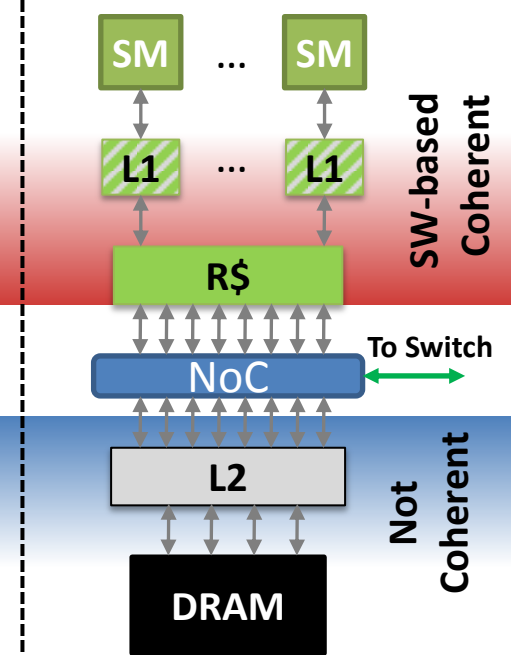
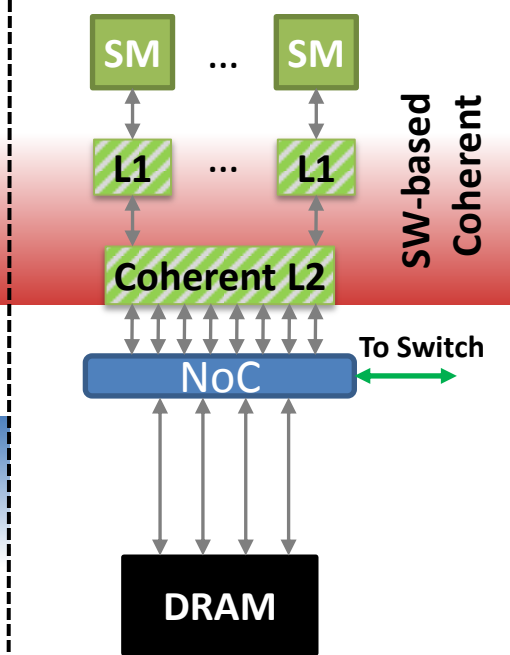


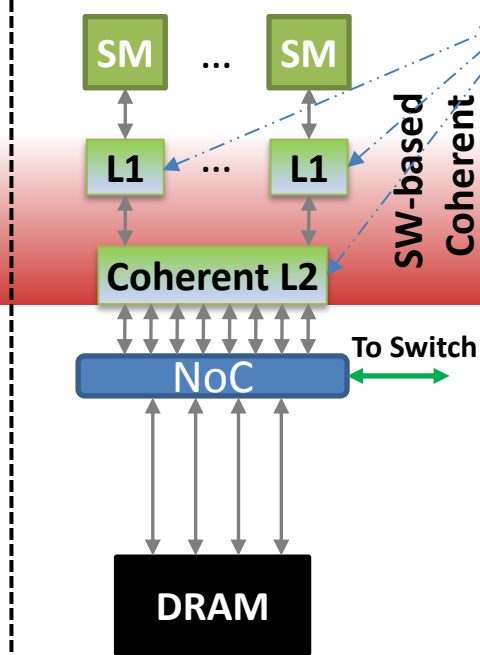
(a) Mem-Side Local Only L2



(b) Static R\$



(c) Shared Coherent L1+L2



(d) NUMA-Aware L1+L2

### Cache Partitioning Algorithm

- 0) Allocate  $\frac{1}{2}$  ways for local and  $\frac{1}{2}$  for remote data
- 1) Estimate incoming inter-GPU BW and monitor local DRAM outgoing BW
- 2) If inter-GPU BW is saturated and DRAM BW not
  - *RemoteWays++* and *LocalWays--*
- 3) If DRAM BW is saturated and inter-GPU BW not
  - *RemoteWays--* and *LocalWays++*
- 4) If both are saturated
  - *Equalize allocated ways (++ and --)*
- 5) None of them is saturated
  - *Do nothing*
- 6) Go back to 1) after *SampleTime* cycles