



Студијски програм: Мастер академске студије информатике			
Назив предмета: АРХИТЕКТУРА СПЕЦИЈАЛИЗОВАНИХ ПРОЦЕСОРА ЗА ЗАКЉУЧИВАЊЕ			
Статус предмета: Изборни на свим модулима мастер академских студија информатике			
Број ЕСПБ: 6			
Услов: Уписан одговарајући семестар			
Циљ предмета Упознати студенте са савременим архитектурама акцелератора за извођење вештачке интелигенције, са акцентом на digital in-memory computing парадигму коју примењује компанија D-Matrix . Развити разумевање техничких изазова у погледу латенције, енергетске ефикасности и скалабилности у односу на стандардне приступе.			
Исход предмета Након успешно савладаног предмета, студент ће бити способан да разуме основне принципе <i>in-memory</i> и <i>near-memory</i> рачунања, да објасни архитектуру и начин рада D-Matrix AI акцелератора (Corsair), као и да упоређи различите архитектуре AI акцелератора као што су TPU, GPU и DIMC (Digital In-Memory Computing). Такође, студент ће бити у стању да симулира основне функционалности DIMC система и да анализира перформансе и ограничења савремених AI акцелератора.			
Садржај предмета <i>Теоријска настава</i> Увод у AI акцелераторе – мотивација, потребе и изазови; Енергетска ефикасност и латенција у инференцији великих модела; Фон Нојманско уско грло и прелазак ка <i>in-memory</i> архитектурама; Принципи <i>digital in-memory</i> рачунања; Алгоритми са свешћу о реткости (<i>sparsity-aware</i>) и уштеда ресурса; Архитектура D-Matrix Corsair акцелератора; Алати за моделирање и симулацију AI акцелератора; <i>Benchmarking</i> LLM модела на различитим архитектурама; Перспективе развоја и примене DIMC решења <i>Практична настава</i> Упознавање са окружењем и алатом за симулацију (NumPy, PyTorch); Симулација множења матрица у меморији (DIMC); Разлика између класичног и <i>in-memory</i> сабирања; Моделовање ретких матрица и оптимизација операција; Увод у HDL моделовање – једноставан DIMC модул; Верификација DIMC модула у Vivado симулатору; Бенчмаркирање инференције LLM модела – CPU vs GPU vs симулација DIMC; Коришћење отвореног кода и алата за акцелераторе (ONNX, Triton); Пројектни рад: дизајн и симулација акцелератора за један слој мреже; Анализа перформанси: латенција, проток, потрошња; тудија случаја: D-Matrix приступ и архитектура; Припрема и презентација пројеката.			
Литература 1. Hennessy, J. L., & Patterson, D. A., <i>Computer Architecture: A Quantitative Approach</i> EI, ISBN: 9780128119051 2. V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, <i>Efficient Processing of Deep Neural Networks</i> . San Rafael, CA, USA: Morgan & Claypool Publishers, 2020. ISBN: 978-1-68173-831-4. 3. https://www.d-matrix.ai/			
Број часова активне наставе		Теоријска настава:	2
		Практична настава:	1 + 1
Методе извођења наставе Проблемски-оријентисана настава, практична настава, самостални рад студената, консултације. Комбинација класичне наставе са е-учењем и уз одговарајућу литературу. Практична настава се обавља у виду лабораторијских вежби у рачунарским учионицама, на којима студенти самостално или уз помоћ асистената решавају реалне проблеме из области оптимизације.			
Оцена знања (максимални број поена 100)			
Предиспитне обавезе	70 поена	Завршни испит	30 поена
колоквијуми	30	усмени испит	30
семинар-и	40		