Job Posting:175029 - Position: W26 Intern Researcher - AI Computing System 175029

Co-op Work Term Posted:	2026 - Winter
App Deadline	11/11/2025 09:00 AM
Application Method:	Through UBC Science Co-op
Posting Goes Live:	11/04/2025 11:50 AM
Job Posting Status:	Approved

ORGANIZATION INFORMATION

Organization	Huawei Canada
Address Line 1	19 Allstate Pky
City	Markham
Postal Code / Zip Code	L3R 5A4
Province / State	Ontario
Country	Canada

JOB POSTING INFORMATION

Placement Term	2026 - Winter
<b> Job Title <b>	W26 Intern Researcher - AI Computing System 175029
Position Type	Co-op Position
Job Location	Vancouver, BC
Country	Canada
Duration	8 or 12 months
Work Mode	To be confirmed
Salary Currency	CAD
Salary	0.0 per hour for 0 Major List
Salary Range $	$78,000 to $150,000

Job Description

About the team:

The Advanced Computing and Storage Lab, currently a part of the Vancouver Research Centre, aims to explore adaptive computing system architectures to address the challenges posed by flexible and variable application loads in the future. It assists in ensuring the stability and quality of training clusters, constructs dynamic cluster configuration strategy solvers, and establishes precision control systems to create stable and efficient computing power clusters. One of the lab's goals is to focus on key industry AI application scenarios such as large model training/inference, based on key technologies like low-precision training, multi-modal training, and reinforcement learning, responsible for bottleneck analysis and the design and development of optimization solutions, thereby improving training and inference performance as well as usability.

About the job:

- Aiming at key industry AI application scenarios such as large model training and inference, this role focuses on advancing performance, efficiency, and usability of AI systems on the Ascend platform. The work involves low-precision training, multimodal optimization, reinforcement learning, and training resource optimization to address system bottlenecks and deliver next-generation AI capabilities.
- Responsible for design and development of optimization solutions for AI training and inference systems, with a focus on FP8 optimization, RL-driven training agents, multimodal reinforcement learning or next-generation multi-modal understanding &

generation.
•Combine AI algorithm requirements with system-level architectural optimization in computing, I/O, scheduling, and precision control to improve performance.
•Build stable, efficient AI training clusters, leveraging dynamic cluster configuration and precision control to ensure scalability and reliability.
•Develop software frameworks, operator libraries, acceleration libraries, and system-level optimizations for NPU platforms to accelerate large-model AI training.
•Drive innovation in optimizing large-model training and inference with low-precision training, parallel strategy tuning, and reinforcement learning.
•Grasp the latest research progress and technological trends in AI computing cluster architecture design, training acceleration, and inference acceleration across academia and industry to strengthen the competitiveness of AI computing cluster systems.
The target annual compensation (based on 2080 hours per year) ranges from $78,000 to $150,000 depending on education, experience and demonstrated expertise.

Job Requirements

About the ideal candidate:

• Ph.D or Masters student in Computer Science, Computer Engineering majors in artificial intelligence, computer science, software, automation, electronics, communications, robotics, etc.
•Familiar with the common model structures of large models such as Deepseek and Llama, and have basic technical accumulation in large model training and inference optimization in the fields of LLM, MoE, multimodality, etc.
•Familiar with the hardware architecture and programming system of AI accelerators such as GPU/NPU, and have experience in optimizing AI systems with coordinated software and hardware cores.
•Those with any of the following experience is an asset:

1. 1) Solid programming foundation, familiar with Python/C/C++ programming languages, good architecture design and programming habits
2. 2) Ability to work independently and solve problems, good at communication, willing to cooperate, keen on new technologies, good at summarizing and sharing, and like hands-on practice
3. 3) Experience in the development of AI training frameworks and AI reasoning engines, or algorithm hardware and related experience
4. 4) Strong research capabilities in new technologies and new architectures, can quickly track and gain insights into the most cutting-edge AI technologies in the industry, and lead the continuous leadership of system architecture innovation.

Citizenship Requirement N/A

APPLICATION INFORMATION

Application Procedure Through UBC Science Co-op
Special Application Instructions

Please apply both through SCOPE and through the employer's website.

Application Link: Huawei Technologies Canada Co., Ltd. - Engineer - Cloud Networking
Applications are accepted on a rolling basis and the posting may be expired at any time by the employer as submissions are received.
Students should submit their applications as soon as they are ready.