ECE284 FA25 Final Progress Report

1. Fill the table below. This helps us to understand your current progress status.

Item	Current Status	Status during Poster Presentation	Note
Part1	Complete	Complete	Verification with Relu was done after the presentation.
Part2	Complete	Complete	base version also includes dual mem buffer and continuous acc / relu SFP
Part3	In Progress (80% Done)	In Progress	Same signal is used to propagate weight and psum_in from one tile to the next . The data which is propagated is decided based on the mode (1- OS and 0- WS) and whether the PE is in execute or load mode. Testbench is also developed, the weights and activation loading is verified in both modes.
Structured Pruning	Complete	Complete	Achieved a accuracy of 90.04 after structured pruning with sparsity of

			0.498 with 50% model size reduction
Activation Sparsity Analysis	Complete		Average activation sparsity was 68.3%, with some layers exceeding 98% zeros. This implies that over two-thirds of MAC operations can be skipped using sparsity-aware gating without accuracy loss.
continuous accumulation	Complete		on-the-fly accumulation calculation instead of running at the end.
dual mem buffer	Complete		Improvement for on-the-fly accumulation: reducing from 9 SRAM sets (1 per KIJ) to 2 SRAM sets (odd and even KIJ)
dual-core (parallel horizontal tile)	Complete		Tiling 2 corelets in 1 core for 2/4 bit switchable SIMD to enable fitting of entire 16ICx16OC 2-bit activation convolution operation. May also just use 1 tile similarly to the original version.