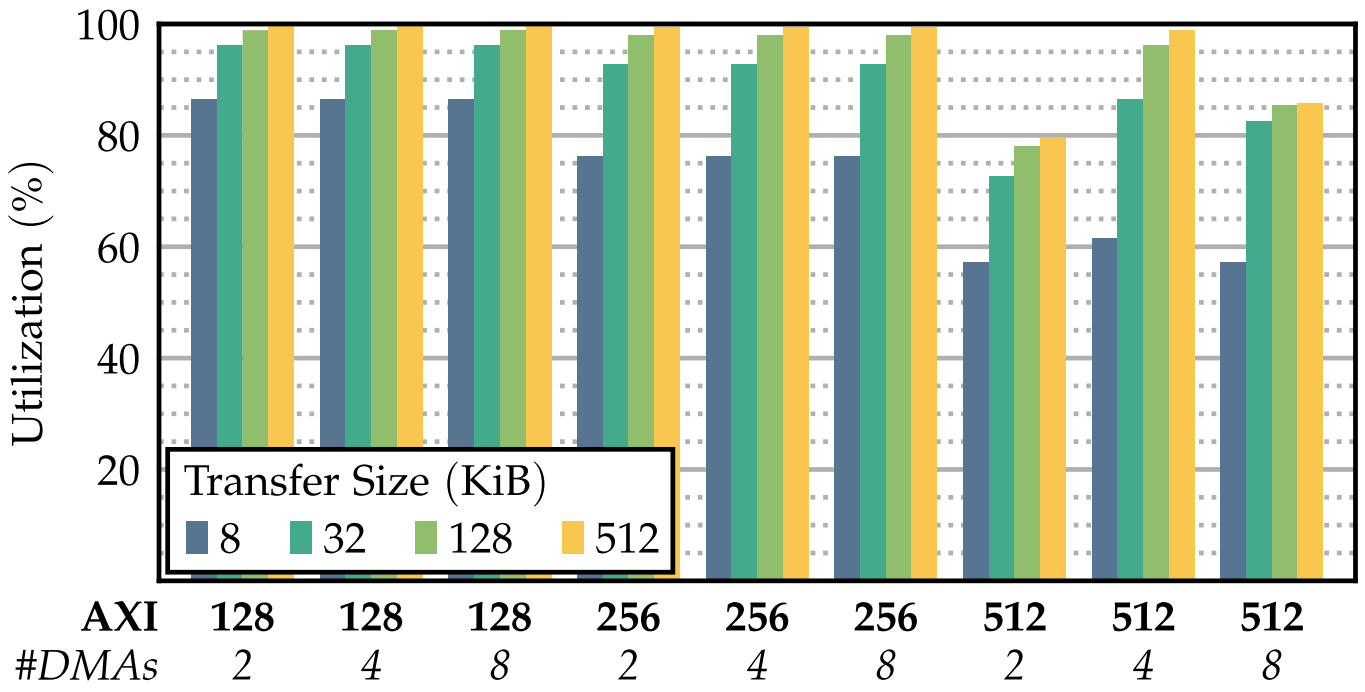
ETH zürich

MemPool meets Systolic

Samuel Riedel
Matheus Cavalcante
Prof. Luca Benini

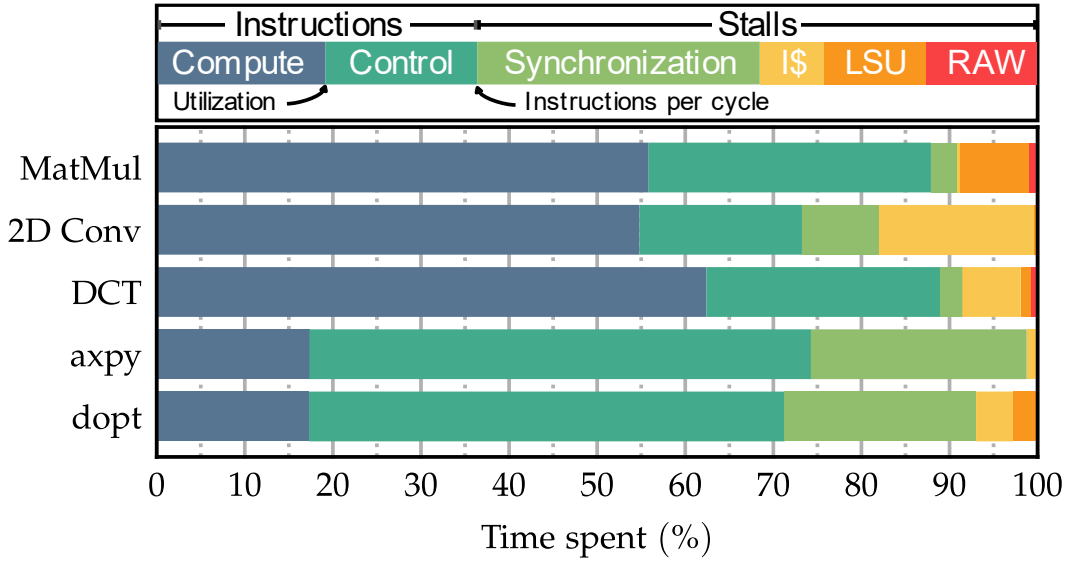
DMA Benchmarking

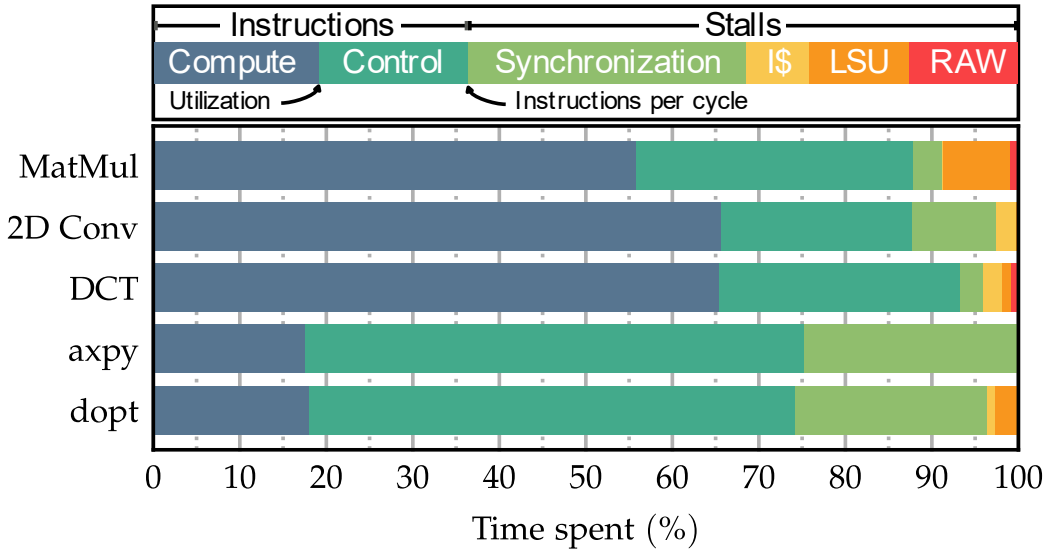


- Measure the DMA's performance for different configurations
- Consistently measure a high utilization for narrow AXI widths
  - Even for small transfers
- Only the 512-bit case shows a clear distinction between the number of DMA backends
  - The configuration with four backends per group achieves 98% utilization

Kernel Benchmarks

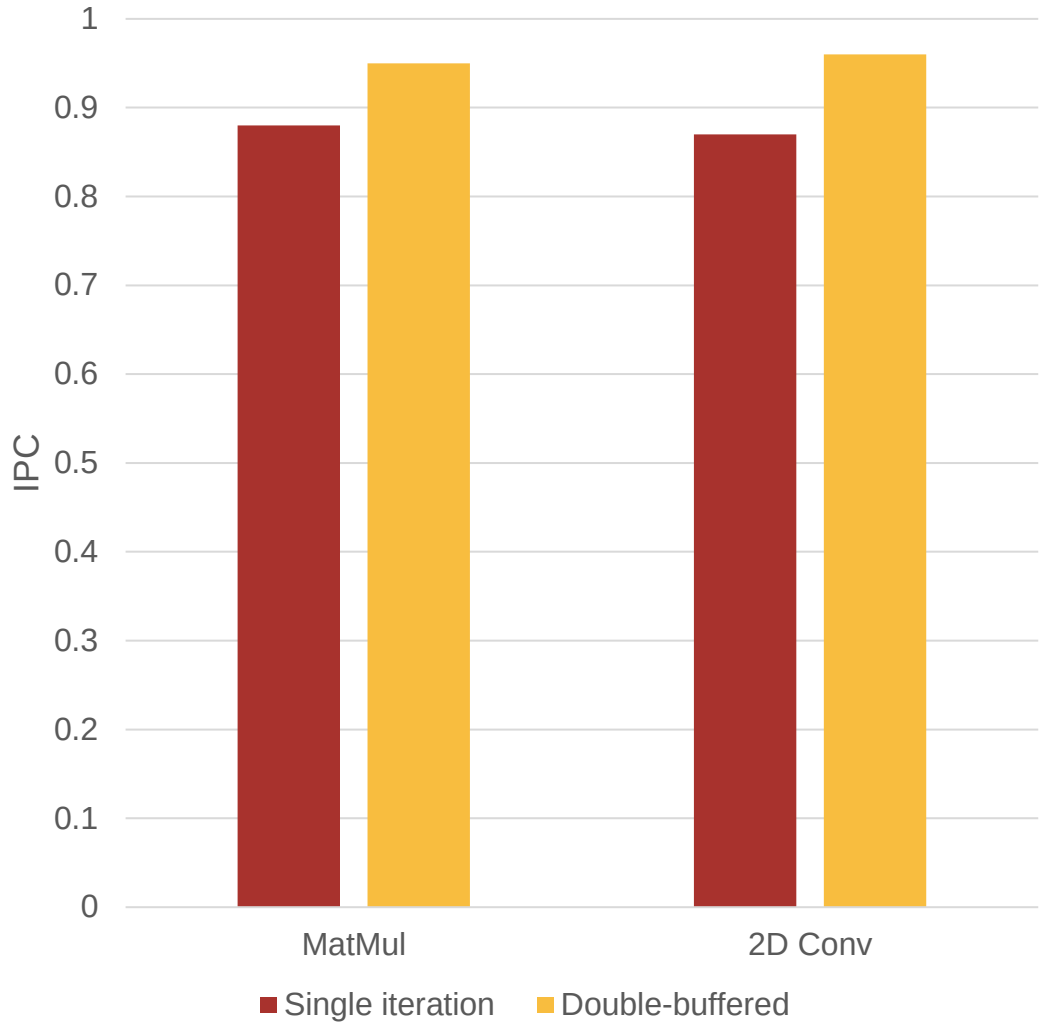
- Where do we lose time?
  - Load-store architecture: Control vs compute instructions
  - Manycore system: Synchronization overhead
  - Nonidealities: Architectural stalls
    - Instruction
    - Interconnect
    - RAW
- Measured with **cold** caches, one iteration
- Measured with **hot** caches



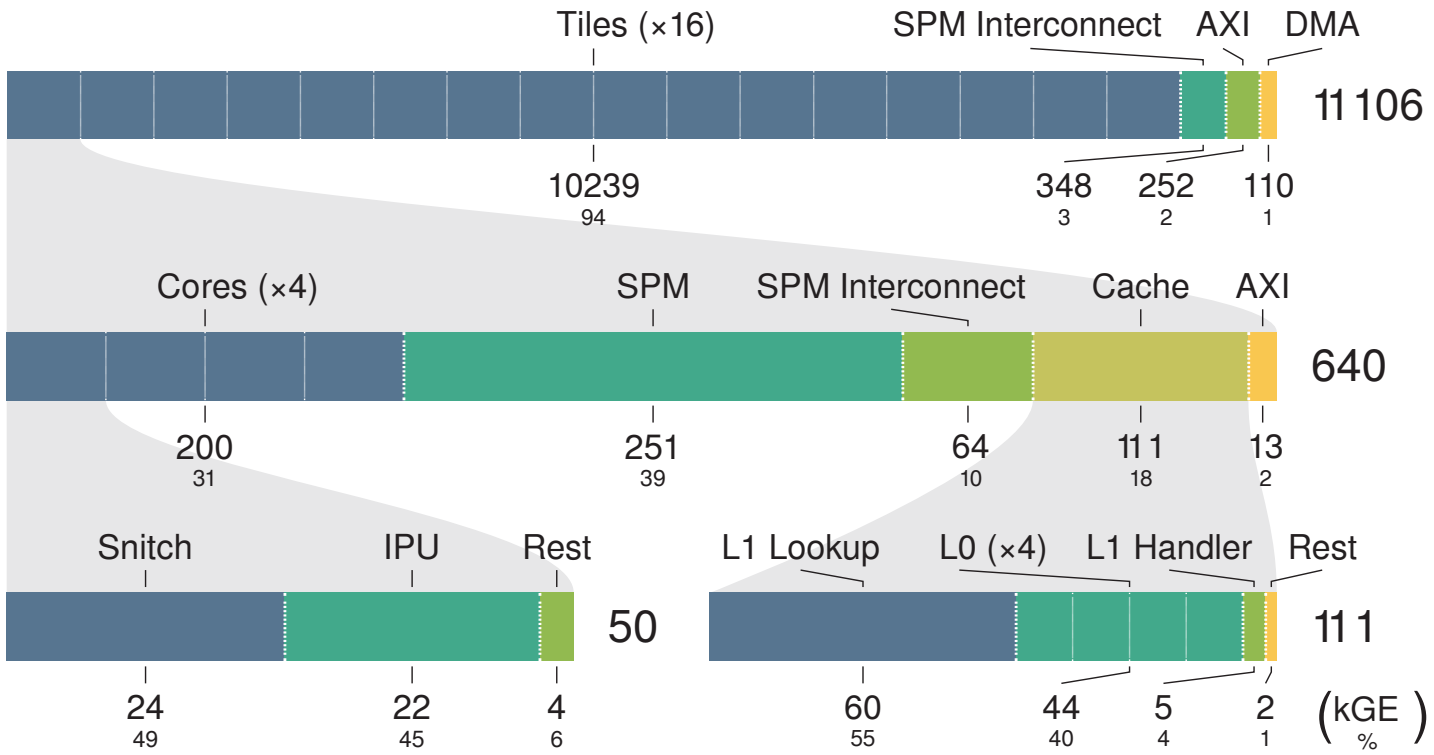


Double-buffered benchmarks

- Implement benchmarks in a double-buffered fashion
  - Axpy, dotp is memory bound
  - Matmul & 2D convolution work very well
    - 0.95/0.96 IPC
    - 307/369 OP/cycle
    - 8/10 % speedup
- Still analyzing and extracting those results



Physical implementation results



Systolic MemPool

- Cleaning up software queues implementation
  - Improved parametrization
- Work in progress:
  - Cleaning up the 2D convolution
  - Merging the Queue push/pop adapter
    - Will be merged as a separate unit that can be instantiated for the systolic MemPool