

ABSTRACT

Today's Graphic Processing Units (**GPUs**) are not only good for gaming and graphics processing their highly parallel structure is predestined for a range of complex algorithms. They offer a tremendous memory bandwidth and computational power. Contrary to Central Processing Units (**CPUs**), **GPUs** are accelerating quickly and advancing at incredible rates in terms of absolute transistor count. Implementing a massively parallel, unified shader design, its flexibility and programmability makes the **GPU** an attractive platform for general purpose computation. Recent improvements in its programmability, especially high level languages (like C or C++), **GPUs** have attracted developers to exploit the computational power of the hardware for general purpose computing.

Several **GPU** programming interfaces and Application Programming Interfaces (**APIs**) represent a graphics centric programming model to developers that is exported by a device driver and tuned for real time graphics and games. Porting non-graphics applications to graphics hardware means developing against the graphics programming model. Not only the difficulties of the unusual graphics centric programming model but also limitations of the hardware makes development of non-graphics applications a tedious task.

Therefore NVIDIA Corporation developed the Common Unified Device Architecture (**CUDA**) that is a fundamentally new computing architecture that simplifies software development by using the standard C language. Using **CUDA** this thesis will show on the basis of a massively parallel application in which extent **GPUs** are suitable for general purpose computation. Special attention is paid to performance, computational concepts, efficient data structures and program optimization.

The result of this work is the demonstration of feasibility of General Purpose Computation on **GPUs** (**GPGPU**). It will show that **GPUs** are capable of accelerating specific applications by an order of magnitude.

This work will represent a general guideline for suggestions and hints as well as drawbacks and obstacles when porting applications to **GPUs**.

ACKNOWLEDGMENTS

Put acknowledgments here.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1	Aims & Objectives	3
1.2	Initial Time plan	3
2	LITERATURE REVIEW	7
2.1	General Purpose Computing on Graphics Hardware	7
2.2	Parallel Programming & Thinking	8
	Programming Libraries & Languages	9
2.3	Embarrassingly Parallel Algorithms	10
2.3.1	Computations which Map Well to GPUs	10
2.3.2	Raytracing	11
2.3.3	Photon Mapping	11
2.3.4	Multiple Precision Arithmetic	11
2.3.5	High Dynamic Range	12
2.3.6	Genetic Algorithms	12
2.3.7	Chaos Theory	13
2.3.8	Image Processing	13
2.4	Summary & Conclusion	14
3	PARALLEL PROCESSING WITHGPUS	15
3.1	Parallel Architectures	15
3.2	The Tesla Architecture	16
3.3	Common Unified Device Architecture (CUDA)	18
3.4	CUDA Programming Model	19
3.5	A Simple Example	20
3.6	Porting Strategy for GPUs	24
4	FEASIBILITY STUDY	27
4.1	An Overview of Ray Tracing	28
4.2	Ray Casting	28
4.3	Performance tuning	29
4.4	Demystifying the Myths	32
5	MEAN SHIFT	33
5.1	Density Estimation	33
5.2	Kernel Density Estimation	34
5.3	Kernel and their Properties	36
5.4	Mean Shift	38
5.4.1	Density Gradient Estimation	39

5.4.2	Mean Shift Method	41
5.5	Filtering & Segmentation	42
5.5.1	Mean Shift Filtering	43
5.5.2	Mean Shift Segmentation	45
5.6	Summary	47
6	MEAN SHIFT ALGORITHM ANALYSIS & DESIGN	49
6.1	Profiling the Original Code	50
6.2	Amdahl's law	51
6.3	Data & Task Parallelism	53
6.3.1	Task Parallelism	54
6.3.2	Data Parallelism	54
6.4	Data Flow	55
6.5	The next phase	56
6.6	The Algorithm Structure	57
6.7	Programming Model	58
6.8	Thread Batching	58
6.9	Device Memory space	59
6.10	Warps	60
6.11	Program Flow	60
6.12	Summary	61
7	MEAN SHIFT IMPLEMENTATION	63
7.1	Reference Implementation	63
7.2	The Host Part	64
Progress & Result Verification		65
7.3	The Device Part	65
7.4	Summary	67
8	OPTIMISATION STRATEGIES	69
8.1	Test & Benchmark Configuration	69
8.2	Offload Compute Intensive Parts	70
8.3	Global Memory & Coalescing	71
8.4	Division Instruction Optimisation	72
8.5	Execution Configurations	73
8.6	Native Data Types	74
8.7	Avoid Branch Divergence	74
8.8	Shared Memory	75
8.9	Know the Algorithm	75
	Attractor	76

8.10	Unrolling Loops and Multiplications	79
8.11	Summary	80
9	PERFORMANCE & SCALABILITY	81
9.1	Varying the Image-size	81
9.1.1	Linearity	82
9.2	Multiple gpus	83
9.3	Over-clocking the gpu	84
9.3.1	Increasing the Memory Clock	85
9.3.2	Increasing the Core Clock	86
Multiple gpus over-clocked		87
9.4	Final Speedup	88
10	CONCLUSIONS & FURTHER WORK	91
10.1	Further Work	92
Fermi		93
BIBLIOGRAPHY		95
GLOSSARY		103
ACRONYMS		103
A	INTERIM REPORT	109

INTRODUCTION

In recent years GPUs have moved from fixed pipeline processors to fully programmable processors. This evolution has attracted many developers to do GPGPUs. The GPUs have devoted more silicon (transistors) for computing engines rather than for control engines like caches, branch prediction, coherency protocols compared to traditional CPUs. This incredible computing power made algorithms with an high arithmetic density run by an order of magnitude faster than on CPUs. Speedups of $100\times$ – $400\times$ faster than the CPU were stunning but only a few people understand why such speedups are possible and why only a couple of algorithm can attain such speedups.

Many developers in these days are faced with multicore processors and have to implement or extend existing algorithms to take full advantage of the processing power of such cores. The CPU manufacturers are facing fundamental problems when increasing performance only by frequency. In former times higher frequency meant higher performance but a paradigm shift took place now the new way is going multicore which means higher performance. Moore's law says that for every 2 years the number of transistors on a chip will double. The transistors will be dedicated to more cores rather than for bigger chips as in the past. What does it mean to developers? They have to think in parallel, not only for two or four cores but rather for 16 or 32 cores. They have to assure that their code is scaling over many cores over many generations of CPU chips. There are several parallel programming languages and middleware to help developers to program in parallel but a quasi standard has not been established.

There are several programming platforms that are favoured in certain fields respectively system levels. On thread or loop-level there is OpenMP a programming interface for parallelisation on shared-memory systems. Going a level higher to distributed systems that exchange messages during parallel execution there is Message Passing Interface (MPI) which is a quasi standard for message passing. There are further libraries but none of them is really suited for heterogeneous computing systems.

Heterogeneous computing systems are different than common computing systems. For developers it is crucial to understand the underlying

architecture and programming models. They have to find the right tool (**A P I**, middleware, ...) to exploit parallelism. Pitfalls and drawbacks often arise when studying the new system and developers have to adapt and deal with it. Furthermore its not only the use of the right tools its more the parallel thinking which is new to most of them. Parallel thinking and spotting concurrency in algorithms is the key for acceleration. Those aspects have to be considered when trying to do **G P G P U**.

As stated before there are several programming models respectively **A P I S** available for programming the **G P U**. Most of them are based on a graphics **A P I** and are exporting a C++ interface to the developer abstracting the underlying hardware. The biggest problem with such **A P I S** is that there are a lot of them and every has its pros and cons. Thats why the focus will be on **A P I S** directly supplied by the manufacturers of the **G P U S**.

At the time of writing this thesis there were several myths about **G P G P U**. Developers are baffled when they hear they have to implement the algorithm with a graphics **A P I** or it is not possible to do scatter operations to the memory. There exist further myths that discourages developers from computing on **G P U S** that should be demystified.

New technologies have a higher learning curve than existing and its always the question if its feasible and if its even possible to do general purpose computing on **G P U S**.

The starting point will be a review of what has been done followed by a feasibility study which will demystify the myths and cover major obstacles and drawbacks when doing computations on a **G P U**. Further experiments will show and reflect the computational capabilities especially in the view of general purpose programs.

The **G P U** compared to the **C P U** is very powerful and inexpensive. It has a much higher memory bandwidth and computational throughput. Not only that it is fast it is accelerating quickly. Recent improvements made the **G P U** as well flexible and programmable. But there are limitations and difficulties. It is a not only the new programming model or architecture the massively parallel nature of a **G P U** discomforts many developers.

The thesis will therefore start with a background chapter to show the relevance of the topic. After establishing the knowledge the reader will get a first introduction into the architecture and programming model of **G P U S**. Equipped with the knowledge a first algorithm will be ported to

the GPU for experiments and a first hands on.

This preparation will be helpful when getting to the main part where the actual algorithm is presented. An extensive analysis followed by the design and implementation of the algorithm will complete the theoretical part.

The following chapter will then go deeper into hardware specific stuff where the algorithm will be highly optimised for the GPU. The optimisation will be verified in a subsequent chapter where the performance and scalability of the algorithm will be presented.

Last but not least a summary and future work will top the thesis off.

1.1 AIMS & OBJECTIVES

This thesis will cover all the topics to understand the architecture, programming model, drawback and pitfalls when doing GPGPU. The GPU is a highly parallel processor with thousands of threads and a peak performance of several hundred Giga FLoating point Operations Per Seconds (GFLOPS).

By the means of an application which will be ported to the GPU the general workflow will be shown and various procedure models examined. It will be presented that often traditional software engineering principles do not apply to high performance, parallel computing. For this work a NVIDIA GPU will be used together with CUDA that is an extension to C for parallel programming of GPUs.

The remainder of the thesis will give some in depth background to the topic and expose with programming models and the architecture of GPUs. Furthermore the development of a parallel implementation of an algorithm will be examined step by step. In this context software analysis and design principles will be shown that fit to parallel programming. A feasibility study will cover major obstacles and show how to avoid them.

Finally an application for segmenting an image will be implemented and presented in all aspects to the reader.

1.2 INITIAL TIME PLAN

For this thesis the following milestones were defined and listed below. Each milestone will have a short description which tasks have to be

accomplished to complete a mile stone. With every milestone a chapter respectively a section of the master thesis will be written which corresponds to the current carried out milestone. So the development of the milestones goes hand in hand with editing of the thesis.

M1: RELATED WORK IDENTIFIED (2/14/2009) The next step to take is to identify related work, conducting the literature survey. This survey should show the demand of GPGPU and used algorithms. Furthermore it will help to understand the architecture and programming models. After gaining enough knowledge about the field its time to get the hands on the GPU.

M2: ALGORITHMIC VIEW UNDERSTOOD (2/28/2009) Equipped with the knowledge from the latter milestone it is time to understand how the architecture works by examining algorithms already mapped to the GPU. This step is crucial as it shows how algorithms map to and work on the GPU and a good understanding of the environment can help later on for choosing the right algorithm for porting, analysis, design and implementation.

M3: FEASIBILITY STUDY (3/14/2009) The feasibility study will be used for the first steps into GPGPU. A low featured ray tracer will be used for examining what the hurdles and obstacles are when developing for the GPU. The knowledge will be used for choosing the right algorithm.

M4: APPLICATION, ALGORITHM CHOSEN (3/21/2009) After completion of this milestone the algorithm is chosen to be ported to the GPU. So the next milestones will be concentrated only on this algorithm and not more generally.

M5: ANALYSIS, SOLUTION PROPOSED (4/4/2009) Here a solution will be proposed and outstanding issues will be identified. There are several myths about GPGPU which will be stated here as well, because there are affecting the design of the application and cannot be discarded.

M6: EXPERIMENTS (4/14/2009) To exclude all eventualities a set of experiments will be undertaken. This will as well solve all myths.

M7: DESIGN, APPLICATION PORTED (5/14/2009) This milestone will cover the design of the application with all constraints, obstacles and hurdles that are involved with GPGPU. In this task the application will be implemented and run on the GPU. A good implementation will need several iterations where several aspects will be tuned: coalesced memory access, load balancing and many more.

M8: DISCUSSION (5/22/2009) Another important part is the discussion of the results and methodologies applied which will be done in this milestone.

M9: MASTER THESIS EDITED (5/30/2009) Here the thesis will be recapped and briefly outlined. Conclusions will be drawn and future work presented. A critical summary and conclusion will be provided as well.

LITERATURE REVIEW

There is a lot of literature that is describing and showing GPGPU. The literature of old algorithms ported to the GPU focuses how to exploit a graphics API to do general purpose computing. With the rise of Software Development Kits (SDKs) specially for GPGPU more and more literature emerges with various algorithms ported to the GPU.

The aim of this chapter is to provide, through selective reference to some of the literature, a clearer understanding why GPGPU is a topic nowadays, the miscellaneous programming environments and lastly the wide field of GPU computing. Several implementations of algorithms on the GPU are presented and evaluated. The chapter is divided in several parts. The subject of the first part is a general view about the move to parallel processing and obstacles and problems that arise when doing parallel programming. Building up on the first part, several programming environments that enable GPGPU are presented and discussed. Succeeding a part that describes the algorithms that are suited for computing on the GPU. The remainder of this chapters shows some already implemented algorithms to the GPU and shows the wide field of application.

2.1 GENERAL PURPOSE COMPUTING ON GRAPHICS HARDWARE

The first thing to answer is why should someone do GPGPU anyway. For the most people CPUs are just enough. They do not demand on high computational power and on a high bandwidth. Still there is paradigm shift taking place and this can not be neglected. As stated in [40] and in [41] CPU manufacturers are facing problems which they cannot overcome just by increasing the frequency. The problems are known as *The Memory Wall*[37], *The Frequency Wall* and *The Power Wall*. The only way seen by CPU makers is currently to go multicore. Intel e.g. went multicore with their new *Core Micro-architecture* for consumer products and even a GPU replacement, *Larrabee* [47]. International Business Machines Corporation (IBM), Toshiba and Sony developed the *Cell Broadband Architecture* a 9 core chip [43]. Sun Microsystems developed the *Niagara* CPU a multi-core general purpose processor. It has eight

in-order cores, each of them capable of executing four simultaneous threads [26]. A further processor from Sun is the *Sun Rock* a 16-core 32-thread plus 32-scout-thread CPU developed for high throughput and high single-thread performance [54].

The GPUs went years ago to multicore and multithreading compared to CPUs. The GPUs are maybe the kind of processor where CPUs are heading to in terms of multithreading and raw performance. Forecasts project that every two years the amount of cores can double. The multi-core approach may be the answer to the problems stated above but this yields to another thing the *parallel programming problem* [42]. User will only benefit from this growth if software can make use of all the cores and software can only benefit from it if developers can leverage concurrency in software. Developers have to look at their code to identify computational intensive operations and identify how those operations can benefit from concurrency. Sutter[52] stated that “The free lunch is over”, developers cannot count only anymore on CPU throughput they have to get their hands on concurrent programming requirements, pitfalls, styles and idioms [52].

2.2 PARALLEL PROGRAMMING & THINKING

Many developers learned about the single-threaded von Neumann model and are not familiar with parallel code which is subject to errors such as deadlocks and live-locks, race conditions and many more. Parallel programming is difficult and there are several paradigms to make life easier for developers.

On of these is the data-parallel paradigm. Where one is not trying to assign different subtasks to separate cores rather assigning an individual data element to a separate core for processing [7]. The three-dimensional (3D) rendering, an embarrassingly data-parallel problem, has driven the GPU evolution which makes the GPU a perfect target for data-parallel code. There are several fine-grained or data-parallel programming environments that leverage the GPU and other parallel architectures for general purpose computing.

Programming Libraries & Languages

Beyond graphics APIs like Cg, High Level Shading Language (HLSL), OpenGL or DirectX several libraries and languages were introduced to make GPGPU possible. Many of them are based on the graphic APIs but have backends to many other libraries.

They have one thing in common, they enable stream computing on the GPU. The first language is Brook[6] for GPU it extends American National Standards Institute (ANSI) C with concepts of stream programming. Brook stream are nothing else than arrays on which the computation is done in parallel. The program that is doing some operations on streams is called kernel. At run-time the kernels and streams are automatically mapped to the fragment processor and to texture memory.

A similar concept follows Sh [35]. Sh is embedded in C++ and allows to program GPUs for graphical and general purpose computations. The research was commercialised and RapidMind Inc. was formed. RapidMind delivered with its programming environment not only support for GPUs but also for other massively parallel processors like the Cell. Recently Intel acquired RapidMind to merge it into there Intel Ct Technology.

For scientific visualisation there is Scout [36], it allows run-time mapping of mathematical operations over data sets to the GPU, for general purpose computing it is rather limited as it focus is on visualisation of data sets.

Another library is Glift [29] which is a template library for a wide range of data structures. It operates with C++, Cg and OpenGL development environments.

All the above libraries and languages are all based on a graphics API. The functions are all graphics centric and not all are suited for GPGPU. Since these functions have to accommodate to the graphics pipeline not all hardware features are directly exposed to the developer they are encapsulated and additionally the graphics pipeline puts up some constraints to the programming flow [40].

Therefore ATI Technologies Inc. (ATI) and NVIDIA Corporation (NVIDIA) developed programming environments specially for GPGPU. From ATI came Close To Metal (CTM) a low level programming interface exposing previously unavailable low level functionality , direct access to the native instruction set and memory, for GPGPU. ATI stand point

was that the community should contribute the high level API and tools. NVIDIA has gone a different way, they offered a complete programming environment (SDK) for GPGPU namely known as CUDA [5] with many tools for programming and debugging.

Recently ATI released as well a full featured programming environment, the Stream SDK. Nevertheless as of writing of the thesis this SDK was not available and hence not considered.

The focus of this work will be on CUDA. CUDA is one of the environments which is not based on a graphics library and officially released by NVIDIA for their GPUs. CUDA is a minimal extension to C and C++ programming languages. The technique employed by CUDA is Single Program Multiple Data (SPMD). Tasks are split up and run simultaneously on multiple threads (cores) with different input [39].

2.3 EMBARRASSINGLY PARALLEL ALGORITHMS

Before even digging into the wide field of algorithms and difficult problems in high performance computing one has to understand the hardware architecture of the GPUs to make a decision if an algorithm can be mapped on GPUs. A pretty good overview over the NVIDIA GPU gives the NVIDIA CUDA programming guide [49].

A little bit outdated but still of interest is the GeForce 6 Series GPU Architecture [25] which gives an overview how the GPU fits into the whole system, what a fragment processor, vertex processor or what textures are. To have a even deeper look into GPUs the article [17] is highly recommended.

2.3.1 *Computations which Map Well to GPUs*

It is important to understand that GPUs are good at running computer graphics and algorithms which mimic or have the attributes of computer graphics in terms of data parallelism and data independence. Not only that similar computations are applied to streams of many data elements (vertices, fragments, ...) but also the computation of each element is completely or almost completely independent [22]. Such types of algorithms are often called embarrassingly parallel algorithms where subtasks rarely or never communicate to each other.

Another important fact for a algorithm is the *arithmetic intensity*. The arithmetic intensity is the ratio of computation to bandwidth or formally:

$$\text{arithmetic intensity} = \text{operations} / \text{words transferred}.$$

This fact is important because the increase of computational throughput is faster than the memory throughput which leads to the problem known as *The Memory Wall*. The GPU memory systems are architected to deliver high bandwidth, rather than low-latency, data access. As such computations that benefit most of the GPUs have a high arithmetic intensity [22]. The next sections will represent some algorithms which are ported completely or to some extent to the GPUs.

2.3.2 Raytracing

Ray Tracing [2] is an embarrassingly parallel algorithm which could fit well to GPUs. A work about raytracing on massively parallel computers [28] is already done. There were several implementations already done for the GPU e.g.

2.3.3 Photon Mapping

Raytracing has a local illumination model. To generate more realistic effects like caustics, diffuse / glossy indirect illumination and more a more sophisticated model has to be used. Global illumination like *Photon Mapping* [24] can create all the effects that raytracing cannot. There are implementations of photon mapping on GPUs [45] which are developed with graphics apis and not with CUDA. Anyhow since photon mapping is heavily using a kd-tree it would be a major effort to develop an efficient data structure which has the same functionality as a kd-tree. Nevertheless *Photon Mapping* is a embarrassingly parallel algorithm that could fit very well to the GPU.

2.3.4 Multiple Precision Arithmetic

Another interesting field which demands high computational power is number theory. The most common/known application is asymmetric cryptography. To decipher messages that are cyphered with an asymmetric algorithm one needs superior computational power. An overview

over common algorithms gives [27]. All algorithms have one thing in common they need a multiple precision library to represent numbers with 200 and more digits. A good overview gives [53].

The idea could be to implement some of the factoring algorithms to the GPU. The only thing needed is the multiple precision library. In [27] the GNU Multiple Precision (GMP) library was used to implement the algorithms. So the multiple precision arithmetic is another candidate for algorithms that can run on the GPU. There are several implementations but they are rather limited and implement a specific algorithm with multiple precision but not a generic multiple precision library [59, 1].

2.3.5 High Dynamic Range

Image processing is always a candidate for embarrassingly parallel algorithms. A pretty new algorithm to enhance images is tone mapping [34]. Tone mapping is the compression of dynamics in High Dynamic Range (HDR) pictures. There are several algorithms for tone mapping: Mantiuk [33], Reinhard [46], Durand [16], Fattal [18] and many more. All of this algorithms are present in the pfs-tools library written by Krawczyk [34]. As this document was written Krawczyk was already implementing a part of the pfs-tools on GPU but not publishing it.

2.3.6 Genetic Algorithms

Parallel genetic algorithms are usually implemented on parallel machines but fine-grained parallel genetic algorithms can be mapped to GPUs [57]. In [30] its shown what kind of genetic algorithm map well to GPUs and how the work and communication is handled. It is pretty easy to implement simple genetic algorithms on the GPU but with increasing complexity one has to consider many more things: load balancing, communication pattern, dynamic memory allocation, resolving of recursion Another paper [56] compares genetic algorithms implemented on CPUs and GPUs and shows that the latter is much more effective than the former. All genetic algorithms have one thing in common the core algorithm. Once effectively implemented on the GPU the core algorithm is extended with definitions like the population, selection, recombination and mutation to solve a specific problem. The skill here is to choose

the right definitions and not more the efficient implementation of the core algorithm.

2.3.7 Chaos Theory

Another interesting field is chaos theory. Especially the visualisation of chaos. The maybe most famous visualisation of chaos is *The Fractal Flames Algorithm*. Fractal flames are a member of iterated function system class of fractals created by Scott Draves [14]. He uses a rather complicated set of functions in the system to generate stunning visualisation of the iterative process of the system. The next release will have support for the GPU. Thats why it was firstly discarded but the research about chaos led to other interesting papers respectively books.

There is no need to use approx. 21 function for a system to generate visually appealing pictures of chaos. Sprott showed in [51] that even with very simple functions one can create patterns in chaos. This patterns are called *Strange Attractors* and are the visualisation of chaotic behaviour. There are created by iterating a simple equation some million times.

Pickover shows in his book [44] how to create patterns from a variety of sources. He shows how to create nice looking patterns from fourier analysis, acoustic, chemistry and many more. Anyhow all of these equations or differential systems have on thing in common no matter how complicated the system is the core algorithm is to iterate a specific equation with correct input numbers to create chaos. The algorithm is heavily computational bound which makes it a good candidate for porting to the GPU.

2.3.8 Image Processing

Many image processing algorithms are embarrassingly parallel as one can calculate filters on individual pixels without considering the neighbouring pixels. Bruyns and Feldman showed by example how image processing is done on the GPU by a canonical example, limb finding [4]. Another paper shows how to do convolution, diffusion, wavelet or tone mapping [38]. Furthermore there are several universities that deal with GPU image processing and publishing algorithms and reports.

2.4 SUMMARY & CONCLUSION

The GPU is intruding into more and more fields in scientific, mathematical, financial and many more fields. This is happening because people are demanding more and more computational power and the GPU can satisfy this demand. As it was shown in the previous chapters the GPU is used in many ways and there is a high demand on exploiting the GPU even farther. With the development of new native programming environments (CUDA, Stream SDK) even more developer will be attracted to use this new architecture.

But all of this reports showed that with the old tools the programming is and was error prone and difficult. Only some specialist that understood the graphics part in extensive and the algorithm to be ported could implement a efficient solution.

This is where this thesis will attach. Taking the new programming environments and trying to port a sample application to the GPU.

PARALLEL PROCESSING WITH GPUS

In recent years more and more parallel hardware appeared on the computing market with different architecture specialised for different problems. One of the recent and most discussed architectures is the GPU. The major player in the field are ATI and the NVIDIA corporation that both offer SDKs and GPUs for general purpose computing. Beside GPUs there are a lot other companies that offer parallel hardware architectures listed in Chapter 2. Following a brief introduction to parallel architectures important to specify and identify the sort of parallelism offered by GPUs and other hardware.

3.1 PARALLEL ARCHITECTURES

There are several ways how operations in a system can be executed in parallel.

- Single Instruction Single Data (**SISD**) This is the common von Neumann model, which is a system which exploits no parallelism.
- Single Instruction Multiple Data (**SIMD**) operations inside a functional unit where one operation is simultaneously applied to a vector of 2–16 components. Many modern CPUs have now SIMD units but every manufacturer is naming them differently (e.g. Altivec, Vector Multimedia eXtension (**VMX**), Matrix Math Extensions (**MMX**), Streaming SIMD Extensions (**SSE**)).
- Multiple Instruction Multiple Data (**MIMD**) threads, cores, processors each executing on its own data. Many parallel architectures can be mapped onto the MIMD architecture.

There is as well the Multiple Instruction Single Data (**MISD**) classification which is a uncommon architecture but can be used e.g. in fault-tolerance systems where different units generate a result on the same data and they must all agree on the result. For common parallel architectures its not used. By far the most used classification for parallel systems is MIMD that can be further divided into two more classifications.

- **SPMD** threads within a core where one program is running on the core and different threads are executing with different data. A combination

of the SIMD and SPMD model is used on modern GPUs.

- Multiple Programs Multiple Data (MPMD) threads executing a different program on different data. The MPMD threads reside either within a single core (hardware threads, hyper-threading) or in distinct cores of a processor or lastly in different processors.

The above classification is based upon the number of concurrent instruction and data streams available in a system. The classification comes from Flynn in [19]. Today there is still no satisfactory characterisation of the different types of parallel systems. Flynn's taxonomy was a first step, today there are much more characterisations that refine the previous mentioned.

To characterise the GPU one needs two classifications. First of all in each GPU there are threads that are grouped together which are executing one instruction on different data. These SIMD groups are further grouped to a SPMD system where all of these, there can be many of them, are executing the same program. On the basis of NVIDIA's Tesla architecture a more detailed look at a GPU will be given in the next sections.

3.2 THE TESLA ARCHITECTURE

The Tesla architecture announced 2007 and developed by NVIDIA is the first GPU highly specialised for raster operations and more important for general purpose computing. Formerly GPUs had fixed-function pipelines and separate processing units with no ability for programmability to the vertex and fragment stages of the pipeline. In recent years manufacturers of GPUs added more and more programmability to the different stages of the pipeline and at the same time general purpose computing capabilities [32]. Furthermore manufacturers introduced the Unified Shader Model (USM) that unifies the processing units allowing for better utilisation of GPU resources. The resources needed by different shaders varies greatly and the unified design can overcome this issue by balancing the load among vertex, fragment and geometry functionality [8].

The Figure 3.1 shows the Tesla architectures. As mentioned above the new GPU architectures are a radical departure from traditional GPU design. The Tesla 8 Series has 16 multiprocessors. Each multiprocessor

is composed of 8 streaming processors, 128 processors in total. Each streaming multiprocessor has 16 Kilobyte (KB) of shared memory a L1 cache attached and has access to a texture unit. A streaming processor consists of a scalar Arithmetic Logic Unit (ALU) and performs floating point operations. 32 streaming processor build up a SIMD unit in NVIDIA terms called warp, in that one instruction is executed. The Tesla C870 has 1.5 Gigabyte (GB) Graphics Double Data Rate v3 (GDDR3) graphics memory, 519 GFLOPS of peak performance and 77 GB/s peak memory bandwidth. Many massively data-parallel algorithms can be run sufficiently on this specialised architecture [8]. Programming for this architecture is done with CUDA the C language extension which will be covered in the next section.

THE MAINSTREAM GPU There are two solutions for GPU computing from NVIDIA. Firstly the *Tesla 8 Series* which is specifically designed for high-performance computing. The *Tesla 8 Series* is available as an Peripheral Component Interconnect Express (PCI-E) add-in card, desk-side computing system and as a one rack unit (1U) computing system. All this systems have no display port and hence they can exclusively be used as a computing nodes.

The second solution are the mainstream *GeForce 8 Series* card that is mainly used for gaming. The architecture of these both solutions is the same (USM) they differentiate only in the magnitude of some performance characteristics (memory size, peak memory bandwidth, peak GFLOP). The advantage of a mainstream GPU is that the visualisation can be done on the same card and the low price.

For the purpose of this thesis a *GeForce 8800 GTS 512* card will be used. The main performance characteristics are shown in Table 3.1 and Table 3.2. For reference the *Tesla C870* card is shown as well.

Card	Memory Size [MB]	Bandwidth [GB/s]	Clock [MHz]
Tesla C870	1500	77	1600
8800 GTS	512	62.1	1940

TABLE 3.1 Memory performance characteristics for Tesla and GeForce

The most significant difference between the two cards is the peak

bandwidth. There are several algorithms which rely on a high bandwidth which would increase the run time on as 8800 GTS card even if it has a higher peak floating point operations per second as seen in Table 3.2.

Card	No. of Stream Processors	GFLOPS	Clock [MHz]
Tesla C870	128	519	1536
8800 GTS	128	624	1625

TABLE 3.2 Computing performance characteristics for Tesla and GeForce

Every card of the *GeForce 8 Series* can be programmed with CUDA an API for GPU computing. What CUDA is and how to use it will be explained in the next section.

3.3 COMMON UNIFIED DEVICE ARCHITECTURE (CUDA)

In 2007, NVIDIA introduced an extended ANSI C programming model and software environment the compute unified device architecture. The reason why CUDA was born is that parallelism is increasing rapidly with Moore's law¹ and the challenge is to develop parallel application software that scales transparently with the number of processor cores. The main goals when CUDA was developed were that it scales to hundreds of cores, thousands of parallel threads and it allows heterogeneous computing (CPU + GPU). All this considerations led to the result that CUDA runs on any number of processors without recompiling an the parallelism applies to both CPUs and GPUs [5].

CUDA is C with minimal extensions and defines a programming and memory model. There are three key abstractions in CUDA, a hierarchy of thread groups, shared memories and barrier synchronisation [49] that are exposed to the developer. CUDA uses extensive multithreading where threads express fine-grained instruction, data and thread parallelism that are grouped into thread blocks which express coarse grained data and task parallelism. The developer has to rethink about his algorithms to be aggressively parallel. The problem has to be split into independent coarse sub-problems and at finer level into fine-grained sub-problems that can be solved cooperatively [49].

¹ Processor count is doubling every 18 - 24 months

CUDA has been accepted by many developers which can be seen by the huge amount of already developed software and contributions to the CUDA zone².

A brief look shows the CUDA computing sweet spots[5]

- High arithmetic intensity (Dense linear algebra, PDEs, n-body, finite difference, ...)
- High bandwidth (Sorting, database, sequencing, virus scanning, genomics, ...)
- Visual computing: (Graphics, image processing, tomography, machine vision, ...)
- Computational modelling, science, engineering, finance, ...

This is just a small snapshot of algorithms that can be used with CUDA on GPUs. For a more extensive list and speedups compared to high-end CPUs see the webpage of NVIDIA CUDA zone.

3.4 CUDA PROGRAMMING MODEL

The CUDA programming model exposes the graphics processor as a highly multithreaded coprocessor. The GPU is viewed as a compute device to the host that has its own device memory and runs many threads in parallel.

Applications are accelerated by executing data parallel portions of the algorithm on the GPU as *kernels* which run in parallel on many threads. There are some major differences between CPU and GPU threads. A GPU needs thousands of threads for full efficiency where CPUs only need a few of them. GPU threads have a very little creation overhead and are extremely lightweight compared to CPU threads.

THREAD BATCHING A kernel is executed as a grid of thread blocks where data memory space is shared by all threads. A thread block is a batch of threads that can cooperate with each other by synchronising their execution³ and efficiently share data through the low latency shared memory. Two threads from two different blocks can cooperate with

² <http://www.nvidia.com/cuda>

³ For hazard free shared memory access

atomic functions through the global memory. The identification of a thread is accomplished through block and thread ids which are assigned to each thread at creation time.

BLOCK AND THREAD IDS Every thread and block has a unique id. As a result of this each thread can decide what data to work on. For every block there is a assigned id in one-dimensional (1D) or two-dimensional (2D) layout. Thread ids can be accessed either with 1D, 2D or 3D coordinates similar to multidimensional arrays. It simplifies memory addressing when processing multidimensional data(e.g image processing, matrix multiplication or solving partial differential equations on volumes). The data can reside in several levels of the device memory.

DEVICE MEMORY SPACE The memory space is a hierarchy of several memory types that can be accessed per thread, block, grid and the host. The threads have access to all memory levels beginning with the read-write (RW) registers, local, shared, global, read-only (RO) texture and constant memory. The grids have only access to global, constant and texture memory. Whereas the CPU can RW global, constant and texture memories.

Global, constant and texture memory have long latency accesses. They reside off-chip where registers local⁴ and shared memory resides on-chip. The global, constant and texture memory are mainly used for communication of RW data between host and device where the contents is visible to all threads. As mentioned above texture and constant memory can be written by the host where constants and data are initialised.

3.5 A SIMPLE EXAMPLE

This simple example will show the structure of an CUDA program. The executing kernel will do some easy calculations on the data provided, load the data into shared memory and write back the results to the global memory.

A CUDA program has a specific structure where the major parts are described in this paragraph. The first thing to do is to initialise the device and some auxiliary variables. The Listing 3.1 shows the initialisation.

⁴ Not true for older GPUs chips where local data is spilled out to global memory

```

1 CUT_DEVICE_INIT(argc, argv);
2
3 uint32_t num_threads = 32;
4 uint32_t mem_size = sizeof(float) * num_threads;

```

LISTING 3.1 Hardware initialisation

Since the GPU is attached to the PCI-E bus the host has no direct access to the global, constant and texture memory and has to transfer the data back and forth with the Direct Memory Access (DMA) engine of the device. This is accomplished through the CUDA API calls that initiate the transfer. Before any transfer can be done one has to allocate memory on the host and on the device for input and output data. This is shown in listing Listing 3.2.

```

1 // allocate host memory
2 float* h_idata = (float*) malloc(mem_size);
3 // allocate device memory
4 float* d_idata; cudaMalloc((void**) &d_idata, mem_size);
5 // allocate device memory for result
6 float* d_odata; cudaMalloc((void**) &d_odata, mem_size);
7 // allocate mem for the result on host side
8 float* h_odata = (float*) malloc( mem_size);
9 // copy host memory to device
10 cudaMemcpy(d_idata, h_idata, mem_size, cudaMemcpyHostToDevice);

```

LISTING 3.2 Data transfer of data

After setting up the input data the setup execution parameters are defined that are used to startup the kernel. The `grid(1, 1, 1)` statement defines a multi-dimensional array of grids $x = 1, y = 1, z = 1$ whereas the `threads(num_threads, 1, 1)` defines a multi-dimensional array of threads $x = num_threads, y = 1, z = 1$ which are actually 1D arrays. Listing 3.3 shows the call of the kernel with its input and output data.

```

1 // setup execution parameters
2 dim3 grid( 1, 1, 1);
3 dim3 threads(num_threads, 1, 1);
4

```

```

5 // execute the kernel
6 kernel<<< grid, threads, mem_size >>>(d_idata, d_odata);

```

LISTING 3.3 Execution of the Kernel

Attention must be paid to the both statements that define the array for the grid and the thread block. The size of the array or the amount of threads dependent on three factors, (1) register usage, (2) shared and (3) local memory usage per thread. Accumulated per grid they are not allowed to pass a certain size. Calculations how the existing resources are used can easily be calculated with the *CUDA occupancy calculator*. The performance of the application depends on how good/bad the existing resources are exploited.

If everything went well the host can copy the data from device memory to host memory and check, visualise or store the calculated values. Listing 3.4 shows the last steps before exiting the program.

```

1 // check if kernel execution generated and error
2 CUT_CHECK_ERROR("Kernel execution failed");
3
4 // copy result from device to host
5 cudaMemcpy(h_odata, d_odata, sizeof( float ) * num_threads,
6           cudaMemcpyDeviceToHost);
7
8 // cleanup memory free(x), free(y), free(z) ...
9 CUT_EXIT(argc, argv);

```

LISTING 3.4 Retrieving of the Results

The previous listings showed the host code and how to launch a kernel on the device. The Listing 3.5 shows the device code portion. There are several qualifiers that define which function is compiled for which processing unit. The `__global__` qualifier specifies that this function is run on the device and hence compiled for the GPU where the `__host__` qualifier specifies that this function is only run on the host and not on the device. There are more function specifiers that can be looked up in [49].

For data there are as well qualifiers where one can specify where the data is located, either in constant, global or shared memory. In Listing 3.5

the `__shared__` qualifier is used. The device will use the shared memory to preload the data for faster access.

```

1 #include <stdio.h>
2
3 #define SDATA(index) CUT_BANK_CHECKER(sdata, index)
4 // Simple test kernel for device functionality
5 __global__ void kernel( float* g_idata, float* g_odata)
6 {
7     // shared memory, the size is determined by the host application
8     extern __shared__ float sdata[];
9
10    // access thread id
11    const unsigned int tid = threadIdx.x;
12    // access number of threads in this block
13    const unsigned int num_threads = blockDim.x;
14
15    // read in input data from global memory
16    // use the bank checker macro to check for bank conflicts
17    // during host emulation
18    SDATA(tid) = g_idata[tid];
19    __syncthreads();
20
21    // perform some computations
22    SDATA(tid) = (float) num_threads * SDATA( tid);
23    __syncthreads();
24
25    // write data to global memory
26    g_odata[tid] = SDATA(tid);
27 }
```

LISTING 3.5 CUDA device code

After loading the data the kernel just multiplies the thread-id with the number of threads and saves the result back to global memory where the host can pick up the result.

3.6 PORTING STRATEGY FOR GPUS

This section is a high-level overview of the porting process to not only the GPU it can be rather seen as general porting process to parallel hardware. It includes strategies that help through the porting process. It breaks down the porting process into a series of discrete steps.

The steps included here are not the only path through the porting process because the porting process is flexible and has to be adapted slightly for every hardware.

In general, one will follow these steps when porting an algorithm to the GPU

- Algorithm complexity study. Profile the code, find the most compute intensive parts analyse the parts of the algorithm which could be parallelised.
- Data layout/locality and Data flow analysis. Look for independent task or independent data to exploit task or data parallelism.
- Experimental partitioning and mapping of the algorithm and program structure to the architecture.
- Develop CPU control, CPU scalar/multicore code
- Develop CPU control, partitioned GPU scalar code (Communication, synchronisation, latency handling)
- Transform GPU scalar code to GPU threaded code, multi GPU code. Exploit all available cores and devices on the system.
- Re-balance the computation / data movement.
- Other optimisation considerations (load balancing, bottlenecks...)

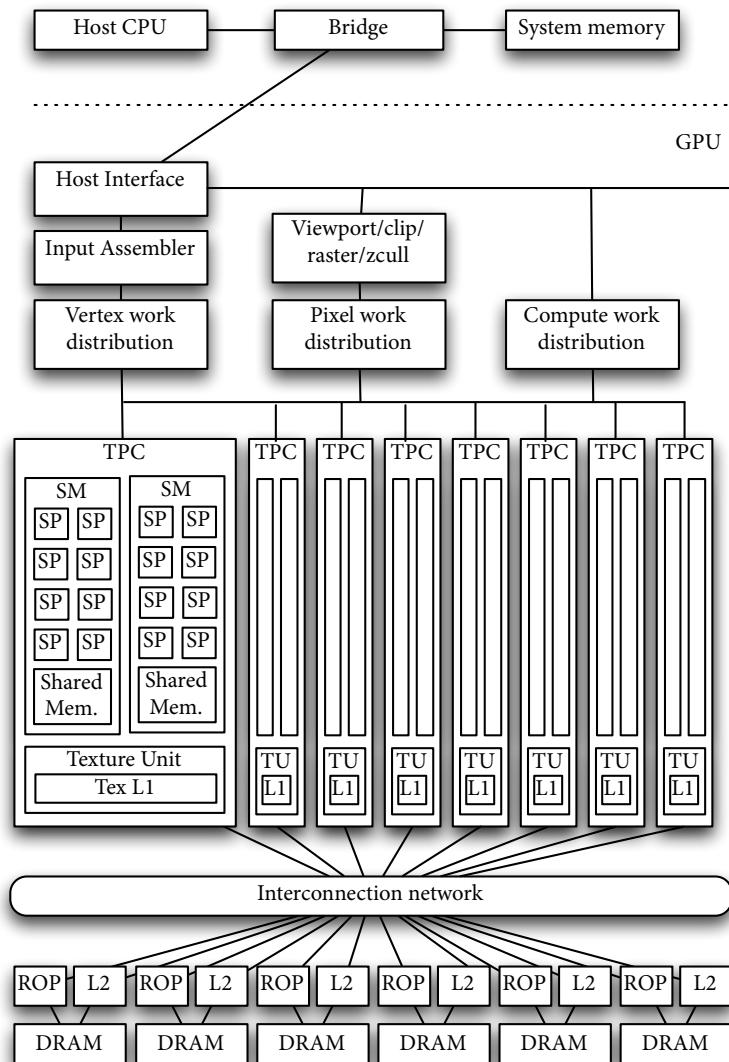


FIGURE 3.1 Tesla Architecture

FEASIBILITY STUDY

Entering a new field in computing without enough experience on the platform it is always advisable to get a better knowledge about the hardware and software before designing anything without proper knowledge. A feasibility study can give an overview if a project is technically feasible and a way to gain some knowledge about the complete eco system. Furthermore the study can answer first questions and help demystify myths.

There are several myths about GPGPU which will be examined and explained Listed below some myths in no particular order.

1. GPU programs are written with a graphics API and layered on top of graphics
2. GPUs can only do a gather and no scatter memory access
3. GPUs are power-inefficient
4. GPUs don't do real floating point math
5. GPUs in average one can only exploit about 10% of the peak performance for general purpose computing
6. GPUs are very wide SIMD machines on which branching is impossible, with 4-wide vector registers

The most interesting myth here is that in average one can only exploit about 10% of the peak performance for general purpose computing, which in turn would lead to that GPUs are power-inefficient in average. Another thing to keep an eye on is the statement that GPUs do not do real floating point math which would exclude many communities (High Performance Computing (HPC), Physics, Astronomy, ...) from spending time in doing GPGPU.

Therefore the first step before choosing an algorithm or put further energy into research of GPUs is to make the feasibility study. The study will examine the myths and show solutions to the problems respectively statements. Furthermore it will show whether the technology, software eco system exists for building applications on GPUs and how difficult it will be to build.

*Many HPC,
Physics,
Astronomy,...
benchmarks and
algorithms rely on
floating point
math*

By the means of a ray tracer the study will show which steps have to be undertaken to gain high performance with an parallel algorithm and show which architectural points have to be considered when designing an application.

4.1 AN OVERVIEW OF RAY TRACING

Ray tracing is a technique for realistic image synthesis. Ray tracing as the name says traces light rays generated from an imaginary camera to their points of origin. Ray tracing is based upon a physical, mathematical model behind light, which facilitates to render photo realistic images.

Ray tracing can be seen as an extension to ray casting. Therefore its easier to understand ray tracing if the base concept of ray casting are understood. The next section will introduce ray casting¹.

4.2 RAY CASTING

Ray casting was first introduced by Appel [2] he developed some techniques for a shading machine for rendering of solids.

First of all it is important to understand the concept of rays. A ray is the path of a particle of light (photon) extending from the eye into the scene [20]. The path is a thin, straight line used to model a beam of light. Each ray can be seen as a *feeler* that reaches the scene and finds out which objects are visible and what colour the object has at a specific point. Rays are the fundamental element of any ray tracer.

The representation of light on the screen is organised in so called pixels. A pixel is a point sample not a little geometric square. This misconception is widespread and it is an issue that strikes right at the root of correct image computing and the ability to correctly integrate the discrete and the continuous [50].

The colour of a given pixel is the colour of the light that passes from the object, through the associated pixel into the eye [23].

For each pixel on the screen a ray is cast from the eye through the pixel into the virtual world. Then for each object it is checked if the ray intersects any of them. If there are several objects in a scene intersected by the ray the shortest distance to the intersection point is the one that

*The ray scans or
rasters the scene
that is why a ray
is often called a
feeler*

¹ For an in depth description of ray tracing on a parallel machine see Krnjajic [28]

is visible to the eye. All other intersection are behind the nearest object and not visible respectively occluded by the visible object. The colour at that point is the accumulated contribution of intensities radiated from all light sources. This colour is given to the pixel through which the ray passed. Ray casting does not consider light reflected or transmitted by other objects in contrast to ray tracing.

Ray casters and ray tracers spent most of their time calculating intersections of rays with different objects. Whitted [55] estimates that anywhere from 75% to over 95% of rendering time is spent in intersection tests. For example an image with 640 pixel width and 480 pixels height, for a total of 307200 pixels, with a medium complex scene with 100 objects results in 30720000 intersection tests. There are well documented ray tracing accelerating techniques not only to decrease the number of intersection tests per ray but also to decrease the number of rays.

In the study the standard ray casting algorithm will be used only, which means that every ray will be intersected with every object to have a high arithmetic density and not to be limited by the memory bandwidth.

Further experiments to extend the basic ray casting algorithm to ray tracing will be discarded this will only add more complexity to the algorithm and will not help in solving the essential problems.

4.3 PERFORMANCE TUNING

After the initial port of the ray tracer to CUDA, the first goal was to make it run on the GPU, it was time to fine tune the run configuration. For any CUDA application it is crucial to take a look at register, shared memory and local memory usage. The performance of the application depends on how good/bad the existing resources are exploited.

The examination of the ray traced showed that the initial port used 48 registers, 28 bytes shared memory and 96 byte constant memory per thread. As shown in Section 3.5 the number of launched threads, blocks and grids is dependent on the three factors mentioned above.

Each Streaming Multiprocessor (SM) has 8192 registers which means one could have $8192/48 = 170$ threads to fully exploit the resources 8 blocks should be launched so $170/8 = 21$ threads per block which is by far too low. It is easily spotted that the application is limited by registers per SM. These calculations can easily be done with the CUDA *occupancy*

It is always good to have a thread count which is a multiple of 32. A warp consists of 32 threads and the scheduler can easier schedule the threads

calculator supplied with the SDK.

After fiddling around with compiler switches especially the switch that limits the used registers `-maxrregcount=x` the application used only 14 register which meant we could have a block-size of 8, $8 \times 8 = 64$ threads per block. When reducing the amount of registers used, one has to consider that registers which are additionally needed are allocated from the local memory rather than the register file. Local memory is located in the global memory which has a high latency (200–400 cycles) and registers often referenced have a big impact on performance. The application can become memory bound.

The Table 4.1 shows the execution configuration used for the sample runs.

blockSize.x	blockSize.y	gridSize.x	gridSize.y	Regs./Thread
8	8	96	96	14

TABLE 4.1 Execution configuration.

To summarise, just by compiling the application and fiddling with the run parameters one can achieve, in this case, a speedup of about 8 times. See Table 4.2 for several runs varying object count and image size.

Image	Objects	CPU GFlops	GPU GFlops	Speedup
768	338	0.37	3.17	8.54
768	450	0.37	3.22	8.6
768	840	0.37	3.23	8.74
512	338	0.37	3.19	8.53
512	450	0.37	3.18	8.5
512	840	0.37	3.18	8.64
256	338	0.37	3.03	8.09
256	450	0.37	3.03	8.44
256	840	0.37	2.95	7.96

TABLE 4.2 Comparison between CPU and GPU

As it can be seen in Table 4.2 the application runs 8 times faster but when looking at the GFLOPs values there are far beyond that what a

GPU could achieve². The application is only using 0.5% of the peak performance. All performance values are gathered with the NVIDIA Performance Profiler available as well with the SDK.

The reasons for this ineffective usage of processing power can be seen in 4.3a and Table 4.4. The application is heavily memory bound. The consequence of reducing the register usage is heavy access to local memory, e.g. there are over 328000000 stores issued to the memory. Furthermore the code is using many branches which is actually reflected by the value 22578032.

Load	Store	Sum	Divergent
63294179	328180560	22578032	26834
(A) Local memory		(B) Branches	

TABLE 4.3 Local memory and branches

Another point is the access to global memory. The GPU is able to do coalesced reads and writes when possible otherwise the access is serialised. The application was not ported with coalesced memory reads and writes in mind and hence the application is loading almost everything incoherent. See Table 4.4 for the values.

Load Incoherent	Load Coherent	Store Incoherent	Store Coherent
1115657863	67961	2506752	0

TABLE 4.4 Global memory loads and stores.

The ray tracer could be extended or optimised in many ways but it will be leaved as is. The main point of the feasibility study was to get a feeling for developing on GPUs and to spot major drawbacks and obstacles. Keeping the lessons learned here in mind the development of the main application will be easier. The following section will demystify some but not all myths stated in the beginning.

² The used GPU is a G92 chip with peak 620 GFLOPS

4.4 DEMYSTIFYING THE MYTHS

There were several myths about GPGPU see Chapter 4 that are simply wrong or need to be proven. Beginning with Item 1 this statement is simply wrong. Using CUDA or the ATI SDK *Close To Metal* no developer is anymore forced to use graphics APIs. Furthermore the abandoned graphics API for GPGPUs makes it possible to gather and scatter to the memory, which resolves myth Item 2.

The GPUs are able to do real floating point math. The myth Item 4 comes from a time where floats were only 16 or 24 bit and had no support for *Nan*, *Rounding to zero* and so on. The situation changed completely with the recent development and the switch to the *Unified Device Architecture* where SIMD processing was discarded in favour of more single independent threads.

The last two myths Item 3 and Item 5 are hard to demystify. If one has a algorithm which fits excellently to the GPU the full power of the GPU can be exploited and hence it is power efficient. It depends heavily on the algorithm. The feasibility study has shown that one can achieve easily a speedup but considering the inefficient use of the resources and the power dissipation it could be a slow down investing so much power for so little return.

MEAN SHIFT

In low level computer vision tasks like filtering, segmentation or edge detection, the analysis of data is often not done on the original images. Features like colours are rather projected into a feature space where they can be more easily analysed. The analysis of the feature space can find interesting attributes of the image like edges or segments.

The feature space has to be smoothed before analysis. Feature spaces originate from real images therefore they are composed of several components from different distributions. The basic approach of a mixture model, a mixture model is a probabilistic model for density estimation using a mixture distribution, is not efficient enough to estimate the density satisfactorily of such complex, arbitrarily comprised densities. The discontinuity preserving smoothing is therefore accomplished with kernel based density estimators. Kernel based density estimators are making no assumptions about densities and hence can estimate arbitrary densities.

The maxima of a feature space correspond to the searched components like the edges of an image. Gradient based methods of feature space analysis are using gradients of the probability density function to find the maxima. Such methods are complex because they need among other things a estimation of the probability of density.

Mean shift is an alternative to the gradient based methods as it is easier to calculate then to estimate the probability of density and then to calculate the gradient. The mean shift vector points to the same direction as the gradient of gradient based methods. Furthermore the *mean shift* vector has a adaptive size and is non parametric. There is no need to supply a step size compared to the other methods. Mean shift a robust approach toward feature space analysis was originally introduced by Comaniciu and Meer [10].

5.1 DENSITY ESTIMATION

In probability and statistics it is known that for different tasks there exist more or less suitable features. In Duda et al. [15] are giving an example of a classification of two different fish types. Features like the length

and brightness are there observed that fit for the task. It is of course possible to find more descriptive features for the fish like the amount of fins but in image processing it would be very expensive and difficult to count such feature. In image analysis specially in real time applications it is important to find suitable features for the task, but also features which are easily visually identified. Colour observations are because of their simplicity and for the eye easily to gather important features. The colour features can have components from the Red Green Blue (RGB) or Gray value colour space. Furthermore there are several other colour spaces that could be observed like the Hue Saturation Value (HSV) or the L* u* v* (LUV) colour space with one luminance and two chromatic components. The LUV colour space is often used in computer graphics because of its attribute to be a perceptually uniform colour space.

With a finite set of observations follows a finite feature space. The main point of *mean shift* is to find the maxima in the feature space. The maxima of a feature space are all important for *mean shift* applications (filtering, segmentation, ...) as the distributions or discontinuities map to clusters or edges of the image.

The *mean shift* method is based on the gradient method. For the gradient estimation a function is upon estimations of discrete observations in the feature space. For this kernel density estimators are used also known as *Parzen Window* method.

5.2 KERNEL DENSITY ESTIMATION

Kernel density estimation is a method to estimate an unknown density distribution with finite observations of a point in the sample space. The result of such a procedure is a probability density function that describes the density of probability at each point in the sample space. To estimate the density of a point $x = x_1, \dots, x_d, \dots, x_D \in \mathbb{R}^D$ in a D dimensional feature space, N observations x_i^N with $x_N \in \mathbb{R}^D$ within a search window that is centred around point x have to be observed. The search window with radius h is the bandwidth of the used kernel. The probability density in point x is the mean of probability densities that are centred in the N observations x_i^N .

The effect of different bandwidth parameters h (search window radius) is shown in Figure 5.1. The example shows a kernel density estimation

with five observations $x = 5, 1, -1, -4, -5$ and a gaussian kernel. The total density estimation is the sum of each kernel at a observation, here shown for three bandwidths. With bigger bandwidth h the density estimation becomes smoother.

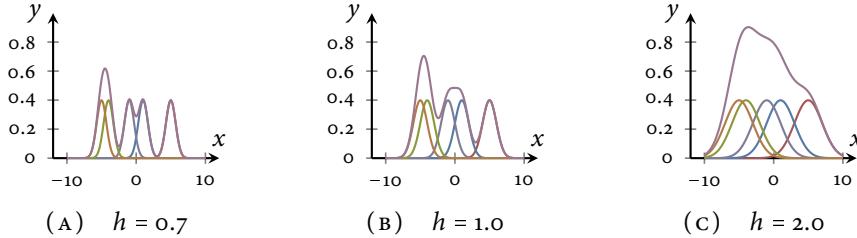


FIGURE 5.1 Effect of bandwidth selection

Kernel density estimation is a non parametric method, although some parameters exists like the search window radius. Non parametric methods are making no assumptions about the density of probabilities. The strength of such methods is that they are not limited to just one probability but they can deal with arbitrary coupled/joined probabilities. With an infinite number of observations the non parametric methods can reconstruct the density of the original probabilities.

To derive the kernel density estimator in point $x \in \mathbb{R}^D$ first of all some definitions have to be made. Let x be a random variable and N observations x_1^N with $x_n \in \mathbb{R}^D$ given. The kernel density estimator $\hat{f}(x)$ in a point $x \in \mathbb{R}^D$, with a kernel $K(x)$ and a $D \times D$ bandwidth matrix \mathbf{H} is

$$\hat{f}(x) = \frac{1}{N} \sum_{n=1}^N K_{\mathbf{H}}(x - x_n) \quad (5.1)$$

where

$$K_{\mathbf{H}}(x) = \frac{1}{\sqrt{|\mathbf{H}|}} K\left(\frac{x - x_n}{h}\right) \quad (5.2)$$

Since a full parametrized matrix \mathbf{H} would lead to very complex estimates only a single bandwidth parameter, the window radius will be

regarded. With the simplification $\mathbf{H} = h^2 \mathbf{I}$ the Equation 5.1 can be written as

$$\hat{f}(x) = \frac{1}{Nh^D} \sum_{n=1}^N K\left(\frac{x - x_n}{h}\right). \quad (5.3)$$

The kernel density estimator is valid for several kernels and successive considerations will deal with several kernel the Equation 5.3 will be formatted into a more generic form. For this transformation the definition of a kernel and profile of the kernel is needed. The following definition of a kernel and its profile is from Cheng [9]. The norm $\|x\|$ of x is a non negative number so that $\|x\|^2 = \sum_{d=1}^D |x_d|^2$. A $K : \mathbb{R}^D \rightarrow \mathbb{R}$ is a known as a kernel, when there is a function $k : [0, \infty] \rightarrow \mathbb{R}$ the profile of the kernel, so that

$$K(x) = c_{k,D} k(\|x\|^2) \quad (5.4)$$

where K is radial symmetric, where k is non negative, not increasingly and piecewise continuous with $\int_0^\infty k(r)dr < \infty$. $c_{k,D}$ is a positive normalisation constant so that $K(x)$ integrates to 1.

Now the kernel density estimator from Equation 5.3 can be transformed into a new equation. The two indices K and h are representing which kernel and which radius are used for the density estimator. With the profile notation k , where Equation 5.3 is inserted into Equation 5.2, the Equation 5.3 is transformed to

$$\hat{f}_{h,K}(x) = \frac{c_{k,D}}{Nh^d} \sum_{n=1}^N k\left(\left\|\frac{x - x_n}{h}\right\|^2\right) \quad (5.5)$$

5.3 KERNEL AND THEIR PROPERTIES

The following section will introduce three univariate profiles and their associated multivariate radial symmetric kernels.

From the Epanechnikov profile

$$k_E x = \begin{cases} 1 - x & 0 \leq x \leq 1 \\ 0 & x > 1 \end{cases}, x \in \mathbb{R} \quad (5.6)$$

follows a radial symmetric kernel

$$K_E(x) = \begin{cases} \frac{1}{2} c_D^{-1} (D+2) (1 - \|x\|^2) & \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}, x \in \mathbb{R}^D \quad (5.7)$$

where c_D is the Volume of the D dimensional globe. The epanechnikov kernel is used often as it minimises the Mean Integrated Squared Error (MISE) [11]. The Figure 5.2a shows the epanechnikov kernel. The derivative of the kernel is a uniform profile.

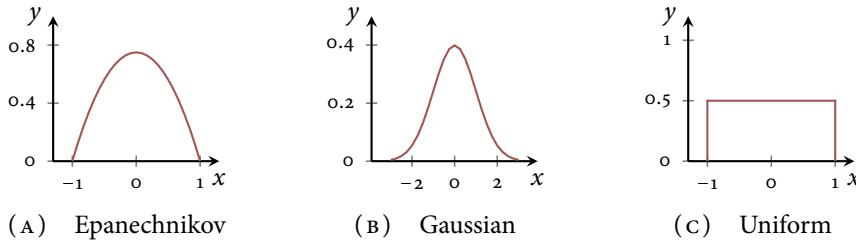


FIGURE 5.2 Kernel density estimators

From the normal profile

$$k_N(x) = \exp\left(-\frac{1}{2}x^2\right) \text{ where } x \geq 0, x \in \mathbb{R} \quad (5.8)$$

follows the normal kernel

$$K_N(x) = (2\pi)^{-D/2} \exp\left(-\frac{1}{2}\|x\|^2\right), x \in \mathbb{R}. \quad (5.9)$$

The normal distribution as every other kernel with infinite support, is often capped for finite support. Finite support is important for the convergence. Capping the normal kernel can be accomplished by multiplying it by a uniform kernel where the inner part of the normal kernel is cut out and weighted with 1 and the outer part is set to 0. The derivative of the normal profile is again a normal profile.

From the uniform profile

$$k_U(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}, x \in \mathbb{R} \quad (5.10)$$

follows the uniform kernel

$$K_U(x) = \begin{cases} 1 & \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}, x \in \mathbb{R}^D \quad (5.11)$$

which is a hyper unit ball in the origin.

Assuming that a derivative of a profile $k(x)$ exists for all $x \in [0, \infty)$, follows a new profile $g(x)$. Now a new kernel $G(x)$ can be defined as

$$G(x) = c_{g,D} g(\|x\|^2) \quad (5.12)$$

where $c_{g,D}$ a normalizing constant and $K(x)$ the *shadow kernel* of $G(x)$. The term *shadow kernel* was introduced in the context of mean shift in [9]. The mean shift vector of a kernel points to the same direction as the gradient of the shadow kernel (See Section 5.4.1).

5.4 MEAN SHIFT

In gradient based methods first the gradient s calculated and then the kernel is shifted by a vector with a specific length in direction of a maximum of the probability. The magnitude/length that is the step size of the vector has to be chosen. The problem of such gradient based methods is the choice of a suitable step size. The run time of such algorithms depends heavily on the right choice of the step size. If the step size is too large the algorithm diverges and choosing a too small step size the algorithm becomes very slow. Convergence is only guaranteed for infinitesimal step sizes. There are several complex procedures for finding the right step size, see Comaniciu and Meer [10].

In the case of mean shift there are no additional procedures needed to choose the step size. The magnitude/length of the mean shift vector is the step size which is adaptive regarding the local gradient of the density of probability. Because of this adaptive nature the mean shift algorithm converges (Proof see Comaniciu and Meer [10]).

The advantage of the mean shift method contrary to gradient based methods is that the step size has not to be chosen by hand and the gradient has not to be calculated. It can be shown that the mean shift vector is pointing to the same direction as the gradient and that it moves along the gradient to the maxima that can be seen in Equation 5.18.

Hence it is sufficient to calculate the more efficient mean shift vector rather than the gradient.

Given are N D -dimensional observer feature vectors x_i^N with $x_n \in \mathbb{R}^D$ and a kernel G at the point $x = x_1, \dots, x_d, \dots, x_D \in \mathbb{R}^D$ in the feature space with search window radius h is

$$m_{h,G}(x) = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \quad (5.13)$$

the D -dimensional *mean shift* vector. The N observations are weighted by the means of kernel G , summed and normalised with the total sum. The *shift* is the difference between the weighted *mean* and x , thus the name *mean shift*.

As pointed out in Section 5.2 the main objective of density estimation is to efficiently find the maxima of a distribution in feature space. The maxima of a function f are located at the positions where the gradient $\nabla f(x) = 0$. With the above stated attribute that the *mean shift* vector always moves along the direction of the gradient, it is a elegant solution for finding the maxima of a distribution without estimating the density.

5.4.1 Density Gradient Estimation

The usage of a differentiable kernel allows one to write the density gradient estimator as the gradient of the density estimator

$$\hat{\nabla} f_{h,K}(x) \equiv \nabla \hat{f}_{h,K}(x) = \frac{2C_{k,D}}{Nh^{D+2}} \sum_{n=1}^N (x - x_n) k'\left(\left\|\frac{x - x_n}{h}\right\|^2\right) \quad (5.14)$$

where the inner part and a part of the prefactor originate from the differentiation of $k\left(\left\|\frac{x-x_n}{h}\right\|\right)$

$$\begin{aligned} \frac{\delta}{\delta x} k\left(\left\|\frac{x - x_n}{h}\right\|^2\right) &= \left(\left\|\frac{x - x_n}{h}\right\|\right)' k'\left(\left\|\frac{x - x_n}{h}\right\|^2\right) \\ &= 2(x - x_n) \left(\frac{1}{h}\right) k'\left(\left\|\frac{x - x_n}{h}\right\|^2\right) \\ &= \frac{2}{h^2} (x - x_n) k'\left(\left\|\frac{x - x_n}{h}\right\|^2\right). \end{aligned} \quad (5.15)$$

Using $g(x) = -k'(x)$ and with Equation 5.12 a new kernel $G(x)$ with profile $g(x)$ can be defined. Transforming Equation 5.14 with the new profile $g(x)$ the gradient of the density estimator becomes

$$\hat{\nabla}f_{h,K}(x) = \frac{2c_{k,D}}{Nh^{D+2}} \sum_{i=1}^N (x_n - x) g\left(\left\|\frac{x-x_n}{h}\right\|^2\right) \quad (5.16a)$$

$$= \frac{2c_{k,D}}{Nh^{D+2}} \left[\sum_{n=1}^N g\left(\left\|\frac{x-x_n}{h}\right\|^2\right) \right] \left[\frac{\sum_{i=1}^N x_i g\left(\left\|\frac{x-x_n}{h}\right\|^2\right)}{\sum_{n=1}^N g\left(\left\|\frac{x-x_n}{h}\right\|^2\right)} - x \right]. \quad (5.16b)$$

The first term of Equation 5.16b conforms with the density estimator $\hat{f}_{h,G}(x)$ for kernel G (compare with Equation 5.5 for kernel K) except for a factor whereas the second term is the difference between the centre of the, with kernel G weighted centre of observation and the centre x of the kernel window which conforms to the *mean shift* vector from Equation 5.13. To localise the maxima with *mean shift*, the maxima are the roots of the gradient, it firstly has to be shown that the *mean shift* vector is moving along the direction of the gradient. Inserting $\hat{f}_{h,G}(x)$ and $m_{h,G}(x)$ into Equation 5.16b follows

$$\hat{\nabla}f_{h,K}(x) = \frac{2c_{k,D}}{h^2 c_{g,D}} \hat{f}_{h,G}(x) m_{h,G}(x) \quad (5.17)$$

transformed to $m_{h,G}(x)$ follows

$$m_{h,G}(x) = \frac{1}{2} h^2 c \frac{\hat{\nabla}f_{h,K}(x)}{\hat{f}_{h,G}(x)}, \text{ whereas } c = \frac{c_{g,D}}{c_{k,D}}. \quad (5.18)$$

The denominator of Equation 5.18 is the normalising factor that originates from the density estimator with kernel G in x and the numerator is the gradient density estimator with kernel K . In fact the *mean shift* vector is proportional to the gradient which means it is adaptive. Kernel K is the shadow kernel of kernel G . The term shadow kernel was firstly introduced by Cheng [9].

Let be

$$mi_{h,K}(x) = \frac{\sum_{i=1}^n x_i k\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n k\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \quad (5.19)$$

the D -dimensional mean of the observations x_1^N from \mathbb{R}^D weighted with kernel K and a window radius h . Then is K the shadow kernel to kernel G if the *mean shift* vector with kernel G

$$m_{h,G}(x) = mi_{h,G}(x) - x = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \quad (5.20)$$

lies in the gradient density estimator direction with kernel K

$$\hat{f}_{h,K}(x) = \frac{c_{k,D}}{Nh^d} \sum_{n=1}^N k\left(\left\|\frac{x-x_n}{h}\right\|^2\right) \quad (5.21)$$

In the following sections gradients will not more be considered as it was shown in Equation 5.18 that the *mean shift* vector moves along the same direction as the gradient. Instead of estimating a density with a kernel density estimator with kernel K and then calculating the gradient now one can achieve the same solution with the *mean shift* and the differentiation K' of kernel K . The next section will continue with the actual *mean shift* method.

5.4.2 Mean Shift Method

The *mean shift* vector moves in direction of the maximal slope of the density, it defines a path to the maximum. The *mean shift* method is described by the following iterations:

1. Choose a window radius h_n for the kernel density estimator in Equation 5.5
2. Choose a start position $y_1 \in \mathbb{R}^D$ for the kernel window
3. Calculate the *mean shift* vector $m_{h,G}(y_j)$ and shift the kernel window G
4. Repeat 3. until convergence

ALGORITHM 5.1 Mean Shift Method

The first step of the algorithm is to choose the window radius or the bandwidth h_n of the kernel in the observation x_n . The bandwidth can be

chosen adaptively or can be fixed. With a fixed bandwidth the density estimator in Equation 5.5 works with identical scaled kernels in every observation or is adapted in every observation. In literature there are several ways described to perform a adaptive Mean Shift (see e.g. Li1 and Klette [31]).

The second step is to choose a start position $y_1 \in \mathbb{R}^D$ in the feature space where to start the mean shift algorithm.

The third step involves shifting the kernel window with the mean shift vector $m_{h,G}(y_j)$. The new position $y_{j+1} \in \mathbb{R}^D$ of the search window in the feature space is calculated with:

$$y_{j+1} = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} \quad j = 1, 2, \dots \quad (5.22)$$

where $y_j \in \mathbb{R}^D$ the old position of the kernel window is. The algorithm is convergent and if a stationary point is also a convergence point, the point is moved by a small random vector and the algorithm is applied again to the shifted point. This way one can guarantee that the found point is really a convergence point. The convergence attribute and the small trick stated above are described by Comaniciu and Meer in [10].

5.5 FILTERING & SEGMENTATION

The primary use of *mean shift* is filtering (smoothing) and segmentation. A colour image can be seen as a 2-dimensional matrix $I \times J$ with N 3-dimensional vectors. The pixels in the image $x_n, n = 1, \dots, N$ consist of a spatial part $x_n^r = (i, j) \in I \times J$ and a part with a colour range $x_n^f = (r, g, b)$ so $x_n = (x_n^r, x_n^f) \in \mathbb{N}^5$ for $n = 1, \dots, N$. Other colour spaces like the LUV colour space could be used as well. The euclidean metric is assumed for both spaces.

When applying *mean shift* for filtering and segmentation applications like in Comaniciu and Meer [10] a joined feature space is used for the spatial and range components of a pixel. The colour space used in filtering and segmentation is the LUV colour space because of its feature of linear mapping. In a joint D -dimensional feature space ($D = 5$ colour images, $D = 3$, gray tone images) the different attributes of each space have to be equalised by normalisation. Therefore is the joint kernel written as a

product of two kernels with window radius h_r for the spatial part and h_f for the colour range part

$$K_{h_r, h_f}(x) = \frac{C}{h_r^2 h_f^p} k\left(\left\|\frac{x^r}{h_r}\right\|^2\right) k\left(\left\|\frac{x^f}{h_f}\right\|^2\right) \quad (5.23)$$

where p is the colour dimension of the image, $p = 1$ gray tone and $p = 3$ colour. With the parameter $h = (h_r, h_f)$ the window radius one can specify the size of the kernel and thereby specify the resolution of the maximum search.

5.5.1 Mean Shift Filtering

Smoothing or filtering with mean shift or bilateral filters have the advantage that discontinuity like edges are preserved. Applying a simple smoothing a weighted average of the neighbours in both space and in colour range are considered which systematically excludes pixels across the discontinuity from consideration.

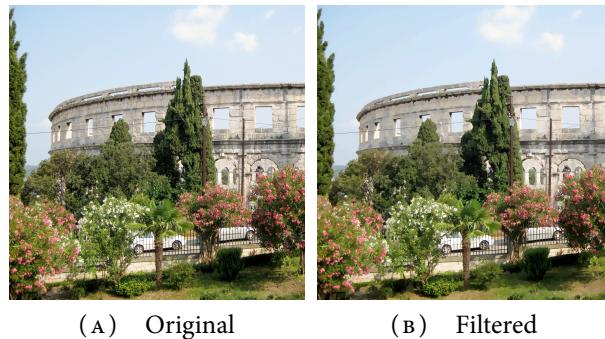


FIGURE 5.3 Mean shift filtering with parameters $(h_r, h_f) = (6.5, 7)$ applied to a colour image

The pixels $x_n, n = 1, \dots, N$ of the image converge toward their local

density maximum applying the filtering Algorithm 5.2.

```

Data:  $x_n = (x_n^r, x_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional RGB pixels
Data:  $c_n = (c_n^r, c_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional LUV pixels
Data:  $z_n = (z_n^r, z_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional filtered pixels
Data:  $o_n = (o_n^r, o_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional output pixels

for  $n = 1, \dots, N$  do
| convert  $x_n$  from RGB to LUV colour space
|  $c_n = x_n$ 
end

for  $n = 1, \dots, N$  do
| initialise  $m = 1$  and  $y_n = c_n$ 
| while not converged do
| | calculate  $y_{n,m+1}$  according to  $y_{j+i} = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}$ 
| | end
| | assign  $z_n = (x_n^r, y_{conv}^f)$ 
end

for  $n = 1, \dots, N$  do
| convert  $z_n$  from LUV to RGB colour space
|  $o_n = z_n$ 
end
```

ALGORITHM 5.2 Mean Shift Filtering

All pixels that converge to the same um are lying in the basin of attraction of this maximum. The filtered image points z_n , $n = 1, \dots, N$ keep there spatial coordinates but obtain the colour values off the convergence point. The convergence points are found by moving the kernel window with *mean shift* in direction of maximum slope of the spatial range feature space. The sample Figure 5.3 was smoothed with the algorithm of Algorithm 5.2. A uniform kernel with $h = (6.5, 7)$ was used.

5.5.2 Mean Shift Segmentation

Image segmentation is a method to partition an image into homogeneous regions. Searched are regions with similar colours like a wall, lawn and clothes. The found areas are associated with the same colour values. The segmentation can be seen as a strong smoothing where the edges are preserved.

The segmentation algorithm is an extension of the *mean shift* filtering algorithm. After applying the filter and all convergence points are found, clusters are built out of them. All convergence points that are closer than h_r in the spatial domain and that are closer than h_f in the range domain are grouped together. In the end all points are labeled after their cluster assignment.

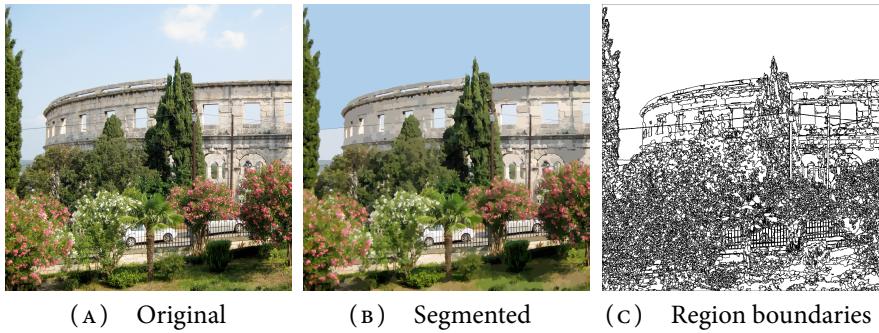


FIGURE 5.4 Mean shift segmentation with parameters $(h_r, h_f, M) = (6.5, 7, 20)$ applied to a colour image

The Algorithm was applied on the image Figure 5.4a. A uniform kernel with a kernel radius of $h_r = 6.5$ and $h_f = 7$ was used. M is a parameter for the last step of the algorithm where regions where the pixel count is smaller than M are purged. M defines the smallest significant feature size. Figure 5.4 shows the result of a segmentation. After the image segmentation follows a edge detection by examining the cluster regions Figure 5.4c.

In filtering as well as in segmentation algorithm a fixed sized window size is used. In Comaniciu and Meer [10] is noted that for segmentation the algorithm is not heavily dependent on the choice of the kernel parameters where as there are application where the window size matters.

The segmentation part of the Algorithm 5.3 has no significant impact on the run time.

```

Data:  $x_n = (x_n^r, x_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional RGB pixels
Data:  $c_n = (c_n^r, c_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional LUV pixels
Data:  $z_n = (z_n^r, z_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional filtered pixels
Data:  $o_n = (o_n^r, o_n^f)$ ,  $n = 1, \dots, N$  the  $D$ -dimensional output pixels

for  $n = 1, \dots, N$  do
| convert  $x_n$  from RGB to LUV colour space
|  $c_n = x_n$ 
end

for  $n = 1, \dots, N$  do
| initialise  $m = 1$  and  $y_n = c_n$ 
| while not converged do
| | calculate  $y_{n,m+1}$  according to  $y_{j+i} = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}$ 
| | end
| | assign  $z_n = (x_n^r, y_{conv}^f)$ 
end

for  $n = 1, \dots, N$  do
| delineate in the joint domain the clusters  $\{C_p\}_{p=1,\dots,P}$  by grouping
| together all  $z_n$  which are closer than  $h_s$  in the spatial domain and  $h_r$ 
| in the range domain.
end

for  $n = 1, \dots, N$  do
| assign  $L_n = \{p | z_i \in C_p\}$ 
end

eliminate spatial regions containing less than  $M$  pixels

for  $n = 1, \dots, N$  do
| convert  $z_n$  from LUV to RGB colour space
|  $o_n = z_n$ 
end
```

ALGORITHM 5.3 Mean shift segmentation

5.6 SUMMARY

Mean shift is a elegant method to analyse feature spaces without gradient based methods. Using gradient based methods first an density estimation has to be defined. With this estimation a gradient is calculated and the algorithm is moving with a choosable step size towards a maximum. Mean shift converge to the same maximum but contrary to gradient based density estimation it is adaptive which means no step size has to be chosen. Depending on the chosen kernel for mean shift, the derivative that is used in the density estimation, can be very easily computed. Mean shift is an alternative way to compute the density estimation without choosing a step size.

MEAN SHIFT ALGORITHM ANALYSIS & DESIGN

Before any porting can be started the developer has to find out if there are multiple activities or tasks which can run simultaneously which expose exploitable concurrency. The developer has to find concurrency either by decomposing the data or tasks. By this decomposition one wants to solve bigger problems in less time as several processing units can solve different parts of the problem.

But before any analysis is started one has to know if the problem is large enough and if the resulting speedup justifies all the effort that is expended on making a parallel version out of it. In case of mean shift which is used for several things like filtering, segmentation, pattern recognition and real time tracking one can deduce that for bigger images or many images the computation time climbs fast as the dsize or the number of images rises. To have a clue how the mean shift algorithm behaves with big pictures several run times were recorded. For the analysis of the mean shift algorithm a ready to use Edge Detection and Image SegmentatiON System (EDISON) was used that was profiled and modified and parallelized for the purpose of this thesis.

The EDISON¹ system which was developed by the authors Comaniciu and Meer of [10] offers functionality to filter, segment and detect edges in images.

The Figure 6.1 shows results of CPU run times of the EDISON application dependant on image size. The vertical axis shows the run time in seconds and the horizontal axis shows the side length in pixels of the quadratic Figure 5.3a. Considering the numbers in the result one can see that the run times grow linear to the image sizes. The mean shift algorithm has linear complexity, hence its complexity can be written as $O(n)$. Doubling the side length of the quadratic picture the run time increase by a factor of 4. See e.g. the run times for side length: (l=256, t=9) and (l=512, t=36).

But this is obvious, as stated in Chapter 5 for each pixel of the image the mean shift vector has to be calculated, and if one increases the number of pixels by a number n we have to deal with n times longer run time. Such consideration have to be done in the run-up to have a clue to which

¹ <http://www.caip.rutgers.edu/riul/research/code/EDISON/index.html>

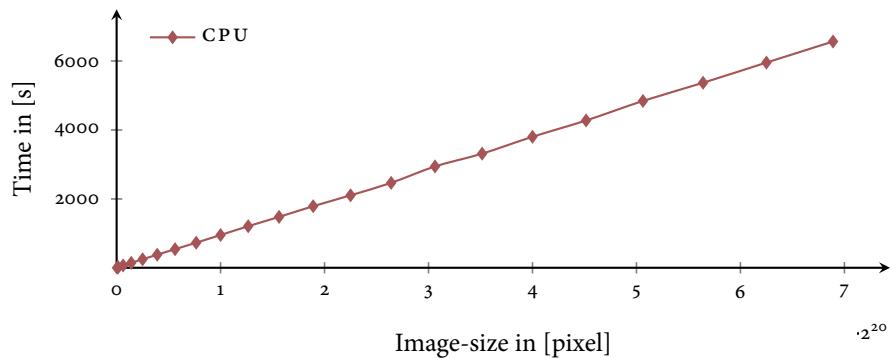


FIGURE 6.1 CPU run time with increasing image size

point the algorithm can be accelerated. The problem has to be well understood.

The next important step is to find parts which can be offloaded to an accelerator as the GPU. A first good point for such parts is to make a profiling of the application to find out the most computationally intensive parts. It makes no sense to parallelise functions which contribute only 1% to the run time.

6.1 PROFILING THE ORIGINAL CODE

A good start point is to take a profiler and generate a profile of the functions, recording their run time and call history. In this case a statistical profiler is used which operates by sampling. The samples are taken from the hardware performance counters which every modern CPU has built-in. OProfile² a system wide profiler for Linux was used to examine the run time of the EDISON program. The Table 6.1 shows the run time analysis of EDISON.

Before taking a closer look at the table, the columns have to be explained. The column *Total* shows how long the function plus their child function were executing. Whereas the column *Self* shows how long the function spent executing itself without the execution time of their child functions. The column *Symbol* shows the function that was executed.

Now having a look at the table there is one function which is executing

² <http://oprofile.sourceforge.net/news/>

Self (%)	Total (%)	Symbol
0.00	100.00	Segment()
0.00	100.00	meanShift(int)
0.00	99.30	Filter(...)
0.40	99.10	NonOptimizedFilter(...)
0.60	98.80	LatticeMSVector(double*, double*)
98.20	98.20	uniformLSearch(double*, double*)
0.00	0.40	FuseRegions(float, int)
0.10	0.30	BuildRAM()
0.00	0.20	TransitiveClosure()

TABLE 6.1 EDISON run time analysis

98.2% of the run time, the function *uniformLSearch(double*, double*)*. In this function the algorithm tries to find feature points that fall into the search window with radius h (see Section 5.4). This is the starting point as it is the most computationally intensive and the focus for parallelisation. There is a way to calculate to which extent the particular function can be parallelised and which speedup one can expect. Speedup is the ratio of run time when executing a program on a single processor to the run time when using n processors. Given T_1 the run time of a program on a single processor and T_n the run time of the same program on n processors then

$$S(n) = \frac{T_1}{T_n} \quad (6.1)$$

is a measure for the speedup. One familiar law to calculate the how much speedup can be obtained through parallelism is Amdahl's law.

6.2 AMDAHL'S LAW

Amdahl's law says supposing that 80% of a computation can be parallelised and 20% can't, then even if the 80% that is parallel were run on an infinite number of processors the highest speedup that one can achieve is 5. Generally speaking if a fraction p of a computation can be run in parallel and the rest must run serially, Amdahl's law upper bounds the speedup by $1/(1-p)$.

The speedup of a application according to Amdahl is given by

$$S_{max}(n) = \frac{1}{(1-p) + \frac{p}{n}} \quad (6.2)$$

where 1 is the execution unit time of the old computation, $(1-p)$ the inherently serial part and p/n the parallel part divided by n the number of processing units. When $n \rightarrow \infty$ then the execution time approaches the time for executing the sequential program fraction. So no matter how many processors one adds to the system it will at least execute as long as the sequential program fraction, which is an upper bound of speedup. A important addition is that the sequential program fraction or serial processing percentage is relative to the overall execution time using a single processor. It is independent of the number n of processors Shi [48].

Assuming that the mean shift filtering can be parallelized with any number of processing units one can calculate the maximum speedup achievable according to Amdahl. Taking the parallel processing percentage from Table 4.2 for the filtering step:

$$p = 99,3\% = 0.993p$$

we get

$$S_{max}(n) = \frac{1}{(1-0.993)} = \frac{1}{0.007} = 142.857$$

Applying the calculation with $n = 128$, the used GPU has 128 processors one can achieve a speedup of:

$$S_{max}(128) = \frac{1}{(1-0.993) + \frac{0.993}{128}} = \frac{1}{0.007 + 0.0077578125} = 67.76$$

One can see that even for very small serial processing percentage the speedup is not very high. Therefore Gustafson proposed a alternative formulation where the fractions are now dependent on n . He assumed that for larger problem, the fraction of a program to parallelisation increases. Thats why Gustafson's law is often referred as a scaled speedup measure and Amdahl's as a non scaled speedup measure. Given the

serial s' and parallel time p' a single processor would require $s' + p' \times n$ time to finish the execution. The speedup is then given by:

$$S_{max}(n) = \frac{s' + p' \times n}{s' + p'} = n + (1 - n) \times s' \quad (6.3)$$

Applying the calculation with $n = 128$, the used GPU has 128 processors one can achieve a speedup according to Gustafson of:

$$S_{max}(128) = 128 + (1 - 128) \times 0.007 = 127.11$$

It looks like that Amdahl's & Gustafson's law are in contradiction but looking more precise on both definitions one can see that both laws employ different definitions for the fraction of the serial and parallel execution times. Amdahl uses non scaled percentage and Gustafson scaled percentage. Mathematically they are equal just two different formulations. The scaled percentage can be transformed to a non scaled percentage where Gustafson's law gives the same results as Amdahl's law [48].

In summary one can expect more than a 100 fold speedup when parallelising the filter step of the mean shift algorithm. The next steps will focus on analysing how the mean shift filter can be decomposed to take advantage of multiple processors.

6.3 DATA & TASK PARALLELISM

There are two ways to decompose an algorithm to take advantage of parallelism. The first way is to decompose the algorithm into several tasks that can run independently. Each task is performing independently different calculation on the data. This characteristic is known as task parallelism. To know if two task can run independent one can take Bernstein conditions and evaluate them.

Let P_i and P_j be two tasks of a program P . For P_i let I_i be the input and O_i the output data and for P_j let I_j be the input and O_j be the output data, then P_i and P_j are independent if they satisfy the following conditions [48]:

$$I_j \cap O_i = \emptyset \quad (6.4a)$$

$$I_i \cap O_j = \emptyset \quad (6.4b)$$

$$O_i \cap O_j = \emptyset \quad (6.4c)$$

The first two formulations Equation 6.4a and Equation 6.4b are stating that the input of one task has to be independent from the other task. Otherwise one task would have to wait for the output of the other task to continue. Furthermore not only the inputs vs. the outputs have to be independent but also the outputs of each task have to be independent. If the above conditions cannot be applied to a program P than it is likewise that race conditions and poor performance can arise through e.g. excessive synchronisation overhead.

6.3.1 Task Parallelism

Equipped with the knowledge how to identify independent parallel tasks, the mean shift segmentation algorithm Algorithm 5.3 can now be examined. Having a look at the algorithm steps one can easily see that every task is violating every condition stated above. The filtering step input data is dependent on the RGB to LUV conversion output data. The segmentation step input data is dependent on the filtering output data. So it does not matter how big an image is or if it is a colour or grayscale image the mean shift segmentation algorithm will always perform all from each other dependent calculations as described in Algorithm 5.3. The longest path of dependent calculations is the critical path. For the mean shift algorithm there exists no shorter path.

In summary one can discard the idea of task parallelism for the mean shift segmentation algorithm. Thats why the focus of parallelisation is now on data parallelism.

6.3.2 Data Parallelism

The second way of decomposing an algorithm is data decomposition. The data is decomposed into chunks on which similar operations are being applied in such a way that the different chunks can be operated on concurrently. The focus in data parallelism is on the data structures which define the problem and not at the tasks.

Focusing on data and having a look at the most computational intensive part (the filtering step) one can see that each pixel of an image can be calculated independently. The input and output pixels are independent and furthermore it doesn't matter at which pixel one starts, the filtering

step is deterministic, starting from the same input will lead to the same output.

Another important aspect is how often the subtasks (where each subtask is calculating one pixel) have to synchronise or communicate. An algorithm exhibits fine-grained parallelism if the subtasks have to communicate many times, coarse-grained parallelism if the subtasks do not communicate many times. In this case the algorithm even exhibits embarrassingly parallel parallelism. The subtasks never have to communicate or synchronise. Each subtask takes one or a chunk of pixels applies the filtering steps and writes back the result. The intermediate steps where the mean shift vector is moved over the spatial space are also independent for each pixel.

The approach in parallelising the mean shift algorithm is to use a data decomposition where each subtask is filtering a chunk or a pixel of the entire image. Depending on the number of processing units one can choose an appropriate chunk size to exploit every unit.

If the algorithm would only exhibit fine-grained parallelism the next steps would be to identify groups to simplify the job of managing dependencies and have a look at the ordering to satisfy constraints among tasks. Luckily here one has only to deal with a embarrassingly parallel algorithm and can move to the next step, identify how data is accessed and shared among subtasks.

6.4 DATA FLOW

It is important to understand that inefficient data access can lead to very poor performance on every processing unit especially on a GPU. Inefficient data access leads to an algorithm that is memory bound which means it doesn't matter how much computational power the processing unit has it will be bound to the memory performance (See Chapter 2 for *arithmetic intensity*). Therefore its crucial to understand the data and optimise it for access. If done incorrectly tasks may get invalid data or it could lead to excessive synchronisation overhead.

In Section 6.3.1 it was shown that the filtering step is dependent on the colour conversion step where the output data of the colour conversion step becomes the input data of the filtering step. Since this is done sequentially there is no need to take care of some synchronisation. This

scheme is the same for the filtering and segmentation step. Since the filtering step is done in parallel and the segmentation needs the output data from the filtering step a barrier after the filtering step is needed.

Now to the interesting part the parallel filtering step. For the following considerations the viewpoint will be a single processing unit. It is firstly enough to consider only one unit and one pixel at a time since the GPU exhibits a SPMD programming model and secondly the mean shift algorithm is embarrassingly parallel. It is later easy to broaden the examination to a cluster of processing units and chunks of pixels if its needed.

A processing unit firstly reads the corresponding pixel $x_n = (x_n^r, x_n^f)$ from the image. Then it moves the mean shift vector to the next pixel on the path to the convergence point $z_n = (x_n^r, y_{conv}^f)$ by finding the feature vector which falls into the search window. This new pixel is the basis for the next iteration—calculation of the mean shift vector. On average of about 15–20 iterations the mean shift vector converges to the mean at the convergence point [58]. The whole path where the mean shift vector is moving on is unpredictable. The vector can move over the whole image or a few pixels, this is an important fact to consider when mapping the algorithm to the GPU because random access to the memory can decrease the memory performance significantly. At last the convergence point is saved to the $z_n = (z_n^r, z_n^f)$ data structure.

The segmentation step as stated above needs the data for the filtering step. The only parallel portion of the algorithm is the filtering step so no further considerations will be made.

6.5 THE NEXT PHASE

So far the problem has been extensively analysed. The most compute intensive parts have been identified and a preliminary speedup has been calculated according to Gustafon's law. Furthermore the problem has been decomposed into tasks that can execute concurrently using task respectively the data decomposition. Additionally to the decomposition the dependencies between processing units has been identified.

The items mentioned above will guide the design of the algorithm in the next phase. Therefore it is important to have clearly understood the algorithm and have done the decomposition in the right way.

The decomposition has to be suitable for the target platform. Since the GPU is a S P M D architecture, which means one program the mean shift algorithm is operating on many data the pixels. Above it was shown that the way of data decomposition is preferable because (1) the pixels can be mapped 1:1 to a processing unit on the GPU and (2) there are enough pixel to keep the GPU busy. Furthermore the GPU is a number cruncher and algorithms doing heavy math or calculations can profit from the sheer computational power.

The decomposition has to be flexible, efficient and simple. A flexible setup means that the amount of independent task can vary and the configuration how they are arranged is as well selectable. This means it can be adapted to from one to many processing units, the load can evenly be distributed among the processing units which leads to good efficiency and scalability and the decomposition can be parameterised in terms of units and problem size. Additionally weak dependency in the decomposition makes load balancing easier. Simple in this case means minimising the synchronisation and communication overhead through the decomposition, which is in this case non-existent.

The concurrency in the mean shift algorithm is now exposed and the algorithm can now be taken into the next phase, designing the algorithm.

6.6 THE ALGORITHM STRUCTURE

The next step involves the structuring of the algorithm to refine the design going further into detail and move it closer to a program that can run on the GPU. The main concern here is how to map the concurrency identified in the previous chapters to the processing units of a GPU. Usually one wouldn't go to much into hardware details in this phase because the design should be flexible and portable, but without considering any hardware details of the GPU the algorithm is unlikely to run efficient. Therefore each design decision made is based upon observance of hardware and programming model (CUDA) features. The strategy taken here is start from high level elements of the programming model and go ever further into detail until the point is reached where the algorithm can be implemented.

6.7 PROGRAMMING MODEL

In Section 3.4 it was explained that applications are accelerated by executing data parallel portions of the algorithm on the GPU as *kernels* which run in parallel on many threads exposing the SPMD nature of a GPU. Additionally in Section 6.3.2 the parallel portion of the algorithm was identified and hence the filtering step will be implemented as a CUDA kernel where it operates on different data simultaneously. There is a strong dependency between the steps of the algorithm shown in Section 6.3 that have to be considered here as well. Before and after the filtering a colour conversion from RGB to LUV and back is needed. These steps are not very compute intensive but why are they considered here?

An example is taken to have some numbers to juggle with. Supposing the algorithm is split in 3 parts and the main part is executing 100 seconds and the two other parts are executing 2 seconds. Additionally supposing that the main part can be accelerated by a factor of 100 which means the main part takes now 1 second, the most computational intensive part is not more the main part but rather the remaining two parts. That means no matter how much faster the main part will be the program will need always at least 4 seconds + accelerated main part to execute. The solution to this problem would be to accelerate the two remaining parts as well and that's what will be done for the mean shift algorithm. The two colour space conversion steps will be as well offloaded to the GPU. With bigger and bigger images the CPU will need more and more time to convert from one colour space to another.

All the other portions will be run on the CPU because there are not(-yet) a significant part of the run-time and are mostly post processing steps for the actual algorithm. The CPU will prepare, provide and gather the data and results from the GPU.

6.8 THREAD BATCHING

Fine grained, data parallel threads are the key concept of CUDA. Several of these threads are grouped together to a thread block. Thread blocks are again grouped to so called grids and kernel is as a grid of thread blocks where data memory space is shared by all threads. The grid and

the thread block can be organised either as a 1D, 2D or as a 3D array, dependent on the problem structure adjusting to the state of the original data. This means that e.g. if the problem is 3D one could use the 3D configuration or as in the mean shift algorithm case where the problem is a 2D image than obviously one should use the 2D array configuration.

Therefore the kernel used will be a 2D grid of 2D thread blocks. The grid represents the whole image and the thread blocks are the image chunks that are updated in parallel where a single thread is working on a single pixel. The exposed parallelism is directly mapped to the problem.

For an efficient execution several x, y combinations for the 2D array of the thread blocks should be evaluated. The grid size in x and in y is fixed since that are the dimensions of the analysed image.

Since the algorithm is embarrassingly parallel there is no need to look for communication and synchronisation mechanisms in CUDA.

6.9 DEVICE MEMORY SPACE

The memory space is a hierarchy of several memory types that can be accessed per thread, block, grid and the host. The first most important memory is the global memory that is the main data exchange place between the CPU and GPU. For the further considerations other memory types are not taken into account because (1) there are implementation specific and (2) using different memories for the algorithm would not change the original nature of organisation and decomposition. They can be later used to increase performance and data throughput.

Coming back to global memory it is mainly used to read in the input data and store the results which can be later either visualised using a graphics library or transferred back to the CPU where it can be further processed.

Therefore the CPU will run all preprocessing steps to the image and transfer it to the global memory on the GPU. After the filtering step where the GPU is reading continuously from the global memory for each iteration of the filtering step the CPU will fetch the data and post process it. Further interaction between CPU and GPU are non-existent.

6.10 WARPS

On the GPU 32 threads build up a SIMD unit in NVIDIA terms called warp, in that the same instruction is executed at a time but on different data where on a CPU the threads are scheduled independently.

The consequence out of this is that if there is a branch the execution of the two branch paths, when both are executed, are serialised. The only branch that can happen in the mean shift algorithm is that several threads have found their basin of attraction and others are still calculating. In this case there will be no serialisation happen rather there are threads that are idle.

This branching or dividing of a warp into idle and active threads is impossible to circumvent. Because each image has different densities and the locations of the densities vary there have to be threads that are idle, those e.g. that are close to a density maximum and finish the calculation in few iterations, and threads that are active, those that were further apart of the density maximum and need a higher number of iterations. This means threads in a warp that are gone to idle state will stay idle and active threads will do their calculations no rearranging of threads or reassignment of data will be done.

In summary it can be said that each thread is executing the same program. The execution time of each thread highly depends on the location of the thread in the grid and the location of the density maximums. The different execution times are visible directly in the final iteration count.

6.11 PROGRAM FLOW

The point of view will be single thread since all threads are doing the same. After the CPU has loaded the picture it preprocesses it by converting the RGB pixel values to LUV. After the conversion the CPU transfers via DMA the image to the GPU.

A thread in the grid having a unique id can now fetch the corresponding pixel according to the id. Equipped with the colour data the thread can now calculate the mean shift vector until convergence. While calculating the thread will access various pixels in the image that are read-only in the global memory. The found convergence point is saved to a new buffer in the global memory.

Finally the CPU fetches the convergence points, converges the image from the LUV colour space to the RGB space. The post processing step involves the segmentation and pruning of regions smaller than a specified variable. The result is a segmented image.

6.12 SUMMARY

After an extensive analysis it was no problem to put the single parts together to form a program. The design phase usually does not consider any hardware but in this case where an efficient algorithm should be mapped to the GPU it is crucial to know the architecture and the features the hardware offers. Furthermore it is important to have a look at the data. A good design follows from good data design. In this case a data decomposition was chosen and hence the program had to accommodate to this circumstance. The data decomposition was not chosen arbitrarily it was chosen with the ulterior motive that the data fits 1 : 1 to the hardware features.

In summary, data is more important than code, because crappy data design leads to low bandwidth usage and the computational units starve because the data is not in place (memory bound). Writing a good algorithm means to understand the data.

MEAN SHIFT IMPLEMENTATION

The EDISON system that was already used in Chapter 6 for analysis was the base for the implementation respectively the porting process. EDISON is build around the mean shift method and offers functionality to filter, segment images and edge detection. The complete software is rather complex having a scripting engine, Graphical User Interface (GUI) and different versions of the filtering step, where each step is introducing some technique to accelerate the filtering. Furthermore the filtering was construed to analyse feature spaces with arbitrarily dimensions, different kernels, weight maps and so on.

Since the focus of this work is on image segmentation only the filtering functions and some segmentation helper functions were taken from EDISON as reference. The image loading, saving, pre-, post processing and the filtering part, which is the most compute intensive part with 99.1% (from Table 6.1) of the run-time, were completely rewritten whereas the clustering, labelling and elimination of spatial regions (see Algorithm 5.3) with in sum 0.9% of the run-time were adopted without change.

As mentioned before there are three algorithms of the filtering step. For the purpose of this work the standard mean shift algorithm was chosen because (1) it is the most compute intensive and (2) it has the best quality output. The other two versions of the algorithm induce a quality loss and use structures to accelerate the algorithm that are not well suited for the GPU. Additionally the compute power of a GPU is more visible with the first version.

7.1 REFERENCE IMPLEMENTATION

The first step was to do a reference implementation on the CPU. The CPU programming framework has better debugging and logging facilities than a GPU. It is not possible to do a system call like `printf`, `fopen` on a GPU which makes debugging hard and logging rather impossible.

The idea behind the reference implementation was to have a running code which is doing the segmentation in the *right* way. The advantage of this is that when porting to the GPU there is always a reference to which

the results of a GPU run can be compared too. This way it is assured that the algorithm is doing the right thing. It will be more crucial when optimisation techniques were used. It has to be verified that they are not breaking the algorithm. How this reference implementation was applied can be seen in Section 7.2. Another positive effect is that now there is a implementation against the GPU, where the run-times of each implementation can be compared and speedups calculated.

The rest of this chapter is split into two parts. The first part describes the host part, its that part which the CPU is executing and the second part describes the device part where the CUDA kernels were run.

7.2 THE HOST PART

The host part takes care of preparing and data provisioning. The first thing to do is to load the image into a appropriate format into the memory. Remembering the lessons learned from the feasibility study in Chapter 4 about global memory access the loaded image should be arranged in such a way that threads getting the data can do that in a coalesced manner.

Luckily the CUDA helper functions library “*cutil*” has image loading function that can align the data in memory for efficient access. The functions in question is `cutLoadPPM4ub()` which reads the Portable Pixel Map (PPM) file which is in RGB (0–255) format and pads a zero to the fourth byte. The effect of this padding is that coalescence and better performance is achieved by accessing 1, 2 or 4 consecutive memory locations simultaneously (see Chapter 8 for more optimisations).

Another thing to consider is that the execution configuration should be configurable and further parameters that are crucial for performance like the number of GPUs, thread configuration, device selection, ...are implemented as command colour line parameters.

The CPU is responsible for calling the CUDA kernels. In this implementation there are three of them, the first one is a kernel that is converting from RGB to LUV colour space, the second is the filtering kernel and the last one is converting from LUV back to RGB. The colour space conversion kernels will not be considered further because there are rather simple and can be looked up in any good computer graphics book. The attention is on the filtering kernel that is explained in the device part (Section 7.3).

Before any kernels are started several input and output buffers were allocated in the **CPU** and **GPU** memory space. Alignment of the buffers is crucial for high bandwidth transfers that's why the allocating functions from **CUDA** are aligning the buffers dependent on the data type per default. Right before the kernel execution the input data is transferred to the **GPU** buffers and after execution the results are transferred back to the **CPU** via the **GPU DMA** engine. This doesn't mean that for each kernel the **CPU** is transferring the data back and forth, this is done only one time. After the first kernel finishes execution the data can reside in the memory of the **GPU** for the next kernel. Memory stays consistent across kernel calls and the subsequent kernel needs the preprocessed data from the previous kernel. This coupling and strong dependency of the kernels or steps was already described in Section 6.3 and effected the design Chapter 6. When the last finishes execution the host can fetch the results for saving and result verification.

Progress & Result Verification

As mentioned in Section 7.1 there exists a **CPU** and a **GPU** implementation of the algorithm. When executing the program it is possible to supply a parameter which triggers the result verification. The images saved by the **CPU** and **GPU** are compared pixel by pixel. For this comparison the great tools from ImageMagick a software suite to create, edit, and compose bitmap images, are used.

The compare program can mathematically and visually annotate the difference between an image and its reconstruction. The tool emphasises pixels that are different between the **CPU** and **GPU** version with a semi-transparent red overlay, whereas a white overlay de-emphasises pixels that are the same between the two versions. An example of such an comparison can be seen in Figure 7.1.

This way it was assured that any modifications or optimisations were done correctly and not changing the natural behaviour of the algorithm.

7.3 THE DEVICE PART

For the device part only the filtering step will be further described. As said before the colour conversion is rather simple and can be looked

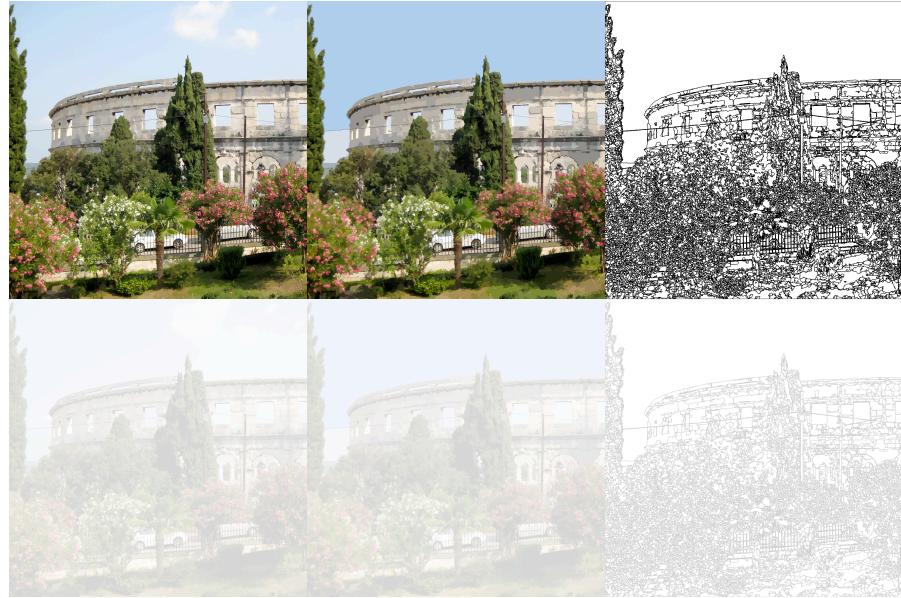


FIGURE 7.1 The first row shows the reference pictures computed by the **CPU**, (1) the filtered, (2) the segmented image and (3) the edges. Below the annotations made by ImageMagick. In this case every pixel computed by the **GPU** matches the reference

up in any good computer graphics book. It can be used with minor modifications on the **GPU**.

The kernel is a single source-code image that runs on each thread on the **GPU**. The first thing that is done in the kernel is to obtain a unique identifier. The identifier is obtained with special functions calls from the hardware. This unique identifier allows different threads to make different decisions during kernel execution. The assignment of the unique id to the hardware threads is deterministic. Thread zero is always the first hardware thread, one can be for sure that actions which have to be done by a certain thread are really done by the corresponding thread in the hardware.

The unique id is as well important for data distribution. In Section 6.8 it was shown that one thread is responsible for a single pixel. So there are so many threads as there are pixels in the image where the threads are numbered from 0-(number of pixels - 1). Every thread having its unique id which is at the same time the position in the array of pixels can now

easily fetch the pixel from global memory for computation.

Each thread takes the LUV values and calculates the mean shift vector until convergence. When a thread finishes execution it saves the convergence point to a second buffer to the same place in the array (unique id) where the thread fetched the input pixel.

For the first shot no other memories were used to improve performance. The first implementation was just to make it work. The next chapter will dig deeper into the hardware and optimise the code to run efficiently. The performance and scalability is evaluated in a subsequent chapter.

7.4 SUMMARY

After an extensive analysis and a good design the effort to implement the algorithm was not so big. The most difficult part was the implementation of the mean shift method in the reference implementation. It had to be verified that the algorithm is working and doing the right stuff.

After that it was crucial to get in touch with CUDA and all dependent libraries and to know what function is exactly doing and how to use it.

Merging the reference implementation with CUDA was pretty easy as CUDA is just an extension to C and only some keywords were introduced that have to be used in the right way.

The programming flow of first moving data to the GPU, processing and then getting the data back to the host was already known, from implementing a ray tracer on the Cell processor. Pretty all accelerators have a similar programming flow.

Luckily there was the reference implementation which was used for result verification not only for the filtering part but as well for the other two kernels where a bug was easily spotted in form of red pixels thanks to the great tools of ImageMagick.

8

OPTIMISATION STRATEGIES

The following chapter will present optimisation strategies which were used to accelerate the mean shift image segmentation. The presented strategies are not only valid for CUDA, they can be applied to many parallel machines which are build after the shared memory or SPMD model. The chapter is divided in several sections. Each section will present an optimisation. There are major and minor optimisation that are examined where mostly major optimisation give higher speedup whereas minor optimisation give lower speedup.

Many of the optimisations are hardware centric, which means that one has to have a good knowledge about the architecture to understand which steps can be undertaken to accelerate the algorithm and which are even possible. Each section will cover the architectural specialty which led to perform the specific optimisation. The first optimisation of the next section is a general rule when trying to accelerate an algorithm, offload compute intensive parts. The succeeding sections respectively optimisations are in chronological order and where applied in that order, identifiable by the ever increasing speedups.

8.1 TEST & BENCHMARK CONFIGURATION

For measuring the run-time the CUDA timers are used. They have a resolution of approximately half a microsecond. The timings are measured on the GPU, and hence they are operating system independent. Further benchmark metrics like bandwidth, coalesced or uncoalesced access to global, shared and local memory or time spent in device functions were reported by cudaprof a profiler for CUDA programs. OProfile was used on the host side for profiling the program.

For the measurement a image with a dimension of 256×256 pixel was used and with the following parameters: $\text{sigmaS} = 7$, $\text{sigmaR} = 6.5$ and $\text{minRegion} = 20$ (minRegion is used for pruning regions smaller than 20 pixel).

8.2 OFFLOAD COMPUTE INTENSIVE PARTS

As shown in Chapter 6 before any analysis is started one has to know if the problem is large enough and if the resulting speedup justifies all the effort that is expended on making a parallel version out of it. This way the most compute intensive parts are identified and these functions are candidates for offloading.

In this case the most compute intensive part is 98.2% of the complete run-time and hence a valuable part to offload. All other parts have no significant part of the run-time but when optimising it can happen that these other part grow in terms of run-time and the before most compute intensive part with a very high percentage drops to a low percentage and hence the percentages of all other parts increase and become very compute intensive.

In other cases the run-time is often distributed among several parts and in this case it should be tried to offload as much as possible of the computational parts to an accelerator as the GPU. In the first place the focus will be on the filter part that is executing in total 99.1% of the run-time.

For the following considerations about speedup the Equation 6.1 will be used to calculate the speedup.

In Chapter 6, Chapter 7 it was shown how to design and to implement the algorithm to offload the important parts to the GPU and the first naive implementation of the mean shift filter in CUDA resulted in a speedup of

$$S = \frac{T_{CPU}}{T_{GPU}} = \frac{9616,77 \text{ ms}}{1932 \text{ ms}} = 4,97762$$

times faster than the CPU version. The reference time was generated from EDISON. It reports timings for filtering and segmentation. The first implementation was just a naive implementation without regarding the hardware restrictions and without exhausting the hardware potential. In Section 6.4 it was shown that in every iteration of calculating the mean shift vector the algorithm fetches colour values for every pixel contained in the search window. The search window is defined by sigmaS the spatial window and sigmaR the range window. For now, only sigmaS is important since that value spans the spatial window over the pixel.

8.3 GLOBAL MEMORY & COALESCING

In Section 6.4 it was shown that the search window is moving unpredictable over the image the main memory is accessed randomly. The issue that raises up is that only a fraction of the bandwidth is utilised. Therefore it is advisable to club or coalesce the access to the global memory.

The GPU can coalesce global memory reads and writes in as few as one transaction, or two transactions in the case of 128-bit words when certain access requirements are met. These requirements are e.g. that a specific access pattern is realised, the address of a 64-byte segment is 64-byte aligned and further more. For a complete list see Chapter 3 of [12]. This requirements are typical for high speed memory not only valid for GPUs rather for many parallel architectures like the Cell or the Niagara CPU.

To satisfy one requirement, the access pattern, it is advisable to use the *float4* data type. It ensures that data is accessed on consecutive memory locations and hence it can be grouped together. After rewriting the algorithm the new speedup was

$$S = \frac{9616,77 \text{ ms}}{1454 \text{ ms}} = 6,614$$

The speedup was really not very satisfying but the GPU has another mechanism which could be used to speedup the algorithm. The input image is read only and the results are written to another structure. Furthermore read pixels from the image are reused in the calculation when the search window is shifted and the access is rather random.

What now proves advantageous is the ability of CUDA to use texture memory for computing. Texture memory can be more faster than device memory because it is first of all read only and it uses a cache for faster access. The image data stored in a texture is optimised in that way that locality is exploited. The texture is perfect match for the algorithm as random access do not harm the bandwidth so much compared to global memory and pixels read in for one search window can be reused for the next search window as they are cached.

Switching to *textures* (glossary textures) yield to a huge speedup of

$$S = \frac{9616,77 \text{ ms}}{250 \text{ ms}} = 38,46708$$

Texture memory is great for algorithms where random access to memory occurs and data read in are used more often than once. Where it is predictable how much and how (access pattern) the data is read in global memory in conjunction with shared memory can be advantageous. But as said at the beginning of this section certain requirements have to be met to achieve a high bandwidth.

8.4 DIVISION INSTRUCTION OPTIMISATION

Every hardware instruction is taking a specific amount of clock cycles to complete. For single-precision floating-point arithmetic instructions like add, multiply or multiply-add take 4 clock cycle to complete. Whereas the single-precision floating-point division takes about 36 clock cycles, which is $9\times$ slower than the common arithmetic instructions. Therefore divisions should be avoided as much as possible to increase instruction throughput.

Having a look at Algorithm 5.2 it can be seen that in the calculation of the mean shift vector divisions occur. The denominator is the size of the search window. In this implementation the size of the search window is not changing and hence a constant. This circumstance can be exploited and the reciprocal can be calculated. The reciprocal can then be substituted by the division and denominator to produce faster multiplications.

For example if it looks like this

```

1 // Determine if inside search window
2 // Calculate distance squared of sub-space s
3 float dl = (luv.x - yj_2) / sigmaR;
4 float du = (luv.y - yj_3) / sigmaR;
5 float dv = (luv.z - yj_4) / sigmaR;
```

LISTING 8.1 Expensive division

one can calculate $rsigmaR = 1.0/\sigma_{\text{R}}$ in advance and everywhere where a division by σ_{R} occurs replace it into a multiplication. The resulting code is then

```
1 // Determine if inside search window
```

```

2 // Calculate distance squared of sub-space s
3 float d1 = (luv.x - yj_2) * rsigmaR;
4 float du = (luv.y - yj_3) * rsigmaR;
5 float dv = (luv.z - yj_4) * rsigmaR;

```

LISTING 8.2 Division turned into fast multiplication

With this technique the kernel code is free of divisions except only one that cannot be eliminated since the value is changing throughout the calculations. This optimisation yielded a speedup of

$$S = \frac{9616,77 \text{ ms}}{193 \text{ ms}} = 49,82782$$

8.5 EXECUTION CONFIGURATIONS

It is important to check the best execution configuration. Each execution configuration has its benefit. Some can exploit the bandwidth to the global memory some can benefit from the cache of the texture. Furthermore it is important to keep the hardware busy to hide memory latencies. Busy means that one should design the application to use threads and blocks so that the hardware can balance the work across the multiprocessors.

A good execution configuration can hide latency arising from register dependencies , provide optimal computing efficiency and facilitate coalescing to global memory. Not every execution configuration is possible because resources are limited and too many threads can eat up one resource where the other resource is not used. The resource meant here are the shared, local memory and the register usage. See the CUDA programming guide [49] for a in depth description about the effects of execution configuration.

Several test showed that the best execution configuration for this algorithm is 2×64 threads and the new speedup resulting out of this new configuration was

$$S = \frac{9616,77 \text{ ms}}{171 \text{ ms}} = 56,23842$$

Another positive effect of a configuration that is multiple of the warp size is that the scheduler can work more efficiently because a warp consists out of 32 threads.

8.6 NATIVE DATA TYPES

The CUDA Programming Guide [49] highly recommends the use of the *float* type for single-precision floating-point code. Additionally it is recommended to use the single-precision floating-point mathematical functions. One should use the *float* data type where ever possible. The GPUs are highly optimised for floating point calculations. After converting all integer calculations to floating point operations another speedup was achieved.

$$S = \frac{9616,77 \text{ ms}}{165 \text{ ms}} = 58,28345$$

On current GPU hardware integer operations can be very expensive in terms of clock cycles. The situation can change with future GPU architecture check against *Fermi* the new architecture from NVIDIA [13].

8.7 AVOID BRANCH DIVERGENCE

On a architecture with such high throughput of calculations per cycle it is preferred to calculate values rather then generating them through *if* and *else* statements. Further problems arise when control flow depends on the thread id which means that they are divided into groups that are executing different execution paths. In this case the execution gets serialised. This can happen when a thread has found a density maximum and all the other threads are still calculating. When the active threads finish the calculation they converge back to the same execution path as the idle threads and they finish together.

Therefore the remaining *if* and *else* statements where arranged in such a way so that a thread is exiting early from the loop and avoiding unnecessary calculations. This leads of course to higher branch divergence but the execution time is lower.

The speedup achieved from this optimisation was

$$S = \frac{9616,77 \text{ ms}}{115 \text{ ms}} = 83,62408$$

8.8 SHARED MEMORY

The optimisation guides state that one should use shared memory to avoid redundant transfers from global memory. Shared memory is much faster than global and local memory assumed that there are no bank conflicts. The GPU shared memory is divided into so called banks for simultaneous access. If two threads are accessing the same bank the access is serialised and hence can be very slow if not only two but far more threads are accessing the same bank [12].

In Section 6.4 it was shown that the search window is moving unpredictable over the image and hence preloading of data into the shared memory is not possible. Furthermore if two or more search windows from different threads overlap it is obvious that the access to the pixel would lead to bank conflicts. After many tryouts this optimisation was discarded.

8.9 KNOW THE ALGORITHM

This section is not related to the hardware more related to the software part of the algorithm. After all the other optimisations done it was obvious that there was not more room for many optimisations that target the hardware. The focus went on to the algorithm and its peculiarity when filtering the image.

The idea was to examine how many iterations in average the algorithm needs to filter the image. In the original code of EDISON there is variable named iterationCount which limits the iteration count to 100 per pixel. The comment in the source code says “*Iteration count is for speed up purposes only – it does not have any theoretical importance*” that brought up a question, why it is needed anyway when the mean shift algorithm converges [10]. For this the source code was modified in that way that for each pixel the iteration count was reported. A precise look at the generated numbers showed that some pixels have iteration counts of 100.

To have an clue if its worth to continue the research into this direction several tests were done. Looking at the iteration count it sticks out that most of the pixel are finished after about 10 iterations, it was strange to see that several pixel were iterating till 100.

The approach taken was to limit the iteration count not to 100 but to 10 and 50. Setting the limit to 10 and executing the CPU and GPU version the speedup was

$$S = \frac{7369 \text{ ms}}{29 \text{ ms}} = 83,62408$$

Setting the limit to 50 and executing the CPU and GPU version again the speedup was

$$S = \frac{10373 \text{ ms}}{82 \text{ ms}} = 126,5$$

In the case where the limit is 10 the GPU could achieve a speedup of $254,10344 \times$ because all multiprocessors are busy then and are not idle as in the case when the limit is 100 and just a few threads are calculating and the rest idles. The algorithm is as fast as the slowest thread which is iterating until 100.

A candidate with iteration count of 100 is picked up from the iteration list and examined further. For this the magnitude squared (the length) of the mean shift vector is analysed over the iterations. When the mean shift vector approaches the convergence point the magnitude becomes smaller and smaller until it falls under a small number. In this implementation this small number is $\epsilon = 0.01$. Then it can be said that the mean shift vector has reached the basin of attraction.

Attractor

Dynamical systems evolve to a specific set called an attractor after many iterations. Dynamical systems are described either by differential equations or by difference equations. A difference equation is a specific type of recurrence relation where each term of a sequence is defined as a function of the preceding terms.

Such an attractor can be a point (fixed point), curve (limit cycle), manifold or a complicated set with a fractal structure known as a strange attractor, see Sprott [51] for impressive pictures of strange attractors.

But not only dynamical systems can lead to such an attractor, furthermore the effect of quantisation, the limited precision of the float data type can lead to an attractor.

The above terminology comes from chaos theory where the behaviour of dynamical and iterative functions systems are described. In Section 2.3.7 several books and articles were presented for further information.

The chaos theory was mentioned because it can happen very often without knowledge that some algorithms are iterating and iterating and the stopping criterion is never reached. In the mean shift algorithm case the upper bound was set to 100 without mentioning why. If the algorithm converges an that is proven, why do we have a upper limit anyhow? Because in this case there are attractors as well. All the densities where the mean shift vector moves to are attractors, fixed points. Small perturbations of the vector lead to the same density.

But there are also attractors where the mean shift vector is moving to and it is not the final density.

```

1 10992 - 0.00000 - 1
2 10992 - 2.52029 - 2
3 10992 - 2.52029 - 3
4 10992 - 2.52029 - 4
5 10992 - 2.52029 - 5
6 ...
7 10992 - 2.52029 - 100

```

LISTING 8.3 Magnitude squared of pixel 10992

Examining the sequence of the magnitude squared for pixel 10992 in Listing 8.3 one can easily see that after some iterations the magnitude is fixed to 2.52029. The algorithm is stuck at this value and will not move away from this attractor. An attractor with a single value is a fixed point.

At the beginning of the section it was shown that if the iteration count of each pixel is dropped to a lower level the performance increases. Further examination showed that several pixel have attractors that were limit cycles. Which means that the calculation of the magnitude squared is periodic. The period in this case is from a period-2 up to a period-8 limit cycle.

For visual feedback the source image in Figure 5.4a was taken and for each pixel the iteration count was visualised. The Figure 8.1 shows from blue to red, where red values represent high iteration count, the iterations for each pixel in a 128×128 big picture.

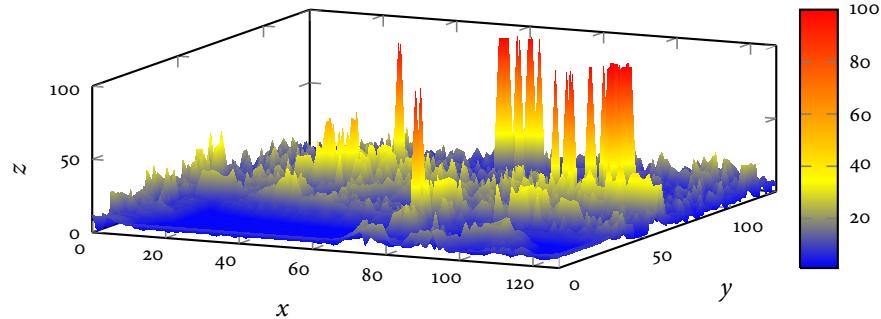


FIGURE 8.1 Iteration count for each pixel

It can be seen that there are several pixels that iterate to 100. To circumvent this problem a simple limit cycle detection algorithm was implemented to prevent that several threads are blocking the termination of the algorithm.

In the first place a period-4 limit cycle detection algorithm was implemented where a vector was saved with the last four values and compared to the new calculated value. If any value from the vector matches the new value the thread terminates and finishes the computation.

The period-4 limit cycle detection led to a speedup of

$$S = \frac{9616,77 \text{ ms}}{99 \text{ ms}} = 97,13909$$

Extending the limit cycle detection to period-8 limit cycle detection a speedup of

$$S = \frac{9616,77 \text{ ms}}{87 \text{ ms}} = 110,53758$$

was achieved. Surely one cannot be sure to have limit cycles with a higher period but the first problem is the register pressure. Because the compiler has to keep the saved values of the previous iterations in the registers for comparison the registers cannot be used for the other

calculations. The CUDA compiler tries to keep as much as possible in the registers. Secondly comparisons are not very fast as computation and having to compare 16 or even 32 values makes no sense because than the prevention of limit cycles would cost more performance than just iterating to 100. The following series of Figure 8.2, Figure 8.3, Figure 8.4 show visually the effect of limit cycle detection and how all iterations to 100 vanish.

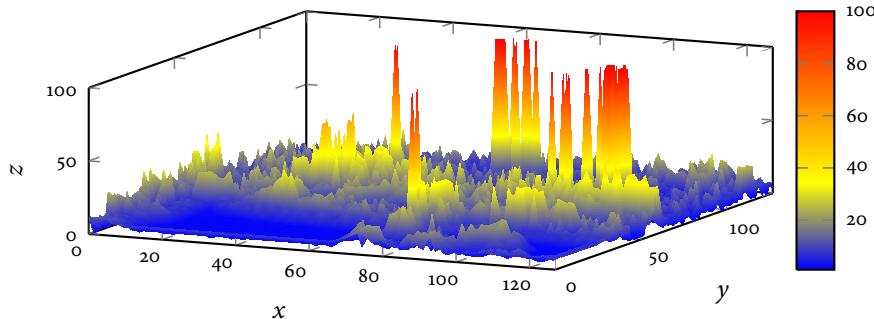


FIGURE 8.2 Period-0 limit cycle detection

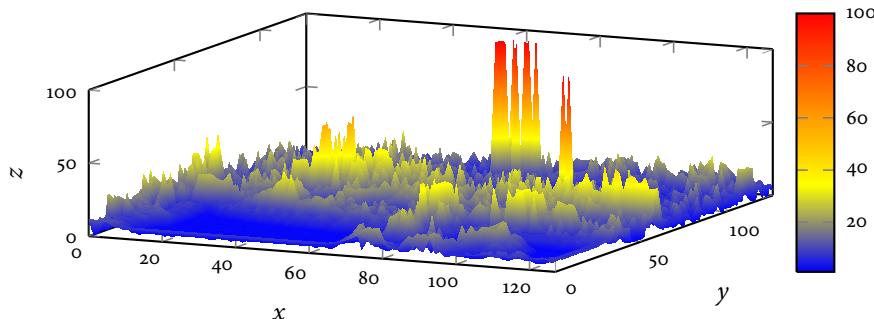


FIGURE 8.3 Period-4 limit cycle detection

8.10 UNROLLING LOOPS AND MULTIPLICATIONS

The last improvement that was made is to unroll loops and complicated multiplications. The idea behind this is that the compiler has more chances to schedule the instructions and reorder them in a proper way.

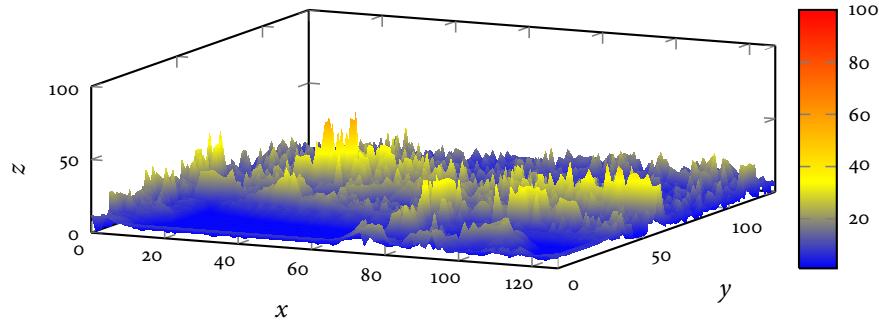


FIGURE 8.4 Period-8 limit cycle detection

This is important because there are some instructions, e.g. `fmadd`, that are executing two floating point operations rather than just one floating point operation per 4 clock cycles and unrolled multiplications are more *clear* to the compiler.

After unrolling the multiplications and loops a final speedup of

$$S = \frac{9616,77 \text{ ms}}{73 \text{ ms}} = 131,73657$$

was achieved.

8.11 SUMMARY

The naive implementation has already led to an speedup of $5\times$. But putting some knowledge and energy into the optimisation one can achieve a speedup of $131,73657\times$. The optimisation was not only done on hardware but also on the algorithm side. It is not enough to know some programming language very good the hardware on which its run is more important. No programming language will expose all hardware features directly to the developer, the developer has to know about the hardware and its features. Furthermore it is crucial to understand the data and how it is accessed. The biggest speedup was achieved when going from global memory to texture memory. The access of data fits perfectly the texture memory and hence the algorithm can take full advantage of it.

After all the optimisations and a recent speedup it is now important to look after the performance and scalability of the algorithm. The next chapter is going to tackle this.

PERFORMANCE & SCALABILITY

To evaluate the performance and the scalability of the implemented mean shift algorithm several benchmarks will be made. Firstly the algorithm is tested against different Image-sizes. After that a multiple GPU version is evaluated and tested with the same Image-sizes as in the previous test. Lastly the hardware parameters of the GPU like the core clock and memory clock are manipulated to get a clue if the algorithm is memory or computational bound. In the case that the algorithm is memory bound one should try to decrease the communication and if the algorithm is computational bound a restructuring of the algorithm could help here. All in all its interesting to see how an algorithm behaves at different circumstances. In general all runs were performed 20 times in a row and the mean was taken as the final result.

9.1 VARYING THE IMAGE-SIZE

The first benchmark varies the Image-size from 128×128 pixels to 2688×2688 pixels. The image side length is incremented by 128 pixels. The Figure 9.1 shows the GPU run-time and the speedup compared to the CPU run-time. The CPU run time was not included in Figure 9.1 because the range of CPU values is $100\times$ larger and the GPU values would not be visible. Its interesting to see that the speedup is increasing faster at the

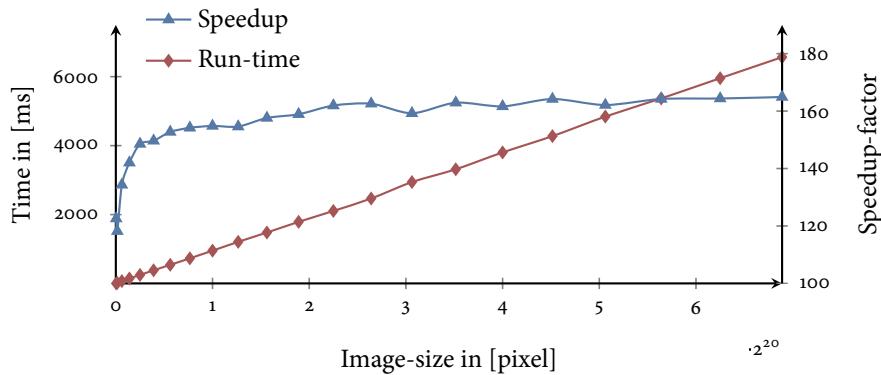


FIGURE 9.1 GPU run-time and speedup with increasing image-size

beginning and than to level off at about 160. This is no surprise because every software respectively hardware has a ramp-up phase. This ramp-up phase, where no computation is performed, often involves allocation and initialisation of data and hardware and additionally on this platform the movement of data from the host to the GPU buffers. With smaller Image-sizes the time for the ramp-up phase is a significant amount of the complete run-time. With bigger and bigger Image-sizes the amount of the ramp-up phase to the complete run-time becomes smaller and hence the speedup increases.

9.1.1 Linearity

Another fact to consider is how the algorithm is scaling with increasing image sizes. In Chapter 6 it was shown that the algorithm exposes linear complexity $O(n)$ and it is interesting to see how the implemented algorithm behaves in terms of linearity. An algorithm is linear when the two metrics for measuring the performance here the Image-size and the run-time are proportional. The Image-size is proportional to the run-time if they have a constant ratio. If one variable increases by a factor x than the proportional variable increases by that factor too.

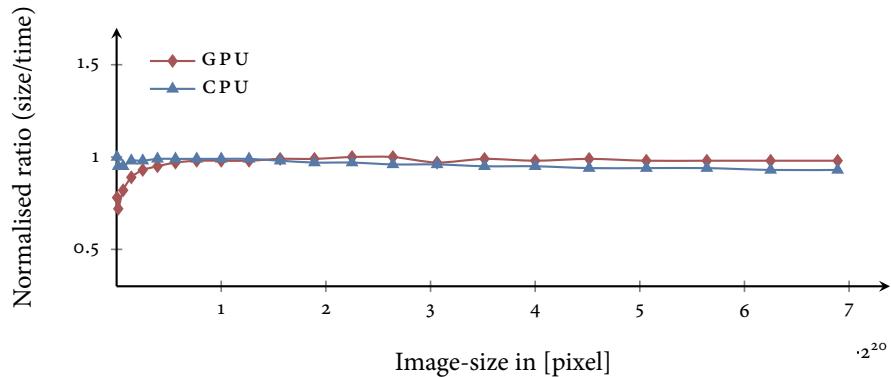


FIGURE 9.2 Normalised size/time ratios for the CPU and GPU

The Figure 9.2 shows the normalised Image-size per time ratios in the interval $[0, 1]$. As it can be seen the ratio for the CPU and the GPU are increasing similar to the speedup because of the before mentioned ramp-up phase. The ratio increases until it reaches the right endpoint

of the interval. If the Image-size is doubled the run-time is doubled (Image-size \propto run-time) which means the algorithm scales linearly.

9.2 MULTIPLE GPUS

The previous section analysed how the implemented algorithm scales over the Image-size. Another important attribute is how the algorithm scales if another GPU is added for computation. A second identical GPU was used to generate the results shown in Figure 9.3.

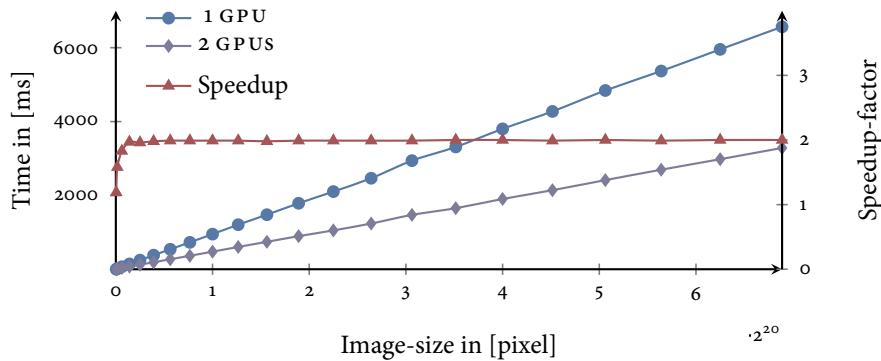


FIGURE 9.3 Multiple GPUs run-time and speedup with increasing image-size

As stated in the previous sections the ramp-up phase is responsible that the algorithm is not scaling with small Image-sizes. Its again clearly identifiable in the Figure 9.3. It's safe to assume that the multiple GPU run-times are proportional to the Image-size as in Figure 9.2 seen for one GPU since the run-times are $2\times$ faster but the ratio will nevertheless still be a constant.

This section showed that the algorithm scales perfectly with multiple devices and it would be no surprise to see a speedup of $3\times$ if a third GPU were added to the system. Additionally the Figure 9.4 shows the speedups of 1 GPU and 2 GPUs compared to the CPU run-times.

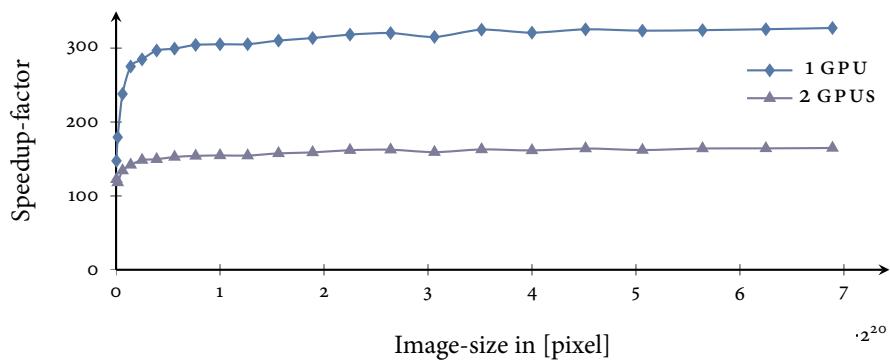


FIGURE 9.4 GPU speedups compared to the CPU run-times with increasing image-size

9.3 OVER-CLOCKING THE GPU

The GPU has three clocks that can be over-clocked to achieve higher performance. The first is the core clocks, the second clock is the memory clock and the third clock is the shader clock where the shader clock typically moves synchronically with the core as they work on a set ratio. For example the used *GeForce 8800 GTS 512* features a 650 Megahertz (MHz) core, a 1620 MHz shader and a 970 MHz memory clock. This means this particular graphics card uses a core to shader clock multiplier of 2.5×. Increasing the core clock means increasing the shader clock at the same time. Therefore only the core and memory clock will be considered for the experiments.

The reason why someone would over-clock a GPU is at first hand to increase the performance and on the second hand to find out if an algorithm is computational or memory bound by experiment.

Therefore several experiments will be undertaken. One experiment will involve the change of the core clock, second the change of the memory clock and lastly the combination of both clocks. All experiments were done with a free tool, *NVClock*¹. NVClock is a small utility that allows to over-clock NVIDIA based GPUs and adjust the fan speed. This is important since the GPU will run out of specification and could overheat. The experiments will begin with changing the memory clock and

¹ <http://www.linuxhardware.org/nvclock/>

evaluation of the results.

9.3.1 Increasing the Memory Clock

For the first experiment the memory clock will be changed by steps of 100 MHz. As a starting point the default 972 MHz of the memory clock will be used and the core clock will be fixed at the default of 648 MHz. Experiments have shown that the maximal achievable memory clock is 1101.600 MHz. All clocks were read out from hardware with *NVClock*. Additionally it must be noted that all frequency settings were not accepted one-on-one which means an increase per software of 100 MHz doesn't mean an increase of 100 MHz in hardware due to hardware restrictions. The Table 9.1 shows the frequency set in software and the resulting real frequency in hardware.

Software [MHz]	Hardware [MHz]
970	972
980	987
990	994
1000	999
1100	1102

TABLE 9.1 Memory clock frequencies set in software and the resulting hardware frequencies

The test was performed with three different Image-sizes: 256×256 pixel, 1024×1024 pixel and 2048×2048 pixel. The Figure 9.5 shows the run-time dependent on the memory clock.

A straight line indicates that the algorithm is not memory bound. It does not make any difference how fast the memory one make the run-time will not decrease. Additionally it can be concluded that the algorithm is computational bound. This will be proven in the next section when the core clock is manipulated. The image sizes 1024×1024 and 2048×2048 pixel show the same behaviour. For the complete list of numbers see Table 9.2.

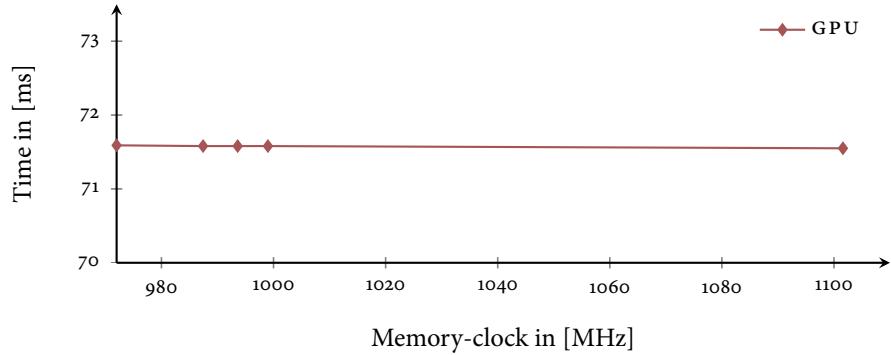


FIGURE 9.5 Run-time with increasing memory clock on an image with 256×256 pixel

MEMCLK	256	1024	2048
972	71.59	954.31	3796.36
987.43	71.58	955.15	3796.79
993.6	71.58	955.07	3795.24
999	71.58	955.32	3798.35
1101.6	71.55	954.96	3795.84

TABLE 9.2 Various memory clocks and the corresponding run times

9.3.2 Increasing the Core Clock

For the second experiment the core clock will be changed by steps of 100 MHz and as a dependency the shader clock as well. As a starting point the default 648 MHz of the core clock will be used and the memory clock will be fixed at the default of 972 MHz because the previous section has shown that the algorithm is not memory bound and the increase of the memory clock has no effect on the run-time. Experiments have shown that the maximal achievable core clock is 783 MHz. All clocks were read out from hardware with *NVClock*. The Table 9.3 shows the frequency set in software and the resulting real frequency in hardware.

The test was performed with three different Image-sizes: 256×256 pixel, 1024×1024 pixel and 2048×2048 pixel. The Figure 9.6 shows the run-time dependent on the core clock. Additionally the normalised

Software [MHz]	Hardware [MHz]
650	648
660	663
670	670
680	679
690	691
700	702
710	713
720	720
730	729
740	738
750	756
760	756
770	771
780	783

TABLE 9.3 Core clock frequencies set in software and the resulting hardware frequencies

ratio between the differences of clock and ratio is depicted as well.

It's visible that the run-time is decreasing proportionally with the increase of the core-clock. This means that the algorithm is heavily computational bound. The run-time is determined in this implementation solely by the speed of the GPU. One way to get over this barrier is to (1) minimize the instructions executed by the GPU by restructuring the code or (2) adopting another algorithm for finding feature point that fall into the search window, see Chapter 10 for enhancements that could be made to accelerate the search. Additionally, the Image-sizes 1024×1024 and 2048×2048 pixel show the same behaviour. For completeness the next section shows the case when two GPUs are over-clocked.

Multiple gpus over-clocked

It's no surprise to see almost a identical diagram compared to the previous section. As the algorithm scales well with multiple GPU (see Section 9.2) and with the frequency of the core clock one can assume to see the same sequence of run-times just cut by half. The Figure 9.7

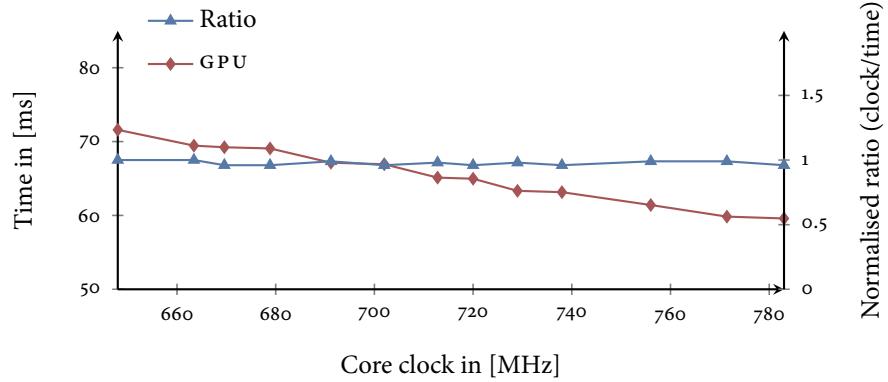


FIGURE 9.6 Run-time with increasing core clock and fixed memory clock on an image with 256×256 pixel with one GPU

shows that, additionally attention should be paid to the shape of the curve as its the same as in Figure 9.6.

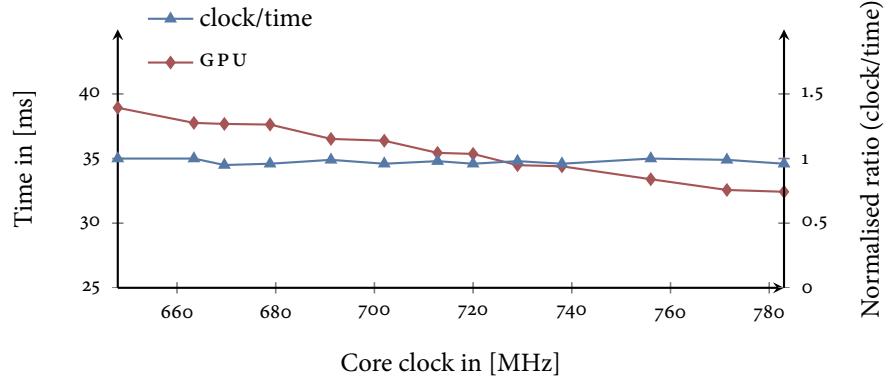


FIGURE 9.7 Run-time with increasing core clock and fixed memory clock on an image with 256×256 pixel with two GPU

9.4 FINAL SPEEDUP

The previous sections showed the how the algorithm compares to the CPU over different work-sizes. Disregarding the ramp-up phase a single GPU achieved a speedup of about $160\times$. Computing with two GPUs

the resulting speedup was of about $320\times$. This showed that a suitable algorithm can achieve reasonable speedups compared to traditional CPUS.

Additionally the scalability over the work-sizes, an additional GPUs and frequency were analyzed and depicted. The result is clear, the algorithm scales linearly over all three before mentioned parameters.

The speedup for an image of 2688×2688 pixel segmented by one overclocked GPU is

$$S_{single}(n) = \frac{T_{CPU}}{T_{GPU}} = \frac{1082570, 14 \text{ ms}}{5465, 28 \text{ ms}} = 198,08136 \quad (9.1)$$

and the final speedup by two overclocked GPUs is

$$S_{multi}(n) = \frac{T_{CPU}}{T_{GPU}} = \frac{1082570, 14 \text{ ms}}{2736, 14 \text{ ms}} = 395,65597 \quad (9.2)$$

That's an impressive number about $400\times$ faster than the CPU version. Of course one has to take into account that the used algorithm is highly suitable for the GPU. Nevertheless this chapter showed to which extent algorithms can be accelerated on the GPU.

CONCLUSIONS & FURTHER WORK

The field of GPU computing is nowadays very broad. Starting from dense linear algebra, n-body, finite difference over graphics, image processing to computational modelling in science, engineering and finance the GPU is existent in all this fields. A lot of algorithms profited from using the GPU as an accelerator and many will profit too. There is a high demand on performance and computational power for scientific algorithms. Developers search for new architectures to solve problems in shorter time and the GPU is one promising candidate because of its incredible raw power.

New architectures meant new programming models and in the case of the GPU the first programming model was not build for general purpose computing and developers were accidentally forced to use 3D APIs (OpenGL¹, DirectX²) to accomplish task like matrix multiplication or fast fourier transformation. These APIs implied some limitations and raised some myth for the new era of GPGPU when the switch from fixed pipelines to fully programmable pipelines happen.

Developers were concerned about the programming models will still have the limitations as the former had. The feasibility study showed that all the odds and ends are not more valid for the new architectures and programming models. This new sophisticated programming environments, like CUDA for the NVIDIA GPUs, offer a complete, easy and new interface to the hardware to exploit parallelism as much as possible that make GPGPU more feasible.

With the advent of these new programming environments more and more developers are using GPUs for general purpose computing. Many of them come from a world where everything is sequential and struggle with thinking in parallel. Developers often only needed to compile their code with another compiler to port the code to another architecture. For example porting code from x86 to PowerPC or x86 to Scalable Processor ARChitecture (SPARC) was easy enough if it was standard C or C++ without any special operating system calls. But the situation for parallel processors like the Cell, Niagara and here the GPU is completely different.

¹ <http://www.opengl.org>

² <http://www.microsoft.com>

One has to identify the critical parts and explicitly offload tasks to the GPU. It is obvious that this is not the only step to do. Furthermore there are other facts which have to be considered, like load-balancing, deadlocks, concurrency, cache coherency and other. A workflow was presented throughout this thesis by the means of an algorithm how all this necessary steps were undertaken.

The porting step is not the last when considering high performance code on a GPU. The next big thing is to optimise the code to some extent to run efficiently on the hardware. The main cause of inefficient code is bad data design. It is the biggest problem of an algorithm when trying to attain maximum performance. If the algorithm is memory bound and not computational bound the limiting factor or bottleneck is the bandwidth to the memory. Good data design is essential for good, performant code. The data and how it is used must be known in order to effectively design and optimise a application for a GPU or other highly parallel systems like the Cell processor [28]. A simple computationally bound algorithm can be easily designed in a way that it becomes memory bound if data arrangement and access is neglected by the developer. A good data design is essential for any massively parallel system.

10.1 FURTHER WORK

For the mean shift algorithm there exist several acceleration techniques. These techniques could be applied to the algorithm implemented in this thesis to accelerate the algorithm to some extent. There were several papers that describe the techniques in depth. All of them have one thing in common, they degrade the quality of the segmented image whereas the algorithm used throughout this thesis is the most accurate but also the slowest. But not all techniques imply a very high quality loss, often not seen by the human eye and can be used for tracking or cutout filters for image processing.

Further work could also extend the existing algorithm to segment not only pictures but also videos where tracking is a big field of interest. For image processing software different filters could be implemented like “paint by numbers” or “comic effect” and many more. Photoshop Creative Suite 4 (cs4) from Adobe e.g., is making heavy use of the GPU already and has a plugin architecture for developing custom filters.

Fermi

Fermi is a announced but yet not released new groundbreaking architecture from NVIDIA. Many of the limitations with the current hardware like non Error-Correcting Code (ECC) memory, single kernel, no C++ support, small main memory, register pressure and so on are being improved or completely eliminated. It features more shared memory, register, streaming multiprocessors and many more especially a design to meet the requirements of the HPC and general purpose computing communities. It would be worthwhile to run the algorithm on this new architecture to see how the promised scaling without recompiling is really doing. But for more performance it should be recompiled and adapted to all the new features available.

The problem with new hardware architectures is they have to be accepted by the developers to succeed. One way to achieve this is to have an easy programming model which adapts easily to common apis and existing systems. There are rarely branches of industry which want to, or are able to rewrite their code for bleeding edge architectures like the cell or GPUs if the outcome is only 2x faster code. In this case many just extend their cluster with another bunch of cheap systems. A factor of 10 is an undocumented limit when to invest in other, faster hardware.

OpenCL, CUDA and other programming model make a promising start in the field of massively parallel computers. More and more developers will or have to adapt to parallel architectures and parallel programming. Such programming models make the life easier and can help to understand parallelism.

BIBLIOGRAPHY

- [1] IkkJin Ahn, Michael Lehr, and Paul Turner. Image processing on the gpu. Technical report, University of Pennsylvania, February 2005. (Cited on page 12.)
- [2] Arthur Appel. Some techniques for shading machine renderings of solids. In *AFIPS 1968 Spring Joint Computer Conference Proceedings*, volume 32, pages 37–45, 1968. (Cited on pages 11, 28, and 119.)
- [3] A. J. Bernstein. Program analysis for parallel processing. *IEEE Trans. on Electronic Computers*, EC-15:757–762, October 1966. (Cited on page 53.)
- [4] Cynthia Bruyns and Bryan Feldman. Image processing on the gpu: a canonical example. Technical report, University of California Berkeley, 2005. (Cited on page 13.)
- [5] Ian Buck, Paulius Micikevicius, Scott Morton, John Stone, Jim Phillips, and Patrick Legresley. Supercomputing 2008 tutorial intro. URL http://www.gpgpu.org/sc2008/M02-01_Intro.pdf. (Cited on pages 10, 18, 19, and 112.)
- [6] Ian Buck, Tim Foley, Daniel Horn, Jeremy Sugerman, Kayvon Fatahalian, Mike Houston, and Pat Hanrahan. Brook for gpus: stream computing on graphics hardware. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 777–786, New York, NY, USA, 2004. ACM Press. (Cited on page 9.)
- [7] Chas. Data-parallel computing. *Queue*, 6(2):30–39, 2008. ISSN 1542-7730. doi: <http://dx.doi.org/10.1145/1365490.1365499>. URL <http://dx.doi.org/10.1145/1365490.1365499>. (Cited on pages 8 and 117.)
- [8] Shuai Che, Michael Boyer, Jiayuan Meng, David Tarjan, Jeremy W. Sheaffer, and Kevin Skadron. A performance study of general-purpose applications on graphics processors using cuda. *Journal of Parallel and Distributed Computing*, In Press, Corrected Proof. doi: <http://dx.doi.org/10.1016/j.jpdc.2008.05.014>. URL <http://dx.doi.org/10.1016/j.jpdc.2008.05.014>

- dx.doi.org/10.1016/j.jpdc.2008.05.014. (Cited on pages 16, 17, and 111.)
- [9] Yizong Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, August 2002. doi: 10.1109/34.400568. URL <http://dx.doi.org/10.1109/34.400568>. (Cited on pages 36, 38, and 40.)
 - [10] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002. doi: 10.1109/34.1000236. URL <http://dx.doi.org/10.1109/34.1000236>. (Cited on pages 33, 38, 42, 45, 49, and 75.)
 - [11] Dorin Comaniciu and Peter Meer. Distribution free decomposition of multivariate data. In *SSPR ’98/SPR ’98: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 602–610, London, UK, 1998. Springer-Verlag. ISBN 3-540-64858-5. URL <http://portal.acm.org/citation.cfm?id=673222>. (Cited on page 37.)
 - [12] Nvidia Corporation. Nvidia cuda c programming best practices guide. Technical report, 2009. (Cited on pages 71 and 75.)
 - [13] Nvidia Corporation. Nvidia’s next generation cuda compute architecture: Fermi. 2009. (Cited on page 74.)
 - [14] Scott Draves and Erik Reckase. The fractal flame algorithm. Technical report, November 2008. (Cited on pages 13 and 121.)
 - [15] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000. ISBN 0471056693. URL <http://www.amazon.com/exec/obidos/redirect>tag=citeulike07-20&path=ASIN/0471056693>. (Cited on page 33.)
 - [16] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *SIGGRAPH ’02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266, New York, NY, USA, 2002.

- ACM Press. doi: <http://dx.doi.org/10.1145/566570.566574>. URL <http://dx.doi.org/10.1145/566570.566574>. (Cited on pages 12 and 120.)
- [17] Kayvon Fatahalian and Mike Houston. Gpus: a closer look. *Queue*, 6(2):18–28, 2008. ISSN 1542-7730. doi: <http://dx.doi.org/10.1145/1365490.1365498>. URL <http://dx.doi.org/10.1145/1365490.1365498>. (Cited on pages 10 and 118.)
- [18] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. In *SIGGRAPH ’02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 249–256, New York, NY, USA, 2002. ACM. ISBN 1-58113-521-1. doi: <http://dx.doi.org/10.1145/566570.566573>. URL <http://dx.doi.org/10.1145/566570.566573>. (Cited on pages 12 and 120.)
- [19] M. Flynn. Some computer organizations and their effectiveness. *IEEE Trans. Comput.*, C-21:948+, 1972. URL http://en.wikipedia.org/wiki/Flynn's_taxonomy. (Cited on page 16.)
- [20] Andrew S. Glassner. *An introduction to Ray Tracing*. Academic Press, 1989. (Cited on page 28.)
- [21] John L. Gustafson. Reevaluating amdahl’s law. *Commun. ACM*, 31(5):532–533, 1988. ISSN 0001-0782. doi: <10.1145/42411.42415>. URL <http://dx.doi.org/10.1145/42411.42415>. (Cited on page 52.)
- [22] Mark Harris. Mapping computational concepts to gpus. In *SIGGRAPH ’05: ACM SIGGRAPH 2005 Courses*, New York, NY, USA, 2005. ACM. doi: <http://dx.doi.org/10.1145/1198555.1198768>. URL <http://dx.doi.org/10.1145/1198555.1198768>. (Cited on pages 10, 11, 118, and 119.)
- [23] Donald Hearn and M. Pauline Baker. *Computer Graphics*. Prentice Hall, 1994. (Cited on page 28.)
- [24] Henrik W. Jensen. *Realistic Image Synthesis Using Photon Mapping*. AK Peters, Ltd., July 2001. ISBN 1568811470. URL <http://www.amazon.ca/exec/obidos/redirect?1568811470>

- tag=citeulike09-20&path=ASIN/1568811470. (Cited on pages 11 and 119.)
- [25] Emmett Kilgariff and Randima Fernando. The geforce 6 series GPU architecture. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Courses*, New York, NY, USA, 2005. ACM. doi: <http://dx.doi.org/10.1145/1198555.1198767>. URL <http://dx.doi.org/10.1145/1198555.1198767>. (Cited on pages 10 and 118.)
 - [26] P. Kongetira, K. Aingaran, and K. Olukotun. Niagara: a 32-way multithreaded sparc processor. *Micro, IEEE*, 25(2):21–29, 2005. doi: <http://dx.doi.org/10.1109/MM.2005.35>. URL <http://dx.doi.org/10.1109/MM.2005.35>. (Cited on pages 8 and 117.)
 - [27] Zvonko Krnjajic. Untersuchung und implementierung von algorithmen zur zerlegung großer zahlen. Technical report, University of applied sciences esslingen, June 2005. (Cited on pages 12, 119, and 120.)
 - [28] Zvonko Krnjajic. A raytracer implementation as an example for a cell broadband engine optimized application architecture. Master’s thesis, University of Applied Sciences Esslingen, July 2006. URL #. (Cited on pages 11, 28, 92, and 119.)
 - [29] Aaron Lefohn, Joe M. Kniss, Robert Strzodka, Shubhabrata Sen-gupta, and John D. Owens. Glift: Generic, efficient, random-access gpu data structures. *ACM Transactions on Graphics*, 25(1):60–99, January 2006. URL <http://graphics.cs.ucdavis.edu/~lefohn/work/glift/>. (Cited on page 9.)
 - [30] Jian-Ming Li, Xiao-Jing Wang, Rong-Sheng He, and Zhong-Xian Chi. An efficient fine-grained parallel genetic algorithm based on gpu-accelerated. In *Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference on*, pages 855–862, 2007. doi: <http://dx.doi.org/10.1109/NPC.2007.108>. URL <http://dx.doi.org/10.1109/NPC.2007.108>. (Cited on pages 12 and 120.)
 - [31] Fajie Lii and Reinhard Klette. Adaptive mean shift-based clustering. Technical report. URL <http://www.mi.auckland.ac.nz/tech-reports/MItech-TR-20.pdf>. (Cited on page 42.)

- [32] Paweł Maciol and Krzysztof Banas. Testing tesla architecture for scientific computing: the performance of matrix-vector product. volume 3, 2008. (Cited on pages 16 and 111.)
- [33] Rafal Mantiuk, Karol Myszkowski, and Hans P. Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Percept.*, 3(3):286–308, 2006. ISSN 1544-3558. doi: <http://dx.doi.org/10.1145/1166087.1166095>. URL <http://dx.doi.org/10.1145/1166087.1166095>. (Cited on pages 12 and 120.)
- [34] Rafal Mantiuk, Grzegorz Krawczyk, Radosław Mantiuk, and Hans P. Seidel. High dynamic range imaging pipeline: perception-motivated representation of visual content. volume 6492. SPIE, 2007. URL <http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=PSISDG006492000001649212000001&idtype=cvips&gifs=yes>. (Cited on pages 12 and 120.)
- [35] Michael McCool and Stefanus Du Toit. *Metaprogramming GPUs with Sh*. AK Peters Ltd, 2004. ISBN 1568812299. URL <http://portal.acm.org/citation.cfm?id=1024226>. (Cited on page 9.)
- [36] Patrick McCormick, Jeff Inman, James Ahrens, Jamaludin M. Yusof, Greg Roth, and Shareen Cummins. Scout: a data-parallel programming language for graphics processors. *Parallel Comput.*, 33(10-11):648–662, September 2007. ISSN 0167-8191. doi: <10.1016/j.parco.2007.09.001>. URL <http://dx.doi.org/10.1016/j.parco.2007.09.001>. (Cited on page 9.)
- [37] Sally A. McKee. Reflections on the memory wall. In *CF ’04: Proceedings of the 1st conference on Computing frontiers*, New York, NY, USA, 2004. ACM Press. ISBN 1581137419. doi: <http://dx.doi.org/10.1145/977091.977115>. URL <http://dx.doi.org/10.1145/977091.977115>. (Cited on pages 7 and 117.)
- [38] Rahul Narain. Image processing on gpus. Technical report, University of North Carolina Chapel Hill, NC 27599-3175, March 2007. (Cited on page 13.)

- [39] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda. *Queue*, 6(2):40–53, 2008. ISSN 1542-7730. doi: <http://dx.doi.org/10.1145/1365490.1365500>. URL <http://dx.doi.org/10.1145/1365490.1365500>. (Cited on pages 10 and 118.)
- [40] Owens, D. John, Luebke, David, Govindaraju, Naga, Harris, Mark, Kruger, Jens, Lefohn, E. Aaron, Purcell, and J. Timothy. A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, 26(1):80–113, March 2007. ISSN 0167-7055. doi: <http://dx.doi.org/10.1111/j.1467-8659.2007.01012.x>. URL <http://dx.doi.org/10.1111/j.1467-8659.2007.01012.x>. (Cited on pages 7, 9, and 117.)
- [41] John Owens. Streaming architectures and technology trends. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Courses*, New York, NY, USA, 2005. ACM. doi: <http://dx.doi.org/10.1145/1198555.1198766>. URL <http://dx.doi.org/10.1145/1198555.1198766>. (Cited on pages 7 and 117.)
- [42] David Patterson, Krste Asanovic, Kurt Keutzer, and And. Computer architecture is back - the berkeley view on the parallel computing landscape. Technical report, January 2007. (Cited on pages 8 and 117.)
- [43] D. Pham, S. Asano, M. Bolliger, M. N. Day, H. P. Hofstee, C. Johns, J. Kahle, A. Kameyama, J. Keaty, Y. Masubuchi, M. Riley, D. Shippy, D. Stasiak, M. Suzuki, M. Wang, J. Warnock, S. Weitzel, D. Wendel, T. Yamazaki, and K. Yazawa. The design and implementation of a first-generation cell processor. pages 184–592 Vol. 1, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1493930. (Cited on pages 7 and 117.)
- [44] Clifford A. Pickover. *Computers, Pattern, Chaos, and Beauty: Graphics from an Unseen World*. Dover Publications, Incorporated, 2001. ISBN 0486417093. URL <http://portal.acm.org/citation.cfm?id=516853>. (Cited on pages 13 and 121.)
- [45] Timothy J. Purcell, Craig Donner, Mike Cammarano, Henrik Wann Jensen, and Pat Hanrahan. Photon mapping on

- programmable graphics hardware. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*, pages 41–50. Eurographics Association, 2003. ISBN 1-58113-739-7. (Cited on pages 11 and 119.)
- [46] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, 2002. ISSN 0730-0301. doi: <http://dx.doi.org/10.1145/566654.566575>. URL <http://dx.doi.org/10.1145/566654.566575>. (Cited on pages 12 and 120.)
- [47] Larry Seiler, Doug Carmean, Eric Sprangle, Tom Forsyth, Michael Abrash, Pradeep Dubey, Stephen Junkins, Adam Lake, Jeremy Sugerman, Robert Cavin, Roger Espasa, Ed Grochowski, Toni Juan, and Pat Hanrahan. Larrabee: a many-core x86 architecture for visual computing. In *SIGGRAPH ’08: ACM SIGGRAPH 2008 papers*, pages 1–15, New York, NY, USA, 2008. ACM. doi: <http://dx.doi.org/10.1145/1399504.1360617>. URL <http://dx.doi.org/10.1145/1399504.1360617>. (Cited on pages 7 and 117.)
- [48] Yuan Shi. Reevaluating amdahl’s law and gustafson’s law, October 1996. URL <http://www.cis.temple.edu/~shi/docs/amdahl/amdahl.html>. (Cited on pages 52 and 53.)
- [49] NVIDIA. *NVIDIA CUDA Programming Guide 2.0*. 2008. (Cited on pages 10, 18, 22, 73, 74, 112, 116, and 118.)
- [50] Alvy Ray Smith. A pixell is not a little square. *Microsoft Technical Memo 6*, 1995. (Cited on page 28.)
- [51] Julien C. Sprott. *Strange Attractors: Creating Patterns in Chaos*. M & T Books, har/dsk edition. ISBN 1558512985. URL <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/1558512985>. (Cited on pages 13, 76, and 121.)
- [52] Herb Sutter. The free lunch is over a fundamental turn toward concurrency in software. March 2005. URL <http://www.gotw.ca/publications/concurrency-ddj.htm>. (Cited on page 8.)

- [53] Andrew Thall. Extended-precision floating-point numbers for GPU computation. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Research posters*, New York, NY, USA, 2006. ACM. ISBN 1-59593-364-6. doi: <http://dx.doi.org/10.1145/1179622.1179682>. URL <http://dx.doi.org/10.1145/1179622.1179682>. (Cited on pages 12 and 119.)
- [54] Marc Tremblay and Shailender Chaudry. A third-generation 65nm 16-core 32-thread plus 32-scout-thread cmt sparc[®] processor. *International Solid-State Circuits Conference*, 2008. (Cited on page 8.)
- [55] Turner Whitted. An improved illumination model for shaded display. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Courses*, New York, NY, USA, 2005. ACM. doi: <http://dx.doi.org/10.1145/1198555.1198743>. URL <http://dx.doi.org/10.1145/1198555.1198743>. (Cited on pages 29 and 119.)
- [56] Man-Leung Wong, Tien-Tsin Wong, and Ka-Ling Fok. Parallel evolutionary algorithms on graphics processing unit. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 3, pages 2286–2293 Vol. 3, 2005. doi: <http://dx.doi.org/10.1109/CEC.2005.1554979>. URL <http://dx.doi.org/10.1109/CEC.2005.1554979>. (Cited on pages 12 and 120.)
- [57] Qizhi Yu, Chongcheng Chen, and Zhigeng Pan. *Parallel Genetic Algorithms on Programmable Graphics Hardware*, volume 3612, pages 1051–1059. July 2005. (Cited on pages 12 and 120.)
- [58] Kai Zhang, James T. Kwok, and Ming Tang. Accelerated convergence using dynamic mean shift. In *ECCV (2)*, pages 257–268, 2006. (Cited on page 56.)
- [59] K. Zhao, X. Chu, M. Z. Wang, and Y. Jiang. Speeding up homomorphic hashing using gpus. In *2009 IEEE International Conference on Communications*, pages 1–5. IEEE, June 2009. doi: <10.1109/ICC.2009.5199483>. URL <http://dx.doi.org/10.1109/ICC.2009.5199483>. (Cited on page 12.)

GLOSSARY

CLOCK	Electronic pulses which are emitted periodically, usually by a crystal device, to synchronize the operation of circuits in a computer. Also known as clock signals.	84
SHADER	A computational unit on the GPU.	84

ACRONYMS

1D	one-dimensional	20, 21, 58
1U	one rack unit	17
2D	two-dimensional	20, 58, 59
3D	three-dimensional	8, 20, 58, 59, 89
ALU	Arithmetic Logic Unit	17
ANSI	American National Standards Institute	9, 18
API	Application Programming Interface	i, 2, 7, 9, 10, 18, 21, 27, 32, 89
ATI	ATI Technologies Inc.	9, 10, 15
CPU	Central Processing Unit	i, 1, 2, 7, 8, 12, 15, 18–20, 24, 30, 49, 50, 58–61, 63–66, 70, 71, 75, 76, 81–84, 87, 88, 102, 103, 107, 110, 112
CS4	Creative Suite 4	90
CTM	Close To Metal	9
CUDA	Common Unified Device Architecture	i, 3, 10, 11, 14, 17–21, 29, 32, 57–59, 64, 65, 67, 69–71, 73, 74, 78, 89, 91, 100–102, 104, 108, 116
DMA	Direct Memory Access	21, 60, 65
ECC	Error-Correcting Code	91
EDISON	Edge Detection and Image SegmentatiON System	49–51, 63, 70, 75

G B	Gigabyte	17
GDDR3	Graphics Double Data Rate v3	17
GFLOP	Giga FLoating point Operations Per Second	3, 17, 30, 31
GMP	GNU Multiple Precision	12
GPGPU	General Purpose Computation on GPUS	i, 1–3, 7, 9, 10, 27, 32, 89
GPU	Graphic Processing Unit	i, 1–4, 7–22, 24, 27, 29–32, 50, 52, 53, 55–61, 63–67, 69–71, 74–76, 81–84, 86–91, 93, 94, 99–105, 107–116
GUI	Graphical User Interface	63
HDR	High Dynamic Range	12
HLSL	High Level Shading Langauge	9
HPC	High Performance Computing	27, 91
HSV	Hue Saturation Value	34
IBM	International Business Machines Corporation	7
KB	Kilobyte	17
LUV	$L^* u^* v^*$	34, 42, 44, 46, 54, 58, 60, 61, 64, 66
MHZ	Megahertz	84–86
MIMD	Multiple Instruction Multiple Data	15
MISD	Multiple Instruction Single Data	15
MISE	Mean Integrated Squared Error	37
MMX	Matrix Math Extensions	15
MPI	Message Passing Interface	1
MPMD	Multiple Programs Multiple Data	16
NVIDIA	NVIDIA Corporation	9, 10, 16
PCI-E	Peripheral Component Interconnect Express	17, 21
PPM	Portable Pixel Map	64
RGB	Red Green Blue	34, 44, 46, 54, 58, 60, 61, 64

RO	read-only	20
RW	read-write	20
SDK	Software Development Kit	7, 10, 14, 15, 31, 32
SIMD	Single Instruction Multiple Data	15–17, 27, 32, 60, 95, 101
SISD	Single Instruction Single Data	15
SM	Streaming Multiprocessor	29
SPARC	Scalable Processor ARChitecture	89
SPMD	Single Program Multiple Data	10, 15, 16, 56–58, 69
SSE	Streaming SIMD Extensions	15
USM	Unified Shader Model	16, 17, 101
VMX	Vector Multimedia eXtension	15

LIST OF FIGURES

FIGURE 3.1	Tesla Architecture	25
FIGURE 5.1	Effect of bandwidth selection	35
FIGURE 5.2	Kernel density estimators	37
FIGURE 5.3	Mean shift filtering with parameters $(h_r, h_f) = (6.5, 7)$ applied to a colour image	43
FIGURE 5.4	Mean shift segmentation with parameters $(h_r, h_f, M) = (6.5, 7, 20)$ applied to a colour image	45
FIGURE 6.1	CPU run time with increasing image size	50
FIGURE 7.1	The first row shows the reference pictures computed by the CPU, (1) the filtered, (2) the segmented image and (3) the edges. Below the annotations made by ImageMagick. In this case every pixel computed by the GPU matches the reference	66
FIGURE 8.1	Iteration count for each pixel	78
FIGURE 8.2	Period-0 limit cycle detection	79
FIGURE 8.3	Period-4 limit cycle detection	79
FIGURE 8.4	Period-8 limit cycle detection	80

FIGURE 9.1	GPU run-time and speedup with increasing image-size	81
FIGURE 9.2	Normalised size/time ratios for the CPU and GPU	82
FIGURE 9.3	Multiple GPUs run-time and speedup with increasing image-size	83
FIGURE 9.4	GPU speedups compared to the CPU run-times with increasing image-size	84
FIGURE 9.5	Run-time with increasing memory clock on an image with 256×256 pixel	86
FIGURE 9.6	Run-time with increasing core clock and fixed memory clock on an image with 256×256 pixel with one GPU	88
FIGURE 9.7	Run-time with increasing core clock and fixed memory clock on an image with 256×256 pixel with two GPU	88
FIGURE A.1	Tesla Architecture	111

LIST OF TABLES

TABLE 3.1	Memory performance characteristics for Tesla and GeForce	17
TABLE 3.2	Computing performance characteristics for Tesla and GeForce	18
TABLE 4.1	Execution configuration	30
TABLE 4.2	Comparison between CPU and GPU	30
TABLE 4.3	Local memory and branches	31
TABLE 4.4	Global memory loads and stores	31
TABLE 6.1	EDISON run time profile	51
TABLE 9.1	Memory clock frequencies set in software and the resulting hardware frequencies	85
TABLE 9.2	Various memory clocks and the corresponding run times	86
TABLE 9.3	Core clock frequencies set in software and the resulting hardware frequencies	87

LIST OF LISTINGS

LISTING 3.1	Hardware initialisation	21
LISTING 3.2	Data transfer of data	21
LISTING 3.3	Execution of the Kernel	21
LISTING 3.4	Retrieving of the Results	22
LISTING 3.5	CUDA device code	23
LISTING 8.1	Expensive division	72
LISTING 8.2	Division turned into fast multiplication	72
LISTING 8.3	Magnitude squared of pixel 10992	77
LISTING A.1	Hardware initialization	114
LISTING A.2	Data transfer of data	114
LISTING A.3	Execution of the Kernel	115
LISTING A.4	Retrieving of the Results	115
LISTING A.5	CUDA device code	116



INTERIM REPORT

INTRODUCTION

In recent years graphic processing units (GPUs) have moved from fixed pipeline graphics processors to a fully programmable processor. This evolution has attracted many developers to do general purpose computing on GPUs. GPUs have devoted their silicon (transistors) for computing engines rather than for control engines like caches, branch prediction, coherency protocols and more. This incredible computing power made algorithms with an high arithmetic density run by an order of magnitude faster than on central processing units (CPUs). Speedups of 100x faster than the CPU were stunning but only a few people understand why such speedups are possible and why only a couple of algorithm can attain such speedups.

This thesis will cover all the topics to understand the architecture, programming model, software ecosystem, drawbacks and pitfalls when doing general purpose computing on GPUs. The GPU is a highly parallel processor with thousands of threads and a peak performance of ~600 GFlops (G92 core¹).

Many developers in these days are faced with multicore processors and have to implement or extend existing algorithms to take full advantage of the processing power of such cores. CPU manufacturer are facing fundamental problems when increasing performance only by frequency. In former times higher frequency meant higher performance but a paradigm shift took place now the new stigma is more cores means higher performance. Moore's law says that for every 2 years the amount of cores on a chip will double. What does it mean to developers? They have to think in parallel, not only for two or four cores but rather for 16 or 32 cores. They have to assure that their code is scaling over many cores over many generations of CPU chips. There are several parallel programming languages and middleware to help developers to program in parallel but a quasi standard has not been established.

By the means of an application which will be ported to the GPU the

¹ http://www.nvidia.com/page/ geforce_8800.html

general workflow will be shown and various procedure models examined. It will be presented that often traditional software engineering principles do not apply to high performance, parallel computing. For this work a Nvidia GPU will be used together with CUDA (Common Unified Device Architecture) that is an extension to C for parallel programming of GPUs.

The remainder of the thesis will give some in depth background to the topic and expose with programming models and the architecture of GPUs. Furthermore the development of a parallel implementation of an algorithm will be examined step by step. In this context software analysis and design principles will be shown that fit to parallel programming. A feasibility study will cover major obstacles and show how to avoid them.

Finally an application for visualization of complex and chaotic dynamical systems will be implemented and presented in all aspects to the reader.

BACKGROUND TO THE PROJECT

There are many opinions about parallel programming. On the one hand some say that parallel programming is hard and error prone and on the other hand with the right tools and programming languages it is a piece of cake. This work will examine some facts and thoughts about parallel programming and show how one can identify parallelism, avoid race conditions and tune for performance. Additionally a modern programming method will be introduced that makes programming graphics processors easy. The following sections will cover some major topics which are the basic modules of this thesis. Beginning with the architecture of modern graphics processors it will be shown how these GPUs can be programmed and at last a simple example implemented in CUDA.

The Tesla Architecture

The Tesla architecture announced 1999 and developed by NVIDIA is the first GPU highly specialized for raster operations and more important for general purpose computing. Formerly GPUs had fixed-function pipelines and separate processing units with no ability for programmability to the vertex and fragment stages of the pipeline. In recent years manufacturers of GPUs added more and more programmability to the

different stages of the pipeline and at the same time general purpose computing capabilities [32]. Furthermore manufacturers introduced the USM that unifies the processing units allowing for better utilization of GPU resources. The resources needed by different shaders varies greatly and the unified design can overcome this issue by balancing the load among vertex, fragment and geometry functionality [8].

FIGURE A.1 Tesla Architecture

The figure Figure 3.1 shows the Tesla architectures. As mentioned above the new GPU architectures are a radical departure from traditional GPU design (USM). The Tesla 8 Series has 16 multiprocessors. Each multiprocessor is composed of 8 streaming processors, 128 processors in total. Each streaming multiprocessor has 16 KB of shared memory and a L1 cache attached and has access to a texture unit. A streaming processor consists of a scalar ALU and performs floating point operations. 32 streaming processor build up a SIMD unit (warp) in that one instruction is executed. The 8800 GTX has 768 MB of graphics memory, 620 GFlops of peak performance and 86 GB/s peak memory bandwidth. Many massively data-parallel algorithms can be run sufficiently on this specialized architecture [8]. Programming for this architecture is done with CUDA (Common Unified Device Architecture) the C language extension which will be covered in the next section.

Common Unified Device Architecture (CUDA)

In 2007, NVIDIA introduced an extended ANSI C programming model and software environment the Compute Unified Device Architecture (CUDA). The reason why CUDA was born is that parallelism is increasing rapidly with Moore's law² and the challenge is to develop parallel application software that scales transparently with the number of processor cores. The main goals when CUDA was developed were that it scales to 100s of cores, 1000s of parallel threads and it allows heterogeneous computing (CPU + GPU). All this considerations led

² Processor count is doubling every 18 - 24 months

to the result that **CUDA** runs on any number of processors without recompiling an the parallelism applies to both **CPUs**and **GPUs**[5].

CUDA is C with minimal extensions and defines a programming and memory model. There are three key abstractions in **CUDA**, a hierarchy of thread groups, shared memories and barrier synchronization [49] that are exposed to the developer. **CUDA** uses extensive multithreading where threads express fine-grained instruction, data and thread parallelism that are grouped into thread blocks which express coarse grained data and task parallelism. The developer has to rethink about his algorithms to be aggressively parallel. The problem has to be split into independent coarse sub-problems and at finer level into fine-grained sub-problems that can be solved cooperatively [49].

CUDA has been accepted by many developers which can be seen by the huge amount of already developed software and contributions to the **CUDA** Zone: www.nvidia.com/cuda. A brief look shows the **CUDA** computing sweet spots[5]:

- High arithmetic intensity (Dense linear algebra, PDEs, n-body, finite difference, ...)
- High bandwidth Sorting, database, sequencing, virus scanning, genomics, ...)
- Visual computing: (Graphics, image processing, tomography, machine vision, ...)
- Computational modeling, science, engineering, finance, ...

This is just a small snapshot of algorithms that can be used with **CUDA** on GPUs. For a more extensive list and speedups compared to high end CPUs see www.nvidia.com/cuda and www.gpGPU.org.

CUDA Programming Model

The **CUDA** programming model exposes the graphics processor as a highly multithreaded coprocessor. The graphics processor is viewed as a compute device that is a coprocessor to the host that has its own device memory and runs many threads in parallel.

Applications are accelerated by executing data parallel portions of the algorithm on the GPU as *kernels* which run in parallel on many threads. There are some major differences between CPU and GPU threads. A

GPU needs thousands of threads for full efficiency where CPUs only need a few of them. GPU threads have a very little creation overhead and are extremely lightweight compared to CPU threads.

THREAD BATCHING A kernel is executed as a grid of thread blocks where data memory space is shared by all threads. A thread block is a batch of threads that can cooperate with each other by synchronizing their execution³ and efficiently sharing data through the low latency shared memory. Two threads from two different blocks can cooperate with atomic functions through the global memory. The identification of a thread is accomplished through block and thread ids which are assigned to each thread at creation time.

BLOCK AND THREAD IDS Every thread and block have a unique id. As a result of this each thread can decide what data to work on. For every block there is assigned an id in 1D or 2D layout. Thread ids can be accessed either with 1D, 2D or 3D coordinates similar to multidimensional arrays. It simplifies memory addressing when processing multidimensional data. For example image processing, matrix multiplication or solving partial differential equations on volumes and so on. The data can reside in several levels of the device memory.

DEVICE MEMORY SPACE The memory space is a hierarchy of several memory types that can be accessed per thread, block, grid and the host. The threads have access to all memory levels beginning with the read/write (rw) registers, local, shared, global, read only (ro) texture and constant memory. The grids have only access to global, constant and texture memory. Whereas the CPU (host) can (rw) global, constant and texture memories.

Global, constant and texture memory have long latency accesses. They reside off chip where registers local⁴ and shared memory reside on-chip. The global, constant and texture memory are mainly used for communication of (rw) data between host and device where the contents is visible to all threads. As mentioned above texture and constant memory can be written by the host where constants and textures are initialized.

³ For hazard free shared memory access

⁴ Not true for older GPUs where local data is spilled out to global memory

A Simple Example

This simple example will show the structure of an CUDA program. The executing kernel will do some easy calculations on the data provided, load the data into shared memory and write back the results to the global memory.

A CUDA program has a specific structure where the major parts are described in this paragraph. The first thing to do is to initialize the device and some auxiliary variables. THe Listing A.1 shows the initalization.

```

1 int main( int argc, char** argv) {
2     CUT_DEVICE_INIT(argc, argv);
3
4     uint32_t num_threads = 32;
5     uint32_t mem_size = sizeof(float) * num_threads;
6
7     ...

```

LISTING A.1 Hardware initalization

Since the GPU is attached to the PCIe bus the host has no direct access to the global, constant and texture memory and has to transfer the data back and forth with the DMA engine of the device. This is accomplished through the CUDA api calls that initiate the transfer. Before any transfer can be done one has to allocate memory on the host and on the device for input and output data. This is shown in listing Listing A.2.

```

1 // allocate host memory
2 float* h_idata = (float*) malloc(mem_size);
3 // allocate device memory
4 float* d_idata; cudaMalloc((void**) &d_idata, mem_size);
5 // allocate device memory for result
6 float* d_odata; cudaMalloc((void**) &d_odata, mem_size);
7 // allocate mem for the result on host side
8 float* h_odata = (float*) malloc( mem_size);
9 // copy host memory to device
10 cudaMemcpy(d_idata, h_idata, mem_size, cudaMemcpyHostToDevice);

```

LISTING A.2 Data transfer of data

After setting up the input data the setup execution parameters are defined that are used to startup the kernel. The `grid(1, 1, 1)` statement defines a multi-dimensional array of grids $x = 1, y = 1, z = 1$ whereas the `threads(num_threads, 1, 1)` defines a multi-dimensional array of threads $x = num_threads, y = 1, z = 1$ which are actually $1D$ arrays. Listing Listing A.3 shows the call of the kernel with its input and output data.

```

1   // setup execution parameters
2   dim3 grid( 1, 1, 1);
3   dim3 threads(num_threads, 1, 1);
4
5   // execute the kernel
6   kernel<<< grid, threads, mem_size >>>(d_idata, d_odata);

```

LISTING A.3 Execution of the Kernel

If everything went well the host can copy the data from device memory to host memory and check, visualize or store the calculated values. Listing Listing A.4 shows the last steps before exiting the program.

```

1   // check if kernel execution generated and error
2   CUT_CHECK_ERROR("Kernel execution failed");
3
4   // copy result from device to host
5   cudaMemcpy(h_odata, d_odata, sizeof( float) * num_threads,
6             cudaMemcpyDeviceToHost);
7
8   // cleanup memory free(x), free(y), free(z) ...
9   CUT_EXIT(argc, argv);
10 }

```

LISTING A.4 Retrieving of the Results

The previous listings showed the host code and how to launch a kernel on the device. The listing Listing A.5 shows the device code portion. There are several qualifiers that define which function is compiled for which processing unit. The `__global__` qualifier specifies that this function is run on the device and hence compiled for the GPU where the

`__host__` qualifier specifies that this function is only run on the host and not on the device. There are more function specifiers that can be looked up in [49].

For data there are as well qualifiers where one can specify where the data is located, either in constant, global or shared memory. In listing Listing A.5 the `__shared__` qualifier is used. The device will use the shared memory to preload the data for faster access.

```

1   #include <stdio.h>
2
3   #define SDATA(index) CUT_BANK_CHECKER(sdata, index)
4   // Simple test kernel for device functionality
5   __global__ void kernel( float* g_idata, float* g_odata)
6   {
7       // shared memory
8       // the size is determined by the host application
9       extern __shared__ float sdata[];
10
11      // access thread id
12      const unsigned int tid = threadIdx.x;
13      // access number of threads in this block
14      const unsigned int num_threads = blockDim.x;
15
16      // read in input data from global memory
17      // use the bank checker macro to check for bank conflicts during host
18      // emulation
19      SDATA(tid) = g_idata[tid];
20      __syncthreads();
21
22      // perform some computations
23      SDATA(tid) = (float) num_threads * SDATA( tid);
24      __syncthreads();
25
26      // write data to global memory
27      g_odata[tid] = SDATA(tid);
28  }
```

LISTING A.5 CUDA device code

After loading the data the kernel just multiplies the thread-id with the number of threads and saves the result back to global memory where the host can pick up the result.

INITIAL SURVEY

The first thing to answer is why should someone do general purpose computing on GPUs (GPGPU) anyway. For the most people CPUs are just enough. They do not demand on high computational power and on a high bandwidth. Still there is paradigm shift taking place and this can not be neglected. As stated in [40] and in [41] CPU manufacturers are facing problems which they cannot overcome just by increasing the frequency. The often cited *Walls* are first, *The Memory Wall*[37], *The Frequency Wall* and *The Power Wall*. The only way seen by CPU makers is currently to go multicore. Intel e.g. went multicore with their new *Core Microarchitecture* for consumer products and even a GPU replacement, *Larrabee* [47]. IBM, Toshiba and Sony developed the *Cell Broadband Architecture* a 9 core chip [43]. SUN developed the *Niagara* CPU a multicore general purpose processor. It has eight in-order cores, each of them capable of executing four simultaneous threads [26].

Compared to CPUs GPUs went years ago to multicore and multithreading. GPUs are maybe the kind of processor where CPUs are heading to in terms of multithreading and raw performance. Forecasts project that every two years the amount of cores can double. The multicore approach may be the answer to the problems stated above but this yields to another thing the *parallel programming problem* [42]. User will only benefit from this growth if software can make use of all the cores. Many developers learned about the single-threaded von neumann model and are not familiar with parallel code which is subject to errors such as deadlocks and livelock, race conditions and many more. Parallel programming is difficult and there are several paradigms to make life easier for developers.

One of these is the data-parallel paradigm. Where one is not trying to assign different subtasks to separate cores rather assigning an individual data element to a separate core for processing [7]. 3D rendering, an embarrassingly data-parallel problem, has driven the GPU evolution which makes the GPU a perfect target for data-parallel code. There are

several fine-grained or data-parallel programming environments that leverage the GPU for general purpose computing (Brook, Sh, RapidMind, ...).

The focus of this work will be CUDA⁵. CUDA is the only environment which is not based on a graphics library and officially released by NVIDIA for their GPUs. CUDA is a minimal extension to C and C++ programming languages. The technique employed by CUDA is single process, multiple data (SPMD). Tasks are split up and run simultaneously on multiple threads (cores) with different input [39].

Choosing a Fast Algorithm

Before even digging into the wide field of algorithms and difficult problems in high performance computing one has to understand the hardware architecture of the GPU to make a decision if an algorithm can be mapped on GPUs. A pretty good overview over the NVIDIA GPU gives the *NVIDIA CUDA Programming Guide* [49]. A little bit outdated but still of interest is *The GeForce 6 Series GPU Architecture* [25] which gives an overview how the GPU fits into the whole system, what a fragment processor, vertex processor or what textures are. To have an even deeper look into GPU the article [17] is highly recommended.

COMPUTATIONS WHICH MAP WELL TOGPUS It is important to understand that GPUs are good at running computer graphics and algorithms which *mimic* or have the attributes of computer graphics in terms of data parallelism and data independence. Not only that similar computations are applied to streams of many data elements (vertices, fragments, ...) but also the computation of each element is completely or almost completely independent [22]. Such types of algorithms are often called embarrassingly parallel algorithms where subtasks rarely or never communicate to each other.

Another important fact for an algorithm is the *Arithmetic Intensity*. The Arithmetic Intensity is the ratio of computation to bandwidth or formally:

$$\text{arithmetic intensity} = \text{operations} / \text{words transferred}.$$

⁵ www.nvidia.com

This fact is important because the increase of computational throughput is faster than the memory throughput which leads to the problem known as *The Memory Wall*. GPU memory systems are architected to deliver high bandwidth, rather than low-latency, data access. As such computations that benefit most of the GPU have a high arithmetic intensity [22]. The next sections will represent some algorithms which could fit to GPUs.

RAY TRACING Ray Tracing [2] is an embarrassingly parallel algorithm which could fit well to GPUs. The author has a extensible knowledge of Ray Tracing on massively parallel computers [28]. Thats why Ray Tracing was first considered for porting to the GPU. Unfortunately there were several implementations already done for the GPU. Nevertheless equipped with all the knowledge about Ray Tracing and how to split up the work, arrange the data on a parallel machine to run efficiently the Ray Tracing algorithm [55] will be used for initial benchmarks and the feasibility study.

PHOTON MAPPING Ray Tracing has a local illumination model. To generate more realistic effects like caustics, diffuse / glossy indirect illumination and more a more sophisticated model has to be used. Global illumination like *Photon Mapping* [24] can create all the effects that Ray Tracing cannot. There are implementations of photon mapping on GPUs [45] which are developed with graphics apis and not with CUDA. Anyhow since photon mapping is heavily using a kd-tree it would be a major effort to develop an efficient data structure which has the same functionality as a kd-tree. Nevertheless the first candidate for porting to the GPU is *Photon Mapping*.

MULTIPLE PRECISION ARITHMETIC Another interesting field which demands high computational power is number theory. The most common/known application is asymmetric cryptography. To decipher messages that are cyphered with an asymmetric algorithm one needs superior computational power. An overview over common algorithms gives [27]. All algorithms have one thing in common they need a multiple precision library to represent numbers with 200 and more digits. A good overview gives [53].

The idea was to implement some of the factoring algorithms to the GPU. The only thing needed is the multiple precision library. In [27] the Gnu Multiple Precision (GMP) library was used to implement the algorithms. So the multiple precision arithmetic is another candidate for porting to the GPU.

HIGH DYNAMIC RANGE Image processing is always a candidate for embarrassingly parallel algorithms. A pretty new algorithm to enhance images is *tone mapping* [34]. Tone mapping is the compression of dynamics in High Dynamic Range (HDR) pictures. There are several algorithms for tone mapping: Mantiuk [33], Reinhard [46], Durand [16], Fattal [18] and many more. All of this algorithms are present in the pfs-tools library written by Krawczyk [34]. As this document was written Krawczyk was already implementing the pfs-tools on GPU but not publishing it. Furthermore there is a complete editor written for hdr image processing which is running on the GPU. So HDR was discarded but there are several other algorithms in image processing which are considered as candidates. Some algorithms in no particular order: segmentation, tracking, filtering, ... and so on. This is put as another candidate to the list of the possible algorithms for porting.

GENETIC ALGORITHMS Parallel genetic algorithms are usually implemented on parallel machines but fine-grained parallel genetic algorithms can be mapped to GPUs [57]. In [30] its shown what kind of genetic algorithm map well to GPUs and how the work and communication is handled. It is *pretty* easy to implement simple genetic algorithms on the GPUs but with increasing complexity one has to consider many more things: load balancing, communication pattern, dynamic memory allocation, resolving of recursion Another paper [56] compares genetic algorithms implemented on CPUs and GPUs and shows that the latter is much more effective than the former. All genetic algorithms have one thing in common the core algorithm. Once effectively implemented on the GPU the core algorithm is extended with definitions like the population, selection, recombination and mutation to solve a specific problem. The skill here is to choose the right definitions and not more the efficient implementation of the core algorithm. So this topic excluded from the candidate list.

CHAOS THEORY Another interesting field is chaos theory. Especially the visualization of chaos. The maybe most famous visualization of chaos is *The Fractal Flames Algorithm*. Fractal flames are a member of iterated function system class of fractals created by Scott Draves [14]. He uses a rather complicated set of functions in the system to generate stunning visualization of the iterative process of the system. The next release will have support for the GPU. Thats why it was firstly discarded but the research about chaos led to other interesting papers respectively books.

There is no need to use approx. 21 function for a system to generate visually appealing pictures of chaos. Sprott showed in [51] that even with very simple functions one can create patterns in chaos. This patterns are called *Strange Attractors* and are the visualization of chaotic behaviour. There are created by iterating a simple equation some million times.

Pickover shows in his book [44] how to create patterns from a variety of sources. He shows how to create nice looking patterns from fourier analysis, acoustic, chemistry and many more. Anyhow all of these equations or differential systems have on thing in common no matter how complicated the system is the core algorithm is to iterate a specific equation with correct input numbers to create chaos. The algorithm is heavily computational bound which makes it a good candidate for porting to the GPU.

Summary

After examining many possible candidates which can be accelerated with an GPU, some of them were already in progress, some other maybe not feasible in the timeframe of a master thesis. Nevertheless one algorithm has to be chosen and this specific one will be ported to the GPU.

By the means of this example the typical data parallel programming approach will be shown and it will be illustrated how to think in parallel.

AIMS AND OBJECTIVES

One aim of this master thesis is to show whether graphic cards are suitable for general purpose computing. With an algorithm example that is running slow on a CPU the process of porting to a GPU is examined.

As a result, the algorithm will be running by a magnitude faster on the GPU.

Another aim is to understand the architecture and programming models of GPUs. Therefore the algorithmic view to GPUs will be presented and computational concepts and efficient data structures explained. Understanding these key concepts makes one understand the architecture of GPUs.

Dissasociation from sequential algorithms and thinking in parallel will be one major discipline after working through this thesis. One will be able identify algorithms that are well suited for GPUs and design efficient data flow and high performing code.

Assessment of algorithms, knowledge about GPU architecture, program flow and data access patterns make one able to rate if a GPU is suitable for general purpose computing.

EXPERIMENTAL/INVESTIGATIVE METHODS TO BE ADOPTED

To estimate if an algorithm is suitable for running on the GPU first of all the hardware architecture has to be understood. Since GPUs are *completely* different than CPUs one has to understand how simple things like loops, arithmetic, data access and program flow are managed on such hardware. The most interesting part here is the fact that GPUs are able to hold up to approx. 12000 threads in flight. Understanding the hardware is the key for high performance computing on GPUs.

To get an impression how one can/should write code for GPU a feasibility study will be done that will show answers to statements stated above and further questions.

Further questions arise when reading scientific papers about general purpose computing on graphics hardware. Often there are statements which are not applicable to current hardware generation. These limitation arise from the fact that the GPUs are implementing a graphics pipeline that is *normally* used for rasterization and rendering. The developers were confronted with an graphics API and a graphics pipeline which thinks of triangles and pixel rather than floats and integers. As a result the developers had to think about how to rotate, scale or translate a triangle to perform a calculation in an algorithm. There are several myths and truths about general purpose computing on graphics hard-

ware which will be answered with the evolution of the feasibility study.

Since general purpose computing on graphics hardware is not more or less 'C' programming with some extension, the best way to learn programming for GPU is to learn by doing. Therefore several algorithms (Box Filter, Gaussian Blur, HDR, ...) and further more will be studied. This way one will get enough experience to structure an initial code and have a clue about how to split the work apart, offload computational intensive parts to the GPU and let the general purpose processor handle the I/O and input handling.

Simple micro, synthetic⁶ benchmarks are often used to describe the hardware by their peak numbers. Either gflops achievable or the bandwidth between the processing unit and the memory. They can easily show the basic conditions which one has to face with. Several benchmarks are already available in the SDK so only a few have to be developed. One critical point in general purpose computing is called arithmetic density. Which describes how many arithmetic operations are executed before another load is needed to fetch data. Micro benchmarks can easily show how arithmetic density effects performance, which in conclusion can show how performant an application will run on the GPU.

The steps above will lead to a pretty good understanding of the hardware which leads to the next step the searching and identifying a suitable algorithm to port to the GPU. A critical survey of algorithms that are ported to GPU and those that are not ported will be done. There are many high performance algorithms out there which take *forever* to complete. If it's chemistry, physics, biology, astronomy and mathematics all of them can be ported to a GPU to some extent. It is important to fetch the applicable one and by means of that algorithm to show the analysis and design process of high performance programs. There are several things to consider when designing for parallel hardware which will be examined and explained in the responsible chapters.

TIME - PLAN

For the project *General Purpose Computing on Graphics Hardware* the following milestones were defined and listed below. Each milestone

⁶ Synthetic benchmarks mimic a particular type of workload on a component or system. Whereas application benchmarks run real world programs

will have a short description which tasks have to be accomplished to complete a mile stone. With every milestone a chapter respectively a section of the master thesis will be written which corresponds to the current carried out milestone. So the development of the milestones goes hand in hand with editing of the thesis.

M0: INTERIM REPORT (1/31/2009) For this milestone following tasks were done, the conduct of the initial survey and the writing of the interim report. The initial survey will give a coarse overview over the field and related work.

M1: RELATED WORK IDENTIFIED (2/14/2009) The next step to take is to identify related work, conducting the literature survey. This survey should show the demand of GPGPU and used algorithms. Furthermore it will help to understand the architecture and programming models. After gaining enough knowledge about the field its time to get the hands on the GPU.

M2: ALGORITHMIC VIEW UNDERSTOOD (2/28/2009) Equipped with the knowledge from the latter milestone it is time to understand how the architecture works by examining algorithms already mapped to the GPU. This step is crucial as it shows how algorithms map to and work on the GPU and a good understanding of the environment can help later on for choosing the right algorithm for porting, analysis, design and implementation.

M3: FEASIBILITY STUDY (3/14/2009) The feasibility study will be used for the first steps into GPGPU. A low featured ray tracer will be used for examining what the hurdles and obstacles are when developing for the GPU. The knowledge will be used for choosing the right algorithm.

M4: APPLICATION, ALGORITHM CHOOSEN (3/21/2009) After completion of this milestone the algorithm is chosen to be ported to the GPU. So the next milestones will be concentrated only on this algorithm and not more generally.

M5: ANALYSIS, SOLUTION PROPOSED (4/4/2009) Here a solution will be proposed and outstanding issues will be identified. There are several myths about GPGPU which will be stated here as well, because there are affecting the design of the application and cannot be discarded.

M6: EXPERIMENTS (4/14/2009) To exclude all eventualities a set of experiments will be undertaken. This will as well solve all myths.

M7: DESIGN, APPLICATION PORTED (5/14/2009) This milestone will cover the design of the application with all constraints, obstacles and hurdles that are involved with GPGPU. In this task the application will be implemented and run on the GPU. A good implementation will need several iterations were several aspects will be tuned: coalesced memory access, load balancing and many more.

M8: DISCUSSION (5/22/2009) Another important part is the discussion of the results and methodologies applied which will be done in this milestone.

M9: MASTER THESIS EDITED (5/30/2009) Here the thesis will be recaped and briefly outlined. Conclusions will be drawn and future work presented. A critical summary and conclusion will be provided as well.

DELIVERABLES OR SPECIFIC OUTCOMES

The main outcome of this work is to show how GPUs can be utilized for general purpose computing. There is a high demand on performance and computational power for scientific algorithms. Developers search for new architectures to solve problems in shorter time. One promising new architecture or hardware is the GPU. Since years GPUs were only used for rendering. The switch from fixed pipelines to programmable pipelines let developers think about using GPUs for general purpose computing. Accidentally they were forced to use 3D apis (OpenGL⁷, DirectX⁸) to accomplish task like matrix multiplication or fast fourier

⁷ <http://www.opengl.org>

⁸ <http://www.microsoft.com>

transformation. Nowadays developer use CUDA or middleware like RapidMind to develop software for (NVIDIA) GPUs. It will be shown that CUDA is highly efficient and easy to program which makes general purpose computing more feasible.

By the means of a application that is computationally intensive it will be shown that traditional software engineering principles often do not apply to high performance, parallel computing. Data access is the biggest problem of an algorithm to attain maximum performance. If the algorithm is memory bound and not computational bound the limiting factor or bottleneck is the bandwidth to the memory. Good data design is essential for good (performant) code. The data and how it is used must be known in order to effectively design and optimize a application for a GPU or other highly parallel systems like the Cell processor (Meine Diplomarbeit link). A simple computationally bound algorithm can be easily designed in a way that it becomes memory bound if data arrangement and access is neglected by the developer. The outcome of all the considerations will be another approach how to tackle software design on such massively parallel architecture.

Furthermore a workflow will be presented that shows how to port sequential code to a massively parallel architecture. Software developer often only need to compile their code with another compiler to *port* the code to another architecture. x86 -> PowerPC, x86 -> MIPS ... but the situation for parallel processors like the cell, niagara and here the GPU is completely different. One has to identify the critical parts and explicitly offload tasks to the GPU. It is obvious that this is not the only step to do. Furthermore there are other facts which have to be considered, like load-balancing, deadlocks, concurrency, cache coherency and other. A workflow which can be adapted to different parallel machines is another outcome of this work.

Last but not least is the outlook on future hardware development and programming models. The problem with new hardware architectures is they have to be accepted by the developers to succeed in industry. One way to achieve this is to have an easy programming model which adapts easily to common apis and existing systems. There are rarely branches of industry which want to, or are able to rewrite their code for bleeding edge architectures like the cell or GPUs if the outcome is only 2x faster code. In this case many just extend their cluster with another bunch of

cheap systems. A factor of 10 is an undocumented limit when to invest in other, faster hardware.