Jovan Stojkovic

Curriculum Vitae

Computer Science Department, UIUC
✉ *jovans2@illinois.edu*
https://jovans2.github.io/

Research Interests

I have been working on novel hardware and software abstractions in data-center architectures with the focus on cloud computing data platforms and deployment paradigms, such as microservices and serverless computing. I enjoy building systems and exploring ways to make them fast, reliable, and efficient in a holistic manner: from the hardware architecture up to the platform and application level.

Education

- August 2020 – Present — **University of Illinois at Urbana Champaign**.
  PhD in Computer Science
  Advisor: *Professor Josep Torrellas*
  GPA: *4.0/4*
  Passed Qualifying Exam: *October 2021*

- 2016 – 2020 — **School of Electrical Engineering, University of Belgrade, Serbia**.
  B.S. in Electrical and Computer Engineering
  Awards: *Best student of Computer Engineering and Information Theory Department for years: 2017, 2018, 2019 and 2020*
  GPA: *9.89/10*

- 2012 – 2016 — **High School Bora Stankovic, Vranje, Serbia**.
  Applied Sciences and Mathematics
  Award: *Best student*
  GPA: *5/5*

Awards and Honors

- April 2024 — **Young Researcher at Heidelberg Laureate Forum (HLF)**, *Selected as one of the 200 young researchers in computer science and mathematics worldwide invited to attend the 11th HLF*.
- January 2024 — **IEEE Micro Top Picks Honorable Mention in Computer Architecture**, *Awarded to the most significant papers in computer architecture published in the previous year*.
- 2022-2023 — **Invited Talks**, *Research seminars at Uber, Microsoft, IBM; Annual Review meeting for ACE Center for Evolvable Computing; Networked Systems Seminar at Cornell*.
- October 2022 — **Invitation for the Workshop on the Future of Computer Architectures (FOCA)**, *Present my research at IBM Research, Yorktown Heights, NY*.
- 2022-2023 — **Student Travel Grants**, *International Symposium on Computer Architecture (ISCA '23), International Symposium on High-Performance Computer Architecture (HPCA '23), International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23, '22), International Symposium on Microarchitecture (MICRO '22)*.
- April 2022 — **Kenichi Miura Award - Excellence in High Performance Computing**, *Department of Computer Science, University of Illinois at Urbana-Champaign*.
- April 2020 — **Best Research Artifact Award**, *International Conference on Information Processing on Sensor Networks (IPSN '20)*.

- 2017, 2018, 2019, 2020 **Best Student of Computer Engineering and Information Theory Department Award**, *School of Electrical Engineering, University of Belgrade*.
- July 2016 **Bronze medal in 48th International Chemistry Olympiad**, *Tbilisi, Georgia*.
- Sep. 2015 **Silver medal in 10th Annual Bios Olympiad**, *St. Petersburg, Russia*.
- 2016 – 2020 **Stipend for Young Scientist/Researcher, from Serbian Ministry of Education, Science and Technological Development**.
- 2010 – 2016 **Numerous Gold, Silver and Bronze Awards at National Level Competitions in Physics, Chemistry, Biology and Mathematics**.
- Dec. 2016 **Serbian Academy of Sciences and Arts Award for Results at International Olympiads**.
- Dec. 2016 **Serbian Chemistry Society Award**.
- July 2017 **Dositeja Award from Young Talent Fund of the Republic of Serbia**.
- January 2015 **St. Sava Award for the Best Student of Pcinja District**.

Publications

J. Stojkovic, N. Iliakopoulou, T. Xu, H. Franke, J. Torrellas, "EcoFaaS: Rethinking the Design of Serverless Environments for Energy Efficiency", *To Appear in Proceedings of the 51th International Symposium on Computer Architecture (ISCA)*, June, 2024.

J. Stojkovic, P. Misra, I. Goiri, S. Whitlock, E. Choukse, M. Das, C. Bansal, J. Lee, H. Qiu, R. Zimmermann, S. Samal, B. Warrier, R. Bianchini, "SmartOClock: Workload- and Risk-Aware Overclocking in the Cloud", *To Appear in Proceedings of the 51th International Symposium on Computer Architecture (ISCA)*, June, 2024.

J. Stojkovic, C. Liu, M. Shahbaz, J. Torrellas, "μManycore: A Cloud-Native CPU for Tail at Scale", *In Proceedings of the 50th International Symposium on Computer Architecture (ISCA),* ***Selected as an IEEE Micro Top Picks Honorable Mention***, June, 2023.

J. Stojkovic, T. Xu, H. Franke, J. Torrellas, "MXFaaS: Rethinking Resource Sharing in Serverless Environments for Parallelism and Efficiency", *In Proceedings of the 50th International Symposium on Computer Architecture (ISCA)*, June, 2023.

J. Stojkovic, T. Xu, H. Franke, J. Torrellas, "SpecFaaS: Accelerating Function-as-a-Service Applications with Speculative Function Execution", *In Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, February, 2023.

J. Stojkovic, N. Mantri, D. Skarlatos, T. Xu, J. Torrellas, "Memory Efficient Hashed Page Tables", *In Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, February, 2023.

J. Stojkovic, D. Skarlatos, A. Kokolis, T. Xu, J. Torrellas, "Parallel Virtualized Memory Translation with Nested Elastic Cuckoo Page Tables", *In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, March, 2022.

G. Lan, Z. Liu, Y. Zhang, T. Scargill, **J. Stojkovic**, C. Joe-Wong, M. Gorlatova, "Edge-assisted Collaborative Image Recognition for Mobile Augmented Reality", *ACM Transactions on Sensor Networks*, February, 2022.

Z. Liu, G. Lan, **J. Stojkovic**, Y. Zhang, C. Joe-Wong, M. Gorlatova, "CollabAR: Edge-assisted Collaborative Image Recognition for Mobile Augmented Reality," *In Proceedings of the International Conference on Information Processing on Sensor Networks (IPSN)*, April, 2020. – **Best Research Artifact Award**

J. Stojkovic, M. Misic, J. Protic, "Collaboration Network Analysis of Scientific Production at UB-SEE", *In 27th Telecommunications Forum (TELFOR)*, November 2019.

Workshops, Posters, Demo

J. Stojkovic, E. Choukse, C. Zhang, I. Goiri, J. Torrellas, "Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference", *In 9th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2 '24) in conjunction with ASPLOS'24*, April 2024.

J. Stojkovic, T. Xu, H. Franke, J. Torrellas, "UniCache: The Next 700 Caches for Serverless Computing", *In 5th International Workshop on Cloud Intelligence/AIOps (AIOps '24) in conjunction with ASPLOS'24*, April 2024.

N. Stojkovic, **J. Stojkovic**, "OasisRPC: Hiding the Overheads of RPCs in Microservice Environments", *In 6th Young Architect Workshop (YArch'24) in conjunction with ASPLOS'24*, April 2024.

J. Stojkovic, C. Liu, M. Shahbaz, J. Torrellas, "Hardware Design for Core Harvesting in Microservice-Heavy Clouds", *In SRC TECHCON Conference*, September 2023.

J. Stojkovic, C. Liu, M. Shahbaz, J. Torrellas, "Hardware Support for Efficient and Secure Resource Harvesting in the Cloud", *In 5th Young Architect Workshop (YArch'23) in conjunction with ASPLOS'23*, March 2023.

J. Stojkovic, T. Xu, H. Franke, J. Torrellas, "Super Scalar Clouds", *In 7th Workshop on the Future of Computing Architecture (FOCA'22)*, November 2022.

J. Stojkovic, J. Torrellas, "Nested Elastic Cuckoo Page Tables", *NSF Arch-1 Workshop*, March 2022.

J. Stojkovic, Z. Liu, G. Lan, C. Joe-Wong, M. Gorlatova, "Demo: Edge-assisted Collaborative Image Recognition for Augmented Reality," *In ACM Conference on Embedded Networked Sensor Systems (SenSys)*, November 2019.

Research Experience

- May – August 2024 — **Intern, Azure System Research, Microsoft Research**, *Advisors : Dr Chaojie Zhang and Dr Esha Choukse*, Energy Efficient High-Performance LLM Inference Server, Redmond, WA.
- May – August 2023 — **Intern, System Innovations, Microsoft Research**, *Advisors : Dr Pulkit Misra and Dr Inigo Goiri*, Virtual Machine Overclocking in the Cloud, Redmond, WA.
- May – August 2022 — **Intern, Hybrid Cloud, IBM Research**, *Advisor : Dr Hubertus Franke*, Efficient and Performant Serverless Computing, Thomas J. Watson Research Center, NY.
- August 2020 – Present — **Research Assistant at University of Illinois at Urbana-Champaign**, *Advisor : Professor Josep Torrellas*, Rethinking Architecture and OS for Modern Virtualization Technologies.
- May – July 2019 — **Duke ECE REU Program**, *Advisor : Professor Maria Gorlatova*, Edge Computing Platforms for the IoT and Collaborative AR.
- 2018 – 2020 — **School of Electrical Engineering, University of Belgrade**, *Advisor : Professor Marko Misic*, Social and Collaboration Networks Analysis.

Mentoring Experience

- 2022-Present — **Nikoleta Iliakopoulou**, 2^{nd} *year PhD student at UIUC*, Infrastructure for Efficient LLM Serving.
- 2023-Present — **Abraham Farrell**, 1^{st} *year PhD student at UIUC*, Hardware Design for Cloud Workloads.
- 2024-Present — **Alan Andrade**, 1^{st} *year master's student at UIUC*, Software Caches for Serverless Workloads.
- 2023-2023 — **Krut Patel**, 1^{st} *year master's student at UIUC*, Graph Analytics on Serverless Platforms.
- 2022-2023 — **Chunao Liu**, 1^{st} *year master's student at Purdue*, CPU Architecture for Microservice Workloads.
- 2022-2023 — **Feiran Qin**, 4^{th} *year undergraduate student at Shanghai Tech*, SW Design for FaaS Workloads.

Teaching Experience

- Spring 2023 — **Guest Lecture**, *UIUC*, CS 533 Parallel Computer Architectures.
- Fall 2022 — **Guest Lecture**, *UIUC*, CS 534 Energy Efficient Computer Architectures.
- Spring 2020 — **Undergrad TA**, *University of Belgrade*, Operating Systems, Object-oriented Programming.
- Fall 2019 — **Undergrad TA**, *University of Belgrade*, Computer Architecture, Algorithms and Data Structures, Fundamentals of Databases, Concurrent and Distributed Programming.

- Spring 2019 **Undergrad TA**, *University of Belgrade*, Computer Networks, Probability and Statistics, Operating Systems, Object-oriented Programming.
- Fall 2018 **Undergrad TA**, *University of Belgrade*, Computer Architecture, Algorithms and Data Structures.
- Spring 2018 **Undergrad TA**, *University of Belgrade*, Lab Exercises in Fundamentals of Electrical Engineering.

Service Experience

- 2021-Present **Graduate Student Ambassador**, *Computer Science Department at UIUC*, Help recruiting students and guiding the admitted students to feel welcome in the beginning of PhD studies.
- 2023-Present **Compilers, Architecture, and Parallel Computing (CAP) Seminar Organizer**, *Computer Science Department at UIUC*, Organize events, lead discussions and invite speakers.
- 2023-Present **Meet a Senior Student (MaSS)**, *Computer Architecture Conferences (ISCA, ASPLOS)*, Guide early-career students with their research and navigating them through the PhD process.

Projects

- **Hardware Support for Microservices in the Cloud (UIUC)** : Current processors are not designed for microservices, an emerging cloud-computing paradigm. Contrary to long-running monolithic applications, microservice environments execute short functions that only interact with one another via remote procedure calls and are subject to stringent tail-latency constraints. During the third year of my PhD, I designed 1) an architecture optimized for cloud-native environments that minimizes unnecessary architecture and removes contention hot-spots that degrade the tail latency (μManycore, ISCA'23), and 2) a processor architecture that enables efficient and secure resource harvesting in the cloud (μHarvest, YArch'23). During the fourth year of my PhD, I have been working on 1) a hardware-software co-design that improves micro-architectural resource utilization in microservice/serverless environments, and 2) a hardware cache coherence protocol that enables servers running microservices to scale to thousands of cores.
- **High-Performant Serverless Computing (UIUC and IBM)** : Current serverless workloads exhibit multiple levels of overheads, including cold start, virtualization, RPC, and the need to persist temporal outputs in remote storage. To make the matter worse, the overheads have significant cascading effects for applications that orchestrate multiple functions through control and data dependencies, which is a common practice in complex real-world applications. During the second and third year of my PhD, I redesigned serverless platforms by 1) applying data and control speculation techniques (SpecFaaS, HPCA'23), 2) creating a novel container abstraction that ensures high cpu, memory and I/O resource utilization while minimizing the response time (MXFaaS, ISCA'23), and 3) providing a systematic approach for mapping the unique characteristics of serverless workload to the design and organization of distributed caching scheme. During the fourth year of my PhD, I have been working on 1) a design of energy efficient serverless system (EcoFaaS, in submission), and 2) an efficient coherence protocol for distributed software caches in serverless environments.
- **Leveraging Parallelism in Virtualized Address Translation (UIUC)** : In spite of nearly twenty years since the inception of virtualization hardware, address translation still introduces substantial performance overhead in virtualized systems. A major reason why address translation has high overhead is because page tables are currently organized in a multi-level tree that is accessed in a sequential manner. This organization is called radix page tables. During my first year of PhD, I redesigned the virtual memory subsystem in virtualized environments to improve its 1) performance by exploiting available memory level parallelism (Nested ECPTs, ASPLOS'22) and 2) memory efficiency by breaking the hashed page table into multiple small allocation units while maintaining the same performance through carefully chosen design decisions (ME-HPTs, HPCA'23).
- **Exploring the Design Space of Virtual Memory: Performance, Memory, Energy Efficiency, Security and Hardware Heterogeneity (UIUC)** : I did an extensive study of the virtual memory subsystem in virtualized environments and conducted sensitivity analyses on helper hardware structures. My results suggest that the current radix organization of page tables is not scalable with a large memory footprint workload. Moreover, I reproduced literature solutions to avoid hardware-based address translation and instead used compiler and runtime protection mechanisms. I also explored the virtual memory design for GPUs and available prefetching

opportunities inherently exposed by the SIMT model of execution. Finally, I conducted a study on the security implications of different page table and TLB organizations with existing attack and defense schemes.

- **Edge Computing Platform for Collaborative Augmented Reality (Duke)** : I built a platform that allows multiple users' related images, captured with Android phones running Google ARCore, to be processed jointly on an edge server, improving user's quality of object recognition. Additionally, I improved performance by reusing cached results from the database. I also did comparative benchmarking of two image recognition tools (AWS Rekognition and Yolo) in terms of quality and time efficiency. I went further to explore possible trade-offs. The project was presented at Duke's REU Symposium and published at IPSN and SenSys.

Technical Skills

Programming languages: C, C++, Java, Python, Golang, Scala, Kotlin, Arduino, Assembly (x86 and ARM)
Software stacks: *container environments* (Docker, Kubernetes), *serverless platforms* (KNative, OpenWhisk), *LLM serving* (vLLM, FasterTransformer), *storage systems* (Azure Blob Storage, MongoDB, Redis, Memcached)
Hardware platforms: GPUs, FPGAs, Arduino Uno, Raspberry Pi
Architecture simulators: SST, Simics, gem5, QEMU, Pin, ChampSim

Miscellaneous

I speak Serbian, English and French. I have been playing tennis for sixteen years and I coached young players. During my primary education I played violin and was member of scouts. I was member of many non-governmental organizations such as Red Cross, Nexus and others. I love dogs and nature.