FIK3221 Networked Systems for Machine Learning 7.5 credits

Nätverkssystem för maskininlärning

This is a translation of the Swedish, legally binding, course syllabus.

Establishment

Course syllabus for FIK3221 valid from Spring 2025

Grading scale

P, F

Education cycle

Third cycle

Specific prerequisites

Knowledge in advanced Internet technique, 7.5 higher education credits, equivalent completed course IK2215. Knowledge and skills in programming in C++, Java or Python, 6 higher education credits, equivalent completed course DD1310-DD1319/DD1331/DD1337/DD100N/ID1018.

Language of instruction

The language of instruction is specified in the course offering information in the course catalogue.

Intended learning outcomes

After passing the course, the student should be able to:

- describe and analyze an example of a service using virtualization of network functions (NFV)
- list and analyze an example of a service using virtualization of network functions (NFV)
- explain and differentiate the important advantages of remote direct memory access (RDMA) and how it operates
- analyze methods for performing network I/O directly to/from graphics processors (GPU)
- explain methods to improve inference latency in detail
- describe and analyze the role of the load balancing for servers
- describe and analyze examples of current research problems with serving machine learning workloads in data centers
- apply the knowledge from the course to analyze your research domain, demonstrating its practical use and impact
- analyze the connections between the course material and your own research, emphasizing their significance
- argue for the validity of these connections, providing clear and evidence-based reasoning

Course contents

Network functions. Virtualisation. Kernel bypass technologies (e.g. DPDK) for networks with over 100 gigabits per second. Offloading to Smart Network Interface Cards (SmartNIC). Fast networking
with little or no CPU intervention using remote direct memory access (RDMA). Network aspects of machine learning inference using graphics processing units (GPUs). Load estimation and load balancing. Request for dispatch and scheduling. Efficient, large-scale machine learning inference.
Inference with large language models (LLM).

Examination

- PRO1 - Project assignments, 2.5 credits, grading scale: P, F
- SEM1 - Paper summaries, 2.5 credits, grading scale: P, F
- TEN1 - Written exam, 2.5 credits, grading scale: P, F

Based on recommendation from KTH's coordinator for disabilities, the examiner will decide how to adapt an examination for students with documented disability.

The examiner may apply another examination format when re-examining individual students.

If the course is discontinued, students may request to be examined during the following two academic years.

Ethical approach

- All members of a group are responsible for the group's work.
- In any assessment, every student shall honestly disclose any help received and sources used.
- In an oral assessment, every student shall be able to present and answer questions about the entire assignment and solution.