Bastian Hagedorn

Einsteinstraße 62
48149 Münster, Germany
✉ *hagedornbastian@gmail.com*

Professional Experience

since 2020 **Senior Deep Learning Compiler Engineer**, *NVIDIA*, Würselen, Germany.
As part of NVIDIA's Deep Learning Compilation Group, I develop novel compilation techniques which will allow to generate high-performance programs for the rapidly changing landscape of specialized hardware accelerators.

University Education

2016 – 2020 **PhD degree in computer science**, *University of Münster*, Münster, Germany.
Supervisor: Prof. Sergei Gorlatch, Prof. Michel Steuwer (University of Glasgow)
Thesis: *High-Performance Domain-Specific Compilation without Domain-Specific Compilers.*

2014 – 2016 **Master of Science in computer science (with distinction)**.
University of Münster, Germany

2011 – 2014 **Bachelor of Science in computer science with a minor in mathematics**.
University of Münster, Germany

Awards and Grants

08/2020 **PhD thesis**, *honored with the highest possible grade*, **Summa cum laude**.

09/2019 **Heidelberg Laureate Forum 2019**, *Invited Participation*, Invited to attend this highly prestigious forum with laureates of ACM A.M. Turing Award, Abel Prize, Fields Medal and Nevanlinna Prize. **Only 100 spaces** available for computer science.

12/2018 **NVIDIA Graduate Fellowship 2019**, *50.000$*, One of ten recipients world wide, selected as the only student from a European university.

02/2018 **Best Paper Award**, *ACM International Symposium on Code Generation and Optimization (CGO) 2018*, Selected as the best paper out of 103 submissions.

since 2016 **Four Travel Grants**, *HPC-Europa3, HiPEAC, 2x EuroLab-4-HPC*, total approx. 12.500€.

Research Visits

11/2019 **Visiting researcher (2 weeks)**, *University of Glasgow*, Glasgow, UK.
During this visit, I prepared a publication about Elevate. We especially investigated how to implement TVM's scheduling language in terms of composable rewrite rules using Elevate.

12/2018 – 06/2019 **Deep Learning Compiler Engineer (7 months)**, *NVIDIA*, Würselen, Germany.
During this time, I extended the IR developed during my Internship in the US, focusing on code generation targeting NVIDIAs tensor cores in order to accelerate low-precision deep learning applications on GPUs.

07/2018 – 09/2018 **Deep Learning Compiler Engineer Intern (3 months)**, *NVIDIA*, Redmond, WA, USA.
During this internship, I was working on an embedded DSL, IR and compiler for optimizing deep learning applications. I focused especially on CUDA code generation for highly optimized matrix multiplication algorithms. My work on this project simplified managing the complex compute and memory hierarchy on modern GPUs and enabled using program synthesis for design space exploration.

02/2018 – 04/2018 **Visiting researcher (2 months)**, *University of Glasgow*, Glasgow, UK.
Funded by HPC-Europa3
During this visit, I investigated the implementation of performance portable HPC applications with Lift including the automatic fusion of multiple compute kernels using rewrite rules. I focused on geometric multigrid methods. The Irish Centre for High-End Computing (ICHEC) supported my visit by providing access to their GPU hardware.

07/2017 – 09/2017 **Visiting researcher (2 months)**, *University of Edinburgh*, Edinburgh, UK.
Funded by HiPEAC
During this visit, I combined modern auto-tuning techniques with the current Lift code generator. I also evaluated Lift's functional compilation approach compared to state-of-the-art polyhedral compilation. A paper describing the results of this and our previous collaborations has won the *best paper award* at the prestigious ACM International Symposium on Code Generation and Optimization (CGO) [8]

02/2017 – 03/2017 **Visiting researcher (2 months)**, *University of Edinburgh*, Edinburgh, UK.
Funded by EuroLab-4-HPC
During this visit, I extended the Lift compiler, developed at the University of Edinburgh, to enable automatic exploration of stencil-specific optimizations.

04/2016 – 05/2016 **Visiting researcher (2 months)**, *University of Edinburgh*, Edinburgh, UK.
Funded by EuroLab-4-HPC
During this visit, I extended the Lift compiler to enable the generation of high-performance stencil code for GPUs.

09/2015 **Visiting researcher (3 weeks)**, *HUST University*, Wuhan, China.
Funded by the EC's 7th Framework Programme MONICA for accelerating the transfer and deployment of research knowledge between European countries and China. During this visit, I implemented an experimental setup for SDN-based multicast, and prepared a research paper on this topic [10]

Publications

2020 [1] **B. Hagedorn**, J. Lenfers, T. Koehler, X. Qin, S. Gorlatch, and M. Steuwer. "Achieving High-Performance the Functional Way - A Functional Pearl on Expressing High-Performance Optimizations as Rewrite Strategies". In: *25th ACM SIGPLAN International Conference on Functional Programming (ICFP), Virtual Event, August 24-26, 2020*. ACM, 2020.

[2] **B. Hagedorn**, A. S. Elliott, H. Barthels, R. Bodik, and V. Grover. "Fireiron: A Data-Movement-Aware Scheduling Language for GPUs". In: *29th International Conference on Parallel Architectures and Compilation Techniques, PACT 2020, Virtual Event, October 3-7, 2020*. (accepted for publication). IEEE, 2020.

[3] **Bastian Hagedorn**, A. S. Elliott, H. Barthels, R. Bodik, and V. Grover. "Fireiron: A Scheduling Language for High-Performance Linear Algebra on GPUs". In: *arXiv preprint: 2003.06324 [cs.PL]* (2020).

[4] **Hagedorn, Bastian**, J. Lenfers, T. Koehler, S. Gorlatch, and M. Steuwer. "A Language for Describing Optimization Strategies". In: *arXiv preprint: 2002.02268 [cs.PL]* (2020).

[5] T. Remmelg, **B. Hagedorn**, L. Li, M. Steuwer, S. Gorlatch, and C. Dubach. "High-Level Hardware Feature Extraction for GPU Performance Prediction of Stencils". In: *Proceedings of the Annual Workshop on General Purpose Processing Using Graphics Processing Unit (GPGPU)*. 2020.

[6] L. Stoltzfus, **B. Hagedorn**, M. Steuwer, S. Gorlatch, and C. Dubach. "Tiling Optimizations for Stencil Computations Using Rewrite Rules in Lift". In: *ACM Transactions on Architecture and Code Optimization (TACO)* 16.4 (2020), 52:1–52:25, Rank B.

2019 [7] P. M. Phothilimthana, A. S. Elliott, A. Wang, A. Jangda, **B. Hagedorn**, H. Barthels, S. J. Kaufman, V. Grover, E. Torlak, and R. Bodik. "Swizzle Inventor: Data Movement Synthesis for GPU Kernels". In: *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS*. 2019.

2018 [8] **B. Hagedorn**, L. Stoltzfus, M. Steuwer, S. Gorlatch, and C. Dubach. "High Performance Stencil Code Generation with Lift". In: *Proceedings of the 2018 ACM/IEEE International Symposium on Code Generation and Optimization, CGO 2018, Vösendorf / Vienna, Austria, February 24-28, 2018*. Ed. by J. Knoop, M. Schordan, T. Johnson, and M. F. P. O'Boyle. (**Best Paper Award, Most cited paper of CGO'18**). ACM, 2018, 100–112, Rank A.

2017 [9] **B. Hagedorn**, M. Steuwer, and S. Gorlatch. "A Transformation-Based Approach to Developing High-Performance GPU Programs". In: *Perspectives of System Informatics - 11th International Andrei P. Ershov Informatics Conference, PSI 2017, Moscow, Russia, June 27-29, 2017, Revised Selected Papers*. Ed. by A. K. Petrenko and A. Voronkov. Vol. 10742. Lecture Notes in Computer Science. Springer, 2017, 179–195, Rank B.

2016 [10] T. Humernbrum, **B. Hagedorn**, and S. Gorlatch. "Towards Efficient Multicast Communication in Software-Defined Networks". In: *2016 IEEE 36th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. June 2016, pp. 106–113. DOI: 10.1109/ICDCSW.2016.15.

2015 [11] M. Haidl, **B. Hagedorn**, and S. Gorlatch. "Programming GPUs with C++14 and Just-In-Time Compilation". In: *Parallel Computing: On the Road to Exascale, Proceedings of the International Conference on Parallel Computing, ParCo 2015, 1-4 September 2015, Edinburgh, Scotland, UK*. 2015, pp. 247–256.

[12] F. Stahl, A. Godde, **B. Hagedorn**, B. Köpcke, M. Rehberger, and G. Vossen. "High Quality Information Delivery: Demonstrating the Web in Your Pocket for Cineast Tourists". In: *Proceedings of the BTW 2015*. 2015, pp. 667–670.

2014 [13] F. Stahl, A. Godde, **B. Hagedorn**, B. Köpcke, M. Rehberger, and G. Vossen. "Implementing the WiPo architecture". In: *E-Commerce and Web Technologies*. Springer, 2014, pp. 1–12.

Presentations

03/2020 Invited Talk: *Elevate - A Language for Describing Optimization Strategies.*
Microsoft Research, Cambridge, UK

01/2020 Invited Talk: *Elevate - A Language for Describing Optimization Strategies.*
NVIDIA, Redmond WA, USA

04/2019 Talk: *High Performance Matrix Multiplication using Tensor Cores.*
NVIDIA, Redmond WA, USA

03/2019 Talk: *Elevate - A Language for Expressing Optimization Strategies.*
Scottish Programming Language Seminar (SPLS), University of St. Andrews, UK

12/2018 Talk: *High Performance Code Generation for Matrix Multiplication and Machine Learning.*
NVIDIA, Redmond WA, USA

10/2018 Talk: *High Performance Geometric Multigrid Operations in Lift.*
HPC-Europa3 Transnational Access Meeting (TAM), Edinburgh, UK

04/2018 Talk: *High Performance Stencil Code Generation with Lift.*
Workshop on Compilers for Parallel Computing (CPC), Dublin, Ireland

04/2018 Invited Talk: *High Performance Stencil Code Generation with Lift*.
Dependable Systems Group, Heriot-Watt University Edinburgh, UK

04/2018 Tutorial: *Lift: Performance Portable Parallel Code Generation via Rewrite Rules*.
International Symposium on Performance Analysis of Systems and Software (ISPASS), Belfast, UK

03/2018 Talk: *High Performance Stencil Code Generation with Lift*.
Scottish Programming Language Seminar (SPLS), University of Glasgow, UK

02/2018 Talk: *High Performance Stencil Code Generation with Lift*.
International Symposium on Code Generation and Optimization (CGO), Vienna, Austria

02/2018 Invited Talk: *High Performance Stencil Code Generation with Lift*.
Research Group on Compiler and Architecture Design, University of Edinburgh, UK

03/2017 Invited Talk: *Performance Portable Stencil Code Generation with Lift*.
Research Group on Compiler and Architecture Design, University of Edinburgh, UK

Research Projects

I have been actively contributing to the following research projects.

since 07/2019 **Elevate**, *A Language for Describing Program Transformations*.
Ongoing research, *https://github.com/elevate-lang*
I am leading the development of Elevate, a language enabling the specification of composable program transformations. Using simple combination and traversal operators, Elevate allows to build domain-specific program optimizations as semantics preserving rewrite rules for arbitrary ASTs.

since 09/2018 **Fireiron**, *A Scheduling Language for High-Performance Linear Algebra on GPUs*.
Ongoing research
I am leading the development of Fireiron, a scheduling language, IR and compiler for the generation of high-performance linear algebra kernels on GPUs. Fireiron is a tool for experts, simplifying the development and experimentation with optimizations required for achieving highest performance on modern GPUs using specialized hardware units such as NVIDIA's Tensor Cores.

2016 – 2019 **Lift**, *A Novel Approach to Achieving Performance Portability on Accelerators*.
www.lift-project.org
I am one of the main contributors focusing on implementing stencil computations in Lift. I extended the functional Lift IR and enabled the generation of efficient OpenCL kernels for stencil-based applications. The Lift project is a novel approach to generate high-performance OpenCL kernels from high-level functional programs.

04/2015 **PACXX**, *Programming Accelerators with C++*.
I developed an LLVM analysis pass for the PACXX compiler and ported HPC applications to the PACXX programming model resulting in a publication [11]. PACXX is a unified HPC programming model for programming accelerators (GPUs etc.) using pure C++ by implementing a custom compiler (based on the LLVM framework) and a runtime system.

Reviewer

2020 CGO 2020 artifact evaluation chair
ASPLOS 2020 artifact evaluation committee
2019 CGO 2019 artifact evaluation committee
2018 CGO 2018 artifact evaluation committee
LCTES 2018 artifact evaluation committee

since 2016 I have been active as an external reviewer for the following conferences and journals: *Principles and Practice of Parallel Programming (PPoPP), the International Parallel and Distributed Processing Symposium (IPDPS), the International Journal of Parallel Programming (IJPP)*, the *Journal of Supercomputing*, the journal *Concurrency and Computation: Practice and Experience*, the Journal of Applied Geophysics (APPGEO), the Parallel Computing Technologies (PaCT), the Parallel Computing Conference (ParCo)

Teaching

Winter '20 Teaching assistant for the course: *Parallel Systems*

Winter '20 Supervised a student project: *Developing a CUDA backend for Lift targeting TensorCores*

Winter '19 Teaching assistant for the course: *Operating systems*

Winter '19 Teaching assistant for the course: *Introduction to programming with Java and Haskell*

Summer '18 Course design and Lecturer: *Introduction to programming with C and C++*

Summer '18 Teaching assistant for the course: *Parallel Programming: Multi-Core and GPU*

Winter '17 Teaching assistant for the course: *Operating systems*

Winter '17 Teaching assistant for the course: *Introduction to programming with Java and Racket*

Summer '17 Course design and Lecturer: *Introduction to programming with C and C++*

Summer '17 Supervised a student project: *Automatic program optimization for modern many-core systems*

Winter '16 Teaching assistant for the course: *Operating systems*

Winter '15 Student assistant for the course: *Operating systems*

Summer '15 Student assistant for the course: *Computer architectures*

Winter '14 Student assistant for the course: *Operating systems*

Supervised Undergraduate and Master Students

12/2018 Johannes Lenfers (Master): *Implementing Compiler Auto-Tuning Strategies for Design Space Exploration of Lift Programs*

02/2018 Bastian Köpcke (Master): *Efficient GPU Code Generation for FFT Computations in Lift*

01/2018 Martin Lücke (Master): *Efficient Implementation and Optimization of Geometric Multi-Grid Operations in the Lift Framework*

03/2018 Clemens Hesse-Edenfeld (Undergraduate): *Integrating Performance Models for Stencil Computations in Lift*

03/2018 Alexander Dirk Holthaus (Master): *Development of an Analytical Tool for Visualizing and Optimizing Memory Accesses in GPU Kernels*

03/2018 Maurice Heine (Undergraduate): *Implementation of a Visualization Tool for Lift Programs*