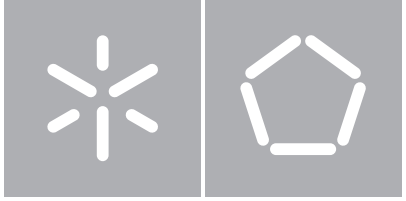
Universidade do Minho
Escola de Engenharia

Miguel Branco Palhas

An Evaluation of the GAMA/StarPU Frameworks for Heterogeneous Platforms: the Progressive Photon Mapping Algorithm

Setembro de 2013



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Miguel Branco Palhas

An Evaluation of the GAMA/StarPU Frameworks for Heterogeneous Platforms: the Progressive Photon Mapping Algorithm

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho realizado sob orientação de

Professor Alberto Proença
Professor Luís Paulo Santos

Setembro de 2013

Anexo 3

DECLARAÇÃO

Nome

Miguel Branco Palhas

Endereço electrónico: pg19808@alunos.uminho.pt Telefone: 910 565 249 /

Número do Bilhete de Identidade: 13473816

Titulo dissertação □/tese □

An Evaluation of the GAMA/StarPU Frameworks for Heterogeneous Platforms: The Progressive Photon Mapping Algorithm

Orientador(es):

Professor Alberto Proença

Professor Luis Paulo Santos Ano de conclusão: 2013

Designação do Mestrado ou do Ramo de Conhecimento do Doutoramento:

Mestrado em Engenharia Informática

Nos exemplares das teses de doutoramento ou de mestrado ou de outros trabalhos entregues para prestação de provas públicas nas universidades ou outros estabelecimentos de ensino, e dos quais é obrigatoriamente enviado um exemplar para depósito legal na Biblioteca Nacional e, pelo menos outro para a biblioteca da universidade respectiva, deve constar uma das seguintes declarações:

1. É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE/TRABALHO APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

2. É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA TESE/TRABALHO (indicar, caso tal seja necessário, nº máximo de páginas, ilustrações, gráficos, etc.), APENAS PARA EFEITOS DE INVESTIGAÇÃO, , MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

3. DE ACORDO COM A LEGISLAÇÃO EM VIGOR, NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA TESE/TRABALHO

Universidade do Minho, 01 / 09 / 2013

Assinatura: Miguel Branco Palhas

Agradecimentos

Ao meu orientador, Alberto Proença, por uma excelente orientação, quer na disponibilidade completa durante todo o ano, quer no rigor exigido e análise ao trabalho efetuado. Ao meu co-orientador, Luís Paulo Santos, pelos desafios lançados no início do trabalho, e pela disponibilidade que demonstrou sempre que solicitado. Deixo ainda um agradecimento aos professores que mais me marcaram positivamente neste percurso, nomeadamente os professores Alberto Proença, Luís Paulo Santos, Rui Mendes, António Ramires Fernandes, Jorge Sousa Pinto, Mário Martins e José Carlos Ramalho.

Aos colegas do LabCG, que me acolheram durante este ano num ótimo ambiente de trabalho, que foi essencial para esta dissertação. Em especial ao Roberto Ribeiro, pela ajuda e discussões ao longo do ano, que foram uma grande contribuição para o trabalho.

Aos meus colegas de trabalho André Pereira e Pedro Costa, por todas as discussões, ajuda mútua e companheirismo durante estes dois anos de mestrado.

A todos os membros do CeSIUM, núcleo que me acolheu como uma segunda casa durante todo o meu percurso académico.

Ao grande grupo de amigos que formei nesta universidade, cuja lista é demasiado grande para enumerar aqui, que me acompanharam nestes que foram os melhores anos da minha vida, e me deram apoio durante os momentos de maior aperto. Sem eles este trabalho não teria sido possível.

Aos colegas e amigos da GroupBuddies, em especial ao Roberto Machado, pela ajuda e compreensão demonstrada sempre que a minha ausência foi necessária.

Por fim, um especial agradecimento à minha mãe e ao meu irmão, pelo suporte e compreensão dada a minha acrescida ausência durante o ano.

Work funded by the Portuguese agency FCT, *Fundação para a Ciência e Tecnologia*, under the program UT Austin | Portugal.

Abstract

Recent evolution of high performance computing moved towards heterogeneous platforms: multiple devices with different architectures, characteristics and programming models, share application workloads. To aid the programmer to efficiently explore these heterogeneous platforms several frameworks have been under development. These dynamically manage the available computing resources through workload scheduling and data distribution, dealing with the inherent difficulties of different programming models and memory accesses. Among other frameworks, these include GAMA and StarPU.

The GAMA framework aims to unify the multiple execution and memory models of each different device in a computer system, into a single, hardware agnostic model. It was designed to efficiently manage resources with both regular and irregular applications, and currently only supports conventional CPU devices and CUDA-enabled accelerators. StarPU has similar goals and features with a wider user based community, but it lacks a single programming model.

The main goal of this dissertation was an in-depth evaluation of a heterogeneous framework using a complex application as a case study. GAMA provided the starting vehicle for training, while StarPU was the selected framework for a thorough evaluation. The progressive photon mapping irregular algorithm was the selected case study. The evaluation goal was to assert the StarPU effectiveness with a robust irregular application, and make a high-level comparison with the still under development GAMA, to provide some guidelines for GAMA improvement.

Results show that two main factors contribute to the performance of applications written with StarPU: the consideration of data transfers in the performance model, and chosen scheduler. The study also allowed some caveats to be found within the StarPU API. Although this have no effect on performance, they present a challenge for new coming developers. Both these analysis resulted in a better understanding of the framework, and a comparative analysis with GAMA could be made, pointing out the aspects where GAMA could be further improved upon.

Resumo

Uma avaliação das frameworks GAMA/StarPU para Plataformas Heterogéneas: O algoritmo de Progressive Photon Mapping

A recente evolução da computação de alto desempenho é em direção ao uso de plataformas heterogéneas: múltiplos dispositivos com diferentes arquiteturas, características e modelos de programação, partilhando a carga computacional das aplicações. De modo a ajudar o programador a explorar eficientemente estas plataformas, várias frameworks têm sido desenvolvidas. Estas frameworks gerem os recursos computacionais disponíveis, tratando das dificuldades inerentes dos diferentes modelos de programação e acessos à memória. Entre outras frameworks, estas incluem o GAMA e o StarPU.

O GAMA tem o objetivo de unificar os múltiplos modelos de execução e memória de cada dispositivo diferente num sistema computacional, transformando-os num único modelo, independente do hardware utilizado. A framework foi desenhada de forma a gerir eficientemente os recursos, tanto para aplicações regulares como irregulares, e atualmente suporta apenas CPUs convencionais e aceleradores CUDA. O StarPU tem objetivos e funcionalidades idênticos, e também uma comunidade mais alargada, mas não possui um modelo de programação único

O objetivo principal desta dissertação foi uma avaliação profunda de uma framework heterogénea, usando uma aplicação complexa como caso de estudo. O GAMA serviu como ponto de partida para treino e ambientação, enquanto que o StarPU foi a framework selecionada para uma avaliação mais profunda. O algoritmo irregular de *progressive photon mapping* foi o caso de estudo escolhido. O objetivo da avaliação foi determinar a eficácia do StarPU com uma aplicação robusta, e fazer uma análise de alto nível com o GAMA, que ainda está em desenvolvimento, para forma a providenciar algumas sugestões para o seu melhoramento.

Os resultados mostram que são dois os principais factores que contribuem para a performance de aplicação escritas com auxílio do StarPU: a avaliação dos tempos de transferência de dados no modelo de performance, e a escolha do escalonador. O estudo permitiu também avaliar algumas lacunas na API do StarPU. Embora estas não tenham efeitos visíveis na efici-

encia da framework, eles tornam-se um desafio para recém-chegados ao StarPU. Ambas estas análisos resultaram numa melhor compreensão da framework, e numa análise comparativa com o GAMA, onde são apontados os possíveis aspectos que o este tem a melhorar.

Contents

1 Introduction **1**

1.1 Context 1

1.2 Motivation & Goals 4

1.3 Document Organization 5

2 Computing Background **7**

2.1 Parallel Computing 7

2.2 The GPU as a Computing Accelerator 8

2.2.1 Fermi Architecture 9

2.2.2 Kepler Architecture 10

2.3 MIC Architecture 11

2.4 Heterogeneous Platforms 12

2.5 Heterogeneity and the Future 13

3 Frameworks for Heterogeneous Platforms **15**

3.1 GAMA 15

3.1.1 Memory Model 16

3.1.2 Programming and Execution Model 17

3.1.3 The Polu Case-Study 19

3.2 StarPU 21

3.2.1 Terminology 21

3.2.2 Task Scheduling 22

3.2.3 Dependencies 23

3.2.4 Virtual Shared Memory 24

3.2.5 Multi-threading 24

3.2.6 API 25

3.2.7 Performance Model 25

3.2.8 Task Granularity 26

3.3 Comparison 27

3.3.1 Usability 27

3.3.2 Scheduling 28

3.3.3 Memory Management 28

4 Case Study: the Progressive Photon Mapping Algorithm 31
4.1 Ray Tracing 31
4.1.1 Overview 32
4.1.2 The Rendering Equation 33
4.2 Photon Mapping 33
4.2.1 Algorithm 34
4.2.2 Applications 35
4.3 Progressive Photon Mapping 36
4.3.1 First Step: Ray Tracing 37
4.3.2 Next Steps: Photon Tracing 37
4.3.3 Radiance Estimation 38
4.4 Stochastic Progressive Photon Mapping 40
4.5 A Probabilistic Approach for Radius Estimation 40
4.6 Summary 42

5 Implementation Approaches of the Algorithm 43
5.1 Data Structures 44
5.2 Computational Tasks 46
5.3 Implementation 47
5.3.1 Original 49
5.3.2 CPU 50
5.3.3 GPU 51
5.3.4 MIC 51
5.3.5 StarPU 52

6 Profiling Results 59
6.1 Testing Environment 59
6.2 Testing Methodology 59
6.2.1 Input Scene 61
6.3 Performance Results Without a Framework 61
6.3.1 CPU 61
6.3.2 GPU 62
6.4 Performance Results With StarPU 63
6.4.1 Scheduler Impact 63
6.4.2 Performance with Accelerators 65
6.4.3 Calibration 66
6.4.4 Overall Performance Comparison 68
6.4.5 Concurrent Iterations 68

7 Conclusions **71**

7.1 Future Work 73

Glossary

AVX Advanced Vector Extensions

BRDF Bidirectional Reflectance Distribution Function

BVH Bounding Volume Hierarchy

CPU Central Processing Unit

CU Computing Unit

CUDA Compute Unified Device Architecture

DMA Direct Memory Access

DSP Digital Signal Processor

GAMA GPU And Multi-core Aware

GPU Graphics Processing Unit

GPGPU General Purpose GPU

HEFT Heterogeneous Earliest Finish Time

HetPlat **Het**erogenous **Plat**form

ISA Instruction Set Architecture

ILP Instruction Level Parallelism

MIC Many Integrated Core

MPI Message Passing Interface

NUMA Non-Uniform Memory Architecture

OpenACC Open Accelerator API

OpenCL Open Computing Language

OpenMP Open Multi-Processing

PCIe Peripheral Component Interconnect Express

PM Photon Mapping

PPM Progressive Photon Mapping

PPMPA Progressive Photon Mapping with Probabilistic Approach

SPPM Stochastic Progressive Photon Mapping

SPPMPA Stochastic Progressive Photon Mapping with Probabilistic Approach

QBVH Quad Bounding Volume Hierarchy

QPI Quick Path Interconnect

SIMD Single Instruction, Multiple Data

SIMT Single Instruction, Multiple Threads

SM Streaming Multiprocessor

SMX Kepler Streaming Multiprocessor. A redesign of the original SM

SSE Streaming SIMD Extensions

hUMA heterogeneous Uniform Memory Access

List of Figures

2.1 Overview of the Fermi Architecture 9
2.2 Overview of the Kepler Architecture 10
2.3 Overview of the MIC Architecture 11
2.4 Example diagram of a HetPlat 13

3.1 GAMA memory model 17
3.2 StarPU access mode 24
3.3 StarPU dependency management caveat 30

4.1 Ray Tracing 32
4.2 High Level Flowchart of Photon Mapping 34
4.3 Overview of the first step of Photon Mapping 35
4.4 Overview of the second step of Photon Mapping 35
4.5 Specular to Diffuse to Specular path (SDS) 36
4.6 Subsurface Scattering 36
4.7 High Level Fluxogram of Progressive Photon Mapping 37
4.8 Overview of Progressive Photon Mapping 38
4.9 Radius Reduction after a Photon Tracing step 39
4.10 Fluxogram of SPPM 40
4.11 Fluxogram of PPMPA 41
4.12 Fluxogram of PPMPA with concurrent iterations 41

5.1 Diagram of SPPMPA computational tasks and execution order 48
5.2 Dependency graph of an iteration of SPPMPA 54

6.1 Input scenes 61
6.2 CPU implementation measurements 62
6.3 GPU implementation measurements 63
6.4 StarPU implementation, CPU-only 64
6.5 StarPU implementation, CPU sequential vs sequential schedulers 64
6.6 Avg. iteration time of the different schedulers with GPU devices 66
6.7 Calibration process starting with an empty performance model 67

6.8 Best cases for each different implementation and scheduler 68
6.9 Speedup with concurrent iterations 69

Chapter 1

Introduction

1.1 Context

Heterogeneous platforms are increasingly popular for high performance computing, with an increasing number of supercomputers taking advantage of accelerating devices in addition to the already powerful traditional CPUs, to provide higher performance at lower costs. These accelerators are not as general-purpose as a conventional CPU, but have features that make them more suitable to specific, usually highly parallel tasks, and as such are useful as coprocessors that complement the work of conventional systems.

Moore's law [1, 2] predicted in 1975 that the performance of microprocessors would double every two years. That expectation has since driven the development of microprocessors. The number of transistors, and high clock frequencies of today's microprocessors is near the limit of power density, introducing problems such as heat dissipation and power consumption. Facing this limitations, research focus was driven towards multi-core solutions.

This marked the beginning of the multi-core era. While multi-core systems were already a reality, it was not until this point that they reached mainstream production, and parallel paradigms began to emerge as more general-purpose solutions.

In addition to regular CPUs, other types of devices also emerged as good computational alternatives. In particular, the first GPUs supporting general purpose computing were introduced by NVidia early this century.

These devices gradually evolved from specific hardware dedicated to graphics rendering, to fully featured general programming devices, capable of massive data parallelism and performance, and sometimes provide lower power consumptions. They enable the acceleration

of highly parallel tasks, being more efficient than CPUs on specific tasks, but also more specialized. The usage of GPUs for general computing has been named General Purpose GPU (GPGPU), and has since become an industry standard. As of 2013, over 30 of the TOP500's[1] list were powered by GPUs. This increased usage is motivated by the effectiveness of these devices for general-purpose computing.

Other types of accelerators recently emerged, like the recent Intel Many Integrated Core (MIC) architecture, and while all of them differ from the traditional CPU architecture, they also differ between themselves, providing different hardware specifications, along with different memory and programming models.

Development of applications targeting these coprocessor devices tends to be harder, or at least different from conventional programming. One has to take into account the differences of the underlying architecture, as well as the programming model being used, in order to produce code that is not only correct, but also efficient. And efficiency for one coprocessor might have a different set of requirements or rules that are inadequate to a different one. As a result, developers need to take into account the characteristics of each different device they are using within their applications, if they aim to fully take advantage of them. Usually, the task of producing the most efficient code for a single platform is very time consuming, and requires thorough understanding of the architecture details, In addition, the inherent parallel nature of these accelerators introduces yet another difficulty layer.

Each device can also be programmed in various ways, ranging from programming models such as CUDA, OpenCL or `pthreads`[2] to higher level libraries like OpenMP or OpenACC. Each of these provides a different method of writing parallel programs, and has a different level of abstraction about the underlying architecture.

The complexity increases even further when it is considered that multiple accelerators might be simultaneously used. This aggravates the already existing problems concerning workload distribution, scheduling, and communication / data transfers.

Recent studies [3, 4] also show that overall speedups when using accelerators should not be expected to be as high as initially suggested. These studies show that, while the measured speedups of multiple applications ported to GPUs were real, the actual reason was not the better overall performance of the device, but actually the poorly optimized original CPU code. Actually, when code is better designed, similar speedups can be obtained in traditional CPUs. This indicates that accelerators should not be regarded as the only source of high computational power, but rather as an additional resource, and the whole system should be appropriately used for better efficiency.

[1] A list of the most powerful supercomputers in the world, updated twice a year (http://www.top500.org/)
[2] POSIX Threads: The standard thread management library for most UNIX systems

Current coprocessor devices are most commonly used as accelerators (in the context of general-computing), in a system where at least one CPU manages the main execution flow, and delegates specific tasks to the remaining computing resources. A system that uses different computing units is commonly referred to as an heterogeneous platform, here referred to as a HetPlat. These systems become particularly noticeable in the TOP500 ranking, where an increasing number of top-rated systems are heterogeneous.

Much like the phenomenon seen at the start of the multi-core era, a new paradigm shift must happen to efficiently use a HetPlat. An even greater level of complexity is introduced, since one has to consider not only the multiple different architectures and programming models being used, but also the distribution of both work and data. Current **Het**erogenous **Plat**forms (HetPlats) are distributed systems, since each computing accelerator device usually has its own memory hierarchy. As much as a given task may be fast on a given device, the required data transfers to offload such task may add an undesirable latency to the process, and is currently one of the highest performance bottlenecks of this approach.

Even within a single device, memory hierarchy usage can have a large impact on performance. In a NUMA system, although each device can transparently access all memory nodes, access times will be dependent on where the requested data is pinned. Performance problems arise from this if one considers the multiple CPU devices as one single multi-core CPU. Instead, the topology of the system can be considered when assigning tasks to each individual processing unit, and data transferred accordingly, to avoid expensive memory transactions.

Code efficiency is also becoming extremely volatile, as each new system that emerges usually requires architecture-specific optimizations, making previous code obsolete in terms of performance. There is an increasing need for a unified solution that allows developers to keep focuses on the algorithmic issues, and automate these platform-specific issues, which present a barrier to the development of code targeting HetPlats.

Several frameworks have been developed in recent years to address these issues and to allow developers to abstract themselves from the underlying system. These frameworks usually manage the multiple resources of the system, treating both CPUs and coprocessors as generic computing devices that can execute tasks, and employ a scheduler to efficiently distribute data and workload. Memory management is also a key factor, with memory transfers playing a significant role in today's coprocessor efficiently.

Among these frameworks it is worth to mention MDR [5], Qilin [6], StarPU [7] and GAMA [8]. This dissertation focuses mostly on StarPU, with an overview and a comparative assessment with GAMA.

These frameworks tend to encapsulate work by explicitly use the terms of task and data dependencies, and employ a task scheduler to assign data and workloads to the available resources. The scheduler is considered one of the key features of these frameworks. It may take into account multiple different factors to decide when and where to run the submitted tasks. These factors can range from the architectural details of the detected resources, to the measured performance of each task on each device, which can be supported by a history-based performance model.

1.2 Motivation & Goals

HetPlats and associated development frameworks, can still be seen as a recent computing environment, especially when considering the volatility and constant evolution of computing systems. As such, there is still much to develop when it comes to the efficient usage of a HetPlat.

GAMA is a recent framework under development at University of Minho and University of Texas at Austin, which aims to provide tools for developers to deploy dynamic applications, that efficiently run on these high performance computing platforms [8]. This framework is still under development, and currently supports only x86-64 CPUs and CUDA-enabled GPUs. It is somewhat inspired in a similar framework, StarPU, but with an emphasis on the scheduling of irregular algorithms, a class which presents extra problems when dealing with workload scheduling. Although GAMA has been deeply tested for a wide variety of kernels to validate the correctness and efficiency of its memory and execution model, it currently lacks a more intensive assessment, with a more robust and realistic application. Small kernels have a wide range of applications, are deeply studied and optimized, and are a good source for an initial analysis on the performance results of the execution model. However, when considering a real and more resource intensive application, where possibly multiple tasks must share the available resources, other problems may arise. Therefore, a more realistic evaluation of the framework requires a richer set of robust test cases, to complement the performance evaluations made so far.

The initial goal of this dissertation was to perform a quantitative and qualitative analysis of the GAMA framework, applied to a large scale algorithm, to validate its effectiveness, and identify possible soft-spots, especially when compared to other similar frameworks. This would be done through the implementation of a real case study, namely the progressive photon mapping algorithm, used in computer graphics. However, a stable and reliable version of the GAMA framework was not available during the time slot for this dissertation work and the qualitative evaluation of the framework was shifted to a competitor framework, StarPU.

StarPU is an older project, presenting a more polished product, with the same overall goal of efficiently managing heterogeneous systems, but with a different philosophy and approach to the problem.

StarPU is analysed here through the implementation of a computationally intensive algorithm, providing the same analysis on the performance and usage of the framework, while serving as groundwork for future work to assert the effectiveness of GAMA. Additionally, a comparative analysis is also made, in order to establish where each framework excels, and what features a future release of GAMA might require to be competitive against similar approaches with similar goals.

Overall, the work presented here consists on an analysis and a quantitative evaluation of GAMA and StarPU, with the later being used for the implementation of a case study algorithm. Framework-less implementations were also developed to establish the baseline for profiling analysis. The analysis of StarPU through the case study helps in performing a comparison with GAMA, both from the usability point of view, but also in terms of performance. Final conclusions indicate the downsides of each framework, as well as their strengths, and can serve as suggestions for future improvements of the GAMA framework.

1.3 Document Organization

Chapter 2 provides background information relevant to fully contextualize the reader about the technologies and issues being studied. Chapter 3 introduces the two analysed frameworks, and explains their purpose, features and the methodologies behind them.

Chapter 4 presents the progressive photon mapping algorithm, the selected case study, to evaluate the frameworks. An initial background on ray tracing and its evolution is given, followed by the presentation of the algorithm, and its evolutions in the context of this work.

Chapters 5 and 6 focus on the actual work developing the case study, particularly the implementation using StarPU, and its subsequent profiling. Initial scalability results and their analysis are presented along with some considerations regarding the performance of StarPU, as well as the issues found along the way.

Finally, Chapter 7 presents the final conclusions of this work, and Section 7.1 leaves suggestions of future work on relevant topics that could not be fully covered during this dissertation.

Chapter 2

Computing Background

2.1 Parallel Computing

Traditional computer programs are written in a sequential manner. It is natural to think of an algorithm as a sequence of steps that can be serially performed to achieve a final result. This has actually been the most common programming paradigm since the early days of computing, and it synergizes well with single processor machines. Optimizations related to instruction-level parallelism, such as out-of-order execution, pipelining, branch prediction or superscalarity, were mostly handled by the compiler or the hardware itself, and transparent to the programmer. Vector processing was also a key performance factor, allowing the same instruction to be simultaneously applied to a data vector, rather than one at a time (commonly referred to as Single Instruction, Multiple Data (SIMD)).

But in the beginning of the XXI century, the development of computational chips shifted from a single faster core perspective, to a multi-core one. The evolution of single-core processors was already reaching its peak, and was slowing down due to the increasing difficulty in reducing transistor size or increasing clock frequencies, while introducing or aggravating other problems, such as heat dissipation, which becomes harder with the increased complexity of a chip. The solution was to move to a multi-core perspective, coupling more cores in the same chip, to share the workload and allow overall computational capabilities to keep evolving.

This has allowed hardware development to keep in conformance with Moore's Law. And while it was a necessary step from a hardware's perspective, this has important implications in software development. In order for an application to take advantage of multi-core technology, it needs to be broken into smaller tasks, that can be independently executed, usually with some form of synchronization and communication between them. Writing parallel al-

gorithms requires an adaptation to this new paradigm, as a sequential line of execution does not provide efficient results in platforms that support parallel execution of several threads or processes.

Writing parallel algorithms is generally not a trivial task compared their sequential counterpart. Several classes of problems are introduced to the programmer such as deadlocks, data races and memory consistency. Some of these problems may cause applications to behave unexpectedly under certain conditions. That, along with the fact that multiple execution lines are being processed in a sometimes very loose order, is also what makes the debugging of these applications much harder.

This is not helped by the fact that current development environments are still mostly unequipped to aid the programmer in such tasks. Support for debugging and profiling is still somewhat outdated in various cases, as should be expected from a paradigm that has not become mainstream until recent years.

2.2 The GPU as a Computing Accelerator

With the increasing demand for highly data-parallel algorithms, and the growing amount of data to process, hardware development started shifting towards the goal of solving that problem. Initially, that started with the increased support for vector instructions in common CPUs, and the SIMD model. This allowed a single instruction to operate on a set of elements at once, effectively achieving a kind of parallelism which is extremely useful in data-parallel applications.

This data-parallel model is also behind the architecture of GPUs, but at a much higher degree. While the concept is the same (applying the same instruction to a set of values, instead of a single value at a time), the architecture of a GPU, particularly a CUDA-enabled device, relies on the usage of several simple cores, grouped in multiple Streaming Multiprocessors, to achieve higher degrees of parallelism, and process massive amounts of data. These features makes GPUs more suitable for tasks with a high degree of data-parallelism, and not the ideal candidate for more irregular problems, where its more difficult to find parallelization points in order to take advantage of the SIMD model.

Although the hardware of a GPU is still tightly coupled with graphics processing and rendering, there have also been several advances in its usage as a general computing device (GPGPU).

2.2.1 Fermi Architecture

The Fermi architecture was an important milestone of GPUs technology, as it was one of the first generations targeted directly towards GPGPU and high performance computing, rather than purely graphics rendering. The first Fermi devices were released in 2010, and include more efficient support for double precision floating point number when compared to earlier generations. Fermi devices also included a GDDR5 memory controller with support for Direct Memory Access (DMA) through the PCIe bus, and up to 16 Streaming Multiprocessor (SM), for a total of up to 512 CUDA cores.

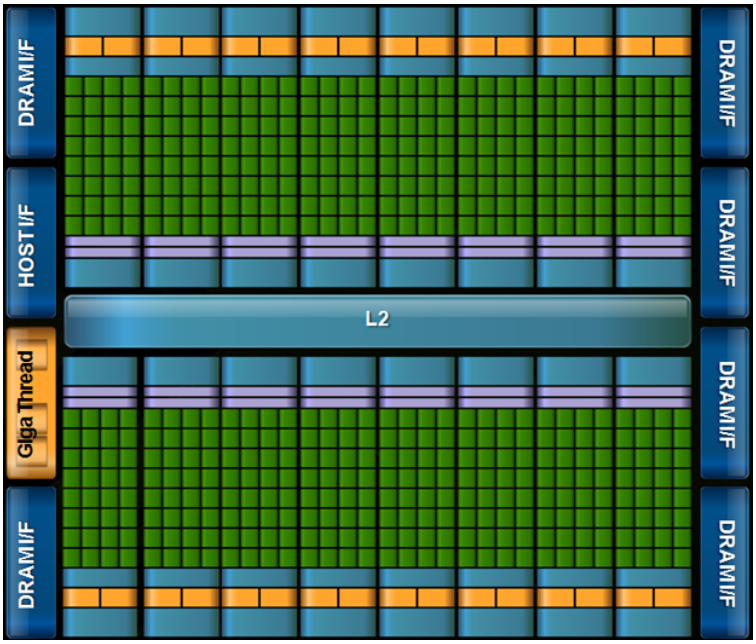


Figure 2.1: Overview of the Fermi Architecture (from NVidia documentation)

This architecture is backed by a hardware-based thread scheduler, within each SM, that attempts to feed the execution unit with threads grouped in blocks of 32, or *warps*. Since the scheduling is made directly via hardware, the switch between threads is nearly free, at least when compared with software scheduling on a CPU. As a result, this strategy works better when the total amount of threads competing for resources is much higher than the amount of execution units, allowing for the latency of memory accesses to be hidden away by instantly scheduling a different *warp*, effectively hiding memory latency while still keeping execution units busy. This is very different from CPU scheduling policies, where switching between threads requires a context switch, which takes considerably longer, making that approach not as feasible as for a GPU.

2.2.2 Kepler Architecture

Kepler, the follow-up generation to Fermi, is in many ways similar to its predecessor. One of the most notorious change is the increase in the total amount of available CUDA cores, going up to 2880 in high-end devices due to the redesign of the Streaming Multiprocessor, now called SMX, each one with 192 CUDA cores. Kepler works at lower frequencies than Fermi, due to the removal of shader frequency, a compromise to make room for the extra CUDA cores. The entire chip now works based on the same core frequency. Overall, individual core efficiency is lowered, but the global system becomes more efficient.

Figure 2.2: Overview of the Kepler Architecture (from NVidia documentation)

The programming model has been extended with the addition of dynamic parallelism, allowing a CUDA thread to spawn new threads, a feature not possible with previous NVidia devices. This is a relevant feature for irregular algorithms. With Kepler we can also invoke multiple kernels for a single GPU, transparently dividing them between the available SMXs.

This shows a clear evolution over the previous generation, and a response to the increasing demand for highly parallel computational power provided by GPUs.

2.3 MIC Architecture

Many Integrated Core (MIC) [9] is an architecture proposed by Intel as a competitor to GPUs for general purpose high performance computing. This alterative device employs 61 cores with a micro-architecture similar to *x86*, with a 1GHz frequency, and up to 16GB of GDDR5 memory, as well as two levels of cache. The last level is interconnected via a bidirectional ring among all cores, effectively sharing memory creating a last level cache with over 30MB. The MIC is based on the *x86* architecture that is used in common CPUs. Extensions to the micro architecture provide support for 64-bit addressing, and SIMD instructions are possible using 512-bit registers. However, instruction sets such as Streaming SIMD Extensions (SSE) or Advanced Vector Extensions (AVX) are not compatible.

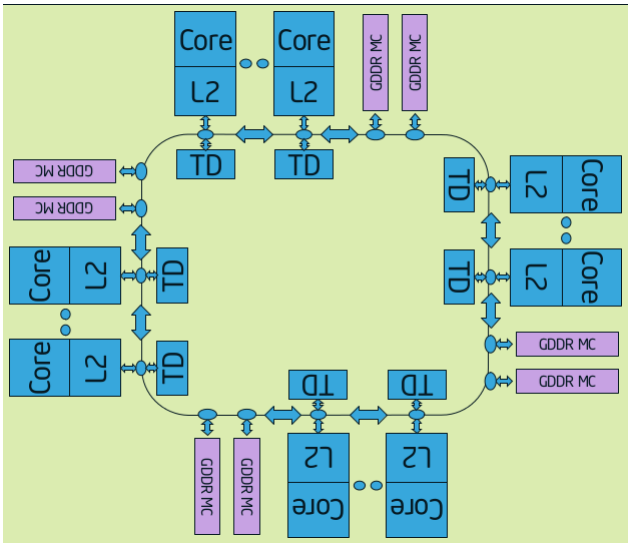


Figure 2.3: Overview of the MIC Architecture (from www.theregister.co.uk)

A MIC device internally uses a Linux-based micro operating system, thus being able to run independently from the rest of the system, as opposed to other accelerating devices. This allows the device to operate in one of three possible modes:

Native the device is used as an independent system, with applications running entirely on it;

Offload the host system uses the device as a coprocessor, by offloading specific tasks to it. This is the usual mode accelerators such as CUDA devices work;

Message Passing By using MPI, the device can be used simply as another node in the network.

Initial claims suggest that this architecture has the ability of providing performance improvements to existing applications, with little to no effort from the programmer. The com-

patibility (to some degree) with *x86* makes this somewhat possible, although some reports have questioned this claim, particularly when attempting to running native *x86* application originally target at Xeon CPUs [10].

2.4 Heterogeneous Platforms

By combining multiple devices such as CPUs and GPUs, the obtained platform can obtain a much higher theoretical peak performance. But in practical terms, it becomes much harder, if not impossible at all, to achieve performance levels near this peak value.

On a homogeneous machine, the actual peak performance is limited by additional factors such as pipeline bubbles[1], or memory access times, which is aggravated in memory bound algorithms [11].

When dealing with multiple devices with disjoint memory address spaces, an additional layer of complexity is introduced, as these different devices need a mean of communication between each other. Usually, a HetPlat, such as the one represented in Figure 2.4 can be seen as a distributed memory system, since each device has its own memory hierarchy. Communication is thus needed for synchronization between tasks, and for all required data transfers.

In the case of GPUs, communication is done via a PCIe bus. Although this technology has greatly evolved over the last recent years, this still proves to be a potential bottleneck for accelerating applications with a GPU.

Even disregarding PCIe devices, a single computational node can also be composed of multiple CPU sockets, connected to each other via a bus such as Quick Path Interconnect (QPI). This bus connects not only the multiple sockets but also the memory banks possibly associated with each one, as these types of machines are generally associated with a NUMA memory hierarchy, where each CPU is directly connected to a different memory bank. While access to all banks is shared between all CPUs, access costs are not uniform. They depend on the bus bandwidth used, and also on the actual path a request has to make from the socket it originated from until it reaches the desired memory bank.

While this is a completely transparent process to the programmer, it can be another source of performance degradation, if most or all data might ends up pinned to a single memory bank, requiring all other sockets to share access to it. This is commonly disregarded by programmers, who end up treating these systems as if their main memory is unified. It is

[1] A delay in the instruction pipeline, required to solve data, structure or control hazards, and limiting Instruction Level Parallelism (ILP)

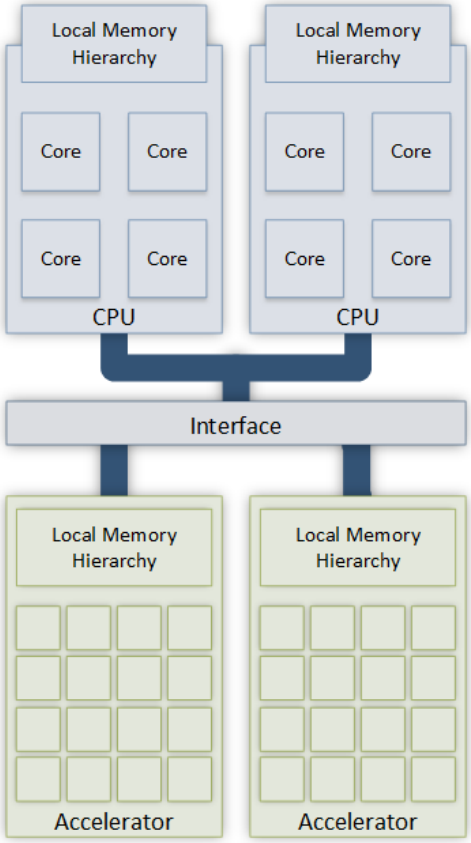


Figure 2.4: Example diagram of a HetPlat

possible to control the affinity of both data and task in a NUMA system. For instance, the `hwloc`[12] library provides access to the topology of the system, as well as an API to control their bindings, effectively allowing control over each individual CPU and data bank.

However, libraries such as `hwloc` are very low-level ones, thus difficult to learn and use properly. For complex problems, worrying about low-level optimizations such as memory and core affinity can lead to over-optimization, making the application much more complex, less portable and harder to debug or modify. Instead, these libraries should be used as a control layer for other libraries or frameworks to work on top of, and allowing a more developer-friendly access to the desired features.

2.5 Heterogeneity and the Future

As stated before, current HetPlats are usually composed with CPUs and GPUs communicating through a PCIe slot. Newer technologies such as the Intel MIC also use this interconnect technique. Both these accelerators have their own memory hierarchy, different from the host CPU, and the latter one is actually *x86*-based, making it architecturally more

similar to ordinary CPUs. However, as history has shown, this cannot be assumed as the global model for heterogeneous computation.

As an example, power efficiency has recently become a more prominent factor in technological evolution. Judging from that, it should be expected that short term evolution would yield more power efficient devices, rather that just providing a higher peak performance or increasing the number of cores. Power efficiency can also be regarded as a relevant factor by a scheduling framework, by making decisions not only based on overall execution time of the application, but also based on its costs in terms of energy consumption. Other metrics can be though of, and their importance increased or decreased, depending on technological trends at each time.

In fact, in early 2013 AMD has already announced a new approach to heterogeneous computing, called hUMA, which effectively implements a shared memory system between the host CPU and a GPU [13]. This eliminates the communication required between both devices through the PCIe bus, which can lead to significant advances in performance. However, such as system will also be incompatible (or at least, not adequate) to any framework, or programming methodology, that assumes a distributed memory system, and the need of PCIe communication.

All these factors emphasize the fact that optimizations based solely on architectural details (for example, software pre-fetching, core affinity and memory affinity) are not desirable to be coupled with an application if is desirable for it to perform well in future platforms, instead of becoming less performant. Such optimizations create tight dependencies between a program's performance, and the actual architecture(s) being used during development. So, it becomes desirable that such issues be handled by a more generic tool, built in a modular way, so that components can be plugged, unplugged or updated, and applications easily portable, while maintaining its efficiency.

Chapter 3

Frameworks for Heterogeneous Platforms

The challenge of efficiently scheduling the workload of an application and its associated data across an entire heterogeneous system is an ambitious one. Usually scheduling involves history based sampling, and memory management. Some degree of complexity is added when trying to have the framework manage data and workload, which makes it sometimes difficult to use with existing code, and keep compatibility with external libraries

Among the several frameworks that have been proposed to target this goal are GAMA and StarPU, which are here presented in more detail. While both of them have common points in their design goals (GAMA is to some degree inspired by StarPU), they followed slightly different approaches to manage HetPlats. In the case of GAMA, a case study is also presented in this chapter. For StarPU, which was the chosen framework fore a more deep analysis, the case study used was the progressive photon mapping algorithm, which is presented later in Chapters 4 and 5

3.1 GAMA

The GPU And Multi-core Aware (GAMA) is a framework to aid computational scientists in the development or porting of data-parallel applications to heterogeneous computing platforms. Currently HetPlat support includes only systems composed of *x86-64* CPU cores and one or more CUDA-enabled GPU devices.

GAMA provides an abstraction of the hardware platform, attempting to free the programmer from the workload scheduling and data movement issues across the different resources. In GAMA, an application is composed of a collection of jobs, each defined as a set of tasks applied to a different input data set. Every job shares the same global address space, instead

of directly using the private memory of each device.

One of the main focuses of GAMA is to efficiently execute irregular applications, which are particularly harder to make estimations on. In an irregular application, input data sets and their sizes cannot be used to make assumptions about how the task will perform under a similar, but not equal set of input. This does not hold true for regular applications, which is what makes their scheduling easier. As such, irregular applications can be more difficult to extract parallelism from, especially when workload is distributed, as is the case with HetPlats and their frameworks.

3.1.1 Memory Model

GAMA uses an unified programming model that assumes a hierarchy composed of multiple devices (both CPUs and GPUs), where each device has access to a private address space (shared by all computing units, or cores, within that device), and a distributed memory system between devices. The framework assumes that multiple computing units of an individual device can cooperate and communicate through a shared memory space, and that the underlying programming and execution model of that device provides synchronization mechanisms (barriers, atomics and memory fences). To abstract the distributed memory model that is used between devices, GAMA introduces a global address space. Figure 3.1 illustrates how GAMA understands the memory hierarchy of a HetPlat.

Memory Consistency

Communication between memory spaces of different devices is expensive due to the need of synchronization and data transfers between the host CPU and the devices. Due to this, a relaxed consistency model is used, which enables the system to optimize data movements between devices, offering the developer a single synchronization primitive to enforce memory consistency.

Software Cache

Some applications require safe exclusive access to specific partitions of a data set. To address this issue, a software cache shared between devices was implemented. This ensures that the required data is as close to the device as possible, taking advantage of the local memory of each device. It also provides a safeguard mechanism in combination with the global memory system, to ensure that each device has a copy of a specific partition, when

requested by the application. Additionally, the cache local copies on the device shared memory space use semantically correct synchronization primitives within the device.

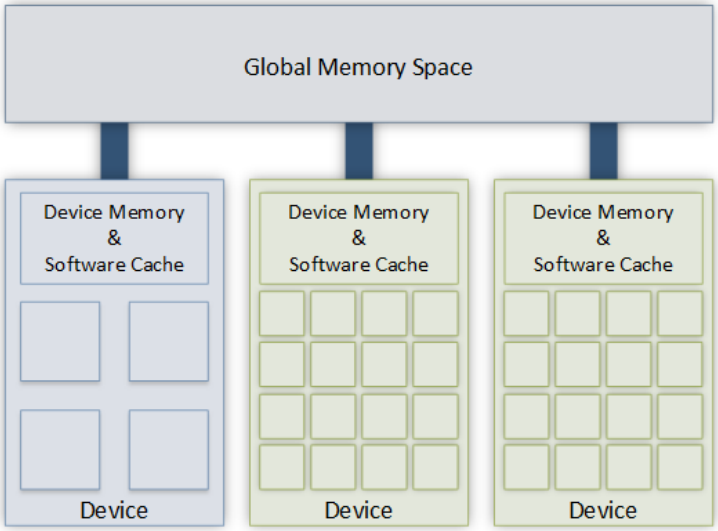


Figure 3.1: GAMA memory model, an extension to the one shown in Figure 2.4

3.1.2 Programming and Execution Model

To better understand the programming and execution model employed in GAMA, some key concepts must be introduced:

Computing Unit (CU)

In GAMA, a Computing Unit is an individual execution unit, capable of executing a general-purpose application. In the context of a CPU, a Computing Unit represents a single core, while on a GPU, in the current implementation represents a single Streaming Multiprocessor (SM). Thus the terms CU and CPU-core may be used with the same meaning.

Device or Worker

Represents a collection of Computing Units that share some level of memory (e.g. the CPU cores on the same machine, or the SMs of a single GPU).

Host

The group of all devices within a single computational node. Currently GAMA supports only a single-host, taking advantage of multiple computational nodes. As such, the host represents the top-most hierarchy layer in GAMA's execution model

Domain

A global view of a particular data structure that enables developers to access any memory location using the global address space, and hiding the complexity of the underlying memory system. At the application level, the user is able to define filters of partial views of a single domain, allowing the system to identify the required communication primitives and enforce the global address space, the memory consistency model, and cache and synchronization mechanisms.

Job

A tuple associating data domains with the corresponding computations related to it (the computational kernel), and a specialized dicing function that defines the best strategy for job granularity, recursively splitting the job into smaller tasks across the data domains. This dicing function is somewhat analogous to the ability of defining task granularity with tools such as OpenMP, but it can employ more flexible solutions, to account for the irregularity of the algorithms.

Kernel

The computation associated with a job. In a best-case scenario, a computational kernel can be mapped directly to a given device simply with the help of the toolkit supporting that device. In most cases, however, the kernel needs to be tailored to a specific device's programming model. This is achievable by extending the job description with the addition of the specialized kernel for a specific device. This feature also enhances the programming model by enabling developers to tailor specific computational kernels for each platform, taking advantage of architecture-specific features.

The organization of the execution model between computing units, Devices and Hosts ensures that a consistent model can be implicitly assumed, where CUs within the same device share a common address space, allowing the usage of device-specific synchronization mechanisms to manage the coordination of concurrent executions within that device.

An application in GAMA is a collection of jobs submitted to a run-time system for scheduling among the available computational resources. Dependencies among jobs can be specified with explicit synchronization barriers. The main goal of the runtime scheduler is to reduce the overall execution time of any given application.

Scheduling

The scheduler uses information provided by each job in order to determine the best scheduling policy, based on current runtime states, as well as execution history. If the granularity of a job is too coarse to enable a balanced scheduling policy, GAMA will recursively

employ the dicing function of a job to adjust the task granularity to the capabilities of the devices.

Internally, GAMA uses a variant of the Heterogeneous Earliest Finish Time (HEFT) scheduling algorithm [14], which uses the computation and communication costs of each task, in order to assign every task to a device in such a way that minimizes the estimated finish time of the overall task pool. This variant of HEFT instead attempts to make a decision every time it is applied to the task pool, so that tasks on the multiple devices take the shortest possible time to execute [15].

3.1.3 The Polu Case-Study

Preliminary tests of the GAMA capabilities were performed in the early training stages, which included well studied cases, such as the SAXPY and the Barnes-Hut algorithms.

Later, an implementation of a small, first order finite volume method was implemented using GAMA, using the previously implemented versions of that same algorithm for comparison references. These versions included a sequential implementation, and two parallel ones, one with OpenMP, and another with CUDA. The details of the algorithm are described in more detail below.

The `polu` application, computes the spread of a material (e.g. a pollutant) in a bi-dimensional surface through the course of time. This surface is discretely represented as a mesh, composed mainly of edges and cells. The input data set contains data on the mesh description, the velocity vector for each cell and an initial pollution distribution.

`Polu` has already been the subject of a parallelization study in [16], which described the incremental work where the application was improved from a sequential implementation, first through a process of analysis and sequential optimization, and then subject to parallelization using two techniques, a shared memory CPU implementation with OpenMP, and a GPU implementation with CUDA.

The Algorithm

The `polu` algorithm is a first order finite volume method, where each mesh element only communicates directly with its first level neighbours in the mesh, a typical case of a stencil computation. The algorithm is very irregular in terms of memory access patterns, since the mesh input generator, `gmsh`, suffers from deep locality issues, turning memory accesses ordered by cells or edges close to random.

The execution of the algorithm consists of looping through two main kernels, advancing in time until an input-defined limit is reached. These two kernels are:

`compute_flux`
: In this step, a flux is calculated for each edge, based on the current pollution concentration of each of the adjacent cells. A constant value, the *Dirichlet* condition, is used for the boundary edges of the mesh, replacing the value of the missing cell. This flux value represents the amount of pollution that travels across that edge in that time step.

`update`
: With the previously calculated fluxes, all cell values are updated, with each cell receiving contributions from all the adjacent edges. After this, one time step has passed.

Implementation

To run the algorithm using the framework, both kernels had to be re-written, following the GAMA model of jobs. Data structures were also re-written to make use of the facilities provided by GAMA to allow memory to be automatically handled by the global address space. This presents an obviously large amount of development work, since almost everything had to be re-written according to the GAMA rules. However, it has to be taken into account the fact that this additional work also had to be performed in the previous implementations studied in [16], since most of the original code was not suitable for efficient parallelization.

From this, one initial consideration can already be made about the framework, in the sense that the effort required to parallelize an application following the GAMA rules might be too high, if a given application was already written for a parallel environment. Since specific data structures and job definitions need to be used, this may hamper the adoption of GAMA by already implemented solutions, unless the performance advantages are significant enough to justify the additional effort.

Study limitations

Several restrictions apply to the input generation for this algorithm. In particular, the utility required to generate a mesh with arbitrary resolution has an estimated complexity of $O(N^3)$, which prevented large enough test cases to be generated. The largest available input contained only around $400,000$ cells, and represented a total memory footprint of just over $40MB$, which is extremely small, and does not allow a good enough analysis

on resource usage. With such a low resource occupancy, the scheduling policy employed by GAMA will most likely assign all the workload to a single device, as the cost of data transfers, and the low execution time for each kernel for such a small data set would not justify otherwise. Additionally, this being a typical stencil, each iteration requires a barrier, allowing no execution of two simultaneous iterations, which would be an additional way of improving parallelism.

Knowing this, any result obtained by profiling the `polu` application under these conditions would not provide a correct insight about the algorithm, or about the framework, and as such, these results are not presented here. The `polu` test case still served as an initial basis to gain some insight into GAMA, and to better prepare the implementation of a more complex case study, the progressive photon mapping algorithm.

3.2 StarPU

StarPU [7] is a unified runtime system consisting on both software and a runtime API that aims to allow programmers of computational intensive applications to more easily exploit the power of available devices, supporting CPUs and GPUs.

Much like GAMA, this framework frees the programmer of the workload scheduling and data consistency inherent from a HetPlat. Task submissions are handled by the StarPU task scheduler, and data consistency is ensured via a data management library.

However, one of the main differences comes from the fact that StarPU attempts to increase performance by carefully considering and attempting to reduce memory transfer costs. This is done using history information for each task and, accordingly to the scheduler's decision of where a task shall be executed, asynchronously prepare data dependencies, while the system is busy computing other tasks. The task scheduler itself can take this into account, and determine where a task should be executed by taking into account not only the execution history, but also the estimation of data transfers latency.

3.2.1 Terminology

StarPU uses a well defined terminology to describe its libraries and API:

Data Handle

References a memory block. The allocation of the required space, and the possibly

required memory transfers to deliver information to each device can be completely handled by StarPU;

Codelet
Describes a computational kernel that can be implemented in one or more architectures, such as CPUs, CUDA or OpenCL. It also stores information about the amount and type of data buffers it should receive;

Task
Is defined as the association between a codelet and a set of data handles;

Partition
The subdivision of a data handle in smaller chunks, according to a partitioning function, which can be user defined;

Worker
A processing element, such as a CPU core, managed by StarPU to execute tasks;

Scheduler
The library in charge of assigning tasks to workers, based on a well defined scheduling policy.

3.2.2 Task Scheduling

The framework employs a task based programming model. Computational kernels must be encapsulated within a task. StarPU will handle the decision of where and when the task should be executed, based on a task scheduling policy, and the available implementations for each task. A task can be implemented in multiple ways, such as CPU or CUDA. Multiple implementations for the same device type can also be used. This allows StarPU to automatically select the appropriate implementation even between different CPU micro architectures (i.e. some CPUs might perform better with an implementation that uses SIMD extensions). When submitting a task, StarPU will use the selected scheduler to select which of the available implementations will be used. The decision varies from scheduler to scheduler, but can take into account information such as the current availability of each resource, the performance model already obtained for that task, and estimations regarding the required data transfers to solve dependencies.

Data handled by a task is automatically transferred as needed between the various processing devices, ensuring memory consistency and freeing the programmer from dealing directly with scheduling issues, data transfers and other requirements associated with it.

Previous work by the StarPU development team indicates that one of the most important issues with scheduling is about obtaining an accurate performance model for the execution time of a task [17, 18]. This is increasingly difficult when data transfers, which the team regards as a key issue, are taken into account, as shown in the latter paper. In it, a data-prefetching implementation for GPUs is present, and asynchronous data request capability is introduced as part of the StarPU library, with the goal of preventing computing units from being stalled waiting for data.

3.2.3 Dependencies

StarPU automatically builds a dependency graph of all submitted tasks, and keeps them in a pool of *frozen tasks*, passing them onto the scheduler once all dependencies are met.

Dependencies can be implicitly given by the data handled by the task. Each task receives a set of buffers, each one corresponding to a piece of data managed by StarPU data management library, and will wait until all the buffers from which it must read are ready.

This includes the possible data transfers that are required to meet dependencies, in case different tasks that depend on the same data are scheduled to run on different computational nodes. StarPU will automatically make sure the required data transfers are made between each task execution to ensure data consistency.

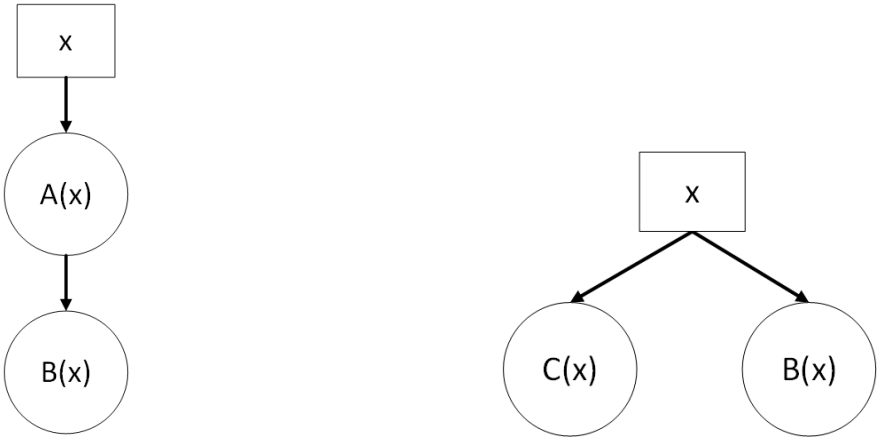
In addition to implicit data dependencies, other dependencies can be explicitly given in order to explicitly force the execution order of a given set of tasks.

Data Access Modes

Each data dependency that is explicitly defined in a task can have a different access mode. Data can be used in read-only, write-only or read-write mode. This access mode does not serve the purpose of ensuring memory correctness. It is used to soften task dependencies by using a *Multiple Readers / Single Writer* model in dependency calculation.

This model describes a type of mutual exclusion pattern where a data block can be concurrently accessed by any number of reader, but must be exclusively accessed by a writer. StarPU uses this concept to further optimize data dependency calculations. If multiple scheduled tasks depend on the same data handle, but only with reading access, then that dependency should not block the multiple tasks from running concurrently (see Figure 3.2). Temporary copies of the data can be created, possibly on different computing units, and later discarded, since a read-only buffer is assumed to remain unchanged at the end of a

task.



(a) A writes to x. B must wait for it to finish to solve the dependency

(b) Neither B nor C write to x, so they can be ran concurrently

Figure 3.2: Access mode illustration; Task A requires read/write access to x, while tasks B and C require read-only access

3.2.4 Virtual Shared Memory

The approach used by StarPU when it comes to memory management is simpler than the model employed by GAMA. The purpose is the same: to automatically manage memory allocations and transfers on all devices. This not only frees the programmer from the work of manually managing memory between tasks, but it also has the potential to lower the cost of such operations. StarPU manages memory by forcing the user to declare data handles to their data. These handles are used as arguments for tasks, allowing the scheduler to allocate and transfer all required data buffers to the correct computing unit prior to the task execution.

3.2.5 Multi-threading

In order to have an accurate view of the topology of the system, the framework can optionally use the `hwloc` library[1] [12] to detect the structure of the architecture, including all CPU sockets, NUMA nodes and cache hierarchy.

A tree is created representing the complete hierarchy of the system. The latest version of the framework also introduce support for parallel tasks, with the concept of combined workers. These workers exist in cases where the system has multiple computing units in the

[1] `hwloc`, or Portable Hardware Locality, is a package that provides access to an abstract topology of modern computating architectures.

same node (such as CPU sockets where multiple cores share a cache hierarchy). In these situations, and if StarPU is using a parallel-task-aware scheduler (currently only `pheft` and `peager` exist), it is possible to specify the maximum degree of parallelism for a function. A combined worker can then be assigned to such tasks using, for example, OpenMP to take advantage of the multiple available cores.

3.2.6 API

Two API versions are available: the high-level, `pragma`-based[2] API, and a low-level version. The `pragma`-based API exposes StarPU's main functionality with a set of directives that can be added to the code, to embed StarPU within it. It is particularly suited for less experienced developers, or developers who simply need to focus completely on the algorithmic issues with less or no knowledge on the underlying parallel hardware being used.

The directives can be easily disabled, restoring the application to its original, StarPU-free nature, which also makes it ideal to add StarPU into already existing code, or applications that require a large degree of portability.

The low-level version is a more verbose one, which is actually internally used by the high-level one, and provides a greater degree of control over what StarPU does.

This trade-off is not uncommon, with many existing libraries besides StarPU supporting both API levels. High level versions are designed to remove complexity and accelerate development time. They are often a subset of an underlying low level version, delivering only the more common features. More experienced developers should be able to achieve better results with a lower level API, with the cost of additional development time.

3.2.7 Performance Model

Most schedulers available with StarPU are history based. This relies on the programmer to configure a performance model for each defined codelet, in order to allow the framework to identify it, and use on-line calibration. Calibration works automatically for every task that is not yet calibrated for a given device. StarPU will ensure that a minimum of 10 executions of that task will be processed before using the performance model. Calibration results will be stored in a centralized directory, and inform StarPU about how each codelet behaves on each different device, with a certain input size.

[2]Directives inserted within the code, to instruct the compiler about how to process input. Can be used to extend the compiler and provide additional features, functionality or optimizations

Once a good calibration has been obtained, StarPU can schedule tasks more efficiently, depending also on the chosen scheduling policy. The list of the available policies is the following:

eager The default scheduler, using a single task queue, from which workers draw tasks. This method does not allow to take advantage of asynchronous data prefetching, since the assignment of a task is only done when a given worker requires so;

prio Similar to eager, but tasks can be sorted by a priority number;

random Tasks are distributed randomly, according to the estimated performance of each worker; Although extremely naive, this has the advantage of allowing data prefetching, and minimizing scheduler decision times;

ws (Work Stealing) Once a worker becomes idle, it steals a task from the most loaded worker;

dm (Deque Model) Performs a HEFT-like strategy, similarly to GAMA (see Section 3.1.2). Each task is assigned to where the scheduler thinks its termination time will be minimized;

dmda (Deque Model Data Aware) Similar to **dm**, but also taking into account data transfer times;

dmdar (Deque Model Data Aware Ready) Similar to **dmda**, but tasks are sorted per-worker, based on the amount of already available data buffers for that worker;

pheft (parallel HEFT) similar to **heft** (which is deprecated in StarPU, with **dmda** working in a very similar way), with support for parallel tasks;

peager similar to **eager**, with support for parallel tasks.

Additionally, schedulers are built as a pluggable system, allowing developers to write their own scheduling policies if desired, and easily use them within StarPU.

3.2.8 Task Granularity

The granularity of a task in GAMA can be automatically defined and dynamically readjusted with the use of a dicing function, which recursively adjusts it to find the best case scenario for each particular device. This feature is not available in StarPU, where granularity has to be defined manually by the programmer.

The API gives the ability to subdivide a data handle into smaller children, based on a partitioning function. This partitioning can be defined as a simple vector or matrix block partitioning, but more complex and custom methods can be defined.

After partitioning, each children can be used as a different data handle. This means that in order to operate in chunks of data at a time, one has to first partition data according to a user defined policy, and later submit multiple individual tasks, using each children individually.

3.3 Comparison

Given that HetPlats can suffer major changes in the future, due to the constant technological evolution, a highly important feature of an application or framework is its modularity, so that individual features can be updated to meet the requirements of the constantly evolving computing platforms. StarPU seems to use this philosophy to some extent, with modules such as the scheduler itself being completely unpluggable and rewritable by a developer. It also provides the ability to assign user defined functions to make decisions within the framework, such as how to partition data.

3.3.1 Usability

From a developer's point of view, StarPU, being written in *C* provides an a clear but somewhat outdated API, in some aspects resembling `UNIX`-like libraries. More modern languages such as *C++*, in which GAMA is written provide less verbose and more structured code. The choice for the *C* language is possibly related to better compatibility and portability, but seems to somehow limit the language, and even expose some unexpected behaviours of the framework (see Sections 3.3.3 and 3.3.3).

Even though GAMA does not yet provide a solid API to work from, but a less clear architecture model, it can still be considered harder to work on, although there is plenty of room for improvement, and once development advances and reaches a more stable position, it can have the conditions to be a much more usable framework than the low-level one provided by StarPU.

An important aspect to consider is that GAMA makes the usage of external libraries much more difficult, due to the encapsulation of every data within the wrappers for its global memory system. As a result, libraries that implement their own data structures may have incompatibilities. Kernel calls can also be a problem, since an external library

that attempts to invoke a CUDA kernel will not do so through GAMA's facilities, thus not taking advantage of its scheduling, and possibly also creating incompatibilities, since GAMA expects to manage all data and workload. StarPU is not as restrict as GAMA in this subject, at least for the low-level API. Since it consists of a more low-level usage of the framework, the developer gains more control, and can more freely operate with the underlying data, and manually invoke kernels (leaving to StarPU only the job of deciding in which device to execute them), making the usage of external code a more viable possibility, although not without its limitations.

3.3.2 Scheduling

Both StarPU and GAMA employ a variant of the HEFT algorithm for scheduling on heterogeneous systems, although StarPU gives much more emphasis to the latency caused by memory transactions, and can also support additional scheduling policies to be used.

GAMA has the ability to recursively change the granularity of a task to adapt it to the characteristics of a device. This is presented as an important step by GAMA to automate decisions by the scheduler. Without this, granularity has to be manually defined, by sub-dividing the domain in arbitrarily sized chunks and process each one as an individual task. This is not only a cumbersome task for the developer, but also a possible weakness, as the ideal task granularity is not equal from system to system, or algorithm to algorithm, and may not be easy to determine without intensive testing and manual tuning.

This seems an extremely important feature in GAMA, at least from the development point of view, as the task of finding the ideal granularity is thus automated by the framework. The fact that StarPU does not provides a similar feature can be a limitation for the developer, which has to manually divide the data set (i.e. using partition functions) and invoke each sub task as an individual one.

3.3.3 Memory Management

Type Safety

Perhaps one of the most important limitations of StarPU's API is the fact that its task submission methods are not type-safe[3]. By definition, the C language (in which StarPU is implemented and exposed to the programmer) is only type safe to a certain extent, since

[3]The extent to which a language discourages or prevents type errors

workarounds are often used in the language to achieve results similar to polymorphism or runtime casting.

This is the case with StarPU, resulting in an API that can be mistakenly used by the programmer. Each codelet defined in a program specifies how many data buffers its tasks will depend on, and their access modes. However, since data buffers are used only through their data handles, which are completely generic, no explicit type checking is made to ensure that the correct types of data handles is passed, and that they are properly received within the task. For example, a task may be expecting to receive buffers X and Y of completely different sizes and types. But on task submission, their order might be reversed, resulting in runtime errors which might be extremely difficult to trace.

The main result of this is a weak task submission API, since it can easily lead to runtime errors. More experienced developers might have enough understanding to easily identify these problems. But technical issues such as type safety should not have to be addressed by the developer, as they can pose serious problems to development time, but could be easily identified by a compiler.

Consistency

StarPU ensures consistency of all data assigned to it via data handles, but only within tasks managed by its scheduler. The issue here is that, while the API allows the creation of a data handle associated with an already allocated data structure (usually pinned to main system memory), it is not ensured that tasks will write to that actual memory, and not a StarPU-managed copy. This has the side effect of consistency not being ensured outside the context of a task. Writing to a buffer via conventional methods can thus be considered dangerous.

It is also impossible to change the size of a data buffer once it has been assigned to StarPU. When this is a requirement, it is necessary to destroy and redefine the data handle, which is not a particularly efficient method, and may introduce additional bottlenecks when used between task submissions.

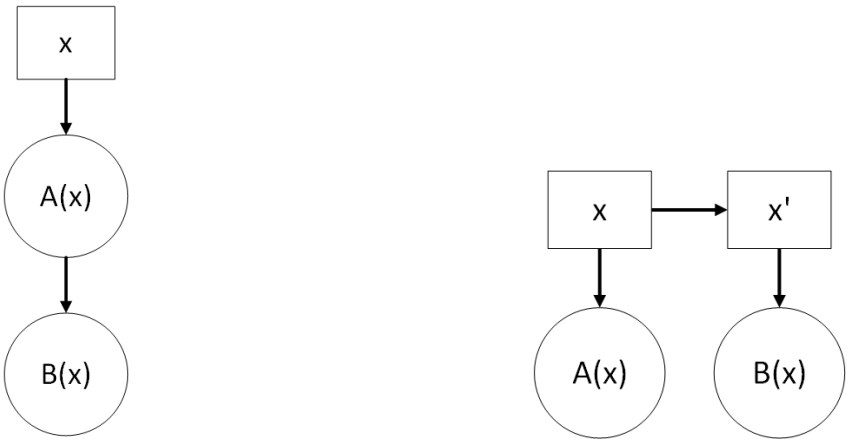
On these subjects, GAMA seems to provide more powerful capabilities. All data that is to be managed by GAMA's unified memory system must be encapsulated in a wrapper that processes all accesses to it. While it is not documented how GAMA actually behaves when data is accessed outside of a job, this model should provide a more transparent solution for consistent memory access.

Data Access Modes

Another caveat of the usage of data handles is that it can lead to incoherent results, again due to possible and easily made developer mistakes. StarPU relies on the estimation of data transfer times to decide where to schedule a task to. An important factor of this comes from the access modes required for each data buffer in a task, as explained in Section 3.2.3.

If a data buffer is declared as read-only within a task, it is assumed to be unchanged by that same task, thus other tasks depending on it will already have that dependency met. The caveat here is that a read-only buffer is only so for the framework, but not actually read-only for the programmer, or for the C compiler itself, and data can actually be overwritten for a read-only buffer. Although that is probably related to a development mistake, it can easily happen nonetheless.

Adding to this is the fact that StarPU might create additional copies of data buffers to solve dependencies faster across multiple devices. Figure 3.3 exemplifies this problem. Tasks A and B have read-only access to the data buffer x, although as explained, both of them can write to it at will. If both tasks are scheduled to the same device (Figure 3.3a), they will be executed in order, and StarPU will reuse the initially existing memory, without the need to create temporary copies of the buffer. In this case, if task A writes to x, task B will later see these changes, since memory is shared.



(a) A and B are executed in order, and the same buffer for x is re-used

(b) A and B are scheduled to different devices, creating a temporary copy of x

Figure 3.3: Illustration of a possible mistake due to dependency management. Both A and B have read-only access to x

In Figure 3.3b, the two tasks are scheduled to different devices. Since StarPU assumes x to be read-only, a copy of it, x' can immediately be created in the additional device. In this case, both tasks will run concurrently, with their own local copy of x. Task A will still write changes to this buffer, but task B will not see them.

Chapter 4

Case Study: the Progressive Photon Mapping Algorithm

The selected case study is one of many algorithms from the ray tracing family, which are typically used in computer graphics for the purpose of realistic image rendering, by computing scene illumination with as much accuracy as possible, attempting to approximate the rendering equation, which is the basis for these algorithms. This chapter presents an overview of ray tracing algorithms, as well as the rendering equation, and follows with the description of the photon mapping algorithm, and its various evolutions until reaching the case study for this dissertation, which corresponds to the combination of all evolutions presented here.

4.1 Ray Tracing

Ray tracing is the designation of a family of algorithms that simulate ray interactions with an environment to determine the visibility of two points in space. Generally, this technique is used to compute light interactions within a three-dimensional scene, and obtain a two-dimensional image with a rendering of that scene. A high degree of visual realism can be achieved by these techniques, but also with greater computational costs.

Ray tracing algorithms are capable of simulating most optical effects, such as reflections, refractions, and scattering, although the actual set of simulated effects, and their overall quality differ between the multiple ray tracing algorithms that have already been developed.

Large efforts have been made recently to improve ray tracing techniques, with several of them being performance-related, by using parallel architectures such as GPUs, which have

always been associated with computer graphics. Some efforts are already showing advances in the area of real time ray tracing, which was unfeasible until very recently [19, 20].

4.1.1 Overview

The first use of ray tracing algorithms was with the introduction of the later called ray casting technique by Arthur Appel in [21]. The idea consisted of casting one ray from an origin point (usually called the eye, or camera) through each pixel of the image, and find the closest object in the scene that blocks the path of the ray. The material properties and light effects in the scene would then determine the final colour of the pixel.

Later advances (e.g. [22]), introduced the recursive ray tracing algorithm. In traditional ray casting, each ray would end after the first hit. This was a limiting factor that prevented the rendering to deal with reflections. Recursive ray tracing solves that, by recursively tracing more rays after each hit. When a ray hits a surface, it can generate new reflection or refraction, depending on the properties of the material hit. Shadow rays can also be cast in the direction of the light sources. If a shadow ray is blocked by an opaque object before reaching the light source, it means that surface is not illuminated by that light source, and thus is shadowed. This technique is illustrated in Figure 4.1

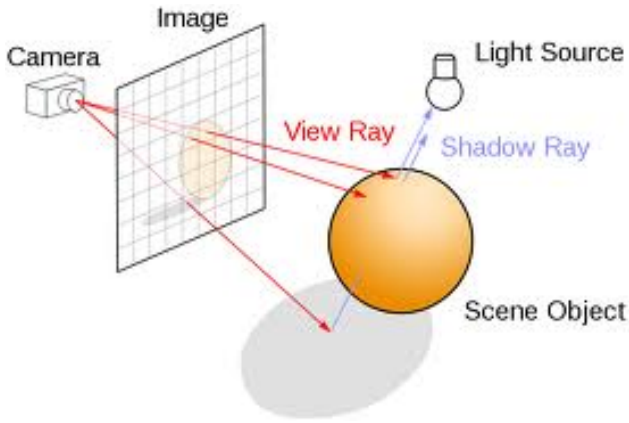


Figure 4.1: Ray Tracing (from Wikipedia)

It is also not uncommon to approximate the rendering equation by using combinations of more than one method, such a Ray Tracing, Radiosity or Metropolis Light Transport [23, 24]. Each method attempts to simulate the travel of light particles across the scene, and model the various interactions with the environment, but with different approaches, advantages and limitations, so the combination of multiple method can combine the advantages of each one.

Traditional ray tracing methods work by simulating light particles traveling from the eye into the scene, being reflected and refracted until they reach a light source. Nowadays, most

ray tracing rendering algorithms are in fact bidirectional, having photons shot from the light sources and their interactions computed in addition to regular eye paths.

Radiosity follows an opposite approach, and simulates the path light takes from the light source, until it reaches the eye. It only deals with diffuse interactions, however, and is commonly used in combination with other techniques.

4.1.2 The Rendering Equation

The realistic simulation of illumination of an environment is a complex problem. In theory, a simulation is truly realistic when it completely simulates, or closely approximates, the rendering equation.

This equation, first proposed in 1986 [25], is based on the laws of conservation of energy, and describes the total amount of emitted light from any given point, based on incoming light and reflection distribution. The equation is presented in Equation (4.1)

$$L_s(x, \vec{\omega}_r) = L_e(x, \vec{\omega}_r) + \int_{\Omega} f_r(x, \vec{\omega}_i, \vec{\omega}_r) L_i(x, \vec{\omega}_i)(\vec{\omega}_i \cdot \vec{n}) d\omega_i \tag{4.1}$$

In short, the equation defines the surface radiance $L_s(x, \vec{\omega}_r)$, leaving the point x in the direction $\vec{\omega}_r$. This is given by $L_e(x, \vec{\omega}_r)$, which represents light self-emitted by a surface, and $L_i(x, \vec{\omega}_i)$, which is the radiance along a given incidence direction. f_r is the Bidirectional Reflectance Distribution Function (BRDF) and Ω represents the semi-sphere of incoming directions centered in x.

4.2 Photon Mapping

Photon mapping is a ray tracing method that intends to approximate the rendering equation, first proposed as a global illumination technique in 1996 [26], and works as a two-pass algorithm, working as an extension to ray tracing that allows the efficient computation of caustics[1] and indirect illumination of surfaces.

Unlike other algorithms such as Path Tracing or Metropolis Light Transport, this is a biased rendering method, meaning that a finite for a finite number of traced photons, the resulting estimation will always be different from the correct solution. However this can be

[1] The light effects caused by light reaching a diffuse surface after being reflected or transmitted by a specular one

worked around by increasing the size of the photon map structure used in the process, or by using variations of this technique, such as Stochastic Photon Mapping.

4.2.1 Algorithm

The original approach to photon mapping consists simply on a two step algorithm. One step to generate a structure with the illumination information for the scene, called the photon map, and a second step to trace rays from the camera to interact with the scene and the photon map, contributing to the final pixels of the generated image.

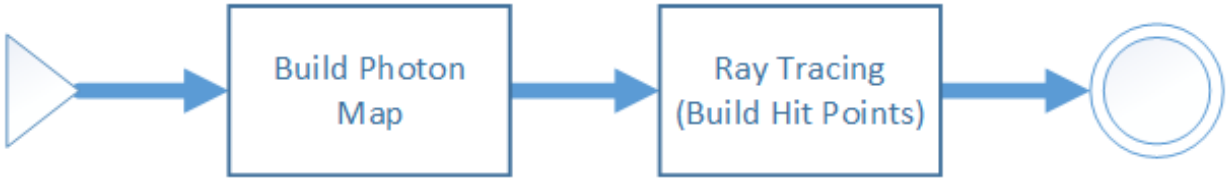


Figure 4.2: High Level Flowchart of Photon Mapping

First Step: Photon Map Construction

In this step, a large number of photons must be traced, starting from the existing light sources in the scene (see Figure 4.3). The tracing of a photon is just like the tracing of a regular ray, interacting with the scene according to BRDF function, in a process similar to path tracing. Every hit by a photon is stored in a structure called a photon map. The original proposal [26] uses two different photon maps, the extra one being a higher density one used for the rendering of caustics. This is done by emitting paths towards specular objects in the scene, and storing them in the photon map as diffuse surfaces. The usage of this extra photon map is not, however, required for the implementation of the algorithm, and serves only as a way of providing additional quality in the rendering of caustics. As such, it was not considered during the implementation in this work.

The photon map structure generated serves as an approximation of the light within the entire scene. As another optional extension, shadow photons can also be cast during this step, which will reduce the amount of shadow rays necessary in the second step to correctly reproduce shadows.

Second Step: Rendering

For the final image render, Monte Carlo ray tracing [27] can be used to traced rays from the camera position into the scene, as illustrated in Figure 4.4. During this step, the

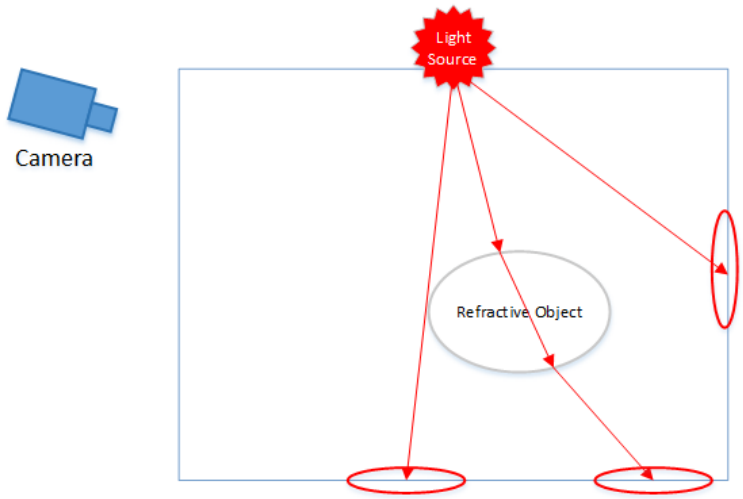


Figure 4.3: Overview of the first step of Photon Mapping

information in the photon map structure can be used to estimate the radiance of a surface, by using the N nearest photons to the hit point, that are within a sphere of radius r centered in the hit point x. The radius r is then used to estimate the surface area covered by the sphere, which is approximated to πr^2.

The photon map proves useful during this step, not only to increase performance, but also to allow the modeling of some light effects that are not present or are inefficient to process without such a structure.

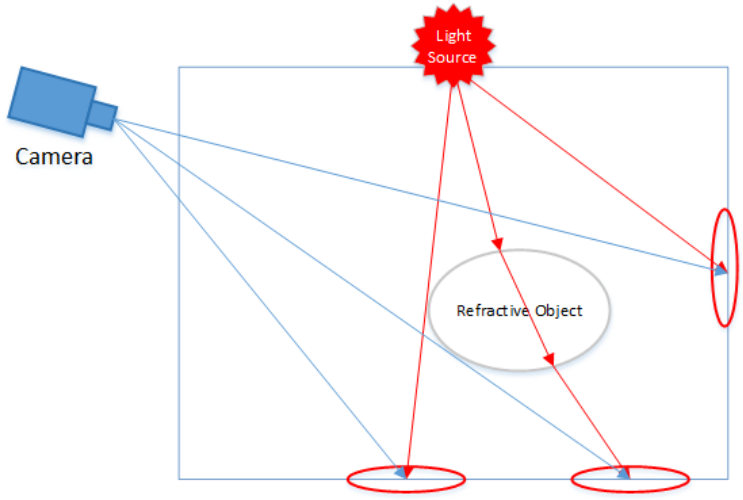


Figure 4.4: Overview of the second step of Photon Mapping

4.2.2 Applications

One particular case where photon mapping provides advantages over other methods is when light is being transported from a light source travels along a specular-to-diffuse path

before reaching the eye (LSDE path), such as the one illustrated in Figure 4.5, before reaching the eye. This is what is commonly known as a caustic, such as, for example, the shimmering light seen of the bottom of a pool, or any light source enclosed in glass. This scenario is very common since most artificial light sources are enclosed in glass, but is particularly hard to simulate, particularly when the light source is small, making the sampling probability very low when using Monte Carlo methods.

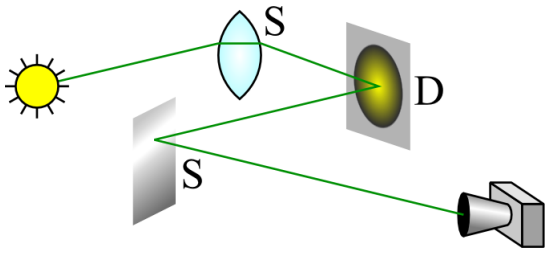


Figure 4.5: Specular to Diffuse to Specular path (SDS)

Another example of a typically hard to simulate effect is Subsurface Scattering, which is observed when light enters the surface of a translucent object, is scattered when interacting with the material, and finally exits the surface at a different point. Generally, light will be reflected several times within the surface before backing out an angle different from the one it would have take had it been reflected by the surface. This is visible in materials such as marble, or skin, and can be seen in Figure 4.6.

Figure 4.6: Subsurface Scattering (from [28])

Both of these can be simulated well by Photon Mapping algorithms, although a high amount of caustics will hinder performance considerably.

4.3 Progressive Photon Mapping

The main problem with the original proposal for photon mapping is that the quality of the final result is limited by the size of the photon map, which in turn has its size limited

by the amount of memory available.

Since effects such as caustics are simulated by directly using the information from the photon map, it is necessary to use a very large number of photons in order to avoid noise in the rendering. Thus, the overall accuracy is limited by the available memory. In other words, the accuracy of photon mapping is not only computationally bounded, but also memory bounded, while usual unbiased methods are only computationally bounded.

[29] proposes a progressive approach to photon mapping, which makes it possible to simulate global illumination, including the effects provided by traditional photon mapping, with arbitrary accuracy, and without being limited by memory. This is done by using a multi step algorithm instead of a two step one, where the first pass consists of a ray tracing step to capture a collection of hit points in the scene, and later multiple photon tracing steps are processed iteratively, with each new iteration improving the accuracy of the result in order to converge to an accurate solution, but without storing photon maps from previous iterations.

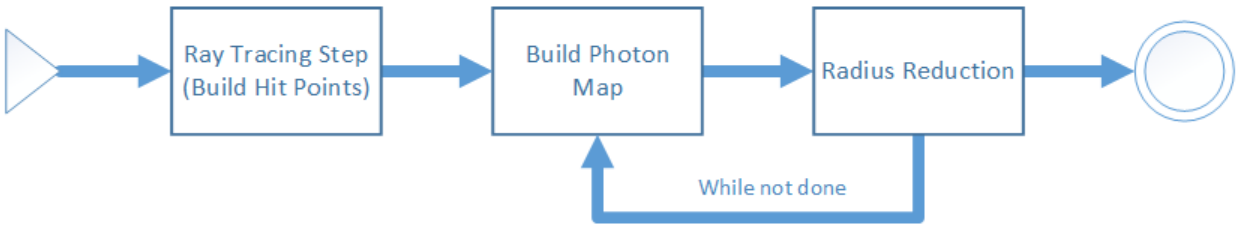


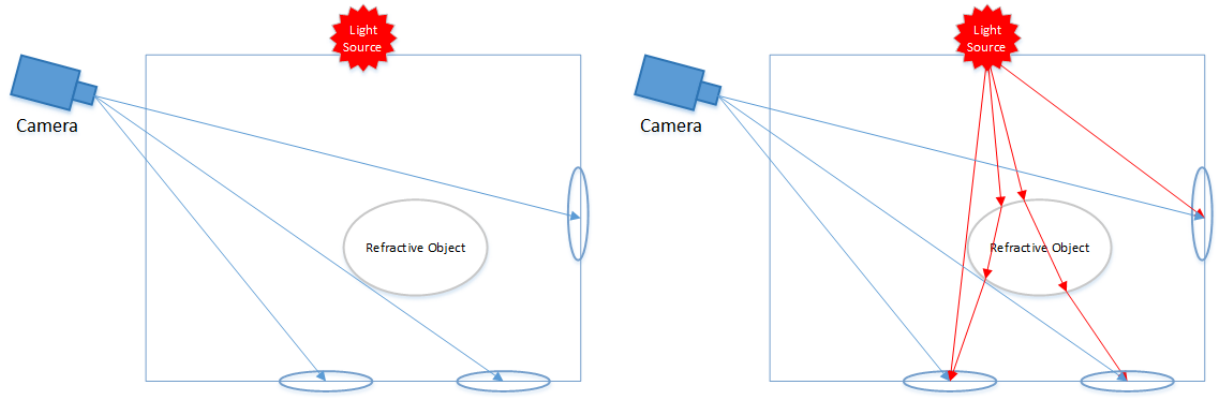
Figure 4.7: High Level Fluxogram of Progressive Photon Mapping

4.3.1 First Step: Ray Tracing

The first step is a standard ray tracing step, used to find all surfaces visible in the scene. Rays are traced through each pixel of the image plane, in order to find all visible surfaces in the scene. For each ray, all hits with a non-specular component in the BRDF function are stored. This includes storing the hit location (x), the ray direction($\vec{\omega}$) and the pixel it originated from. Additionally, data for the progressive estimation is also included, such as a radius, intercepted flux, and number of photons within the defined radius.

4.3.2 Next Steps: Photon Tracing

After the initial ray tracing step, an iterative process begins. Each iteration, a given number of photons is traced into the scene, building a photon map. At the end, all hit points stored from the initial step are processed, to find all the photons in the map that are



(a) First step: Ray tracing **(b)** Next steps: Photon Tracing

Figure 4.8: Overview of Progressive Photon Mapping

within the radius of that hit point. These photons are used to refine the estimate of the illumination of the hit point.

Once this contribution is computed, the photon map is no longer needed, and can be discarded, before the next iteration repeats the process.

This provides two key advantages over the original, two-step approach. The total amount of photons traced is not limited by memory, but only by the amount of iterations that are computed. An arbitrary number of iterations can be used, without requiring any additional memory at all, and resulting in a better quality result. Also, after each photon pass an image can be rendered, and the progressive result can be shown while the image is progressively improved, instead of having to wait for the entire algorithm to finish.

4.3.3 Radiance Estimation

Traditional photon mapping estimates radiance by using the density of photons, given by Equation (4.2). This is based on locating the nearest N photons within a sphere of radius $R(x)$. The surface are is assumed to be flat, and the surface area is approximated to $\pi R(x)^2$. In progressive photon mapping, using this estimation may result in different iterations having different estimations for the same hit point.

$$d(x) = \frac{N}{\pi R(x)^2} \tag{4.2}$$

To solve this, the estimations from each iteration can be averaged to obtain a more accurate estimate. However, the final result will not be more detailed than the result of each individual photon map, which is not desirable. Also, as the radius $R(x)$ is constant

throughout the iterations, small details within that radius cannot be correctly solved, making the overall accuracy limited by the size of each individual photon map.

The solution to this, also proposed in [29] consists of combining the estimate from each photon map in such a way that the final estimation will converge to a correct solution. The key technique is to reduce the radius r at each hit point, for every new photon map computed.

Radius Reduction

Assuming each hit point has a radius $R(x)$, and that $N(x)$ photons have already been accumulated in it, after a new photon tracing step, resulting in $M(x)$ new photons within the radius $R(x)$, the new photon density $\hat{d}(x)$ can be given by Equation (4.3)

$$\hat{d}(x) = \frac{N(x) + M(x)}{\pi R(x)^2} \tag{4.3}$$

The radius reduction step is about computing a new, smaller radius $\hat{R}(x)$ for each hit point, such that the amount of photons within the new radius $\hat{N}(x)$ is greater that the amount of photons that was present in the previous radius. This ensures that the final result is increasingly more accurate, and converges to a correct solution. The radius reduction is illustrated in Figure 4.9.

The proposed approach in [29] simplifies this by using a parameter α to control the fraction of photons to keep after an iteration. Therefore, $\hat{N}(x)$ can be given by Equation (4.4).

$$\hat{N}(x) = N(x) + \alpha M(x) \tag{4.4}$$

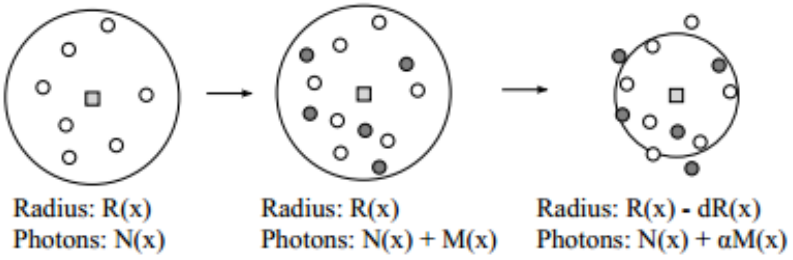


Figure 4.9: Radius Reduction after a Photon Tracing step

The final radius $\hat{R}(x)$ can be computed by combining Equations (4.2) to (4.4), and as shown by the original work on the subject, can be given, for each single hit point, by Equation (4.5).

$$\hat{R}(x) = R(x) - dR(x) = R(x)\sqrt{\frac{N(x) + \alpha M(x)}{N(x) + M(x)}} \tag{4.5}$$

4.4 Stochastic Progressive Photon Mapping

Progressive photon mapping still does not address the problem of computing the average radiance of an unknown region. While progressive photon mapping allows the computation of the estimated radiance on a given point x stored as a hit point, it does not allow the estimation of a different unknown point. This is a problem when trying to simulate distributed ray tracing effects, such as motion blur or depth-of-field.

The solution proposed in [30] presents a new formulation for the progressive radiance estimation, allowing the computation of the correct average radiance over a region.

In practice, the implementation of that formulation consists only in generate a new set of hit points after each photon pass (see Figure 4.10). The local statistics for each new hit point is taken directly from the previous hit point for that same pixel. Original progressive photon mapping generates a set of hit points, and then iteratively uses new photon maps to converge to a correct solution, based on those same hit point. This stochastic approach does not reuse the hit points, but only their local statistics. The results in the proposed work show that this solution provides better results for scenes with complex illumination, and including distributed ray tracing effects, such as motion blur, depth-of-field and glossy interactions.

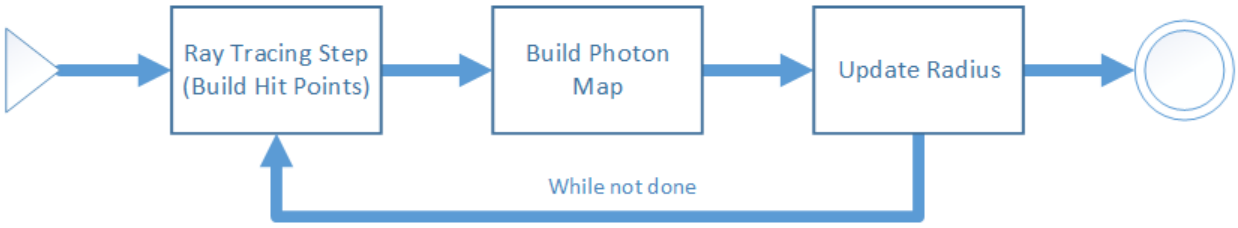


Figure 4.10: Fluxogram of SPPM

4.5 A Probabilistic Approach for Radius Estimation

Another evolution of photon mapping and progressive photon mapping comes from a probabilistic approach for the estimation of photon radius for each iteration, first presented in [31]. The proposed solution, much like original progressive photon mapping, is capable of of computing global illumination without bias, and with no theoretical limit in the amount of photons, allowing an arbitrary number of iterations to be computed.

The new formulation, called PPMPA, includes a probabilistic approach that does not require local photon statistics to be stored. It is shown in the original work that each different photon mapping step of the progressive photon mapping approach can be performed with complete independence from other steps, by using a probabilistic model to compute an estimation of the photon radius for each iteration, instead of gradually reducing it after each photon tracing step (see Figure 4.11.

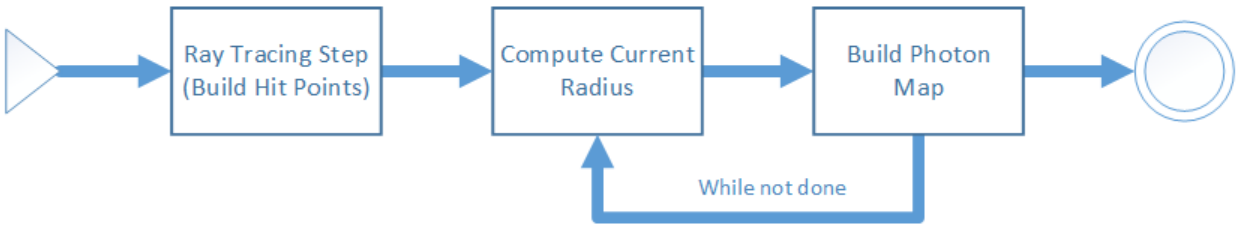


Figure 4.11: Fluxogram of PPMPA

In summary, the probabilistic analysis in the original work shows that for a photon mapping step i, the radius for for a hit point for that step, r_i, can be estimated by Equation (4.6)

$$r_i^2 = r_1^2 \left(\prod_{k=1}^{i-1} \frac{k+\alpha}{k}\right) \frac{1}{i} \tag{4.6}$$

The biggest benefit of this is that the radius computation kernel is not dependent on previous iterations, allowing for multiple photon mapping steps to be concurrently computed, as shown in Figure 4.12.

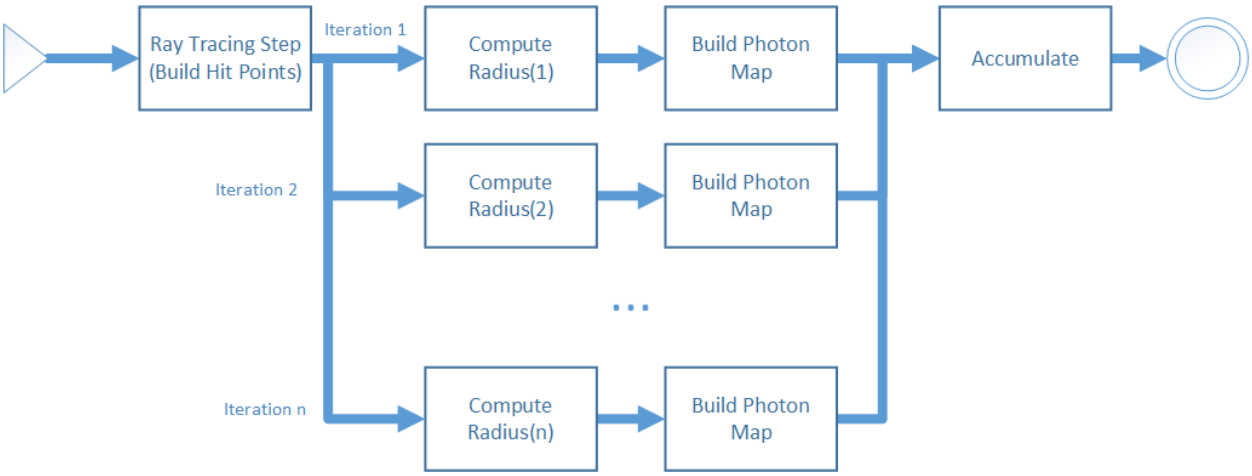


Figure 4.12: Fluxogram of PPMPA with concurrent iterations

The result is a memoryless algorithm that does not require the maintenance of intermediate statistics and allows the possibility of computing multiple iterations, or photon mapping steps, in parallel.

4.6 Summary

Several techniques were presented, from the basic Photon Mapping algorithm, following a progressive approach that enables arbitrary accuracy without being memory limited. Further improvements include a stochastic version that enables additional accuracy in the final result, and a new formulation of the radius estimation that removes dependencies between iterations, effectively allowing their concurrency.

The final result of this is the algorithm Stochastic Progressive Photon Mapping with a Probabilistic Approach, here shortly called SPPMPA, which is the case study employed during the rest of this work.

Chapter 5

Implementation Approaches of the Algorithm

The case study was built based on an already existing implementation, available at the beginning of the project. This implementation provides the initial Progressive Photon Mapping method, described in Section 4.3, the stochastic extension presented in Section 4.4, and the probabilistic approach for radius reduction in Section 4.5. Support for both CPU and CUDA rendering is also included, although CUDA support was actually not fully completed until a later version. It was implemented on top of the LuxRender project, an open source, physically based and unbiased rendering engine. The source code of LuxRender provides an ideal basis of data structures to implement a rendering algorithm such as photon mapping, and this was exploited by the author of this implementation.

The implementation used here was also based on those same data structures, and other code from LuxRender. The algorithms themselves for the ray tracing and photon mapping steps, radiance estimation, and radius reduction were based not only on the theoretical research work already presented in Chapter 4, but also on the already available implementation

This was helpful to speed up development time, by working with already existing code for the same algorithm, but also to serve as a validation tool, to assert whether the final solution, and the individual algorithms within it, produced a correct result.

The final implementation developed for this work was an adaptation and an improved version of the original one provided at the start of the project. Several approaches were made available. The first two use CPUs with OpenMP and GPUs with CUDA, respectively, and are used mostly for the sake of comparison of results and profiling. Later approaches consist of using the StarPU framework to handle task management. A native MIC implementation was also attempted, although it proved unsuccessfully due to limitations of the platform

regarding existing code.

It should be noted that only two of the presented algorithms were taken in consideration during this work, namely:

PPM (Progressive Photon Mapping)
Corresponds to the original proposition for progressive photon mapping, described in Section 4.3.

SPPMPA (Stochastic Progressive Photon Mapping with Probabilistic Approach)

Extends the initial PPM solution to include both the stochastic version and the probabilistic approach for radius reduction.

There was no attempt to implement any of the intermediate versions (SPPM or PPMPA).

There were also attempts to port the original implementation to run on the MIC platform, whose details are also described in this chapter.

5.1 Data Structures

Most of the structures used throughout all the implementations were based on the ones used in the original implementation, which in turn was heavily supported on the source code of LuxRender, namely:

Basic Geometry
The basic geometric structures, such as vector, normals, 3D points and triangles;

Meshes
A collection of vertices, edges and faces that define the shape of a 3D object in a scene;

Scene
The full description of the 3D scene to render, including all meshes, materials and light descriptions associated with it; This scene is read initially by the LuxRender library, which handles the parsing of all data files associated with the scene (mesh descriptions, materials, light sources, textures, etc.);

Bounding Volume Hierarchy
A tree structure used to index spatial objects, in this case the objects within the scene,

in order to reduce the number of required ray intersection operations; The actual implementation used, a Quad Bounding Volume Hierarchy (QBVH), is an extension of a regular BVH, optimized for a low memory footprint, and SIMD computations [32, 33]; This structure is created at startup by LuxRender, and is used to spatially index all elements in the scene, allowing for faster computations of ray intersections.

In addition to these structures that LuxRender already provided, additional data structures were also required for the implementations described later in this chapter:

Pointer-free Scene

The original scene structures available with LuxRender relied heavily on pointer based structures, which was not adequate for GPU computations. So a custom solution was required in order to store scene information in a compact manner, easily transferable and usable by a GPU

Hit Points

A data structure was required to store information about hit points position, direction vector, the pixel that it originated from, and about accumulated photon radiance; In practice this was actually split into two different data structures, the first one storing only static information about the hit point, such as position and origin point, and the last one to store incident radiance; This separation allowed a more efficient memory usage, since the two different components are read and written to at different points during the algorithm;

Lookup Table

An acceleration structure used to index the hit points in order to quickly find all relevant hit points to update after a photon trace; This corresponds to all hit points within a given radius of the hit point of each photon, or in other words, all the hit points that photon will contribute to; The structure is implemented with an hash table that spatially indexes the hit points by dividing the 3D space into a grid, where each cell references all hit points that intersect it, based on the current radius;

Ray Buffer

In the original implementation, the total number of photons traced every photon pass was not directly processed at once; Instead, a ray buffer was used to process a specific, pre-configured number of photons at a time, independently from the total number of photons to process during that step; This was mostly to increase coherence on the CPU by processing a smaller batch of consecutive rays at a time;

5.2 Computational Tasks

Throughout all implementations, computations were divided across several tasks with more or less the same duties. This was a concern from the beginning to ensure that all versions remained similar to each other, facilitating the comparison between them.

The approach used during development was mostly inspired by the original implementation. This section presents a list of all the computational tasks that were already present in the original implementation, and that were reimplemented and reworked during the development of the case study:

1. **Initialize Seeds**
   This was not part of the algorithm itself, but was necessary to initialize a seed buffer used for the random number generation process required by the following tasks. The random generation is done using a Tausworthe Generator [34]. Throughout the algorithm, it was required to keep a buffer were each seed was stored. When a random number generation with a given index was required, the respective seed was rewritten to provide the next number in the random sequence. For this process, an initialization of this buffer was necessary, using each index in the buffer as the initial seed.

2. **Generate Eye Paths**
   In this step, eye paths were initialized, based on camera position.

3. **Advance Eye Paths**
   Following the initialization of the eye paths, this task would compute their interactions with the scene, while building the set of hit points required by photon mapping. This is analogous to the Ray Tracing step in the photon mapping algorithms presented in Chapter 4.

4. **Update Radius**
   This is a small but necessary task that computes the hit point radius for the subsequent photon mapping step. When using the probabilistic approach, this is equivalent to the computation of Equation (4.6), presented in Section 4.5.

5. **Rebuild Lookup Table**
   As explained in Section 5.1, a lookup table is used to index all hit points. This task would build that structure after all hit points were generated by the **Advance Eye Paths** task.

6. **Generate Photon Paths**
   Similarly to **Generate Eye Paths**, this was used to generate a set of photon paths, leaving the light sources into the scene.

7. Advance Photon Paths

During this step, the initialized photon paths were traced within the scene, updating the local accumulated flux of each hit point they interact with along the way.

8. Accumulate Flux

To finish the photon mapping stage, this additional step was separated from the previous one, to compute the final radiance of each hit point.

9. Update Frame Buffer

This maps all hit points, and their final radiance values to the corresponding pixels on the screen, creating a temporary buffer of the image generating during the current photon mapping step.

10. Update Film

The temporary frame buffer was merged into a film structure, that represents the final image computed so far by all photon mapping steps. This directly represents the rendered 2D image, and was used to directly display the image if a live preview window was enabled, and to optionally save the current state of the render in an image file.

These tasks map almost directly into the theoretical SPPMPA algorithm. Tasks **2** and **3** represent a ray tracing step, where the hit points are computed. This is the first step in the multi pass progressive photon mapping algorithm presented in Section 4.3, and also the first step of each iteration in the stochastic approach Section 4.4. Tasks **4**, **6**, **7** and **8** are the equivalent of a photon tracing step, which corresponds to a full iteration in the original progressive approach. Figure 5.1 gives an overview of the entire algorithm (except input and output tasks), along with the dependencies between tasks that prevent them from running concurrently.

The remaining tasks are not specified by the photon mapping techniques, but are required for computational purposes. Task **5** represents an important step to make sure that access to each hit point is done efficiently when tracing the photons. By using a lookup table to index hit points, the average complexity in accessing a single hit point is lowered to $O(1)$.

5.3 Implementation

This section presents an overview of the differences and challenges between the multiple versions used.

While StarPU is the actual object of study in this project, initial development was focused on building a CPU-only, implementation, similarly to the original version. Following that,

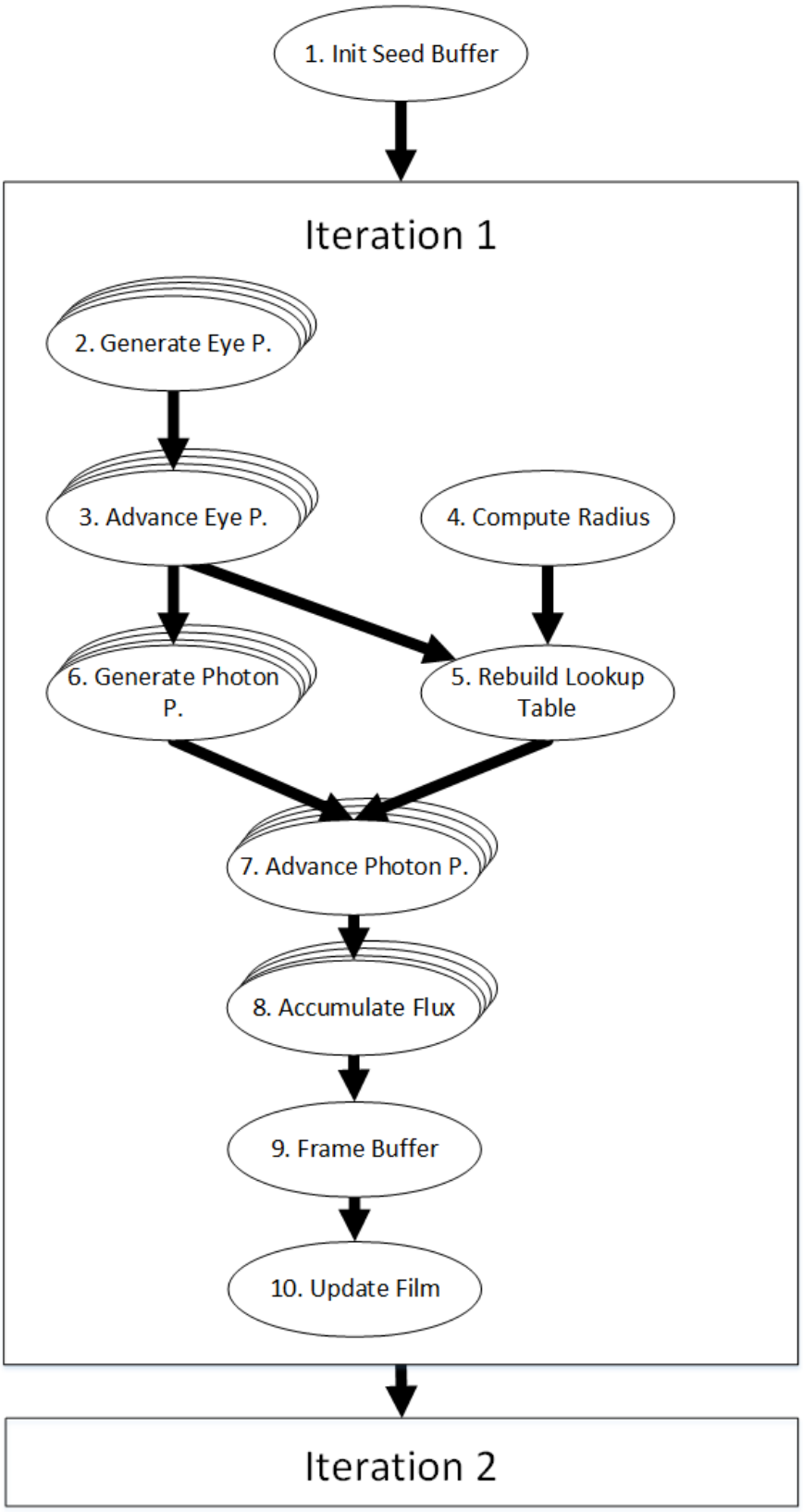


Figure 5.1: Diagram of the SPPMPA computational tasks and execution order. Tasks **2**, **3**, **6**, **7**, **8** can be parallelized. All other are sequential. The SPPMPA algorithm actually allows independent iterations, which is not represented in this diagram.

an similar approach was followed to build a CUDA version, while still avoiding to use StarPU or any HetPlat management system.

The goal of this approach was to have a functional reference for comparison that shared as much of the functionality as possible with a future implementation that takes advantage of HetPlats by using StarPU as the task manager.

All computational code was kept encapsulated to be reusable by future implementations.This helps to speed up the development time of future versions, and contributes to fairer comparisons between versions.

5.3.1 Original

The original implementation provides a few different versions of the algorithm, based on the multiple extensions on photon mapping described in Chapter 4. However, only PPM and SPPMPA are considered.

In this approach, tasks are encapsulated within a CPUWorker class and a CUDAWorker class, which implements the ray tracing and photon mapping steps using OpenMP and CUDA, respectively. While the initial PPM version, due to implicit dependency limitations (without the probabilistic approach for radius estimation, each photon mapping iteration is dependent on the previous one), can only run a single worker at a time, the later version (SPPMPA) allows the instantiation of multiple workers. In that case, each worker will share access to a centralized structure that keeps track of how many iterations have been finished in total.

In practice, the SPPMPA version provided support for running one CPUWorker and two CUDAWorkers[1]. Since the CPUWorker uses OpenMP internally to take advantage of multi-threading, this approach effectively takes advantage of the full power of a multi-core machine with at most two CUDA devices.

When this implementation was first available, however, the CUDA implementation was not yet fully finished, with only the implementation of the photon tracing steps being run on a GPU. This means that the performance when using CUDA is limited by the necessary data transfers required during each iteration. Particularly, when using SPPMPA version, where new hit points are generated after each step, following the stochastic progressive photon mapping extension, an even greater amount of communication is required, since the GPU has to wait for the new hit points before starting the computation of a new photon pass.

[1]This was not a limitation of the implementation. Actually, extending it to support more than two CUDA devices would be completely straightforward. However, there was no interesting in doing so, as all test machines used throughout the project had available at most two CUDA devices.

This versions was used only for validation of later implementations, and an in-depth study of its performance was not considered interesting, as similar versions were to be produced, but with the advantage of being further structured and optimized, and not relying on a third-party code base.

5.3.2 CPU

The first implementation to be developed was one consisting only on the implementation of each task, using OpenMP for parallelism within each individual task. Of all tasks shown in Section 5.2, it should be noted that the following tasks cannot be parallelized, and so are only implemented in sequential code:

- Update Radius (kernel 4)
- Rebuild Lookup Table (kernel 5)
- Update Frame Buffer (kernel 9)
- Update Film (kernel 10)

All other tasks (which in practice represent all the actual code for both the ray tracing and photon tracing steps) were fully parallelized, as each ray can be independently intersected. The only exception to this is with the **Advance Photon Paths** task. Since a hit point might be simultaneously hit by multiple photons, a small critical section was required for every hit point update. That critical section, however, represents a very small portion of the entire task.

In this implementation, some of the original code from LuxRender was also employed. Particularly, the intersection code for a ray, which traverses the accelerating structure indexing the scene, a Quad Bounding Volume Hierarchy (QBVH), searching for the next ray hit. This intersection code is implemented using SSE intrinsic functions, meaning that SSE code was hard coded, instead of being compiler generated. This makes sure that the accelerating structure takes advantage of the optimizations for the original BVH structures presented in [32].

Initially only the PPM version of the algorithm was developed later, whose code was lather adapted to implement SPPMPA since evolving from basic PPM to one of its extensions is relatively straightforward. The tasks themselves remain almost identical, and most of the changes are related to when and how those tasks are actually called. The PPM version was

used just for the initial stages of development and a first step towards the final SPPMPA version.

Even though the later algorithm theoretically allows multiple iterations to be run in parallel, this is not supported in the CPU-only implementation. This mimics the behaviour of the original implementation from Section 5.3.1, which was the intended result. The main goal was to have an application as much similar as possible to the original CPU version, while still sharing task implementations with the future StarPU version.

5.3.3 GPU

Following the CPU implementation, it was also desirable to produce a CUDA based implementation. Like the previous one, this served the purpose of producing an implementation similar to the original, but having task code shared with the future StarPU version, and minimize any details that would be different about the implementation.

The goal of this version is to port as much as possible of the CPU task code to CUDA. so most of the tasks described in Section 5.2 were implemented in CUDA. However, as explained before (in Sections 5.2 and 5.3.2), some of the tasks are not parallelizable, and consequently, not adequate to massively parallel devices. This means that these tasks were kept running on the CPU. It can be argued that the cost of offloading these tasks back to the CPU can be slower than executing them sequentially on the GPU, since the required data transfers can be the dominating factor. However, this was also the decision made in the original approach of the algorithm, which this dissertation attempts to approximate as much as possible. Due to that, implementing these tasks sequentially on GPU was considered, but left for possible future work as it was not a priority. The obvious consequence of this is that a full iteration of the algorithm is not capable of running entirely on a GPU, requiring memory transfers in between to solve data dependencies.

The most problematic drawback of this decision is about the **Rebuild Lookup Table** task. Running this task on CPU will likely have a very noticeable impact on performance, since it requires to transfer the generated hit points from the GPU to the CPU, and later copy the generated hash table back to the GPU.

5.3.4 MIC

The initial message transmitted by Intel regarding the new MIC Architecture explains that the device is intended to provide performance to applications much like any other

accelerator, without the need to learn a new programming model. It is also explained that existing code bases should have little problem compiling and running natively on it, providing additional performance for already existing CPU code. If proved right, this would be a huge step forward in coprocessor technologies, as the usage of different programming models (such as CUDA) is currently one of the blocking factors of their usage, due to learning difficulties, and incompatibilities with existing code.

Thus, some efforts were put into attempting to port the original implementation to compile and run natively on a MIC device. Other execution modes (offload or message passing) would require a rewrite of the program, and as such, would not provide the ease of usage claimed by Intel.

However, this was later abandoned as the code for the original implementation (as well as the later CPU implementation produced for this work) proved to actually be incompatible with the device. This is mostly due to the implementation of the QBVH accelerating structure (discussed in Section 5.1). This structure is coded using compiler SSE intrinsics[2] for the intersection functions of rays with the scene. These intrinsics render the intersection code completely incompatible, requiring a complete rewrite to remove coupling with SSE functions, and use other vectorization methods. Such would require a larger refactoring effort to port the implementation, which was not towards the original goals of this dissertation. As a result, this implementation was abandoned. Other factors, such as difficulties regarding compatibility with external libraries, which usually also have to be compiled natively for the MIC were also encountered, which would have increased the difficulty if a port was to be done.

5.3.5 StarPU

For the final implementation, using the StarPU framework, most of the remaining work consisted on reusing the existing code for each computational tasks, and submit them as tasks to be scheduled by StarPU. The development process of the previous implementations (CPU and CUDA) resulted in a very modular solution, with very little coupling[3] between tasks implementation and the algorithm and scheduling code being used.

[2] intrinsics: functions available in a programming language that are actually implemented by the compiler. More specifically, SSE intrinsics expose the SSE instruction set directly in the language

[3] The degree of dependency between the multiple modules of a system. Tight coupling tends to difficult refactoring of one module without requiring subsequent changes to dependant modules, difficulting code

Early Decisions

An early decision for this implementation was to consider only the low-level API, and not the `pragma`-based one (see Section 3.2.6). The reason for this is due to the high-level version being an early product, still being in earlier development stages, and not being fully capable of providing the full set of features of StarPU.

Additionally, the actual documentation for the framework is almost entirely focused on the low-level functions, making it easier getting up to speed and understand its usage.

An additional decision that was enforced by the existing implementations is related to the used task scheduling policy. Since the CPU implementation of all parallelizable tasks was based on OpenMP, it was intuitive to approach the problem by using the capabilities of parallel tasks and combined workers of StarPU (see Section 3.2.5). The only drawback that comes from this is that only the parallel-aware task schedulers (`pheft` and `peager`) are capable of parallelizing CPU tasks. This means that when using a non-parallel-aware scheduler, CPU tasks such as **Advance Photon Paths** will increase in cost.

Data Management

The first step for this implementation was to refactor data management, letting StarPU handle all necessary data for the algorithm. One exception was made to this, regarding the input information for the 3D scene. This information is stored in a somewhat complex structure, as opposed to all other dynamic data used throughout the photon mapping algorithm, which consists only on vectors whose size can be static and predetermined.

A small change was necessary in the lookup table build process. This task previously generated a dynamically sized structure, since it is dependent on the number of photons that intersect the scene within the current radius of each hit point, which cannot be predetermined.

One solution for this would be to only register the lookup table in a StarPU data handle only after its generation is complete, and the size can be determined. This is not a desirable solution, as it would introduce a barrier on that point of the iteration, and forcing all future tasks to be submitted only once the lookup table build process is complete. This would prevent StarPU from having knowledge on future tasks, and preventing it from asynchronously prepare data buffers to solve dependencies, increasing the latency caused by the imposed barrier.

This seems to be a rather harsh limitation of StarPU, as irregular sized structures are

commonly used. However, an alternative solution is possible for this specific problem. The cell size used for the hash grid is based on the photon radius for the current iteration, in such a way that a cell will never have a width, height or depth greater than the current radius. With that any given hit point will always intersect at most 8 cells. Thus, it can be determined that the maximum hash table size can be set as $8 * \#hit\_points$, for any iteration.

Following this constraint, the hash table structure was refactored to be a fixed-size one, allowing StarPU to handle it without the need for barrier, and allowing future tasks to be submitted at will, using the lookup table data handle as a regular data buffer and dependency.

Task Submission

The other main change required is to wrap tasks around StarPU API calls. All data for each iteration is assigned to an individual data handle. No allocations are ever done manually during the main loop, making all memory managed by StarPU. While in the CPU and CUDA version, only a single iteration is considered at a time, so task invocation is made synchronously, here all tasks are submitted asynchronously, and dependencies are implicitly given by the data handles required by each task. These dependencies are represented in Figure 5.2.

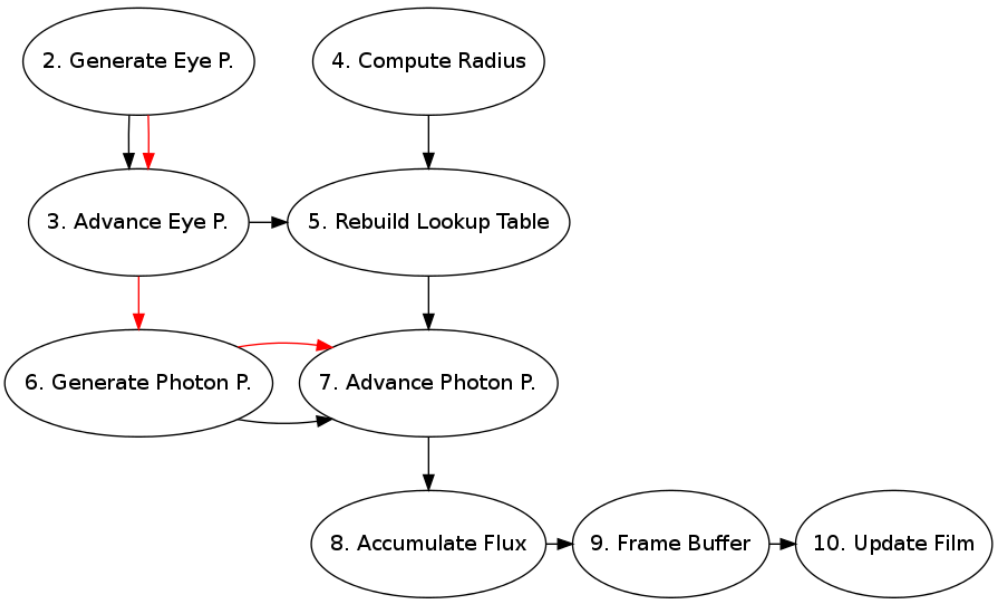


Figure 5.2: Dependency graph of an iteration of SPPMPA. Red arrows represent dependencies related to the seed buffer, which are not imposed on the algorithm itself, but only a limitation of the implementation

It can be seen from the graph that some task concurrency is possible, although limited. It is expected that the most time consuming tasks are **Advance Eye Paths** and **Advance**

Photon Paths, as they compute scene intersections. **Rebuild Lookup Table** can also be considered, especially when other tasks of the same iteration are run on a GPU. This will require synchronization memory transactions to solve the dependencies of this task. Prefetching cannot be employed here by StarPU, as these dependencies are not available until the immediately previous tasks are computed. This could provide a large bottleneck for the performance of each individual iteration.

One of the limiting factors of these dependencies is related to the number generation method employed, which relies on a read-write seed buffer, with each new value requested updating the corresponding seed. As can be seen in the graph, only one relevant dependency is actually created by this, between **Advance Eye Paths** and **Generate Photon Paths**. These tasks should be considered independent, but share a dependency on the seed buffer, and thus cannot execute concurrently as it would be expected. Other dependencies are created from this limitation (shown in red in the graph), but they do not represent a problem, since their elimination would not introduce any new concurrency possibilities.

A solution for this would be to change the random number generation method, to one that would not require intermediate seed memory to be passed between each task. However, that was not attempted, as it would require additional effort in changing the algorithm, as well as the previous implementations (if coherence between them was to be kept), and choosing an adequate and efficient new method. The final solution actually came as a consequence of enabling concurrency between iterations, explained in the next section.

Enabling Concurrent Iterations

Since the high amount of dependencies between tasks does not allow a good degree of concurrency within a single iteration, efforts were focused in allowing the execution of multiple iterations concurrently. This is one of the main reasons SPPMPA was the only one focused on, as other versions without the probabilistic radius estimation would introduce dependencies between each iteration, preventing their parallelization.

The initial approach to port the implementation to StarPU relied on data handles being declared at the start of the rendering process, and released at the end.

As shown in Listing 1, data handles are created and kept during the whole rendering process. In practice, this means that each iteration will depend on the same data buffers as the previous one, even though they are to be completely re-written, and their previous values ignored. This is not desirable, as it prevents concurrency between iterations, just like it happens due to the seed buffer dependency previously explained.

```
void render() {
  starpu_data_register(seeds, ...)
  starpu_data_register(eye_paths, ...)
  starpu_data_register(hit_points, ...)

  starpu_insert_task(codelets::init_seeds, seeds);

  int iteration = 0;
  while(iteration < config.max_iters) {
    starpu_insert_task(codelets::generate_eye_paths, eye_paths, seeds);
    starpu_insert_task(codelets::advance_eye_paths,  eye_paths, seeds, hit_points);

    ... // all other tasks for this iteration

    iteration++;
  }

  starpu_data_unregister(seeds);
  starpu_data_unregister(eye_paths);
  starpu_data_unregister(hit_points);
}
```

Listing 1: The beginning of the main rendering loop, with global data handles

An alternative solution is to declare the handles in-loop, as shown in Listing 2.

With this method, each iteration declares its own copy of the required data. StarPU provides the `starpu_data_unregister_submit` API call, which instructs the library that the given data buffer can be discarded as soon as existing tasks depending on it are finished. Since data is local to each iteration, this can actually be considered a more intuitive way to approach the problem.

However, an additional problem arose from this approach. Due to the asynchronous nature of the tasks, the program actually submits every single iteration to the scheduler, meaning that multiple copies of the `eye_paths` and `hit_points` buffers will be immediately requested. For a large enough number of iterations, this resulted in memory problems, and eventually program failures. Since StarPU does not provide any method to control this, the limitation had to be imposed manually, by inserting a barrier every few iterations (with the actual number being a configurable value).

```
void render() {
  starpu_data_register(seeds, ...)

  starpu_insert_task(codelets::init_seeds, seeds);

  int iteration = 0;
  while(iteration < config.max_iters) {
    starpu_data_register(eye_paths, ...)
    starpu_insert_task(codelets::generate_eye_paths, eye_paths, seeds);
    starpu_data_register(hit_points, ...)
    starpu_insert_task(codelets::advance_eye_paths,  eye_paths, seeds, hit_points);
    starpu_data_unregister_submit(eye_paths);

    ... // all other tasks for this iteration

    starpu_data_unregister_submit(hit_points);

    iteration++;
  }

  starpu_data_unregister(seeds);
}
```

Listing 2: The beginning of the main rendering loop, now with in-loop data handles

Chapter 6

Profiling Results

Initial analysis was focused on studying the scalability of both CPU and CUDA versions without the help of any framework. This provided a reference to evaluate the overhead of using the framework. When using the StarPU framework, different schedulers were tested, particularly **peager** and **pheft** due to their awareness of combined workers, which allows OpenMP parallelization within CPU tasks. **dm** and **dmda** were also tested, to analyse the impact of memory transfers in the performance model of a scheduler.

6.1 Testing Environment

All tests were performed within the SeARCH[1] cluster, particularly using the most recent generation of hardware, in the node 711 (fully described in Table 6.1).

All tests were compiled with GCC 4.6.2 (latest major version with full CUDA support), the boost library 1.49.0, and version 5.0 of the official CUDA compiler. The latest available version of StarPU was used, 1.1.0rc2, as well as the hwloc 1.7 library for hardware topology, which StarPU internally uses.

6.2 Testing Methodology

Measurements were done only in the algorithmic section of the program, disregarding any input and output operations as well as initial setup of StarPU and other libraries. All implemented approaches (CPU, GPU and the multiple StarPU options) were analysed, with

[1] http://search.di.uminho.pt

CPU device:	**Intel Xeon E5-2670**
# CPUs:	2
# Cores p/CPU:	8
# Threads p/Core:	2
Clock frequency:	2.66 GHz
L1 cache:	32 KB + 32 KB
L2 cache:	256 KB
L3 cache(shared):	20 MB
RAM:	64 GB
CUDA Device 0:	**Kepler K20m**
# SMX:	13
# CUDA-cores p/ SMX	192 (2496 total)
Clock frequency:	706 MHz
L1 cache (p/SMX):	64 KB
L2 cache shared:	1.25 MB
Global memory:	5GB
CUDA Device 1:	**Tesla M2090**
# SM:	16
# CUDA-cores p/ SMX	32 (512 total)
Clock frequency:	1301 MHz
L1 cache (p/SM):	64 KB
L2 cache (shared):	0.75 MB
Global memory:	5GB

Table 6.1: Hardware description of the SeARCH computational node 711

the first two serving mostly for comparison with the framework.

In the main rendering loop, the time for each full iteration of the SPPMPA algorithm was measured (see Figure 5.1, which explains how iterations are organized). This was done by obtaining the time-stamp at the beginning of each iteration, and at the end of the last task of the same iteration, which also works well when using concurrent iterations. The full iteration time includes the required data transfer times, but more granular values were obtained, by measuring the time for each invoked task of each iteration, which does not account for data transfers. Presented results show only the average time of each iteration for the execution (with at least 100 iterations each). Full calibration was done prior to any StarPU test (except the calibration test shown in Section 6.4.3). Since support for concurrent iterations was not fully implemented until later in the development stages, most results focus only on the approach of using only a single iteration at a time.

Since the entire main loop consists solely on task invocations, it was assumed that the difference between total iteration time and the sum of each task within the iteration corresponds to idle periods when the framework is waiting for resources or dependencies, or performing data transfers.

For each measurement, only the time spent in the main rendering function was considered, discarding any input and output time spent by the program. A minimum of 10 executions were made for each measurement, for which the 3 best executions within at most 5% of each other were considered. When comparing results, the average time for each iteration of the main loop of SPPMPA was the base value to use (with each test running at least 20 iterations).

Whenever CUDA was employed, the CUDA Occupancy Calculator[2], as well as manual tuning, were used to find the correct block size used for each computational kernel.

6.2.1 Input Scene

For simplicity, input reading was left to the LuxRender library, relying on the existing structures and parser to read all the data to render a scene. This limits all testing to the available scenes shipped with LuxRender as samples. From these scenes, only three were selected, namely, **kitchen**, **cornell** and **luxball**, shown in Figure 6.1.

(a) cornell **(b)** kitchen **(c)** luxball

Figure 6.1: Input scenes

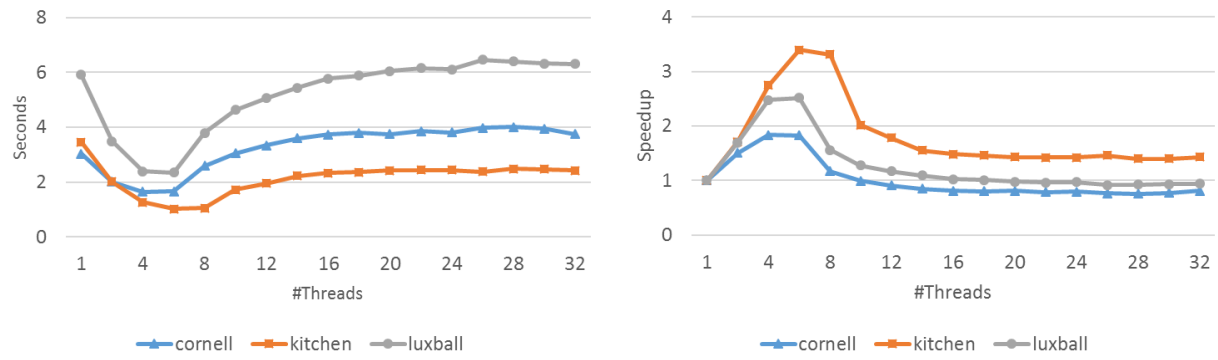
6.3 Performance Results Without a Framework

6.3.1 CPU

For the CPU implementation, earlier analysis of the original implementation showed relatively poor scalability. The measured results are not presented here due to different algorithms being employed (since only the PPM version was available at the time for that implementation), and no assumptions could be made about code quality, as explained in

[2]A spreadsheet by NVidia that helps estimating the ideal block size for a given kernel

Section 5.3.1. Instead, a scalability analysis was made on the actually implemented CPU version, shown in Figure 6.2.



(a) Avg iteration time

(b) Speedup *vs* sequential version

Figure 6.2: CPU implementation measurements

It can be seen that the code scales fairly up to 6 threads, which is close to the point at which a single CPU socket is filled, at which point performance starts to degrade. For **luxball** and **cornell** scenes, performance with 16 threads (same amount as physical CPU cores in the test machine) actually barely differs from the sequential approach.

Scalability is only fair when a single socket is used. The algorithm is clearly memory bounded, especially considering the fact that the test machine uses a NUMA architecture. This pins all memory allocated by the master thread (such as the input scene, which is heavily used throughout the algorithm) to one of the sockets, leaving the other with slower access times to such memory.

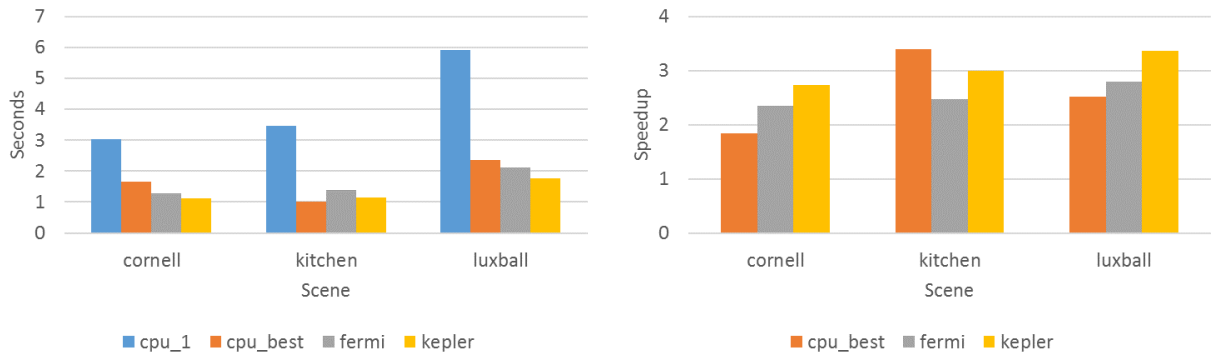
Memory affinity tools such as `hwloc` could prove useful here, for example, by creating multiple copies of the input scene, and pinning each one to each NUMA node. Each socket would then benefit from faster accesses to its own memory bank.

6.3.2 GPU

When analysing the GPU implementation, tests were made on both available GPUs, and compared against the base sequential CPU implementation. The best execution time on CPU was also included for comparison.

It should be taken into account that, as explained in Section 5.3.3, this is an almost GPU-only implementation, and still requires a certain amount of in-loop memory transfers, and some CPU computation to generate the lookup table. Since this was not implemented on GPU, it should add an overhead to the execution. Still, Figure 6.3 shows the implementation

is able to outperform the CPU in two of the three cases, achieving a speedup of around 3 when compared to the sequential approach.



(a) Avg iteration time **(b)** Speedup vs sequential version

Figure 6.3: GPU implementation measurements

6.4 Performance Results With StarPU

In order to accurately profile the performance obtained by using the framework, measurements included not only the relevant tests to compare overall execution times with previous approaches, but also tests to analyse the impact of using the framework, and its behaviour in different conditions.

6.4.1 Scheduler Impact

Figure 6.4 shows the overhead of using the framework to schedule tasks instead of directly invoking them. Execution times were measured with different schedulers, with only CPU tasks, and compared against the best CPU times obtained without the framework. No major speedups are expected here, since the framework should itself create a significantly high overhead, but it should be interesting to see if delegating the task of choosing OpenMP thread pool size to StarPU, rather than manually tuning it, directly impacts performance.

This was mostly a concern due to the scalability problems observed in Section 6.3. If StarPU, using one of the less smart schedulers, would opt to eagerly use all available CPUs, performance would degrade as seen in Figure 6.2. This was indeed the observed result with the **peager** scheduler. **dm** and **dmda** also degrade performance down to around 4x (which roughly corresponds to worst-case scenarios observed on CPU implementation), but that is to be expected since these schedulers do not support parallel tasks. Since no accelerators are being used here, this results in all tasks being ran sequentially.

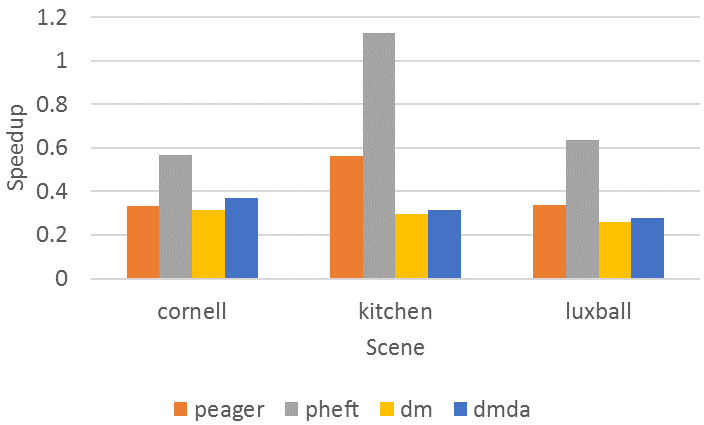


Figure 6.4: StarPU implementation, CPU-only

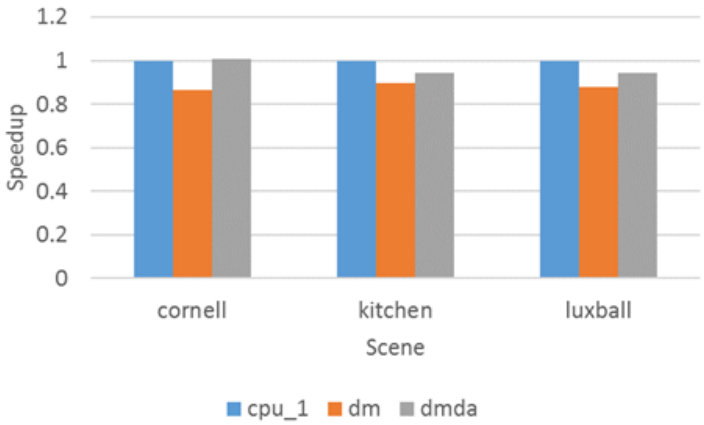


Figure 6.5: StarPU implementation, CPU sequential vs sequential schedulers

Additionally, Figure 6.5 shows a comparison of the three sequential versions, namely the framework-less CPU version with 1 thread, and the StarPU version with the **dm** and **dmda** schedulers using only CPU devices. **dm** performs slightly worse than **dmda** in all cases. Since this is a CPU-only test, it is executed using only the main system memory, so no memory transfers are actually required. The small performance gain from **dmda** can thus be attributed to the ability of asynchronously allocating requested memory needed in future tasks.

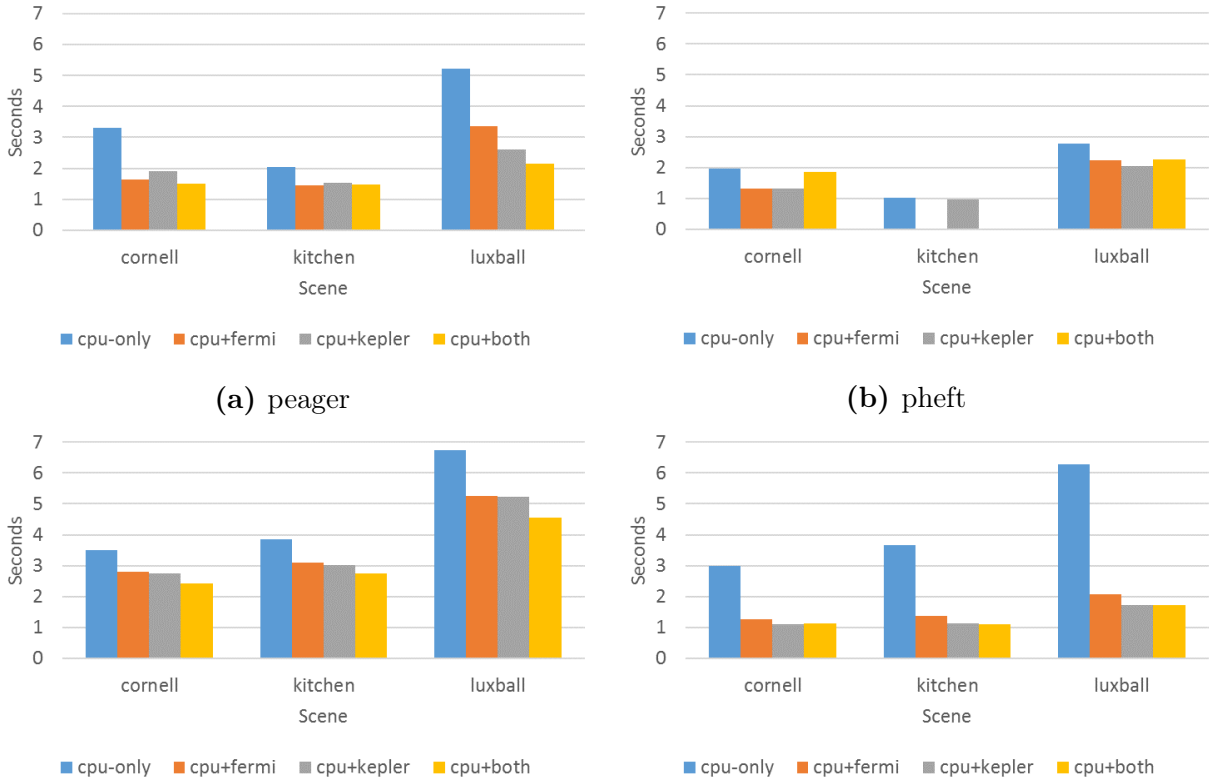
6.4.2 Performance with Accelerators

To test how GPU devices influenced the algorithm, measurements were made for both individual GPU, and with both available GPUs for StarPU to use. When using accelerators, all schedulers are able to speedup the implementation when compared to the best CPU times (which were achieved with around 6 threads). However, as seen in Figure 6.6, the gain difference between each scheduler is noticeable: **dmda**, which performs poorly on CPU due to sequentializing tasks sees the largest improvement. This is expected since it is essentially a comparison between sequential CPU tasks with massively parallel GPU tasks. The fact to note here is that **dm** has much worse evolution under the same conditions. It should be noted here that only a single iteration is being processed at a time. This limits the amount of parallelism available, which results in small performance gains when going from one to two GPUs. The only gain that can be extracted from that is by concurrently executing multiple tasks of the same iteration, one on each device. As seen in Figure 5.2, the number of dependencies is a limiting factor for this.

This enforces the fact that memory transfers are extremely important to take into account by the scheduler, as this is the only difference between the two.

As for **peager** and **pheft**, their gains are not as large, since the CPU code was already parallelized, but it is relevant to note that the smarter **pheft** seems to be outperformed once the whole set of devices is used. This is a consequence of the low iteration level-parallelism available. Since each task is fully scheduled to a given device, and multiple task dependencies exist within an iteration, it is difficult to efficiently take advantage of multiple devices simultaneously, in which case eagerly selecting the best device can be considered a faster and more efficient choice.

Unfortunately, when using **pheft** with the Fermi device, memory errors would constantly be raised, so it was not possible to finish those tests successfully. This is most likely due to problems with this particular scheduler, which is still under development by the StarPU team, and thus cannot be assumed to be fully functional.



(a) peager **(b)** pheft

(c) dm **(d)** dmda

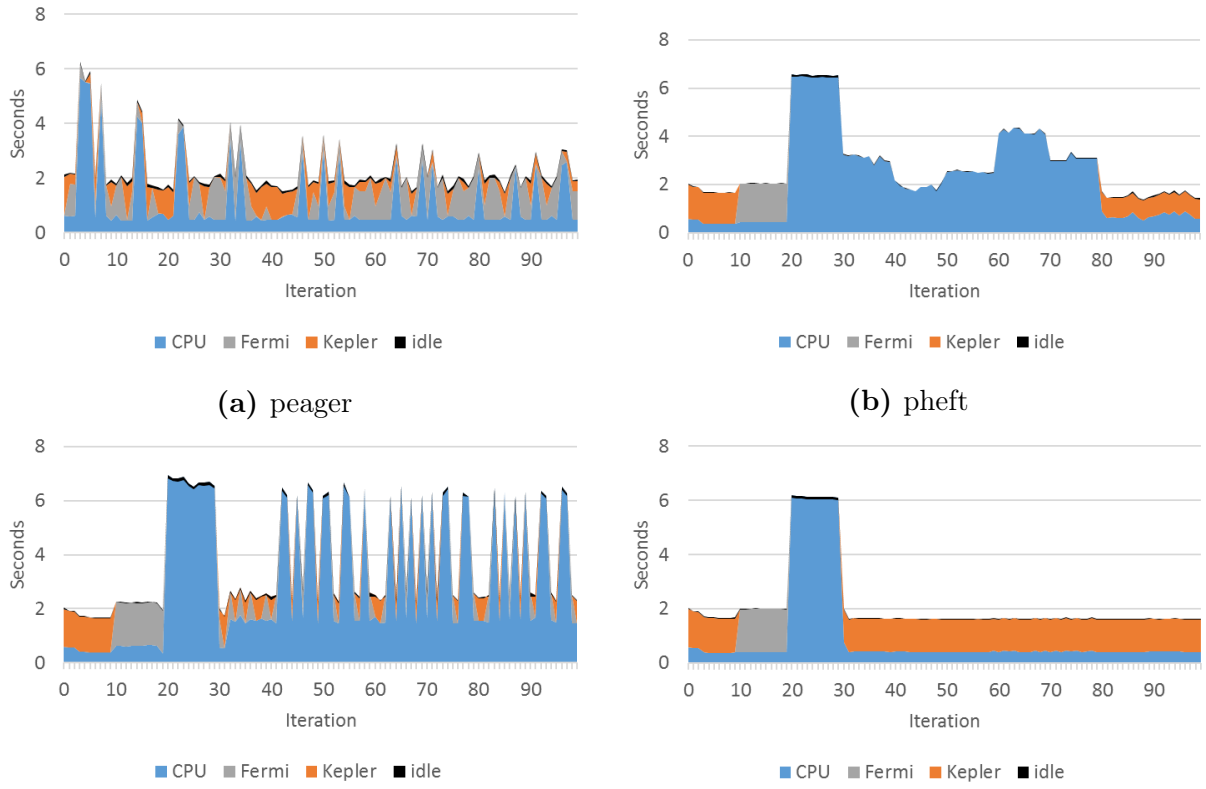
Figure 6.6: Avg. iteration time of the different schedulers with GPU devices

6.4.3 Calibration

StarPU builds a performance model for each task by calibrating them on the first executions. Once 10 executions are profiled, calibration is finished for that task and the performance model can start to be used. Figure 6.7 shows how calibration affects each scheduler. Each case consists of running 100 iterations of the algorithm with no performance model to start with (deleted prior to the execution). This performance model is then build in the initial stages of the algorithm, until a good amount of sampling is obtained (at least 10 executions for each different task on each different device). Results show very different behaviours between schedulers.

peager, being a more naive algorithm, does not actually perform calibration, and does not seem to evolve very efficiently. A drop in execution time can be seen across time, but since all devices are eagerly used whenever possible, this results in more workload being assigned to the CPUs, which have worse performance when compared to gpus, as explained in previous sections. It should also be noted that **peager** was found to always employ 14 OpenMP threads for each task, assigning it all available cores. This is not the best choice, as it was noted in Section 6.3 that tasks do not scale beyond 6 threads. The apparent drop in CPU time that can be observed in Figure 6.7a is not due to better thread pool size selection,

but rather the result of more tasks being offloaded to GPUs.



(a) peager **(b)** pheft

(c) dm **(d)** dmda

Figure 6.7: Calibration process starting with an empty performance model

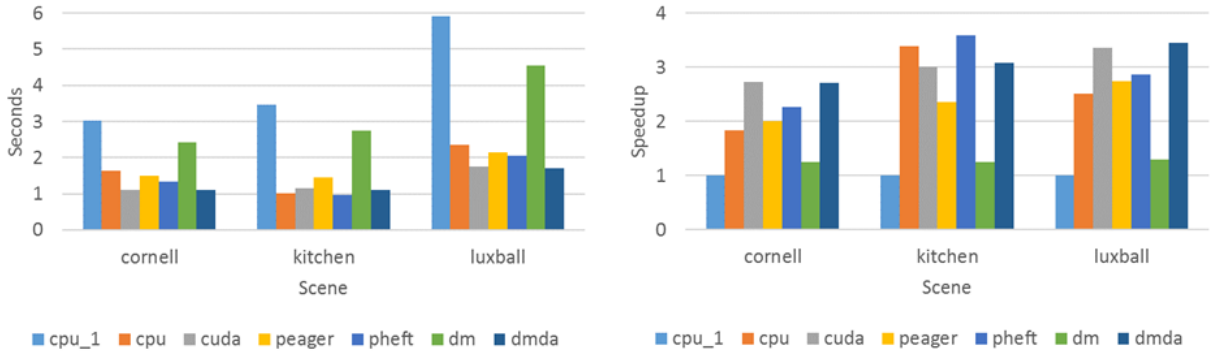
For the remaining schedulers, calibration can be clearly noticed in the first 30 iterations of each one. Each calibration starts by running 10 iterations almost entirely on the Kepler device (an entire iteration would not be possible, since some tasks are CPU-only), following the Fermi device, and finally the CPU devices.

After the initial calibration, **pheft** continues to calibrate the CPU implementation, by successfully choosing different thread pool sizes, until the best one is found. After this process, the initial calibration is finished, and the scheduler progresses normally.

Dm and **dmda** do not support combined workers, so CPU calibration is restricted to using only 1 CPU thread at a time. Since **dmda** is a data-aware version of the original **dm**, it deals more efficiently with data transfers. This being the only difference, **dmda** seems to perform much better, offloading most computations to the Kepler device, while **dm** still runs some iterations on the CPU. This is a result of **dm** not being able to asynchronously perform required data transfers, thus GPU task execution costs (including the required communication) are much higher, ending up with a significant part of it being kept on the CPU.

6.4.4 Overall Performance Comparison

The best case scenario for each approach is shown in Figure 6.8. This serves to show the impact of StarPU with each different scheduler against framework-less solutions.



(a) Avg. iteration time **(b)** Speedup

Figure 6.8: Best cases for each different implementation and scheduler

	Sequential	CPU	GPU	**peager**	**pheft**	**dm**	**dmda**
cornell	3.03	1.65	1.11	1.51	1.33	2.42	1.12
kitchen	3.46	1.02	1.15	1.46	0.96	2.75	1.12
luxball	5.91	2.35	1.76	2.15	2.06	4.55	1.71

Table 6.2: Avg Iteration time for all versions

	CPU	GPU	**peager**	**pheft**	**dm**	**dmda**
cornell	1.83	2.73	2.01	2.27	1.25	2.71
kitchen	3.39	2.30	2.36	3.60	1.26	3.09
luxball	2.52	3.36	2.75	2.87	1.30	3.45

Table 6.3: Avg speedup for all versions

6.4.5 Concurrent Iterations

With the employed approach, task-level parallelism is limited. The **dmda** does not support combined workers, greatly lowering efficiency of CPU tasks. **pheft** does support this, but its not a data aware scheduler, meaning that data transfers are not considered when assigning tasks. As a result, performance with StarPU is limited when using processing a single iteration. The attempted solution was to allow the execution of a variable number of concurrent iterations, in order to take advantage of multiple devices without the limitation of dependencies. Results shown in

Results show in Figure 6.9 indicate this was not a successful approach, as the best speedup achieved was slightly above 2, when using between 16 to 32 concurrent iterations. Further

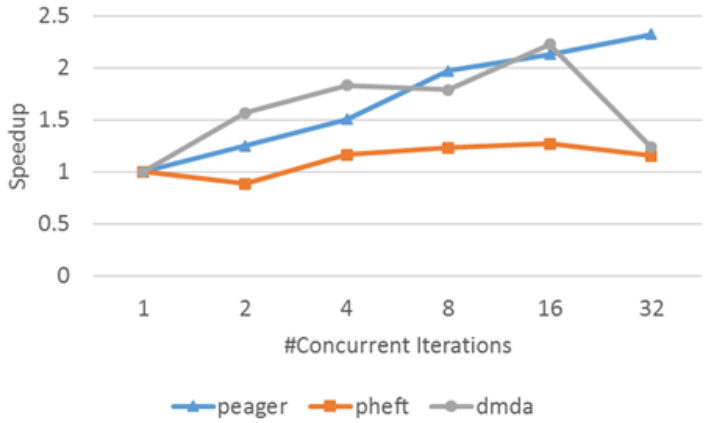


Figure 6.9: Speedup with concurrent iterations

analysis of the results indicated that when using this approach, large contention points appeared at the end of each iteration, which consists of 2 computational tasks that have to be ran sequentially on CPU (see Figure 5.1). These tasks essentially perform memory operations, to aggregate the result of the individual iteration into the final aggregated image. This works as a barrier, since two iterations cannot be merged concurrently to the final image, slowing all iterations in that stage.

One possible alternative that could be attempted to seek better results was an approach based on data partitioning rather than concurrent iterations. Since multiple iterations seem to create excessive memory contention, partitioning data using the StarPU API, and manually defining task granularity might be an alternative way to extract more parallelism from a single iteration, since each sub task can be assigned to different devices.

Chapter 7

Conclusions

This dissertation presents an analysis of two emerging frameworks that aim to ease the process of implementing or porting parallel applications to heterogeneous platforms. GAMA was tested with a small case study, using a finite volume method based on an already existing application to compute the spread of a material in a surface. A more in-depth analysis was made of the StarPU framework, using a more robust algorithm as a case-study. The selected algorithm was the progressive photon mapping, a ray tracing technique, along with two extensions proposed for it: a stochastic approach that better estimates radiance accross unknown regions, and a probabilistic approach for radius estimation that makes it possible to run independent and concurrent photon mapping iterations. Framework-less versions were implemented (particularly in CPU and GPU devices) for comparison purposes, and StarPU approaches were tested against various of the available schedulers.

GAMA is a relatively new product, still under development at University of Minho and University of Texas at Austin, that presents some promising features not present in competitor frameworks. StarPU, being a more developed product, provides a more solid API, backed by a relatively large user-base, but still with some problems to be solved.

Design issues in the StarPU API allow simple developer errors to cause unexpected and hard-to-debug behaviour. This is a problem not directly related to the performance of the framework, but to its usage. With a user base composed not only of experienced developers, but also of scientists with less low-level knowledge about parallel programming practices or heterogeneous platforms, error-prone products may be a blocking factor. Additionally, using StarPU still requires some amount of knowledge regarding parallel computing. Questions such as "Is it worth the effort to implement a given task using an accelerator?" or "What scheduling policy best suits a given algorithm?" must be answered during implementation. From that comes that developers or scientists without a great understanding of such issues

won't be able to take the best of the framework. The high-level StarPU API can be useful to solve these difficulties, but it does not seem to provide a large enough subset of the full range of StarPU features (although no actual hands-on test was done to assert this).

GAMA's API is still not as solid, but it seems to present a more consistent solution, in terms of API and programming model. This can be partially related to the use of $C++$ instead of plain C.

For a practical analysis, GAMA was tested with a small test case, to gain some knowledge about the framework and understand its usage. For the implementation of a more robust case study, the StarPU framework was the choice. The presented progressive photon mapping algorithm was implemented, using two extensions to it, namely the stochastic progressive photon mapping, and the probabilistic approach for radius estimation. This provided an algorithm with several possibilities to explore parallelism.

Several versions were produced, starting with two simple, framework-less targeting CPUs and GPUs, respectively. An already available implementation was used as the basis for validating the correctness of the algorithm in those cases. Later, the same code of the two previous versions was re-used in a new implementation using the StarPU framework.

Profiling results of these versions showed confirmed the assertion made by the StarPU team about data transfers being a key factor for the scheduler to make a decision. The **dmda** showed the best results in most cases, except when using no accelerators, due to the restriction of not supporting parallel workers.

Results also indicate that performance with StarPU is greatly dependent on the selected scheduling policy, of which there are several available. While GAMA currently only provides a single, HEFT-like policy, profiling shows that this is not a one-fits-all solution, since less smart policies such as eager loading can provide better results under certain conditions. Thus, the fact that StarPU allows pluggable schedulers to be selected, and even programmed, can come as both a blessing and a drawback, depending on the degree of control one requires of the performance of an application. The requirement of manually choosing the best scheduler can sometimes be a tedious task, but it can also lead to better performance results. This trade-off between control and simplicity is not uncommon, and must be considered carefully when developing a product such as StarPU or GAMA.

Another factor that greatly differs between the two frameworks is the ability to control task granularity. StarPU requires the developer to manually divide tasks and data, and submit each one individually. GAMA follows a more robust approach, attempting to automatically adjust task granularity to each device. This still requires some intervention from the developer, as the function required by GAMA to divide the data must be manually

defined. But it prevents the developer from having to manually calibrate task granularity, which can be a difficult task, given the heterogeneity of the system.

Even though GAMA shows promising features, some extensions could be considered when comparing it to StarPU. The most promiment factors are the **dmda** scheduler, the modularity of the framework, that can allow different components such as the scheduler to be replaced or changed, and the more consistent API. The definition of a unified programming and execution model can make it easier for developers to not have to worry about different architectural details (unless desired), but it can also present a new barrier for new developers, who will have to deal with yet another programming model in order to use the framework. Additionally, this model is one of the factors that makes it difficult for GAMA to maintain compatibility with existing libraries and applications, making it a product targeted only at newly written parallel programs.

7.1 Future Work

While this dissertation focused mostly on the implementation of a case study in StarPU, a similar effort should be made to produce a similar implementation with GAMA, to actually compare the two in terms of performance. Without such implementation, only a more shallow comparison could be made, regarding mostly the features, usability, and a few problems with each solution. It would also be interesting to test the usability of the `pragma`-based API of StarPU

Other possible points of improvement on top of this work are more related to the produced implementation. The first point is related to the random number generation, which could be further improved by using a different random number generator, that would not require an intermediate buffer, thus eliminating one dependency between tasks.

In addition, a new approach to task parallelization could be attempted, which did not depend on OpenMP, and as such would allow an efficient usage of other StarPU schedulers without support for combined workers, which are still under development by the framework's team.

Finally, due to the observed results when attempting to use concurrent iterations to exploit more parallelism, a different approach might prove more viable, by using data partitions to split the domain, and submit multiple child tasks instead of a larger one, with granularity having to be manually controlled and tuned. This method might be a more efficient solution to extract parallelism from the application.

Bibliography

[1] G.E. Moore. "Cramming More Components Onto Integrated Circuits". In: *Proceedings of the IEEE* 86.1 (1998), pp. 82–85. ISSN: 0018-9219. DOI: 10.1109/JPROC.1998.658762. URL: http://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf.

[2] Gordon E Moore. "Progress in digital integrated electronics". In: *Electron Devices Meeting, 1975 International*. Vol. 21. IEEE. 1975, pp. 11–13.

[3] V.W. Lee et al. "Debunking the 100X GPU vs. CPU myth: an evaluation of throughput computing on CPU and GPU". In: *ACM SIGARCH Computer Architecture News*. Vol. 38. 3. ACM. 2010, pp. 451–460.

[4] Rajesh Bordawekar, Uday Bondhugula, and Ravi Rao. "Believe it or not!: mult-core CPUs can match GPU performance for a FLOP-intensive application!" In: *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. ACM. 2010, pp. 537–538.

[5] Michael D Linderman et al. "Merge: a programming model for heterogeneous multi-core systems". In: *ACM SIGOPS operating systems review*. Vol. 42. 2. ACM. 2008, pp. 287–296.

[6] Chi-Keung Luk, Sunpyo Hong, and Hyesoon Kim. "Qilin: exploiting parallelism on heterogeneous multiprocessors with adaptive mapping". In: *Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on*. IEEE. 2009, pp. 45–55.

[7] Cédric Augonnet et al. "StarPU: a unified platform for task scheduling on heterogeneous multicore architectures". In: *Concurrency and Computation: Practice and Experience* 23.2 (2011), pp. 187–198.

[8] João Barbosa. *GAMA framework: Hardware Aware Scheduling in Heterogeneous Environments*. Tech. rep. Computer Science Dept., University of Texas at Austin, Sept. 2012.

[9] Intel. *Intel Xeon Phi Coprocessor Developer's Quick Start Guide*. URL: http://software.intel.com/en-us/articles/intel-xeon-phi-coprocessor-developers-quick-start-guide.

[10] NVidia Steve Scott. *No Free Lunch For Intel MIC (Or GPU's)*. URL: http://blogs.nvidia.com/blog/2012/04/03/no-free-lunch-for-intel-mic-or-gpus/.

[11] S. Williams, A. Waterman, and D. Patterson. "Roofline: an insightful visual performance model for multicore architectures". In: *Communications of the ACM* 52.4 (2009), pp. 65–76.

[12] François Broquedis et al. "hwloc: A generic framework for managing hardware affinities in HPC applications". In: *Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference on*. IEEE. 2010, pp. 180–186.

[13] Peter Bright. *AMD's "heterogeneous Uniform Memory Access" coming this year in Kaveri*. URL: http://arstechnica.com/information-technology/2013/04/amds-heterogeneous-uniform-memory-access-coming-this-year-in-kaveri/.

[14] Haluk Topcuoglu, Salim Hariri, and Min-you Wu. "Performance-effective and low-complexity task scheduling for heterogeneous computing". In: *Parallel and Distributed Systems, IEEE Transactions on* 13.3 (2002), pp. 260–274.

[15] A. Mariano. "Scheduling (ir)regular applications on heterogeneous platforms". Master Thesis. Gualtar, 4715 Braga: University of Minho, Sept. 2012.

[16] Miguel Palhas and Pedro Costa. *A Finite Volume Case Study From An Industrial Application*. Course Report, MSc Informatics Engineering. Tech. rep. University of Minho, 2012.

[17] Cedric Augonnet et al. "Data-aware task scheduling on multi-accelerator based platforms". In: *Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on*. IEEE. 2010, pp. 291–298.

[18] Cédric Augonnet, Samuel Thibault, and Raymond Namyst. "Automatic calibration of performance models on heterogeneous multicore architectures". In: *Euro-Par 2009–Parallel Processing Workshops*. Springer. 2010, pp. 56–65.

[19] Steven G Parker et al. "Optix: a general purpose ray tracing engine". In: *ACM Transactions on Graphics (TOG)* 29.4 (2010), p. 66.

[20] Niklas Huss. "Real Time Ray Tracing". PhD thesis. Linnaeus University, Faculty of Science, Engineering, School of Computer Science, Physics, and Mathematics, 2004.

[21] Arthur Appel. "Some techniques for shading machine renderings of solids". In: *Proceedings of the April 30–May 2, 1968, spring joint computer conference*. ACM. 1968, pp. 37–45.

[22] Turner Whitted. "An improved illumination model for shaded display". In: *ACM SIGGRAPH 2005 Courses*. ACM. 2005, p. 4.

[23] J.R. Wallace, M.F. Cohen, and D.P. Greenberg. *A two-pass solution to the rendering equation: A synthesis of ray tracing and radiosity methods*. Vol. 21. 4. ACM, 1987.

[24] E. Veach and L.J. Guibas. "Metropolis light transport". In: *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co. 1997, pp. 65–76.

[25] J.T. Kajiya. "The rendering equation". In: *ACM SIGGRAPH Computer Graphics* 20.4 (1986), pp. 143–150.

[26] H.W. Jensen. "Global illumination using photon maps". In: *Rendering Techniques* 96 (1996), pp. 21–30.

[27] Henrik Wann Jensen et al. "Monte Carlo ray tracing". In: *ACM SIGGRAPH*. 2003.

[28] Randima Fernando. *GPU gems: programming techniques, tips, and tricks for real-time graphics*. 2004.

[29] T. Hachisuka, S. Ogaki, and H.W. Jensen. "Progressive photon mapping". In: *ACM Transactions on Graphics (TOG)*. Vol. 27. 5. ACM. 2008, p. 130.

[30] T. Hachisuka and H.W. Jensen. "Stochastic Progressive photon mapping". In: ().

[31] C. Knaus and M. Zwicker. "Progressive photon mapping: A probabilistic approach". In: *ACM Transactions on Graphics (TOG)* 30.3 (2011), p. 25.

[32] Holger Dammertz, Johannes Hanika, and Alexander Keller. "Shallow bounding volume hierarchies for fast SIMD ray tracing of incoherent rays". In: *Computer Graphics Forum*. Vol. 27. 4. Wiley Online Library. 2008, pp. 1225–1233.

[33] Martin Stich, Heiko Friedrich, and Andreas Dietrich. "Spatial Splits in Bounding Volume Hierarchies". In: *Proc. High-Performance Graphics 2009*. 2009.

[34] Robert C Tausworthe. "Random numbers generated by linear recurrence modulo two". In: *Mathematics of Computation* 19.90 (1965), pp. 201–209.