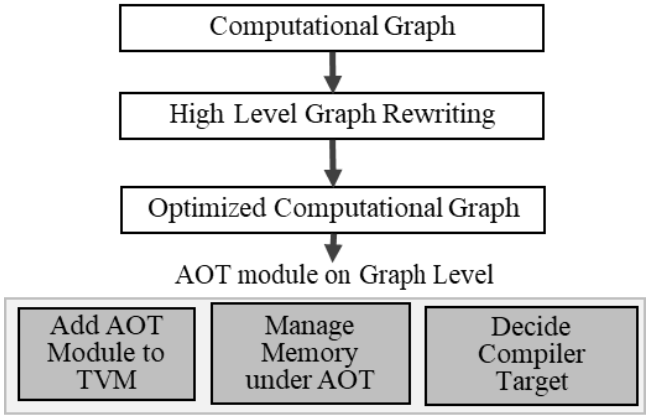
Towards Optimized Tensor Code Generation for Deep Learning on Sunway Many-Core Processor

Mingzhen LI, Changxi LIU, Jianjin LIAO, Xuegui ZHENG, Hailong YANG , Rujun SUN, Jun XU, Lin GAN, Guangwen YANG, Zhongzhi LUAN, Depei QIAN

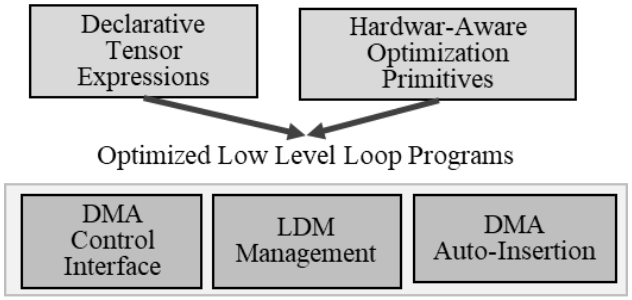
Frontiers of Computer Science, DOI: 10.1007/s11704-022-2440-7

Problems & Ideas

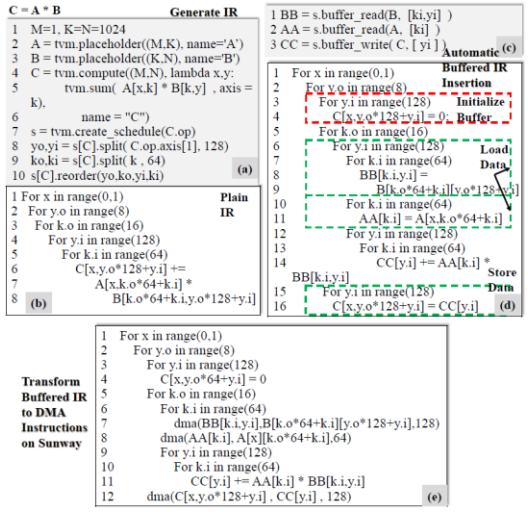
- Problems of optimized tensor code generation on Sunway:
  - The unique compilation environment and architecture features prevent a naive adoption of DL compiler (e.g., TVM) to Sunway.
  - Needs the support of ahead-of-time code generation for MPE/CPE.
  - Needs automatic DMA control, LDM management, and parallelization.
- Ideas: Propose swTVM, a deep learning compiler extending TVM to support optimized tensor code generation for deep learning on Sunway processor automatically.



AOT module at graph level



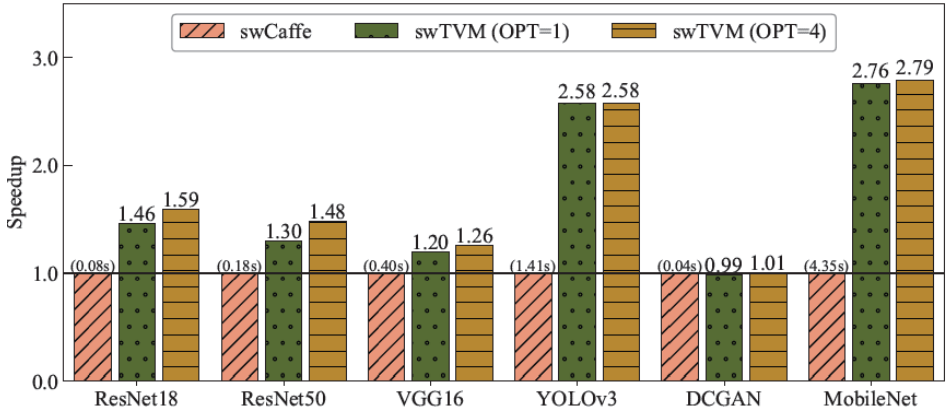
Optimization and code generation



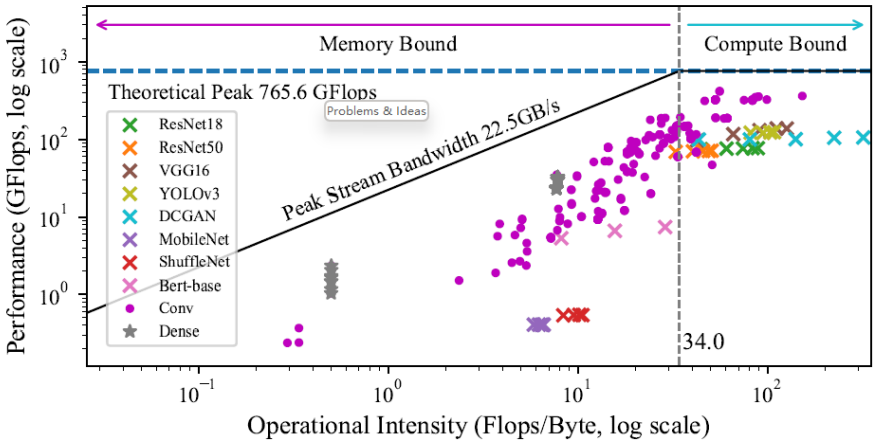
A matmul example

Main Contributions

- Contributions:
  - We implement the ahead-of-time code generation, that produces different compilation targets for MPE and CPE as well as manages the function calls between MPE and CPE efficiently; and we manage the intermediate memory space for each tensor operation globally;
  - We apply several optimizations to the tensor operations regarding the unique architecture features on Sunway, including a DMA control interface, a LDM management mechanism, and DMA instruction inserting mechanism.



The end-to-end performance of swTVM compared to the state-of-the-art DL framework swCaffe.



The roofline analysis of swTVM on eight representative DL models when bs=1.