Motivation

Since Graphics Processing Units (GPUs) have gotten a significant improvement in performance, it is widely used as co-processors of Central Processing Units (CPUs) not only in traditional desktop system, but also in High-Performance Computing (HPC) field. There are four major parts in GPUs which are named shader core system, cache system, DRAM system and On-Chip Network (a.k.a NoC) respectively. To enable the growing number of threads concurrently running in GPUs, it is imperatively to design a NoC that can support a large amount of on-chip communication cost-effectively. There are two exactly same physical networks exist in GPU, one of two carry memory or L2 cache access traffic travelling from shader cores to memory controllers which called request network, the other network carry memory or L2 cache data information which travel back from memory controllers to shader cores, corresponding to request network, this is called reply network. Request data usually just contains 20 to 25 bytes memory address information but reply data at least includes 128 bytes (same size as cache line) which originally stored in main memory or L2 cache. So reply traffic is nearly 5 times of request traffic, which intuitively means that reply network is much more crowded than request network. But actually, Figure 1 shows that request network latency is much higher than reply network latency. This is because a cache line usually needs consume 9 cycles (under 16 bytes network bandwidth) to be fully injected into reply network, which means that memory controllers can only send one cache line every 9 cycles, but request packets just need to take 1 or 2 cycles to pop from request network into memory controllers, so memory controllers will eventually cannot accept request data anymore and thus request data will be stalled in request network worsening, we call this situation “backpressure”. In order to solve backpressure problem, reply network must be accelerated so that memory controller can send reply cache line more than current design and thus consuming more request data to remit request network traffic pressure and reduce overall network latency which packets travel in NoC.