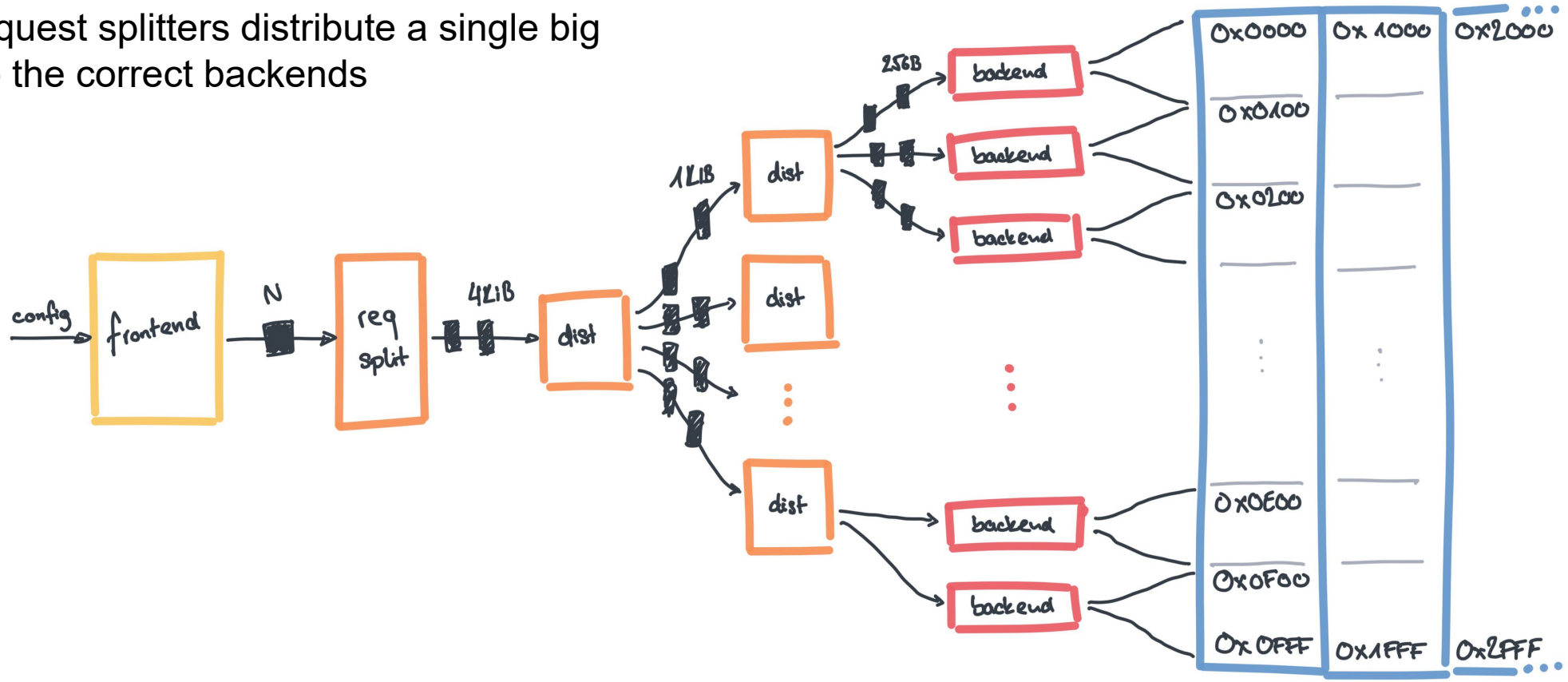
ETH zürich

# MemPool meets Systolic

Samuel Riedel
Matheus Cavalcante
Prof. Luca Benini

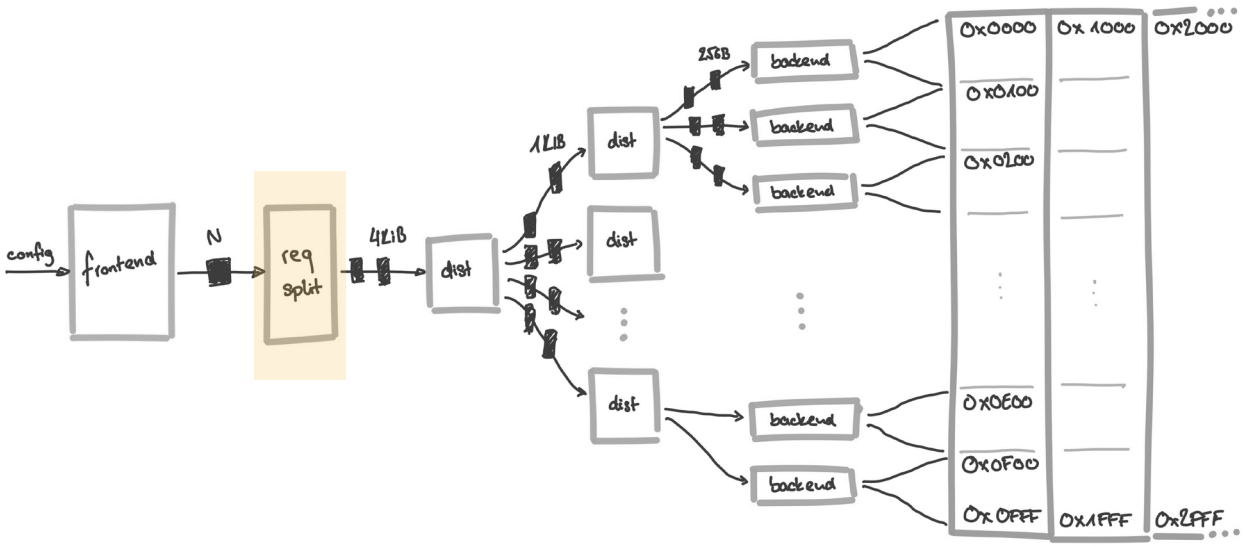
# DMA

- Implemented preliminary DMA
  - Configurable number of data movers per group
  - Tree of request splitters distribute a single big transfer to the correct backends



# DMA: What is missing

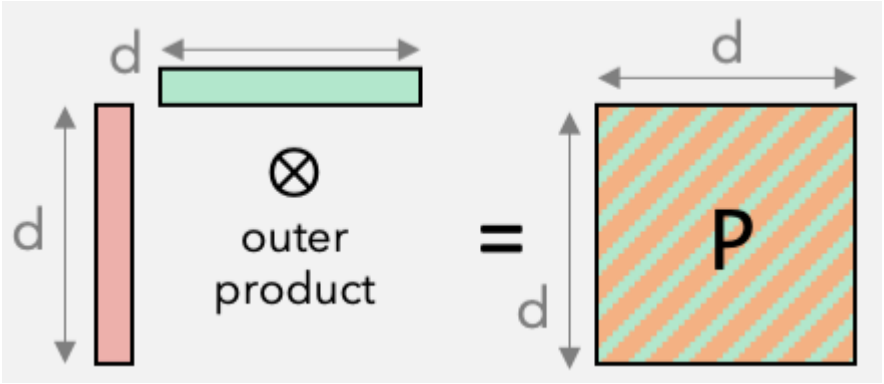
- We care about:
  - Request splitting
  - Not impacting the critical path
- Will be done later
  - Handling corner cases of not-word-aligned transfers
  - Multiple outstanding requests
  - Fast notification/communication between cores and DMA



# Matrix Multiplication

- Go from a 2x2 kernel to 4x4
  - Requires all 30 available registers
  - Fully written in assembly
- Benchmark a 256x256 matmul on 256 cores
  - Including a final barrier

- 143 MACs/cycle
- 56 % MAC unit utilization (ideal is 64%)
- 65 GOPS/W



# Convolution

- Start reusing data
  - Compute four output pixels per iteration
  - Fully written in assembly
- Benchmark a 1024x96 image on 256 cores
  - Including a final barrier

- 168 MACs/cycle
- 66 % MAC unit utilization (ideal is 77%)
- 91 GOPS/W