대학원 인공지능학과 & AI융합네트워크학과

Artificial Intelligence & AI Convergence Network Colloquium

TITLE Challenges of Building Inference Chips for Data Centers

When : 2021.11.3.(WED) P.M.4:30~

Where : Zoom

링크 https://zoom.us/j/99320661895?pwd=Yy9Xa0lYZW1CT2hXRHlKUzhGNkpsZz09

회의 ID: 993 2066 1895, 암호 : 3898

Speaker : Hanjoon Kim(FuriosaAI CTO)

Abstract : According to papers from hyperscale datacenters[1][2], the demand for deep learning inference in data centers is growing rapidly. While energy efficiency is important to reduce TCO (Total Cost of Ownership), high performance is also essential to serve large models in production. Hyperscalers, on the other hand, emphasized the importance of programmability and flexibility for inference accelerators to track DNN progress[1]. In order to build a production accelerator for all these challenging requirements, instead of building a chip that is optimized for a specific model, the architecture should expose the raw ability to maximize parallelism and energy efficiency of DNN models to the software with well-defined abstraction. Software stack should also exploit every parallelism and energy efficiency for each operator and model. To accomplish such cross-layer optimizations within algorithm, architecture, and software, small and excellent teams must communicate deeply and closely, and design methodologies and the infrastructures must support these communication structures.

[1] Norman P. Jouppi et al., Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product, ISCA'21

[2] Michael Anderson et al., First-Generation Inference Accelerator Deployment at Facebook, https://arxiv.org/abs/2107.04140

BIO : Hanjoon Kim is co-founder and CTO of FuriosaAI Inc. He is leading AI chip development, setting the engineering direction and technology vision. Prior to FuriosaAI Inc, he lead the development of memory-centric accelerator architecture targeting hyperscale datacenter at Samsung Memory. He holds a PhD in Computer Science from KAIST.

Contact : 정보통신대학 소프트웨어학과 김상훈 교수
(sanghoonkim@ajou.ac.kr)

☎ 세미나 문의 : 031 - 219 - 3647, 3898

아주대학교

BK21 FOUR