelLoopState** object. Here is an example.

```
Parallel.For(0, N, (i, loopState) =>
{
  // ... do some work ...
  if (/* stopping condition is true */)
  {
    loopState.Break();
    return;
  }
}
```

Calling the **Break** method of the **ParallelLoopState** object begins an orderly shutdown of the loop processing. Any iterations that are running as of the call to **Break** will run to completion. However, you may want to check for a break condition in long-running loop bodies and exit that iteration immediately if a break was requested.

To see if another iteration running in parallel has requested a break, retrieve the value of the parallel loop state's **IsStopped** property. A value of **true** indicates that a break has been requested.

During the processing of a call to the **Break** method, iterations with an index value less than the current index will be allowed to start (if they have not already started), but iterations with an index value greater than the current index will not be started. This ensures that all iterations below the break point will be completed.

Because of parallel execution, it is possible that more than one iteration may call **Break**. In that case the lowest index will be used to determine which iterations will be allowed to start after the break occurred.

The **Parallel.For** method returns an object of type **ParallelLoopResult**. You can find out if a loop terminated with a break by examining the values of two of the loop result properties. If the **IsCompleted** property is **false** and the **LowestBreakIteration** property returns an object whose **HasValue** property is **true**, then you know that the loop terminated by a call to the **Break** method. You can query for the specific index using the loop result's **LowestBreakIteration** property. Here is an example:

```
int N = ...
var result = new double[N];

var loopResult = Parallel.For(0, N, (i, loopState) =>
{
   if (/* stopping condition is true */)
   {
      loopState.Break();
      return;
   }
   result[i] = DoWork(i);
});

if (!loopResult.IsCompleted &&
       loopResult.LowestBreakIteration.HasValue)
{
   Console.WriteLine("Loop encountered a break at {0}",
                     loopResult.LowestBreakIteration.Value;
}
```

The **Break** method ensures that data up until a particular iteration index value will be processed. Depending on how the iterations are scheduled, it may be possible that some iterations with a higher index value may have been started before the call to **Break** occurs.

There are also cases such as unordered searches where you want the loop to stop as quickly as possible after the stopping condition is met. The difference between "break" and "stop" is that with stop no attempt is made to execute loop iterations less than the stopping index if they have not already run. To stop a loop in this way, call the **ParallelLoopState** class's **Stop** method instead of the **Break** method. Here is an example of how to test to see if a stop occurred.

```
if (!loopResult.IsCompleted &&
    !loopResult.LowestBreakIteration.HasValue)
{
   Console.WriteLine("Loop was stopped");
```

```
}
```

The index value of the iteration that caused the stop is not available.

You cannot call both **Break** and **Stop** during the same parallel loop. You have to choose which of the two loop exit behaviors you want to use. If you do call both **Break** and **Stop** in the same parallel loop, an exception will be raised.

External Loop Cancellation

In some scenarios you may want to cancel a parallel loop from the outside, for example, in response to a request from the user interface. In this variation, use a cancellation token derived from a **CancellationTokenSource** instance. Here is an example.

```
void DoLoop(CancellationTokenSource cts)
{
  CancellationToken token = cts.Token;

  var options = new ParallelOptions
                        { CancellationToken = token };

  try
  {
    Parallel.For(0, N, options, (i) =>
    {
      // ... do some work ...

      // ... optionally check to see if cancellation happened
      if (token.IsCancellationRequested)
      {
         // ... optionally exit this iteration early
         return;
      }
    });
  }
  catch (OperationCanceledException ex)
  {
     // ... handle loop cancellation
  }
}
```

When the caller of the **DoLoop** method is ready to cancel, it invokes the **Cancel** method of the **CancellationTokenSource** that was provided as an argument to the **DoLoop** method. The parallel loop will allow currently running iterations to complete and then throw an **OperationCanceledException**. No new iterations will be started after cancellation begins.

If external cancellation has been signaled *and* your loop has called either the **Break** or the **Stop** method of the **ParallelLoopState** object, then a race occurs to see which will be recognized first. The parallel

loop will either throw an **OperationCanceledException** or it will terminate using the mechanism described in the previous section, but not both.

Exception Handling

If an iteration of the loop throws an unhandled exception, a parallel loop no longer begins any new iterations. By default, iterations that are executing at the time of the exception, other than the iteration that threw the exception, will be allowed to complete. When they have finished, the parallel loop will throw an exception in the context of the thread that invoked it.

Long-running iterations may want to test to see if an exception is pending in another iteration. They can do this with the **ParallelLoopState** class's **IsExceptional** property. This property returns **true** if an exception is pending.

Because more than one exception may occur during parallel execution, exceptions are grouped using an exception type called an **AggregateException**. The **AggregateException** class has an **InnerExceptions** property that contains all of the exceptions that occurred during the execution of the parallel loop.

Exceptions take priority over external cancellations and terminations of a loop initiated by calling the **Break** or **Stop** methods of the **ParallelLoopState** object.

Special Handling of Small Loop Bodies

If the body of the loop is very small and each iteration is expected to take the same amount of time, you may find that you achieve better performance by partitioning the iterations into larger units of work.

The reason for this is that there are two types of overhead that are introduced when processing a loop: the cost of synchronizing between worker threads and the cost of invoking a delegate method. In most situations these are negligible, but with very small loop bodies they can be significant.

Partitioning divides data into sets of non-overlapping regions called partitions; partitions are allocated to available processors.

The partition-based syntax is more complicated than standard parallel loop implementations, and when the amount of work in each iteration is large (or of uneven size across iterations), it may not result in better performance. Generally, you would only use the more complicated syntax after profiling or in the case that loop bodies are extremely small and the number of iterations large. Here is an example.

```
int[] result = new int[N];
Parallel.ForEach(Partitioner.Create(0, N),
    (range) =>
    {
       for (int i = range.Item1; i < range.Item2; i++)
       {
         // very small, equally sized blocks of work
         result[i] = i * i;
       }
    });
```

The **Partitioner** class has several static methods that create objects in order to control how parallel loops subdivide their work. In this example you can think of the result of the **Create** method as an object that acts like an instance of **IEnumerable<Tuple<int, int>>**. In other words, **Create** returns a collection of tuples (unnamed records) with two integer field values. Each tuple represents a range of values that should be processed in a single iteration of the parallel loop. By grouping iterations into ranges, you can avoid some of the overhead of a normal parallel loop. The number of ranges that will be created depends on the number of cores in your machine. The default number of ranges is approximately four times the number of cores.

If you know how big you want your ranges to be, you can use a special overload of the **Partitioner.Create** method that allows you to specify the size of each range. Here is an example.

```
int[] result = new int[1000000];
Parallel.ForEach(Partitioner.Create(0, 1000000, 50000),
    (range) =>
    {
       for (int i = range.Item1; i < range.Item2; i++)
       {
         // small, equally sized blocks of work
         result[i] = i * i;
       }
    });
```

In this example, each range will span 50,000 index values. In other words, for a million iterations the system will use twenty parallel iterations (1,000,000/ 50,000). These iterations will be spread out among all the available cores.

Choosing a suitable partitioning strategy may make it possible to see performance improvements even with extremely small loop bodies, as long as the number of iterations is high enough. It's possible that very small loop bodies with a small number of iterations may not benefit from parallel execution.

Custom partitioning is an extension point in the API for parallel loops. You can implement your own partitioning strategies. This is covered in more detail in the Implementation Details section of this chapter.

Small Numbers of Iterations

If you have a very small number of iterations, you may want to invoke them in parallel as a list. Here is an example.

```
Parallel.Invoke(() => DoWork(0),
                () => DoWork(1),
                () => DoWork(2));
```

The **Parallel.Invoke** method takes a **params** array of **Action** delegates and invokes each in parallel. The method returns when all of the actions have completed.

Controlling the Degree of Parallelism

Although it is usually recommended to let the system manage how iterations of a parallel loop are mapped to your computer's cores, in some cases you may want to specify how many threads should be used to process a particular parallel loop. Reducing the degree of parallelism is often used in performance testing to simulate less capable hardware. Increasing the degree of parallelism to a number larger than the number of cores can be appropriate when iterations of your loop spend a lot of time waiting for I/O operations to complete.

You can control the maximum number of worker threads used by specifying the **MaxDegreeOfParallelism** property of a **ParallelOptions** object. Here is an example.

```
var options = new ParallelOptions()
                      { MaxDegreeOfParallelism = 2};
Parallel.For(0, N, options, i =>
{
  // ... do some work ...
}
```

This loop will run using at most two worker threads.

You can also do this for PLINQ queries by setting the **WithDegreeOfParallelism** property of a **ParallelQuery<T>** object. Here is an example.

```
IEnumerable<T> myCollection = // ...
myCollection.AsParallel()
            .WithDegreeOfParallelism(8)
            .ForAll(obj => /* do work */);
```

This query will run with a maximum of eight worker threads.

If you specify a larger degree of parallelism you may also want to use the **ThreadPool** class's **SetMinThreads** method so that these threads are created without delay. If you don't do this, the thread pool's thread injection algorithm may limit how quickly threads can be added to the pool of worker threads that is used by the parallel loop. It may take more time than you want to create the required number of threads.

Anti-pattern: Be careful if you use parallel loops for I/O-bound workloads. If the I/O wait times are long, you may experience an unbounded growth of worker threads due to a hill-climbing heuristic used by the .NET **ThreadPool** class's thread injection logic. This heuristic steadily increases the number of worker threads when the threads of the current pool are blocked for long periods of time. You can limit the degree of parallelism as a way to prevent this from happening.

Note: The **Parallel** class and PLINQ work on slightly different threading models in the .NET 4 Framework. PLINQ uses a fixed number of threads to execute a query; by default, it uses the number

of logical cores in the machine, or it uses the value passed to the **WithDegreeOfParallelism** method if one was specified.

Conversely, by default the **Parallel.ForEach** and **Parallel.For** methods can use a variable number of threads. The number of threads may grow if some iterations take a long time.

Anti-pattern: Arbitrarily increasing the degree of parallelism puts you at risk of *processor oversubscription*, a situation that occurs when there are many more compute-intensive worker threads than there are cores. Tests have shown that in general the optimum number of worker threads for a parallel task equals the number of available cores divided by the average fraction of CPU utilization per task. In other words, with four cores and 50% average CPU utilization per task, you need eight worker threads for maximum throughput. Increasing the number of worker threads above this number begins to incur extra cost from additional context switching without any improvement in processor utilization.

On the other hand, arbitrarily restricting the degree of parallelism can result in *processor undersubscription*. Having too few tasks misses an opportunity to effectively use the available processor cores. You might restrict the degree of parallelism if you know that other tasks in your application are also running in parallel.

In most cases, the built-in load balancing algorithms of the .NET Framework are the most effective way to manage these tradeoffs. They coordinate resources among parallel loops and other tasks that are running concurrently.

Anti-pattern: Be extremely careful about specifying an increased degree of parallelism in server applications, such as those running on ASP.NET. In the server applications, multiplying the number of threads needed by the thousands of users sending incoming requests may overload even the most powerful server.

Using Thread-Local State in a Loop Body

Occasionally you will need to maintain thread-local state during the execution of a parallel loop. For example, you might want to use a parallel loop to initialize each element of a large array with random values. The .NET Framework's **Random** class does not support multi-threaded access. Therefore, you need a separate instance of the random number generator for each thread. Here is an example of how to do this using one of the overloads of the **Parallel.ForEach** method. The example uses a **Partitioner** object to decompose the work into relatively large chunks, since the amount of work performed by each step is small and there are a large number of steps.

```
int numberOfSteps = 10000000;
double[] result = new double[numberOfSteps];

// ForEach<TSource, TLocal>(
```

```
//        OrderablePartitioner<TSource> source,
//   ParallelOptions parallelOptions,
//   Func<TLocal> localInit,
//   Func<TSource, ParallelLoopState, long, TLocal, TLocal> body,
//   Action<TLocal> localFinally);
```

```
Parallel.ForEach(
     // source
     Partitioner.Create(0, numberOfSteps),

     // parallelOptions
     new ParallelOptions(),

     // localInit
     () => { return new Random(); },

     // body
     (range, loopState, _, random) =>
     {
       for (int i = range.Item1; i < range.Item2; i++)
         result[i] = random.NextDouble();
       return random;
     },

     // localFinally
     null);
```

The **Parallel.ForEach** loop will create a new instance of the **Random** class for each of its worker threads. This instance will be passed as an argument to each partitioned iteration. Each partitioned iteration is responsible for returning the next value of the thread-local state. In this example, the value is always the same object.

> **Anti-pattern**: Be careful that your loop body does not contain hidden dependencies. The example of trying to share an instance of a class such as **Random** that is not thread safe across parallel iterations is an example of a subtle dependency.

Using a Custom Task Scheduler for a Parallel Loop

In some cases you may want to substitute custom task scheduling logic for the default task scheduler that uses **ThreadPool** worker threads. This variation occurs, for example, when Single-Threaded Apartment (STA) threads, such as those used to invoke legacy COM components, must be used to perform the body of the parallel loop. By default, Multithreaded Apartment (MTA) worker threads from the .NET thread pool are used to execute parallel loops. You could also use this variation as a way to ensure a processor or node affinity on computers with non-uniform memory architecture (NUMA). These are just a few of the advanced scenarios where a custom task scheduler might be needed.

To specify a custom task scheduler set the **TaskScheduler** property of a **ParallelOptions** object. Here is an example.

```
TaskScheduler myScheduler = ...
var options = new ParallelOptions()
                      { TaskScheduler = myScheduler};
Parallel.For(0, N, options, i =>
{
  // ... do some work using tasks managed by myScheduler ...
}
```

For more information on task schedulers, see the Implementation Notes section in chapter 3, Parallel Tasks.

> **Note:** It isn't possible to specify a custom task scheduler for PLINQ queries. If you need a custom scheduler for a PLINQ query, you must implement a custom class that derives from **ParallelQuery<T>**.

Mixing the Parallel Class and PLINQ

PLINQ queries are instances of the **ParallelQuery<T>** class. This class provides the **IEnumerable<T>** interface, so it is possible to use a PLINQ query as the source collection for a **Parallel.ForEach** loop. This is not recommended.

Instead, you should use PLINQ's **ForAll** extension method for the **ParallelQuery<T>** instance. PLINQ's **ForAll** extension method performs the same kind of iteration as **Parallel**.**ForEach**, but it avoids wasteful merge and repartition steps that would otherwise be required in this case.

Here is an example of how to use the **ForAll** extension method.

```
var q = (from d in data.AsParallel() ... select d => F(d));
q.ForAll(item =>
{
  // ... Process item
});
```

Parallel Loops with Custom Iteration

Sometimes you want to apply a parallel loop to data structures that do not have standard iterators. For example, consider a binary tree.

```
class Tree<T>
{
  public Tree<T> Left, Right;
  public T Data;
}
```

You can implement a custom iterator for the **Tree<T>** class.

```
public static IEnumerable<Tree<T>> Iterate<T>(Tree<T> root)
{
    var queue = new Queue<Tree<T>>();
    queue.Enqueue(root);
    while (queue.Count > 0)
```

```
    {
        var node = queue.Dequeue();
        yield return node;
        if (node.Left != null) queue.Enqueue(node.Left);
        if (node.Right != null) queue.Enqueue(node.Right);
    }
}
```

The custom iterator is a powerful addition to the **Parallel.ForEach** method.

```
Tree<T> myTree = ...

Parallel.ForEach(Iterate(myTree), node =>
{
   // ... process node in parallel
});
```

Note: A more advanced variation would create partitions based on subtrees. For example, you could implement a **Partitioner** object that divided the tree into subtrees whose roots corresponded to nodes of a certain depth in the original tree. This would be especially efficient if, for example, the memory locality of subtrees improved memory cache performance.

Overriding the Default Behavior of IList<T>

In certain rare cases, the parallel loop's default handling of the **IList<T>** type may not be what you want. This can occur when the **IList<T>** implementation has unfavorable random-access performance characteristics or when random access causes a list with lazy-loading semantics to load all list values into memory.

This variation shows you how to override **Parallel.ForEach**'s default handling of a source that provides the **IList<T>** interface.

The **Parallel.ForEach** method requires its sources to provide the **IEnumerable<T>** interface; however, it also checks to see if its source provides the **IList<T>** interface. In most cases, using **IList<T>** to access elements of the collection will result in more efficient partitioning strategies, since it provided random (that is, indexed) access to the items in the collection. In contrast, **IEnumerable<T>** only supports access by walking the collection using the **MoveNext** method to retrieve successive elements. In almost all cases, the default behavior results in better performance.

A few types that provide **IList<T>** do so in a way that makes indexing an expensive operation. For these types, **MoveNext** is a better accessor. You can use a **Partitioner** object to force **Parallel.ForEach** to use the **IEnumerable<T>** interface, even if **IList<T>** is available. Here is an example.

```
IEnumerable<T> source = ...;

// Will always use source's IEnumerable<T> implementation.
Parallel.ForEach(Partitioner.Create(source),
                 item => { /*... do work ... */ });
```

Note: The **System.Data.Linq.EntitySet<TEntity>** class is an example of a type that shows better performance if **IEnumerable<T>** is used for parallel iteration. This is due to the type's lazy loading semantics.

Collections with Thread Affinity Requirements

In some rare cases a collection's implementation of the **MoveNext** method may have *thread affinity*. This means that the **MoveNext** method must always be called from a specific thread such as a UI thread. This situation can arise with objects provided by Windows Forms or Windows Presentation Foundation or with objects that pull data from the object model of a Microsoft Office application.

Parallel loops, including PLINQ queries, run by default in worker threads of the .NET Framework thread pool. However, there is a way to marshal elements of the collection from the required threading context to the threads that are executing the loop iterations. Here is an example:

```
// run from thread that is allowed to call MoveNext on source
static void ForEachWithEnumerationOnMainThread<T>(
                          IEnumerable<T> source, Action<T> body)
{
  var collectedData = new BlockingCollection<T>();
  var loop = Task.Factory.StartNew(() =>
                Parallel.ForEach(
                    collectedData.GetConsumingEnumerable(),
                    body));
  try
  {
    foreach (var item in source) collectedData.Add(item);
  }
  finally { collectedData.CompleteAdding(); }
  loop.Wait();
}
```

The **Parallel.ForEach** loop and the sequential **foreach** loop execute concurrently. The two loops are connected by a special concurrent collection that pipes values from one thread to another. The parallel loop may take advantage of more than one worker thread. The sequential **foreach** loop will only run in the same thread as the method that calls it does.

This variation will only run faster if the body of the **Parallel.ForEach** loop is significantly slower than the **MoveNext** method of the collection. Normally, this will be the case, but you should not expect the same level of performance as you would if **MoveNext** were allowed to be called from within the parallel loop.

Design Notes

The parallel loop pattern emphasizes decomposition by data, not tasks. It is designed to scale to any number of cores.

The implementations of the parallel loop pattern in the .NET Framework's parallel extensions and in the C++ Parallel Patterns Library contain sophisticated algorithms for dynamic work distribution. These loops automatically adapt to the workload and to a particular machine.

The parallel loop pattern expresses only potential parallelism, but does not guarantee it. This allows performance to scale from single-core scenarios to many cores.

Mechanisms, such as locks, are still needed to protect concurrent modifications for programs that use shared memory. In general, avoid writing to variables that are shared by iterations of the loop. When this is unavoidable, overloads of the **Parallel.For** and **Parallel.ForEach** methods offer abstractions such as thread-local state and a thread finalizing phase that help with synchronization.

Some parallel loops may be more compute-bound than others. The .NET Framework adapts dynamically to this situation by allowing the number of worker threads for a parallel loop to change over time. The load balancing algorithm also uses a partitioning technique where the size of units of work increases over time. This approach processes both small and large ranges with minimal overhead.

Implementation Notes

[TBD – will include description of writing a custom partitioner, approximately 3 pages]

Related Patterns

Related patterns:

> **Note:** The parallel loop pattern is sometimes called the map pattern, especially when the operation returns a value. In this case, the elements of the input collection are "mapped" or transformed into other values. The use of the term map is especially common in functional languages such as F#.

SPMD – Distributed systems (MPI, Batch & SOA)

Master/Worker – Task centric (TPL or PPL)

Fork/Join – Thread centric (TPL or PPL)

Data Parallelism – Algorithm strategy pattern

Exercises

1. Which of the following problems could be solved using the parallel loop techniques taught in this chapter? Sorting an array of numbers. Putting the words in each line in a text file in alphabetic order. Adding together all the numbers in one collection, to obtain a single sum. Adding numbers from two

collections pairwise, to obtain a collection of sums. Counting the total number of occurrences of each word in a collection of text files. Finding the word that occurs most frequently in each file in a collection of text files.

2. Choose a suitable problem from exercise 1. Code three solutions, using a sequential loop, a parallel loop, and PLINQ.

3. In the credit review example, what is the type of account ? What is the type of account.AllAccounts ? Of account.AllAccounts.AsParallel() ?

4. Is it possible to use a PLINQ query as the source for a Parallel.ForEach loop? Is this recommended? Explain your answers.

5. Do a performance analysis of the credit review example code on the Codeplex site. Use command line options to independently vary the number of iterations (the number of accounts) and the amount of work done in the body of each iteration (the number of months in the credit history). Record the execution times reported by the program for all three versions, using several different combinations of numbers of accounts and months. Repeat the tests on different computers with different numbers of cores, and with different execution loads (from other jobs).

6. Modify the credit review example from Codeplex so you can set MaxDegreeOfParallelism. Observe the effect of different values for this option on execution time, when running on machines with different numbers of cores.

7. When is it appropriate to write a custom partitioner? (Hint: review the binary tree example).

Further Reading

The examples in this book use features and libraries of recent C# versions. Some helpful books are Bishop (2008) and Albahari and Albahari (2010).

The book by Mattson, et al. (2004) is a survey of software design patterns for parallel programming that is not specialized for a particular language or library.

Many of the cases in the Variations section of this chapter are from Toub (2009).

3 Parallel Tasks

THIS CHAPTER IS TBD

4 Parallel Aggregation with Map-Reduce

THIS CHAPTER IS TBD

5 Futures and Continuations

In chapter 3 you saw how the parallel task pattern allowed you to fork the flow of control in a program. In this chapter, you will see how control flow and data flow can be integrated with two patterns known as futures and continuations. These patterns allow you to schedule tasks that can accommodate data flow constraints.

A *future* is a task that returns a value. Instead of explicitly waiting for the task to complete, you simply ask the task for its result when you are ready to use it. If the task has already finished by the time you do this, then its result is waiting for you and is immediately returned. If the task is running but has not yet finished, then the current thread blocks until the result value becomes available. If the task has not yet started, the task will be executed in the current thread context. In other words, querying for the result of a future integrates asynchronous control flow with data flow. Futures are also a good way of expressing the principle of potential parallelism: decomposing a sequential operation with futures may result in much faster execution, and in the worst case will be no slower than the serial case.

A future is a task that returns a value. Futures are implemented in the .NET Framework by the **Task<T>** class.

The **Task<T>** class in the .NET Framework supports the futures pattern. The type parameter **T** gives the type of the task's result. If the parallel tasks described in chapter 3 are asynchronous *actions*, then futures are asynchronous *functions*. (Recall that actions don't return values but functions do.)

C++ Note: PPL does not provide a futures implementation. However, the documentation in MSDN shows you how to use tasks to implement this pattern for use in your own code.

Anti-pattern: As described in chapter 3 Parallel Tasks, tasks, including futures, cannot always be run inline. In these cases, , the latency of a task-based implementation will be more variable than a sequential implementation due to the strict first-in first-out scheduling of tasks that may not be inlined. There are workarounds available if you experience unacceptable scheduling latency with futures. See the "Implementation Details" section of chapter 3, "Parallel Tasks" for more information about task inlining and the scheduling techniques used by the default task scheduler.

A *continuation*, or *continuation task*, is a task that automatically starts when another task, known as its antecedent, completes. In most cases the antecedent is a future whose result value is used as input to the continuation. An antecedent may have more than one continuation, and a continuation may have more than one antecedent.

A continuation is a task that automatically starts when a specified antecedent task finishes.

Continuations represent the nested application of asynchronous functions. In some ways, continuations are like callback methods—in both cases you register an operation that will automatically be invoked at a specified point in the future.

Unlike other kinds of tasks, continuations are never inlined into the current thread context.

The Problem

The futures and continuations patterns are the solution to some very common problems. When you think about the parallel task pattern described in chapter 3, you see that in many cases the purpose of a task is to calculate a result. In other words, asynchronous operations often act like functions. Of course, tasks can also do other things, such as reordering values in an array, but calculating new values is common enough to warrant a pattern tailored for it.

Futures

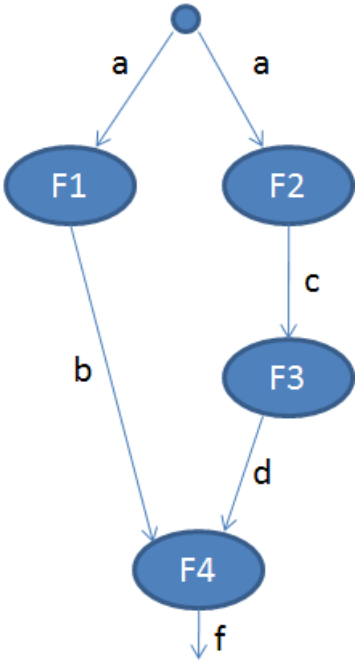
To better understand the problem, consider the following example, taken from the body of a sequential method:

```
var b = F1(a);
var c = F2(a);
var d = F3(c);
var f = F4(b, d);
return f;
```

Suppose that the functions **F1**, **F2**, **F3** and **F4** are all CPU-intensive computations that interact with each other using arguments and return values instead of updating shared variables in memory.

You want to distribute the work across the available CPUs, and you want your code to run correctly no matter how many CPUs are available on your platform. When you study the example, you can see that **F1** can run in parallel with **F2** and **F3** but that **F3** can't start until **F2** is finished. The possible orderings become apparent when you look at the function calls as a graph.

A Task Graph for Calculating "f"



The nodes of the graph are the functions **F1**, **F2**, **F3** and **F4**. The incoming arrows for each node are the inputs required by the function, and the outgoing arrows are values calculated by each function.

Futures and continuations are an easy way to introduce asynchrony that is constrained by these kinds of data flow dependencies. Here's an example that shows how it works. The code assumes for simplicity that the values being calculated are integers and that the value of variable **a** has already been supplied, perhaps as an argument to the current method.

```
Task<int> futureB = Task<int>.Factory.StartNew(() => F1(a));
int c = F2(a);
int d = F3(c);
int f = F4(futureB.Result, d);
return f;
```

The **Result** property returns a pre-calculated value immediately or waits until the value becomes available.

This code creates a future that begins to asynchronously calculate the value of F1(a). On a multicore system **F1** will run in parallel with the current thread. This means that **F2** can begin executing without waiting for **F1**. The function **F4** will execute as soon as the data it needs becomes available. It doesn't matter whether **F1** or **F3** finishes first, since the results of both functions are required before **F4** can be invoked. (Recall that the **Result** property does not return until the future's value is available.) Note that the calls to **F2**, **F3** and **F4** do not need to be wrapped inside of a future, since a single additional asynchronous operation is all that is needed to fully exploit the potential parallelism of this example.

Of course, you could equivalently have put **F2** and **F3** inside of a future, as shown here.

```
Task<int> futureD = Task<int>.Factory.StartNew(
```

```
                                                () => F3(F2(a)));
  int b = F1(a);
  int f = F4(b, futureD.Result);
  return f;
```

It doesn't matter which branch of the task graph shown in the figure is run asynchronously.

An important point of this example is that exceptions that occur during the execution of a future will be thrown by the **Result** property. This makes exception handling easy, even in cases with many futures and complex chains of continuations.

Futures defer exceptions until the **Result** property is read.

Continuations

Another very common case occurs when one asynchronous operation invokes a second asynchronous operation and passes data to it. This is described by the continuations pattern.

For example, if you wanted to update the user interface with the result produced by the function **F4** from the previous section, you could use the following code.

```
TextBox myTextBox = ...;

var futureB = Task.Factory.StartNew<int>(() => F1(a));
var futureD = Task.Factory.StartNew<int>(() => F3(F2(a)));

var futureF = Task.Factory.ContinueWhenAll <int, int>(
                 new[] { futureB, futureD },
                 (tasks) => F4(futureB.Result, futureD.Result));

futureF.ContinueWith((t) =>
  myTextBox.Dispatcher.Invoke(
       (Action)(() => { myTextBox.Text = t.Result.ToString(); }))
            );
```

This code structures the computation into four tasks, two futures and two continuations.

The **ContinueWhenAll<T, S>** method of the **Task.Factory** object allows you to create a continuation that depends on more than one antecedent task. The **ContinueWith** method creates a continuation task with a single antecedent. The system understands the ordering dependencies between continuations and their antecedent tasks. It makes sure that continuations will only be started after their antecedent tasks have completed.

The first task calculates the value of **b**. The second task calculates the value of **d**. These two tasks may run in parallel. The third task calculates the value of **f**. It may run only after the first two tasks are complete. Finally, the fourth task takes the value calculated by **F4** and updates a text box on the user interface.

The Adatum Financial Dashboard

Let's look at an example of how the futures and continuations patterns can be used in an application. The example shows how you can run computationally intensive operations in parallel in an application that uses a graphical user interface (GUI).

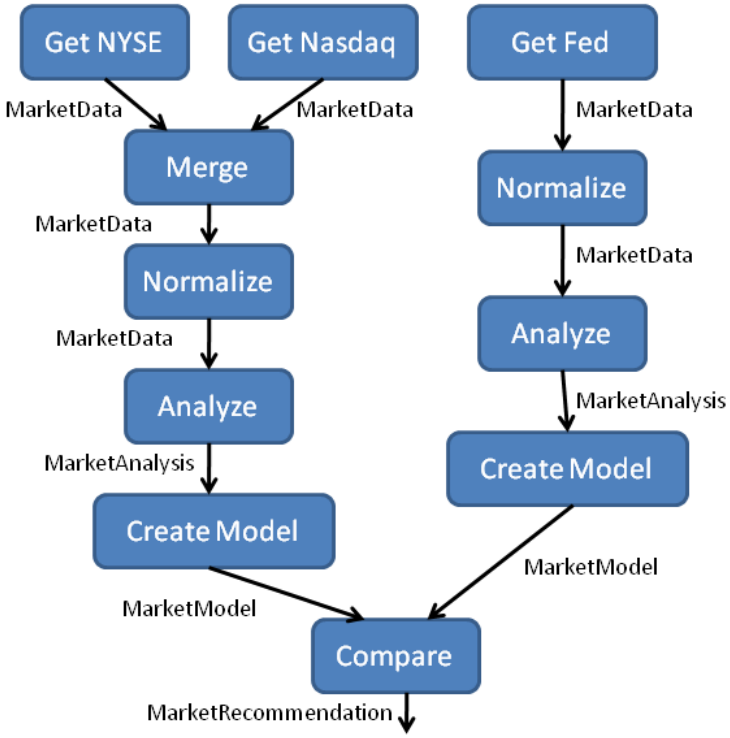
Adatum is a financial services firm that provides a financial dashboard application to its employees. The application, known as the Adatum dashboard allows employees to perform analyses of financial markets. The dashboard application runs on an employee's desktop workstation. The analyses it performs are computationally intensive, but there is also some I/O latency, since the Adatum Dashboard application collects input data from several sources over the network. Employees expect the application to remain responsive regardless of computational load and I/O latencies. Blocking the UI for more than a fraction of a second would be unacceptable.

The application gathers market data from several sources and then merges the data sets together. The application normalizes the merged market data and then performs an analysis step. After the analysis, it creates a market model. It also performs these same steps for Fed historical market data. After the current and historical models are ready, the application compares the two models and makes a market recommendation of "buy," "sell" or "hold."

> **Note:** The Adatum dashboard analyzes historical data, not a stream of real-time price data.

You can visualize these steps as a graph. This is shown in the following diagram.

Adatum Dashboard Tasks



The tasks in this diagram communicate by specific types of business objects, such as **MarketData**, **MarketAnalysis**, **MarketModel** and **MarketRecommendation**. These are implemented as .NET classes in the Adatum dashboard application.

> **Note:** You can download the complete source code for the Adatum dashboard application from the CodePlex site http://parallelpatterns.codeplex.com. The application consists of three parts: the business object definitions, an analysis engine and the user interface.

The Dashboard's Analysis Engine

The Adatum dashboard's **AnalysisEngine** class could use either parallel or sequential modes of execution to produce a market recommendation.

The sequential process is shown in the following code.

```
public MarketRecommendation DoAnalysisSequential()
{
   return CompareModels(
          new[]
          {
               CreateModel(
```

```
                    AnalyzeData(
                       NormalizeData(
                           MergeMarketData(
                              new [] { LoadNyseData(),
                                       LoadNasdaqData() }))),
                CreateModel(
                    AnalyzeData(
                       NormalizeData(
                              LoadFedHistoricalData())))
                      });
}
```

You can see how ordinary nested method invocations correspond to the data dependencies among the analysis phases shown in the figure. The result of the computation is a **MarketRecommendation** object. Each of the nested method calls returns data that becomes the input to the operation that invokes it. When you use method invocations in this way, you are limited to sequential execution.

The parallel version uses futures and continuations. It creates futures for each of the operational steps and connects them to each other with continuations. You can look online for the complete example, but here are some highlights.

```
public AnalysisTasks StartParallelAnalysis()
{
    // Create each task
    loadNyseData = Task<MarketData>.Factory.StartNew(...);

    // ...

    return new AnalysisTasks()
    {
        LoadNyseData = loadNyseData,
        LoadNasdaqData =  ... ,
        MergeMarketData =  ... ,
        NormalizeMarketData =  ... ,
        LoadFedHistoricalData =  ... ,
        NormalizeHistoricalData =  ... ,
        AnalyzeMarketData = ... ,
        ModelMarketData =  ... ,
        AnalyzeHistoricalData =  ... ,
        ModelHistoricalData =  ... ,
        CompareModels =  ...
    };
}
```

In the parallel case, the application uses an asynchronous operation implemented by the **StartParallelAnalysis** method.

The **StartParallelAnalysis** method returns an **AnalysisTasks** object that contains futures for each of the operations in the task graph. For example, the **CompareModels** property contains a future that returns

a market recommendation value. The future is implemented as a **Task<MarketRecommendation>** object. To get the final result of the analysis, use the following code.

```
CancellationTokenSource cts = new CancellationTokenSource();
CancellationToken token = cts.Token;
AnalysisEngine ae = new AnalysisEngine(token)
AnalysisTasks tasks = ae.StartParallelAnalysis();
MarketRecommendation recommendation = tasks.CompareModels.Result;
```

You don't need to wait until the final answer is ready before exploring some of the early results. If you are interested in examining partial results that might be available before the entire analysis finishes, you can do this. For example, you can examine the information produced by the task that loads the NYSE market data from the network using the following code.

```
MarketData nyseData = tasks.LoadNyseData.Result;
```

Now let's look at how each of the tasks is created.

Loading External Data

The methods that gather the external data from the network are long-running, I/O intensive operations. Unlike the other steps, they are not particularly CPU intensive. Most of their time is spent waiting for I/O operations to complete. You create these tasks using a factory object and you use an argument to specify that the tasks are I/O intensive and of long duration.

```
Task<MarketData> loadNyseData =
   Task<MarketData>.Factory.StartNew(
      () => LoadNyseData(),
      TaskCreationOptions.LongRunning);

Task<MarketData> loadNasdaqData =
   Task<MarketData>.Factory.StartNew(
      () => LoadNasdaqData(),
            TaskCreationOptions.LongRunning);
```

Note that the **Task<MarketData>.Factory** object creates futures that return values of type **MarketData**. The **TaskCreationOptions.LongRunning** enumerated value tells the task library that the operations are not CPU-intensive and are expected to run for a long time. To prevent underutilization of CPU resources, the task library may choose to run tasks like these in additional threads.

> Use "long running" tasks for tasks that are not CPU-intensive, such as long-running I/O operations.

Merging

The merge operation takes inputs from both the **loadNyseData** and the **loadNasdaqData** tasks. It is a continuation that depends on two antecedent tasks, as shown in the following code.

```
Task<MarketData> mergeMarketData =
    Task.Factory.ContinueWhenAll<MarketData, MarketData>(
        new[] { loadNyseData, loadNasdaqData },
```

```
        (tasks) => MergeMarketData(
            from t in tasks select t.Result));
```

The **ContinueWhenAll<T, S>** method of the **Task.Factory** object allows you to create a continuation that gets data from more than one antecedent task. Once the **loadNyseData** and **loadNasdaqData** tasks have completed, the anonymous delegate given as an argument is invoked. At that point the **tasks** parameter will be an array of antecedent tasks.

The **MergeMarketData** method takes an array of **MarketData** objects as its input. The LINQ expression **from t in tasks select t.Result** maps the input array of futures into an array of **MarketData** objects by getting the **Result** property of each future.

Normalizing

After the market data is merged, it undergoes a normalization step.

```
Task<MarketData> normalizeMarketData =
     mergeMarketData.ContinueWith(
       (t) => NormalizeData(t.Result));
```

The **ContinueWith** method creates a continuation task with a single antecedent. The continuation gets the result value from the task referenced by **mergeMarketData** variable and invokes the **NormalizeData** method.

Analysis and Model Creation

After the market data is normalized, the application performs an analysis step. This takes an object of type **MarketData** as input and returns an object of type **MarketAnalysis**.

```
Task<MarketAnalysis> analyzeMarketData =
    normalizeMarketData.ContinueWith(
      (t) => AnalyzeData(t.Result));

Task<MarketModel> modelMarketData =
    analyzeMarketData.ContinueWith(
      (t) => CreateModel(t.Result));
```

Analysis and model creation are two additional examples of continuations with a single antecedent task.

Processing Historical Data

The application also creates a model of historical data. The steps used to create the tasks are similar to those for current market data. However, because these steps are performed by tasks, they may be run in parallel if data dependencies and hardware resources allow it.

Comparing Models

```
Task<MarketRecommendation> compareModels =
   Task.Factory.ContinueWhenAll<MarketModel,
MarketRecommendation>(
        new[] { modelMarketData, modelHistoricalData },
```

```
        (tasks) => CompareModels(
           from t in tasks select t.Result));
```

The “compare models” task compares the current and historical market models and produces the final result.

The Dashboard’s User Interface

Futures and continuations are also used in the Adatum dashboard’s user interface. It uses the Windows Presentation Foundation (WPF) framework.

The Adatum dashboard user interface is designed so that the result of each analysis step can be viewed by the user as the computation progresses. There are individual buttons for each of the steps. As each result becomes available, the buttons are individually enabled. It is also possible to cancel the analysis from the user interface.

The application uses continuations instead of registering callbacks. This has the advantage of not requiring the analysis layer to refer to any types defined by the user interface.

Here’s a walkthrough of how the notification works.

```
public class MainWindowViewModel :
                            INotifyPropertyChanged, IDisposable
{
  // ...

  void OnRequestCalculate()
  {
     // ...

     this._cancellationTokenSource =
                          new CancellationTokenSource();
     CancellationToken token =
                        this._cancellationTokenSource.Token;

     // Start the analysis
     var analysisEngine = new AnalysisEngine(token);
     AnalysisTasks tasks =
                    analysisEngine.StartAnalysisParallel();

     AddButtonContinuations(tasks, token)
  }

  // ...
}
```

The user interface uses the Model View ViewModel (MVVM) pattern. The main window’s view model has a **Calculate** command that invokes the **OnRequestCalculate** method in response to a user-interface button press.

The cancellation token is passed as an argument to the data analysis object.

Rather than starting background tasks with a method such as **QueueUserWorkItem** from the **ThreadPool** class, the view model asks the analysis engine to create a record that contains tasks corresponding to each independently viewable analysis result.

This architecture demonstrates decoupling. The person who wrote the analysis layer of the application was able to do this without any knowledge of how other parts of the application would use the results of the analysis.

Next, the handler creates user interface-specific continuations with the **AddButtonContinuations** method. This is shown in the following code.

```
public void AddButtonContinuations(AnalysisTasks tasks)
{
  TaskScheduler s =
     TaskScheduler.FromCurrentSynchronizationContext();

  tasks.LoadNyseData.ContinueWith(
    (t) => { this.NyseMarketData = t.Result; }, s);
  tasks.LoadNasdaqData.ContinueWith(
    (t) => { this.NasdaqMarketData = t.Result; }, s);
  tasks.LoadFedHistoricalData.ContinueWith(
    (t) => { this.FedHistoricalData = t.Result; }, s);
  tasks.MergeMarketData.ContinueWith(
    (t) => { this.MergedMarketData = t.Result; }, s);
  tasks.NormalizeHistoricalData.ContinueWith(
    (t) => { this.NormalizedHistoricalData = t.Result; }, s);
  tasks.NormalizeMarketData.ContinueWith(
    (t) => { this.NormalizedMarketData = t.Result; }, s);
  tasks.AnalyzeHistoricalData.ContinueWith(
    (t) => { this.AnalyzedHistoricalData = t.Result; }, s);
  tasks.AnalyzeMarketData.ContinueWith(
    (t) => { this.AnalyzedMarketData = t.Result; }, s);
  tasks.ModelHistoricalData.ContinueWith(
    (t) => { this.ModeledHistoricalData = t.Result; }, s);
  tasks.ModelMarketData.ContinueWith(
    (t) => { this.ModeledMarketData = t.Result; }, s);
  tasks.CompareModels.ContinueWith(
    (t) => {
              this.Recommendation = t.Result;
              this.StatusTextBoxText =
                 (this.Recommendation == null) ?
                 "Canceled" : this.Recommendation.Recommendation;
              this.ModelState = State.Ready;
             }, s);

}
```

The method creates continuations that will automatically run *in the user-interface thread* after each of the tasks finishes and has results ready to view. When the final "compare models" task finishes, the user interface view model will be updated with the final recommendation. Callbacks registered with the view model will notify the user interface of the changes so that these changes will be reflected in the UI.

The result of the **FromCurrentSynchronizationContext** method of the **TaskScheduler** class is a **ThreadScheduler** object that will only allow its tasks to run in the current thread (that is, the user interface thread).

Continuations for User Interfaces

The Adatum dashboard application demonstrates how an application can use continuations to keep the user interface up to date.

Before adopting parallelism, Adatum used background worker threads to handle the computationally intensive parts of applications such as the Adatum dashboard. However, the Adatum dashboard application has some requirements, such as the use of WPF for the user interface, that make continuations more appropriate than background worker threads.

One of the reasons that the futures and continuations patterns works for the Adatum dashboard is because it satisfies the constraints of thread affinity. Some frameworks place this constraint on objects they expose. For example, in WPF you have to run all methods of a user-interface object on the same thread that you used to create that object.

The futures and continuations pattern makes it easy to deal with thread affinity constraints. Continuation tasks can be configured with a task scheduler that only runs the task on a chosen thread. Antecedents of the task do not need to run on the same thread as the continuation. This allowed Adatum's developers to distribute computationally intensive work among many CPUs while allowing the calculated values to appear in the user interface without violating WPF's thread affinity constraint.

Tasks make it easy to satisfy thread affinity constraints.

Task scheduling has been optimized for CPU-intensive operations. It is more efficient to run fewer threads than tasks. With futures and continuations, the amount of parallelism you can achieve is only limited by your design and the number of available CPUs. This eliminates a restriction of .NET 2.0 background worker threads which are limited to a single instance and cannot have data dependencies. Therefore, they are not appropriate when you are merging the results from multiple concurrent calculations.

Anti-pattern: When are continuations used rather than .NET Framework 2.0 background worker threads? Answer: A background worker thread is inappropriate when you are merging the results from multiple concurrent calculations. (Also, possibly, when there is more than one background task needed, regardless of merging.)

Modifying the Graph at Runtime

The code in the analysis engine creates a static task graph. In other words, the graph of task dependencies is reflected directly in the code. By reading the implementation of the analysis engine you can determine that there are a fixed number of tasks with a fixed set of dependencies among them.

The extension of the analysis tasks in the user interface layer is an example of dynamic task creation. This user interface augments the graphs of tasks by adding continuations programmatically, *outside of the context where these tasks were originally created*.

Dynamically created tasks are also a way to structure algorithms used for sorting, searching and graph traversal. See chapter 6, "Recursive Task Parallelism" for examples.

Support for Cancellation

The Adatum dashboard application supports cancellation from the user interface. It does this by calling the **Cancel** method of the **CancellationTokenSource** class. This sets the **IsCancellationRequested** property of the cancellation token to **true**.

The application checks for this condition at various checkpoints. If cancellation has been requested, the operation is aborted.

Variations

So far you've seen some of the most common ways to use futures and continuations to create tasks. Here are some other ways to use them.

Canceling Futures and Continuations

There are several patterns relating to the cancellation of futures and continuations. [TBD] You can handle cancellation entirely from within the task, as Adatum's dashboard does or you can pass cancellation tokens as an option when the tasks are created.

> **Anti-pattern**: If you pass a cancellation token as an argument to a task's constructor (or to the **StartNew** method of the **Task.Factory** object), you should also make sure that every continuation task is passed that same cancellation token when it is created. If you don't do this consistently, you may experience unhandled **OperationCancelled** exceptions.

Handling Exceptions in a Continuation

Using a continuation to handle exceptions thrown by the antecedent can be a useful pattern. Here is an example:

```
var t1 = Task<int>.Factory.StartNew(() => F(token));
```

```
var t2 = t1.ContinueWith<int>((t) =>
  {
    try
    {
       return t2.Result;
    }
    catch (AggregateException ae)
    {
       ae.Handle((e) =>
       {
         if (e is MyException)
         {
            Console.WriteLine("MyException caught: " +
e.Message);
            return true;
         }
         return false;
       });
    }
  });
var t3 = t2.ContinueWith((t) => G(t.Result));

return t3.Result;
```

In this example, the continuation **t2** passes the result of its antecedent to the next continuation, unless it catches an exception of type **MyException**.

Continue When "At Least One" Antecedent Completes

It is possible to invoke a continuation when the first of multiple antecedents completes. To do this, use the **Task<T>.Factory** object's **ContinueWhenAny** method.

```
var t1 = Task<int>.Factory.StartNew(F1);
var t2 = Task<int>.Factory.StartNew(F2);
var t3 = Task<int>.Factory.StartNew(F3);
var t4 = Task<string>.Factory.ContinueWhenAny(new[] { t1, t2, t3
},
                (t) => "The answer is" + t.Result.ToString());
Console.WriteLine(t4.Result);
```

This is useful when the result of any of the tasks will do. For example each task queries a web service which gives the local weather. The first answer is returned to the user.

Converting a .NET Asynchronous Call into a Future

It is possible to convert Asynch calls into futures. [TBD]

Futures and continuations are similar in some ways to asynchronous methods that use the .NET **IAsynchResult** interface. In general, futures are easier to use than **IAsynchResult**. See "Variations and Related Patterns" in this chapter for more information.

Futures are easier to use than **IAsynchResult** and benefit from easier exception handling.

Design Notes

There are several ideas behind the design of the Adatum dashboard application.

Decomposition into Futures and Continuations

The first design decision is the most obvious one: the Adatum dashboard introduces parallelism by means of futures and continuations. This makes sense because the problem space could be decomposed into operations with well-defined inputs and outputs.

Functional style

There are two approaches to synchronizing data between tasks. In this chapter, the examples have used an explicit approach. Data is passed between tasks as parameters, which makes the data dependencies very obvious to the programmer. Alternatively, as you saw in chapter 2, "Parallel Tasks", it is possible for tasks to communicate with side effects on the tasks could act on shared data structures. In this case, and rely on thettasks use control dependencies to block appropriately and control flowfor the (implied) data flow.

You can use control flow constraints and side effects with shared data structures for legacy applications and in cases. In general, explicit data flow is less prone to error.

You can see this by an analogy to functions and subroutines. There is no need for a programming language to support methods with return values. Programmers can always use methods without return values and perform updates on shared global variables as a way of communicating the results of a computation to other components of an application. However, in practice, return values are considered to be much less subject to error.

Similarly, using futures (tasks that return values) can reduce the possibility of error in a parallel program as compared to tasks that communicate results by a modified shared global state. Tasks that return values can often require less synchronization than tasks that globally accessible state variables.

Using a control flow approach removed the overhead of passing data between tasks but makes it much less clear as to what tasks are operating on what data.

The design of the Adatum Dashboard uses the functional style of programming with a focus on operations that communicate using input and output values. This is in contrast to programs that modify structures in place. The functional pattern is well suited to parallel programming. Functional programs are very easy to adapt to multicore environments. Let’s examine a few of the assumptions.

The first is a commitment to scalable sharing of data. This means that futures should generally only communicate with the outside world by means of their return values. In general, they should be as free as possible of side effects such as writes to mutable shared variables. If you do read and write to shared

variables you will need to serialize your program using synchronization objects or be extraordinarily careful when you write to shared state.

> Communicating among tasks by means of arguments and return values scales well as the number of cores increases.

It is also a good practice to use immutable types for return values. .NET strings are a good example of a complex type implemented as an immutable class.

Functional Programming using Value Types

If you use the futures and continuations pattern, you may want to use value types. [TBD]

Implementation Notes

The Adatum Dashboard example introduced **Task<T>** class (i.e., tasks that return values), **Task.ContinueWith**, and **Task.Factory.ContinueWhenAll** methods.

Callout: Comparison with existing .NET asynch UI programming patterns. [TBD]

Related Patterns

There are a number of variations and related patterns for the continuation pattern. [TBD]

Exercises

1. Given this sequential code: var b = F1(a); var d = F2(c); var e = F3(b,d); var f = F4(e); var g = F5(e); var h = F6(f,g); Suppose we will parallelize this code using futures in the style of the first example. Draw the task graph. In order to achieve the most possible concurrency, what is the minimum number of futures we must define? What is the largest number of these futures that can be running at the same time?

2. Modify the BasicFutures sample from Codeplex so that one of the futures throws an exception. What should happen? Observe the behavior when you execute the modified sample.

3. What is the thread affinity constraint imposed by the WPF user interface framework? In the dashboard sample, how does the AddButtonContinuations method satisfy this constraint?

Further reading

Leijen (2009) describes the motivation for including futures in TPL and has references to other work, especially in functional languages.

The NModel framework (2008) provides a C# library of immutable collection types including set, bag, sequence, and map.

6 Dynamic Task Parallelism

THIS CHAPTER IS TBD

7 Pipelines

THIS CHAPTER IS TBD

Appendix A: Supporting Patterns

THIS APPENDIX IS TBD

Appendix B: Debugging and Profiling Parallel Applications

THIS APPENDIX IS TBD

Appendix C: Technology Roadmap

THIS APPENDIX IS TBD

Appendix D: QuickStart Examples

THIS APPENDIX IS TBD

References

[1] Joseph Albahari and Ben Albahari. C# 4 in a Nutshell. O'Reilly, fourth edition, 2010.

[2] Judith Bishop. C# 3.0 Design Patterns. O'Reilly, 2008.

[3] Daan Leijen, Wolfram Schulte, and Sebastian Burckhardt. The design of a task parallel library. In Shail Arora and Gary T. Leavens, editors, OOPSLA 2009: Proceedings of the 24th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applicatons, pages 227{242. ACM, 2009.

[4] Timothy G. Mattson, Beverly A. Sanders, and Berna L. Massingill. Patterns for Parallel Programming. Addison-Wesley, 2004.

[5] NModel software, 2008. http://nmodel.codeplex.com/.

[6] Stephen Toub. Patterns of Parallel Programming: Understanding and Applying Parallel Patterns with the .NET Framework 4 and C#, 2009. Available from MSDN.