MALTE SCHWARZKOPF

MIT CSAIL
32 Vassar Street, 32-G980
Cambridge, MA 02139

people.csail.mit.edu/malte
malte@csail.mit.edu
+1 (781) 484-7008

EDUCATION

University of Cambridge, Cambridge, UK *2009–2016*
Ph.D. in Computer Science.
Dissertation title: *Operating system support for Warehouse-Scale Computing*.
Advisor: Steven Hand.

University of Cambridge, St John's College, Cambridge, UK *2006–2009*
B.A. in Computer Science (superseded by M.A. conferred on May 10, 2013).
First Class Honours (ranked 2nd in year).

RESEARCH INTERESTS

Distributed systems, datacenter systems, operating systems, networking, and security.

RESEARCH PROJECTS

Noria [1] makes it easy to develop high-performance web applications. Noria precomputes the results of SQL queries using streaming data-flow so that reads are fast, and adapts the live data-flow to changing queries. A new data-flow model, *partially-stateful data-flow*, allows Noria to keep only partial operator state and query results, which bounds the data-flow's memory footprint and allows downtime-free data-flow changes when application queries change.

Firmament [4] is a datacenter cluster scheduler that achieves high throughput, low decision latency, and high-quality placement decisions on large clusters. Its centralized design exposes the full cluster state to the scheduling policy (for high-quality placement decisions), and Firmament relies on efficient constraint solvers to lower the amortized cost of each decision (for high throughput), achieving sub-second latency in the common case.

QJUMP [5] demonstrates a readily-deployable approach to achieving bounded-latency communication and reduced network interference in datacenter networks. QJUMP achieves low tail latencies for latency-critical applications (*e.g.*, memcached) that share the network with high-throughput applications (*e.g.*, MapReduce) using only rate limiting and packet prioritization mechanisms available in current switches and network stacks.

Musketeer [6] increases portability and efficiency of "big data" processing pipelines by *decoupling* the developer's workflow expression from the systems used to execute it. For workflows expressed using a declarative frontend (*e.g.*, SQL, vertex-centric BSP), Musketeer chooses the best combination of systems for execution and generates efficient code for them, thus supporting migration to new and more efficient systems at no cost to the workflow developer.

Omega [7] efficiently shares a datacenter's resources between multiple cluster schedulers. Using optimistically concurrent changes to *shared cluster state*, Omega allows independent, purpose-built schedulers to place workloads on the same underlying cluster without explicit mediation. The shared state approach exposes full cluster state to all schedulers, which simplifies their implementation, scales well, and supports previously difficult-to-implement scheduling policies.

CIEL [8] efficiently executes iterative algorithms with data-dependent control flow over "big data". It achieves this by expressing data-parallel computations as *dynamic* data-flow graphs, in which each parallel task can extend the computation without returning control to a driver program.

AWARDS AND HONORS

Best Paper Award at NSDI for QJUMP [5]. *May 2015*
Best Student Paper Award at EuroSys for Omega [7]. *April 2013*
Full Ph.D. scholarship from the St John's College Supplementary Emoluments Fund. *2009–2014*
Scholar of St John's College, Cambridge. *Since 2008*
St John's College Hockin, Hughes, and Wright prizes for academic distinction. *July 2009*
St John's College prize for academic distinction in Computer Science. *July 2008*
Scholar of the German Academic Scholarship Foundation ("Studienstiftung des deutschen Volkes"). *2007–2009*

PUBLICATIONS

Refereed conference papers:

[1] Jon Gjengset, Malte Schwarzkopf, Jonathan Behrens, Lara Timbó Araújo, Martin Ek, Eddie Kohler, M. Frans Kaashoek, and Robert Morris.
"Noria: dynamic, partially-stateful data-flow for high-performance web applications".
In: *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
Carlsbad, California, USA, Oct. 2018.

[2] Shoumik Palkar, James Thomas, Deepak Narayanan, Pratiksha Thaker, Rahul Palamuttam, Parimajan Negi, Anil Shanbhag, Malte Schwarzkopf, Holger Pirk, Saman Amarasinghe, Samuel Madden, and Matei Zaharia.
"Evaluating End-to-End Optimization for Data Analytics Applications in Weld".
In: *Proceedings of the VLDB Endowment* 11.9 (May 2018).

[3] Shoumik Palkar, James J. Thomas, Anil Shanbhag, Deepak Narayanan, Holger Pirk, Malte Schwarzkopf, Saman Amarasinghe, and Matei Zaharia.
"Weld: A Common Runtime for High Performance Data Analysis".
In: *Proceedings of the 8th Biennal Conference on Innovative Data Systems Research (CIDR)*.
Chaminade, California, USA, Jan. 2017.

[4] Ionel Gog, Malte Schwarzkopf, Adam Gleave, Robert N. M. Watson, and Steven Hand.
"Firmament: fast, centralized cluster scheduling at scale".
In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.
Savannah, Georgia, USA, Nov. 2016.

[5] Matthew P. Grosvenor, Malte Schwarzkopf, Ionel Gog, Robert N. M. Watson, Andrew W. Moore, Steven Hand, and Jon Crowcroft.
"Queues don't matter when you can JUMP them!"
In: *Proceedings of the 12th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*.
Oakland, California, USA, May 2015.
Best paper award.

[6] Ionel Gog, Malte Schwarzkopf, Natacha Crooks, Matthew P. Grosvenor, Allen Clement, and Steven Hand.
"Musketeer: all for one, one for all in data processing systems".
In: *Proceedings of the 10th ACM European Conference on Computer Systems (EuroSys)*.
Bordeaux, France, Apr. 2015.

[7] Malte Schwarzkopf, Andy Konwinski, Michael Abd-El-Malek, and John Wilkes.
"Omega: flexible, scalable schedulers for large compute clusters".
In: *Proceedings of the 8th ACM European Conference on Computer Systems (EuroSys)*.
Prague, Czech Republic, Apr. 2013.
Best student paper award.

[8] Derek G. Murray, Malte Schwarzkopf, Christopher Smowton, Steven Smith, Anil Madhavapeddy, and Steven Hand.
"CIEL: a universal execution engine for distributed data-flow computing".
In: *Proceedings of the 8th USENIX Symposium on Networked System Design and Implementation (NSDI)*.
Boston, Massachusetts, USA, Mar. 2011.

Refereed workshop papers:

[9] Ionel Gog, Jana Giceva, Malte Schwarzkopf, Kapil Vaswani, Dimitrios Vytiniotis, Ganesan Ramalingam, Manuel Costa, Derek G. Murray, Steven Hand, and Michael Isard.
"Broom: sweeping out Garbage Collection from Big Data systems".
In: *Proceedings of the 15th USENIX/SIGOPS Workshop on Hot Topics in Operating Systems (HotOS)*.
Kartause Ittingen, Switzerland, May 2015.

[10] Malte Schwarzkopf, Matthew P. Grosvenor, and Steven Hand.
"New Wine in Old Skins: The Case for Distributed Operating Systems in the Data Center".
In: *Proceedings of the 4th Asia-Pacific Systems Workshop (APSYS)*.
Singapore, July 2013.

[11] Malte Schwarzkopf, Derek G. Murray, and Steven Hand.
"The Seven Deadly Sins of Cloud Computing Research".
In: *Proceedings of the 4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*.
Boston, Massachusetts, USA, June 2012.

Journals and magazines:

[12] Malte Schwarzkopf.
"Cluster Scheduling for Data Centers".
In: *ACM Queue* 15.5 (Oct. 2017). (Invited article.)
Also appears in *Communications of the ACM* 61.5 (May 2018), pages 50–53.

[13] Mihir Nanavati, Malte Schwarzkopf, Jake Wires, and Andrew Warfield.
"Non-volatile Storage".
In: *ACM Queue* 13.9 (Nov. 2015). (Invited article.)
Also appears in *Communications of the ACM* 59.1 (Dec. 2015), pages 56–63.

[14] Matthew P. Grosvenor, Malte Schwarzkopf, Ionel Gog, and Andrew W. Moore.
"Jump the queue to lower latency".
In: *USENIX ;login: magazine* 40.2 (Apr. 2015). (Invited article.)

[15] Heidi Howard, Malte Schwarzkopf, Anil Madhavapeddy, and Jon Crowcroft.
"Raft Refloated: Do We Have Consensus?"
In: *SIGOPS Operating Systems Review* 49.1 (Jan. 2015).

In submission:

[16] Frank McSherry, Andrea Lattuada, and Malte Schwarzkopf.
"K-Pg: Shared State in Differential Dataflows".
Oct. 2018.

[17] Nikolaj Volgushev, Malte Schwarzkopf, Andrei Lapets, Mayank Varia, and Azer Bestavros.
"Conclave: Secure Multi-Party Computation on Big Data".
Oct. 2018.

[18] Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh.
"Variance Reduction for Reinforcement Learning in Input-Driven Environments".
Sept. 2018.

Preprint:

[19] Hongzi Mao, Malte Schwarzkopf, Shaileshh Bojja Venkatakrishnan, Zili Meng, and Mohammad Alizadeh.
Learning Scheduling Algorithms for Data Processing Clusters.
arXiv: 1810.01963 [cs.LG].

Work Experience

Postdoctoral Associate, Massachusetts Institute of Technology, Cambridge, MA, USA. *Since February 2016*

- Parallel and Distributed Operating Systems (PDOS) group, Computer Science and AI Laboratory (CSAIL).

Research Assistant, University of Cambridge Computer Laboratory, Cambridge, UK. *January 2014–January 2015*

Acting Director of Studies in Computer Science.
St John's College, Cambridge, UK (sabbatical replacement) *October 2013–August 2014*
Peterhouse, Cambridge, UK (temporary replacement) *April 2014–August 2014*

- Organized small-group teaching and mentored 18 (St John's) and 3 (Peterhouse) undergraduate students.

Software Engineering Intern (PhD).
Google, Inc., Mountain View, CA, USA *June–November 2011*
Google UK Ltd., London, UK *February–May 2012*

- Researched optimistically concurrent cluster scheduler design [7].

Co-founder, Adjutem Software Solutions, Cambridge, UK. *2008–2015*

- Software consultancy developing bespoke student room allocation software for Cambridge colleges.

Engineering Intern, Broadcom Europe Ltd., Cambridge, UK.

- Developed software simulation framework for next-generation video processor's OpenGL shader hardware. *June–September 2008*
- Wrote a profiler for the 3D subsystem of the VideoCore III processor, and the VideoCore assembly emitter component of an OpenGL shader compiler. *June–September 2007*

TEACHING EXPERIENCE

Co-instructor, Distributed Systems Engineering (6.824), MIT EECS. *February–May 2018*

- Co-taught the class with Robert Morris.

Guest lecturer, Distributed Systems Engineering (6.824), MIT EECS. *April 2016*

- Gave guest lecture on Google's Borg cluster management system.

Guest lecturer, Concurrent and Distributed Systems, CST Part IB, University of Cambridge. *February–March 2014*

- Gave three lectures focusing on recent large-scale distributed systems.

Small-group supervisor, University of Cambridge (various colleges).

- Operating Systems *2010–2013*
- Computer Fundamentals *2010, 2012*
- Software Engineering *2012*
- Digital Communications II *2009*
- Advanced Systems Topics *2009*
- Object-Oriented Programming *2009–2012*
- Programming in Java *2009–2012*
- Programming in C and C++ *2009–2011*
- Further Java *2009–2010*

Research mentor, Massachusetts Institute of Technology:

- for graduate students Jon Gjengset and Jonathan Behrens. *Since 2016*
  Jon (co-advised with Robert Morris) and Jonathan (with Frans Kaashoek) work with me on the Noria data-flow system [1]. With submissions based on this work, Jon won the SOSP 2017 Student Research Competition and the runner-up prize in the graduate category of the ACM-wide Student Research Competition in 2018.
- for undergraduate student Samyukta Yagati. *Since September 2018*
  Samyu is adding support for differentially-private aggregations to a "multiverse" database built atop Noria.
- for undergraduate student Joshua Segaran. *Since Februrary 2018*
  Josh is exploring ideas for simplified personal data ownership management compliant with GDPR-style "right to erasure" using a storage backend based on data-flow.
- for undergraduate and masters student Lara Timbó Araújo. *October 2016–January 2018*
  Lara first completed an undergraduate advanced project (6.UAP) on automated query reuse to improve data-flow efficiency in Noria. She then developed the idea of a "multiverse" database for isolating users in a web application backend for her M.Eng. thesis, which I co-advised with Frans Kaashoek. A multiverse database restricts and modifies the data visible in each user's "universe" according to global, declarative security policies. Lara's implementation shows that this can be achieved with moderate overheads using a data-flow system.

Bachelor dissertation project supervisor, University of Cambridge:

- for Joshua Bambrick (with Ionel Gog). *2015–2016*
  "Vilfredo: Optimising cluster resource allocations, one Pareto improvement at a time".
- for Adam Gleave (with Ionel Gog). *2014–2015*
  "Hapi: Fast and scalable cluster scheduling using flow networks".
- for Andrew Scull. *2014–2015*
  "Hephaestus: A Rust runtime for a distributed operating system".
- for James Chicken. *2013–2014*
  "Hydra: Automatic Parallelism Using LLVM".
- for Matthew Huxtable. *2013–2014*
  "Unbuckle: A high-performance key-value store". *Best dissertation award.*

- for Leonard Markert (with Ionel Gog). *2013–2014*
  "Benson: A structured voting extension for an implementation of the Raft consensus algorithm".
- for Bogdan-Cristian Tătăroiu (with Ionel Gog). *2013–2014*
  "The BDFS Distributed File System".
- for Christopher Wheelhouse (with Matthew P. Grosvenor). *2013–2014*
  "Network Trace Visualisation at Scale".
- for Forbes Lindesay. *2012–2013*
  "Wrench: A Distributed Key-Value Store with Strong Consistency".
- for Antanas Ursŭlis. *2012–2013*
  "Cluster-in-a-box: task parallel computing on many-core machines".
- for James Bown. *2010–2011*
  "Osiris: Secure Social Backup". *Specially commended by the examiners*.
- for Valentin Dalibard. *2010–2011*
  "Implementing Homomorphic Encryption". *Specially commended by the examiners*.
- for Sebastian Hollington (with Derek Murray). *2010–2011*
  "Cirrus: Distributing Application Threads on the Cloud". *Specially commended by the examiners*.

Professional Service

Conference program committees: SYSTOR 2019, ICDCS 2019 (Distributed Operating Systems & Middleware Track), ICDCS 2018 (Distributed Operating Systems & Middleware Track), WWW 2017 (Systems & Infrastructure Track), EuroSys 2016 (light).

Workshop program committees: EdgeSys 2019, SFMA 2018, EdgeSys 2018.

External reviewer: SOSP 2017, OSDI 2018.

Other committees: PLDI 2015 Artifact Evaluation Committee; TinyToCS (Tiny Transactions on Computer Science) Program Committee, vol. 2 (2013), vol. 3 (2015), and vol. 4 (2016); EuroSys 2013 shadow PC.

Journal reviewer: ACM Transactions on Computer Systems (2017, 2018), IEEE Transactions on Parallel and Distributed Systems (2018), IEEE Transactions on Computers (2017), SIGCOMM Computer Communication Review (CCR) (2016).

Other activities

Convenor of the Operating Systems reading group at the University of Cambridge Computer Laboratory. *2012–2014*

Vice President of the St John's College Junior Combination Room Committee. *2009–2010*

Chairman of the Student-Run Computing Facility (SRCF) in Cambridge. *2009–2010*

Computing Officer on the St John's College Junior Combination Room Committee. *2007–2009*

President of the Wilkes Society (then "St John's College Computer Science Society"). *2008–2009*

Treasurer and system administrator for the Student-Run Computing Facility (SRCF) in Cambridge. *2008–2009*

Open source software

Nearly all the research systems and tools I developed are available as open-source software.

Noria, a high-performance distributed data-flow system.
https://github.com/mit-pdos/noria

noria-mysql, a MySQL adapter for Noria that exposes a MySQL binary protocol interface to applications.
https://github.com/mit-pdos/noria-mysql

nom-sql, a SQL parser written in Rust.
https://github.com/ms705/nom-sql

taster, a continuous integration tool for performance regression tests on Rust projects.
https://github.com/ms705/taster

Firmament, a cluster manager and scheduler built around high-performance constraint solvers.
https://github.com/camsas/firmament

Poseidon, a Firmament scheduler plugin for Kubernetes (now maintained as a Kubernetes SIG-scheduling project).
https://github.com/kubernetes-sigs/poseidon

Musketeer, a workflow manager for parallel data processing compatible with many backend engines.
https://github.com/camsas/Musketeer

REFERENCES

Prof. M. Frans Kaashoek
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139
USA
kaashoek@mit.edu

Prof. Robert Morris
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139
USA
rtm@csail.mit.edu

Prof. Eddie Kohler
Harvard University
Maxwell Dworkin, Room 327
33 Oxford Street
Cambridge, MA 02138
USA
kohler@seas.harvard.edu

Dr Steven Hand
Senior Staff Engineer
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA
steven.hand@gmail.com

Dr Rebecca Isaacs
Software Engineer
Twitter, Inc.
1355 Market St #900
San Francisco, CA 94103
USA
risaacs@twitter.com