

Prof. Dejan Kostić (<http://dejankostic.com>)

- Wrapping up my ERC Consolidator Grant: "Increasing the Spatial Correlation of Logical Units of Data to Enable an Ultra-Low Latency Internet (ULTRA)"
 - Showed that deliberately delaying packets can improve the performance of backend servers by ~2x ("OrderMatters", Community Award at NSDI 2022)
 - Reduced the CPU utilization of high-performance distributed filesystems (Best Paper Award at SOSP 2021); more info at <https://www.kth.se/blogs/ultra/>
 - 15 patent applications with Ericsson Research
- Getting ready to start my Wallenberg Scholar grant: "Sustainable and Adaptive Inferencing for Democratizing AI"
 - Machine learning inferencing is already unsustainable
 - the cost of simply running ChatGPT is estimated to be more than \$700 000 per day
 - Goal: fundamentally decrease (by a factor of 10) the power consumption of systems that are used for inferencing when using large language models