Qiong Cai

Senior Researcher
Hewlett Packard Labs

445 Oak Grove Ave, Apt 12
Menlo Park
CA 94025, US
+1 650-898-9598
qiong.cai@gmail.com

Personal Information

Date of Birth	15 December,1976
Place of Birth	Shanghai, China
Nationality	Australia
US Visa Status	E3 work visa for Australian
Marital Status	Married, 1 daughter and 1 son

Languages

Mandarin	Native
English	Fluent

Expertise

Memory and storage systems for client and server, especially Intel's Iris Pro chips based on eDRAM, Intel's two-level memory systems based 3D XPoint and HP's The Machine architecture.

Emerging non-volatile memory (NVM) technologies, especially 3D XPoint from Intel and STTRAM

Performance analysis of computer systems (single nodes and distributed systems, x86 or ARM systems, OpenMP/MPI/Spark framework, HPC and graph applications).

The design and implementation of CPU, video accelerator and memory-side accelerator.

Compiler optimizations, especially the middle-end and backend optimizations and Just-In-Time compiler optimizations.

Work Experience

2015–present **Senior Researcher**, *Hewlett Packard Labs*, Palo Alto.
I am working on The Machine (TM) architecture, a rack-scale system with 320TB globally addressable memory pool.

Performance characterization: Performance characterization of HPC, Big Data and Spark-based workloads for TM.

Rack-scale architecture improvement — Architecture recommendation to improve latency and bandwidth of memory fabric for TM.

Memory-side accelerator — Design and implement power-efficient memory side accelerators for a rack-scale system.

2010–2014 — **Manager/Technical Lead**, *Intel Labs*, Barcelona.
I led a team of 6 researchers, working on the client and server platforms based on emerging memory technologies.

Crystal Ridge — Path finding with product groups to define the NVM-based server systems.

3D XPoint — Performance owner of 3D XPoint-based memory controller and path finding for 2nd generation of 3D XPoint.

STTRAM — Path finding with Intel's component research group to define STTRAM-based computer systems.

eDRAM, WideIO2, HBM — Design and implement a set of large cache managemet techniques for high bandwidth memory technologies.

2005–2009 — **Senior Research Scientist**, *Intel Labs*, Barcelona.
I did research on power-efficient microarchitectures for multi-core CPUs and video accelerators.

Video Programmable Accelerator — I was chief architect of a low-power programmable accelerator. The accelerator was one of candidates for a video processor being incorporated into Merrifield SoC.

Power-efficient Processors — I was in a team designing power and thermal-awared microarchitecture techniques for Intel multicore systems. I proposed and implemented several CPU microarchitecture techniques such as thread migration and critical thread characterization. I also proposed a software-hardware co-designed technique to steer instructions in a clustered processor. The whole team designing clustered processors won a Spanish technology award in 2008.

2002–2005 — **Tutor (part time)**, *University of New South Wales*, Sydney.
I gave tutorials of two courses (algorithm and compiler) to junior undergraduate students in UNSW.

Education

2002–2006 — **PhD**, *University of New South Wales*, Sydney, Australia. My research focus is on the compiler optimizations. The title of my PhD thesis is profile-guided redundancy elimination.

2001–2002 — **Bachelor of Science in Computer Science(1st class Honours)**, *University of New South Wales*, Sydney, Australia. The title of Honours thesis is: Speculative partial redundancy elimination in dynamic compilation.

1997–2000 — **Bachelor of Mathematics**, *University of Wollongong*, Wollongong, Australia.

1997–2000 — **Bachelor of Computer Science**, *University of Wollongong*, Wollongong, Australia.

Computing Skills

Languages — C/C++, Java, JavaScript, Perl, python, awk, R

Simulators — Years' experience in writing modules for different Intel in-house x86 and Graphics performance and functional simulators

Compilers — Years' experience in implementing compiler optimizations for different compilers such as GCC for x86, ORC for Itanium, Intel's production compiler (ICC) for x86 and LLVM for x86

Microcode — Years' experience in writing modules by using Intel's internal microcode

Virtual Machines — (1) More than 6 months experience in writing code generator for Transmeta's Code Morphing Software (CMS). (2) Years' experience in writing optimization modules for ORP (Intel's open Java virtual machine). (3) Years' experience in writing modules for Intel's internal hypervisor.

Issued and Filed Patents

2014 Dyer Rolan, Nevin Hyuseinova, Blas Cuesta and Qiong Cai, "Method, Apparatus and System to Cache Sets of Tags of an off-die Cache Memory", filed with patent application number 14/227,940.

2013 Qiong Cai, Dyer Rolan, Blas Cuesta, Ferad Zyulkyarov, Serkan Ozdemir and Marios Nicolaides, "Memory Imbalance Prediction Based Cache Management", filed with patent number 13/793,674.

Blas Cuesta, Qiong Cai, Nevin Hyuseinova, Serkan Ozdemir, Marios Nicolaides and Ferad Zyulkyarov, "Sectored Cache with Hybrid Line Granularity", filed with patent number 13/729,523.

Ferad Zyulkyarov, Nevin Hyuseinova, Qiong Cai, Blas Cuesta, Serkan Ozdemir and Marios Nicolaides, "Method for Pinning Data in Large Cache in Multi-Level Memory System", filed with patent number PCT/US13/32474.

2012 Ferad Zyulkyarov, Qiong Cai, Nevin Hyuseinova and Serkan Ozdemir "System and method for managing persistence with a multi-level memory hierarchy including non-volatile memory", filed with patent number US 13/997,220.

Qiong Cai, Nevin Hyuseinova, Serkan Ozdemir, Ferad Zyulkyarov, Marios Nicolaides and Blas Cuesta, "Adaptive cache replacement policy for a write-limited main memory", filed with patent number 13/626,464.

Serkan Ozdemir and Qiong Cai, "Endurance aware error-correcting code (ECC) protection for Non-volatile memories", file with patent number 13/630,541.

Ferad Zyulkyarov and Qiong Cai, "Persistent Log Operations for Non-Volatile Memory", filed with patent number 13/630,548.

2011 Serkan Ozdemir, Qiong Cai, Ayose Falcon and Nevin Hyuseinova, "Workload-adaptive address re-mapping methodology for improved PCM performance", filed with patent number US 13/995,469.

Nevin Hyuseinova and Qiong Cai, "Sub-block based wear leveling", filed with patent number PCT/US2011/067218.

Nevin Hyuseinova and Qiong Cai, "Page miss handler including wear leveling logic", filed with patent number PCT/US2011/067221.

Nevin Hyuseinova, Qiong Cai, Serkan Ozdemir and Ayose Falcon, "Utility and lifetime based cache replacement policy", filed with patent number PCT/US2011/067213.

Qiong Cai, Jose Gonzalez, Pedro Chaparro Monferre, Grigorios Magklis and Antonio Gonzalez, "Thread migration to improve power efficiency in a parallel processing environment", issued with patent number US 7930574 B2.

2010 Qiong Cai, Jose Gonzalez, Ryan Rakvic, Pedro Chaparro, Grigorios Magklis and Antonio Gonzalez, "Meeting point thread characterization", issued with patent number US 7665000 B2.

Grigoris Magklis, Jose Gonzalez, Pedro Chaparro, Qiong Cai, and Antonio Gonzalez, "Compressing address communications between processors", issued with patent number US 7698512 B2.

Publications

2011-2013 I published 6 internal papers between 2011 and 2013 and gave 4 major internal talks between 2011 and 2013. Due to Intel confidential policy, I cannot list them here. Some of them were filed as patents.

2011 Q. Cai, J. Gonzalez, G. Magklis, P. Chaparro and A. Gonzalez, "Thread Shuffling: Combing DVFS and Thread Migration to Reduce Energy Consumption for Multi-Core Systems", International Symposium on Low Power Electronics and Design (ISLPED), 2011.

2010 R. Rakvic, Q. Cai, J. Gonzalez, G. Magklis, P. Chaparoo and A. Gonzalez, "Thread-Management Techniques to Maximize Efficiency in Multicore and Simultaneous Multithreaded Microprocessors", ACM Transaction on Architecture and Code Optimization, Vol 7, No. 2, 2010.

R. Rakvic, J. Gonzalez, Q. Cai, P. Chaparro, G. Magklis and A. Gonzalez, "Energy Efficiency via Thread Fusion and Value Reuse", IET Computer and Digital Techniques, Vol 4, Issue 2, 2010.

2009 P. Chaparro, J. Gonzalez, Q. Cai and G. Chrysler, "Dynamic Thermal Management using Thin-Film Thermoelectric Cooling", International Symposium on Low Power Electronics and Design (ISLPED), 2009.

2008 Q. Cai, J. Gonzalez, R. Rakvic, G. Magklis, P. Chaparro and A. Gonzalez, "Meeting Points: Using Thread Criticality to Adapt Multicore Hardware to Parallel Regions", International Conference on Parallel Architecture and Compilation Techniques (PACT), 2008.

J. Gonzalez, Q. Cai, P. Chaparro, G. Magklis, R. Rakvic and A. Gonzalez, "Thread Fusion", International Symposium on Low Power Electronics and Design (ISLPED), 2008.

Q. Cai, J. M. Codina, J. Gonzalez and A. Gonzalez, "A Software-Hardware Hybrid Steering Mechanism for Clustered Microarchitectures", 22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2008.

2007 P. Chaparro, J. Gonzalez, G. Magklis, Q. Cai and A. Gonzalez, "Understanding the Thermal Implications of Multi-Core Architectures", IEEE Transactions on Parallel and Distributed Systems, Special Section on CMP Architectures, 18(8), pp. 1055-1065, 2007.

2006 J. Xue and Q. Cai, "A lifetime Optimal Algorithm for Speculative PRE", ACM Transactions on Architecture and Code Optimization, 3(2):115-155, 2006.

J. Xue, Q. Cai and L. Gao, "Partial Dead Code Elimination on Predicated Code Region", Software-Practice and Engineering, 36(15):1655-1685, 2006.

2004 Q. Cai, L. Gao and J. Xue, "Region-based Partial Dead Code Elimination on Predicated Code", 2004 International Conference on Compiler Construction, 2004.

2003 Q. Cai and J. Xue, "Optimal and Efficient Speculative-based Partial Redundancy Elimination", 1st Annual IEEE/ACM International Symposium on Code Generation and Optimization, 2003.

Awards and Scholarships

2014 Received an award from Intel data center group for excellent support to new server memory architecture (Crystal Ridge), the key player and contributor to this architecture and performance analysis.

2014 Received an award from Intel business group for devising and developing a new simulation methodology and toolchain for Iris Pro chips and two-level memory system based on 3D Xpoint memory technology.

2013 Received the division recognition award from Intel Labs for wear leveling work for 3D Xpoint.

2013 Received the division recognition award from Intel Labs for new simulation methodology for two level memory systems..

2012 Received an award from Intel business group for new memory hierarchy path finding for Intel's Crystal Ridge platform.

2008 Received the "Premio Duran Farell de Investigacion Tecnologica" to the best Research Project in Technology in Spain. Project name is Diseño Eficiente de Procesadores Mediante Particionado de Componentes (Efficient Processor Design by Clustering Resources).

2002–2005 The Australian Postgraduate Award (APA).

2002–2005 CSE-Supplementary Award, University of New South Wales

2004–2005 National ICT Australia (NICTA) Award

2002 CSE Summer Scholarship, University of New South Wales

2001 CSE Summer Scholarship, University of New South Wales

2001 MATH Summer Scholarship, University of New South Wales

Professional Activities

Editor Co-Content architects for Intel Technology Journal edition on Memory Resiliency

Paper Reviewer ISCA, MICRO, HPCA, PACT, ASPLOS, CGO, CC

R&D Project Participation Microarquitectura y Compiladores para Futuros Procesadores (microarchitecture and compilation for future processors), TIN2007-61763, Researcher, total funding 352,110 Euros, 2007-2010

Microarquitectura y Compiladores para Futuras Nanotecnologias(microarchitecture and compilation for future nanotechnology), TIN2004-03072, Researcher, total funding 135,000 Euros, 2004-2007

References

Available on request