Institution: The University of Edinburgh

Unit of Assessment: B11 — Computer Science and Informatics

Title of case study: MilePost GCC and compiler research at Edinburgh

1. Summary of the impact

Compiler research at Edinburgh over the last decade has had significant industrial and commercial impact. Early work on pointer conversion is now available in Intel's commercial compilers. Later ground-breaking work on machine-learning based compilation led to the release of MilePost GCC, an enhanced version of the world's widest-used open source compiler supported by IBM. More recent work on parallelism discovery and machine-learning mapping has led to a new ARM Centre of Excellence at Edinburgh.

2. Underpinning research

University of Edinburgh researchers involved in this case study are listed below.

Professor Michael O'Boyle, 1997–date	Professor Christopher Williams, 1998–date
Professor Nigel Topham, 2003–date	Björn Franke, Reader, 2003–date
Christophe Dubach, PhD Edinburgh 2009, Lecturer, RAEng Research Fellow and Intel Honor Fellow	Hugh Leather, PhD Edinburgh 2010, Lecturer, RAEng Research Fellow
Timothy M. Jones, PhD UoE 2006, RAEng / EPSRC Research Fellow. Left UoE 2011.	Grigori Fursin, PhD Edinburgh 2004. Research Assistant. Left UoE 2005.

2.1. Pointer conversion

Embedded systems account for the vast majority of shipped processors and require high performance and energy efficiency at low cost. Until recently the compiler technology for such systems was poor. This was partly due to unconventional processor architectures and the pointer-based structure of the programs. Franke and O'Boyle (2001) developed the first pointer conversion scheme that automatically recovers linear array accesses in digital signal processing applications. This opened up the possibility of applying the large body of literature in high-level transformations to DSP programs for the first time to dramatic effect. Embedded systems are now parallel and multi-core in nature. However, the complex and non-standard memory model of such systems means that they are extremely difficult to program. Franke and O'Boyle (2003) developed the first ever auto-parallelisation approach for multiple address space DSPs. This required the combination of pointer-recovery and a new rank-modifying transformation framework to reconcile location of memory addresses and enable communication optimisation.

2.2. Iterative compilation and auto-tuning via machine learning

Traditional approaches to optimisation rely on static models of program/processor interaction. O'Boyle (1998) was the first to show that such an approach poorly models the interaction and is fundamentally flawed. This led to work in iterative compilation that formulated the transformations available as a formal optimisation space and applied search-based techniques. This work has been widely used and shown to outperform all existing techniques. Iterative compilation and auto-tuning are now standard topics in compiler- and performance-based conferences. Our research work has incorporated machine-learning techniques directly into the search, modelling transformation spaces as Markov processes, which can then be learnt [1]. This has been used to speed up the performance of iterative compilation by an order of magnitude and dramatically improve the performance of Just-In-Time (JIT) compilation. This research has led to the development of compilers that can self-adapt and learn about the optimisation space automatically, outperforming the best hand-tuned compiler-writer heuristics.

2.3. Applying machine learning to compilers and architectures

The machine-learning-based approach has extended beyond compiler optimisation to consider the compiler/architecture design space. Dubach and O'Boyle [3] developed modelling approaches that could simulate and predict the performance of any architecture configuration. This approach was then extended [4] to predict the performance of an optimising compiler on any architecture and finally to automatically generate an optimising compiler for any architecture. In addition, we have developed techniques that dynamically adjust hardware to the predicted best on-line configuration allowing hardware to adapt to workloads, reducing energy consumption [5].

2.4. Innovations in auto-parallelisation

Since 2009, Franke and O'Boyle have developed a unique approach to auto-parallelisation. First they developed a machine-learning-based approach to mapping different forms of parallelism to varying architectures outperforming all existing techniques. In 2009, they developed an innovative approach to determining the best mapping of parallelism with profile-directed discovery of parallelism [2]. This has then been extended to the heterogeneous multi-core space. Franke and Topham's research on parallel JIT compilation [6] has contributed to the scientific and commercial success of the ArcSim dynamic binary translator. Parallel JIT compilation is a novel concept to hide JIT compilation latency and to increase compiler throughput on standard multi-core host machines. This results in unprecedented simulation speeds of single-core and multi-core simulators beyond those of actual speed-optimised silicon implementations of the system under simulation.

3. References to the research

3.1. Publications

1. *Using Machine Learning to Focus Iterative Optimization*. F. Agakov, E.V. Bonilla, J. Cavazos, B. Franke, G. Fursin, M.F.P. O'Boyle, J. Thomson, M. Toussaint, and C.K.I. Williams, Proceedings of the International Symposium on Code Generation and Optimization (CGO '06), pages 295-305, March 2006. (doi: *10.1109/CGO.2006.37*)

2. *Towards a Holistic Approach to Auto-Parallelization: Integrating Profile-Driven Parallelism Detection and Machine-Learning Based Mapping*. Z. Wang, B. Franke and M. O'Boyle, Proceedings of the ACM SIGPLAN 2009 Conference on Programming Language Design and Implementation (PLDI '09), June 2009. Pages 177–187. (doi: *10.1145/1542476.1542496*)

3. *Portable Compiler Optimization Across Embedded Programs and Microarchitectures using Machine Learning*. C. Dubach, T.M. Jones, E.V. Bonilla, G. Fursin and M.F.P. O'Boyle, 42nd IEEE/ACM International Symposium on Microarchitecture (MICRO '09), December 2009. Pages 78–88. (doi: *10.1145/1669112.1669124*)

4. *Partitioning Streaming Parallelism for Multi-cores: A Machine Learning Based Approach.* Z. Wang and M. O'Boyle, In 19th International Conference on Parallel Architectures and Compilation Techniques (PACT '10), September 2010. Pages 307–318. (doi: *10.1145/1854273.1854313*)

5. *Predictive Model for Dynamic Microarchitectural Adaptivity Control*. C. Dubach, T.M. Jones, E.V. Bonilla, and M.F.P. O'Boyle, In 43rd IEEE/ACM International Symposium on Microarchitecture (MICRO '10), December 2010. Pages 485–496. (doi: *10.1109/MICRO.2010.14*)

6. *Generalized Just-In-Time Trace Compilation using a Parallel Task Farm in a Dynamic Binary Translator*. Igor Bøhm, T.J.K. Edler von Koch, S. Kyle, B. Franke, and N. Topham, Proceedings of the 32nd ACM SIGPLAN conference on Programming Language Design and Implementation (PLDI '11), June 2011, San Jose, California, USA. Pages 74–85. (doi: *10.1145/1993498.1993508*)

References [1], [2] and [3] above are most indicative of the quality of the underpinning research.

3.2. Research grants and funding

- EP/G000691 Machine Learning for Thread Level Speculation on Multicore architectures £350,652
- EP/I013539 Dynamic Adaptation in Heterogeneous Multicore Embedded Processors £1,217,557
- EP/H051988 A predictive modelling based approach to portable parallel compilation for heterogeneous multi-cores £494,120
- EP/K008730 PAMELA: A Panoramic Approach to the Many-Core Landscape £4,135,048 (3 partners)
- EU HiPEAC 2 Network of Excellence FP7 c £400,000 2008-2012
- EU HiPEAC 3 Network of Excellence FP7 c £400,000 2012-2016
- EU TETRACOM – technology transfer project c. £100,000

3.3. Awards and fellowships

- Tim Jones, Christophe Dubach, Hugh Leather, Christian Fensch – Royal Academy of Engineering Five-year Research Fellowships
- Christophe Dubach CPHC/BCS Distinguished Dissertation award 2009

4. Details of the impact

4.1. Impact of pointer conversion

Pointer conversion is now available in Intel's commercial *icc* compiler. This was added in 2005, and continues to be used in versions 11.0, 11.1, 12.0, 12.1, and 13.0 of this compiler, released in 2008, 2009, 2010, 2011, and 2012. Intel dominates the desktop and high-end processor market. Research undertaken at Edinburgh is now used to improve code performance on Intel platforms across the world. This is a wide impact since the vast majority of desktop machines are Intel-based: according to *http://www.cpubenchmark.net/market_share.html*, estimates of market share since 2008 show that Intel has between 70% and 73% of the x86 processor market with ARM providing almost all the rest. A smaller scale company CAPS-Enterprise (approximately 20 people) are also known to have implemented this technique in their software tool chain, which is used by Intel in their library development.

4.2. Impact of machine-learning-based approaches

GCC is the most widely used compiler in the world. It is open-source and has a large community of academic and industrial contributors of which IBM is the leader. Working with IBM we developed MilePost GCC, a compiler that automatically learns to optimise [A, B, C]. The learning component is available as a simple plug-in that determines optimisation based on prior knowledge. Uniquely this can access a shared database allowing community-based continuous optimisation. There have been 643 downloads by developers, the number of end users is not known. This work led to the creation of the Collaborative Tuning resource [D], a platform for exchange of best practice in performance optimisation of program code.

4.3. Impact of machine-learning-based approaches

Our work on compiler/architecture co-design in collaboration with the architecture group at the School of Informatics influenced the design of the reconfigurable EnCore processor. The associated ArcSim high-performance architecture simulator is based on the parallel JIT compiler technology developed by us. EnCore and ArcSim are the subject of a separate School of Informatics REF impact case study.

4.4. Impact of auto-parallelisation research

Combining our experience in parallelisation with machine-learning-based optimisation has led to a major breakthrough in the area of auto-parallelisation. This was recognised when ARM made a substantial investment in a heterogeneous parallelism centre of excellence at Edinburgh [E]. This

is ARM's first centre of excellence outside the University of Michigan. The centre funds fundamental research in data centre scale parallelism leading to patentable ARM IP. We are currently jointly working with ARM on an LLVM-based OpenCL compiler based on this work. This work has attracted considerable industrial interest: NVIDIA has made one of our students a fellow for our work on heterogeneous parallelisation while Freescale, Imagination Technology and IBM are collaborating on a variety of projects. Samsung is developing a prototype based on our JIT technology. The pioneering work on profile-directed parallelisation is the on-going subject of commercialisation. The University of Edinburgh and Samsung have signed a collaboration agreement [F] publicised at *http://wcms.inf.ed.ac.uk/icsa/news/samsung-research-collaboration*.

4.5. Details of on-going collaboration arrangements with industrial partners

The ARM centre of excellence has two components: an overarching collaboration agreement, and student project agreements. This allows intellectual property to be jointly created and exploited by all parties. Students have a supervisor at both ARM and Edinburgh. They are paid an enhanced stipend and undertake a three-month internship during their studies.

In 2012 Intel announced expansion of its Intel Doctoral Student Honour Programme into Europe. The University of Edinburgh was one of only three universities in the UK to be selected. In 2012 one of our students Bhargava Rajaram was awarded an Intel PhD fellowship [G]. Christophe Dubach was awarded an Intel Early Career Faculty award: this was the only award made to a UK academic [H].

5. Sources to corroborate the impact

A. MilePostGCC press release. *http://www-03.ibm.com/press/us/en/pressrelease/27874.wss*
B. High-Impact ICT research: “Machine-learning revolutionises software development”. *http://cordis.europa.eu/ictresults/index.cfm?section=news&tpl=article&ID=91208*
C. An open-source machine-learning compiler that intelligently optimizes applications. Dr Dobb’s Software Journal. *http://www.drdobbs.com/open-source/milepost-gcc-now-available/218102130*
D. The Collective Tuning website. *http://ctuning.org*
E. University of Edinburgh and ARM Research Centre of Excellence Framework agreement. This is a commercially sensitive document describing the details of the collaboration agreement between the University of Edinburgh and ARM. Copies can be made available on request.
F. University of Edinburgh and Samsung Research Collaboration agreement. This is a commercially sensitive document which describing the details of the collaboration agreement between the University of Edinburgh and Samsung. Copies can be made available on request.
G. Intel Doctoral Student Honour Programme. *http://www.intel.com/content/www/us/en/education/university/intel-2012-doctoral-student-honor-awardees.html?wapkw=2012+doctoral+student+honor+awardees*
H. Intel University Collaborative Research *https://www.intel-university-collaboration.net/?ai1ec_event=early-career-faculty-awards*

Copies of these web page sources are available at *http://ref2014.inf.ed.ac.uk/impact*