BERT Accelerator Block diagram

Block diagram

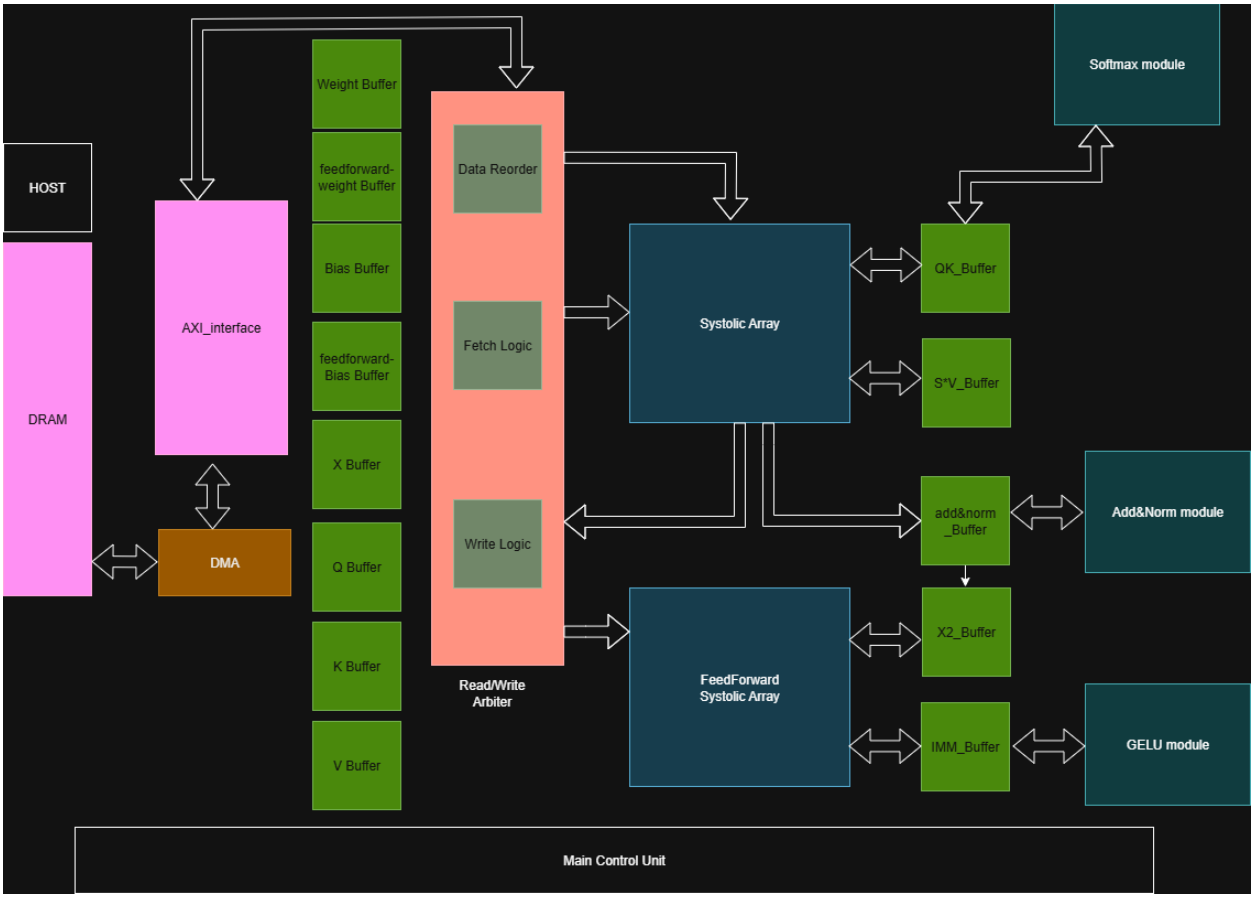

Figure 1: Block diagram

Buffers sizes details:

Buffer	Size
Weight buffer	32*32*2
feedforward-weight Buffer	32*32*2
Bias Buffer	1*768
feedforward-Bias Buffer	1*3072
X Buffer	512*32*2
Q Buffer	512*768
K Buffer	512*768
V Buffer	512*768
QK_Buffer	512*512
S*V_Buffer	512*768
add&norm _Buffer	512*768
X2_Buffer	512*768
IMM_Buffer	512*3072

Note: *2 means double buffering