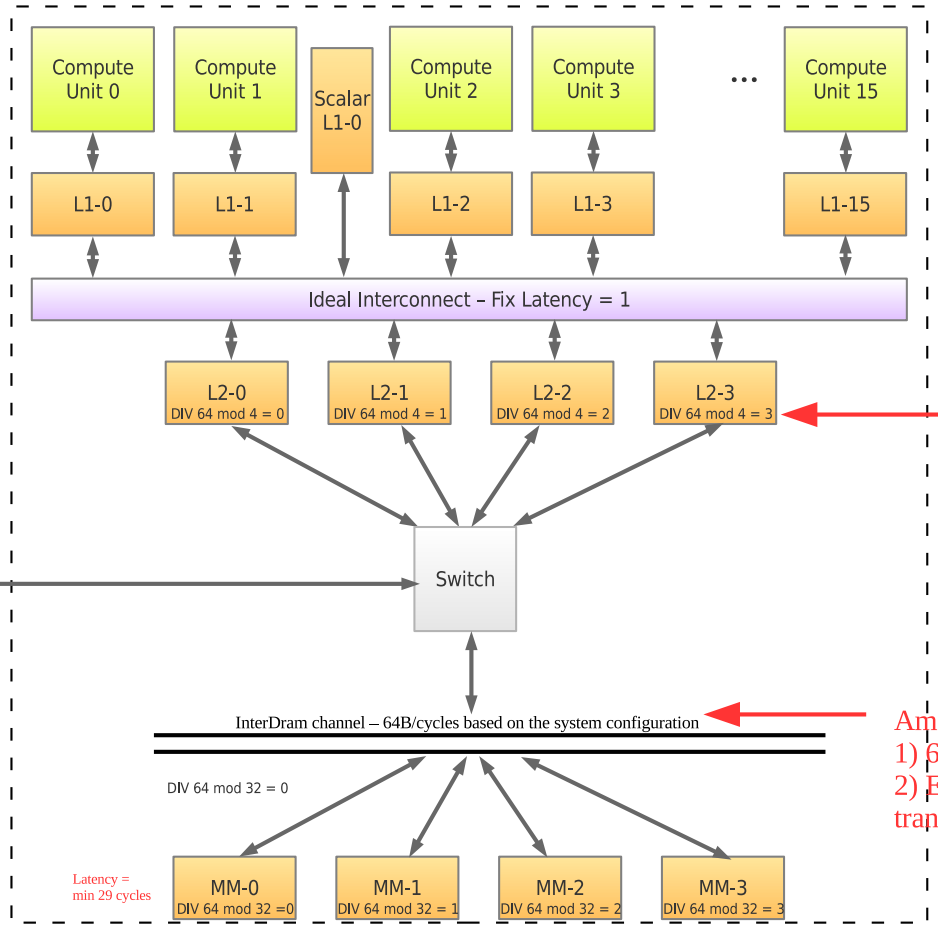
GPU0



1) Why are we assigning 4 L1 to a single L2?
2) Is it the number of L2's, 4? The following link suggests 8.
http://www.neoseeker.com/Articles/Hardware/Roundups/AMD_HD_7870_7850/2.html
3) The L2's are not partitioned. Therfor it is very likely that we will see a lot of data replacements between one certain L2 and the others. Also there are a lot of L2 polutions (due to having mu copies of the same cache-line in multiple L2s which hurts the traffic on the lower network (due to coherence)

Am I right here?
1) 64 byte per cycle in the slides I shared
2) Each controller is stated to have 64B/cycle data transfer. So should we do 4 lanes?

The GPUs are connected to each other through either PCIe bus, NVLink, or another network suggested by Kim.
Kim network is not considering the **Wire Latency** so first course of action is to calculate this latency, and the NVLink.

* Important. The first data copy is also uses one of these networks. So whatever the latency is we should measure the latency of a mem-copy of each Partition over the connections, and add them up to the whole execution time (in a serialize fashion)

GPU1, GPU2, GPU3 , ….

Yifan, would you review this and see if I am missing something? Thank you limitless, Ka