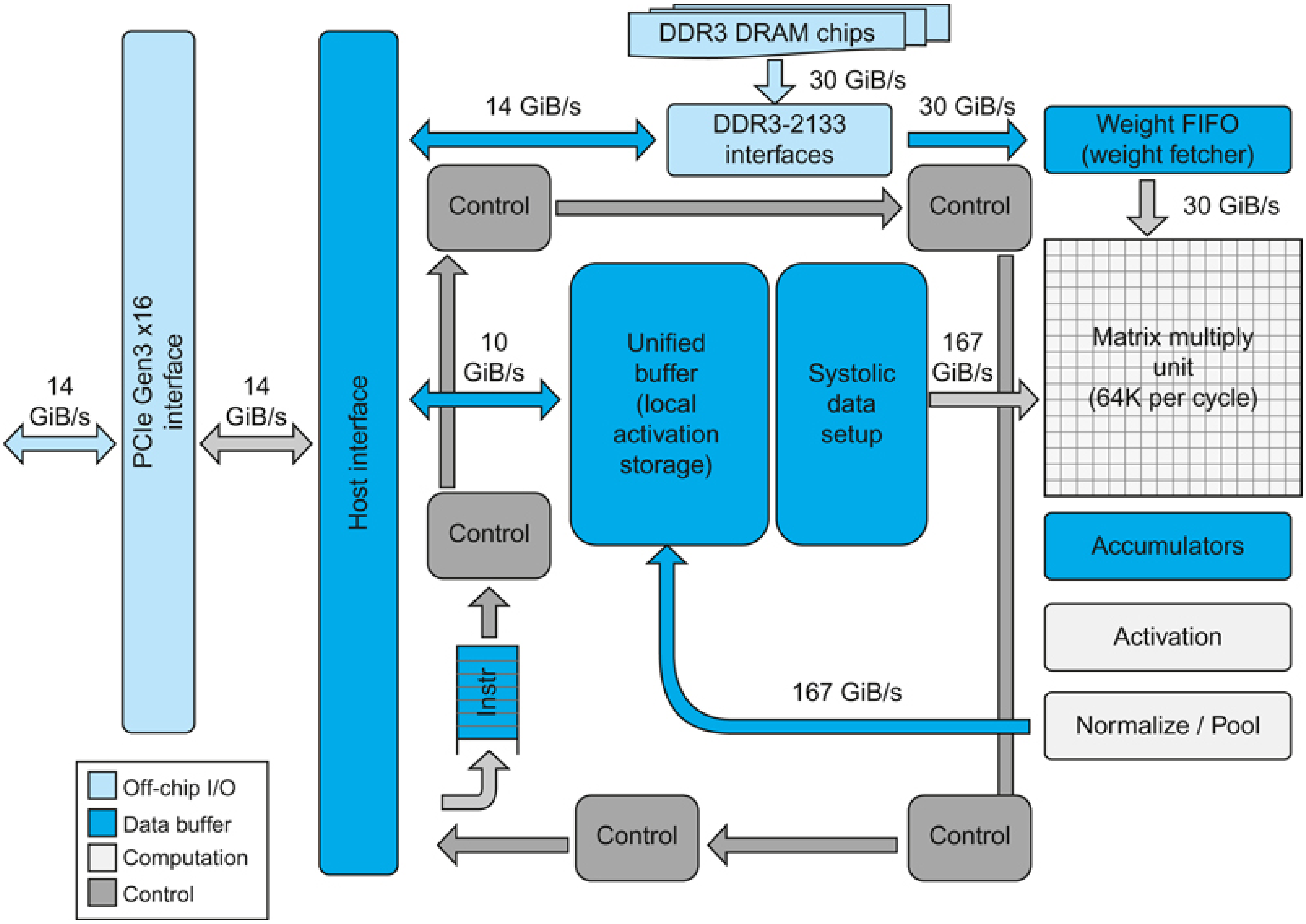

FIGURE 6.13 TPUv1 Block Diagram. The main computation part is the Matrix Multiply Unit in the upper right corner. Its inputs are the Weight FIFO and the Unified Buffer, and its output is the Accumulators. The 24 MiB Unified Buffer is almost a third of the TPUv1 die, and the Matrix Multiply Unit with 65,536 multiple-accumulate ALUs is a quarter, so the datapath is nearly two-thirds of the TPUv1 die. For CPUs, multilevel caches are often two-thirds of the die.