ETH zürich

MemPool meets Systolic

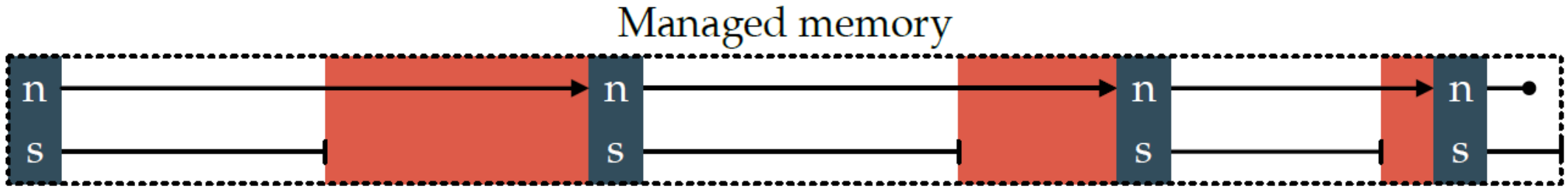
Samuel Riedel
Matheus Cavalcante
Prof. Luca Benini

Systolic array in Software

- Prerequisites
  - Memory allocator
  - Sequential region
  - Software queues
- Matrix multiplication working
  - Start optimizing it

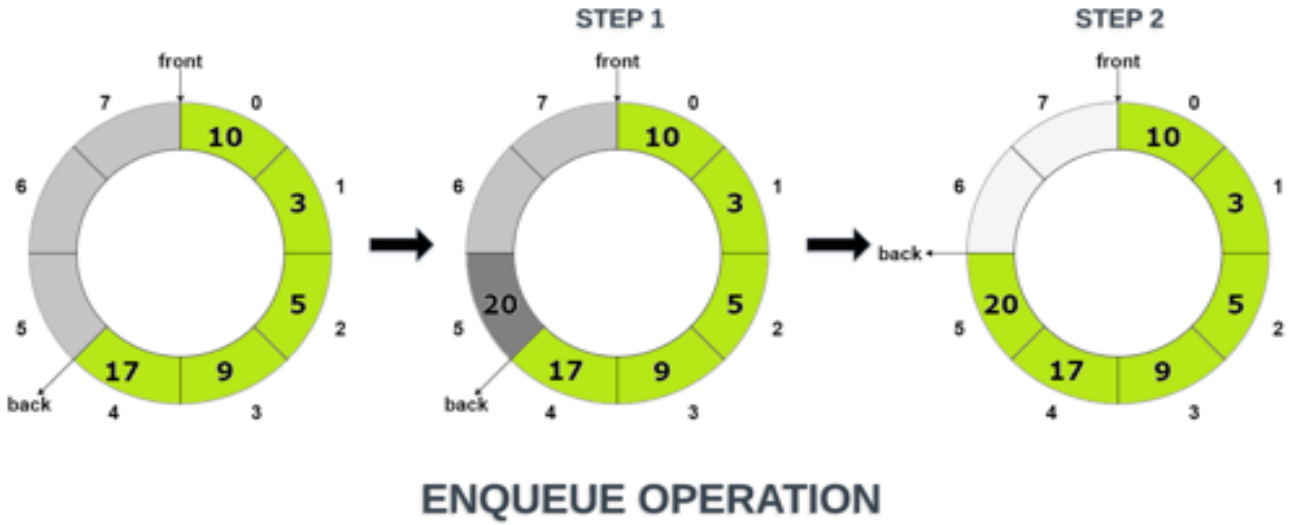
Memory Allocator

- Linked-list based
  - List of free elements with the respective size
  - Currently not-thread safe we could make it blocking
- Allocating in sequential region?
  - Increase sequential size in MemPool
  - Add one allocator per tile



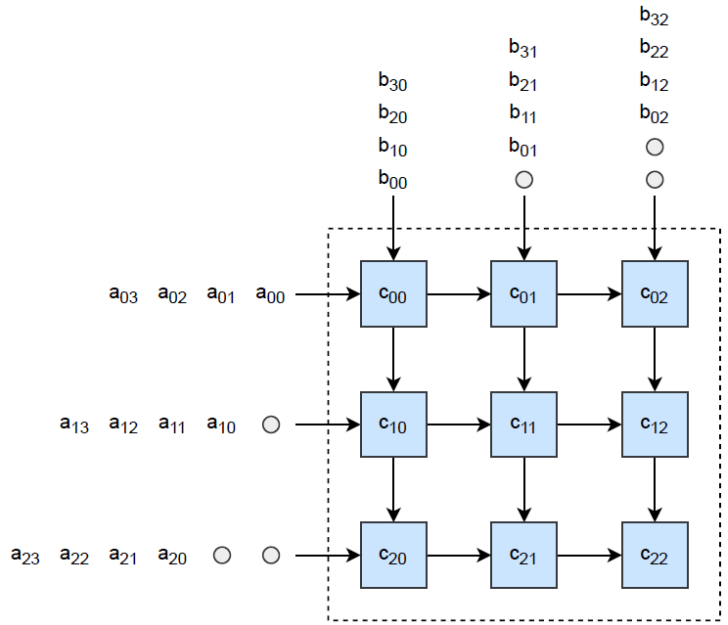
Software Queues

- Array-based/Circular buffer
  - We expect few and a constant number of queue elements
  - Alleviate allocation/free overhead
- Single producer/consumer
  - Not thread-safe for now



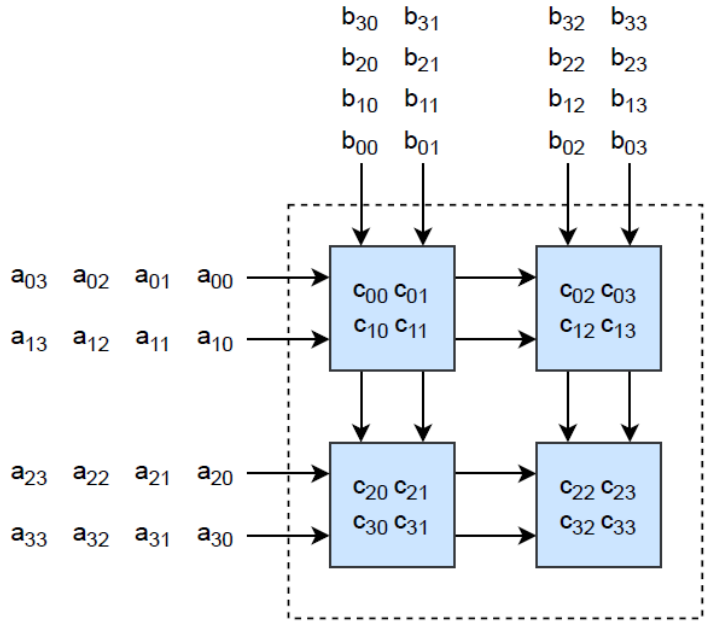
Matrix Multiplication Structure

- Initial setup
  - Allocate two input queues per core
- Each core does:
  - 2 pop
  - 1 MAC
  - 2 push
- A lot of corner cases and boundary conditions
  - Edge and corner cores are special
  - 9 different cores
  - Matmul must be divisible by array size



Matrix Multiplication Performance

- Huge delay due to queue overhead
  - 4 queue operations per MAC
- Make queues local to the consumer
  - Gain roughly 3% speedup
- Try different allocations
  - Cores in one tile are allocated column wise
- Queues pop a single element
  - Let's pop/push multiple elements at once
  - Increase computational intensity
  - TBD



TeraPool: Reaching the TOPS with MemPool

- Bringing the core-count of MemPool to the thousand
  - Easiest solution: just increase the core-count of each tile to 16
  - Each group would have 256 cores:
    - Each group of TeraPool is the size of MemPool

- Currently solving a few issues
  - 1024 cores means 1024 open files containing the trace of each core
    - Too many open files for Linux :)
  - Verilator has a hard time with TeraPool (despite hierarchical Verilation)
    - QuestaSim works, though
  - Synthesis of a TeraPool group takes an unfeasibly long time
    - Add another hierarchy level to MemPool? (With extra latency)
    - Have 8 groups with 128 cores instead of 4 groups with 256 cores?
      - Need to get TeraPool to run before I can evaluate this

ETH zürich

MemPool meets Systolic

Samuel Riedel
Matheus Cavalcante
Prof. Luca Benini