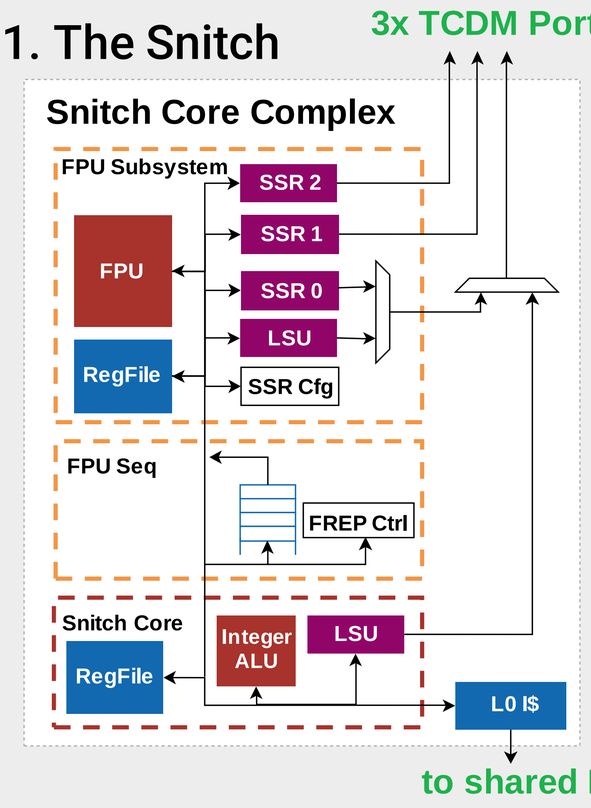
# Parallel Sparse Deep Learning Operators on Lightweight RISC-V Processors

Marco Bertuletti[1], Tim Fischer[1], Yichao Zhang[1], Luca Benini[1,2]
[1]Integrated Systems Laboratory, ETH Zurich; [2]Università di Bologna, Italy;

## 1. The Snitch

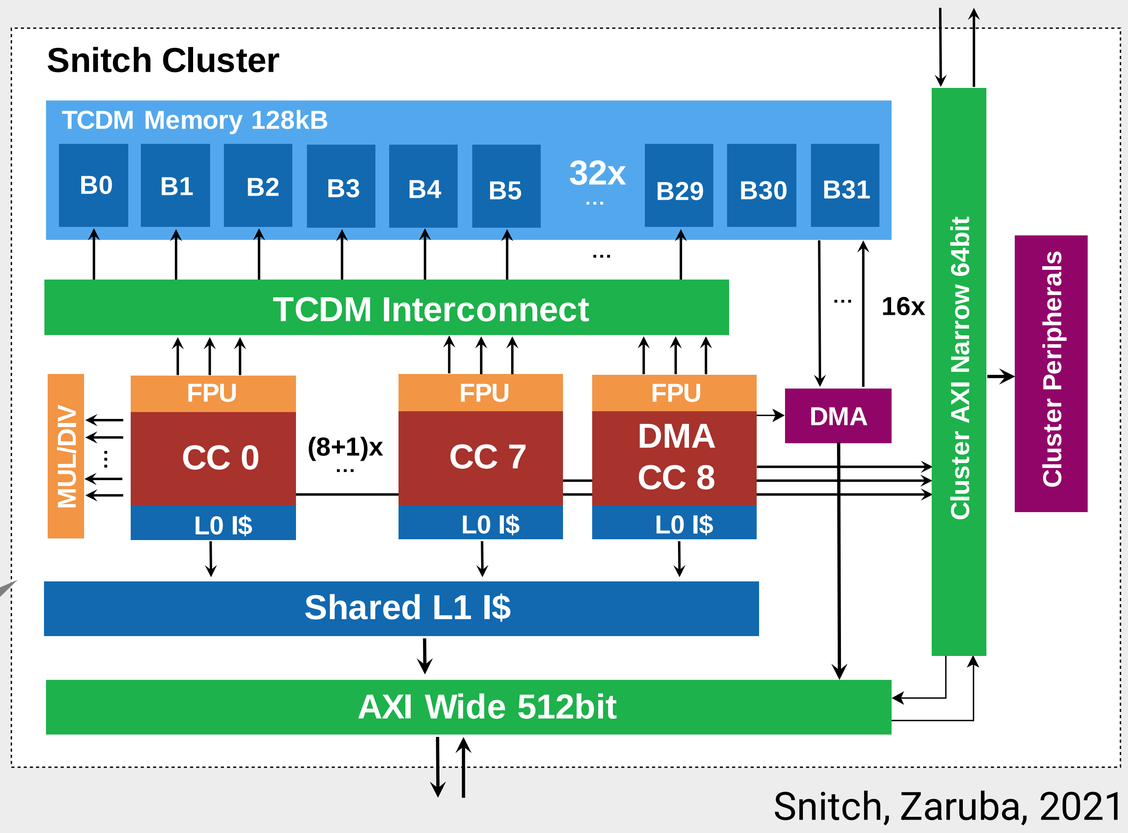



Single-stage pipeline
Support **RISCV32IMAFD**
Double precision FPU

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

**SSRs** allow prefetching of data

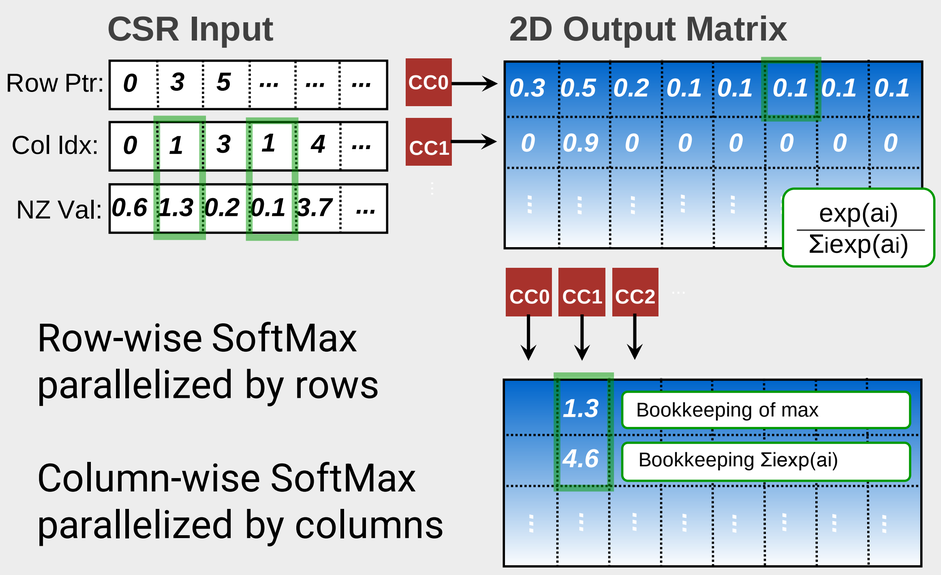
Snitch Cluster: 8 compute PE + **DMA** core




Snitch, Zaruba, 2021

## 2. Parallelization of Sparse kernels

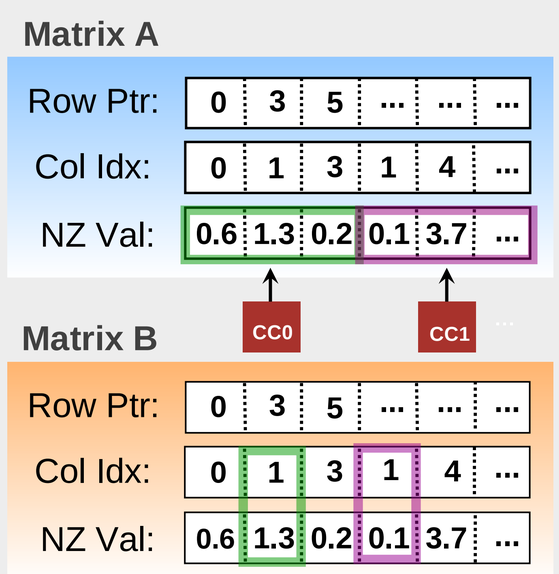
### SoftMax



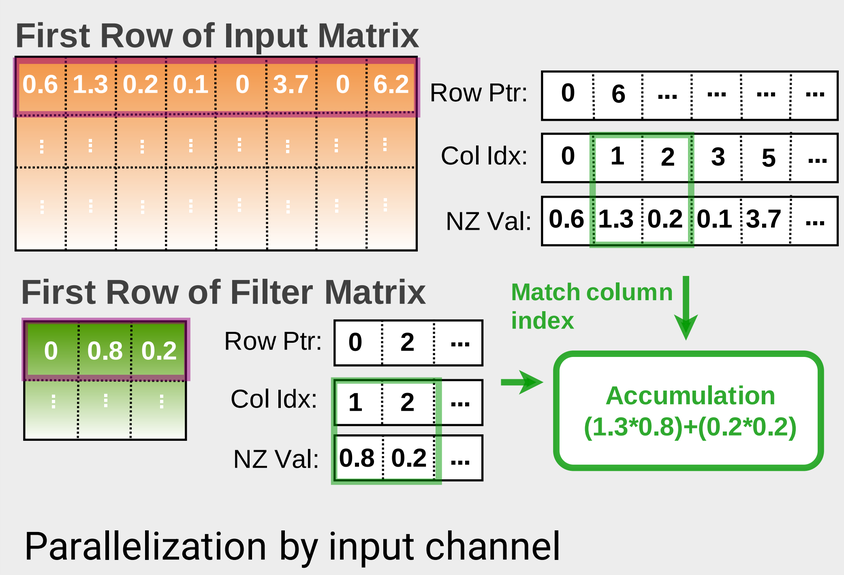

Row-wise SoftMax parallelized by rows

Column-wise SoftMax parallelized by columns

### Matrix Multiplication

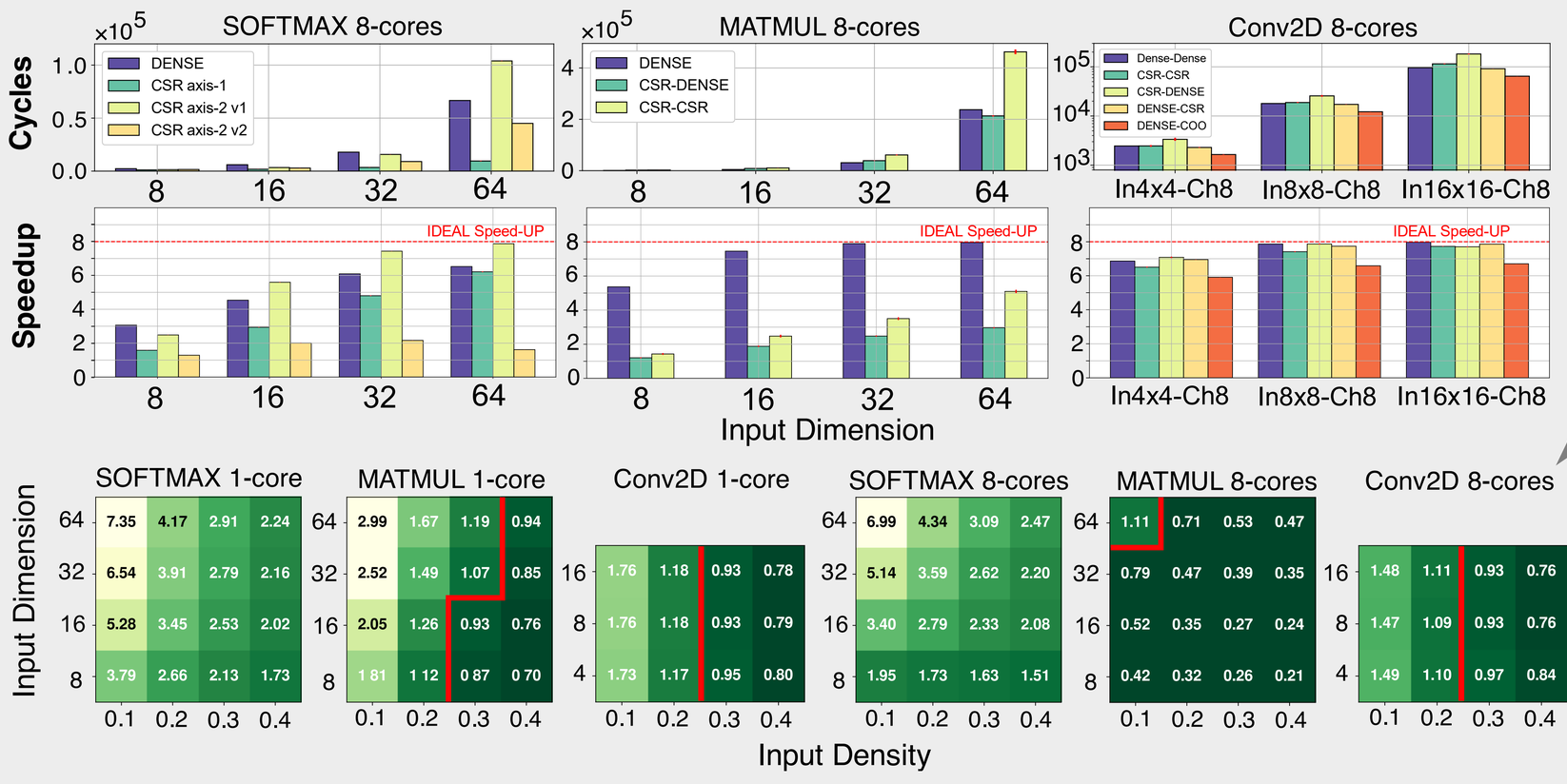



### 2D Convolution




Parallelization by input channel

## 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

When the density of the inputs is high we still benefit from **reduced memory footprint**

Further improvement via prefetching

mbertuletti@iis.ee.ethz.ch pulp-platform.org @pulp_platform