Big Astrophysical Data Visualisation on the MIC architecture

Claudio Gheller[1], Timothy Dykes[2], Mel Krokos[2] and Marzia Rivi[3]

[1] CSCS-ETHZ, Lugano, Switzerland
cgheller@cscs.ch
[2] University of Portsmouth, Portsmouth, U.K.
timothy.dykes@myport.ac.uk, mel.krokos@port.ac.uk
[3] University of Oxford, Oxford, U.K.
rivi@physics.ox.ac.uk

Abstract

....

1 Introduction

The scientific data volume produced by experiments, observations and numerical simulations is increasing exponentially with time. This is true for any scientific domain, but in particular for astrophysics. Next generations of telescopes and antennas are expected to produce enormous amount of data. The Square Kilometer Array project (REF), for instance, will generate an exabyte of data every day, twice the information currently exchanged on the internet on a daily basis and 100 times more information then the CERN LHC (REF) experiment produces. At the same time, computer simulations represent an invaluable instrument for astrophysicists to validate theories and compare to observations through numerical experiments. The ultimate cosmological simulations (REF), performed using sophisticated N-body codes (REF), could describe the details of the evolution of the universe up to the present time, following the behaviour of gravitating matter represented by a hundred billion particles. These runs produce output files whose size is of the order of tens of terabytes each. The fast technological progresss of supercomputing systems will soon lead to simulations producing outputs with size of the order of the petabyte or more.

Size does not represent the only challenge posed by scientific data. Is also essential to effectively extract all the information hidden in the sea of bytes represented by each single data file. Software for data mining and analysis is often highly computationally demanding and ultimately unusable on large datasets.

Visual exploration and discovery represents an outstanding aid to big data processing, e.g. by providing scientists with prompt and intuitive insights enabling them to identify interesting characteristics and thus define regions of interest within which to apply time-consuming methods. Additionally, they can be a very effective way in discovering and understanding correlations, associations and data patterns, or in identifying unexpected behaviours or even errors. Visualization is also an effective means for communicating scientific results not only to researchers but also to members of the general public. However, also visualization tools require high performance computing (hereafter HPC) resources, to fulfil the requirements posed both from data size and from the need of having fast (if not real-time) rendering.

The Splotch software (REF), our ray-casting algorithm for effectively visualizing large-scale, particle-based datasets, addresses these issues, providing high quality graphic outputs, processing data of, ideally, any size, already efficiently exploiting a broad variety of HPC systems:

multi-core processors and multi-node supercomputing systems (REF: Procedia Computer Science, 1(1) pp.1775-1784, 2010), and GPUs (REF1: Astronomical Society of the Pacific Conference Series, 475 (ADASS XXII) pp.103-106, 2013; REF2: in preparation). This paper will describe the work accomplished to enable Splotch to run on the new Intel PHI (REF) accelerator, taking advantage of the Many Integrated Core (hereafter MIC) architecture, which is expected to provide, on suitable classes of algorithms, outstanding performance with power consumption being comparable to standard CPUs. We will present the MIC implementation and optimisations, performance tuning, benchmarks carried out, and the resulting performance measurements, comparing that of an OpenMP based implementation running on multiple cores of a single CPU. A brief overview of this implementation will also be given, referring to [REF] for further details.

2 Splotch Overview

Contents:
- Standard brief overview of Splotch workflow

2.1 MPI and OpenMP Implementations

Contents:
- Describing how OpenMP implementation works for reference when building upon this for the MIC implementation.
- Describing how MPI implementation works for reference when discussing how MPI is used to increase MIC performance.
- Mention GPU implementation with reference, no need for details

3 Splotch on the MIC

The MIC based implementation of Splotch uses both OpenMP and MPI in order to fully exploit the many core programming paradigm necessary to take advantage of this architecture. This section gives a brief overview of the Xeon Phi, and the steps taken to modify and optimise Splotch to make effective use of this hardware.

3.1 Overview of the MIC Architecture

The core ideal behind the MIC micro-architecture is obtaining a massive level of parallelism for high throughput performance in power restricted cluster environments. To this end Intel's flagship MIC product, the Xeon Phi, contains roughly 60 cores on a single chip, dependent on the model. Each core has access to a 512 KB private fully coherent L2 cache, memory controllers and the PCIe client logic provide access to up to 8 GB of GDDR5 memory, and a bi-directional ring interconnect brings these components together. The cores are in-order, however up to 4 hardware threads are supported to mitigate the latencies inherent with in-order execution. The Vector Processor Unit (hereafter VPU) is worthy of note due to the utilisation of an innovative 512 bit wide SIMD capability, allowing 16 single precision (SP) or 8 double precision (DP) floating point operations per cycle, support for fused-multiply-add operations increased this to 32 SP or 16 DP floating point operations per cycle.

The Xeon Phi acts as a coprocessor for a standard Intel Xeon processor connected via PCIe, and there are various modes of execution on a system utilising one or more Xeon Phi coprocessors. It runs the Linux operating system, and so can be seen as a networked node through a virtualised TCP/IP stack over the PCIe bus. This allows a user to log into the node, transfer a program over and run natively, or to use the coprocessor as an MPI process alongside the Xeon. The ability to partition subgroups of processors allows the Phi to run multiple MPI processes at the same time, a technique exploited in [insert subsection number]. In addition, a heterogenous approach is possible, using the coprocessor to accelerate a standard CPU based algorithm by offloading sections of computation to the device using a small series of directives available to both C++ and Fortran.

An advantageous factor of programming for the Xeon Phi, as opposed to other accelerators commonly used in HPC environments such as GPUs, is the similarity of techniques used to exploit parallelism on both the Xeon Phi and regular Xeon processors. Algorithms that already utilise parallel paradigms involving OpenMP, MPI, Intel TBB, or Intel Cilk Plus can often be run on the Xeon Phi with little modification. While further tuning is necessary to fully take advantage of the new hardware, it is not a necessity to reimplement the entire algorithm.

3.2 MIC Implementation

The implementation targeting the MIC is based on the original C++ and OpenMP version of Splotch (see section 2.1). While the core algorithm executes on the Xeon (referred to as the host), data is transferred to the Xeon Phi (referred to as the device) and the majority of the computation is offloaded using the C++ pragma extensions provided by Intel in a syntactic style similar to OpenMP, attention has been paid to optimising parts of the algorithm to better take advantage of the wide SIMD capability of the Phi.

3.2.1 Algorithm

The rendering algorithm can be broken down into a series of phases illustrated by the execution model shown in [figure x]. A double buffered scheme has been implemented using the ability to asynchronously transfer data via a series of signal and wait clauses provided by the Intel extensions. This allows to minimise overhead due to transferring data to the device for processing, and to facilitate the rendering of datasets potentially much larger than the memory capacity available, discussed further in section [section number].

[figure illustrating execution model]

A necessary overhead is incurred from the outset performing various initialisations discussed further in section 4.1. In addition to this, the first chunk of data must be transferred to the device before the double buffering system can begin to compensate for data transfer times. This initial transfer is carried out asynchronously while precomputing render parameters such as those used in the 3d transformation, which will be static throughout the rendering process. The render parameters are copied to the device, which subsequently begins rendering while the next chunk of data is transferred.

The first stage of rendering is a highly parallel 3D transform and colorize performed on a per-particle basis using four OpenMP threads per available core, equating to roughly 240 threads assuming the current process has access to the whole device. [expand on this]

The second stage is not so simply parallelised, leading to more complex solution. A problem inherent in parallel rendering is the appearance of race conditions arising from multiple threads attempting to render to the same pixel. In a scene with potentially billions of particles, it is inevitable that multiple particles, potentially a vast number, will overlap and need to be rendered to the same pixel. [discuss this somewhere else?]

The number of available OpenMP threads are split into groups, allocating an image buffer per group large enough to hold the entire resulting image. Each group is assigned a subsection of the particles where n_particles_per_group = total_particles / n_groups. [Latex up this equation including remainder compensation]. At this point each thread group acts independently, drawing the allocated subsection of particles to the allocated image, finally reducing all images into a single buffer when all particles of the current data chunk have been processed, a system conceptually similar to the MPI Splotch implementation discussed in section 2.1. The number of groups to create and number of threads per group can be passed in as a runtime parameter, and the chosen values should reflect consideration of the available number of threads and thread:core ratio. This is discussed further in section 4.1.

Subsequent to group allocation, each thread group begins independently rendering an allocated subset of particles. This occurs in two phases, a pre render phase and a render phase. In order to allow each thread sole access to a particular set of pixels, and avoid race conditions discussed previously, the image is split into 2 dimensional grid of tiles the number of which is determined by a run time parameter tile_size. The pre-render phase generates a list of particle indices per tile, indicating all the particles whose area of influence overlaps with the tile. In this phase each thread is allocated a subset of the group's allocated particles, and generates a list for each tile resulting in (n_thread * n_tile) lists. A single thread accumulates the per-thread lists to attain a single list per tile. Once all lists have been accumulated, phase two begins. In this phase each thread is allocated a tile, or subset of pixels, and renders all particles in the list associated with that tile. In this way pixels are not shared between threads and concurrent accesses are avoided.

Following the accumulation of each group-specific image into a single buffer, which is retained throughout the entire rendering process, the next chunk of data is processed until all particles have been rendered. This constant buffer is then copied back to the host to be written to file.

3.2.2 Optimisation

In order to appropriately exploit the MIC architecture and attain high performance gains, various optimisation methods were employed. Some more generic methods are applicable to other architectures, such as Intel 64 and IA-32 [REF], while others are specific to the Xeon Phi. For example, cost of dynamic memory allocation on the Phi is relatively high [REF1], in order to minimise unnecessary allocations buffers are created at the beginning of the program cycle and reused throughout. This is not such a large concern in a standard Intel 64 environment where frequent memory allocations, such as might be invoked by repetitive resizing of dynamic arrays, carry little comparative cost.

Often scientific datasets being visualized are very large, it is inevitable that there may at times be more data than available device memory. This raises the issue of waiting for data transfer from main memory which, while reaching roughly 6 GB/s over the 16 channel PCIe 2.0 connection, is an additional overhead to be considered. The possibility of data transferral and computation being carried out asynchronously has allowed for a double buffered method to overlap these stages, minimising the overhead. In addition, the MIC_USE_2MB_BUFFERS environment variable is set to 64K, forcing buffers larger than this value to be allocated in 2MB pages, which improves data transfer rate and can benefit performance by potentially reducing page faults and TLB (translation look-aside buffer) misses. [REF2]

The large 512 bit width SIMD capability of the MIC architecture is addressed through vectorization carried out both automatically by the compiler, and manually using Intel Initial Many-Core Instructions (IMCI) [REF3] . Firstly the core data structure used in Splotch rendering was re-examined, and converted from an array of structures (AoS) to a structure of arrays

(SoA). The AoS method stores a large array of structures, each containing all of the information pertaining to a single particle. While this is useful for encapsulation, it is unsuitable for SIMD processing. [REF4] [Run test with aos vs soa and show results?]

- Use of MPI to partition Phi's cores into sections

4 Performance Analysis

- Contents:
- Performance analysis methods, test hardware etc

4.1 Tuning

- Necessary overhead (mic initilisation, omp thread initialisation - check thread stack size allocations effect on this?)
- Thread grouping, tile size, number of mpi processes used, results

4.2 Scalability

- Performance tests on various sizes of data compared to CPU

5 Discussion and Conclusions

...

5.1 Acknowledgments

...