Institution: University of Oxford

Unit of Assessment: 10 – Mathematical Sciences

Title of case study: Development and implementation of mathematical algorithms enhance performance of software libraries on GPUs

1. Summary of the impact

Many of the top supercomputers use Graphical Processing Units (GPUs) to accelerate scientific computing applications with less energy consumption and lower overall cost. GPUs achieve this by having comparatively large numbers of simple processing elements when compared against CPUs, which have fewer, more sophisticated, elements. However, to take full advantage of GPUs requires quite different algorithms and implementation techniques for mathematical software libraries. Researchers at the University of Oxford have developed a number of such algorithms and implementation techniques over the period 2008-2013, which have been incorporated into software libraries distributed by NAG, NVIDIA and the Apache Foundation and have enhanced the performance up to 150x compared with single thread CPU calculations and 20x relative to multithreaded CPU calculations. These libraries are used by large numbers of application developers worldwide.

2. Underpinning research

In 2007, Professor Mike Giles began research on the use of GPUs for Monte Carlo simulations. This exploited the new capability to use NVIDIA graphics cards for high performance computing (HPC) applications through writing programs using CUDA, NVIDIA's proprietary extension to the computer language C. The performance benefits proved to be substantial, with many of the newly developed algorithms exhibiting a speed-up of over 150x compared with single thread CPU execution [2].

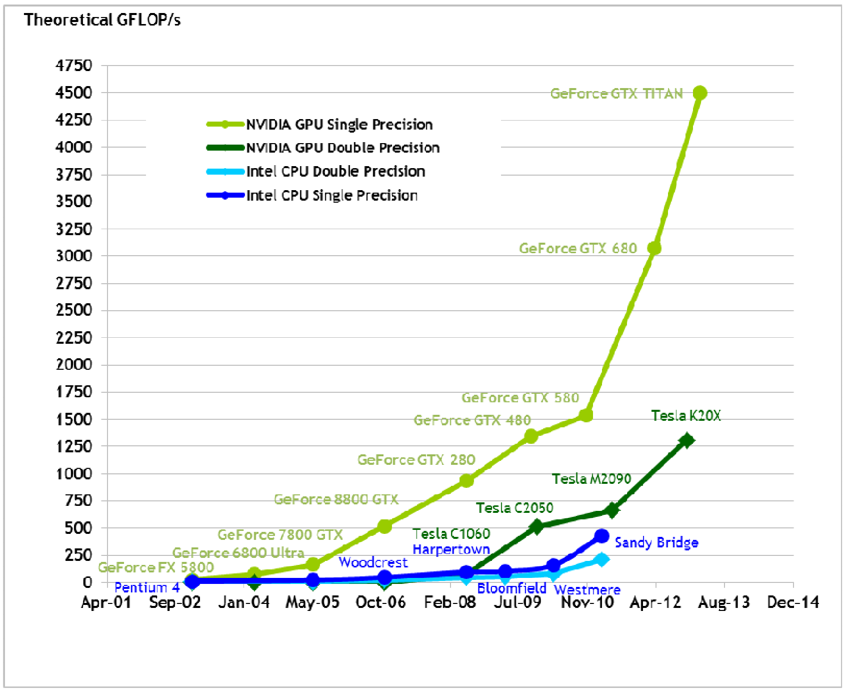

Figure 1: Graph showing the number of theoretical Floating-Point Operations per Second for CPUs and GPUs, clearly showing the enhanced computing power achievable using GPUs (figure taken from NVIDIA's CUDA website).

This led to a collaboration with Professor Chris Holmes at the University of Oxford's Statistics Department and Professor Arnaud Doucet (who was then at University of British Columbia but moved to the University of Oxford in 2012), in which they demonstrated the performance that could be achieved for more challenging statistical applications such as particle filters; the primary challenge is the re-weighting of the particles which is not easily parallelised [1].

Part of Giles' research programme involved the massively parallel implementation of random number generators, including both L'Ecuyer's mrg32k3a pseudo-random generator, and Sobol's quasi-random generator. This work is documented in [2] which has co-authors from both NAG and NVIDIA; both companies have adopted these generators in their respective random number libraries (see below).

One standard method of converting uniform random numbers into Normal random numbers is through inverting the Normal cumulative distribution function. This is a simple affine transformation of the inverse error function which is a standard function of many mathematical libraries, but the standard way in which it is approximated performs very poorly on GPUs because of their vector computing nature. This led to Giles developing a new approximation which is detailed in [3].

Another key component in many engineering and scientific applications is sparse matrix-vector multiplication. It is easy to implement this efficiently on CPUs, but much harder on GPUs due to the very limited amount of level 1 cache (memory) available to the large number of compute threads. Giles addressed this in a novel way by using multiple compute threads to cooperate to compute each one of the output elements, resulting in a factor 2x speedup compared with NVIDIA's existing CUSPARSE implementation [4].

3. References to the research

* [1] A. Lee, C. Yau, M.B. Giles, A. Doucet, C.C. Holmes. 'On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods'. *Journal of Computational and Graphical Statistics*, 19(4): 769-789, 2010.
DOI: 10.1198/jcgs.2010.10039 (Google Scholar: 94 citations, Web of Knowledge: 17 citations)

* [2] T. Bradley, J. du Toit, M.B. Giles, R. Tong, P. Woodhams. 'Parallelisation techniques for random number generators'. pp.231-246 in GPU Computing Gems, Emerald Edition, Morgan Kaufmann, 2011. ISBN: 0123849888

* [3] M.B. Giles. 'Approximating the erfinv function'. pp.109-116 in GPU Computing Gems, Jade Edition, Morgan Kaufmann, 2011. ISBN: 978-0-12-385963-1

[4] I. Reguly, M.B. Giles. 'Efficient sparse matrix-vector multiplication on cache-based GPUs' IEEE Refereed Proceedings of Innovative Parallel Computing Conference, 2012.
DOI: 10.1109/InPar.2012.6339602

The three asterisked outputs best indicate the quality of the underpinning research. [1] is in a major international journal, while [2] and [3] are chapters in research monographs instigated by NVIDIA.

4. Details of the impact

Since 2008, the University of Oxford's work on GPUs has had both economic impact and impact on practitioners and professional services, via improvement of existing software and the provision of consulting services. The beneficiaries are NAG, NVIDIA, the Apache Foundation, and the large number of people who use the improved software produced by these companies.

The impact has been achieved through the transfer of software developed by Giles as part of his research. This has gone into libraries developed and maintained by NAG, NVIDIA, and Apache. With NAG and NVIDIA, this came about through long-standing research collaborations and personal contacts. In the case of the Apache Foundation (which develops open-source software),the organisation asked for Giles' software as a result of reading his papers.

The first impact of this research was on the Numerical Algorithms Group (NAG) and through their subsequent dissemination of the software to the financial industry. NAG is an Oxford-based world leader in the development of mathematical computer software libraries, whose products are used worldwide in both academia and industry. NAG and Giles have long-standing connections, including the recent development of a wholly new GPU-based library [B] targeted at the needs of the finance industry. The Vice President of Sales at NAG states [A] "*Thanks to your contributions, we developed the GPU random number generation library quite quickly; I think you were responsible for over half of the original code before it went into our quality assurance process*". He further states that "*It* [the NAG GPU library] *is being used by two major banks and two others have purchased related consultancy services and training….We have also benefited indirectly from this project, for example one major Tier 1 bank made a significant licence upgrade of ~£100,000 and this upgrade only became possible by NAG establishing new contacts within the bank through our GPU work. It is important that we are seen by our customers as being at the cutting edge of scientific computing research, and our work in areas like the GPU library is key to that and does help us to keep existing customers and bring in new ones; over the past 3 years the percentage of top banks who use our software has increased to 60%.*" NAG report [C] that the French bank BNP Paribas reported excellent speed up results (between 150x and 240x relative to a single-threaded CPU simulation, depending on the number of simulations).

Giles has contributed fundamental software components to two of the NVIDIA's mathematical libraries for sparse linear algebra (CUSPARSE) and random number generation (CURAND) [D]. Both are integral to many scientific applications using NVIDIA's GPUs, without which the scientific applications would be unable to take advantage of the acceleration hardware in the leading supercomputers. The inverse error function routine (erfinv) is now part of their standard mathematics library, while the Sobol quasi-random generator is part of the CURAND random number generation library. The implementation of the mrg32k3a pseudo-random generator is based on [2] referenced above, and the sparse matrix-vector product (spMV) routine was put into the CUSPARSE library to replace the previous version developed internally by NVIDIA. Both of these algorithm libraries were developed and analysed at Oxford. The spMV routine is also a foundation for NVIDIA's new NVAMG solver, an algebraic multigrid solver which is the basis for a new GPU version of Ansys' Fluent computational fluid dynamics (CFD) software, in turn probably the leading commercial CFD solver worldwide. This illustrates the layered nature of software development, with high-level packages addressing specific applications layered on top of lower-level, more generic, more fundamental software.

The Senior Manager for CUDA Libraries and Algorithms at NVIDIA states in his support letter [D] "*It is clear that high quality efficient software libraries are important to users implementing their algorithms on our hardware, and if we did not have our CUDA libraries then NVIDIA would not hold the powerful position within HPC which it does. This is well illustrated by the fact that the Titan system at Oak Ridge National Laboratory, the top supercomputer in the world according to the Top500 list, is based on our new Kepler GPUs. Other indications, quoting from our CEO's keynote presentation in the 2013 GTC conference* [E], *are that in 2012 we sold 100M CUDA-capable GPUs, and the CUDA development kit, including all of the libraries, was downloaded 1.6M times…..Although it is hard to quantify the impact, indications of our appreciation of the impact of your work are that we made you one of our inaugural CUDA Fellows in 2008 (there are still only 11 worldwide and you are the only one in the UK), and we made Oxford University a CUDA Centre of Excellence (CCoE) in 2012*". This award included a donation of $100k to support undergraduate research internships, and hardware donations with a value of approximately another $100k, with further donations likely in future years.

Finally, the inverse error function approximation software has also been adopted by the Apache Software Foundation for its Java-based Apache Commons Math library [F]. This has widespread use across numerous sectors, where it is used to convert uniformly-distributed random numbers into normally-distributed random numbers for stochastic simulations written in Java, in application areas as diverse as financial option pricing, biochemical reaction modelling, engineering uncertainty quantification, and the simulation of groundwater flow in nuclear waste repositories.

5. Sources to corroborate the impact

[A] Letter from Vice President of Sales, Numerical Algorithms Group (NAG), dated 12 June 2013, detailing the significance of GPU computing and Giles' influence on GPU computing at NAG. Copy held by the University of Oxford.

[B] Information on GPU computing on the NAG website which mentions Mike Giles by name: http://www.nag.co.uk/numeric/GPUs/index.asp

[C] NAG presentation detailing performance results generated by BNP Paribas http://www.nag.co.uk/numeric/gpus/FinanceNVIDIA.pdf

[D] Letter from the Senior Manager, CUDA Libraries and Algorithms, NVIDIA, dated 27 March 2013, detailing Giles' contribution to CUDA libraries and their impact on NVIDIA. Copy held by the University of Oxford

[E] 2013 presentation by NVIDIA CEO Jen-Hsun Huang. http://www.ustream.tv/recorded/30095793 confirms the number of downloads of the library

[F] Email from Independent Contributor to Apache Commons Math library, confirming their use of Giles' implementation of erfinv. Copy held by University of Oxford