Metrics for Evaluating MatMul Arch Options

- Arithmetic Intensity (total int/fp muladd ops per instruction)
- # of RegFile read and write ports (e.g. low-end: 3R1W, high-end:8R4W, hpc: 3R1W?)
  - # RF Cycles = max(total RF rd's, total RF wr's)
  - # of MulAdd's - by default, assume no perf limit below that implied by # Regfile rd/wr ports
- # of memory read accesses and max access size
  - # Read Accesses = # of wide contiguous memory read accesses
  - Assume all the (fewer) write accesses can overlap execution with read accesses
  - Pack/unpack overhead - by default, assume handled for free by new matrix ld/st instructions
- Eval Metrics
  - Arithmetic Intensity / # Read Accesses
  - Arithmetic Intensity / # RF Cycles

A Couple of Simple Option Comparisons

- Assumptions
  - High-performance Apps CPU - VLEN=DLEN=512 and loads up to DLEN=512b-wide
  - Int8 -> int32 matmul (64 int8's/vreg, 16 int32's/vreg)
  - C=AxB+C with M x K x N matmul with M=N (i.e. square result matrix, and max M=N=16)
  - Maximize use of vector regfile to maximize data reuse and arithmetic intensity
  - Maintain C in regfile while load new A's and B's and do new matmul's
- Option 1 (16 x 32 x 16 matmul):
  - C:16x16 stored in 16 vreg's, A:16x32 stored in 8 vreg's, B: 32x16 stored in 8 vreg's
  - AI = 16K, # Read Accesses=32, # RF Cycles=32
  - AI / Read Accesses = **512**, AI / # RF Cycles = **512**
- Option 2A (8 x 64 x 8 matmul):
  - C:8x8 stored in 4 vreg's, A:8x64 stored in 8 vreg's, B: 64x8 stored in 8 vreg's
  - AI = 8K, # Read Accesses=16, # RF Cycles=20 (note: ~½ muladd hardware of Option 1)
  - AI / Read Accesses = **512**, AI / # RF Cycles = **410**
- Option 2B (four 8 x 64 x 8 matmul's):
  - Four C:8x8 stored in 16 vreg's, A:8x63 stored in 8 vreg's, four B: 64x8 stored in 8 vreg's (i.e. do 4 matmuls and 4 B loads, reusing A)
  - AI = 32K, # Read Accesses=40, # RF Cycles=56 (read A from RF once) (note: ~½ muladd hardware of Option 1)
  - AI / Read Accesses = **820**, AI / # RF Cycles = **585**
  - Assuming 2x Rd/cyc and 8R4W and 2.5 GHz => **4 Tops** (limited by # Read Accesses)