

Bachelorarbeit

Modernes paralleles C++ für Hardwarebeschleuniger

Technische Hochschule Ulm

Fakultät Informatik

Studiengang Informatik

vorgelegt von

Erwin Kenner

April 2022

Erstgutachter: Prof. Dr. Georg Schied

Zweitgutachter: Prof. Dr. Phillipp Graf

Firmenbetreuer:

Christian Aßfalg,

Continental ADC Automotive Distance

Control Systems GmbH

Eigenständigkeitserklärung

Diese Abschlussarbeit wurde von mir selbständig verfasst. Es wurden nur die angegebenen Quellen und Hilfsmittel verwendet. Alle wörtlichen und sinngemäßen Zitate sind in dieser Arbeit als solche kenntlich gemacht.

Biberach an der Riß, den <Datum>, <Unterschrift>

Abstract

In Automobilen werden Hardwarebeschleuniger verwendet um Algorithmen zu beschleunigen. Die Verwendung solcher externen Hardwarebeschleuniger ist in Simulationen nicht wünschenswert. In Simulationen ist es das Ziel Algorithmen mit GPUs zu beschleunigen. Mit Hilfe der in C++17 eingeführten Parallel STL Algorithmen ist es möglich die Parallelisierung dem Compiler zu überlassen. Das Nvidia Toolkit unterstütz mit seinem eigenen C++ Compiler die Parallel STL Algorithmen und macht ihre Ausführung auf GPUs möglich. In dieser Arbeit werden anhand eines Beispiels aus der Automobilbranche die Möglichkeiten der Parallel STL Algorithmen in Verbindung mit dem Nvidia Toolkit erfasst. Die erhaltenen Ergebnisse zeigen Verschlechterungen der Laufzeiten, die durch Optimierung reduziert werden können. Dennoch ist die parallelisierte Anwendung langsamer. Es zeigt sich, dass die Möglichkeiten der Parallel STL in Verbindung mit dem Nvidia Toolkit begrenzt sind.

Inhaltsverzeichnis

A. Einführung 4

I. Motivation 4

II. Zielsetzung 6

III. Kapitelübersicht 7

B. Grundlagen 8

I. Paralleles Programmieren 8

1. Taskparallelität und Datenparallelität 8

2. Parallel STL 9

II. GPU und die Nvidia HPC SDK 16

1. Graphics Processing Unit 16

2. CUDA 17

3. Nvidia High Performance Computing Software Development Kit 21

III. (Algorithmus) 22

C. Lösungsansatz / Versuchsaufbau 23

I. Anforderungsanalyse / IST Zustand / 23

II. Lösungsansätze / Probleme 29

1. Beschreibung möglicher Lösungsansätze 29

2. Evaluierung der Lösungsansätze 29

III. Ausgewählter Ansatz und Durchführung 30

1. Detailanalyse der Polynomkoeffizienten Berechnung 30

2. Übersetzung in Per-Pixel-Berechnung 35

3. Parallelisierung der Per-Pixel-Berechnung. 39

D. Ergebnisse 44

I. Validierung der Ergebnisse und Korrektheit der Implementierung 44

II. Übersicht der Ergebnisse 45

III. Ergebnisdiskussion und Erkenntnisse 47

1. Bewertung der Ergebnisse 47

2. Gesammelte Erfahrungen und Erkenntnisse 47

E. Zusammenfassung 49

F. Literaturverzeichnis 50

G. Anlagen 52

# Einführung

## Motivation

Software für Anwendungen in der Automobilbranche muss eine Reihe von Test und Kriterien erfüllen um für den Straßenverkehr zugelassen zu werden. Eine dieser Kriterien ist, dass die Software in einem Testfahrzeug mehrere Hunderttausend Kilometer gefahren sein muss. Fahrten dieser Testfahrzeuge können auch simuliert werden. Die in der Simulation zurückgelegten Kilometer werden ebenfalls gewertet. In Fahrzeugen werden Hardwarebeschleuniger verwendet um die Performance von Rechen- oder Speicherintensiven Algorithmen zu verbessern. Diese Hardwarebeschleuniger sind speziell für die Art ihrer Aufgabe konstruiert. In Simulationen ist diese Verwendung von mehreren externen Hardwarebeschleunigern nicht wünschenswert, da die für die Simulation benötigte Hardware nicht portabel ist. Die Hardware ist eine spezifische Lösung für ein Problem und ist nutzlos nachdem das Projekt beendet wurde. Um Hardware effizienter nutzen zu können braucht es eine generalisierte Hardware, welche nicht auf eine Art von Problem spezifiziert ist. Generalisierte Hardware ist in den meisten Rechnern verbaut. Die Graphics Processing Unit (GPU) ist eine solche generalisierte Hardware. Wenn Simulationen auf Rechenknoten mit eingebauten GPUs ausgeführt werden, können diese zur Beschleunigung der Algorithmen verwendet werden. Da die Simulation auf generalisierter Hardware läuft kann sie nachdem das Projekt beendet wurde für neue Projekte weiterverwendet werden.

Dennoch können die Algorithmen nicht ohne weiteres einfach so auf GPUs ausgeführt werden. Der vorher auf den speziellen Hardwarebeschleunigern ausgeführte Code muss für die Ausführung auf GPUs angepasst werden. Dies kann je nach Art des Algorithmus sehr aufwendig sein. Der C++17 Standard stellte Parallelisierungsmöglichkeiten vor, welche es dem Nutzer ermöglichen Parallelisierung der Standard Library Algorithmen anzufordern. Nvidia, einer der größten Entwickler von Grafikprozessoren und Chipsätzen, unterstützt diese „Parallel Algorithms“. Mit den C++17 Algorithmen und den von Nvidia entwickelten Compilern soll die Kompilierung für GPUs vereinfacht sein.

## Zielsetzung

Das Ziel dieser Arbeit ist es Erfahrungswerte im Umgang mit den „Parallel Algorithms“ zu sammeln. Um diese Erfahrungswerte sammeln zu können soll mit Hilfe der C++17 „Parallel Algorithms“ ein im Bereich Automotive verwendeter Algorithmus parallelisiert werden. Der zu parallelisierende Algorithmus ist ein „dense optical flow“ Algorithmus und wird zur Berechnung des Optischen Flusses einer Videosequenz verwendet. Der parallelisierte Algorithmus soll dann auf einer GPU ausführbar gemacht werden. Hierfür wird das Nvidia Toolkit verwendet, welches nötige C++ Compiler für die Generierung von GPU-Code zur Verfügung stellt. Nachdem der Algorithmus auf einer GPU ausführbar gemacht wurde, sollen Performanceeinflüsse ermittelt werden. Während der Parallelisierung des Algorithmus und der Portierung auf eine GPU sollen mögliche Limitierungen in der Nutzung von C++17 „Parallel Algorithms“ oder des Nvidia Toolkits erfasst und bewertet werden.

## Kapitelübersicht

Diese Arbeit ist in mehrere Kapitel aufgeteilt. Es folgt eine Übersicht über die Inhalte und Anordnung dieser Kapitel:

1. **Grundlagen**

In diesem Kapitel werden Themen und Begriffe eingeführt, auf denen das Lösungskonzept aufbaut. Hierbei werden Themen zu Parallelem Programmieren, GPUs, Nvidia HPC SDK und zum verwendeten Algorithmus behandelt.

1. **Lösungsansatz**

Dieses Kapitel beinhaltet die Analyse der Gegebenheiten vor der Parallelisierung des Algorithmus. Außerdem werden zwei Lösungsansätze vorgestellt um den Algorithmus zu Parallelisieren. Die Ansätze werden bewertet und es wird Ansatz ausgewählt. Dieser Ansatz wird im Detail verfolgt, aufkommende Probleme beschrieben und eine endgültige Lösung des Ansatzes präsentiert.

1. **Ergebnisse**

In diesem Kapitel wird die erarbeitete Lösung validiert und die aus der Lösung erhaltenen Ergebnisse dargestellt und bewertet.

1. **Zusammenfassung**

Es werden erarbeitete Erkenntnisse, aufgetretene Probleme und erreichte Ergebnisse zusammengefasst. Außerdem wird ein Ausblick auf mögliche weitere Schritte gegeben.

# Grundlagen

In den Grundlagen werden Themen und Begriffe eingeführt, welche für das Lösungskonzept relevant sind. Themen sind hierbei paralleles Programmieren, GPUs, die Nvidia HPC SDK und der verwendete Algorithmus.

## Paralleles Programmieren

Beim parallelen Programmieren gibt zwei Arten von Parallelität. Die Taskparallelität und die Datenparallelität.

### Taskparallelität und Datenparallelität

Parallelität kann erreicht werden indem Ausführungsschritte eines Programms voneinander gekapselt und dann gleichzeitig ausgeführt werden. Hierbei wird jedem Ausführungsschritt ein Thread oder manchmal auch ein ganzer Prozessorkern zugewiesen. Diese Art der Parallelität wird Taskparallelität genannt. Um Taskparallelität zu erreichen muss darauf geachtet werden, keine Racing Conditions oder Deadlocks einzuführen. Außerdem sollte ein Prozessor möglichst effizient ausgelastet werden um maximale Performance zu erreichen. Um dies möglich zu machen werden Ausführungsschritte oft in kleinere Unterschritte aufgeteilt um sie besser auf die Anzahl von Threads und Prozessorkernen zu verteilen. Je nach Natur des Programms kann sich dies aber durchaus als schwierig oder unmöglich herausstellen. (Boyd, 2008)

Die zweite Art der Parallelität wird Datenparallelität genannt. Hierbei werden Threads Datenelemente zugewiesen. Anstatt die Ausführungsschritte des Programms aufzuteilen, wird auf die Menge der zu bearbeitenden Daten geachtet. So werden die einzelnen Datenelemente, welche meist unabhängig voneinander sind, auf verfügbare Threads auf den Prozessorkernen verteilt. [1]

Datenparallelität spielt vor allem eine Rolle in Programmen welche sich mit Bild- und Videoverarbeitung befassen. In diesen Bereichen werden oft tausende von Datenelementen verarbeitet, welche unabhängig voneinander erfasst und bearbeitet werden können.

### Parallel STL

#### Execution Policies

Mit C++17 wurden mehrere Algorithmen der Standard Library mit sogenannten „execution policies“ erweitert. Diese ermöglichen es Standard Algorithmen zu parallelisieren. Viele Algorithmen besitzen nun Überladungen, welche Objekte dieser „execution policies“ akzeptieren.

Zu den „execution policies“ gehören:

* sequenced\_policy
* parallel\_policy
* parallel\_unsequenced\_policy

[2]

Wird das Objekt „seq“ der sequenced\_policy Klasse als Parameter übergeben, wird die Funktion in richtiger Reihenfolge (sequenziell) ausgeführt. Dennoch darf der Compiler den Algorithmus parallelisieren, wenn dies für das Programm nicht sichtbar ist.[3]

Beim übergeben des „par“ Objekts der parallel\_policy Klasse wird dem Algorithmus signalisiert, dass dieser die Ausführung auf mehrere Threads aufteilen darf. Dennoch werden die Operationen innerhalb eines Threads sequenziell abgearbeitet.[3]

Wird das „par\_unseq“ Objekt der Klasse parallel\_unseq\_policy übergeben, wird die Ausführung auf mehrere Threads aufgeteilt und vektorisiert.[3]

#### Standard Algorithmen

Die C++ Standard Template Library (STL) ist eine Menge von Template Klassen, welche häufig verwendete Programmier- und Datenstrukturen und Funktionen zur Verfügung stellt. Es ist eine Bibliothek von Klassen, Algorithmen und Iteratoren. [4]

Die Algorithmen der C++ STL sind Funktionen, die von der Algorithms library definiert werden. Diese Funktionen agieren dann auf einem Bereich von Elementen. Beispiele hierfür sind das Sortieren, Suchen, Zählen oder Modifizieren von Elementen einer Liste oder eines Arrays.[4]

Die Parallel STL ist ein kleiner Teil dieser Algorithms Library. Dieser mit C++17 dazugekommene Teil beinhaltet unteranderem die Algorithmen „std::for\_each“, „std::for\_each\_n“, „std::transform“ und „std::transform\_reduce“, welche in dieser Arbeit behandelt werden.

Um die Verwendung dieser Algorithmen darzustellen werden Anwendungsbeispiele gezeigt. Diese beinhalten Lambdaausdrücke, auch Lambdafunktionen genannt. Dies sind anonyme Funktionen welche nicht an einen Bezeichner gebunden sind. Sie eignen sich sehr gut als Funktionen für die Parallel STL Algorithmen.

1. std::for\_each, std::for:\_each\_n

Der Algorithmus „std::for\_each“ wird verwendet um eine Funktion auf alle Elemente eines Bereichs anzuwenden. Er besitzt mehrere Interfaces und das für diese Arbeit wichtige Interface ist in Codeausschnitt B.1 abgebildet.



Codeausschnitt B.1, eines der Interfaces von „std::foreach mit „execution policy“.

„std::for\_each“ akzeptiert als ersten Parameter ein Objekt der Klasse „ExecutionPolicy“ und kann somit parallelisiert werden. Weitere akzeptierte Parameter sind Objekte der Klasse „ForwardIt“. Hierbei handelt es sich um Iteratoren die den Bereich der von „std::for\_each“ behandelt wird bestimmen, indem ein Iterator für das erste und das letzte Element des Bereichs angegeben werden. Als letzten Parameter muss eine Funktion angegeben werden, die für jedes Element des Bereichs ausgeführt wird. [5]

Neben dem „std::for\_each“ Algorithmus, gibt es noch dem „std::for\_each“ sehr ähnlichen Algorithmus „std::for\_each\_n“. Das für diese Arbeit wichtige Interface wird von Codeausschnitt B.2 dargestellt.



Codeausschnitt B.2, eines der Interfaces von „std::for\_each\_n“ mit „execution policy“.

Dieses Interface unterscheidet sich im Wesentlichen nicht besonders vom Interface von „std::for\_each“. Anstatt des Iterators für das letzte Element, wird ein Wert n für die Anzahl der Elemente des zu bearbeitenden Bereichs angegeben. [6]

Als Anwendungsbeispiel, welches die Nutzung von std::for\_each verdeutlicht, wird eine Stencil Operation mit std::for\_each“ implementiert. Stencil Operationen brauchen meist Informationen zu benachbarten Werten in einem zweidimensionalen Array um Berechnungen durchzuführen oder Arraywerte miteinander zu vergleichen. Werden Stencil Operationen mit einem Algorithmus der Parallel STL umgesetzt, kann folgendes Hindernis entstehen. Die verwendete Lambdafunktion besitzt nur eine Referenz auf das benutzte Element, aber hat keinerlei Informationen über die Position des Elements innerhalb des Bereichs. Somit können auch keine Informationen über benachbarte Elemente erhalten werden. Um dieses Hindernis zu umgehen muss mit Hilfe von Pointer Arithmetik der Index des Elements berechnet werden. Dieser kann dann verwendet werden um benachbarte Elemente zu adressieren.

Die Verwendung einer Stencil Operation und die Umgehung eines solchen Hindernisses wird mit Hilfe von Codeausschnitt B.9 erläutert.



Codeausschnitt B.9, Beispielcode einer Stencil Operation mit Parallel STL Algorithmus.

Bei diesem Beispiel handelt es sich um eine vereinfachte Implementierung von Conway's "Game of Life".

Hier werden für jede Zelle in einem Bereich entschieden, ob diese weiterhin existieren werden. Eine Zelle kann nur dann existieren, wenn sie mindestens 3 Zellen als direkte Nachtbaren hat. Zusätzlich werden Felder zu Zellen, wenn das Feld von 3 oder mehr Zellen umgeben ist. Um über den N x N großen Bereich „field“ zu iterieren wird „std::for\_each“ verwendet.

In Zeile 3 des Beispiels findet die Berechnung des Indexes statt. Es wird die Adresse des Bereichs, welches gleichzeitig die Adresse des ersten Elements ist, von der Adresse des gerade verwendeten Elements subtrahiert. Als Ergebnis erhält man den Abstand des Elements zum ersten Element und somit den Index des Elements innerhalb des Bereichs.

Die eigentliche Stencil Operation findet in den Zeilen 12 – 15 statt. Hier werden die Werte benachbarter Zellen mit Hilfe des zuvor berechneten Index summiert.

1. std::transform

Der Algorithmus „std::transform“ hat zwei Varianten. Die Erste appliziert eine einstellige Verknüpfung „unary\_op“ auf einen Bereich festgelegt durch die Iteratoren „first1“ und „last1“. Das Ergebnis wir dann in einem Bereich gespeichert, der vom Iterator „d\_first“ bestimmt wird. Grundsätzlich ähneln die akzeptierten Parameter von „std::transform“ denen von „std::for\_each“. Im Gegensatz zu „std::for\_each“ garantiert „std::transform“ die Reihenfolge der Applizierung des Operators nicht. Das wichtigste Interface von „std::transform“ ist in Codeausschnitt B.3 abgebildet.



Codeausschnitt B.3, ein Interface von „std::transform“ mit „execution policy“.

Die zweite Variante ähnelt der erste, aber ihr Interface ermöglicht es zwei Eingangsbereiche anzugeben. Dies wird möglich gemacht indem das Interface einen weiteren Parameter „first2“ akzeptiert. Dieser Iterator bestimmt den Anfang des zweiten Eingangsbereichs. Hierbei wird die Größe des Bereichs vom ersten Eingangsbereich abgeleitet. Der zweite Aspekt ist der applizierte Operator. In diesem Interface wird eine binäre Verknüpfung angewendet, welche die Elemente der beiden Eingangsbereiche verknüpft. Dieses Interface ist in Codeauschnitt B.4 zu sehen.



Codeausschnitt B.4, ein Interface von „std::transform“ mit „execution policy“ und einem weiteren Iterator.

Bei „std::transform“ ist zu beachten, dass die Reihenfolge der Operationen nicht garantiert ist. Möchte man die Reihenfolge der Operationen garantiert haben, sollte „std::for\_each“ stattdessen verwendete werden. [7]

Um die Verwendung von „std::transform“ zu verdeutlichen wird in einem Beispiel die Helligkeit eines Graustufenbildes mit Hilfe von std::transform geändert. Soll die Helligkeit eines Bildes geändert werden, so muss jeder Pixelwert mit einem Wert skaliert werden. Ist der Wert größer als 1, wird die Helligkeit des Bildes erhöht. Ist der Wert kleiner als 1, wird die Helligkeit gesenkt. Für diese Art der Bildverarbeitung bietet sich der „std::transform“ Algorithmus an. Ein Beispiel hierfür ist in Codeausschnitt B.7 zu finden.



Codeausschnitt B.7, Beispielcode zur Änderung der Helligkeit in einem Graustufenbild „image“.

Das Graustufenbild wird im eindimensionalen Vector „image“ gespeichert. Das Ergebnis wird in einem ebenfalls eindimensionalen Vector „image2“ gespeichert. Durch den Lambda-Ausdruck in Zeile 5 wird jeder Wert „a“ aus dem Vector „image“ mit dem Faktor 3 multipliziert und an passender Stelle in den Vector „image2“ geschrieben.

1. std::transform\_reduce

„std::transform\_reduce“ fungiert im Prinzip genauso wie „std::transform“, aber zusätzlich werden die Ergebnisse der „transform“ Operation noch aufsummiert. Für der Wert der Summe wird ein Initialwert „init“ als Parameter akzeptiert, auf welchen dann die Ergebnisse addiert werden. Es kann jedoch nicht nur eine Summe produziert werden, sondern zum Beispiel auch ein Produkt. Dies macht der Parameter „reduce“ der Klasse „BinaryReductionOp“ möglich. In den Codeausschnitten – sind die zwei wichtigsten Interfaces von „std::transform\_reduce“ dargestellt.



Codeausschnitt B.5, eines der Interfaces von „std::transform\_reduce“ mit „execution policy“.

Wie auch zuvor bei „std::transform“ akzeptiert „std::transform\_reduce“ auch zwei Eingansbereiche. [8]



Codeausschnitt B.6, ein Interfaces von „std::transform\_reduce“ mit „execution policy“ und zweitem Eingangsbereich.

Die Nutzung von „std::transform\_reduce“ wird veranschaulicht, durch die Berechnung des Skalarprodukts zweier Vektoren mit „std::transform\_reduce“. Bei der parallelen Berechnung des Skalarprodukts zweier Vektoren können Racing Conditions eingeführt werden, da für die Summierung der Produkte jeweils der aktuelle Wert der Summe gelesen werden muss bevor die Summe aktualisiert werden kann. Um Racing Conditions zu vermeiden wird deshalb „std::transform\_reduce“ verwendet. Als Beispiel wird in Codeausschnitt B.8 das Skalarprodukt der beiden Vektoren „vec1“ und „vec2“ gebildet und in der Variablen „sum“ gespeichert.



Codeausschnitt B.8, „std::transform\_reduce“ Beispiel zur Berechnung des Skalarprodukts zweier Vektoren.

## GPU und die Nvidia HPC SDK

### Graphics Processing Unit

##### Hardwarearchitektur

Eine CPU ist so entworfen, dass sie eine Sequenz von Operationen so schnell wie möglich ausführen kann. Dahingegen ist die GPU darauf ausgelegt tausende von Operationen gleichzeitig auszuführen. Deshalb unterscheidet sich die Hardwarearchitektur der GPU von der einer CPU.[9]

Die Grundarchitektur einer GPU unterscheidet von der einer CPU in mehreren Aspekten. Während eine CPU aus einer Handvoll von komplexen Prozessorkernen besteht, ist eine GPU mit hunderten einfacheren Prozessorkernen ausgestattet. Diese Kerne besitzen jeweils tausende parallel laufende Hardware Threads.[10] Abbildung B.1 veranschaulicht den Aufbauunterschied.

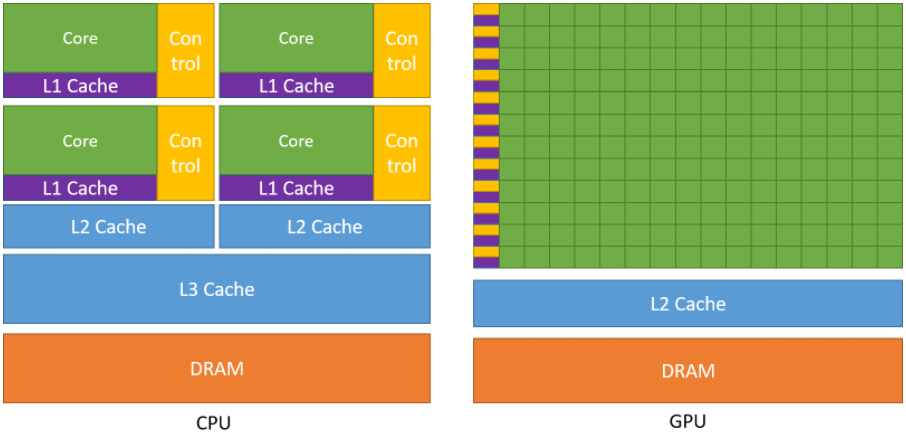


Abbildung B.1, die GPU dediziert mehr Kerne für die Datenverarbeitung als eine CPU. [9]

### CUDA

CUDA ist eine Programmierschnittstelle entwickelt von Nvidia. Diese ermöglicht es dem Programmierer Anwendungssoftware zu schreiben, die ihre Parallelität skaliert um die steigende Anzahl an Prozessorkernen vollends auszunutzen. CUDA stellt drei Abstraktionen bereit, welche es ermöglichen ein Problem in Teilprobleme aufzuteilen. Abstrahiert wird unteranderem die zu Verfügung stehenden Threads. Threads werden zu gleichgroßen Blöcken zusammengefasst. Teilprobleme können dann unabhängig voneinander parallel gelöst werden, indem jedem Teilproblem ein Block mit Threads zugeteilt wird. Die Anzahl von Threads in einem Block ist limitiert durch die Anzahl der Threads die sich auf einem Prozessorkern befinden. Auf aktuellen GPUs kann ein Threadblock bis zu 1024 Threads beinhalten. Threadblöcke sind Teile eines sogenannten Gitters, welches die Threadblöcke organisiert.[9] Abbildung B.2 zeigt die logische Unterteilung von Threads in Blöcke und Gitter.

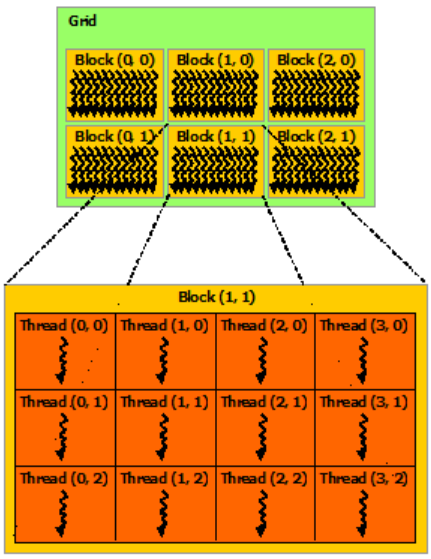


Abbildung B.2, ein Gitter (engl. Grid) ist unterteilt in Blöcke (engl. Blocks) von Threads. [9]

CUDA erlaubt es dem Programmierer bestimmte Funktionen zu definieren, welche dann parallel von CUDA Threads ausgeführt werden. Diese Funktionen werden Kernel genannt. Einem Kernel wird definiert wie viele CUDA Threads diesen ausführen. Hierbei ist die Anzahl der Threads nicht limitiert, da ein Kernel von mehreren Threadblöcken ausgeführt werden kann. Bei der Definition eines Kernels wird die Anzahl der Threads pro Block und die Anzahl von Blöcken pro Gitter angegeben.[9]

Nvidias GPU Architektur ist um eine Anordnung von Streaming Multiprozessoren (SMs) konzipiert. Jeder Streaming Multiprozessor kann einen Threadblock nach dem anderen ausführen. Die Anwendungssoftware wird auf die entsprechende GPU Hardwarearchitektur skaliert, indem die auszuführenden Threadblöcke auf die Anzahl der verfügbaren Streaming Multiprozessoren gleichmäßig verteilt werden. So kann eine GPU mit vier Streaming Multiprozessoren vier Threadblöcke gleichzeitig ausführen, während eine GPU mit zwei Streaming Multiprozessoren in der gleichen Zeitspanne nur zwei Threadblöcke ausführen kann. [9]

Dies wird durch Abbildung B.3 veranschaulicht.

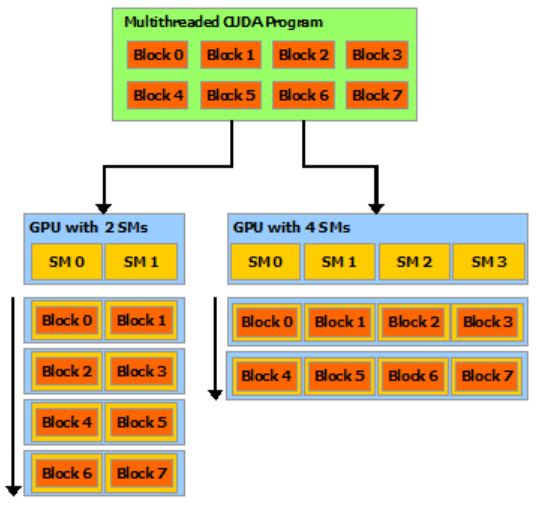


Abbildung B.3, automatische Skalierung durch Aufteilung der Blöcke auf die Verschiedenen Streaming Multiprozessoren. [9]

#### CUDA Unified Memory

Unified Memory bezeichnet einen Speicheradressbereich, auf welchen jeder Prozessor in einem System zugreifen kann. Abbildung B.4 zeigt die funktion dieses Adressbereichs.

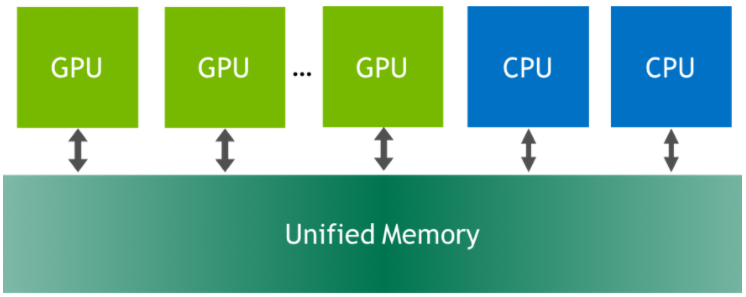


Abbildung B.4, Unified Memory ist von allen Prozessoren abrufbar. [11]

Diese Technologie ermöglicht es Speicher für Daten zu allokieren, welche von sowohl CPU als auch GPU gelesen und geschrieben werden können. Um Unified Memory anzulegen werden aufrufe zu malloc() oder new mit aufrufen zu cudaMallocManaged() ersetzt. cudaMallocManaged() ist eine Funktion zum allokieren von Speicher, welche einen Pointer zugänglich für jeden Prozessor zurückgibt. Greift ein Programm auf Daten zu, die auf diese Art angelegt wurden, kümmert sich die CUDA Systemsoftware und die Hardware um das Verschieben von Speicher Pages auf den Speicher des zugreifenden Prozessors.[12]

Greift die GPU auf Speicher zu der von cudaMallocManaged angelegt wurde werden folgende Schritte ausgeführt. Zuerst werden neue Pages auf der GPU angelegt. Dann werden die Verlinkungen der alten Pages auf der CPU gelöst. Daten werden dann von der CPU zur GPU kopiert und die neuen Pages auf der GPU werden gelinkt. Zuletzt werden die alten CPU-Pages freigegeben.

Wenn neuere GPUs auf eine Page zugreifen, welche sich nicht auf dem lokalen Speicher der GPU befindet, wird beim Übersetzen der Page eine Fault Message generiert. Die GPU kann mehrere solcher Fault Messages gleichzeitig generieren. Der Unified Memory Treiber verarbeitet die faults, beseitigt Duplikate, bringt die Verlinkungen auf den neusten Stand und transferiert die Daten. Wichtig zu erwähnen ist, dass Unified Memory eine Entwicklung von NVIDIA ist und deshalb auf nur auf NVIDIA GPUs verfügbar ist.[11]

### Nvidia High Performance Computing Software Development Kit

Erklären wie nvc++ c++ in cuda kernel übersetzt

Das Nvidia HPC Software Development Kit (SDK) beinhaltet sowohl C, C++ und Fortan Compiler als auch Bibliotheken und Tools für die Verbesserung von Leistung und Portabilität von High Performance Computing (HPC) Anwendungen. [13] Die Nvidia HPC SDK steht zum Zeitpunkt dieser Arbeit nur für Linux-Systeme zur Verfügung.

Der NVIDIA HPC C++ Compiler NVC++ unterstützt den C++17 Standard und wird genutzt um Algorithmen mit „execution policies“ wie „std::execution::par“ oder „std::execution::par\_unseq“ für die Ausführung auf NVIDIA GPUs zu kompilieren. Für die Aktivierung der GPU-Beschleunigung werden keinerlei zusätzliche Spracherweiterungen oder Bibliotheken benötigt.[14]

GPU-Beschleunigung kann aktiviert werden, indem der Kommandozeilenbefehl „-stdpar“ an den Compiler übergeben wird. Die Datenübertragung zwischen dem Host Speicher und dem GPU Speicher geschieht automatisch unter der Verwendung von CUDA Unified Memory. Durch die Verwendung des NVC++ Compilers wird die GPU-Beschleunigung von Parallelem C++ Code vereinfacht. Dennoch gibt es Einschränkung für die Verwendungen von NVC++ zur GPU-Beschleunigung, welche nun genannt werden.[14]

Der NVC++ erkennt automatisch, welche Funktionen für die Ausführung auf einer GPU kompiliert werden müssen. Dies ist aber nur möglich, wenn die Funktionsdefinitionen im selben Source File liegen wie der Aufrufe der Funktionen.[14] Funktionen die auf einer GPU kompiliert werden müssen sind Funktionen welche von dem Parallel STL Algorithmus aufgerufen werden. Da der Algorithmus auf der GPU ausgeführt werden soll, muss für jede Funktion die er aufruft GPU-Code existieren.

NVC++ vertraut auf CUDA Unified Memory um Daten automatisch zwischen CPU und GPU zu übertragen. Dies ist aktuell nur möglich, wenn diese Daten dynamisch auf dem Heap Speicher der CPU allokiert sind und vom NVC++ Compiler kompiliert wurden. Das bedeutet, dass jeder Pointer der dereferenziert und jedes C++ Objekt das referenziert wird muss auf Daten auf dem CPU-Heap referieren. Außerdem müssen diese Daten von einer Anwendung, welche vom NVC++ Compiler kompiliert wurde, allokiert worden sein.[14]

Ebenso wie für verwendete Daten, gibt es Limitierungen bei der Verwendung von Funktionen im Parallel STL Aufruf. So können keine Funktionspointer an einen Parallel STL Aufruf als Parameter übergeben werden. Nur Funktionsobjekte oder Lambdas können übergeben werden. Die letzte wichtige Einschränkung ist, dass Exceptions in Code welcher auf der GPU ausgeführt wird nicht verwendet werden sollten. Der Exceptioncode kompiliert zwar, aber hat nicht das erwarte Verhalten. So werden Catch-Bereiche ignoriert und das werfen einer Exception führt dazu, dass der GPU-Kernel abgebrochen wird. [14]

## **(Algorithmus**)

* > Mehr, Genauer

OpenCV ist eine open-source Library für Computer Vision. Für diese Arbeit wichtig ist das Modul zur Berechnung des Optischen Flusses. Der von OpenCV implementierte Algorithmus zur Berechnung des Optischen Flusses basiert auf dem Paper „Two-Frame Motion Estimation Based on Polynomial Expansion“ von Gunnar Farnebäck.

In diesem Paper stell Farnebäck einen Algorithmus zur Bewegungsabschätzung vor, welcher auf Distributivgesetzen basiert. So wird im Algorithmus zuerst jeweils eine Umgebung (Neighbourhood) für zwei Frames mit quadratischen Polynomen angenähert. Dann wird die Verschiebung mit Hilfe der polynomialen Erweiterungskoeffizienten abgeschätzt. [15]

# Lösungsansatz / Versuchsaufbau

## Anforderungsanalyse / IST Zustand /

Um entscheiden zu können welchen Teil der OpenCV Implementierung am meisten von einer Parallelisierung profitiert, ist es notwendig herauszufinden welche Teile der Implementierung am meisten Zeit für ihre Ausführung benötigen. Diese Ausführungszeit kann durch Komplexität oder Anzahl der Rechenoperationen und die Menge an Speicherzugriffen beeinflusst werden. Danach muss analysiert werden ob Parallelisierung dieses Teils möglich ist. Vor allem sollte auf mögliche Datenparallelität geachtet werden.

Die Implementierung von OpenCV teilt die Berechnung des Optischen Flusses in mehrere Unterfunktionen auf. Messungen sollen zeigen, wie viel Zeit die einzelnen Unterfunktionen jeweils in Anspruch nehmen. Für die Zeitenmessung wird eine Videosequenz, welche in 300 einzelne Framebilder aufgeteilt ist, verwendet. Zeiten werden mithilfe der „chrono::steady\_clock“ Klasse aus der Standard Library von C++ genommen. Diese ermöglicht es Zeitpunkte zu speichern und daraus dann Zeitintervalle zu berechnen.

Um das Input Video einzulesen und den daraus resultierenden Optischen Fluss darzustellen wird das Programm „densFlow“ aus Anhang A verwendet. „denseFlow“ basiert auf dem Beispielcode zur Berechnung von Optischem Fluss der OpenCV Dokumentation. Die Ausführungszeiten folgender Funktionen werden gemessen:

* calcOpticalFlowFarneback()
* FarnebackPolyExp()
* FarnebackUpdateMatricies()
* FarnebackUpdateFlow\_Blur()

Ausgewählt wurden diese Funktionen aus den folgenden Gründen. calcOpticalFlowFarneback wird in „denseFlow“ verwendet um den Optischen Fluss zu berechnen. Der restliche Code aus „denseFlow“ beschäftigt sich mit dem Einlesen und Anzeigen von Videoframes bzw. Flussbildern. FarnebackPolyExp, FarnebackUpdateMatricies und FarnebackUpdateFlow\_Blur werden jeweils von calcOpticalFlowFarneback aufgerufen und werden zur Flussberechnung verwendet.

Es ist zu erwähnen, dass zuvor genannte Funktionen mehr als einmal pro Aufruf von calcOpticalFlowFarneback aufgerufen werden. Die Anzahl der Aufrufe ist abhängig von den gewählten Parametern für calcOpticalFlowFarneback. Unter Verwendung des oben genannten Beispielcodes und der Videosequenz werden die Durchschnittlichen Ausführungszeiten für die Berechnung der Flussbilder ermittelt. Diese Zeiten werden wie folgt berechnet:

a : Anzahl der Funktionsaufrufe pro Aufruf von calcOpticalFlowFarneback.

n : Anzahl der Aufrufe von calcOpticalFlowFarneback Aufrufen.

t : Ausführungszeit eines Funktionsaufrufs.

ist das arithmetische Mittel der Ausführungszeit eines Funktionsaufrufs.

Sind die Ausführungszeiten erfasst, sollte der Code der ausgewählten Funktion genauer betrachtet werden. Besonders sollte auf Schleifen geachtet werden, welche in der Regel einfach zu parallelisieren sind. Schleifen welche über die Dimensionen des Input Videoframes iterieren sind Anzeichen für mögliche Datenparallelität.

Messungen werden auf einer Intel(R) Core(TM) i7-9850H CPU abgenommen. Jeder Input Videoframe besitzt eine Dimension von 768x576 Pixeln. Alle Messergebnisse stammen aus einer einzigen Ausführung von „denseFlow“. Die Messergebnisse der durchschnittlichen Ausführungszeiten werden durch das Balkendiagramm in Abbildung C.1 und der Tabelle C.1 dargestellt.

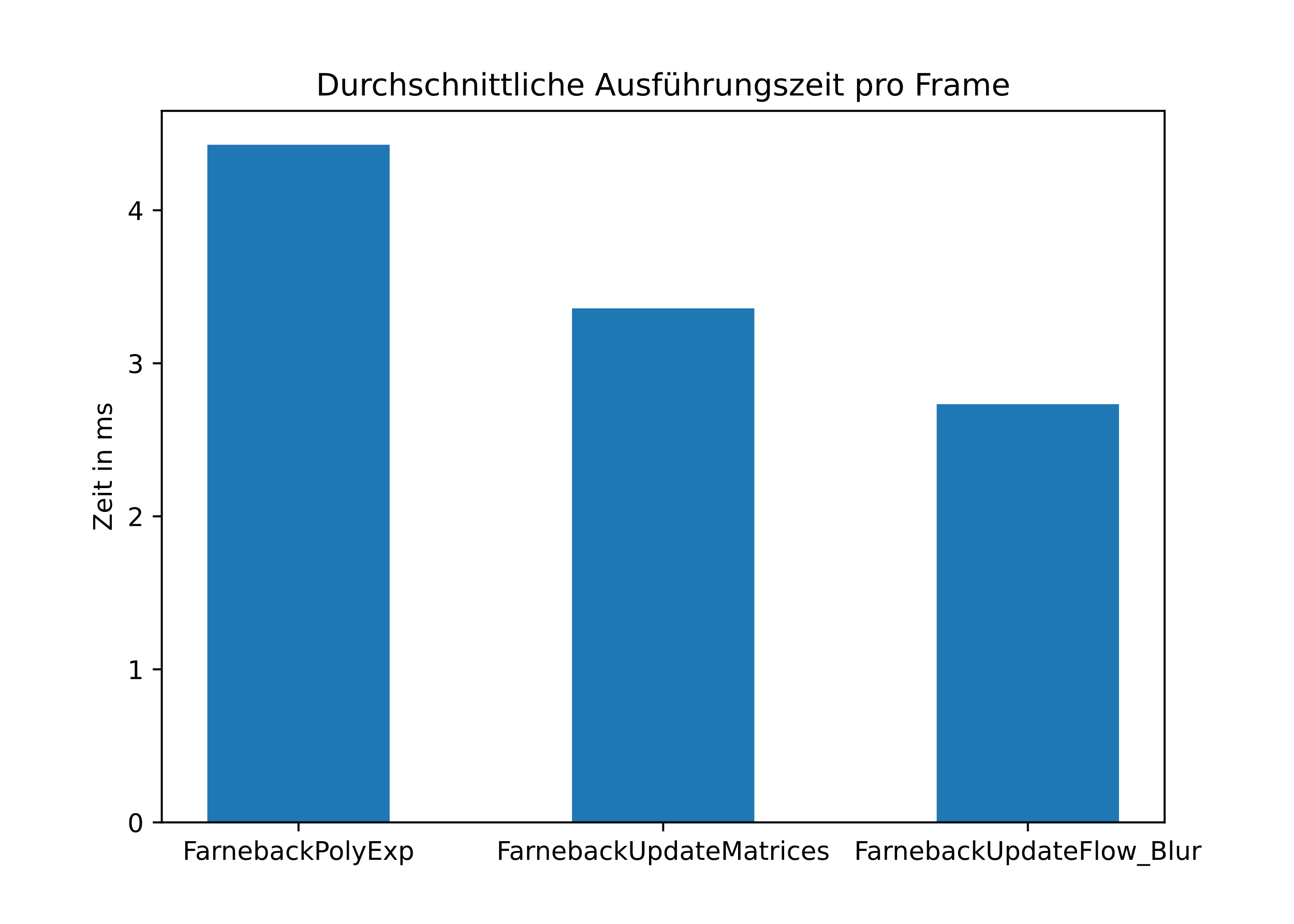


Abbildung C.1, Durchschnittliche Ausführungszeiten pro Frame der Funktionen FarnebackPolyExp, FarnebackUpdateMatrices und FarnebackUpdateFlow\_Blur. Gemessen auf einer Intel(R) Core(TM) i7-9850H CPU.

|  |  |
| --- | --- |
| Funktion | berechnete Ausführungszeit |
| FarnebackPolyExp | 4,428 ms |
| FarnebackUpdateMatrices | 3,359 ms |
| FarnebackUpdateFlow\_Blur | 2,733 ms |

Tabelle C.1, berechnete Ausführungszeiten von FarnebackPolyExp, FarnebackUpdateMatrices und FarnebackUpdateFlow\_Blur in Millisekunden.

Auf den ersten Blick ist FarnebackPolyExp die Funktion mit der längsten durchschnittlichen Ausführungszeit. Dennoch sollte auch in Betracht gezogen werden, dass die Funktionen mehr als einmal von calcOpticalFlowFarneback aufgerufen werden. Deshalb wird neben den durchschnittlichen Ausführungszeiten auch die Verteilung der Ausführungszeit von calcOpticalFlowFarneback auf die einzelnen Funktionen betrachtet und in Abbildung C.2 festgehalten.

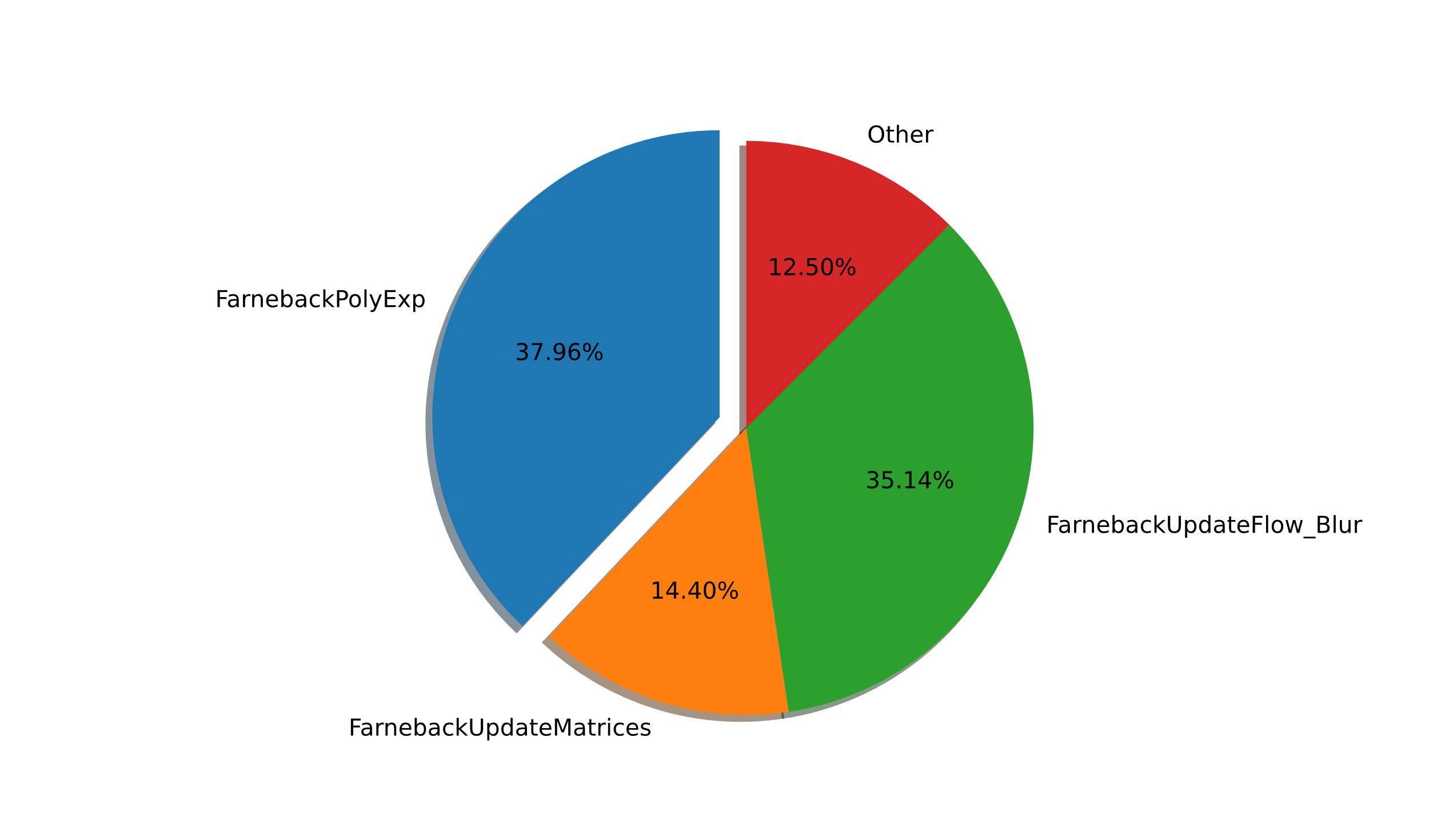


Abbildung C.2, Verteilung der Ausführungzeiten von calcOpticalFlowFarneback auf die Funktionen FarnebackPolyExp, FarnebackUpdateMatrices und FarnebackUpdateFlow\_Blur.

Other bezeichnet hier Code, welcher von calcOpticalFlowFarneback direkt ausgeführt wird und nicht in eine andere Funktion ausgelagert ist. Zu bemerken ist, dass FarnebackUpdateFlow\_Blur einen signifikanten Anteil der Ausführungszeit von calcOpticalFlowFarneback zu verschulden hat. Obwohl FarnebackUpdateFlow\_Blur die niedrigste durchschnittliche Ausführungszeit vorzeigt, hat sie einen ähnlich großen Anteil der Ausführungszeit von calcOpticalFlowFarneback wie FarnebackPolyExp. Dies kann begründet werden durch die unterschiedliche Anzahl der jeweiligen Funktionsaufrufe innerhalb eines Aufrufs von calcOpticalFlowFarneback. Mit den in der Messung verwendeten Parametern von calcOpticalFlowFarneback wird FarnebackPolyExp acht male, FarnebackUpdateMatrices vier male und FarnebackUpdateFlow\_Blur zwölf male aufgerufen.

Auch nach mehreren Ausführungen von „densFlow“ sind keine nennenswerten Änderungen der Messergebnisse zu beobachten. Daraus lässt sich schließen, dass im Durchschnitt die Funktion FarnebackPolyExp die meiste Zeit für ihre Ausführung benötigt.

Aufgrund der erfassten Gegebenheiten wird die Funktion FarnebackPolyExp auf mögliche Parallelisierung geprüft. Hierbei soll besonders Wert auf Datenparallelität gelegt werden.

Als erstes sollen verwendete Schleifen analysiert werden, da diese in der Regel am meisten Zeit in Anspruch nehmen.

FarnebackPolyExp beinhaltet insgesamt sieben Schleifen, welche in Codeausschnitt C.1 zusammengefasst sind. Die steuernden Parameter dieser Schleifen sind „height“, „width“ und n. Bei „height“ und „width“ handelt es sich um Höhe und Breite des Inputframes. Der Parameter n gibt den Radius des Neighbourhoods an in welchem das Polynom angenähert wird. Die komplette FarnebackPolyExp Funktion ist im Anhang zu finden.



Codeausschnitt C.1, Darstellung der genutzten Schleifen in der Funktion FarnebackPolyExp.

Die Schleife in Zeile 132 iteriert über die Höhe des übergebenen Frames. Innerhalb dieser Schleife wird dann mehrmals über die Breite des Frames iteriert. Hier in Zeile 138, Zeile 150 und Zeile 169 zu sehen. Außerdem sind Schleifen vorhanden, die in Bezug zur Variable n ausgeführt werden. Die Variable n beschreibt hier den Radius des verwendeten Neighbourhoods und liegt in realistischen Anwendungen bei Ganzzahlwerten zwischen fünf und sieben.

Wie in Kapitel B-III angesprochen berechnet FarnebackPolyExp die Koeffizienten des Polynoms mithilfe eines Neighbourhoods. FarnebackPolyExp ist soweit optimiert, dass das iterieren über den kompletten Inputframe in einen horizontalen und einen vertikalen Teil aufgeteilt wird. Aufgrund dieser Gegebenheit kann die Schleifenstruktur von FarnebackPolyExp so zusammengefasst werden, dass die Schleifen im Grunde über den ganzen Inputframe iterieren. Berechnungen durch das Neigbourhood um einen Pixelwert herum gemacht. Dies deutet stark auf Datenparallelität hin.

## Lösungsansätze / Probleme

### Beschreibung möglicher Lösungsansätze

Im Ersten Lösungsansatz soll die Reihenfolge der Operationen beibehalten werden um eine mögliche Einführung von Fehlern in der Berechnung zu verhindern. Außerdem soll der Grundgedanke der separablen Faltung beibehalten werden. Da es sich bei den Schleifen von FarnebackPolyExp um einfache for-Schleifen handelt, können diese in Befehle der Allgorithms Library übersetzt und parallelisiert werden. Vor allem Befehle wie „std::for\_each“ oder „std::transform“ sind dafür geeignet.

Der zweite Lösungsansatz greift mehr in die Struktur des originalen Codes ein. Die originale Implementierung von FarnebackPolExp berechnet die Polynomkoeffizienten Bildzeile für Bildzeile und nicht Pixel für Pixel. Da es sich um eine Implementierung von OpenCV handelt, werden viele OpenCV eigene Klassen verwendet. Dies könnte ein Grund für die Aufteilung in Bildzeilen sein. Um die Möglichkeiten einer GPU auszuschöpfen, soll in diesem Ansatz die Berechnung Pixel für Pixel von statten gehen. In dieser Weise wird dann auch von der Datenparallelität Gebrauch gemacht.

### Evaluierung der Lösungsansätze

Ein Hauptproblem des ersten Ansatzes ist die komplexität des orginalen Codes. Wird der originale Code in Befehle der Allgorihms Library übersetz, kann es zu Performance-Einbußen kommen. Für jeden Befehl der Allgorithms Library, der mit einer Execution Policy versehen ist, wird ein Kernel gestartet und auf der GPU ausgeführt. Wenn für jede Zeile des Bildes ein neuer Kernel aufgerufen wird besteht die Möglichkeit, dass mehr Zeit gebraucht wird um die Kernel zu starten als sie auszuführen. Diese Möglichkeit entsteht durch das Synchronisieren der Daten zwischen den Kernelaufrufen. Deshalb ist es sinnvoller nur einen einzigen Kernelaufruf zu haben, welcher jeden Pixel des Bildes verwendet. Der zweite Lösungsansatz macht genau dies, indem nur ein Algorithms Library Befehl aufgerufen wird. Da es Ziel dieses Ansatzes ist die Berechnung für jeden Pixel einzeln zu machen, wird nur ein Algorithms Library befehl benötigt. Da der originale Code signifikant geändert werden muss, wird der zweite Lösungsansatz einen höheren Aufwand bedeuten.

## Ausgewählter Ansatz und Durchführung

Erste Versuche den ersten Lösungsansatz zu implementieren zeigten sehr schlechte Performance. Deshalb wird der zweite Lösungsansatz, den originalen Code in eine Pixel für Pixel Berechnung umzuschreiben und zu parallelisieren, ausgewählt. Hierfür ist folgende Vorgehensweise geplant. Zuerst soll der originale Code in eine Pixel für Pixel Berechnung übersetzt werden. Ist dies gelungen, wird mit Hilfe der von Algorithms Library bereitgestellten Befehle und Execution Policies parallelisiert. Danach soll der parallelisierte Code mit dem NVC++ Compiler der Nvidia HPC SDK kompiliert und gebaut werden um ihn auf der GPU auszuführen.

### Detailanalyse der Polynomkoeffizienten Berechnung

Um die Berechnung in eine Pixel für Pixel Berechnung umzuschreiben, wird im Detail analysiert, wie der originale Code die Polynomkoeffizienten berechnet. Codeausschnitt C.2 zeigt den Beginn der Polynomkoeffizienten Berechnung. Sämtliche Berechnung findet innerhalb der Schleife in Zeile 132 statt. Diese Schleife nimmt die variable y = 0 als Startwert und inkrementiert sie bis zur Höhe des Inputframes. In Zeile 135 wird die variable y dann verwendet um Pointer auf korrespondierenden Zeilen des Inputframes zu erhalten. Der Inputframe ist in der Matrix „src“, welche als Parameter an die Funktion übergeben wurde, gespeichert. In Zeile 136 wird ein Pointer auf die Zeile einer weiteren Matrix „dst“ gespeichert. Auch diese wurde als Parameter übergeben und dient später zur Speicherung der Polynomfaktoren.



Codeausschnitt C.2, Pointer werden verwendet Zeilen des Inputframes referenzieren zu können.

Die Schleife in Zeile 138 des Codeausschnitts C.3 nimmt die variable x = 0 als Startwert und inkrementiert sie bis zur Breite des Inputframes. Genutzt wird sie zum befüllen eines Buffers „row“ mit Initialwerten. Es ist zu beobachten, dass der Buffer für jeden x-Wert drei Werte speichert.



Codeausschnitt C.3, ein Buffer wird mit Initialwerten befüllt.

Direkt im Anschluss folgt in Codeausschnitt C.4 eine Schleife, welche mit „k = 1“ startet und bis zur variable n inkrementiert. Innerhalb dieser Schleife befindet sich in Zeile 149 eine weiter Schleife, die wie in Zeile 138 zuvor über Werte gleich der Breite des Inputframes iteriert. In den Zeilen 146 und 147 werden jeweils Pointer zu Inputframezeilen über und unter der anfangs in Zeile 135 genannten Zeile akquiriert. Hierbei wird der Abstand durch die variable k festgelegt. Dies deutet auf die Berechnung innerhalb eines Neighbourhoods hin. Die eben erwähnte Schleife in Zeile 149 berechnet dann Werte mithilfe der zwei in Zeilen 146 und 147 erhaltenen Inputframezeilen und addiert diese jeweils zu den bereits vorhandenen Werten im „row“ Buffer. Auch hier werden wieder drei Werte für jeden x-Wert berechnet.



Codeausschnitt C.4, der vertikale Teil der Faltung wird für eine Zeile berechnet und im Buffer gespeichert.

Wichtig festzuhalten ist, dass bis zu diesem Punkt Werte aus einem vertikalen Neigbourhood in „row“ zusammengefügt wurden. Außerdem werden die Ränder des Inputframes bei der Akquisition der Inputframezeilen im Neigbourhood ebenfalls beachtet.

„std::max“ und „std::min“ sorgen dafür, dass falls das Neighborhood auf eine Inputframezeile außerhalb des Inputframes zugreifen möchte, die Randzeile des Inputframes stattdessen genommen wird. Somit wird die Randzeile effektiv kopiert und imaginär an den Rand des Inputframes angehängt. Zusammengefasst wurde in vertikaler Richtung das Neigbourhood für jeden x-Wert der Inputframezeile bearbeitet.

Um für die Randbehandlung die gleiche Funktionalität wie bisher bieten zu können, werden mit der Schleife in Codeausschnitt C.5 die ersten drei Werte am Anfang des Buffers n-mal an den Anfang angefügt und die letzten drei Werte des Buffer n-mal an das Ende angefügt.



Codeausschnitt C.5, Randwerte werden kopiert um „out-of-bounds“ Zugriffe zu vermeiden.

Es ist festzuhalten, dass die Randbehandlung sowohl in vertikaler als auch in horizontaler Richtung durch kopieren der Randwerte realisiert ist.

Da der vertikale Teil der Berechnung für eine Inputframezeile abgearbeitet ist, folgt nun in Codeausschnitt C.6 der horizontale Teil. Zeilen 172 – 174 initialisieren Variablen für die zu berechnenden Polynomfaktoren. Die Schleife in Zeile 175 wird verwendet um das horizontale Neighbourhood jedes x-Wertes zu bearbeiten. Die aus diesem Neigbourhood berechneten Werte werden dann in der jeweiligen Faktorvariable gespeichert. Nachdem das Neigbourhood für einen x-Wert abgearbeitet ist werden die endgültigen Polynomfaktoren berechnet und in der Matrixzeile „drow“ der 5-Chanel Matrix „dst“ gespeichert. Dies geschieht in Zeile 187 – 191. Damit ist dann eine Iteration der ersten Schleife aus Zeile 132 beendet.



Codeausschnitt C.6, der Horizontale Teil der Faltung wird abgearbeitet und die Polynomkoeffizienten werden berechnet.

Zusammenfassend wird im originalen Code der Inputframe Zeile für Zeile eingelesen. Dann werden über ein Neigbourhood für jeden Pixel der Inputframezeile drei Werte berechnet. Mit Hilfe dieser drei Werte werden dann die Polynomfaktoren berechnet und in der Zielmatrix gespeichert. Das Schema der Zusammenführung des Neigbourhoods einer Inputframezeile wird durch Abbildung C.3 veranschaulicht.

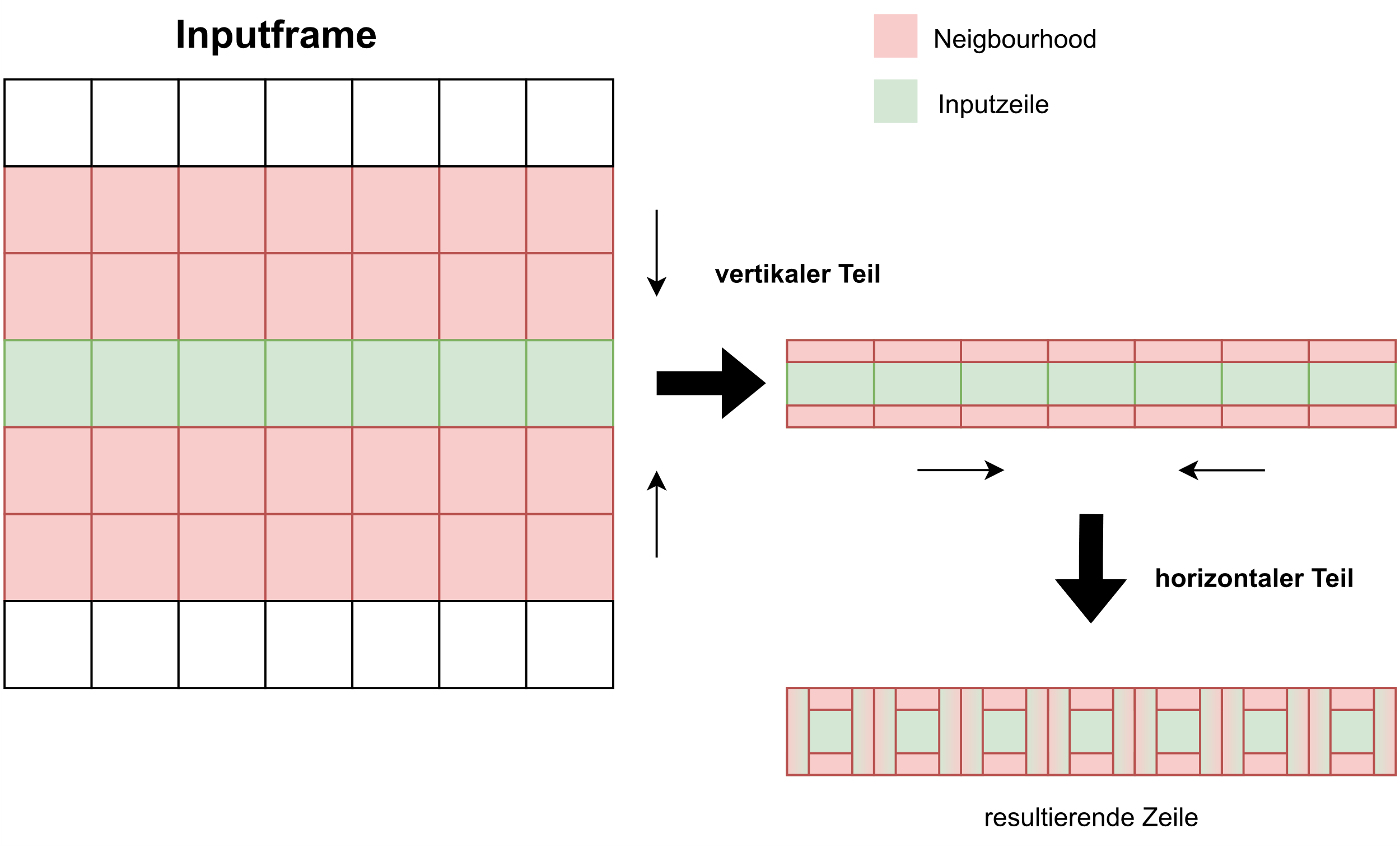


Abbildung C.3, zuerst wird das Neigbourhood der Inputzeile in vertikaler Richtung in eine Zeile zusammengefügt. Danach wird für jeden Wert der Zeile das Neighbourhood in horizontaler Richtung zusammengefügt.

### Übersetzung in Per-Pixel-Berechnung

Das zuvor beschriebene verhalten wird nun in eine Pixel für Pixel Berechnung übersetzt. Der vollständige Code ist in Anhang A zu finden. Die Eigenschaft der Aufteilung der Berechnung in einen vertikalen und horizontalen Teil soll dabei beibehalten werden. Abbildung C.4 veranschaulicht das Schema, in dem das Neibourhood bearbeitet werden soll.

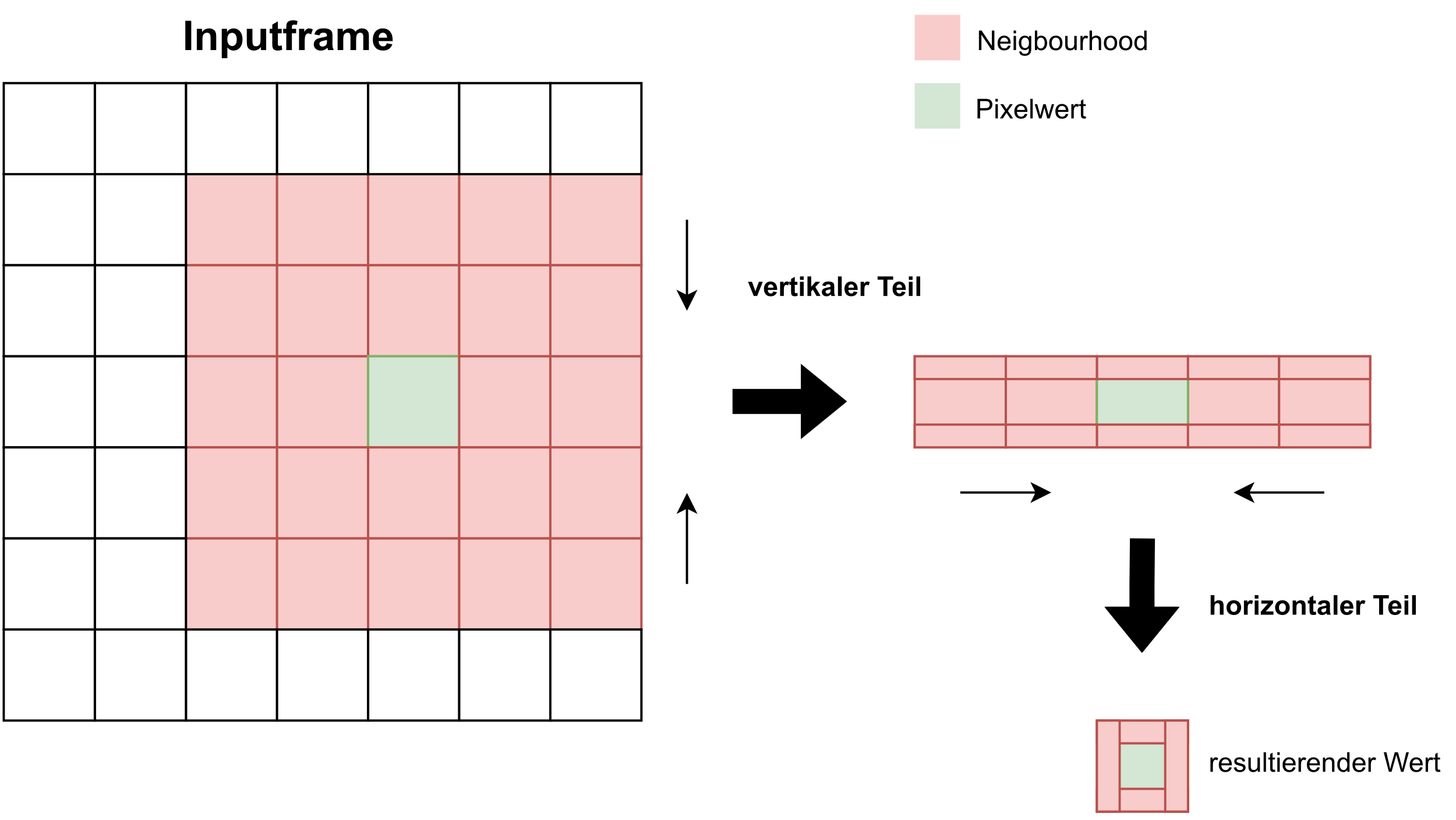


Abbildung C.4, zuerst wird das Neigbourhood in vertikaler Richtung in eine Zeile zusammengefügt. Danach wird diese Zeile in horizontaler Richtung zusammengefügt

Um über den kompletten Inputframe iterieren zu können werden zwei verschachtelte Schleifen benötigt. Die Variablen „height“ und „width“ entsprechen hier der Höhe und der Breite des Inputframes. Außerdem wird ein Buffer benötigt, welcher dreimal so groß ist wie das Neigbourhood, um alle benötigten Werte zwischenspeichern zu können. Codeausschnitt C.7 zeigt den Aufbau dieser Schleifen.



Codeausschnitt C.7, zwei verschachtle Schleifen werden verwendet um über den Inputframe zu iterieren. Außerdem wird ein Buffer eingeführt um benötigte werte Zwischenzuspeichern.

Danach wird der vertikale Teil, veranschaulicht durch Codeausschnitt C.8, des Neigbourhoods um einen Pixel bearbeitet. Dies geschieht durch zwei verschachtelte Schleifen. Die äußere Schleife iteriert über die Breite des Neighbourhoods, welches durch den Term „2\*n+1“ beschrieben werden kann. Außerdem wird eine neue variable „neighX“ eingeführt, welche als x-Koordinate für das Neighbourhood dient. Die innere Schleife wird verwendet um Pixel in vertikaler Richtung innerhalb des Neigbourhoods abzuarbeiten. Zuerst werden neue variablen „neighY0“ und „neightY1“ eingeführt um die y-Koordinaten im Neighbourhood zu bestimmen. Dann werden für jede x-Koordinate im Neighbourhood drei Werte berechnet und im Buffer gespeichert. Wie auch im originalen Code wird die Randbehandlung durch „std::min“ und „std::max“ durchgeführt. Falls das Neighbourhood auf Koordinaten außerhalb des Inputframes zugreifen würde, wird stattdessen die auf die Werte des Randes zugegriffen.



Codeausschnitt C.8, der vertikale Teil der Polynomkoeffizienten Berechnung für einen Pixel.

Nachdem das Neigbourhood eines Pixels im vertikalen Teil auf eine Zeile des Neigbourhoods zusammengefasst wurde, wird nun in Codeausschnitt C.9 der horizontale Teil behandelt. Hierbei werden die Variablen „b1“ – „b6“ verwendet um nach dem Zusammenfassen der Neigbourhoodzeile direkt die Polynomfaktoren berechnen zu können.



Codeausschnitt C.9, der horizontale Teil der Polynomkoeffizienten Berechnung für einen Pixel.

Die in den Variablen „b1“ – „b6“ gespeicherten Werte werden dann, wie im originalen Code zuvor, dazu verwendet die Polynomkoeffizienten für einen Pixel zu berechnen und in der Matrix „dst“ zu speichern. Dies wird in Codeasuchnitt C.10 dargestellt.



Codeausschnitt C.10, die endgültigen Polynomkoeffizienten werden Berechnet und in einer Matrix gespeichert.

### **Parallelisierung der Per-Pixel-Berechnung**.

Die Funktion FarnebackPolyExp ist nun auch als per Pixel Berechnung implementiert. Unter Verwendung der Algorithms Library kann nun mit geringem Aufwand paralleler Code entstehen. Hierfür werden in Codeausschnitt C.11 die zwei äußersten Schleifen, welche über den Inputframe iterieren, durch einen „std::foreach“ Aufruf ersetzt. Da für jedes Pixelelement innerhalb der Schleifen mehr als nur eine Operation benötigt wird, sind „std::transform“ und „std::transform\_reduce“ nicht geeignet. Die übergebene Execution Policy signalisiert, dass der Aufruf parallelisiert werden darf.



Codeausschnitt C.11, „std::for\_each“ wird verwendet um über den Inputframe zu iterieren und macht Parallelisierung möglich.

Innerhalb der Lambda-Funktion befindet sich fast derselbe Code wie innerhalb der ersetzen zweifach verschachtelten Schleifen. Da nun bei jeder Iteration nur ein Pixel des Inputframes zur Verfügung steht und keine Informationen zur Position des Pixels, muss der Inhalt der Lambda-Funktion angepasst werden. Um auf Pixel des Neighbourhoods zuzugreifen werden x- und y-Koordinaten benötigt. Diese werden berechnet indem der Index des Pixels mit Hilfe von Pointer-Arithmetik berechnet wird. Die Speicheradresse auf die der Pointer des Inputframes zeigt wird subtrahiert von der Speicheradresse des aktuellen Pixels. Mit dem dadurch berechneten Index werden x- und y-Koordinate berechnet. Codeausschnitt C.12 zeigt diese Berechnungen.



Codeausschnitt C.12, durch Pointer-Arithmetik wird der Index des verwendeten Pixels brechnet.

Diese Berechnung ist nur dann möglich, wenn der Inputframe konsekutiv im Speicher liegt.

Nachdem der parallele Code implementiert ist, kann dieser für die Ausführung auf einer GPU kompiliert und getestet werden. Hierfür wird der NVC++ Compiler der NVIDIA HPC SDK verwendet.

Versuche den Code mit Hilfe des NVC++ Compilers zu kompilieren ergeben mehrere Probleme. Diese können zurückgeführt werden auf die Verwendung von OpenCV. Der NVC++ Compiler kann die Definitionen der OpenCV Klassen nicht sehen, da die Klassen über die OpenCV Library inkludiert werden. Somit ist die Nutzung von CUDA Unified Memory nicht möglich und der Code kann aufgrund dessen nicht erfolgreich ausgeführt werden.

Die zuvor Beschriebenen Probleme erfordern umfassende Änderung der originalen Implementierung von OpenCV um sie zu lösen. Da die Parallelisierung mit der Parallel STL und die dadurch resultierende Performance im Vordergrund steht, wird die Funktion FarnebackPolyExp von der OpenCV Implementierung gekapselt. Hierfür wird ein neues Programm erstellt, welches zur Berechnung Platzhalterdaten verwendet. Diese Platzhalterdaten sollen einem Videoframe ähneln. Auch die ursprüngliche Hauptfunktionalität soll beibehalten werden.

Um die in „cv::Mat“ Klassen gespeicherten Videoframedaten zu ersetzen, geniert das neu erstellte Programm „polyExp-stl“ Platzhalterdaten und speichert diese in einem eindimensionalen „std::vector“ Array. „std::vector“ wird verwendet, da der Speicher für diesen Vektor auf dem CPU Heap allokiert wird. Dies ist Voraussetzung um die Daten mit CUDA Unified Memory nutzen zu können. Die berechneten Polynomkoeffizienten werden ebenfalls ein einem „std::vector“ gespeichert. Die zuvor erarbeitete Per-Pixel-Implementierung der FarnebackPolyExp Funktion wird von „polyExp-stl“ aufgerufen. Anstatt der Videoframedaten werden nun die Platzhalterdaten als Parameter übergeben. (ig values static)

Erstmals kann nun die neue FarnebackPolyExp Implementierung auf der GPU ausgeführt werden. Erste Performance Messungen sind in Abbildung C.5 dargestellt.

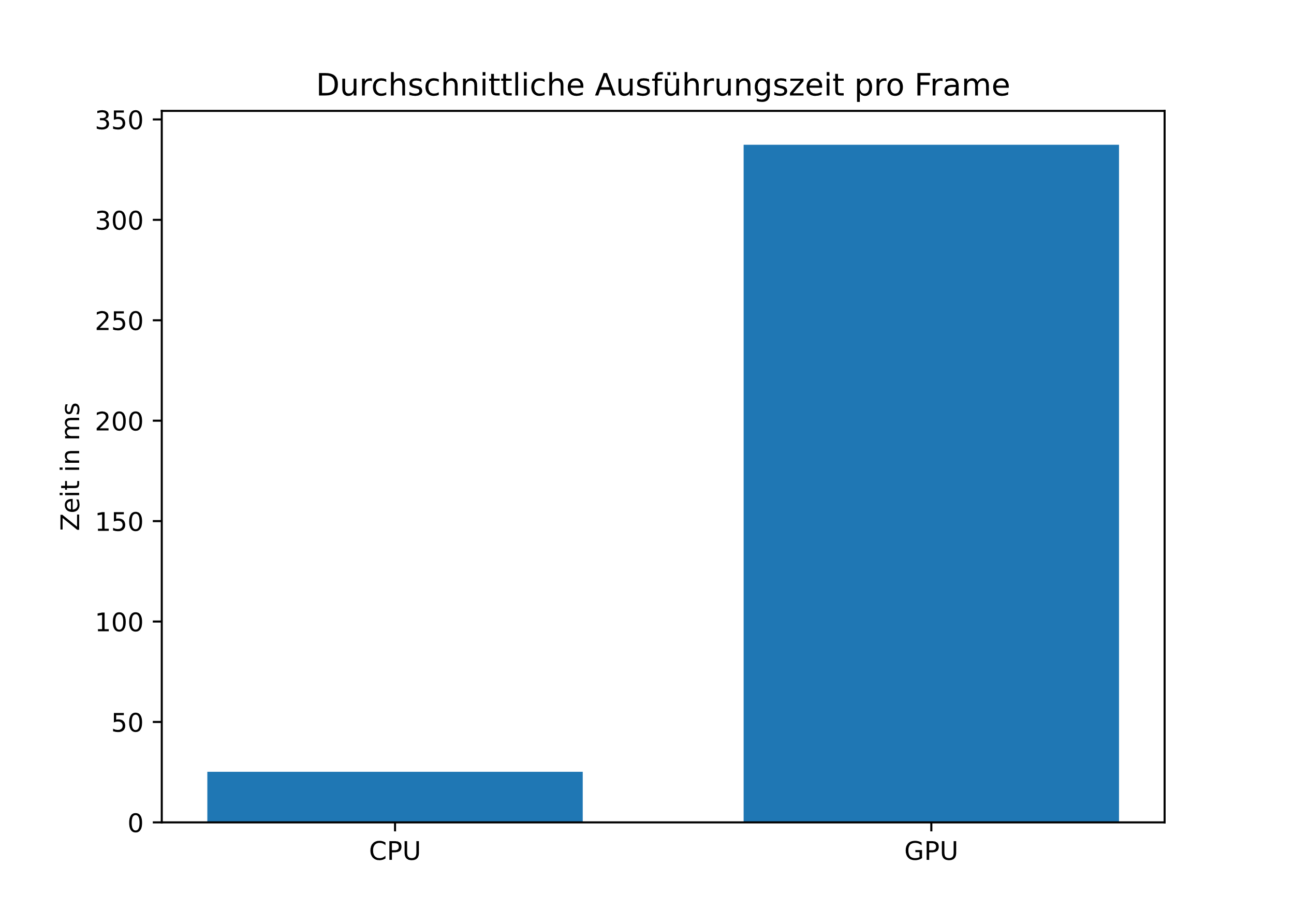


Abbildung C.5, Performance gemessen auf einer Nvidia Quadro P2200 GPU. Das Balkendiagramm zeigt die Durchschnittliche Ausführungszeit von FarnebackPolyExp pro Frame.

Die erhaltenen Messergebnisse zeigen einen deutlichen Unterschied in der Ausführungszeit zwischen CPU und GPU. Um die unerwartet lange Ausführungszeit begründen zu können wird mit Hilfe von einem Profiling-Tool der Nvidia HPC SDK die Ausführung genauer analysiert. Hierfür muss die GPU-Implementierung allerdings auf einer anderen GPU ausgeführt werden, da das Profiling-Tool die zuvor genutzte Hardware nicht unterstütz. Zum Profiling wird eine GPU einer neueren Generation verwendet. Die Profilingergebnisse zeigen eine hohe Anzahl von Speicher- und Ladebefehlen, welche reduziert werden sollen.

Der Letzte Schritt in der Parallelisierung der FarnebackPolyEXP Funktion ist es die Anzahl von benötigten Speicherzugriffen zu reduzieren und somit die bisher erarbeitete Implementierung zu optimieren. Als konkreten Optimierungsschritt soll die Anzahl der verwendeten Zwischenspeicher innerhalb des Kernels reduziert werden. Um die Berechnung der Polynomkoeffizienten effizienter zu machen, wird in Codeausschnitt C.13 auf den Zwischenspeicher für die horizontale Abarbeitung des Neigbourhoods verzichtet. Die Werte des vertikalen Neighbourhoods werden nicht mehr zwischengespeichert. Stattdessen wird jeder erhaltene Wert mit dem richtigen Faktor multipliziert und je nach Koeffizienten und Position innerhalb des Neigbourhoods auf die Variablen der Polynomkoeffizienten addiert oder subtrahiert.



Codeausschnitt C.13, Optimierung durch Verzicht auf Zwischenspeicher und Reduzierung der benötigten Schleifen.

Die CPU-Implementierung, die ursprüngliche GPU-Implementierung und die optimierte GPU-Implementierung werden in Abbildung C.6 verglichen.

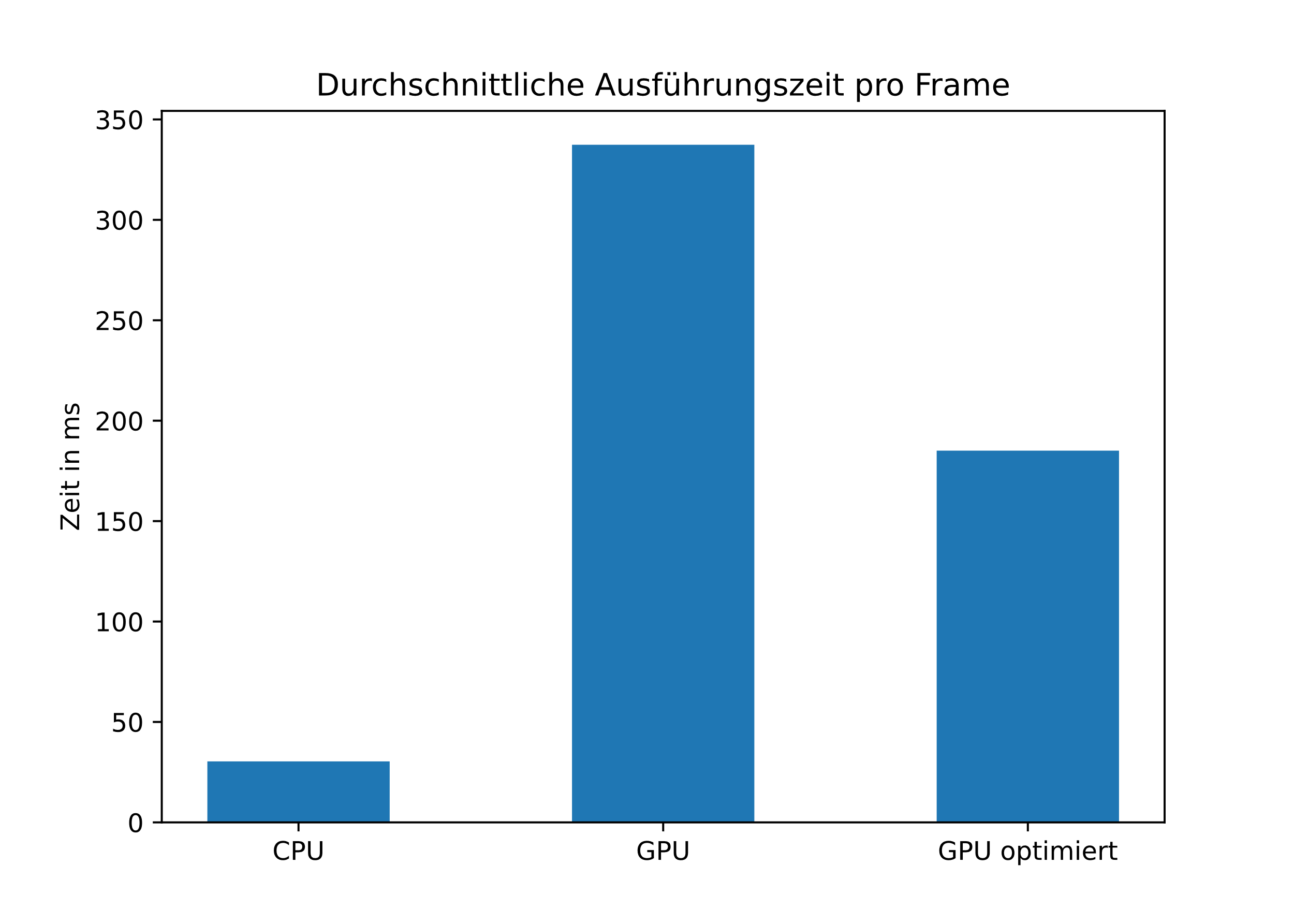


Abbildung C.6, Vergleich der Durchschnittlichen Ausführungszeit zwischen der CPU-Implementierung, der nicht optimierten GPU-Implementierung und der optimierten GPU-Implementierung.

# Ergebnisse

## Validierung der Ergebnisse und Korrektheit der Implementierung

Da die Messergebnisse der verschiedenen Implementierung verglichen werden, müssen die Ergebnisse der Funktionen immer identisch sein. Da es sich bei den berechneten Polynomkoeffizienten um floating-point Werte handelt, können gewisse Abweichungen entstehen. Die Abweichung zwischen den Polynomfaktoren der Original-Implementierung und der Per-Pixel-Implementierung sollten möglichst gering sein.

Um die Korrektheit der FarnebackPolyExp Implementierungen gewährleisten zu können wird der durchschnittliche absolute Fehler der Polynomkoeffizienten aus der Original-Implementierung und den Per-Pixel-Implementierungen berechnet:

A: Polynomkoeffizienten der Original-Implementierung

B: Polynomkoeffizienten der Per-Pixel-Implementierung

b: Summe der absoluten Differenz

d: durchschnittliche absolute Fehler

In Tabelle D.1 sind die Berechneten absoluten Fehler der jeweiligen Implementierungen festgehalten. Alle Implementierung berechnen die Polynomfaktoren korrekt, da die durchschnittlichen absoluten Fehler äußerst gering sind.

|  |  |
| --- | --- |
| Implementierung | Durchschnittliche absolute Fehler |
| Per-Pixel | 0,0 |
| Per-Pixel STL | 0,0 |
| Per-Pixel STL auf GPU | 0,0 |
| Per-Pixel STL auf GPU optimiert | 4,78885410879754e-07 |

Tabelle D.1, Auflistung der berechneten durchschnittlichen absoluten Fehler der erarbeiteten Implementierungen.

## Übersicht der Ergebnisse

Sowohl die Original Implementierung, als auch die Per-Pixel Implementierung wurden auf einem einzigen CPU-Kern ausgeführt. Durch die Einführung eines Parallel STL Algorithmus mit Execution Policy wurde die Per-Pixel Implementierung dann auf mehreren CPU-Kernen ausgeführt. Die Per-Pixel Implementierung mit Parallel STL wurde mit Hilfe des Nvidia NVC++ Compilers für die Ausführung auf einer GPU kompiliert.

Um die Ausführungszeit zu reduzieren, wurde die Implementierung anschließend optimiert. Abbildung D.1 zeigt die erreichten durchschnittlichen Ausführungszeiten der verschiedenen Implementierungen:

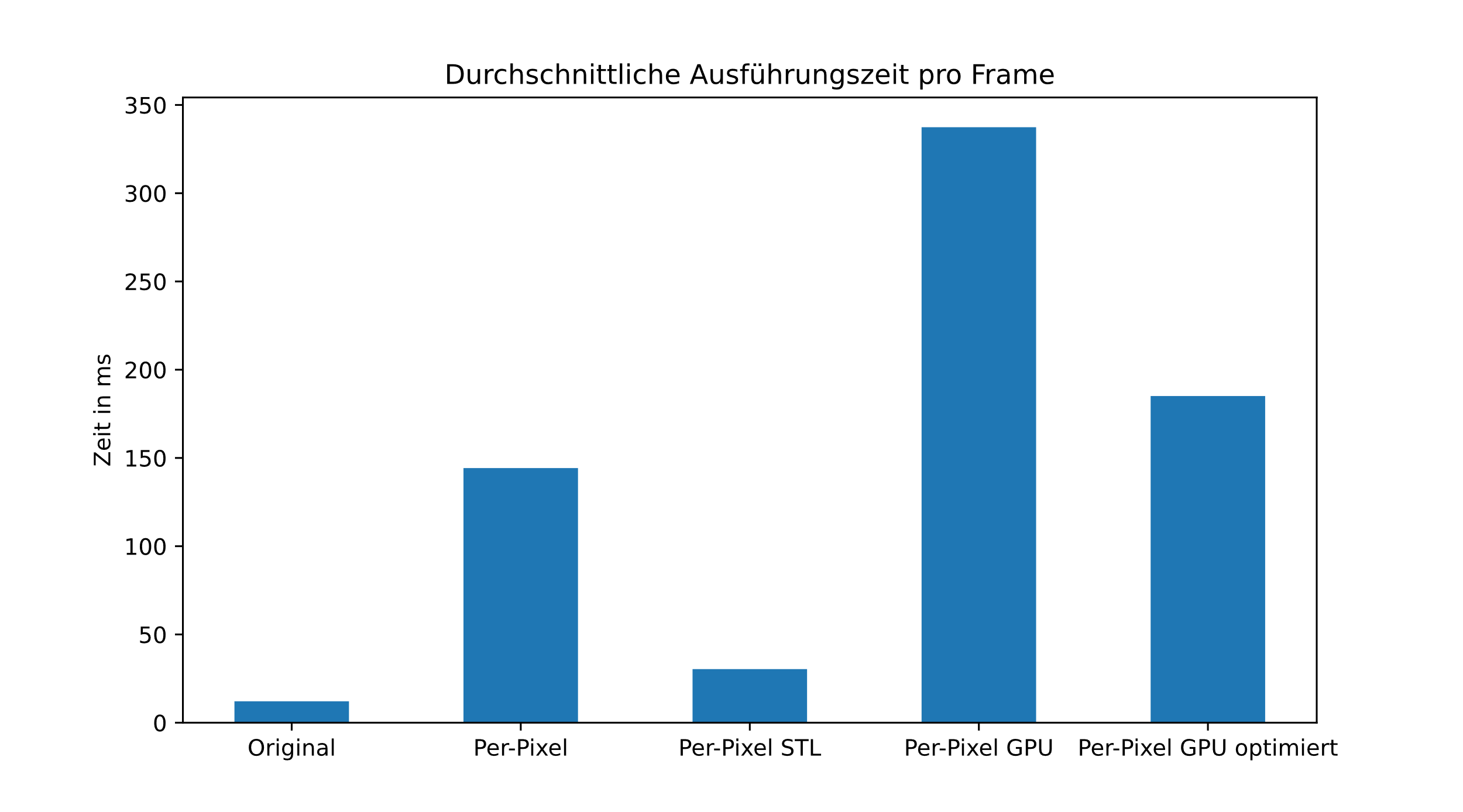


Abbildung D.1, durchschnittliche Ausführungszeiten pro Frame für alle erarbeiten Implementierungen

|  |  |
| --- | --- |
| Implementierung | Durchschnittliche Ausführungszeit |
| Original auf CPU | 12,123 ms |
| Per-Pixel auf CPU | 144,257 ms |
| Per-Pixel STL auf CPU parallelisiert | 30,360 ms |
| Per-Pixel STL auf GPU | 337,379 ms |
| Per-Pixel STL auf GPU optimiert | 185,078 ms |

Tabelle D.1, Auflistung der gemessenen durchschnittlichen Ausführungszeiten in Millisekunden.

Die mit einer GPU parallelisierten Implementierungen haben beide eine höhere durchschnittliche Ausführungszeit als die nicht parallelisierten Implementierungen.

Durch Optimierung der Speichernutzung wurde die Anzahl der benötigten Speicher- und Ladebefehle, zu sehen in AbbildungX, reduziert. Dies führte zu einer Verbesserung in der Ausführungszeit der GPU parallelisierten Implementierung.

Profilingergebnisse vergleich lade befehele

## Ergebnisdiskussion und Erkenntnisse

### Bewertung der Ergebnisse

Der Algorithmus ist nicht komplett parallelisiert, sondern nur ein Teil des Algorithmus welcher den größten Performance Einfluss hat. Die Ausführung auf einer GPU verschlechtert hier die Allgemeine Perfomance des Algorithmus. Die ursprüngliche Implementierung des Algorithmus ist für die Ausführung auf einer CPU gedacht und dementsprechend optimiert. Mit Hilfe der Parallel STL und der Nvidia Compiler ist es möglich Teile dieses CPU-Codes auf einer GPU auszuführen. Die schlechten Ausführungszeiten lassen sich dadurch begründen, dass die GPU-Implementierung nicht für die genutzte GPU Hardware optimiert ist. Durch geringe Optimierung ist die Ausführungszeit schon reduziert worden. Wenn weitere Maßnahmen ergriffen werden um die GPU-Implementierung besser für GPU Hardware zu optimieren, kann sich die Ausführungszeit weiter verbessern. Eine Maßnahme könnte sein eine Art Pipelining einzuführen. So könnten die Daten des nächsten Frames schon auf den GPU Speicher kopiert werden, während die GPU mit der Berechnung der Polynomkoeffizienten beschäftigt ist. Jedoch ist eine solche Art der Optimierung mit der Parallel STL nicht möglich.

### Gesammelte Erfahrungen und Erkenntnisse

#### Erstellung von parallelem Code

Durch die Algorithmen der Parallel STL ist es möglich sequenziellen Code schnell und unkompliziert zu parallelisieren oder von Anfang an parallelen Code zu schreiben. Für die Ausführung auf einer GPU muss auf Limitierungen in Software und Hardware geachtet werden. So kann zum Beispiel die Parallel STL nicht auf GPU Architekturen älter als Pascal verwendet werden. Außerdem beinhaltet der NVC++ Compiler nicht alle Header der Standard Library und fehlende Header müssen extra inkludiert werden.

#### Einfluss auf erstellte Kernel

Werden Algorithmen der Parallel STL auf einer GPU ausgeführt, erstellt der NVC++ Compiler für jeden Algorithmus einen Kernel. Da dies alles Automatisch passiert, hat der Entwickler keinen direkten Einfluss auf Parameter mit denen der Kernel Erstellt wird. Zu diesen Parametern gehören die Anzahl von Threads pro Block oder die Anzahl von Blöcken pro Grid. Der NVC++ Compiler entscheidet diese Anzahl legt sie für den Kernel fest. Außerdem werden sämtliche Datentransfers und Synchronisation von CUDA Unified Memory verwaltet. Somit sind auch hier keinerlei Anpassungsmöglichkeiten gegeben.

#### Compilerbugs und Fehlersuche

Das Kompilieren mit dem NVC++ Compiler lief nicht reibungslos. Es kann durchaus vorkommen, dass die Kompilierung mit internen Compilererrors fehlschlägt. Um zu verstehen warum diese Fehler auftreten sollte der Entwickler Erfahrung mit GPU Programmierung haben, da Fehlermeldungen des NVC++ Compilers nicht Aussagekräftig genug sind um die Fehlerquelle sofort identifizieren zu können. Meist musste geraten und ausprobiert werden welche Teile des Codes zu Fehlern in der Kompilierung führen.

# Zusammenfassung

Abstract wiederholen -> Ergebnisse zusammmenfassen -> Ausblick schaffen

Mit Hilfe der in C++17 eingeführten Parallel STL Algorithmen ist es möglich die Parallelisierung dem Compiler zu überlassen. Das Nvidia HPC Software Development Kit unterstütz mit seinem eigenen C++ Compiler die Parallel STL Algorithmen und macht ihre Ausführung auf GPUs möglich. In dieser Arbeit sind anhand des Beispiels der Berechnung des Optischen Flusses, die Möglichkeiten der Parallel STL Algorithmen in Verbindung mit der Nvidia HPC SDK erfasst worden. Die erhaltenen Ergebnisse zeigen Verschlechterungen der Laufzeiten. Durch Optimierung können diese reduziert werden. Es zeigt sich, dass die Möglichkeiten der Parallel STL in Verbindung mit dem Nvidia Toolkit begrenzt sind. Es ist nicht möglich den aus dem Parallel STL Algorithmus erstellten GPU-Kernel direkt anzupassen. Dadurch können keine weiteren GPU spezifischen Optimierung durgeführt werden. -> Ausblick?

# Literaturverzeichnis

[1] C. Boyd, „Data-parallel computing“ in *ACM SIGGRAPH 2008 classes*, Los Angeles, California, 2008, S. 1, doi: 10.1145/1401132.1401150.

[2] *std::execution::sequenced\_policy, std::execution::parallel\_policy, std::execution::parallel\_unsequenced\_policy, std::execution::unsequenced\_policy - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm/execution\_policy\_tag\_t (Zugriff am: 7. Februar 2022).

[3] M. Voss, R. Asenjo und J. Reinders, „TBB and the Parallel Algorithms of the C++ Standard Template Library“ in *Pro TBB*, M. Voss, R. Asenjo und J. Reinders, Hg., Berkeley, CA: Apress, 2019, S. 109–136, doi: 10.1007/978-1-4842-4398-5\_4.

[4] *Algorithms library - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm (Zugriff am: 7. Dezember 2021).

[5] *std::for\_each - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm/for\_each (Zugriff am: 9. Februar 2022).

[6] *std::for\_each\_n - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm/for\_each\_n (Zugriff am: 9. Februar 2022).

[7] *std::transform - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm/transform (Zugriff am: 9. Februar 2022).

[8] *std::transform\_reduce - cppreference.com.* [Online]. Verfügbar unter: https://en.cppreference.com/w/cpp/algorithm/transform\_reduce (Zugriff am: 11. Februar 2022).

[9] *Programming Guide :: CUDA Toolkit Documentation.* [Online]. Verfügbar unter: https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html (Zugriff am: 1. Februar 2022).

[10] A. R. Brodtkorb, T. R. Hagen und M. L. Sætra, „Graphics processing unit (GPU) programming strategies and trends in GPU computing“, *Journal of Parallel and Distributed Computing*, Jg. 73, Nr. 1, S. 4–13, 2013, doi: 10.1016/j.jpdc.2012.04.003.

[11] NVIDIA Developer Blog, *Maximizing Unified Memory Performance in CUDA | NVIDIA Developer Blog.* [Online]. Verfügbar unter: https://developer.nvidia.com/blog/maximizing-unified-memory-performance-cuda/ (Zugriff am: 14. Februar 2022).

[12] NVIDIA Developer Blog, *Unified Memory for CUDA Beginners | NVIDIA Developer Blog.* [Online]. Verfügbar unter: https://developer.nvidia.com/blog/unified-memory-cuda-beginners/ (Zugriff am: 14. Februar 2022).

[13] NVIDIA Developer, *HPC SDK | NVIDIA.* [Online]. Verfügbar unter: https://developer.nvidia.com/hpc-sdk (Zugriff am: 27. Februar 2022).

[14] *C++ Parallel Algorithms Version 22.2 for ARM, OpenPower, x86.* [Online]. Verfügbar unter: https://docs.nvidia.com/hpc-sdk/compilers/c++-parallel-algorithms/index.html (Zugriff am: 27. Februar 2022).

[15] G. Farnebäck, „Two-Frame Motion Estimation Based on Polynomial Expansion“ in *Scandinavian Conference on Image Analysis*, 2003, S. 363–370, doi: 10.1007/3-540-45103-X\_50.

# Anlagen