technische universität
dortmund

Masterarbeit

Effiziente String-Verarbeitung in Datenbankanfragen auf hochgradig paralleler Hardware

Florian Lüdiger
Juni 2019

Gutachter:
Prof. Dr. Jens Teubner
Henning Funke

Technische Universität Dortmund
Fakultät für Informatik
Datenbanken und Informationssysteme (LS-6)
http://dbis.cs.tu-dortmund.de

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation und Hintergrund

1.2 Aufbau der Arbeit

Kapitel 2

Grundlagen der GPU-Programmierung

Um die in dieser Arbeit vorgestellten Herausforderungen bei der Verarbeitung von String-Daten mit Grafikprozessoren, nachfolgend auch *GPU* genannt, verstehen zu können, ist zunächst ein Verständnis der grundlegenden Eigenschaften aktueller Hardware nötig. Dabei beschränkt sich diese Untersuchung auf die Grafikkarten-Serie Maxwell von NVIDIA, die hier besprochenen Prinzipien lassen sich allerdings auch auf andere GPUs anderer Hersteller übertragen und finden dort ebenfalls Anwendung.

2.1 Grundaufbau einer NVIDIA-Grafikkarte

Der Hauptprozessor eines Computers, auch *Central Processing Unit (CPU)* genannt, arbeitet eher sequenziell schwerwiegende Threads ab, wodurch individuelle Operationen schnell abgearbeitet werden können, ein hoher Durchsatz allerdings schwierig zu erreichen ist. Für die Verarbeitung großer Datenmengen wurden daher spezielle Co-Prozessoren in Form von Grafikkarten entwickelt, die hochgradig parallel arbeiten und somit einen massiven Durchsatz erreichen können. Die *Graphics Processing Unit (GPU)* bildet das Herzstück der Grafikkarte. Sie besteht aus einer hohen Anzahl an Kernen, die zwar individuell eine vergleichsweise geringe Leistung besitzen, allerdings aufgrund ihrer großen Zahl in datenparallelen Anwendungsfällen in Kombination mit einer hohen Speicherbandbreite eine hervorragende Performanz bieten.

Neben der GPU benötigt eine Grafikkarte noch weitere Peripherie, um effizient funktionieren zu können. Zur Speicherung der zu verarbeitenden Daten gibt es eigenständige Speichermodule, die unabhängig vom Hauptspeicher des Computers verwaltet werden. Für die NVIDIA GTX 950, welche im Folgenden als Beispiel genutzt werden soll, beträgt die Größe dieses Speichers 2 GB. Über eine PCI-Express-Anbindung wird die Kommunikation

mit dem Hauptprozessor und die Übertragung der Daten zwischen den Speicherbereichen realisiert.

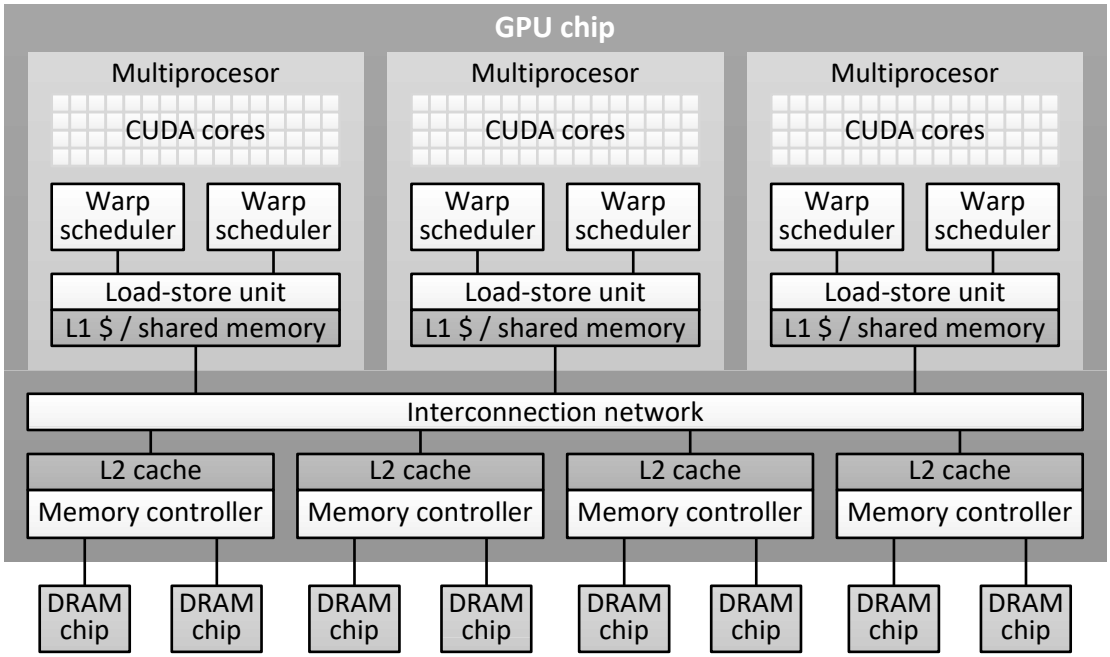

Abbildung 2.1: Architektur einer GPU [10]

Wie in Abbildung 2.1 dargestellt, lässt sich die GPU wiederum in kleinere Module, sogenannte *Streaming Multiprocessors (SM)*, unterteilen, welche jeweils eigenständige Recheneinheiten darstellen. Eine GT X950 besitzt beispielsweise sechs dieser Streaming Multiprocessors, welche sich ebenfalls in kleinere Einheiten unterteilen lassen. Die SM bestehen aus vier unabhängigen Blöcken von Rechenkernen, welche jeweils 32 skalare Recheneinheiten, auch *CUDA-Kerne* genannt, beinhalten. Jeder dieser Blöcke besitzt einen eigenen Scheduler und einige Unterstützungselektronik, sodass diese logisch gesehen ebenfalls unabhängig voneinander arbeiten können. [8] Bei sechs Streaming Multiprocessors mit jeweils vier Blöcken und 32 Recheneinheiten pro Block besitzt die GT X950 also 768 Kerne, welche über eine Programmierschnittstelle angesprochen werden können.

2.2 Scheduling auf GPUs

Um die hohe Anzahl von Kernen innerhalb einer GPU effizient mit Arbeit versorgen zu können, ist es wegen des großen Overheads nicht praktikabel, ein individuelles Scheduling für die einzelnen Recheneinheiten durchzuführen. Aus diesem Grund werden die Threads eines Programms in sogenannte *Warps* zusammengefasst, was damit die kleinste Einheit für das Scheduling bildet. Ein Warp enthält dabei genau 32 Threads, welche in diesem Kontext auch *Lanes* genannt werden. Mehrere Warps werden außerdem zu *Blöcken* zusammengefasst, welche schließlich als Ganzes an einzelne Streaming Multiprocessors zuge-

wiesen werden. Innerhalb eines SM werden Warps ausgetauscht, wenn der vorher aktive Warp beispielsweise auf einen Speicherzugriff wartet, um die dadurch entstehende Latenz zu verstecken.

Über die Anzahl der Threads pro Block und die gesamte Anzahl der Blöcke, ist die Konfiguration des sogenannten *Grids* definiert. Die Grid-Konfiguration nimmt starken Einfluss auf die Ausführungszeit der Software. Beispielsweise kann eine zu geringe Anzahl von Threads pro Block dazu führen, dass eventuell entstehende Latenzen nicht mehr so gut versteckt werden können, da nicht genug Threads innerhalb eines SM vorhanden sind. Eine zu hohe Anzahl von Threads pro Block kann allerdings auch von Nachteil sein, da Hardwareressourcen wie die Speichergröße pro SM gegebenenfalls nicht mehr ausreichen und das Programm nicht mehr korrekt funktioniert. Das Finden der richtigen Parameter gestaltet sich als äußerst schwierig, da die verwendete Hardware ein komplexes Konstrukt mit vielen Faktoren bildet, die auf Änderungen des Grids Einfluss nehmen.

Eine für die Programmierung von GPUs entscheidende Eigenschaft besteht darin, dass die Threads innerhalb eines Warps parallel ausgeführt werden. Ähnlich wie bei dem Prinzip *Single Instruction Multiple Data (SIMD)*, führen die Threads in einem Warp die Instruktionen synchron aus, sodass dieses Prinzip auch *Single Instruction Multiple Threads (SIMT)* genannt wird. Die Trennung in mehrere Threads bietet hierbei den Vorteil, dass eigene Register angesprochen werden können, an unterschiedlichen Stellen im Speicher gelesen werden kann und Threads verschiedene Kontrollflüsse verfolgen können. Prozesse laufen außerdem zwar logisch parallel ab, allerdings muss dies nicht notwendigerweise physikalisch auch so sein, sodass in einigen Fällen eine höhere Leistung erzielt werden kann. Für die optimale Performanz einzelner Operationen sollte allerdings gewährleistet sein, dass die Threads größtenteils synchron ausgeführt werden.

Bei der Verwendung von Branching-Instruktionen kann es vorkommen, dass unterschiedliche Threads verschiedene Kontrollflüsse durchlaufen, was auch als *Divergenz* bezeichnet wird. Da allerdings alle Threads identische Instruktionen ausführen müssen, führt dies dazu, dass sämtliche Threads in einem Warp alle notwendigen Kontrollflüsse durchlaufen und dabei gegebenenfalls das Ergebnis verwerfen, wenn diese sich logisch gesehen in einem anderen Zweig befinden. Alle Threads, für die der aktuell bearbeitete Kontrollfluss nicht relevant ist, werden als inaktiv bezeichnet. Inaktive Threads warten somit lediglich auf die aktiven Threads, bis diese die Arbeit innerhalb ihres Kontrollflusses abgeschlossen haben, sodass an dieser Stelle gegebenenfalls massiv Rechenleistung verschwendet wird. In dieser Problematik liegt der Grund dafür, dass die Verarbeitung von Strings auf Grafikkarten aufgrund ihrer variablen Länge problematisch ist, da die auftretenden Kontrollflüsse divergieren.

2.3 Synchronisation von Threads

Der Compiler und die GPU selbst versuchen innerhalb eines Warps die Anzahl der synchron ausgeführten Operationen zu maximieren, da dadurch eine höhere Leistung erzielt wird [7]. Diese Synchronisation kann allerdings auch explizit durch den Entwickler erfolgen, indem er die dafür vorgesehenen Operationen der Entwicklungsschnittstelle verwendet. Das Verwenden solcher Operationen führt dazu, dass alle Threads an dieser Stelle aufeinander warten müssen.

Diese Methoden können außerdem dazu verwendet werden, Informationen über die anderen Threads zu erlangen und die Zusammenarbeit innerhalb der Warps effektiver zu gestalten. Die Instruktionen werden von der Hardware unterstützt, sodass sie typischerweise sehr effizient ausgeführt werden können. Ein Beispiel für eine solche Operation ist das Auswerten eines Prädikates für alle Threads und anschließend das Erstellen einer Bitmaske, welche das Ergebnis der Auswertung für alle Threads enthält. Ein weiteres Beispiel ist das Generieren einer Maske für alle Threads, die in dem aktuellen Ausführungszweig aktiv sind. Schließlich können noch alle Threads ohne besondere Berechnung synchronisiert werden. Dies ist zum Beispiel nötig, wenn ein Thread aus dem Speicher lesen will, den andere Threads vorher beschreiben und dieser sicherstellen will, dass die Daten fertig geschrieben wurden [6].

2.4 Shared Memory

Eine Kommunikation zwischen Threads innerhalb eines Blocks kann über sogenannten *Shared Memory* geschehen. Dadurch können größere Mengen von Informationen ausgetauscht werden, als dies über die Synchronisations-Operationen effizient möglich wäre. Dieser Speicher ist um einige Größenordnungen schneller als der globale Speicher, da sich dieser direkt auf dem Chip der GPU befindet [4]. Die Speichergröße innerhalb eines Streaming Multiprocessors ist allerdings beschränkt, weshalb die Anzahl der Threads ebenfalls beschränkt ist, sofern eine große Menge Shared Memory von diesen benötigt wird.

2.5 Die CUDA-Programmierschnittstelle für C++

Für eine effiziente Entwicklung der hochgradig spezialisierten Grafikhardware stellt NVIDIA die *CUDA*-Programmierschnittstelle bereit. Diese ermöglicht es, die GPU aus einer Hochsprache wie C++ heraus anzusprechen und durch verschiedene Hilfestellungen leicht ein funktionierendes Programm zu erstellen. Neben vordefinierten Schlüsselwörtern und Syntaxelementen bietet die Entwicklungsumgebung auch einen eigenen Compiler, welcher das erstellte Programm für den Einsatz auf der Grafikkarte optimiert. Für das Verständnis der

Beispiele in dieser Arbeit sollen im Folgenden einige Grundkonzepte des Programmiermodells erläutert werden.

Das Hauptprogramm von CUDA-Programmen besteht aus Code für die CPU, welcher dafür zuständig ist, die Grafikkarte für ihre Aufgabe vorzubereiten und anschließend das Unterprogramm aufzurufen, welches auf der GPU ausgeführt werden soll. Ein solches Unterprogramm wird *Kernel* genannt und besteht im einfachsten Falle aus einer einfachen Funktion, welche durch das Schlüsselwort `__global__` gekennzeichnet wird. In diesem Kontext wird der GPU-Code üblicherweise *Device Code* und der CPU-Code *Host Code* genannt.

Auf die Schnittstellen zur Speicherverwaltung oder zur Festlegung der Grid-Konfiguration aus dem Host Code heraus soll hier nicht weiter eingegangen werden, da für die untersuchten Kriterien lediglich der Device Code interessante Aspekte bietet.

Einem Kernel können verschiedene Parameter wie Zeiger auf Speicherbereiche innerhalb des Grafikspeichers aus dem Hauptprogramm übergeben werden. Zum Durchlaufen eines solchen Speicherbereiches in einem sequenziellen Programm wäre es ausreichend mit einem Index über das Feld zu iterieren und diesen nach jeder Iteration um eins zu erhöhen. Bei einer parallelen Architektur würden so allerdings sämtliche Threads über den gesamten Datensatz laufen, anstatt wie gewünscht den Datensatz auf die einzelnen Threads aufzuteilen. Zu diesem Zweck muss jeder Thread die Informationen darüber haben, welchen globalen Index er innerhalb des Grids hat, um mit dem entsprechenden Element aus dem Datensatz zu beginnen und wie viele Threads in dem Grid vorhanden sind, damit er den entsprechenden Abstand zu dem nächsten zu untersuchenden Element kennt. Innerhalb eines Kernels kann der Thread auf seinen Threadindex (`threadIdx.x`), die Anzahl der Threads in einem Block (`blockDim.x`), seinen Blockindex (`blockIdx.x`) und die Anzahl der Blöcke im Grid (`girdDim.x`) zugreifen. Der globale Index eines Threads berechnet sich somit aus `blockIdx.x * blockDim.x + threadIdx.x` und die Sprungweite ist definiert durch `blockDim.x * gridDim.x`. Die dafür zur Verfügung gestellten Variablen und eine beispielhafte Iteration über zwei Datensätze sind in Listing 2.1 dargestellt.

```
1 __global__
2 void add(int n, float *x, float *y)
3 {
4   int index = blockIdx.x * blockDim.x + threadIdx.x;
5   int stride = blockDim.x * gridDim.x;
6   for (int i = index; i < n; i += stride)
7     y[i] = x[i] + y[i];
8 }
```

Listing 2.1: Beispielhafter CUDA-Kernel zum Iterieren über zwei Datensätze [3]

Mit den vorgestellten Informationen zum Grundaufbau von Grafikprozessoren, der Verwaltung von Threads und der Funktionsweise der C++-Programmierschnittstelle werden die in den nachfolgenden Kapiteln vorgestellten Techniken leicht verständlich sein.

Kapitel 3

Compiled Query Pipelines

In dieser Arbeit werden String-Vergleiche im Kontext des Query Compilers *DogQC* für GPUs untersucht. Dieser basiert auf dem Query Compiler *HorseQC* [2] und wurde für das erleichterte testen von aktuellen Techniken vereinfacht. DogQC erstellt aus einem gegebenen Anfrageplan eine Query Pipeline, die für die Ausführung auf GPUs optimiert ist. Als Grundlage für die späteren Codebeispiele, wird hier die grundsätzliche Funktionsweise des Query Compilers erklärt und die Vorteile dieser Technik erläutert.

Klassische in-memory Datenbanken wie *MonetDB* arbeiten die Operatoren innerhalb eines Anfrageplans nacheinander ab, was auch *Operator At A Time* genannt wird. Dabei wird für den gesamten Datensatz zunächst der erste Operator vollständig ausgeführt, bevor der gesamte Datensatz an den nächsten Operator weiter gegeben wird, bis schließlich der gesamte Anfrageplan abgearbeitet wurde. Der Nachteil dieser Strategie besteht in einer besonders hohen Lese- und Schreiblast für die Zwischenergebnisse der Operatoren, da diese nach jeder Operation im Speicher materialisiert werden müssen. Reicht der begrenzte GPU-Speicher nicht aus, um die Zwischenergebnisse zu speichern, werden während der Berechnung Transfers in den Hauptspeicher des Systems notwendig, um neue Blöcke der Tabelle nachzuladen. Als Konsequenz entstehen massive Flaschenhälse durch die begrenzte Bandbreite.

Das Pipelining-Prinzip, welches bei dem vorgestellten Query Compiler zum Einsatz kommt, besagt, dass der Anfrageplan in Pipelines aufgeteilt wird, welche von den Tupeln immer vollständig durchlaufen werden, bevor das Ergebnis materialisiert wird. Dieses Vorgehen wird *Tuple At A Time* genannt. Die Operatoren innerhalb des Anfrageplans aus Abbildung 3.1 werden zu zwei Pipelines zusammengefasst. In der linken Pipeline wird die dates-Relation gelesen, die Selektion ausgeführt und die Hashtabelle für den Join berechnet. Die rechte Pipeline fasst das Lesen der orders-Relation, die Selektion, die Probe-Operation der Hashtabelle und das Zählen der Ergebnisse zusammen. Technisch werden die Operationen innerhalb der Pipeline zu einem einzigen Operator verschmolzen, indem der Query Compiler für jede Pipeline einen eigenständigen Kernel generiert, welcher mit der CUDA-

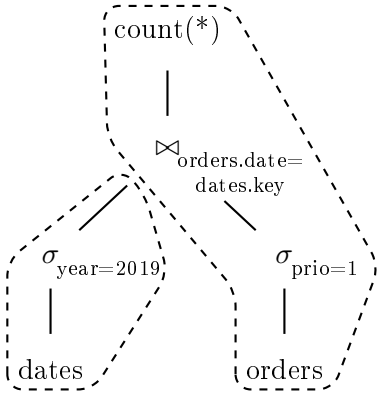

Abbildung 3.1: Beispielplan mit eingezeichneten Pipelines

Schnittstelle ausgeführt werden kann. Der Vorteil des Pipelinings besteht drain, dass die Ergebnisse jedes Operators nicht immer wieder im Speicher materialisiert werden müssen, sondern die einzelnen Tupel stets in den GPU-Registern vorgehalten werden können, bis diese fertig verarbeitet wurden.

In Listing 3.1 wird der Code vorgestellt, welcher von dem Query Compiler für die rechte Pipeline aus dem in Abbildung 3.1 dargestellten Anfrageplan generiert wurde. Hier ist zu erkennen, dass die drei Operationen innerhalb eines Kernels zusammengefasst wurden, wodurch eine Pipeline entsteht. Um jedem Thread eine Menge von Tupeln zuweisen zu können, wird zunächst der globale Index des aktuellen Threads innerhalb des Grids berechnet, damit dieser als Schleifenindex `loop_var` verwendet werden kann. Anschließend wird über alle Elemente aus dem Datensatz iteriert, für die der aktuelle Thread zuständig ist. Die Variable `active` zeigt im Algorithmus an, ob der aktuelle Thread aktiv läuft oder nur darauf wartet, dass die anderen Threads aus seinem Warp ihre Berechnung abschließen.

In einem Schleifendurchlauf, bei dem das Element mit dem Index `loop_var` untersucht wird, wird zunächst die rot markierte Selektion ausgeführt. Ist diese fehlgeschlagen, da die Priorität der Bestellung nicht bei 1 liegt, wird der Thread deaktiviert und im weiteren Verlauf nicht mehr beachtet, bis er ein neues Tupel erhält. Hat sich das Tupel qualifiziert, folgt darauf der Hash Probe, welcher in grün dargestellt ist. Dabei wird das Tupel in der übergebenen Hashtabelle `hashtable_date_key` gesucht und dementsprechend wieder die `active`-Variable angepasst. Schließlich wurde vom Query Compiler noch das Zählen der Ergebnisse umgesetzt, welches hier in gelb hervorgehoben wurde. Dabei wird wieder mithilfe der Synchronisierungsoperation `__ballot_sync` die Anzahl der aktiven Lanes gezählt, welche jeweils ein Element des Ergebnisses repräsentieren. Diese Anzahl wird daraufhin vom ersten Thread innerhalb des Warps auf das Ergebnis addiert.

Nach der Durchführung der gesamten Pipeline von Operationen für das untersuchte Tupel wird der Index erhöht, sodass im nächsten Schleifendurchlauf das nächste Tupel untersucht wird. Falls der neu gewählte Index hinter dem Ende der Daten liegt, hat der

aktuelle Thread seine Arbeit vollständig abgeschlossen und er wird nicht mehr benötigt, sodass die `active`-Variable in Zeile 20 auf `false` gesetzt wird. Wird anschließend mithilfe der `__ballot_sync`-Methode festgestellt, dass sämtliche Lanes inaktiv sind, ist der Datensatz vollständig durchlaufen worden und die Berechnung kann abgeschlossen werden.

```
1  __global__
2  void joinProbePipeline(
3    int *orders_prio,                // priority attribute of orders table
4    int *orders_date,                // date attribute of orders table
5    unique_ht *hashtable_date_key,   // hashtable from other pipeline
6    int *number_of_matches) {        // return value
7
8    // global index of the current thread,
9    // used as the iterator in this case
10   unsigned loop_var = ((blockIdx.x * blockDim.x) + threadIdx.x);
11
12   // offset for the next element to be computed
13   unsigned step = (blockDim.x * gridDim.x);
14
15   bool active = true;
16   bool flush_pipeline = false;
17   while(!flush_pipeline) {
18
19     // element index must not be higher than number of tuples
20     active = loop_var < TUPLE_COUNT_ORDERS;
21
22     // break computation when every line is finished and therefore inactive
23     flush_pipeline = !__ballot_sync(ALL_LANES, active);
24
25     // selection
26     if (active)
27       active = orders_prio[loop_var] == 1;
28
29     // hash join probe
30     if (active)
31       active = hashProbeUnique(hashtable_date_key, HASHTABLE_SIZE,
32         hash(orders_date[loop_var]))
33
34     // count and write
35     numProj = __popc(__ballot_sync(ALL_LANES, active))
36     if (threadIdx.x % 32 == 0)
37       atomicAdd(number_of_matches, numProj);
38
39     loop_var += step;
40   }
41 }
```

Listing 3.1: Generierter Kernel für den Beispielplan

Bei dem hier vorgestellten Verfahren der *Tuple At A Time* Verarbeitung wird klar, dass dieses einen großen Vorteil bei der effizienten Nutzung von schnellem Speicher gegenüber der *Operator At A Time* Verarbeitung bietet. Zwischenergebnisse müssen nicht mehr materialisiert werden, weshalb Tupel in Registern oder Caches vorgehalten werden können und die Operationen für das explizite Materialisieren entfallen. Im Gegensatz zu der *Operator At A Time* Technik lässt sich das Pipelining allerdings nicht so einfach auf eine parallele Verarbeitung mit GPUs übertragen. Wird jedem Thread in einem Warp ein Tupel übergeben, kann es passieren, dass bei beispielsweise einer Selektion dieses Tupel aus der Ergebnisrelation heraus fällt, somit im weiteren Verlauf nicht weiter beachtet werden muss und die entsprechende Lane inaktiv wird. Dieses Problem wird in Kapitel 4.4 aufgegriffen und in Kapitel 5 wird ein Lösungsvorschlag dafür vorgestellt.

Kapitel 4

Einfacher, paralleler String-Vergleich

Um einige Techniken zur String-Verarbeitung in kompilierten Anfragepipelines auf Grafikkarten entwickeln zu können, wird zunächst ein Operator für den einfachen String-Vergleich für den Query Compiler erarbeitet. Ein Vergleich auf Gleichheit stellt dabei die einfachste, sinnvolle Variante von String-Verarbeitung dar, wodurch der Einfluss vieler Eigenschaften von Strings auf die Ausführung entsprechender Operationen leicht untersucht werden kann.

Zunächst wird dazu die Motivation hinter einer derartigen Untersuchung erläutert und eine bestehende Technik zur String-Verarbeitung erklärt. Außerdem wird eine Umsetzung des einfachen String-Vergleichs mittels der CUDA-Schnittstelle ohne spezielle Optimierungen vorgestellt und für einen alternativen Workload für weitere Tests leicht angepasst werden. Schließlich wird beurteilt, ob die Lösung das Potential hat eine optimale Performanz zu bieten und auf einen Nachteil der einfachen Implementierung eingegangen.

4.1 Motivation

In den meisten modernen Anwendungsfällen für Datenbankmanagementsysteme stellt die Verarbeitung von String-Daten einen wesentlichen Bestandteil dar. So wird häufig auf Gleichheit, Ungleichheit, das Enthaltensein eines Teilstrings oder das Erfüllen eines regulären Ausdrucks getestet. Die effiziente Berechnung unterschiedlicher Operatoren auf Zeichenketten ist somit essenziell für das Erreichen eines hohen Durchsatzes und für das Gewährleisten von maximaler Performanz. In diesem Kontext versprechen Grafikkarten durch ihre hochgradig parallele Architektur in der Theorie eine bestmögliche Leistung.

Da die Ausführung von String-Operationen auf Grafikkarten wie in Kapitel 4.4 beschrieben einige Probleme birgt, verwenden bisherige Ansätze zur Verarbeitung von Zeichenketten das Konzept der Dictionaries. Dabei wird eine Tabelle mit allen Strings aufgebaut und zu diesen ein Schlüssel abgespeichert, welcher zusammen mit jedem String in den anderen Tabellen der Datenbank gespeichert wird. Somit können String-Operationen durch andere Operationen auf den Schlüsseln abgebildet werden, wodurch diese eine einheitliche

Struktur und damit ein effizientes Ausführungsmuster auf Grafikkarten erhalten. Das Aufbauen und Verwalten des für diese Technik verwendeten Dictionaries erzeugt vor allem bei Daten, die sich häufig ändern, einen hohen Aufwand, worunter die Leistungsfähigkeit des Gesamtsystems sinkt.

Um diesen Verwaltungsaufwand für eine zusätzliche Datenstruktur zu eliminieren, wird eine Lösung entwickelt, die direkt auf den Zeichenketten arbeitet und trotzdem eine hohe Performanz bietet.

4.2 Umsetzung des einfachen String-Vergleichs

Als Basis für die Untersuchungen wird der String-Vergleich-Operator für den Query Compiler zunächst naiv, also ohne tiefgehende Optimierungen umgesetzt. Mithilfe dieses Operators kann in einer Datenbankabfrage beispielsweise eine Selektion über eine Spalte mit String-Daten durchgeführt werden. Die Anforderung des Operators besteht somit darin, eine Liste von Zeichenketten mit einem vorher spezifizierten String zu vergleichen und zu entscheiden, ob diese identisch sind oder nicht.

Zur Durchführung dieser Operation wird jedem Thread der GPU eine Zeichenkette aus dem Datensatz zugewiesen. Zunächst wird überprüft, ob die Länge des Strings mit der des Suchstrings übereinstimmt, sodass der entsprechende Eintrag direkt verworfen werden kann. Sind die Längen identisch, werden beide Zeichenketten Zeichen für Zeichen durchlaufen und diese an jeder Stelle auf Gleichheit überprüft. Sobald eine Ungleichheit gefunden wurde, wird ein entsprechendes Flag gesetzt und die weiteren Zeichen müssen nicht mehr genauer betrachtet werden.

Sämtliche Threads innerhalb eines Warps werden entsprechend dem Verarbeitungsmodell der GPU parallel abgearbeitet. Somit sind die Positionen, an denen die Strings verglichen werden ebenfalls für alle Threads identisch. Sobald der Vergleichsstring im gesamten Warp vollständig durchlaufen wurde, wird das Zwischenergebnis geschrieben. Sollten alle Threads in dem Warp vorzeitig feststellen, dass keiner der Strings mit dem Suchstring übereinstimmt, wird die aktuelle Untersuchung vorzeitig abgebrochen.

Schließlich wird jedem Thread eine neue Zeichenkette aus dem Datensatz zugewiesen, sodass das Verfahren im weiteren Verlauf wiederholt wird. Sobald der gesamte Datensatz durchlaufen wurde, ist die Berechnung abgeschlossen.

In Abbildung 4.1 ist der Ablauf des Algorithmus innerhalb eines Warps mit drei Threads dargestellt. Es ist erkennbar, an welchen Stellen die Lanes inaktiv werden, wann die Berechnung frühzeitig abgebrochen werden kann und an welchen Stellen das Ergebnis geschrieben wird und neue Daten aus dem Datensatz geholt werden.

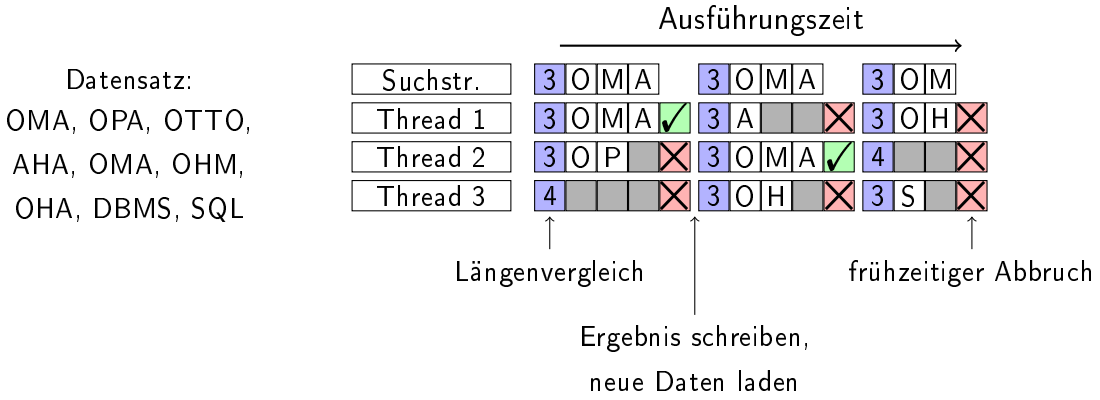

Abbildung 4.1: Funktionsweise des Algorithmus innerhalb eines Warps mit drei Threads

```
1  /* execute previous operators in the pipeline */
2
3  data_length = char_offset[loop_var+1] - char_offset[loop_var] - 1;
4
5  // if string lengths are unequal, discard
6  if (active && data_length != search_length)
7    active = false;
8
9  int search_id = 0;
10
11 // iterate over strings completely or until they don't match anymore
12 while(active && search_id < search_length) {
13   int data_id = search_id + char_offset[loop_var];
14
15   // when strings don't match, inactivate the lane
16   if (active && data_content[data_id] != search_string[search_id])
17     active = false;
18
19   search_id++;
20 }
21
22 /* execute following operators in the pipeline */
```

Listing 4.1: Naive Implementierung einer Selektion von Strings

In Listing 4.1 ist die Implementierung des Operators dargestellt, welcher in einer kompilierten Anfragepipeline die Selektion eines String-Attributs durchführen kann. Dieser würde beispielsweise die rot markierte Selektion in Listing 3.1 ersetzen, falls das *prio*-Attribut in dem Beispiel aus Abbildung 3.1 Strings enthalten würde und alle Bestellungen mit der Priorität *HIGH* gesucht werden würden. Nach Abschluss der Berechnung ist der Wert der `active`-Variable genau dann `true`, wenn der String mit dem gesuchten String übereinstimmt, sodass im Anschluss mit weiteren Operationen fortgefahren werden kann.

Diese Implementierung erwartet, dass einige Daten vorher vom Hauptspeicher in den Speicherbereich der GPU kopiert wurden und dort zur Verfügung stehen. Der Datensatz, der mit dem Vergleichsstring abgeglichen werden soll, besteht aus einer Aneinanderreihung der entsprechenden Zeichenketten ohne Trennzeichen und ist in der Variable `data_content` gespeichert. Damit daraus die ursprünglichen Strings extrahiert werden können, gibt es das Feld `char_offset`, welches Informationen über die Indizes der Einzelstrings innerhalb des Datensatzes enthält. Ebenfalls muss natürlich ein Zeiger auf den Suchstring und dessen Länge in den entsprechenden Parametern `search_string` und `search_string_length` vorhanden sein. Um die Berechnung rechtzeitig vor Speicherüberschreitungen abbrechen zu können, wird schließlich noch die Variable `line_count` benötigt, welche die Anzahl der Zeichenketten im Datensatz beschreibt.

Im ersten Schritt wird für einen String überprüft, ob dessen Länge mit der des Suchstrings übereinstimmt und dieser anderenfalls verworfen. Sind die Längen identisch, wird in der Schleife über beide Zeichenketten iteriert, bis das Ende beider erreicht wurde, oder festgestellt wird, dass ein Zeichen aus dem Vergleichsstring nicht mit dem aus dem Suchstring übereinstimmt. An dieser Stelle fällt die parallele Struktur des Kernels besonders auf, da die Zeichen aller Strings parallel von den Threads durchlaufen werden und diese erst aufhören, wenn der letzte Thread den ihm zugewiesenen Datensatz vollständig durchlaufen hat. Ist die Schleife abgeschlossen oder vorzeitig abgebrochen worden, kann am Zustand der `active`-Variable abgelesen werden, ob die Strings übereinstimmen oder nicht.

4.3 Präfixtest als alternativer Workload

Als zusätzliche String-Operation, für die ein simpler Algorithmus existiert, wurde neben dem exakten String-Vergleich auch ein Präfixtest entwickelt. Dabei soll geprüft werden, ob die Strings aus einer Datenbanktabelle einen vordefinierten Präfix besitzen.

Der Algorithmus arbeitet ähnlich wie das in Kapitel 4.2 vorgestellte Verfahren zum exakten String-Vergleich, mit einem einzigen Unterschied. Für jede Zeichenkette aus der Tabelle wird zunächst geprüft, ob diese länger als der gesuchte Präfix ist, anstatt zu prüfen, ob die Längen identisch sind. Danach verfährt der Algorithmus wie gehabt, sodass wieder beide Strings Zeichen für Zeichen verglichen werden, bis das Ende des Such-Präfixes erreicht ist. An der Implementierung in Listing 4.1 ändert sich dementsprechend fast nichts, es muss nur die Bedingung in Zeile 4 durch `active && data_length >= search_length` ersetzt werden.

Dadurch, dass die Prüfung auf exakte Längengleichheit entfällt, müssen je nach Anwendungsfall viele Zeichenketten weiter durchlaufen werden, da diese nicht schon im ersten Schritt ausgeschlossen werden können, wie bei dem exakten String-Vergleich. Durch diesen Umstand und durch die Tatsache, dass ein Präfixtest in realen Systemen ein häufig gefragter Anwendungsfall ist, ergibt diese Umsetzung einen interessanten alternativen Workload.

4.4 Einschätzung der GPU-Auslastung

Das in diesem Kapitel vorgestellte, naive Verfahren zum einfachen String-Vergleich nutzt die Ressourcen der GPU nicht besonders effizient aus. In Abbildung 4.1 sind in grau die inaktiven Threads dargestellt. Diese sind inaktiv geworden, da erkannt wurde, dass sie nicht mit der gesuchten Zeichenkette übereinstimmen, da sie entweder eine unpassende Länge besitzen oder im Laufe des Vergleichs der einzelnen Zeichen ein Unterschied festgestellt wurde. Die untersuchten Strings aus der Tabelle können völlig unterschiedlich geartet sein, weshalb es häufig vorkommt, dass einige Threads ihre Untersuchung bereits abgeschlossen oder vorzeitig unterbrochen haben, während andere Threads innerhalb des Warps noch lange weiter rechnen müssen. Aufgrund des Programmiermodells von Grafikkarten, laufen die inaktiven Threads weiter synchron zu den aktiven Threads des Warps, dabei wird allerdings das Ergebnis verworfen und diese verrichten keine nutzbare Arbeit mehr.

Im schlimmsten Fall werden 31 Threads aus einem Warp Zeichenketten mit unpassender Länge zugewiesen und einem einzigen Thread eine mit dem Vergleichsstring übereinstimmende Zeichenkette zugeteilt. Somit stellen 31 Threads im ersten Schritt fest, dass die Länge des Strings nicht mit der des Vergleichsstrings übereinstimmt und werden somit inaktiv. Der Thread mit dem passenden String muss allerdings noch über jedes Zeichen iterieren, bevor sich der gesamte Warp neue Strings holen kann. Es kann also sein, dass für die Dauer der Iteration über den Suchstring die GPU nur zu 1/32 ausgelastet ist, weshalb die Performanz dieser Lösung verbesserungswürdig ist.

Inaktiven Threads sollte somit dynamisch neue Arbeit zugewiesen werden, sobald diese inaktiv geworden sind. Da in dem Algorithmus lediglich zwei Zeichen an zwei Positionen verglichen werden und diese Position für jeden Thread unterschiedlich sein kann, könnte ein Thread, sobald er inaktiv geworden ist, sich eine neue Zeichenkette holen und mit dieser weiter arbeiten. Mit diesem Vorgehen könnte zwar eine hohe Auslastung erreicht werden, da die Threads niemals wirklich inaktiv werden können, allerdings funktioniert dies nicht mit dem in Kapitel 3 vorgestellten Pipelining-Modell. Dies liegt daran, dass gegebenenfalls noch beliebig viele andere Operationen in der Pipeline vor dem String-Vergleich durchgeführt werden müssen, bevor eine neue Zeichenkette an den Thread übergeben wird.

Im folgenden Kapitel wird eine weitere Technik vorgestellt, mit der die Auslastung der GPU verbessert werden kann. Diese wird auch im Umfeld einer Anfragepipeline funktionieren und somit für den praktischen Einsatz geeignet sein.

Kapitel 5

Verbesserung des einfachen String-Vergleichs

Zur Verbesserung der Auslastung der GPU bei der Berechnung des einfachen String-Vergleichs wird ein Verfahren vorgestellt, das auch in kompilierten Anfragepipelines eine optimierte Laufzeit verspricht. Dieses stellt durch Verwendung eines Puffers sicher, dass zu jeder Zeit während der Laufzeit die Auslastung eines Warps über einem bestimmten Grenzwert liegt. Das Prinzip wurde im Kontext von kompilierten Anfragepipelines mit SIMD als *consume everything* vorgestellt [5] und im Kontext dieser Arbeit auf Grafikkarten angewendet. Im Folgenden wird das Verfahren als *Lane Refill* bezeichnet, da es die inaktiven Lanes in einem Warp dynamisch wieder auffüllt.

5.1 Funktionsweise des String-Vergleichs mit Lane Refill

Zur Verbesserung des einfachen String-Vergleichs kann das naive Verfahren durch das Lane Refill erweitert werden, indem ein Puffer eingeführt wird, in dem teilweise abgearbeitete Tupel zwischengespeichert werden können. Bei der naiven Umsetzung gibt es drei Stellen, an denen Lanes innerhalb eines Warps inaktiv werden können. Zum einen passiert dies, wenn beim Längenvergleich zu Beginn eine unpassende Länge festgestellt wird und zum anderen wenn beim Vergleich von zwei Zeichen ein Unterschied festgestellt wird. Schließlich werden Lanes ebenfalls inaktiv, wenn diese ihren Vergleich abgeschlossen haben und einen passenden String gefunden haben. Jeweils nach diesen Ereignissen soll nun überprüft werden, ob die Auslastung des Warps noch ausreichend ist, um den Grenzwert zu erreichen. Ist dieser Grenzwert unterschritten, können wiederum zwei Fälle auftreten. Sollten sich im Puffer noch ausreichend Tupel befinden, sodass durch Auffüllen der Grenzwert wieder erreicht ist, so werden die inaktiv gewordenen Lanes mit den Zeichenketten aus den gepufferten Tupeln befüllt und es wird weiter gerechnet. Wurden zuvor zu wenige Elemente gepuffert, sodass der Grenzwert nicht erreicht werden kann, werden die Tupel aus den

aktiven Lanes im Puffer gespeichert und die Pipeline mit frischen Tupeln neu gestartet, damit später neue Zeichenketten für den String-Vergleich bereitstehen.

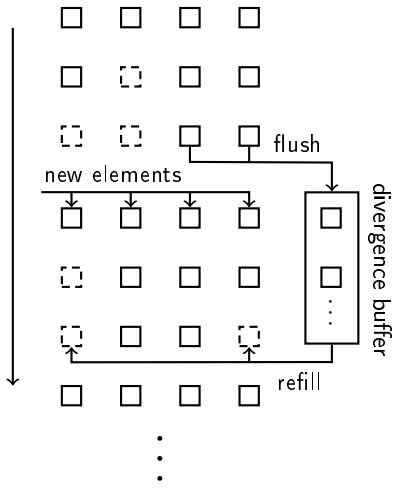

Abbildung 5.1: Funktionsweise des Lane Refill (Quelle: Henning Funke)

In Abbildung 5.1 ist die Funktionsweise der vorgestellten Technik dargestellt. Es ist zu erkennen, dass aktive Tupel in einem schlecht ausgelasteten Warp im Puffer zwischengespeichert werden, was hier als *flush* bezeichnet wird. Im Anschluss fließen neue Tupel über den Weg der Pipeline wieder in den Warp ein, wodurch dieser wieder vollständig ausgelastet ist. Sobald erneut eine Unterauslastung auftritt, werden die zwischengespeicherten Elemente aus dem Puffer in die inaktiven Lanes geladen und der Warp ist wieder effizient ausgelastet.

Die hier vorgestellte Technik ermöglicht es, die Unterauslastung innerhalb der String-Vergleichs-Operation zu verringern, allerdings entsteht dadurch gegebenenfalls eine starke Unterauslastung bei den Folgeoperationen in der Pipeline. Dieser Effekt entsteht, da die Überprüfung der Zeichenketten zu sehr unterschiedlichen Zeitpunkten abgeschlossen sein kann und somit immer mal wieder einige Tupel bereit sind, die Folgeoperationen der Pipeline auszuführen, während viele andere Tupel sich noch innerhalb des String-Vergleichs befinden. Um dieses Problem zu lösen, kann beispielsweise nach dem String-Vergleich eine weitere Puffer-Operation eingeführt werden, welche sicherstellt, dass genug Zeichenketten fertig überprüft wurden, sodass die nachfolgenden Operationen mit einer ausreichenden Auslastung ausgeführt werden.

5.2 Struktur des optimierten String-Vergleichs im Kernel

Um den einfachen String-Vergleich durch Lane Refill zu optimieren, muss die Struktur des Operators im Kernel etwas angepasst werden, was in Listing 5.1 dargestellt wird. Die äußere Schleife ist bereits aus der Grundstruktur einer Pipeline, wie sie in Kapitel 3 vorgestellt

wurde, bekannt und wurde hier für ein leichteres Verständnis aufgegriffen. Innerhalb der Schleife werden wie bei der naiven Implementierung zunächst die Operatoren ausgeführt, die in der Pipeline vor dem String-Vergleich stehen und es wird die Länge der Zeichenkette untersucht und diese gegebenenfalls verworfen.

Um im nächsten Schritt beurteilen zu können, ob eine ausreichende Auslastung besteht, wird zunächst mithilfe einer Synchronisierungsoperation überprüft, wie viele Lanes im Warp aktiv sind. Sind im Puffer und im Warp in Summe noch ausreichend aktive Elemente vorhanden, um den vorher definierten Grenzwert zu überschreiten, so kann weiter mit dem String-Vergleich fortgefahren werden. Sollten die aktiven Lanes alleine nicht ausreichen, um den Grenzwert zu erreichen, werden gegebenenfalls zuvor noch die leeren Lanes mit Tupeln aus dem Puffer wieder aufgefüllt. Der eigentliche Vergleich zweier Zeichen funktioniert hier wieder genau wie bei der naiven Implementierung. Nach dem Vergleich wird geprüft, ob der String vollständig durchlaufen wurde und entsprechend die folgenden Operationen in der Pipeline ausgeführt. An dieser Stelle fällt auf, dass im Gegensatz zu der naiven Implementierung eine Verschachtelung entsteht, da die Folgeoperationen innerhalb der `while`-Schleife des String-Vergleichs-Operators ausgeführt werden.

Ist nach erneuter Zählung der aktiven Lanes die Schleifenbedingung nicht mehr erfüllt, da zu wenige Tupel existieren, um eine ausreichende Auslastung zu gewährleisten, werden die Tupel aus den restlichen, aktiven Lanes in den Puffer geschrieben. Schließlich wird die Pipeline mit frischen Tupeln wieder von vorne gestartet.

```
while(!flush_pipeline) {

  /* execute previous operators in the pipeline */

  // if string lengths are unequal, discard
  if (active && data_length != search_length)
    active = false;

  int numactive = __popc(__ballot_sync(ALL_LANES, active));
  int bufferelements = 0;

  while(bufferelements + numactive > THRESHOLD) {

    if (numactive < THRESHOLD) {

      /* refill empty lanes from buffer in case of underutilization */

      bufferelements = bufferelements - numrefill;
    }

    // when strings don't match, inactivate the lane
    if (active && data_content[data_id] != search_string[search_id])
      active = false;

    search_id++;

    if (search_id == search_length) {

      /* execute following operators in the pipeline */

    }

    numactive = __popc(__ballot_sync(ALL_LANES, active));
  }

  if (numactive > 0) {

    /* flush active lanes to buffer */

    bufferelements += numactive;
    active = false;
  }

  loop_var += step;
}
```

Listing 5.1: Struktur der String-Selektion mit Lane Refill

5.3 Technische Umsetzung der Pufferung

Der Puffer wird mithilfe der CUDA-Programmierschnittstelle als Shared Memory umgesetzt, um eine effiziente Kommunikation zwischen den Lanes zu ermöglichen. Für den einfachen String-Vergleich werden dazu zwei Speicherbereiche benötigt, welche mit dem Schlüsselwort `__shared__` initialisiert werden. In einem Feld wird die Position des untersuchten Strings gespeichert (`current_divergence_buffer`) und in dem anderen Feld wird der Index des Zeichens innerhalb des Strings gespeichert, das als nächstes verglichen werden muss (`search_id_divergence_buffer`). Da der Shared Memory immer auf dem Level eines ganzen Blocks gültig ist, muss dieser so viele Elemente fassen können wie es Threads pro Block gibt. Eine ausführliche Version der Umsetzung befindet sich in Anhang A, hier sollen aber dennoch die Techniken zum Zwischenspeichern und Laden von Elementen kurz beschrieben werden.

```
1  if (numactive < THRESHOLD) {
2    numRefill = min(32 - numactive, bufferelements);
3    numRemaining = bufferelements - numRefill;
4
5    previous_inactive = __popc(~__ballot_sync(ALL_LANES, active) &
         prefixlanes);
6
7    if (!active && previous_inactive < bufferelements) {
8      buf_ix = numRemaining + previous_inactive + bufferbase;
9      search_id = search_id_divergence_buffer[buf_ix];
10     current = current_divergence_buffer[buf_ix];
11     active = true;
12   }
13
14   bufferelements -= numRefill;
15 }
```

Listing 5.2: Befüllen inaktiver Lanes mit Elementen aus dem Puffer

Bei dem in Listing 5.2 dargestellten Befüllen von leer gelaufenen Lanes mit Elementen aus dem Puffer, bleiben die noch aktiven Lanes unberührt. Die inaktiven Lanes müssen zunächst beurteilen, ob diese berechtigt sind, sich ein Tupel aus dem Puffer zuzuweisen. Befinden sich beispielsweise zwei Elemente im Puffer, es gibt aber drei inaktive Lanes, dann erhalten die zwei Lanes mit dem niedrigeren Index ein neues Tupel und die Lane mit höherem Index bleibt weiter inaktiv. Als erstes muss also bestimmt werden, wie viele Lanes vor der betrachteten Lane inaktiv sind. Der Wert `prefixlanes` ist dabei eine Bitmaske mit einem Bit für jede Lane im Warp, bei der für alle vor der aktuellen Lane liegenden Lanes ein Bit gesetzt wurde. Gibt es weniger inaktive vor der betrachteten Lane als Elemente im Puffer, darf sich diese Lane ein neues Tupel aus dem Puffer holen.

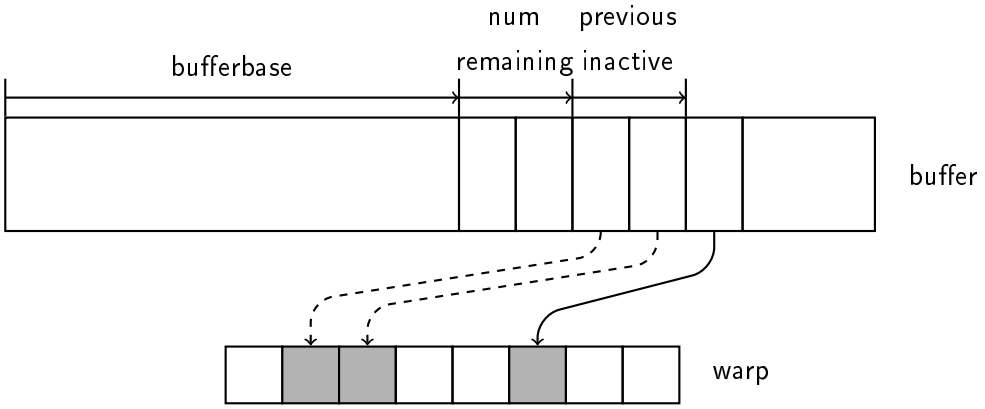

Abbildung 5.2: Berechnung des Indexes für ein Element im Puffer

Um das richtige Element zu erhalten, muss zunächst wie in Abbildung 5.2 dargestellt dessen Index im Puffer errechnet werden, welcher sich aus verschiedenen Offsets Zusammensetzt. `buffer_base` bestimmt dabei die Position des Speicherbereichs, welcher dem aktuellen Warp im Block zugewiesen wurde. Da der Puffer grundsätzlich von rechts abgearbeitet wird, muss noch der Wert `num_remaining` für die Elemente, die im Puffer verbleiben sollen, addiert werden und schließlich noch die Tupel übersprungen werden, die für Lanes mit niedrigerem Index bestimmt sind.

Ist der Index berechnet, können der String und der Suchindex innerhalb des Strings aus dem Puffer geladen werden und die Lane wieder aktiv geschaltet werden.

```
1  if (numactive > 0) {
2    previous_active = __popc(__ballot_sync(ALL_LANES, active) & prefixlanes);
3    buf_ix = bufferbase + bufferelements + previous_active;
4
5    if(active) {
6      search_id_divergence_buffer[buf_ix] = character_index;
7      current_divergence_buffer[buf_ix] = current;
8    }
9
10   bufferelements += numactive;
11   active = false;
12 }
```

Listing 5.3: Auslagern übriger, aktiver Lanes in den Puffer

Das in Listing 5.3 dargestellte Verfahren zum Einlagern übriger, aktiver Lanes in den Puffer funktioniert ähnlich wie das wieder befüllen von Lanes. Zunächst wird wieder der Index berechnet, auf den die aktuelle Lane im Puffer zugreifen kann, was identisch zu dem in Abbildung 5.2 dargestellten Verfahren funktioniert, nur dass in diesem Falle die aktiven Lanes statt der inaktiven Lanes als Offset verwendet werden. Alle verbleibenden, aktiven Lanes speichern daraufhin die Position ihres vorher untersuchten Strings und den

als nächstes zu untersuchenden Index innerhalb des Strings im Puffer ab. Schließlich wird noch die Anzahl der Elemente im Puffer erhöht und der Algorithmus kann anschließend wie gehabt verfahren.

5.4 Reduzierung des Overheads

Ein Nachteil dieses Verfahrens liegt darin, dass durch das Verwalten des Puffers ein erhöhter Overhead entsteht, der die bessere Performanz durch eine gute Auslastung der GPU überschatten könnte. Um diesen Overhead zu verringern, könnte es sinnvoll sein in größerem Zeitabstand die Auslastung zu überprüfen und entsprechend Lanes neu zu befüllen. Um dies zu erreichen, muss wie in Listing 5.4 lediglich der Zeichen-Vergleich in einer kurzen Schleife laufen, sodass dieser ohne Unterbrechung einige male nacheinander ausgeführt wird. Dabei muss natürlich nach jeder Überprüfung auch geschaut werden, ob der String bereits fertig durchlaufen wurde, damit gegebenenfalls die Folgeoperationen in der Pipeline ausgeführt werden können. Durch das Schlüsselwort `#pragma unroll`, ersetzt der Präprozessor die einfache Schleife durch Duplikate des Codes, wodurch eine weitere, kleine Leistungssteigerung erreicht wird.

```
1  #pragma unroll
2  for (int u = 0; u < ITERATION_COUNT; u++) {
3    int data_id = search_id + char_offset[current];
4
5    // when strings don't match, inactivate the lane
6    if (active && data_content[data_id] != search_string[search_id])
7      active = false;
8
9    search_id++;
10
11   if (search_id == search_length)
12     break;
13 }
14
15 if (search_id == search_length) {
16
17   /* execute following operators in the pipeline */
18
19   active = false;
20 }
```

Listing 5.4: Reduzierung des Overheads des Verfahrens

Die in diesem Kapitel vorgestellte Technik zur Verbesserung der Leistungsfähigkeit des einfachen String-Vergleichs liefert einen vielversprechenden Ansatz, um die Ausführung von String-Operationen in Pipelining-Umgebungen auf Grafikkarten effizient zu gestalten, ohne dabei auf Dictionaries zurückgreifen zu müssen. Ob dieses Verfahren die gewünschten Leistungsziele erreichen kann, wird in Kapitel 10 anhand einiger praktischer Anwendungsszenarien ermittelt werden.

Kapitel 6

Grundlagen von regulären Ausdrücken

Kapitel 7

Paralleler Musterabgleich mit regulären Ausdrücken

Aufbauend auf Kapitel 4 wird im Folgenden eine komplexere Operation in Form eines Matchers für reguläre Ausdrücke im Kontext der in Kapitel 3 beschriebenen Compiled Query Pipelines auf GPUs untersucht. Diese Operation stellt wieder eine Selektion über eine Spalte von String-Daten dar, bei der alle Tupel in die Ergebnisrelation übernommen werden, welche dem Muster eines vorgegebenen regulären Ausdrucks entsprechen.

Zunächst wird dazu erläutert, warum das Umsetzen einer solchen Operation im Kontext von Datenbanksystemen relevant ist. Anschließend wird die allgemeine Struktur der Operation dargestellt, gefolgt von der tatsächlichen Umsetzung des Musterabgleichs mithilfe von Automaten. Schließlich werden noch einige alternative Verfahren vorgestellt, welche unterschiedliche Eigenschaften aufweisen und mit der vorgestellten Umsetzung verglichen werden.

7.1 Motivation

Die Verwendung von regulären Ausdrücken eröffnet in vielen Anwendungsfällen verschiedenste Möglichkeiten, String-Daten effizient zu verarbeiten und die Leistungsfähigkeiten des Datenbanksystems voll auszunutzen. Reguläre Ausdrücke stellen ein mächtiges Werkzeug dar, welches den meisten Anwendern von Datenbankmanagementsystemen bekannt ist. Diese sind vor allem relevant, da sie wie in Kapitel **??** beschrieben das hoch effiziente Auswerten komplexer Muster erlauben.

Da viele Datenbankmanagementsysteme keine volle Unterstützung für reguläre Ausdrücke bieten, muss eine entsprechende Selektion nach dem Ausführen des Anfrageplans durch das DBMS manuell im Anwendungsprogramm durchgeführt werden. An dieser Stelle entstehen zahlreiche Probleme, da der Optimierer die Selektion nicht an der optimalen Stelle im Anfrageplan platzieren kann, damit frühzeitig Tupel weg fallen, die nicht in das

Ergebnis aufgenommen werden. Es entsteht also schon beim Datenbankserver eine erhöhte Last durch das unnötige Verarbeiten von Tupeln. Diese werden zusätzlich noch über das Netzwerk zur Anwendung übertragen, wodurch erneut eine erhöhte Netzlast entsteht. Die Anwendung wiederum muss anschließend mit einer potenziell geringeren Leistungsfähigkeit als der Datenbankserver die Selektion durchführen, wodurch an dieser Stelle wieder eine unnötig hohe Last entsteht. Sollten also Anwendungsfälle auftreten, bei denen eine Selektion durch reguläre Ausdrücke gewünscht ist, stellt die Unterstützung dieser Operation durch das Datenbankmanagementsystem einen massiven Vorteil dar.

7.2 Struktur der Operation

Zunächst wird die vorgestellte Operation ohne tiefgreifende Optimierungen implementiert, um damit eine Basis für das Entwickeln einer Verbesserung zu bieten. Für das Verarbeiten des regulären Ausdrucks wird ein deterministischer Automat erstellt, der genau dann akzeptiert, wenn der aktuell überprüfte String in das Muster des regulären Ausdrucks passt.

Das allgemeine Vorgehen der Operation ist ähnlich zu dem in Kapitel 4.2 beschriebenen einfachen String-Vergleich. Jedem Thread der GPU wir ein Tupel zugewiesen, welcher mithilfe des Automaten untersucht werden soll. Nachdem der aktuelle Zustand des DFA auf den Startzustand gesetzt wurde, wird der String Zeichen für Zeichen durchlaufen, wobei der Zustand in jedem Schritt entsprechend der Regeln des Automaten aktualisiert wird. Sobald die Zeichenkette vollständig durchlaufen wurde, wird geprüft, ob sich der Automat in einem akzeptierenden Zustand befindet, in welchem Fall der gesuchte String in das Muster des regulären Ausdrucks passt. Ist am Ende kein akzeptierender Zustand erreicht, oder wurde die Untersuchung vorzeitig abgebrochen, weil ein Fehlerzustand erreicht wurde, wird das Tupel nicht in das Ergebnis übernommen.

Bei der Ausführung wird der gesamte Warp parallel abgearbeitet, wodurch die Positionen, an denen die Strings untersucht werden für alle Lanes identisch sind. Außerdem wird das Ergebnis erst geschrieben, sobald alle Threads ihren String vollständig durchlaufen haben oder einen Fehlerzustand erreicht haben. Schließlich wird jedem Thread eine neue Zeichenkette zugewiesen, mit der das Verfahren wiederholt wird bis sämtliche Tupel abgearbeitet sind.

```
1  /* execute previous operators in the pipeline */
2
3  char *p = data_content + char_offset[loop_var];
4  char *pe = data_content + char_offset[loop_var + 1];
5
6  int cs = machine_start;
7
8  while(active) {
9    cs = singleDfaStep(cs, p);
10
11   p++;
12
13   if (p == pe)    // string completely processed
14     active = false;
15
16   if (cs == 0)    // invalid state reached
17     active = false;
18 }
19
20 active = cs >= machine_first_final;
21
22 /* execute following operators in the pipeline */
```

Listing 7.1: Naive Implementierung einer Selektion mit einem regulären Ausdruck

In Listing 7.1 wird die Implementierung des Operators vorgestellt, der im Kontext der kompilierten Anfragepipelines die Selektion über einen regulären Ausdruck ausführt. Die Datensätze `data_content` und `char_offset` enthalten wie in Kapitel 4.2 die Zeichenketten aus der zu untersuchenden Spalte und die Indizes der einzelnen Tupel innerhalb des Datensatzes.

Zunächst wird die Lane initialisiert, indem ein Zeiger `p` auf den Anfang des zu untersuchenden Strings und der Zeiger `pe` auf das Ende gesetzt wird. Außerdem wird der aktuelle Zustand `cs` auf den Startzustand des Automaten gesetzt.

In der Schleife wird anschließend über den gesamten String iteriert und dabei mithilfe der Methode `singleDfaStep` der Folgezustand des Automaten nach Einlesen des nächsten Zeichens bestimmt. Daraufhin wird überprüft, ob der iterierende Zeiger `p` auf das Ende des Strings `pe` zeigt, in welchem Falle dieser vollständig durchlaufen wurde und die Lane vorerst deaktiviert werden kann. Die Lane wird ebenfalls deaktiviert, wenn der Fehlerzustand `0` erreicht wurde, von dem aus ein Erreichen eines akzeptierenden Zustandes unmöglich ist.

Nachdem sämtliche Lanes ihre Berechnung abgeschlossen haben, wird überprüft, ob der aktuelle Zustand des DFA zu der Gruppe der akzeptierenden Zustände gehört. Ist dies der Fall wird die Lane für folgende Operationen aktiviert, ansonsten bleibt diese deaktiviert, sodass die Folgeoperationen nicht ausgeführt werden müssen.

7.3 Erstellen und Durchlaufen des Automaten

Das vorgestellte Verfahren generiert vor der Ausführung des Anfrageplans einen deterministischen Automaten, welcher beim kompilieren der Anfragepipeline in den Kernel eingebaut wird. Der Automat wird mithilfe des Ragel State Machine Compilers [9] erzeugt, auf dem auch der Code zur Verarbeitung des Automaten basiert.

Ragel ermöglicht es, für das Auswerten eines gegebenen regulären Ausdrucks C-Code zu erzeugen, der automatisch in ein vorgegebenes Rahmenprogramm eingefügt wird. Somit ist es leicht möglich, den Automaten in den Rahmen einer kompilierten Anfragepipeline einzupflegen, es müssen also keinerlei Schnittstellen zu anderen Sprachen oder Konzepten erstellt werden.

Der generierte Code lässt sich in zwei Hauptbestandteile aufteilen. Zum einen wird etwas Code zur tatsächlichen Durchführung des Musterabgleichs generiert, welcher für unterschiedliche Ausdrücke größtenteils identisch ist. Außerdem werden einige Tabellen generiert, welche die tatsächlichen Zustände und Zustandsübergänge des Automaten beinhalten und für jeden regulären Ausdruck neu generiert werden müssen.

```
1  __device__ int singleDfaStep(int cs, char* p) {
2    int _slen;
3    int _trans;
4    const char *_keys;
5    const char *_inds;
6
7    _keys = _machine_trans_keys + (cs<<1);
8    _inds = _machine_indicies + _machine_index_offsets[cs];
9
10   _slen = _machine_key_spans[cs];
11   _trans = _inds[ _slen > 0 && _keys[0] <=(*p) &&
12     (*p) <= _keys[1] ?
13     (*p) - _keys[0] : _slen ];
14
15   return _machine_trans_targs[_trans];
16 }
```

Listing 7.2: Methode zur Durchführung eines DFA-Schrittes

Listing 7.2 zeigt die Methode zur Durchführung eines DFA-Schrittes, welche in Listing 7.1 verwendet wird. Die Implementierung basiert auf dem durch Ragel erzeugten Ausführungscode, welcher mit der *flat*-Einstellung generiert wurde. Dies stellt den simpelsten Code dar, der von Ragel generiert wird, welcher sehr ähnlich zu dem in Kapitel **??** vorgestellten Prinzip ist. Hier werden die ebenfalls von Ragel generierten und in Listing 7.3 dargestellten Felder verwendet, welche den eigentlichen DFA enthalten. In diesem Beispiel handelt es sich um den Automaten, der aus dem Ausdruck `(0|1)*((00)+|001)0` generiert wurde.

```
1  static const char _machine_trans_keys[] = {
2    0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0
3  };
4
5  static const char _machine_key_spans[] = {
6    0, 2, 2, 2, 2, 2, 2
7  };
8
9  static const char _machine_index_offsets[] = {
10   0, 0, 3, 6, 9, 12, 15
11 };
12
13 static const char _machine_indicies[] = {
14   0, 2, 1, 3, 2, 1, 4,
15   5, 1, 6, 2, 1, 4, 5, 1,
16   3, 2, 1, 0
17 };
18
19 static const char _machine_trans_targs[] = {
20   2, 0, 1, 3, 5, 4, 6
21 };
22
23 static const int machine_start = 1;
24 static const int machine_first_final = 5;
```

Listing 7.3: Methode zur Durchführung eines DFA-Schrittes

Da diese Darstellung des Automaten für den Menschen kaum lesbar ist und ein Debugging somit sehr aufwändig werden würde, bietet Ragel außerdem ein Werkzeug zur Visualisierung des Graphen. Abbildung 7.1 zeigt dazu die visuelle Darstellung des erzeugten DFA.

7.4 Alternative Verfahren

Zusätzlich zur *flat*-Einstellung, erlaubt Ragel außerdem eine Generierung von Code, welche für die Verarbeitung regulärer Ausdrücke mit einem großen Alphabet optimiert wurde. Diese Option kann über die *table*-Einstellung abgerufen werden. Im Gegensatz zur *flat*-Einstellung wird das aktuell untersuchte Zeichen nicht als Index für einen Array von Zustandsübergängen verwendet, sondern in diesem Array eine binäre Suche nach der korrekten Transition durchgeführt. Nach Thurston ist die oben beschriebene *flat*-Option generell schneller, sie lässt sich allerdings nur für ein kleines Alphabet anwenden [9]. Da im Datenbankkontext generell eher simplere Ausdrücke zu erwarten sind, ist es für die meisten Anwendungsfälle zwar ausreichend, die einfache und schnelle Implementierung zu wählen, es wäre aber auch interessant zu untersuchen, wie groß der Leistungsverlust mit

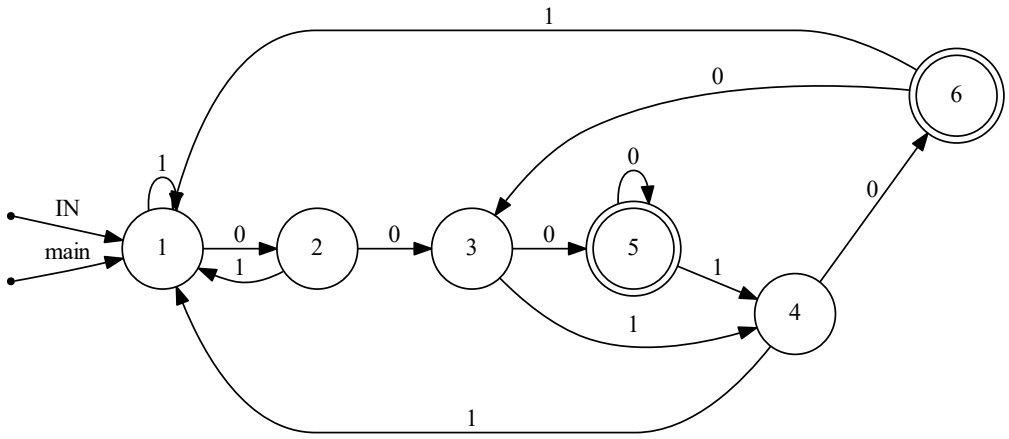

Abbildung 7.1: Visuelle Darstellung des DFAs zum regulären Ausdruck `(0|1)*((00)+|001)0`

der alternativen Implementierung aussieht, falls komplexere Ausdrücke untersucht werden sollen.

Als Alternative zur vollumfänglichen Unterstützung von regulären Ausdrücken bieten viele Datenbankmanagementsysteme den *LIKE*-Operator an, welcher den einfachen String-Vergleich um Platzhalter erweitert. So ist es dem Nutzer möglich anstelle des genauen Suchstrings ein Muster anzugeben, in dem Platzhalter für beliebige Zeichen enthalten sind. Mithilfe dieser simpel umzusetzenden Operation lassen sich viele einfache Abfragen an eine Datenbank modellieren, sodass in vielen Fällen kein richtiger regulärer Ausdruck benötigt wird. Es wäre daher interessant zu sehen, wie sich die Leistungsfähigkeit dieser Operation mit geringerem Funktionsumfang gegenüber der Umsetzung eines vollständigen Matchers für reguläre Ausdrücke verhält.

Kapitel 8

Verbesserung des Verfahrens zum Musterabgleich

Bei dem in Kapitel 7 vorgestellten Verfahren tritt ein ähnliches Problem auf wie bei dem zuvor beschriebenen, einfachen String-Vergleich. Aufgrund der unterschiedlichen Struktur von Strings werden einige Lanes innerhalb eines Warps früher als andere Lanes inaktiv wenn sie einen Fehlerzustand oder das Ende des Eingabestrings erreicht haben. Dies hat eine Unterauslastung des Warps zur Folge, wodurch Rechenleistung verschwendet wird. Aus diesem Grund ist es wünschenswert den inaktiv gewordenen Threads dynamisch neue Arbeit zuzuweisen. Dazu wird das in Kapitel 5 vorgestellte Verfahren zum Einsatz kommen, wodurch im Rahmen der kompilierten Anfragepipelines die Laufzeit optimiert werden kann. Das von Lang et al. entwickelte Verfahren [5] wird im Folgenden weiterhin als *Lane Refill* bezeichnet.

8.1 Struktur des optimierten Musterabgleichs mit Lane Refill

Die grundlegende Funktionsweise des Lane Refill wurde in Kapitel 5.1 beschrieben und ist genau so auch auf den parallelen Musterabgleich anwendbar. Bei dem Verfahren wird eine Lane immer dann inaktiv, wenn der untersuchte String vollständig durchlaufen wurde oder ein Fehlerzustand erreicht wurde und es nicht mehr möglich ist, einen akzeptierenden Zustand zu erreichen. Nach genau diesen Ereignissen muss überprüft werden, ob die gewünschte Auslastung des Warps unterschritten wird und gegebenenfalls die aktuellen Elemente in den Puffer geschrieben oder neue Elemente aus dem Puffer geladen werden.

```
1  while(buffercount + numactive > THRESHOLD) {
2    if (numactive < THRESHOLD) {
3
4      /* refill empty lanes from buffer in case of underutilization */
5
6      bufferelements = bufferelements - numrefill;
7    }
8
9    if (active) {
10     cs = singleDfaStep(cs, p);
11
12     if (cs == 0)    // invalid state reached
13       active = false;
14   }
15
16   p++;
17
18   if (active && p == pe) {    // string completely processed
19     if (cs >= machine_first_final) {  // finishes with accepting state
20
21       /* execute following operators in the pipeline */
22
23     } else {  // finishes with non accepting state
24       active = false;
25     }
26   }
27
28   numactive = __popc(__ballot_sync(ALL_LANES, active));
29 }
```

Listing 8.1: Struktur des Musterabgleichs mit Lane Refill

Die allgemeine Struktur des Kernels ist identisch zu der in Listing 5.1 vorgestellten Struktur des einfachen String-Vergleichs. Um den parallelen Musterabgleich damit umzusetzen, muss die innere Schleife wie in Listing 8.1 dargestellt angepasst werden.

Zunächst wird hier von allen aktiven Lanes ein Schritt im DFA durchgeführt und überprüft, ob ein Fehlerzustand erreicht wurde. Anschließend wird für die vollständig durchlaufenen Strings überprüft, ob diese sich in einem akzeptierenden Zustand befinden und gegebenenfalls die Folgeoperationen der Pipeline ausgeführt.

Die technische Umsetzung der Puffer-Operationen funktioniert identisch zu dem in Kapitel 5.3 beschriebenen Vorgehen, mit dem Unterschied, dass hier der Inhalt der Variablen `p`, `pe` und `cs` im Puffer gespeichert werden. Eine Reduzierung des Overheads durch die Puffer-Operation ist ebenfalls analog zu dem in Kapitel 5.4 vorgestellten Verfahren möglich.

Kapitel 9

Optimierung der Ausführungsparameter

Die Ausführungszeit der in den vorherigen Kapiteln vorgestellten Algorithmen wird stark durch unterschiedliche Parameter beeinflusst. Den größten Einfluss nimmt dabei die Art des Datensatzes, also wie viele Matches er enthält, wie lang die enthaltenen Zeichenketten sind, wie viele Strings im Datensatz vorhanden sind und wie die Matches darin verteilt sind. Wie sehr unterschiedliche Datensätze die Ausführungszeit beeinflussen soll in Kapitel 10 untersucht werden.

Neben dem Aussehen des Datensatzes gibt es noch Parameter, die bei der Ausführung des Algorithmus der GPU übermittelt werden und dort ebenfalls einen starken Einfluss auf die Laufzeit nehmen können. Diese Parameter bestehen wie in Kapitel 2.2 beschrieben in der Anzahl der Threads pro Block (*Block Size*) und der Anzahl der Blöcke im Grid (*Grid Size*). Aus den Parametern setzt sich die *Grid-Konfiguration* zusammen, welche durch das Ausprobieren unterschiedlicher Werte optimiert werden kann.

Um dieses Vorgehen zu veranschaulichen, soll anhand eines Beispiels gezeigt werden, wie eine möglichst optimale Grid-Konfiguration gefunden werden kann und wie unterschiedliche Konfigurationen einen Einfluss auf die Laufzeit nehmen. Abbildung 9.1 zeigt dazu die Laufzeit für den String-Vergleichsalgorithmus in Millisekunden, welcher für unterschiedliche Grid-Konfigurationen auf dem in Kapitel 10.2 vorgestellten Type-Datensatz mit einer Selektivität von 0.25% ausgeführt wurde. Besonders hohe Laufzeiten sind dabei in rot und besonders niedrige Laufzeiten in grün eingefärbt.

Es fällt auf, dass eine niedrige Block Size in Kombination mit einer niedrigen Grid Size zu einer hohen Laufzeit führt. Außerdem gibt es bei einer Grid Size von unter 100 einige Kombinationen, die besonders geringe Laufzeiten aufweisen, allerdings von weniger guten Konfigurationen umgeben sind. Mit einer Grid Size zwischen 2.000 und 200.000 in Verbindung mit einer Block Size zwischen 64 und 512 werden generell ordentliche Laufzeiten

erzielt. Wird allerdings eine Block Size über 512 gewählt, nimmt die Leistungsfähigkeit des Systems wieder ab.

Generell ist die Analyse eines solchen Ergebnisses schwierig, da die Grid-Konfiguration Einfluss auf unterschiedlichste Bereiche der GPU nimmt und der CUDA-Optimierer einige Effekte erfolgreich versteckt. Die hohe Laufzeit des Algorithmus bei einer geringen Anzahl von Threads ist dadurch zu erklären, dass zunächst gar nicht alle Kerne der GPU ausgelastet sind, weil nicht genügend Warps für die Anzahl der SM vorhanden sind. Eine steigende Anzahl von Threads führt zwar dazu, dass alle Kerne ausgelastet sind, allerdings können bei Speicherzugriffen nicht genügend Warps vom Scheduler ausgetauscht werden, sodass wie in Kapitel 2.2 beschrieben die Latenz der Speicherzugriffe erfolgreich versteckt werden könnte. In dem zuvor beschriebenen Bereich, in dem ordentliche Laufzeiten erzielt werden, steht eine ausreichende Anzahl von Threads zur Verfügung, um eventuelle Latenzen zu verstecken und somit die GPU bestmöglich auszulasten. Wie der Anstieg der Laufzeiten bei einer Block Size von über 512 zu begründen ist, wird hier leider nicht klar.

Für die Verwendung des Algorithmus sollte im Allgemeinen eine Grid-Konfiguration aus dem Bereich gewählt werden, welcher eine durchweg ordentliche Performanz erzielt. Dadurch ist die Wahrscheinlichkeit hoch, dass eine Konfiguration gewählt wird, mit der eine Laufzeit erreicht wird, welche bis auf eine kleinere Abweichung dem Optimum entspricht. Die Position des Bereichs ändert sich für unterschiedliche Selektivitäten des Datensatzes nur geringfügig, wie in zahlreichen Tests untersucht wurde und beispielhaft in Anhang B für eine Selektivität von 64% analog zu Abbildung 9.1 dargestellt wird. Aus diesem Grund kann die Grid-Konfiguration schon vor der Ausführung bestimmt werden und eine nah am Maximum liegende Leistungsfähigkeit erzielt werden. Bei den Leistungsmessungen in den folgenden Kapiteln sollte sich allerdings nicht darauf verlassen werden, dass durch diese Heuristik ein nahezu optimaler Wert gefunden wird, weshalb für die Tests jeweils das Optimum aus einer Auswahl von Grid-Konfigurationen bestimmt wird. Dazu werden alle Konfigurationen mit einer Grid Size von {1.000, 2.000, 3.000, 4.000, 6.000, 8.000, 10.000, 20.000, 50.000, 100.000, 150.000, 200.000} und einer Block Size von {32, 64, 96, 128, 160, 192, 224, 256, 384, 512, 640, 768} überprüft und das Optimum der Messungen für die Auswertung der aufgestellten Statistiken gewählt.

Grid Size \ Block Size	32	64	96	128	160	192	224	256	384	512	640	768	896	1024
10	806	506	352	267	242	187	166	146	102	81	76,2	59,9	54	49,8
12	806	406	283	212	185	150	133	117	82,5	65,5	57,9	48,2	43,2	39,7
14	690	366	250	191	165	137	120	106	76,1	61,1	54,7	80,9	67,9	61,3
16	604	318	219	169	144	121	105	92,5	67,1	53,6	47,7	71,1	63,5	55,3
18	537	283	196	150	128	107	93,3	82,4	59,6	47,8	42	63,3	56,8	51,9
20	498	267	187	145	131	102	89,9	80,6	58,8	48,6	75,6	59,1	52,9	49,5
30	344	186	129	99,5	92,6	71,9	64,5	56,5	41,6	60,3	54,6	58,9	52,9	46,3
40	263	145	101	80,5	74	57,7	51,8	45,6	61,2	47,3	60,4	58,9	52,7	47,1
60	185	99,5	71,9	56,4	53,3	41,4	66,7	59,4	41,7	46,3	53,9	49,9	45,2	42,6
80	145	78,7	57,9	45,6	67,7	57	52	46	46,4	46,9	50,4	52,1	48,2	44,6
100	124	74,2	53,3	67,8	65,2	51,6	46,7	57,6	53,2	50,6	53,6	57,9	52	49
150	92	53,5	61,6	51,3	61,2	51,8	46,8	51,7	46,8	48,8	53,4	56	50	46,9
200	111	64,9	51,2	54,8	53,5	51,8	46,7	49,7	47,8	48,4	53,5	54,9	49,6	46,7
300	86,9	52,1	51,4	51	56,1	47,4	47,4	48,4	46,9	47	50,9	54,6	48,7	46,4
400	90,9	53,9	51,7	49,8	54,5	47,2	48,1	48,3	48	47	50,7	53,9	48	46,4
500	84,5	53,5	53,7	54,8	60,2	54,1	53,4	52,3	50,8	51,2	59,9	57,7	51,4	49,1
1000	80,3	55,9	55,8	54,3	58,3	50,4	51,8	51,5	49,6	49	58,1	57,4	49,5	47,4
2000	78,5	56,5	54,3	53,2	54,2	47,9	49,2	49	48,1	46,9	55	54,1	48,5	45,5
3000	77,6	55,9	53,2	50,9	52,9	47,3	47,3	46,6	47	45,7	52,9	54	48	45,1
4000	76,1	54,5	51,6	49,5	49,4	46,1	46	45,5	46,7	44,5	52,1	53,8	50	46,1
6000	75,3	52,6	48,5	45,5	47,4	44,9	44,9	44,5	44,5	43,9	51,9	55,3	49,7	47,8
8000	74,9	50,7	45,4	44,2	46,2	45,5	44	43,4	44,5	44,8	51,8	56,8	51,7	49,2
10000	74,8	49,7	46	45,3	47	45,1	44,5	45	48,2	50,1	58,6	61,1	55,6	52,6
20000	73,2	45,6	44,3	42,1	44,8	43,5	44,3	44,7	48	49,1	57,9	61,2	56,6	54
50000	69,1	47,1	43,4	43,7	44,5	45,2	45,8	46,3	49,5	52,8	57,8	65,7	67	61,5
100000	64,9	47	43,1	43,5	45,4	45	45,9	46,4	49,6	52,5	56,9	66,2	66,8	62,4
150000	65	46,8	43,1	43,5	44,6	45,2	46	46,6	49,3	52,5	57,8	69,6	69	64,9
200000	64,6	46,7	43,2	43,7	44,7	45,5	46,1	47,1	49,5	53,7	59,2	70,2	71,4	67,4

Abbildung 9.1: Laufzeit des naiven String-Vergleichsalgorithmus in ms für den Type-Datensatz mit einer Selektivität von 0.25% unter Verwendung unterschiedlicher Grid-Konfigurationen

Kapitel 10

Evaluation des einfachen String-Vergleichs

In Kapitel 5 wurde eine Technik vorgestellt, von der zu erwarten ist, dass sie die Laufzeit des einfachen String-Vergleichs verbessert, indem die Ressourcen der GPU besser genutzt werden und somit eine erhöhte Auslastung erreicht wird. Da diese Technik allerdings einen gewissen Overhead mit sich bringt, bleibt es noch zu untersuchen, ob diese Technik tatsächlich eine bessere Laufzeit erzielt, oder ob der Mehraufwand so groß ist, dass die erreichten Vorteile von diesem überschattet werden. In diesem Kapitel wird diese Untersuchung anhand realer Arbeitslasten durchgeführt und außerdem überprüft, ob die in Kapitel 5.4 vorgestellte Reduzierung des Overheads eine weitere Leistungssteigerung mit sich bringt.

10.1 Testumgebung

Für die Durchführung der Leistungsmessungen wird der Algorithmus so angepasst, dass dieser lediglich die Anzahl der passenden Zeichenketten zählt und diese am Ende ausgibt. Die Testumgebung entspricht also einer Selektion auf einer Spalte einer Relation und dem anschließenden Zählen der Ergebnisse. Dieses Vorgehen hat den Vorteil, dass das Zählen der Ergebnisse nicht viel Rechenaufwand verursacht und somit die Leistungsmessungen möglichst wenig verfälscht wird. Trotzdem bleibt es aber möglich aufgrund der Ausgabe des Algorithmus beurteilen zu können, ob der Test korrekt durchgeführt wurde.

Sämtliche Tests wurden auf einem Computer durchgeführt, welcher eien NVIDIA GTX 950 verbaut hat und als Betriebssystem Ubuntu 18.04 verwendet.

10.2 Verwendete Workloads und deren Merkmale

In analytischen Anwendungsfällen kommen häufig selektive Filter vor [1], weshalb diese ebenfalls für diese Untersuchungen verwendet werden. Außerdem ist zu erwarten, dass

diese besonders stark vom Lane Refill profitieren werden, da bei einer kleinen Menge von Ergebnissen oftmals eine starke Unterauslastung auftritt.

Der erste verwendete Workload, welcher im Folgenden *Type* genannt wird, wurde aus dem TPC-H-Benchmark entnommen.[1] Hier wird eine Selektion über die Type-Spalte durchgeführt, welche Zeichenketten der Länge 16-25 enthält. Diese bestehen aus den Zeichen *A-Z* und dem Leerzeichen. Für die Untersuchung wurde ein Datensatz mit 90.000.000 Tupeln generiert.

Ein weiterer Workload wurde aus dem Datensatz von DBLP[2] erstellt, welcher die Titel vieler Veröffentlichungen im Informatik-Umfeld enthält. Dazu wurden doppelte Titel entfernt und die übrigen Strings so angepasst, dass diese nur noch Kleinbuchstaben enthalten. Die durchschnittliche Länge der Zeichenketten in diesem Datensatz beträgt 76 Zeichen und es wurde ein Präfix untersucht, welcher 31 Zeichen beinhaltet. Der generierte Datensatz enthält schließlich 21.513.695 Tupel.

Um die unterschiedlichen Selektivitäten für die folgenden Tests zu erreichen wurde ein neuer String entsprechend zufällig verteilt in den Datensatz eingebracht. Die gewünschte Datengröße wurde schließlich erreicht, indem der so generierte Datensatz einige male vervielfacht wurde.

10.3 Vorstellung der Messergebnisse

In den Abbildungen 10.1 bis 10.3 sind die Ergebnisse der durchgeführten Messungen aufgetragen. Hierzu wurden die verschiedenen Algorithmen verglichen und auf ihr Verhalten bei unterschiedlichen Anteilen von passenden Strings im Datensatz untersucht. Als Vergleichsgröße wurde die Laufzeit der Algorithmen verwendet, welche ein möglichst allgemein nutzbares und neutrales Maß darstellt.

10.3.1 Gleichheitstest mit dem Type-Datensatz

In Abbildung 10.1 werden die Ergebnisse eines Tests der vorgestellten Algorithmen bei Verwendung des vorher beschriebenen *Type*-Datensatz ausgewertet. Dabei werden die Laufzeiten der naiven Umsetzung mit denen der verbesserten Variante mit Lane Refill verglichen. Außerdem wurde in der Messung die in Kapitel 5.4 beschriebene Variante zur Reduzierung des Overheads berücksichtigt, welche einmal so durchgeführt wurde, dass immer zwei Zeichen pro Schritt verglichen werden und einmal in Dreierschritten.

Sofort fällt auf, dass eine höhere Anzahl von Matches im Datensatz eine erhöhte Laufzeit für den Algorithmus bedeutet. Außerdem liegt die Laufzeit der naiven Umsetzung bedeutend höher als die der drei optimierten Varianten, welche das Lane Refill-Verfahren

[1] `http://www.tpc.org/tpch/`
[2] `https://dblp.org/`

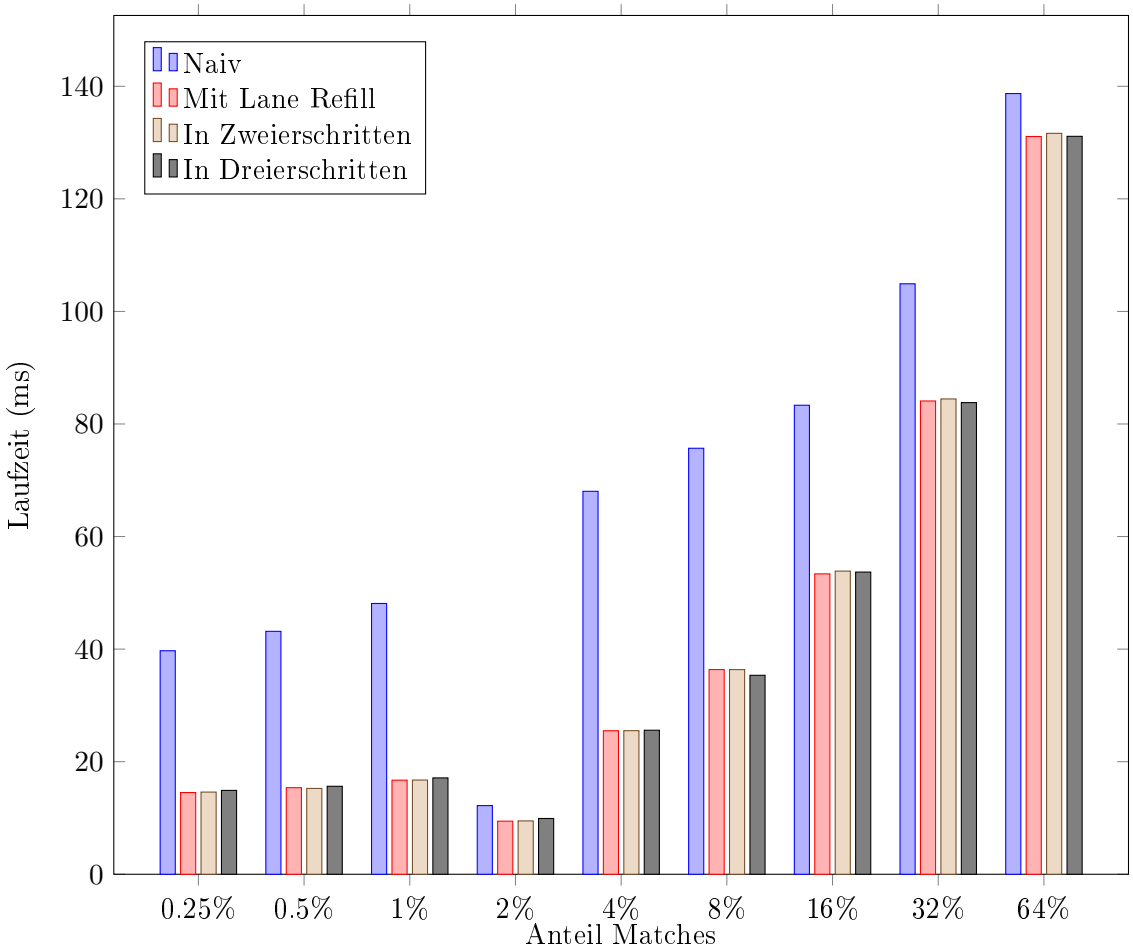

Abbildung 10.1: Laufzeit für Gleichheitstest mit verschiedener Verteilung beim Type-Benchmark

verwenden. Der Algorithmus, welcher das Lane Refill implementiert ist zwischen 43% und 4% schneller als die naive Implementierung. Dabei entsteht der größte Vorteil bei einem geringen Anteil von passenden Elementen, welcher geringer wird, sobald der Datensatz eine höhere Anzahl von Matches enthält. Bis zu einem Anteil von 8% beträgt die Verbesserung der Laufzeit, die durch das Lane Refill-Verfahren erreicht wurde mehr als 50%. Es ist außerdem zu erkennen, dass die Laufzeit der beiden Varianten, welche die Reduzierung des Overheads erzielen sollten, eine nahezu identische Laufzeit erreichen, wie die erste Version, welche Lane Refill verwendet. Schließlich entsteht bei der Messung ein Ausreißer bei einem Anteil passender Matches von 2%, welcher sich durch besonders geringe Laufzeiten für alle Algorithmen auszeichnet.

10.3.2 Präfixtest mit dem Type-Datensatz

Abbildung 10.2 zeigt das Ergebnis eines ähnlichen Tests wie der im letzten Abschnitt beschriebene Test, mit der Ausnahme, dass hier der in Kapitel 4.3 beschriebene Präfixtest untersucht wurde. Der verwendete *Type*-Datensatz blieb dabei unverändert, es wurden

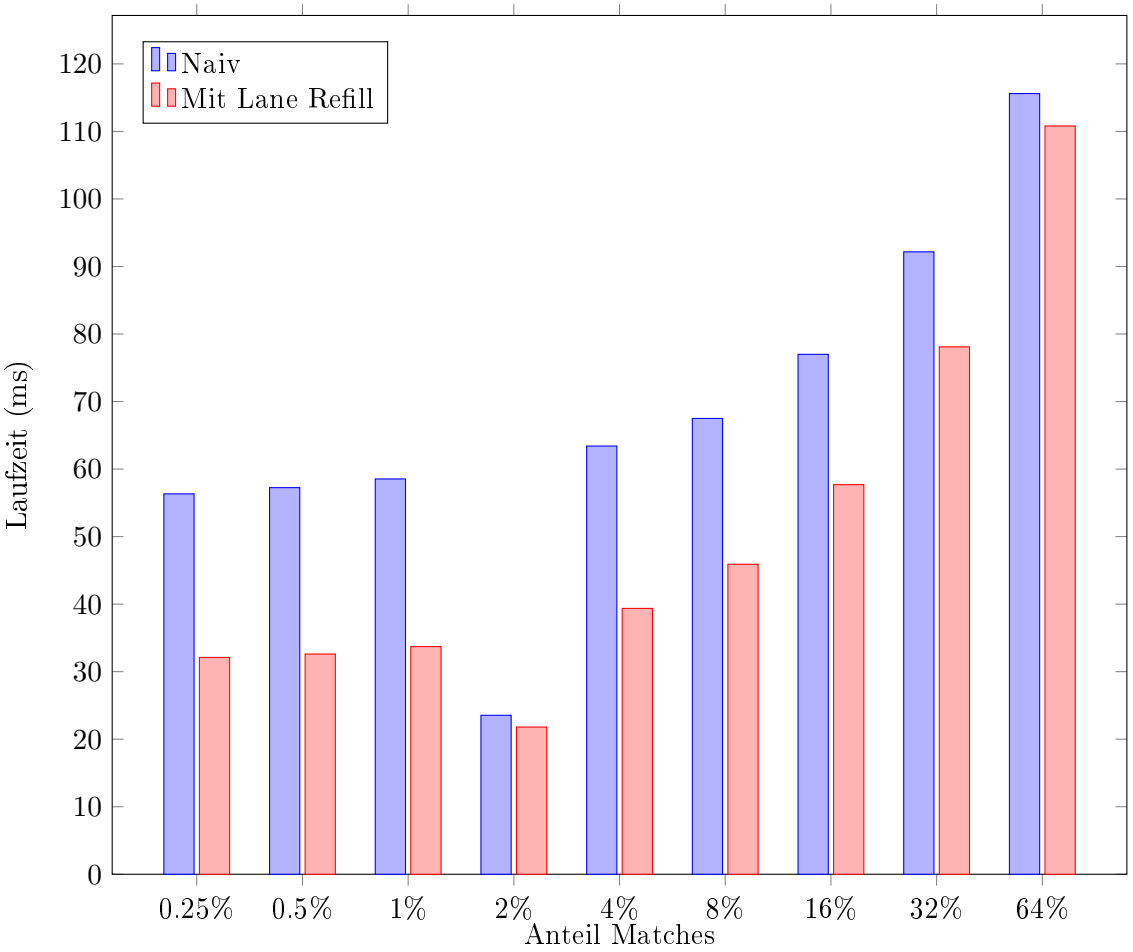

Abbildung 10.2: Laufzeit für Präfixtest mit verschiedener Verteilung beim Type-Benchmark

lediglich die Messungen für die Implementierung zur Reduzierung des Overheads des Lane Refill-Verfahrens ausgelassen.

Wie bei dem vorherigen Benchmark ist zu erkennen, dass eine höhere Anzahl von Matches hier auch eine erhöhte Laufzeit bedeutet und dass die Laufzeiten der naiven Implementierung bedeutend über denen der durch das Lane Refill optimierten Umsetzung liegt. Die verbesserte Variante des Algorithmus erzielt hier eine Reduzierung der Laufzeit zwischen 43% und 4%, wobei auch hier die Differenz der beiden Verfahren geringer wird, wenn ein höherer Anteil von Matches im Datensatz vorhanden ist. Bis zu einem Anteil von 8% zutreffender Strings im Datensatz ist der Vorteil allerdings noch höher als 30%. Für einen Anteil von 2% ist wie im vorherigen Test ebenfalls ein Ausreißer zu erkennen.

10.3.3 Präfixtest mit dem DBLP-Datensatz

In Abbildung 10.3 wird schließlich das Ergebnis eines Benchmarks dargestellt, welcher den Präfixtest mit dem *DBLP*-Datensatz untersucht. Wie in den vorherigen Tests, ist zu erkennen, dass ein höherer Anteil von Matches im Datensatz eine höhere Laufzeit verursacht

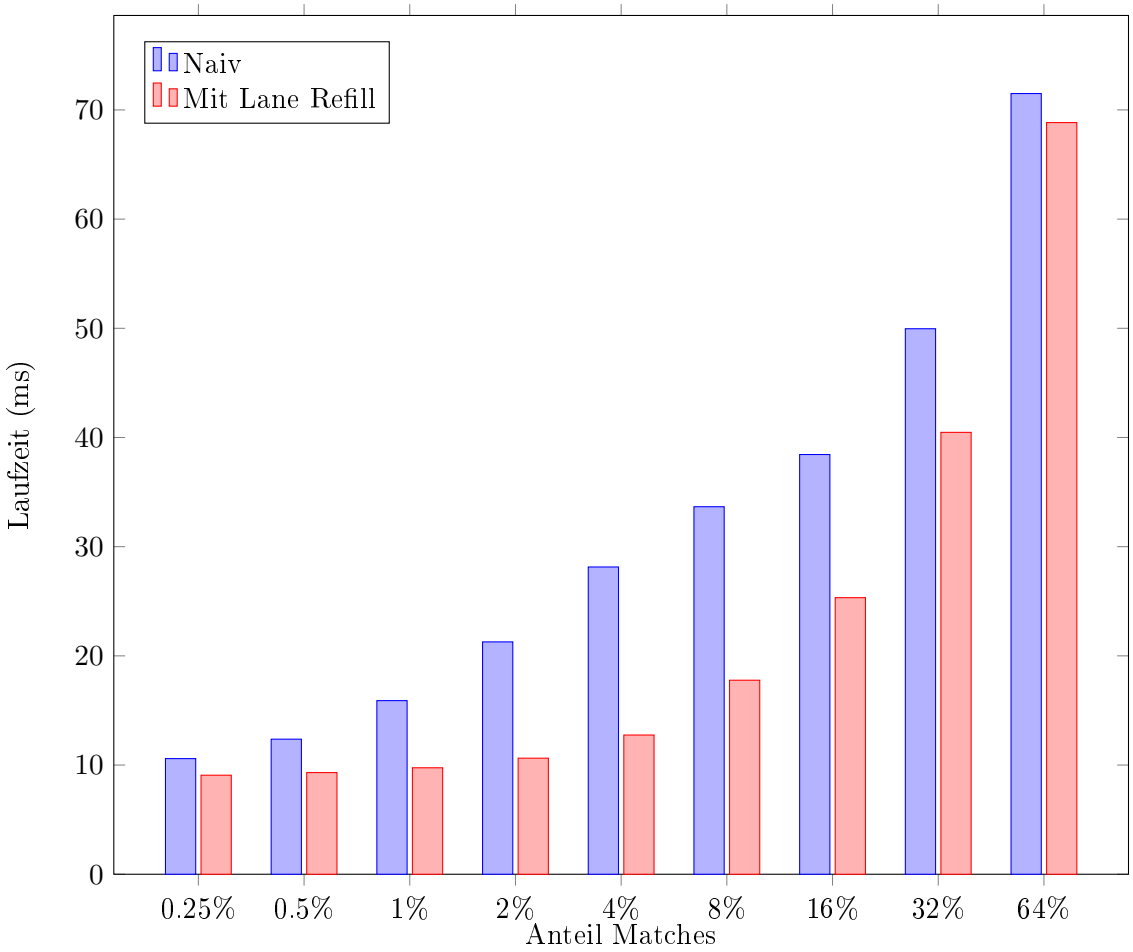

Abbildung 10.3: Laufzeit für Präfixtest mit verschiedener Verteilung beim DBLP-Benchmark

und dass das Einführen des Lane Refill eine deutliche Verringerung der Laufzeiten bewirkt. Diese Verbesserungen liegen zwischen 54% und 3%, wobei die größte Aufspaltung im mittleren Bereich der Selektivität liegt. Zwischen 1% und 16% zutreffender Elemente im Datensatz beträgt die Verbesserung mehr als 30%.

10.4 Diskussion der Ergebnisse

Die Ergebnisse der durchgeführten Tests lassen eindeutige Rückschlüsse auf das Verhalten der zuvor vorgestellten Verfahren für unterschiedliche Anwendungsfälle ziehen. Die geringeren Laufzeiten bei einer niedrigeren Anzahl von Matches lassen sich darauf zurückführen, dass viele untersuchte Strings vorzeitig verworfen werden. Somit ist bei der naiven Umsetzung die Anzahl der Iterationen geringer, welche tatsächlich bis zum Ende des Vergleichsstrings durchlaufen werden müssen. Auch die verbesserte Variante profitiert davon, dass die Zeichenketten nicht vollständig durchlaufen werden müssen, sodass häufiger neue Strings nachgeladen werden können und weniger tatsächliche Arbeit zu erledigen ist.

Der durchgehend erkennbare Leistungsgewinn, der durch das Lane Refill erreicht wird, lässt erkennen, dass das Erhöhen der Auslastung der Warps eine Steigerung der Performanz mit sich gebracht hat. Dies lässt darauf schließen, dass die Verbesserung der Auslastung einen signifikanteren Einfluss auf die Leistung hat, als der durch das Verfahren zusätzlich generierte Overhead. Der geringe Einfluss des Overheads lässt sich außerdem daran erkennen, dass die in Abbildung 10.1 untersuchten Verbesserungen des Lane Refill-Verfahrens keine signifikante verringerung der Laufzeit erbringen. Die Speicherzugriffe, welche für den eigentlichen String-Vergleich durchgeführt werden müssen, haben somit einen bedeutend höheren Einfluss auf die Laufzeit als die Instruktionen, die für das Lane Refill ausgeführt werden.

Ebenfalls in allen Diagrammen erkennbar ist, dass die Leistungssteigerung bei einer höheren Anzahl von Matches geringer wird, was darauf zurückzuführen ist, dass bei einem hohen Anteil zutreffender Elemente keine so starke Unterauslastung bei dem naiven Algorithmus auftritt. Dies liegt daran, dass in einem Warp zu jeder Zeit eine höhere Anzahl von zutreffenden Strings geprüft wird und somit die Zahl der wartenden Lanes geringer wird. Es lässt sich also in diesem Fall gar keine so große Verbesserung mehr durch das Einführen von Lane Refill erreichen, da keine große Unterauslastung vorhanden ist, die es zu beseitigen gilt.

Der Ausreißer bei einem Anteil von 2% zutreffender Elemente im *Type*-Datensatz tritt aufgrund eines Fehlers innerhalb des Datensatzes auf. Es wurde versucht diesen zu beheben, indem der Datensatz analysiert, überprüft und neu generiert wurde, allerdings ließ sich keine Lösung finden, welche den Ausreißer beseitigt hätte. Die restlichen Datenpunkte behalten dennoch ihre Gültigkeit und an ihnen lässt sich das generelle Verhalten der Verfahren einwandfrei ablesen.

Kapitel 11

Evaluation des parallelen Musterabgleichs

11.1 Verwendete Workloads und deren Merkmale

11.1.1 Vergleich der Basisalgorithmen mit dem DBLP-Datensatz

11.1.2 Verbesserung der Algorithmen durch das Lane Refill

11.1.3 Vergleich der optimierten Algorithmen

11.1.4 Einfluss beliebiger Anfangszeichen in dem DBLP-Datensatz

11.1.5 Vergleich verschiedener Automatengrößen mit dem TPC-H-Datensatz

11.2 Vorstellung der Messergebnisse

11.3 Diskussion der Ergebnisse

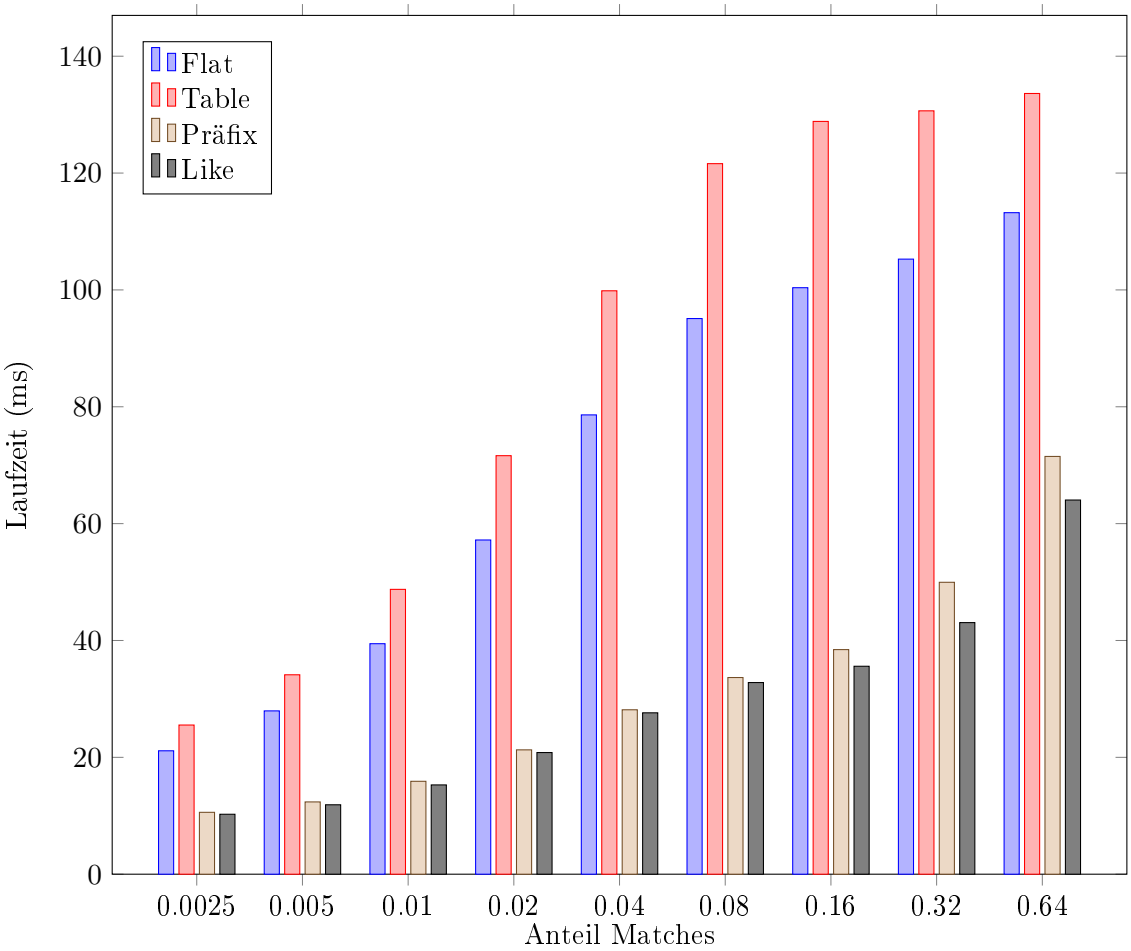

Abbildung 11.1: Laufzeit für Präfixtest mit Basisalgorithmen für den DBLP-Datensatz

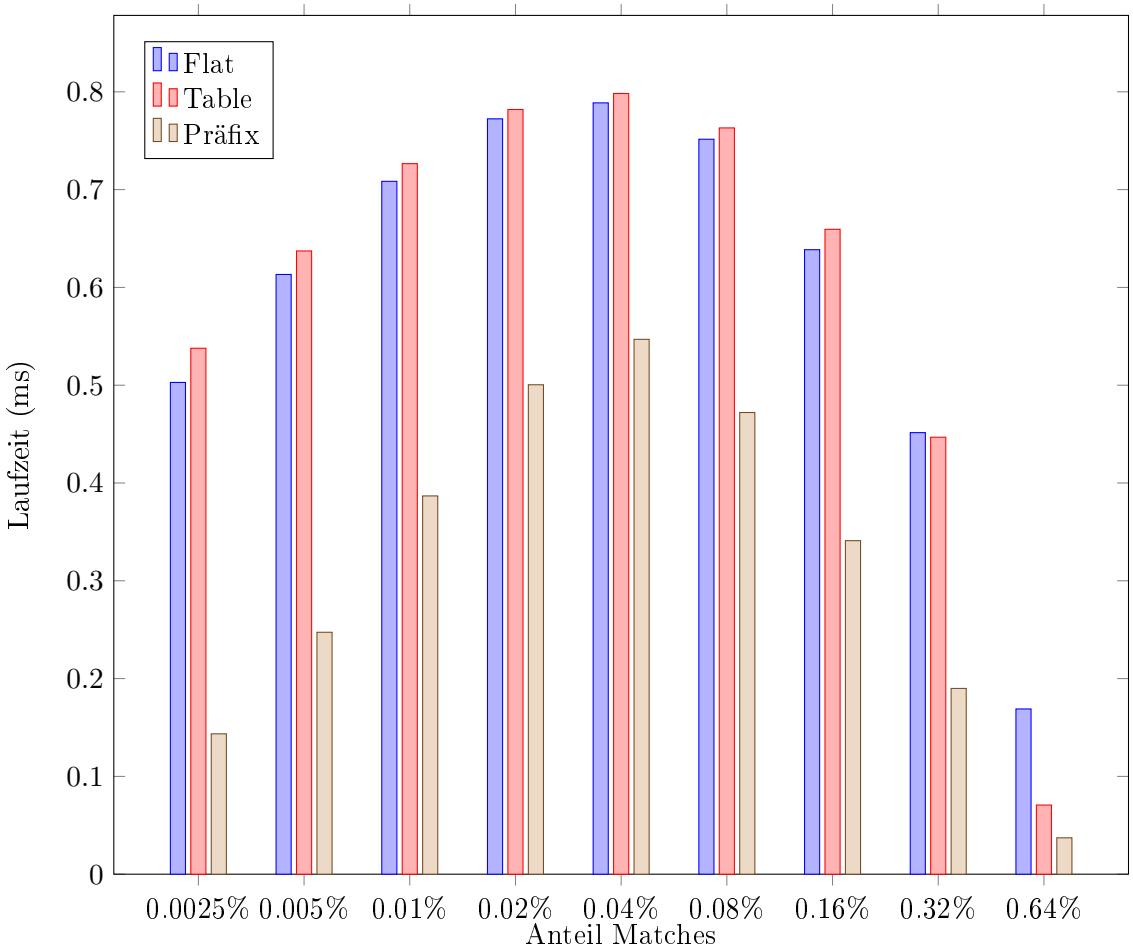

Abbildung 11.2: Verbesserungen der Algorithmen durch das Lane Refill für den DBLP-Datensatz

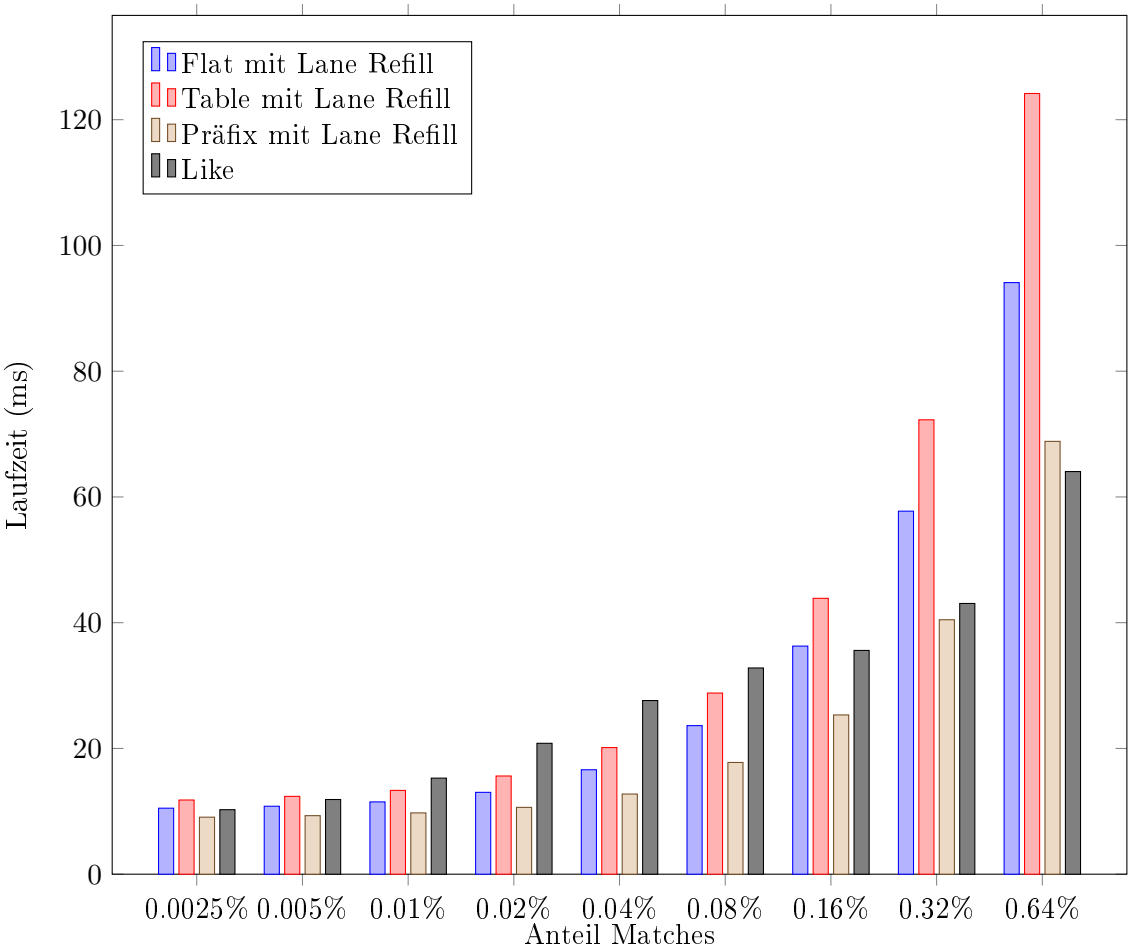

Abbildung 11.3: Laufzeiten der optimierten Algorithmen für den DBLP-Datensatz

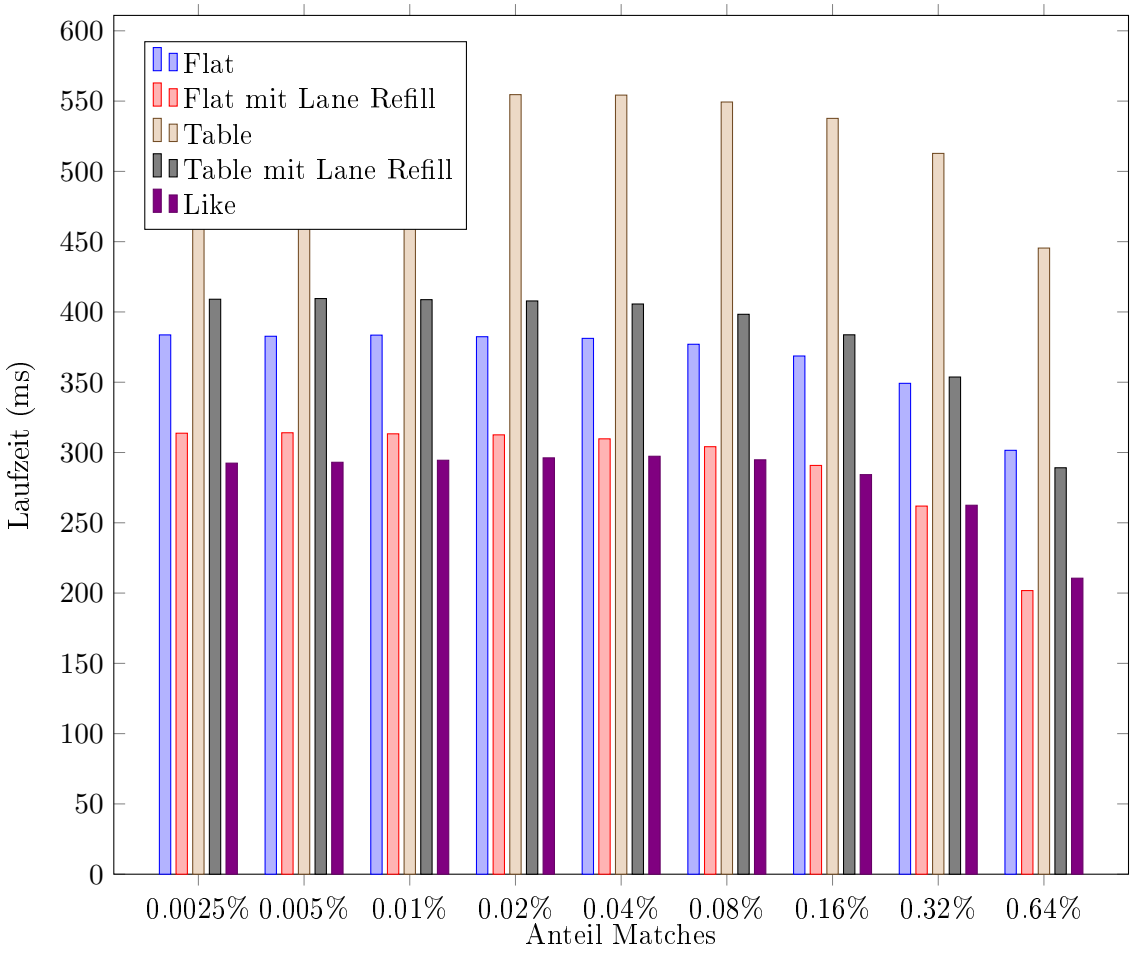

Abbildung 11.4: Laufzeiten für Benchmark, der beliebige Anfangszeichen zulässt

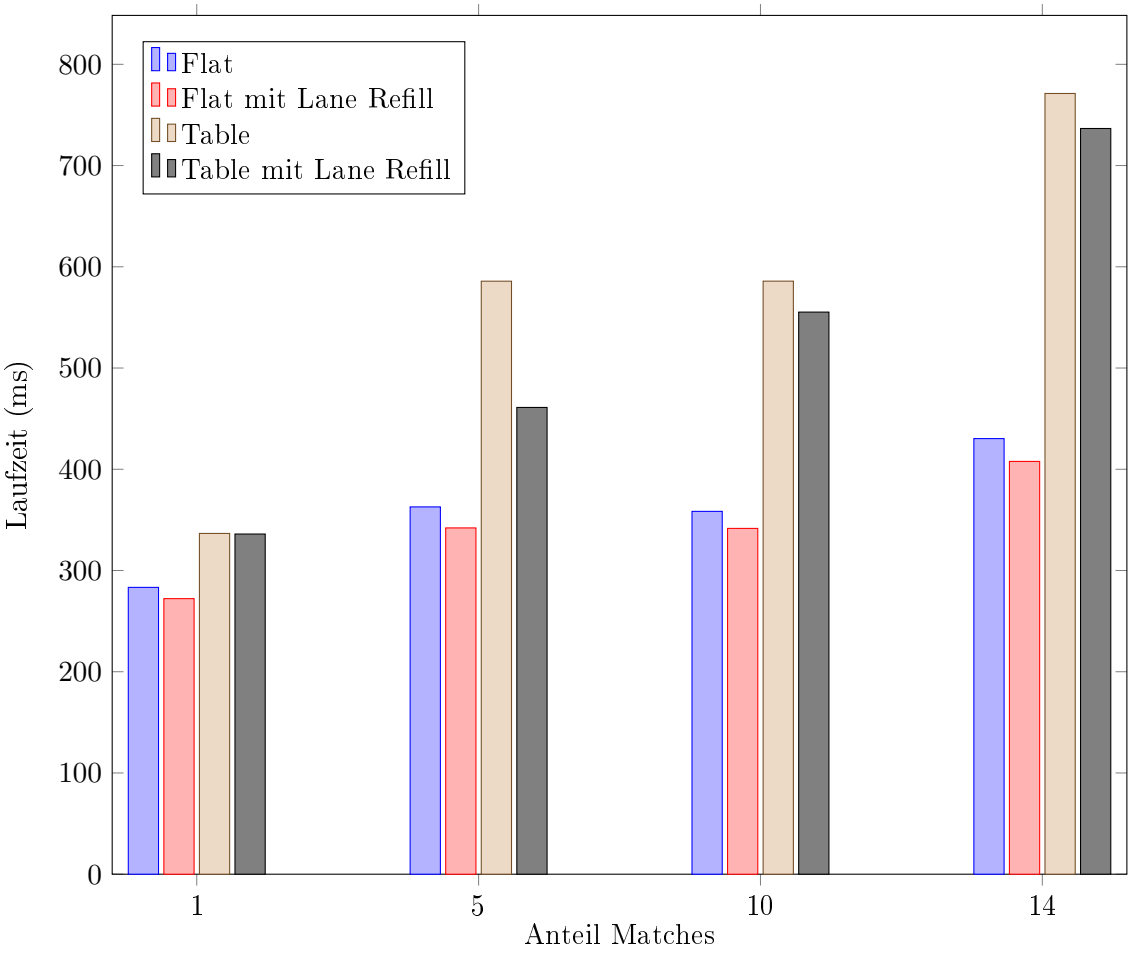

Abbildung 11.5: Laufzeiten für unterschiedliche Automatengrößen mit dem TPC-H Datensatz

Kapitel 12

Ergebnis und Fazit

Anhang A

Umsetzung der String-Selektion mit Lane Refill

```
1  // shared memory for the divergence buffers
2  __shared__ int search_id_divergence_buffer[THREAD_COUNT];
3  __shared__ int current_divergence_buffer[THREAD_COUNT];
4
5  unsigned warpid = (threadIdx.x / 32); // index of warp in block
6  unsigned bufferbase = (warpid * 32);    // buffer offset for warp in block
7  unsigned warplane = (threadIdx.x % 32); // index of lane in warp
8  unsigned prefixlanes = (0xffffffff >> (32 - warplane)); // previous lanes
9  int bufferelements = 0;         // number of elements in buffer
10
11 while(!flush_pipeline) {
12   current = loop_var;
13
14   /* execute previous operators in the pipeline */
15
16   data_length = char_offset[current+1] - char_offset[current] - 1;
17
18   // if string lengths are unequal, discard
19   if (active && data_length != search_length)
20     active = false;
21
22   int numactive = __popc(__ballot_sync(ALL_LANES, active));
23   while(bufferelements + numactive > THRESHOLD) {
24
25     // refill empty lanes from buffer in case of underutilization
26     if (numactive < THRESHOLD) {
27       numRefill = min(32 - numactive, bufferelements);
28       numRemaining = bufferelements - numRefill;
29
30       previous_inactive = __popc(~__ballot_sync(ALL_LANES, active) &
             prefixlanes);
```

```
31
32        if (!active && previous_inactive < bufferelements) {
33          buf_ix = numRemaining + previous_inactive + bufferbase;
34          search_id = search_id_divergence_buffer[buf_ix];
35          current = current_divergence_buffer[buf_ix];
36          active = true;
37        }
38
39        bufferelements -= numRefill;
40      }
41
42      int data_id = search_id + char_offset[current];
43
44      // when strings don't match, inactivate the lane
45      if (active && data_content[data_id] != search_string[search_id])
46        active = false;
47
48      search_id++;
49
50      if (search_id == search_length) {
51
52        /* execute following operators in the pipeline */
53
54        active = false;
55      }
56
57      numactive = __popc(__ballot_sync(ALL_LANES, active));
58    }
59
60    // flush active lanes to buffer
61    if (numactive > 0) {
62      previous_active = __popc(__ballot_sync(ALL_LANES, active) & prefixlanes
          );
63      buf_ix = bufferbase + bufferelements + previous_active;
64
65      if(active) {
66        search_id_divergence_buffer[buf_ix] = character_index;
67        current_divergence_buffer[buf_ix] = current;
68      }
69
70      bufferelements += numactive;
71      active = false;
72    }
73
74    loop_var += step;
75  }
```

Listing A.1: Umsetzung der String-Selektion mit Lane Refill

Anhang B

Laufzeiten für alternative Selektivität des Type-Datensatzes

Grid Size \ Block Size	32	64	96	128	160	192	224	256	384	512	640	768	896	1024
10	2101	1053	725	543	459	386	344	301	217	176	154	146	142	140
12	1750	877	605	454	384	321	288	253	185	153	145	142	139	139
14	1499	779	541	410	353	305	268	233	173	149	143	201	200	184
16	1312	683	474	359	309	268	236	206	155	144	139	186	172	166
18	1166	607	422	320	276	239	211	185	147	141	139	173	155	150
20	1051	547	387	303	258	220	199	177	146	140	195	155	151	147
30	714	384	278	213	186	163	150	145	139	180	142	157	151	148
40	547	303	225	174	154	146	143	139	169	158	165	159	152	152
60	384	213	164	145	144	139	194	179	139	145	152	141	139	139
80	305	175	147	140	187	168	153	155	149	151	146	144	141	143
100	257	153	141	192	162	144	142	164	150	152	142	148	145	144
150	186	141	190	146	167	148	144	158	139	147	144	145	141	140
200	253	190	149	166	142	152	143	154	143	144	144	142	139	140
300	189	148	152	156	154	139	143	148	139	140	140	141	141	139
400	205	166	158	153	142	142	143	143	142	140	141	142	140	140
500	173	143	139	148	139	144	144	142	143	139	139	141	139	141
1000	174	151	140	142	139	140	140	139	140	139	139	141	139	139
2000	160	142	139	140	139	140	140	139	139	142	139	141	139	139
3000	156	140	139	140	142	139	139	139	139	139	139	141	139	139
4000	156	139	139	139	139	142	139	139	139	139	140	141	139	139
6000	157	140	139	139	139	139	139	139	139	139	139	141	139	139
8000	155	139	139	139	139	139	142	139	139	139	139	143	139	139
10000	154	140	139	139	139	139	139	139	139	139	139	141	139	139
20000	153	139	139	139	139	139	139	141	139	139	139	142	143	139
50000	154	139	139	139	139	139	139	139	139	139	139	143	141	140
100000	154	139	139	139	139	139	139	139	141	139	140	144	143	144
150000	154	140	140	139	139	139	139	139	139	140	141	146	145	144
200000	154	139	139	140	139	139	139	139	140	143	143	148	147	147

Abbildung B.1: Laufzeit des String-Vergleichsalgorithmus in ms für den Type-Datensatz mit einer Selektivität von 64% unter Verwendung unterschiedlicher Grid-Konfigurationen

Abbildungsverzeichnis

Literatur

[1] Peter A. Boncz, Thomas Neumann und Orri Erling. "TPC-H Analyzed: Hidden Messages and Lessons Learned from an Influential Benchmark". In: *TPCTC*. 2013.

[2] Henning Funke u. a. "Pipelined Query Processing in Coprocessor Environments". In: *Proceedings of the 2018 International Conference on Management of Data*. SIGMOD '18. Houston, TX, USA: ACM, 2018. DOI: 10.1145/3183713.3183734. URL: http://doi.acm.org/10.1145/3183713.3183734.

[3] Mark Harris. *An Even Easier Introduction to CUDA*. 2017. URL: https://devblogs.nvidia.com/even-easier-introduction-cuda/.

[4] Mark Harris. *Using Shared Memory in CUDA C/C++*. 2013. URL: https://devblogs.nvidia.com/using-shared-memory-cuda-cc/.

[5] Harald Lang u. a. "Make the Most out of Your SIMD Investments: Counter Control Flow Divergence in Compiled Query Pipelines". In: *Proceedings of the 14th International Workshop on Data Management on New Hardware*. 2018. DOI: 10.1145/3211922.3211928. URL: http://doi.acm.org/10.1145/3211922.3211928.

[6] Yuan Lin und Vinod Grover. *Using CUDA Warp-Level Primitives*. 2018. URL: https://devblogs.nvidia.com/using-cuda-warp-level-primitives/.

[7] John Nickolls und David Kirk. "Graphics and Computing GPUs". In: *Computer Organization and Design: The Hardware/Software Interface*. 2009.

[8] *NVIDIA GeForce GTX 980. Featuring Maxwell, The Most Advanced GPU Ever Made*. Techn. Ber. NVIDIA Corporation, 2014.

[9] Adrian Thurston. *Ragel State Machine Compiler*. Techn. Ber. Colm Networks, 2009.

[10] Vasily Volkov. *Understanding Latency Hiding on GPUs*. Techn. Ber. University of California at Berkley, 2016.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 18. Mai 2019

Florian Lüdiger