Parallel Sparse Deep Learning Operators on Lightweight RISC-V Processors

Marco Bertuletti1, Tim Fischer1, Yichao Zhang1, Luca Benini1,2
1Integrated Systems Laboratory, ETH Zurich; 2Università di Bologna, Italy;

1. The Snitch

Single-stage pipeline
Support RISCV32MAFD
Double precision FPU

The FPU Sequencer allows
offloading of the FP instructions
without stalling the integer core

SSRs allow prefetching of data

2. Parallelization of Sparse kernels

SoftMax

Matrix Multiplication

2D Convolution

3. Sparse kernels speedup

SW Implementation of Sparse kernels:
Quasi-ideal parallel speedup

CSR vs DENSE SPEEDUP

x7.3 - x3.0 - x1.7 1 core
x7.0 - x1.1 - x1.5 8 cores

Further improvement via prefetching