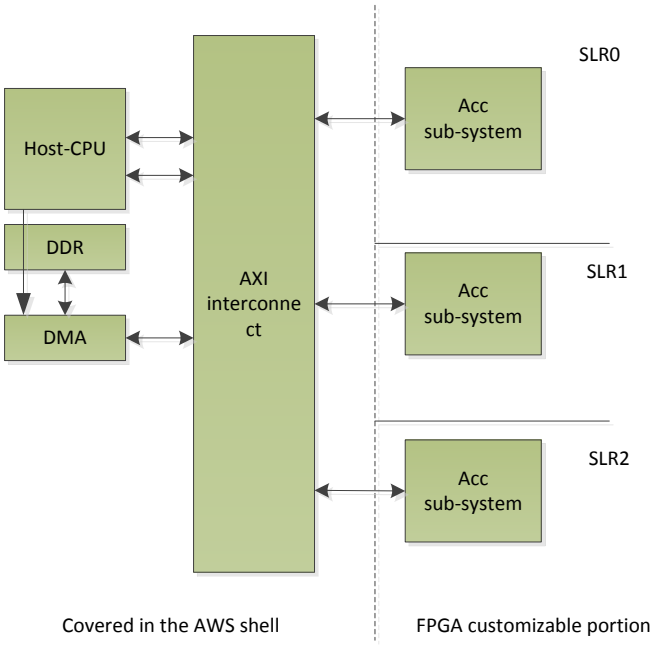
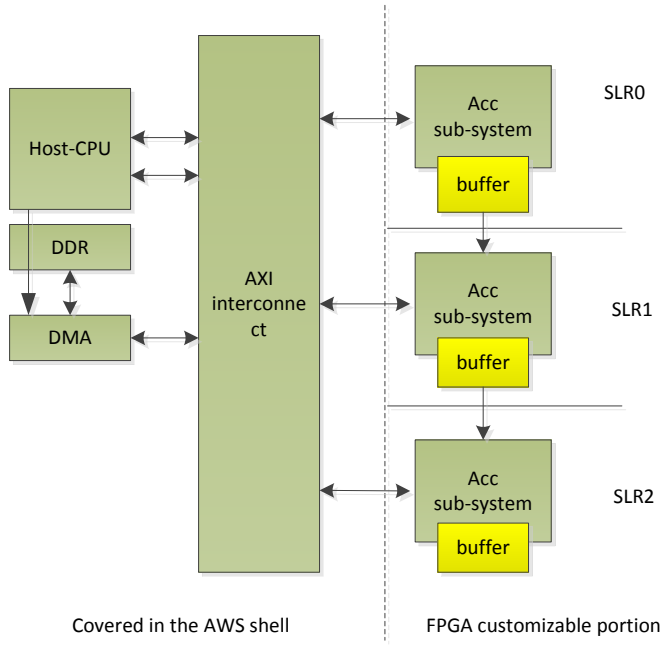


Sullotion 1: all intermediate data are transferred back to program space by DMA, different acc works on different layers of the input model

--- Pro: 1) flexible accelerator scheduling. Accelerators could have no dependency scheduling. 2) less timing issue

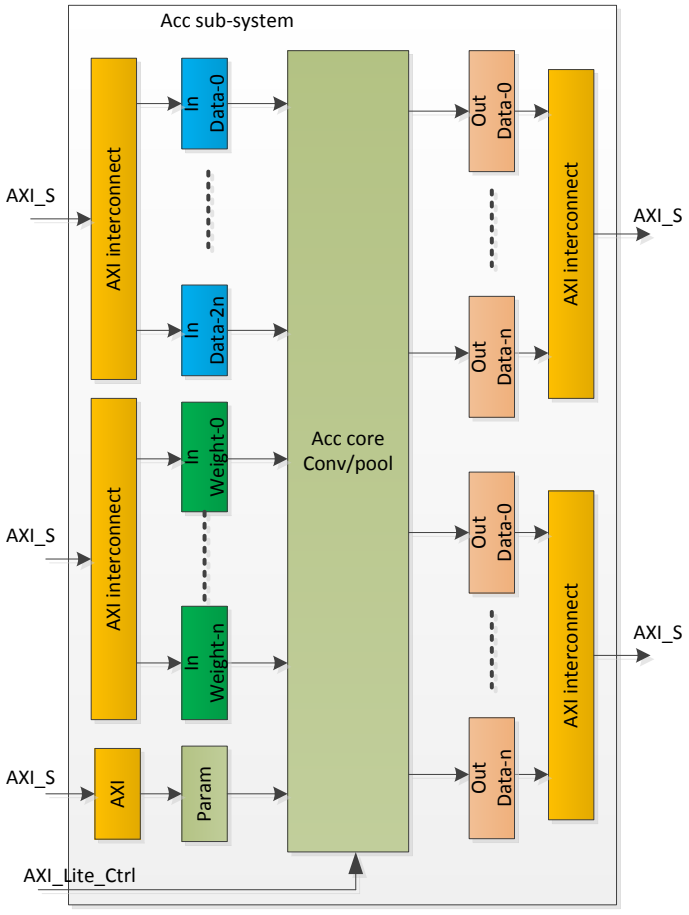
--- Con: Requires higher on/off-chip bandwidth



Sullotion 2: all or most of the intermediate data stored on-chip and transfer between SLRs.

--- Pro: less off-chip bandwidth requirement

--- Con: 1) new buffering tech 2) dependency based scheduling 3) accelerator output port modification to be compatible with FIFOs for on-chip data streaming



* Notes:

1. each of the In Data-n is a memory bank with parameterized width and depth.
2. All the input and output data ports are designed with double buffering, so in/out both have two mem-clusters. Each of the cluster have several memory bank based on the accelerator customization.
3. All the membanks for the port should be organized in continuous mem space.
4. The maximum number of AXI master port is 32, so when the core customization exceeds this number for any of the membanks (in/weight/out), our script need to be aware with it.