Jovan Stojkovic

Curriculum Vitae

Computer Science Department
University of Illinois at Urbana-Champaign
✉ *jovans2@illinois.edu*
https://jovans2.github.io/

Research Interests

- Computer architecture and hardware-software co-design of datacenters.
- Cloud computing and emerging paradigms such as microservices and serverless computing.
- Power, energy, and thermal efficiency of cloud environments.

Education

- August 2020 – Present: **University of Illinois at Urbana-Champaign**.
  PhD in Computer Science **Advisor**: *Professor Josep Torrellas*
  Thesis: *Chasing the "Tail at Scale": Toward Cloud-Native Architectures*
  Awards: *Mavis Future Faculty Fellowship (2024); Kenichi Miura Award (2022)* **GPA**: *4.0/4*
- 2016 – 2020: **School of Electrical Engineering, University of Belgrade, Serbia**.
  B.S. in Electrical and Computer Engineering
  Awards: *Best student in the Department (2017-2020)*
  GPA: *9.89/10*

Awards and Honors

- March 2025: **Best Paper Award**, *IEEE International Symposium on High-Performance Computer Architecture (HPCA'25)*, Las Vegas, NV, USA.
- August 2024: **MICRO PhD Forum**, *Selected to present my PhD research at the PhD Forum*, International Symposium on Microarchitecture (MICRO'24), Austin, TX, USA.
- May 2024: **Mavis Future Faculty Fellowship**, *Selected as one of the MF3 fellows for the 2024-2025 academic year*, The Grainger College of Engineering, University of Illinois at Urbana-Champaign.
- April 2024: **Young Researcher at the Heidelberg Laureate Forum (HLF)**, *Selected as one of the 30 young researchers in CS/mathematics worldwide invited to present their research at the HLF.*
- January 2024: **IEEE Micro Top Picks in Computer Architecture, Honorable Mention**, *Awarded to the most significant papers in computer architecture published in the previous year.*
- 2022-2024: **Invited Talks**, *Gave research seminars at industry (Uber, Microsoft, IBM, RedHat), academia (Cornell, University of Wisconsin-Madison, University of California-Riverside), Annual Meeting of the ACE Center for Evolvable Computing, and vHive community meetup.*
- October 2022: **Workshop on the Future of Computer Architectures (FOCA)**, *Selected to present my PhD research at IBM Research, Yorktown Heights, NY.*
- 2022-2024: **Student Travel Grants**, *ISCA '23, '24; HPCA '23; ASPLOS '22, '23; MICRO '22, '24.*
- April 2022: **Kenichi Miura Award - For Excellence in High Performance Computing**, *Department of Computer Science, University of Illinois at Urbana-Champaign.*
- April 2020: **Best Artifact Award**, *International Conference on Information Processing on Sensor Networks (IPSN '20)*, Sydney, Australia.
- 2017-2020: **Best Student of the Computer Engineering and Information Theory Department Award**, *School of Electrical Engineering, University of Belgrade.*

Publications

C1. **J. Stojkovic**, C. Liu, M. Shahbaz, J. Torrellas, "HardHarvest: Hardware-Supported Core Harvesting for Microservices", *To Appear in Proceedings of the 52nd International Symposium on Computer Architecture (ISCA)*, June, 2025.

C2. **J. Stojkovic**, C. Zhang, I. Goiri, E. Choukse, H. Qiu, R. Fonseca, J. Torrellas, R. Bianchini, "TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms", *To Appear in Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, April, 2025.

C3. **J.Stojkovic**, C. Alverti, A. Andrade, N. Iliakopoulou, T. Xu, H. Franke, J. Torrellas , "Concord: Rethinking Distributed Coherence for Software Caches in Serverless Environments", *In Proceedings of the 31st IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, March, 2025.

C4. **J. Stojkovic**, C. Zhang, I. Goiri, J. Torrellas, and E. Choukse, "DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency", *In Proceedings of the 31st IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, March, 2025. – **Best Paper Award**

C5. **J. Stojkovic**, E. Choukse, E. Saurez, I. Goiri, J. Torrellas, "Mosaic: Harnessing the Micro-architectural Resources of Servers in Serverless Environments", *In Proceedings of the 57th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, November, 2024.

C6. **J. Stojkovic**, N. Iliakopoulou, T. Xu, H. Franke, J. Torrellas, "EcoFaaS: Rethinking the Design of Serverless Environments for Energy Efficiency", *In Proceedings of the 51st International Symposium on Computer Architecture (ISCA)*, June, 2024.

C7. **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", *In Proceedings of the 51st International Symposium on Computer Architecture (ISCA)*, June, 2024.

C8. **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.

C9. **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.

C10. **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.

C11. **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.

C12. **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.

C13. 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.

C14. 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**

C15. **J. Stojkovic**, M. Misic, J. Protic, "Collaboration Network Analysis of Scientific Production at UB-SEE", *In 27th Telecommunications Forum (TELFOR)*, November 2019.

Pre-prints

A1. **J. Stojkovic**, C. Zhang, Í. Goiri, E. Choukse, H. Qiu, R. Fonseca, J. Torrellas, R. Bianchini, "TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms", *CoRR, vol. abs/2501.02600*, 2025

A2. N. Iliakopoulou, **J. Stojkovic**, C. Alverti, T. Xu, H. Franke, J. Torrellas, "Chameleon: Adaptive Caching and Scheduling for Many-Adapter LLM Inference Environments", *CoRR, vol. abs/2411.17741*, 2024

A3. **J. Stojkovic**, C. Zhang, I. Goiri, J. Torrellas, and E. Choukse, "DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency", *CoRR, vol. abs/2408.00741*, 2024

A4. L. Huang, A. Parayil, J. Zhang, X. Qin, C. Bansal, **J. Stojkovic**, P. Zardoshti, P. Misra, E. Cortez, R. Ghelman, I. Goiri, S. Rajmohan, J. Kleewein, R. Fonseca, T. Zhu, R. Bianchini, "Workload Intelligence: Punching Holes Through the Cloud Abstraction", *CoRR, vol. abs/2404.19143*, 2024

Workshop Papers

W1. **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*, April 2024.

W2. **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*, April 2024.

W3. N. Stojkovic, **J. Stojkovic**, "OasisRPC: Hiding the Overheads of RPCs in Microservice Environments", *In 6th Young Architect Workshop (YArch'24) in conjunction with ASPLOS*, April 2024.

W4. **J. Stojkovic**, C. Liu, M. Shahbaz, J. Torrellas, "Hardware Design for Core Harvesting in Microservice-Heavy Clouds", *In SRC TECHCON Conference*, September 2023.

W5. **J. Stojkovic**, C. Liu, M. Shahbaz, J. Torrellas, "Hardware for Efficient and Secure Resource Harvesting in the Cloud", *In 5th Young Architect Workshop (YArch'23) in conjunction with ASPLOS*, March 2023.

W6. **J. Stojkovic**, T. Xu, H. Franke, J. Torrellas, "Super Scalar Clouds", *In 7th Workshop on the Future of Computing Architecture (FOCA'22)*, November 2022.

W7. **J. Stojkovic**, J. Torrellas, "Nested Elastic Cuckoo Page Tables", *NSF Arch-1 Workshop*, March 2022.

W8. **J. Stojkovic**, Z. Liu, G. Lan, C. Joe-Wong, M. Gorlatova, "Edge-assisted Collaborative Image Recognition for Augmented Reality," *In ACM Conference on Embedded Networked Sensor Systems (SenSys)*, November 2019.

Patents

P1. **J. Stojkovic**, H. Franke, "Serverless Computing with Latency Reduction", *US Patent App 18/152,341*, 2024.

P2. **J. Stojkovic**, H. Franke, T.Xu, J. Torrellas, "Serverless Computing Using Resource Multiplexing", *US Patent App 18/049,125*, 2024.

P3. **J. Stojkovic**, H. Franke, A. Buyuktosunoglu, "Power Management in Serverless Computing", *Filed*, 2024.

P4. **J. Stojkovic**, E. Choukse, C. Zhang, I. Goiri, J. Torrellas, "LLM Inference Clusters for Performance and Energy Efficiency", *Filed*, 2024.

P5. **J. Stojkovic**, C. Zhang, I. Goiri, E. Choukse, H. Qiu, R. Fonseca, J. Torrellas, R. Bianchini, "Thermal- and Power-Aware Scheduling for LLM Inference in the Cloud", *Filed*, 2024.

P6. P. Misra, S.Whitlock, E. Choukse, **J. Stojkovic**, I. Goiri, R. Bianchini, "Distributed Overclocking and Risk-Aware Overclocking Management", *Filed*, 2024

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

- 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.
- 2024-Present **JooYoung Park**, 1^{st} *year PhD student at NTU Singapore*, Networking for Serverless Workloads.
- 2023-2024 **Nikoleta Iliakopoulou**, 2^{nd} *year PhD student at UIUC*, Infrastructure for Efficient LLM Serving.
- 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 2025 **Teaching Assistant**, *UIUC*, CS 533 Parallel Computer Architecture.
- Spring 2024 **Guest Lecture**, *UIUC*, CS 533 Parallel Computer Architectures, "Architecture for Datacenters".
- Spring 2023 **Guest Lecture**, *UIUC*, CS 533 Parallel Computer Architectures, "Memory Hierarchies".
- Fall 2022 **Guest Lecture**, *UIUC*, CS 534 Energy Efficient Computer Architectures, "Process Variations".
- 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 recruit students and guide admitted students to feel welcome.
- 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 navigate them through the PhD process.

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