Matthias Springer, Ph.D.

シュプリンガー　マティアス

Profile

I am a principal deep learning compiler compiler engineer at NVIDIA, working on the Tile IR compiler infrastructure. I have been working with and contributing to the open-source MLIR compiler project since 2021. My academic research was concerned with code generation, domain-specific languages (DSLs), programming models and memory access optimizations for SIMD/GPU architectures. I am passionate about compilers, high-performance computing (especially GPUs), programming language design and parallel computing.

Work Experience

Since May 2024 **NVIDIA Switzerland AG**, *Zürich, Switzerland*, Principal Deep Learning Compiler Engineer (IC6).
Working on CUDA Tile IR, an MLIR-based JIT compiler for tile-based GPU programming. Various contributions to MLIR (Multi-Level IR Compiler Framework), in particular the dialect conversion infrastructure (1:N / no-rollback driver, context-aware type conversion) and core interfaces / transformations (region branch, floating-point types etc.).

Mar. 2021 - Apr. 2024 **Google Switzerland GmbH**, *Zürich, Switzerland*, Senior Research Engineer (L5), Google Research.
Core MLIR contributor with >1000 commits. See personal website for RFCs. Contributions to the IREE (Intermediate Representation Execution Environment) Compiler.

- Contributions to core APIs: builders/rewriters, listener infrastructure, greedy pattern rewriter, dialect conversion, operation definition specification (ODS), etc.
- One-Shot Bufferize: Design, implementation and maintenance of an MLIR pass for buffer assignment/management in tensor IR. Supported its adoption in IREE and XLA/TensorFlow (see MLIR Open Design Meeting presentation). Improvements and maintenance of the ownership-based buffer deallocation pass.
- Contributions to the Transform dialect, a scheduling language for compiler transformations: driving rewrite pattern application from transform scripts, core transform ops, integration of existing MLIR transformations, etc.
- Loop Peeling: Implemented a `for` loop peeling and loop body simplification transformation in the SCF dialect.
- Value Bounds Inference: Designed and implemented an infrastructure for computing bounds of integer SSA values.
- Various contributions to the tiling, fusion and vectorization infrastructure in the Linalg and Tensor dialects.
- Various debugging and usability improvements such as runtime op verification, rewrite API verification checks, Graphviz export of IR graphs, pattern application fuzzing, transform dialect interpreter hardening.
- Various improvements and bug fixes in the core MLIR infrastructure (interfaces, vector warp distribution on GPUs, MemRef/SCF/SparseTensor/Tensor/Transform/Vector dialect improvements, etc.)

Oct. 2019 - Feb. 2021 **Google Japan G.K.【グーグル合同会社】**, *Tokyo, Japan*, Software Engineer (L4), ChromeOS Platform.
Worked on ARCVM (Android Runtime for Chrome), a virtual machine that runs Android. Identified and fixed bugs in the ChromeOS/Android camera stacks and worked on overall performance optimizations. Gained experience in virtualization, kernel/OS development and large-scale software development (mainly C++, some Java and Golang).

2017, 2018 **Google LLC**, *United States of America*, Software Engineering Intern (Summer Internship).
2016, 2014 (4 × 3 months)

- [2018; Mountain View, CA; Host: DeLesley Hutchins, Ph.D.] Worked on an auto-batching system for tree-structured RNNs. Implemented linear algebra operations and TPU-specific optimizations with TensorFlow XLA.
- [2017; Mountain View, CA; Host: DeLesley Hutchins, Ph.D.] Worked on LLGTM (Low-level Library for Gradients, Tensors, and Matrices), a deep learning C++ framework for dynamic computation graphs. Implemented Eigen/CUDA kernels and TensorFlow kernel adapters.
- [2016; Seattle, WA; Host: Vijay Menon, Ph.D.] Part of the Dart programming language team. Worked on a Dart-to-Java compiler, focusing on performance optimizations and language interoperability.
- [2014; Boulder, CO; Host: Craig Wright] Worked on a business event process engine for an internal Google payments system, using Megastore, F1/Spanner, Java and Guice.

April 2012 - June 2017 **Tokyo Institute of Technology / Hasso Plattner Institute**, Teaching Assistant.
Teaching assistant for exchange students and courses: Information Literacy (Prof. Morozov, Ph.D.), Mathematics II (Dr. habil. Börner), Software Architecture (Prof. Dr. Hirschfeld), Software Engineering I (Prof. Dr. Hirschfeld)

Aug. 2012 - Oct. 2012 **Senacor Technologies AG**, *Munich, Germany*, Software Engineering Intern.
Developed SOA components and tests (with Java EE, Spring Framework), supporting a merger of two banks.

Aug. 2011 - Oct. 2011 **TNG Technology Consulting GmbH**, *Munich, Germany*, Software Engineering Intern.
Developed plugins for the team's continuous integration software stack (Atlassian JIRA/Confluence, Hudson/Jenkins).

Education

Oct. 2015 - Sept. 2019 **Tokyo Institute of Technology 【東京工業大学】**, *Meguro-ku, Tokyo, Japan*,
Programming Research Group, Department of Mathematical and Computing Sciences
Doctor of Philosophy (Ph.D.). Academic advisor: Prof. Dr. Hidehiko Masuhara.
- *Doctoral thesis:* Memory-Efficient Object-Oriented Programming on GPUs
- *Research areas:* Compilers, Program Optimization, GPGPU, Modularity, Context-oriented Programming
- *Relevant coursework:* Programming Language Design, Practical Parallel Computing, Distributed Computing

Sept. 2014 - Sept. 2015 **Hasso Plattner Institute, University of Potsdam**, *Potsdam, Brandenburg, Germany*,
Master of Science, IT Systems Engineering, overall grade: 1.0 (A+).
- *Master's project:* Spur to go faster: Low-level Functionality in a High-level Language
- *Master's thesis:* Nested Class Modularity in Squeak/Smalltalk
  Project and thesis supervised by Prof. Dr. Robert Hirschfeld, Tim Felgentreff, Tobias Pape
- *Relevant coursework:* VMs and Execution Environments, Context-oriented Programming

Sept. 2013 - June 2014 **University of California San Diego**, *La Jolla, CA, United States of America*,
Visiting student, Department of Computer Science and Engineering, GPA: 4.0.
- Full tuition and living expenses covered by *UC Education Abroad Program* and *DAAD Scholarship*
- *Relevant coursework:* Advanced Compilers (CSE 131/231), Programming Languages (CSE 130/230), Adv. Algorithms (CSE 190/202/203A), Parallel Computation (CSE 260), Database Analytics (CSE 190)

Aug. 2010 - July 2013 **Hasso Plattner Institute, University of Potsdam**, *Potsdam, Brandenburg, Germany*,
Bachlor of Science, IT Systems Engineering, overall grade: 1.0 (A+), rank 1/74.
- *Bachelor's project:* Evolving Applications: Object-migration with Ruby and GemStone
- *Bachelor's thesis:* Inter-language Collaboration in an Object-oriented Virtual Machine
  Project and thesis supervised by Prof. Dr. Robert Hirschfeld, Tim Felgentreff, Tobias Pape
- *Relevant coursework:* Software Architecture, Software Engineering I, Advanced Modularity, Database Systems I/II, Internet and WWW Technologies, Designing Interactive Systems (HCI)

Publications on Google Scholar: `https://scholar.google.com/citations?user=EvHvYtMAAAAJ`

Research Experience

Oct. 2014 - Sept. 2019 **Programming Languages/Compilers**, *Tokyo Institute of Technology / Hasso Plattner Institute*,
Research with Prof. Hidehiko Masuhara and Prof. Robert Hirschfeld.
- DynaSOAr: A lock-free CUDA memory allocator with SOA layout, based on hierarchical bitmaps.
- CompactGpu: Fully parallel GPU memory defragmentation for better vector/cache performance.
- Ikra-Cpp: A C++/CUDA DSL for object-oriented programming with Structure-of-Arrays data layout.
- Ikra-Ruby: A GPGPU library for Ruby (Ruby-to-CUDA compiler) with parallel array operations.
- Matriona: A module system for Squeak/Smalltalk based on class nesting/parameterization.
- ContextAmber: A COP (Context-oriented Programming) library for Amber Smalltalk.
- *Other Projects*: Minor contributions to the RSqueak VM and to the Truffle-based JRuby implementation.

March 2014 - Nov. 2014 **Database Research**, *University of California San Diego (with Prof. Yannis Papakonstantinou)*.
Evaluated algorithms and data structures for queries on compressed data in relational/graph database systems.

May 2011 - Apr. 2013 **Internet Technologies**, *Hasso Plattner Institute, Internet Technologies and Systems Group*.
- SOA Security Lab: A browser-based simulation system for modelling and executing web service scenarios.
- Tele-Lab: A platform for teaching and simulating network security scenarios.

Achievements and Prizes

Feb. 2021 **Seiichi Tejima Doctoral Dissertation Award 【手島精一記念研究賞（博士論文賞）】**.
Honoring authors of outstanding dissertations at Tokyo Institute of Technology. Category: *Information Science*.

July 2019, Nov. 2018, Feb. 2018 **ACM Student Research Competition**, *Association for Computing Machinery*.
- [PLDI 2019 (Phoenix, AZ)] 1st place, graduate category for CompactGpu
- [SPLASH 2018 (Boston, MA)] 1st place, graduate category for SoaAlloc (now called DynaSOAr)
- [CGO 2018 (Vienna, Austria)] 3rd place, graduate category for Ikra-Cpp

April 2018 - Sept. 2019 **Research Fellowship for Young Scientists (JSPS DC2) 【日本学術振興会特別研究員DC2】**.
Japanese government fellowship, covering living expenses and research expenses (18J14726).

Oct. 2015 - Mar. 2018 **Monbukagakusho (MEXT) Scholarship** 【文部科学省奨学金】.
Japanese government scholarship for doctoral students, covering tuition and living expenses.

Oct. 2014 - Sept. 2015 **Hasso Plattner Scholarship**.
One-year scholarship awarded to the best Bachelor graduates of each year, covering living expenses.

Sept. 2013 - May 2014 **German Academic Exchange Service Scholarship (DAAD Jahresstipendium)**.
German government scholarship, covering tuition and living expenses at a North American university.

2007 - 2010 **German Federal Competition in Computer Science (Bundeswettbewerb Informatik)**.
Participated three years in a row. 30/around 1100 participants are invited to the final round.
- [2009/2010] 1st prize in first two rounds, invited to final round *(University of Freiburg)*
- [2008/2009] 2nd prize in first two rounds
- [2007/2008] 1st prize in first two rounds, invited to final round *(Max Planck Institute for CS, Saarbrücken)*

2010/2011 **informatiCup 2011**, *organized by the Gesellschaft für Informatik*, *Bonn, Germany*.
Participated in the first round and in the final round (6/38 teams invited). Solved an optimization problem using Simulated Annealing, Tabu Search and greedy algorithms.

(various) **Academic Honors**, Graduation with distinction (Bachelor's, Master's), 2x Provost's Honors at UCSD.

(various) **Travel Grants**, ECOOP Summer School 2016, SIGPLAN-PLMW (POPL 2017), SIGPLAN-PAC (PLDI 2017), Google MUC Compiler and PL Summit (2017, 2018), ACM SRC (CGO 2018, SPLASH 2018, PLDI 2019).

Adademic Service / Open Source Involvement

Program Committee, CROW 2016, COP (2017, 2018), LLVM Dev (2024, 2025, 2026), EuroLLVM 2025.

Reviewer, APLAS 2016, ARRAY 2017, ECOOP 2020, LASSY 2017, MPLR 2019, SIMPAT.

Student Volunteer, ECOOP (2015, 2016, 2017, 2019), PLDI (2016, 2017, 2018), SPLASH 2018.

Organizing Committee, Student volunteer co-chair at PLDI 2019.

LLVM Project, MLIR Area Team (since Feb. 2025), MLIR Tensor Compiler Working Group, Maintainer of the Bufferization Dialect and the Dialect Conversion Infrastructure, Google Summer of Code 2026 Project Mentor (with Javed Absar), Teacher at the MLIR Schools (Winter 2026, Summer 2025, Summer 2026).

Other Projects

Sept. 2014 - July 2015 **ME310 Global Team-based Product Innovation & Engineering**,
Collaboration between Hasso Plattner Institute and Stanford University, including several exchange visits.
Worked on a design challenge by Audi USA with Stanford mechanical engineering students. Prototyped concepts for communication between pedestrians, passengers and autonomous cars, using design thinking.

June 2013 - Sept. 2013 **Athens for Amber Smalltalk**, *Google Summer of Code 2013 Project (Student Participant), European Smalltalk User Group (ESUG), Mentors: Nicolas Petton, Igor Stasenko*.
Implemented the Athens vector graphics library in Amber Smalltalk, a Smalltalk system running in the web browser, using HTML5 Canvas. Developed a Morphic-like framework for building GUIs with Athens.

Dec. 2012 - June 2013 **MagLev Database Explorer**, *Part of Bachelor's project at HPI, Software Architecture Group*.
Developed an IDE running in a web browser for exploring Ruby/Smalltalk objects persisted in a GemStone/S 64 image, writing Ruby/Smalltalk code and debugging Rails/Sinatra applications interactively.

</> Skills

Programming: C, C++, CUDA, Dart, Go, HTML, Java, LLVM, MLIR, NumPy, Python, Ruby, Smalltalk (Amber, GemStone, Pharo, Squeak), SQL, Visual Basic.

Software Eng.: Continuous Integration, Design Patterns, Git, Mercurial, Scrum, Subversion, TDD, UML.

Personal Information

Languages: German (native speaker), English (CEFR C2, TOEFL iBT score 118/120), Japanese (JLPT N3 147/180).

Peer-reviewed Publications and Talks

Talks at the LLVM Developers' Meetings.

- M. Springer. *Value Forwarding in MLIR: Control-Flow Intf. and Canonicalizations You Get for Free*. LLVM Dev 2026 (Santa Clara, CA).
- M. Springer. *Storing Arrays in MLIR: A Field Guide to Elements Attrs. & Custom Element Types*. LLVM Dev 2026 (Santa Clara, CA).
- M. Springer. *Floating-Point Types in MLIR: Infrastructure, New Types and Dialect Design*. EuroLLVM 2026 (Dublin, Ireland).
- M. Springer, L. Chelini. *CUDA Tile IR: Lessons from a Tile-Centric CUDA Dialect for MLIR*. EuroLLVM 2026, MLIR Workshop.
- M. Springer, M. Böck. *A Faster and Simpler Dialect Conversion Driver without Pattern Rollback*. LLVM Dev 2025 (Santa Clara, CA).
- M. Springer. *Sanitizing MLIR Programs with Runtime Operation Verification*. AsiaLLVM 2025 (Tokyo, Japan).
- M. Springer. *A Listener-based Approach to Discardable Attribute Conversion/Propagation*. EuroLLVM 2025, MLIR Workshop.
- M. Springer. *1:N Dialect Conversion in MLIR*. EuroLLVM 2025 (Berlin, Germany).
- M. Springer. *The State of Pattern-Based IR Rewriting in MLIR*. LLVM Dev 2024 (Santa Clara, CA).
- M. Springer. *Computing Bounds of SSA Values in MLIR*. EuroLLVM 2024 (Vienna, Austria).
- M. Springer, M. Erhart. *MLIR Bufferization: From Tensors to MemRefs*. LLVM Dev 2023 (Santa Clara, CA).

Conference Publications.

- J. Chenxin, H. Masuhara, M. Springer and Y. Cong. *Sanajeh: A DSL for GPGPU Programming With Python Objects*. Annual Conference of the Japan Society for Software Science and Technology (JSSST 2020).
- M. Springer, H. Masuhara. *Massively Parallel GPU Memory Compaction.* International Symposium on Memory Management (ISMM 2019), Phoenix, AZ. 🏆 Best Student Presentation Award.
- M. Springer, H. Masuhara. *DynaSOAr: A Parallel Memory Allocator for Object-oriented Programming on GPUs with Efficient Memory Access.* European Conference on Object-oriented Programming (ECOOP 2019), London, United Kingdom.
- M. Springer, F. Niephaus, R. Hirschfeld, H. Masuhara. *Matriona: Class Nesting with Parameterization in Squeak/Smalltalk.* International Conference on Modularity (Modularity 2016), Málaga, Spain.

Journal Publications.

- M. Springer, H. Masuhara, R. Hirschfeld. *A Layer-based Approach to Hierarchical Dynamically-scoped Open Classes*. IPSJ (Information Processing Society of Japan) Transactions on Programming, Vol. 10, No. 2, March 2017.
- C. Lin, B. Mandel, Y. Papakonstantinou, M. Springer. *Fast In-Memory SQL Analytics on Typed Graphs.* Proceedings of the VLDB (Very Large Data Bases) Endowment, Vol. 10, Iss. 3, Nov. 2016.

Workshop Publications.

- N. Vasilache, O. Zinenko, A. J. C. Bik, M. Ravishankar, T. Raoux, A. Belyaev, M. Springer, T. Gysi, D. Caballero, S. Herhut, S. Laurenzo, A. Cohen. *Structured Operations: Modular Design of Code Generators for Tensor Compilers.* Workshop on Languages and Compilers for Parallel Computing (LCPC 2022).
- M. Springer, Y. Sun, H. Masuhara. *Inner Array Inlining for Structure of Arrays Layout.* Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY 2018), Philadelphia, PA.
- M. Springer, H. Masuhara. *Ikra-Cpp: A C++/CUDA DSL for Object-oriented Programming with Structure-of-Arrays Data Layout.* Workshop on Programming Models for SIMD/Vector Processing (WPMVP 2018), Vienna, Austria.
- M. Springer, P. Wauligmann, H. Masuhara. *Modular Array-based GPU Computing in a Dynamically-typed Language.* Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY 2017), Barcelona, Spain.
- M. Springer, A. Krieger, S. Manilov, H. Masuhara. *dart2java: Running Dart in Java-based Environments.* Workshop on Implementation, Compilation, Optimization of Object-Oriented Languages, Programs and Systems (ICOOOLPS 2017), Barcelona, Spain.
- M. Springer, H. Masuhara, R. Hirschfeld. *Classes as Layers: Rewriting Design Patterns with COP.* International Workshop on Context-oriented Programming (COP 2016), Rome, Italy.
- M. Springer, H. Masuhara. *Object Support in an Array-based GPGPU Extension for Ruby.* Workshop on Libraries, Languages, and Compilers for Array Programming (ARRAY 2016), Santa Barbara, CA.
- M. Springer, H. Masuhara, R. Hirschfeld. *Hierarchical Layer-based Class Extensions in Squeak/Smalltalk.* Workshop on Live Adaptation of Software Systems (LASSY 2016), Málaga, Spain.
- M. Springer, J. Lincke, R. Hirschfeld. *Efficient Layered Method Execution in ContextAmber.* International Workshop on Context-oriented Programming (COP 2015), Prague, Czech Republic.
- F. Niephaus, M. Springer, T. Felgentreff, T. Pape, R. Hirschfeld. *Call-target-specific Method Arguments.* Workshop on Implementation, Compilation, Optimization of Object-Oriented Languages, Programs and Systems (ICOOOLPS 2015), Prague, Czech Republic.

Publications on Google Scholar: `https://scholar.google.com/citations?user=EvHvYtMAAAAJ`

Date: August 6, 2026