Diplomarbeit

Parallelisierung von ungeordneten, paarweisen Berechnungen

David Hofer, Technische Informatik, KT13wIA-F

Kurzfassung  
  
Diese Arbeit behandelt den empirischen Vergleich von Verteilungsmethoden für paarweise, ungeordnete Berechnungen. Durch die Verteilung der Berechnungen, soll eine möglichst gute Auslastung des unterliegenden Berechnungssystems erreicht werden.

Zuerst werden paarweise, ungeordnete Berechnungen im Detail beschrieben um dem Leser das notwendige Fundament zu liefern, auf dem die Arbeit aufbaut.

Darauf folgt eine Auswertung der grundlegenden Probleme bei Parallelisierungsvorgängen

Einleitung

Dieser Abschnitt wird sich damit befassen, Grundlagen zu dem Thema zu erläutern und die Motivation hinter dieser Arbeit zu erklären.   
Anschließend werden zu erwartende Probleme aufgezeigt und sowohl logische als auch strukturelle Herangehensweisen angeboten.   
Die Vorschläge werden nicht im Detail ausgearbeitet, sondern sollen ein besseres Verständnis für das Problem schaffen, auf dem im Laufe der Arbeit aufgebaut wird.  
Strukturelle Konzepte werden oberflächlich gehalten und sollen klären, auf welcher Abstraktionsebene die Lösungen angewendet werden. Es soll ersichtlich werden, wo die Grenzen zum ober- und unterliegenden System liegen.

Motivation

In der Prozessorindustrie zeichnet sich ein starker Trend ab. Anstelle von höheren Frequenzen, setzen Prozessorhersteller auf mehrere Rechenkerne innerhalb ihrer Prozessoren. Dies zwingt auch die Softwareentwicklung zu einem Umdenken.  
Parallelisierung von Standardproblemen sollte nicht mehr die Ausnahme, sondern die Regel sein. Dies ist jedoch oft aufwendig und Parallelisierung von Software kann, bei schlechter Umsetzung, mehr Probleme bereiten, als dadurch gelöst werden.   
Deshalb braucht es generische, skalierbare Lösungen, welche in dieser Arbeit untersucht werden sollen.

“… the rapid increase in single processor performance is likely a thing of the past. Future advancements in computing capability must come from harnessing more cores to a single solution.“

Grundlagen

Paarweise, ungeordnete Berechnungen

An dieser Stelle muss festgehalten werden, dass es sich um einen, vom Autor, gewählten Ausdruck handelt.

Die Art der Berechnung ergibt sich aus der abzählenden Kombinatorik, bei der für eine Anzahl n an Elementen einmalig Berechnungen zwischen den Elementen angestellt werden müssen.  
„Paarweise“ sagt hierbei aus, dass die Berechnung immer für zwei Elementen ausgeführt wird.  
„Ungeordnet“ steht für eine beliebige Reihenfolge der Berechnungen.

Beispiel: **Berechnungen zwischen 3 Elementen { A, B, C }**

Mögliche paarweise, ungeordnete Kombinationen und deren Berechnungen:

AB – Berechnung 1 zwischen den Elementen A und B  
AC – Berechnung 2 zwischen den Elementen A und C   
BC – Berechnung 3 zwischen den Elementen B und C  
  
Es ergeben sich somit **drei** paarweise, ungeordnete Berechnungen.

Für die ungeordnete Kombinatorik gilt folgende Formel für die Anzahl der Berechnungen

Für die paarweise Variante ist k immer 2

Begriffe  
  
Core

Als Cores werden jene Teile des Rechnersystems beschrieben, auf welche die Berechnungen im Endeffekt verteilt werden. Mehrere Cores bilden zusammen den Prozessor.

Dummy

Dummies ersetzen Teile des Systems, um spezielle Testroutinen durchführen zu können. Ein Dummy-Objekt kann entweder zusätzliche Funktionen bereitstellen, um zusätzliche Informationen bereitzustellen, oder besonders geringen Berechnungsaufwand benötigen, um Performancetests für andere Module zu ermöglichen.

Latenz

Die Verzögerungszeit zwischen den Cores. Dauert die Kommunikation, und damit auch die Synchronisation, zwischen den Cores lange, bedeutet dies eine hohe Latenz.

Gegenüberstellung der Algorithmen

Um einen möglichst guten Vergleich anstellen zu können, werden sowohl die Algorithmen, als auch die Systemabstraktionen selbst, möglichst modular implementiert.  
Dies erlaubt eine flexible Kombination der Algorithmus- und Systemmodule wodurch möglichst viele und aussagekräftige Tests durchgeführt werden können.

Zeitmessungen innerhalb der Tests sind immer von Zufall behaftet. Bei Rechenvorgängen, kann zum Beispiel das Betriebssystem die Rechenzeit von Test zu Test unterschiedlich verteilen.  
Solche Zeitmessungen werden deshalb wiederholt durchgeführt werden, um stattdessen einen Mittelwert zu ermitteln, welcher statistische Ausreißer leicht erkennbar macht und in diesen Messungen wesentlich aussagekräftiger ist.

Probleme bei der Verteilung von Berechnungen

Ressourcenzugriff

Wie bei den meisten parallelen Systemen, ist der virtuell gleichzeitige Zugriff auf einzelne Ressourcen problematisch. Es muss davon ausgegangen werden, dass Ressourcen vor gleichzeitigem Zugriff geschützt werden müssen.

Synchronisation

Bei jeder Verteilung muss das Ergebnis der Cores anschließend synchronisiert werden. Dies bedeutet, dass entweder Zwischenergebnisse zwischen den Cores ausgetauscht werden, oder das Endergebnis der Cores zurück in die Applikation fließen. Je nach Algorithmus kann dies mit nur einer oder aber mit sehr vielen Synchronisierungen passieren. Synchronisationen können, je nach System, unterschiedlich lang dauern und wirken sich dementsprechend stark aus, wenn viele Synchronisationen in einem System passieren, welches diese nur sehr langsam durchführen kann.

Overhead

Auch der sogenannte „Overhead“ der Verteilung muss beachtet werden. Dauert die Verteilung der Berechnungen lange, bedeutet dies einen großen Overhead und muss durch eine dementsprechend verkürzte Rechenzeit gerechtfertigt sein.

Mehrfachberechnung

Je nach Verteilungsalgorithmus kann es sein, dass manche Berechnungen mehrfach, auf unterschiedlichen Nodes, durchgeführt werden.  
Für diese Arbeit zählt ausschließlich die Geschwindigkeit als Kriterium, es soll an dieser Stelle jedoch erwähnt werden, dass mit steigender Anzahl der Berechnungen auch der Energieverbrauch steigt, was für Systeme mit beschränktem Energiehaushalt durchaus ein Grund sein kann, einen Verteilungsalgorithmus nicht anzuwenden.

Lösungsansätze

Um qualitative Aussagen über die Verteilungsalgorithmen treffen zu können, werden sowohl die Algorithmen, als auch die unterliegenden Rechensysteme, auf Kennmerkmale untersucht. Auf diese Weise, sollen Gemeinsamkeiten und Unterschiede gefunden werden, welche Vorhersagen, auch für ähnliche Systeme, ermöglichen.

Resource Lock

Jede Ressource wird einzeln vor Zugriffen geschützt. Wenn ein Core eine Berechnung durchführt, werden zuerst beide Ressourcen, welche für die Berechnung benötigt werden, für alle anderen Cores gesperrt. Sobald die Berechnung beendet ist, gibt der Core die Ressourcen wieder frei.



Core 1 sperrt Element 1 und Element 2 und führt die Berechnung zwischen ihnen durch.  
  
Core 2 will die Berechnung zwischen Element 2 und Element 4 ausführen, muss jedoch warten, bis Core 1 das Element 2 wieder freigibt.

Deadlock-Gefahr

Sobald mit zwei oder mehr Semaphoren gearbeitet wird, muss analysiert werden, ob das System in einem Deadlock-Zustand enden kann. Dies kann jedoch bei richtiger Implementierung niemals passieren, denn es werden unter keinen Umständen die gleichen Semaphore von zwei verschiedenen Threads belegt. Das liegt daran, dass dies nur passieren könnte, wenn zwei Threads die gleiche Berechnung zwischen den gleichen Elementen durchführen.  
Im weiteren Sinne bedeutet das, dass ein Deadlock nur auftreten kann, wenn der Algorithmus von vornherein falsch arbeitet.

Diskussion

Diese Art der Verteilung führt je nach Anzahl der Cores zu vielen oder wenigen Zugriffskollisionen. Bei einer Zugriffskollision versucht ein Core auf eine Ressource zuzugreifen, welche derzeit von einem anderen Core verwendet wird. Der Core muss nun warten bis die Ressource freigegeben wird, was zu erheblichen Leistungseinbußen führt, wenn solche Kollisionen oft vorkommen.

Round Robin Tournament Algorithmus

Diese Art der Verteilung wurde eigentlich für das Zuweisen von Spielpartnern, bei zum Beispiel Schachturnieren, entwickelt. Die Besonderheit ist dabei, dass jeder Teilnehmer gegen jeden anderen Teilnehmer genau einmal spielt.   
Die Verteilung sorgt dafür, dass pro „Spielrunde“ die maximale Anzahl an Kombinationen gleichzeitig durchgeführt werden.   
Durch die paarweise Struktur lässt sich dieses System auch auf die Problemstellung dieser Arbeit anwenden.

Vorgehensweise

Die Spieler werden durchnummeriert und „kreisförmig“ angeordnet.  
Jeweils der Spieler, der in der Tabelle oben steht, spielt in der ersten Runde gegen den Spieler, welcher unter ihm steht. Im Beispiel wären das Spieler 1 gegen 14, Spieler 2 gegen 13, usw.



Runde 1

In der zweiten Runde werden alle Spieler um eins weitergerückt, mit der Ausnahme von einem Spieler (im Beispiel Spieler 1).  
Nun spielt wieder die obere Reihe gegen die untere (Spieler 1 gegen 13, Spieler 14 gegen 12, …).

 Runde 2

Dieser Vorgang wird wiederholt bis sämtliche Kombinationen von Spielerpaaren miteinander gespielt haben.

 Letzte Runde

Die Rundenzahl ist abhängig von der Anzahl der Teilnehmer.

Die Anzahl der benötigten Runden beträgt für **N = gerade** **Teilnehmerzahlen**:  
Benötigte Runden = N – 1

Für **N = ungerade** **Teilnehmeranzahl**:  
Benötigte Runden = N

Diskussion

Es wird eine gleichmäßige Auslastung mit verhältnismäßig geringem Overhead erreicht. Der größte Teil der Verteilung kann im Vorfeld berechnet werden und wirkt sich dadurch nur sehr mild im Betrieb aus.  
Der Algorithmus skaliert mit der Anzahl der verfügbaren Cores, wenn die Anzahl der Elemente, welche zu berechnen sind, im gleichen Maß steigen.  
Ein Nachteil sind die vielen Synchronisationen, die zwischen den Berechnungsrunden durchgeführt werden müssen. Die Anzahl der Synchronisierungen steigt mit der Anzahl der Cores linear an. Dies wird bei Systemen mit hohen Latenzen problematisch.

Deepcopy

Es wird für jeden Core eine eigene Kopie der Daten erstellt. Auf jedem Core wird dann ein Teil der Berechnungen ausgeführt. Die Ergebnisse von allen Cores müssen zum Schluss synchronisiert werden.

Diskussion

Je nach Menge der benötigten Daten, kann der initiale Overhead groß sein.  
Sind die Daten jedoch verteilt, kann jeder Core, ohne Synchronisation, seinen Teil der Berechnungen störungsfrei durchführen.  
Die Synchronisation und Kombination der Endergebnisse kann je nach Art der Berechnung aufwendig sein und läuft außerdem nicht parallel ab, was den Vorgang nochmals verlangsamt.

Unterliegende Systeme

Single Thread

Es wird nur ein einziger Thread für die Ausführung verwendet. Somit kann auch nur ein Core des Systems ausgelastet werden.  
Wird verwendet um einen Referenzwert zu erhalten.

Threadspawning

Bei jedem Aufruf werden neue Threads erzeugt, welche dann die auf die Cores verteilt werden.  
Ist die Berechnung abgeschlossen, werden die Threads zerstört und die Ressourcen wieder freigegeben.  
Das Erzeugen von Threads ist nicht billig und es muss daher mit Leistungseinbußen gerechnet werden. Es könnte sich jedoch in einem System, mit sehr begrenztem Speicher, als die einzige Alternative herausstellen.

Actors

Die Threads (Actors) werden initial erstellt und laufen immer parallel mit dem Hauptprogramm mit.  
Sobald Berechnungen durchzuführen sind, werden diese als „Messages“ an die Actors gesendet und abgehandelt.  
Da die Threads permanent sind, sind die benötigten Ressourcen dauerhaft belegt. Dies ist jedoch nicht mit Prozessorlaufzeit gleichzusetzten, da die Actors im Sleep Modus sind, wenn sie nicht aktiv sind.

.NET Threadpool

<https://blogs.msdn.microsoft.com/pedram/2007/08/05/dedicated-thread-or-a-threadpool-thread/>

<https://msdn.microsoft.com/en-us/library/ms973903.aspx>

Abstraktion des unterliegenden Rechensystems  
  
Das Verteilungssystem soll ein austauschbarer Teil des Gesamtsystems sein und zwischen der eigentlichen Applikation und dem unterliegenden Rechensystem arbeiten.  
Dies soll die Applikation unabhängig davon machen, welches unterliegende System schlussendlich die Berechnungen, parallelisiert, durchführt.

|  |  |  |  |
| --- | --- | --- | --- |
| Applikation | | | |
|
|  |  |  |  |
| Verteilungssystem | | | |
|
|  |  |  |  |
| Berechnungssystem | | | |
|

Das Berechnungssystem soll über ein Interface an das Verteilungssystem gekoppelt werden.   
Dadurch kann das Berechnungssystem flexibel ausgetauscht werden und außerdem durch ein Dummysystem ersetzt werden, welches gewisse Eigenschaften des Berechnungssystems simulieren kann. Das Dummysystem kann außerdem dazu verwendet werden, um Tests durchzuführen, die von außerhalb nicht möglich sind.



Ressourcenzugriff  
  
Wie bei den meisten parallelen Systemen, ist der virtuell gleichzeitige Zugriff auf einzelne Ressourcen problematisch. Es muss davon ausgegangen werden, dass Ressourcen vor gleichzeitigem Zugriff geschützt werden müssen.  
Dies muss vom Verteilungssystem übernommen werden, da die Applikation kein Wissen über die Zugriffsweise des unterliegenden Systems hat.

Testsetup

Initialisierung

Die Testsuite wählt zuerst eine Kombination aus Verteilungssystem und Berechnungssytem, welche getestet werden soll. Dabei kann das Berechnungssystem auch Teil der Testsuite sein, um Tests durchführen zu können, welche von außerhalb nicht möglich wären.  
Die Konfiguration des Berechnungssystem beinhaltet zum Beispiel die Anzahl der verwendeten Cores, Verzögerungszeiten oder die Art der Berechnung, welche auf den Input angewendet wird.



Ablauf

Der Input ist abhängig von der Testroutine und kann entweder aus realen Daten bestehen, um Performance zu testen, oder aus Dummyobjekten bestehen, welche zusätzliche Auswertungen ermöglichen.  
Über den Output wird validiert, ob die Berechnungen korrekt durchgeführt wurde.  
Der Status des Berechnungssystems liefert Detailinformationen, zum Beispiel welche Berechnung auf welchem Core ausgeführt wurde.

Output-Validierung

Der Output wird mithilfe von Dummy-Objekten validiert. Dabei führt jedes Element eine Liste von allen Elementen, mit denen es bereits berechnet wurde.  
Die Berechnung auf den Cores, beinhaltet lediglich das Beschreiben dieser Liste.  
Nach dem Abschluss der Berechnungen, werden sämtliche Listen auf Fehler geprüft.

Performance-Messung

Die Performance-Messung wird ohne Output Validierung durchgeführt, da die Prüfung zusätzlichen Rechenaufwand verursacht, welcher die Zeitmessung verfälscht.

Um die Randbedingungen zu eliminieren, wird

Testumgebung

Betriebssystem

Getestet wird auf einem Windows 10 Betriebssystem. Dieses unterstützt nativ das verwendete .NET-Framework. Es handelt sich dabei um kein Echtzeitsystem, was die Varianz der Ergebnisse erhöht.

Prozessor

Es wird auf zwei verschiedenen Prozessoren getestet. Ein Intel i3 mit zwei echten Cores und aktiviertem Hyperthreading (4 Threads) und ein Intel i5 mit vier echten Cores ohne Hyperthreading (ebenfalls 4 Threads).

Links

<https://web.archive.org/web/20150806181915/http://www.zdnet.com/article/intel-why-a-1000-core-chip-is-feasible/>

<https://www.ll.mit.edu/HPEC/agendas/proc08/Day1/16-Day1-Session2-Reilly-abstract.pdf>

<http://www.eetimes.com/document.asp?doc_id=1167932>

<https://www.chesskid.com/article/view/ask-coach-jessica-chess-tournaments>

<http://www.emba.uvm.edu/~jdinitz/preprints/design_tourney_talk.pdf>

<https://arxiv.org/abs/1205.2367>

<https://dspace.mit.edu/bitstream/handle/1721.1/6935/AITR-633.pdf?sequence=2> // Actors