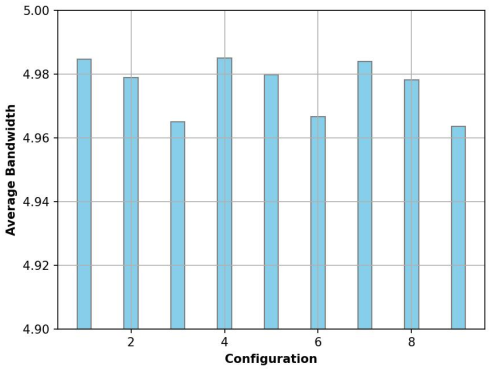

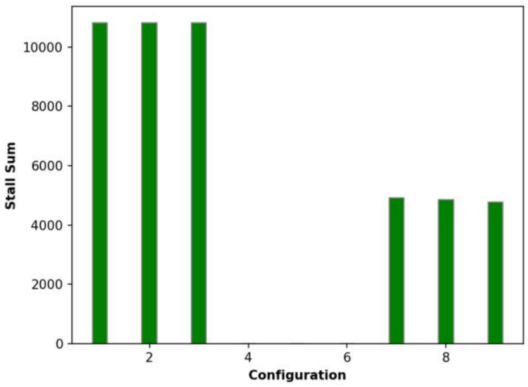

When L1 active_buf_frac is fixed, and we increase the L2 active_buf_frac, the average bandwidth is decreased, because we we more space to serve the systolic array and more data reuse can happen. For the Stall cycles, a balanced partition for the double buffering ensures we not only can store the reused data and also a portion served for the new data coming, this ensures executing on new data would not have waiting cycles for new data.