

《计算平台 xPU 全栈加速研究思考》学术报告

2022年5月14日 15:25, 在CCF 15th ICSS2022大会的算力网络论坛, 将开展《计算平台 xPU 全栈加速研究思考》报告, 欢迎参加!

内容摘要: 针对各种 GPU/DPU/NPU/IPU 等各种新型 xPU 计算平台, 介绍其全栈加速架构和编程模式, 分享当前的研究热点和痛点, 以及高效能服务器和存储技术国家重点实验室异构加速团队在代表性研究热点的创新实践, 探讨启发如何与时俱进的开展产学研界联合学术创新和人才培养问题。

参会方式: Zoom ID: 834 9183 6227 (Room 3),
<https://us02web.zoom.us/j/83491836227>

报告人简介: 阚宏伟, 高效能服务器和存储技术国家重点实验室异构加速首席科学家, 中国矿业大学人工智能研究院兼职教授。主要从事面向云边端的可重构分布式异构加速计算平台的研究, 人工智能原创性基础研究, 聚焦在硬件加速平台、CPU-FPGA-xPU 的全栈加速、加速器资源池化/云化/热迁移等方向, 和人工智能算法模型与特征度量方向。先后主持重大国内外研究课题 6 项, 论文专利 100 余项, 多项成果应用于重大装备或授权国内外 TOP10 知名企业, 并填补行业空白。



xPU全栈异构计算运行机制全景

Host Source Code Device Source Code

OpenCL like Runtime Lib

Execute Program Device Program Device Parallel Runtime Lib

Clang/LLVM

Workgroup Transformation Passes

Device Kernel Translation

Codegen

- ① 首先由fetch模块取回一条指令, 并将该指令, 以及执行该指令的所有线程信息传递给decode模块
- ② decode模块解码指令, 将解码后的指令, 以及执行该指令的所有线程信息传递给issue模块;
- ③ issue模块根据线程信息, 从不同的GPR (通用寄存器) 组中取出对应不同线程的数据, 发送给Execute模块
- ④ Execute模块同时计算所有线程的所有数据, 并将计算结果发送给Commit模块

icache dcache

threads scheduled instruction decode gpr set 1 thread 1 commit result to gpr

Fetch Decode Issue Execute Commit

Core

核数: x86/ARM经典100--300核; Risc-V数万核 (FPGA实现6000核), 各种xPU支持几百万线程, 工作频率小于4G

inspur

13

Registration Program WeChat Group