JOSHUA J. BAKITA

jbakita@cs.unc.edu | https://jbakita.me/

EDUCATION

Doctor of Philosophy in Computer Science, *University of North Carolina at Chapel Hill*   Aug. 2019 - May 2025
Master of Science in Computer Science, *University of North Carolina at Chapel Hill*   Aug. 2018 - May 2019
- Courses of Note: Real-Time Sys, Computer Vision, OS Impls, Data Ctr. Sys & Programming, Data Ctr. Software Dynamics

Bachelor of Science in Computer Science, *University of North Carolina at Chapel Hill*   Aug. 2014 - May 2018
- **Graduated with Honors** | 3.57 Computer Science Course GPA | Fall 2016 Honors Study Abroad in London
- Courses of Note: Digital Logic & Computer Design, 2D Graphics, OS, Computer Security, Data Compression

RESEARCH EXPERIENCE

Dept. Computer Science - UNC Chapel Hill, *Research Assistant with Dr. James H. Anderson*   Aug. 2018 - Current
- Enabling **predictable and efficient GPU sharing** for real-time systems (pubs: RTSS'21, RTAS'19, ECRTS'18, **OSPERT'18**)
  - → Produced a **sub-1%-overhead** GPU core partitioning library and reverse engineered GPU scheduling (pub: **RTAS'23**)
  - → Demonstrated **41% lower mean and 52% lower max** execution times by using my library vs. MPS (pub: **EuroSys'25**)
  - → Revealed that 3 prior real-time GPU use and analysis approaches are unsafe on modern GPUs (pub: **RTAS'24**)
  - → Built transparent GPU memory oversubscription scheme **3x** faster than demand paging in Linux (pub: **RTSS'22**)
  - → My GPU partitioning and reverse engineering tools `libsmctrl` and `nvdebug` have **over 13,000 downloads** (as of Dec'24)
- **Linux kernel contributor**, LITMUS-RT maintainer, and kernel module developer (pubs: Journal of RTS'21)
- Reworked the <u>m</u>ixed-<u>c</u>riticality on <u>m</u>ulticore (MC²) Linux patches for x86 and SMT (pub: **RTAS'21**)
  - → Developed novel ways to isolate, e.g. NUMA-based **O(1) page coloring** in Linux's allocator (O(n) prior state-of-the-art)
  - → Rewrote MC² scheduler to allow for predictable execution times when using SMT, boosting schedulable utilization by 22%
- Constructed techniques to allow real-time systems to use SMT to address **FAA need** (pubs: RTAS'22, ECRTS'19)

General Motors Research, *Research & Development Intern*, Remote   June 2020 - Aug. 2020
- Developed and evaluated lossless GPU parallelization techniques in **CUDA** and **C++** for autonomous car perception CNNs
- Enabled **one device to do the work of two** at comparable latency by reducing context switches

Dept. Computer Science - UNC Chapel Hill, *Undergrad Research Assistant with Dr. Henry Fuchs*   June 2015 - Dec. 2015
- Led a team of 5 undergraduates to apply embedded wearable accelerometers for motion tracking
- Personal contributions in **C** on Arduino and **Java** on Android plus a high-efficiency bluetooth data transmission protocol

TEACHING EXPERIENCE

Dept. Computer Science - UNC Chapel Hill, *Instructor of Record for COMP 211*   Jan. 2023 - May 2023
- Developed curriculum and taught **C**, Linux, and more in *Systems Programming Fundamentals*, the 3rd intro CS course
- Led a team of 6 undergraduate and 1 graduate teaching assistants to teach **154 students**

Dept. Computer Science - UNC Chapel Hill, *Teaching Assistant for COMP 524*   Jan. 2019 - May 2019
- Taught **Haskell**, **Rust**, **Julia**, and **Go** to illustrate fundamental programming language concepts to a class of 150 students
- Graded assignments and tests, taught selected classes, and held up to 20 hours of office hours a week for students
- Honored with "**Teaching Assistant of the Year**" award by CS undergraduates and faculty. Nominated by >30% of class

PUBLICATIONS

J. Bakita, J. H. Anderson, "Hardware Compute Partitioning on NVIDIA GPUs for Composable Systems," *Proceedings of the 20th European Conference on Computer Systems (EuroSys)*, <u>in submission</u>, Apr 2025.

J. Goh, S. Ali, **J. Bakita**, S. Chakraborty, J. H. Anderson, "Concurrent FFT Execution on GPUs in Real-Time," *Proceedings of the 33rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, <u>in submission</u>, Mar 2025.

J. Bakita, J. H. Anderson, "Demystifying NVIDIA GPU Internals to Enable Reliable GPU Management," *Proceedings of the 30th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 394-305, May 2024. Acceptance rate: 23.3%.

J. Bakita, J. H. Anderson, "Hardware Compute Partitioning on NVIDIA GPUs," *Proceedings of the 29th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 54-66, May 2023. **Winner, outstanding paper award.** Acceptance rate: 26.9%.

J. Bakita, J. H. Anderson, "Enabling GPU Memory Oversubscription via Transparent Paging to an NVMe SSD," *Proceedings of the 43rd Real-Time Systems Symposium (RTSS)*, pp. 370-382, Dec 2022. Acceptance rate: 28.9%.

S. Osborne, **J. Bakita**, J. Chen, T. Yandrofski, J. H. Anderson, "Minimizing DAG Utilization by Exploiting SMT," *Proceedings of the 28th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 267–280, May 2022. Acceptance rate: 24.2%.

T. Amert, Z. Tong, S. Voronov, **J. Bakita**, F.D. Smith, and J. H. Anderson, "TimeWall: Enabling Time Partitioning for Real-Time Multicore+Accelerator Platforms," *Proceedings of the 42nd Real-Time Systems Symposium (RTSS),* pp. 455–468, Dec 2021. Acceptance rate: 29.3%.

J. Bakita, S. Ahmed, S.H. Osborne, S. Tang, J. Chen, F.D. Smith, and J. H. Anderson, "Simultaneous Multithreading in Mixed-Criticality Real-Time Systems," *Proceedings of the 27th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 278-291, May 2021. Acceptance rate: 27.2%.

C. Hobbs, Z. Tong, **J. Bakita**, and J. H. Anderson, "Statically Optimal Dynamic Soft Real-Time Semi-Partitioned Scheduling," *Real-Time Systems Journal,* pp. 97-140, 57.1, Jan 2021.

S. Osborne, **J. Bakita**, and J. H. Anderson, "Simultaneous Multithreading Applied to Real Time," *Proceedings of the 31st Euromicro Conference on Real-Time Systems (ECRTS)*, pp. 3:1-3:22, Jul 2019. Acceptance rate: 33.8%.

M. Yang, S. Wang, **J. Bakita**, T. Vu, F.D. Smith, J. H. Anderson, and J.-M. Frahm, "Re-thinking CNN Frameworks for Time-Sensitive Autonomous-Driving Applications: Addressing an Industrial Challenge," *Proceedings of the 25th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pp. 305-317, Apr 2019. Acceptance rate: 25.8%.

M. Yang, N. Otterness, T. Amert, **J. Bakita**, J. H. Anderson, and F.D. Smith, "Avoiding Pitfalls when Using NVIDIA GPUs for Real-Time Tasks in Autonomous Systems," *Proceedings of the 30th Euromicro Conference on Real-Time Systems (ECRTS)*, pp. 20:1-20:21, Jul 2018. Acceptance rate: 33.3%.

J. Bakita, N. Otterness, J. H. Anderson, and F.D. Smith, "Scaling Up: The Validation of Empirically Derived Scheduling Rules on NVIDIA GPUs," *Proceedings of the 14th Annual Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPERT)*, pp. 49-54, Jul 2018. Acceptance rate: 75.0%.

TECHNICAL EXPERIENCE

Waymo, *Software Engineering Intern,* Remote   May 2021 - Aug. 2021
- Worked on the Onboard Compute Optimization Team to improve planner GPU utilization and CPU SIMD efficiency
- Reimplemented edge construction for on-vehicle planner search graph on GPU using **C++** and **CUDA**

Microsoft Corporation, *Software Engineering Intern*, Redmond, WA   May 2018 - Aug. 2018
- Worked in **C++** on the Web Platform Team to help build the Edge browser (**>100 million active users**) + Windows app APIs
- Implemented CSS parser, DOM interface, and **GPU-accelerated rendering** for background-blend-mode and mix-blend-mode (used by .04% and .111% of all web pages per Bing, both are top 55% most used CSS properties)
- Wrote web platform interoperability tests in **HTML**, **CSS**, and **JavaScript** to benchmark and beat Chrome's implementation

Microsoft Corporation, *Software Engineering Intern*, Issaquah, WA   May 2017 - Aug. 2017
- Redesigned architecture for lead unsubscribe, **increasing speed 2x** for customer and protecting from **over $3B** in fines
- Worked in **C#** and **SQL** on enrichment and privacy management systems processing over **16 million leads** weekly
- Upgraded, unified, and simplified logging in lead enrichment and privacy sync systems to enable business alerting

House of Commons of the United Kingdom, *Parliamentary Intern for James Berry MP*, London, UK   Sep. 2016 - Dec. 2016
Capital One Financial Corporation, *Software Engineering Intern (TDP)*, McLean, VA   June 2016 - Aug. 2016
- Refactored free CreditWise tool (**>11 million active users**) to **speed up deployment 10x**, startup by 2x, and testing by ~2x
- Full stack development in **Java**, Spring MVC, **CSS**, **HTML**, and **AngularJS** on Apache Tomcat in **AWS** EC2
- Optimized test workflow, removed all proprietary libraries, and significantly slimmed backend codebase size

Wildfire Games, *Open-Source Game Developer*   June 2013 - Aug. 2014
- Led a team to develop an online multiplayer matchmaking lobby and ranking system based on XMPP for 0 A.D. RTS
- Worked with **C++**, **JavaScript**, and XML client-side, **Python** and **Erlang** server-side

SERVICE

Graduate and Professional Student Government (GPSG), *Senator*   Aug. 2020 - Aug 2024
- Sole elected representative of the graduate students in the Computer Science Department to the university administration
- President Pro Tempore of the Senate, Sept. 2022 - Sept. 2023
- Chair of Senate Finance Committee, Aug. 2021 - Sept. 2022

Computer Science Students Association (CSSA), *President*   Aug 2021 - Aug 2022
- Represented graduate students in the Computer Science Dept. as a voting member of faculty meetings
- Contributed to a **13.6% stipend increase** for graduate students, and a ~25% increase in graduate student event attendance
- Leadership Committee Officer, May 2020 - Aug 2021

UNC Renewable Energy Special Projects Committee (RESPC), *Voting Member*   Sep. 2014 - May 2019
- Collaborated with a committee of over 15 students to **manage over $1M** for renewable energy initiatives on campus

UNC Undergraduate Computer Science Club, *President*   May 2017 - Aug. 2018
Rancho 3M Christian School and Orphanage, *Missions Trip Volunteer*, Guadalupe, Mexico   2010,11,13,15,16,19
Eagle Scout; Boy Scouts of America, *Troop 94*   2007 - 2014

REVIEWING ACTIVITIES

Program Committee Service

    2024 Real-Time Systems Symposium (RTSS). Work-in-Progress and Brief Presentations PC.

    2024 Euromicro Conference on Real-Time Systems (ECRTS). Artifact Evaluation PC.

Conference Subreviewer

    2025 European Conference on Computer Systems (EuroSys).

    2024 Real-Time Systems Symposium (RTSS).

    2024 Euromicro Conference on Real-Time Systems (ECRTS).

    2023 Real-Time Systems Symposium (RTSS).

    2023 International Conference on Real-Time Networks and Systems (RTNS).

    2023 Real-Time and Embedded Technology and Applications Symposium (RTAS).

    2022 International Conference on Real-Time Networks and Systems (RTNS).

    2022 Real-Time and Embedded Technology and Applications Symposium (RTAS).

    2021 Euromicro Conference on Real-Time Systems (ECRTS).

    2021 Real-Time and Embedded Technology and Applications Symposium (RTAS).

Journal Reviewer

    Journal of Systems Architecture.

OTHER ACTIVITIES

Supervised Undergraduate Students

    Alex Georgiev, *(current)*.

    Benjamin Hadad IV, *(now graduating senior)*.

    Saman Sehabi, *(now 1st year MS student at NC State University)*.

    Jingyuan ("Leo") Chen, *(now 3rd year PhD student at Princeton University)*.

Invited Talk, "Unlocking Simple and Efficient Real-Time AI Inference on COTS Hardware via Better GPU Models", *2024 Workshop on Machine-learning enabled safety-Critical systems (WMC)*.

Pitch Presentation, "GPU Predictability and Throughput in Autonomous Systems: Can We Have Both?", *2024 Euromicro Conference on Real-Time Systems (ECRTS)*.

Invited Keynote, "Why HackNC is For You", *HackNC 2022 (Undergraduate Hackathon)*.

Hackathon Winner, "Best Hack Addressing Inequality" and "Best Use of the Blockchain", *HackDuke 2017*.