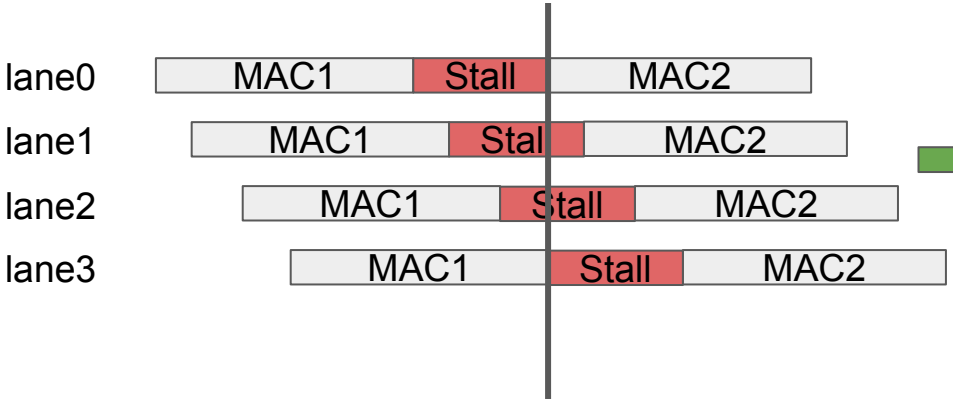
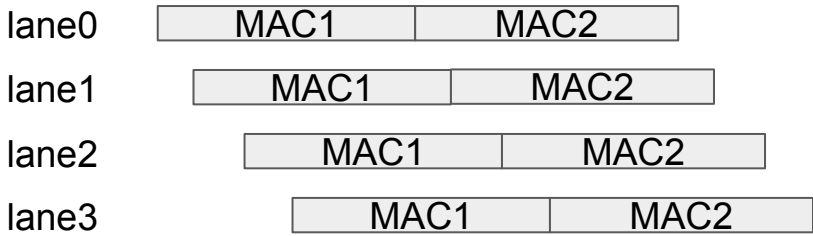
Week 17

Optimization 1

- RAW hazard opt
  - Clear the dependency if MAC1 is finished
  - If MAC2 is writing result, set stall to 0
- Result (64x768x64, 16 lanes):
  - utilization: 74.85% → 89.21%





Optimization 2

- Burst store
  - Before: one AXI request one 32-bit data
  - Now: one AXI request NrLanes 32-bit data
  - Burst Length = NrLanes
  - next_base_addr = cur_base_addr + stride
- Result (64x768x64, 16 lanes):
  - store runtime: 8412 cycles → 318 cycles
  - utilization: 89.21% → 95.88%

Updated Performance

	bc_matmul_opt	bc_matmul	matmul	Change
4 lanes	98.35%	92.28%	65.67%	1.5x
8 lanes	97.94%	85.63%	34.22%	2.8x
16 lanes	95.88%	74.85%	17.03%	5.6x

Other Matrix Sizes

	bc_matmul_opt	matmul	Change
32x32 * 32x32	83.30%	56.97%	1.46x
64x64 * 64x64	92.72%	82.62%	1.12x
128x128 * 128x128	96.02%	94.13%	1.02x

4 lanes

Backend Results (Lane)

	Baseline	New Ara	Change
Area (utilization)	89.30%	89.88%	+0.58%
MAX Frequency	950 MHz (-50.24ps)	804 MHz (-242.29ps)	-15%

Baseline: vector mask insn (28.11.2022)

New Critical Paths

1. bc_valid_i → bc_invalidate_o (in2out)
    a. slack: -242.287
2. bc_valid_i → bc_ready_o (in2out)
    a. slack: -221.211

- Paths have longer latency
    - vmfpu/result_queue_q → sldu_addrgen_operand_o[34]
    - slack: -14.56 → -87.81