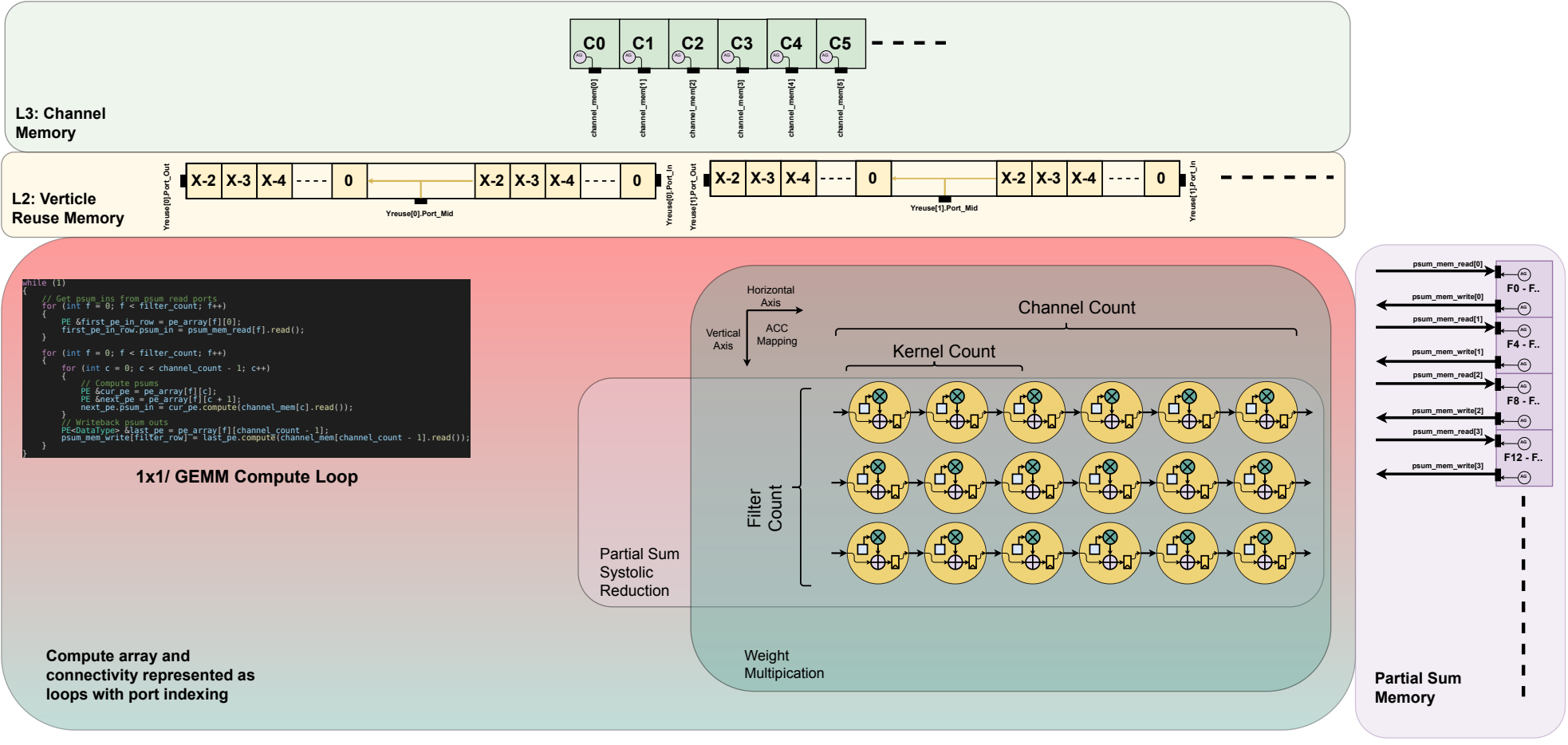
L3: Channel Memory
C0
C1
C2
C3
C4
C5
channel_mem[0]
channel_mem[1]
channel_mem[2]
channel_mem[3]
channel_mem[4]
channel_mem[5]
L2: Verticle Reuse Memory
Yreuse[0].Port_Out
X-2
X-3
X-4
0
Yreuse[0].Port_Mid
Yreuse[0].Port_In
Yreuse[1].Port_Out
Yreuse[1].Port_Mid
Yreuse[1].Port_In
while (1)
{
// Get psum_ins from psum read ports
for (int f = 0; f < filter_count; f++)
{
PE &first_pe_in_row = pe_array[f][0];
first_pe_in_row.psum_in = psum_mem_read[f].read();
}
for (int f = 0; f < filter_count; f++)
{
for (int c = 0; c < channel_count - 1; c++)
{
// Compute psums
PE &cur_pe = pe_array[f][c];
PE &next_pe = pe_array[f][c + 1];
next_pe.psum_in = cur_pe.compute(channel_mem[c].read());
}
// Writeback psum outs
PE<DataType> &last_pe = pe_array[f][channel_count - 1];
psum_mem_write[filter_row] = last_pe.compute(channel_mem[channel_count - 1].read());
}
}
1x1/ GEMM Compute Loop
Horizontal Axis
Vertical Axis
ACC Mapping
Channel Count
Kernel Count
Filter Count
Partial Sum Systolic Reduction
Weight Multipication
Compute array and connectivity represented as loops with port indexing
psum_mem_read[0]
psum_mem_write[0]
psum_mem_read[1]
psum_mem_write[1]
psum_mem_read[2]
psum_mem_write[2]
psum_mem_read[3]
psum_mem_write[3]
F0 - F..
F4 - F..
F8 - F..
F12 - F..
Partial Sum Memory