Ph.D. Thesis Defense

Zhiyan Lu

Computer Science

Friday, February 24

2 p.m. • Rekhi 101 and Zoom

Talk Title: Optimize the memory systems for modern workloads

Abstract: Both new non-volatile memory (NVM) and emerging deep neural networks (DNN) inferences pose different challenges. Maintaining crash consistency exposes memory update operations on the execution critical path. The DNN inference execution on DNN accelerators suffers due to the high memory bandwidth requirement. This proposal will address these two challenges.

The logging operations, required by the crash consistency, lead to severe performance overhead. To reduce the log request persistence time, we propose a load-aware log entry allocation (LALEA) scheme that allocates log requests to the address whose bank has the lightest workload. To address the intra-record ordering issue, we pro- pose to buffer log metadata (BLOM) in a non-volatile ADR buffer until its log can be removed. Moreover, the recently proposed LAD introduces unnecessary logging operations on multicore CPU. To reduce these unnecessary operations, we design two-stage transaction execution(TSTE) and virtual ADR buffers(VADR).

The low response time of DNN inferences and its high computational intensity necessitates the execution of inference on hardware accelerators in data centers. The computing engines of accelerators can complete massive computations quickly. How- ever, the loading of data from off-chip memory takes longer than the computing on them, hurting the performance. This proposal plans to tackle this issue by considering the different memory bandwidth requirements of different DNN models and branches of a DNN model.

Read more and find the Zoom link at blogs.mtu.edu/computing.

Computing [MTU]
Department of Computer Science

1885 Michigan Tech