Converged Web Scale Switching and Routing Becomes a Reality

Cisco Silicon One and HBM Memory Change the Paradigm

Introduction

There are several reasons why routing and web scale switching architectures have diverged over the years, but one of the major ones is the ever-increasing gap between the bandwidth of memories and network devices. This led the industry to create two types of devices – high bandwidth switches with small buffers and low bandwidth routers with large buffers. When planning their systems, network engineers need to select which type of device they should use. Every choice has its advantages, but one couldn't benefit from both worlds, at least not until now.

What's the reason for having two different product lines? What are the technologies that have changed the game? And how does Cisco Silicon One™ use these technologies to help eliminate this hard choice and unify the switch and the router to a single high bandwidth device with a large buffer? This isn't a unidimensional problem, and multiple components and creative solutions were embedded in Cisco Silicon One to unify these two product lines. In the next section, we'll touch on one technology that made it happen, which is the hybrid buffering scheme.

Contents

The history of buffers

All healthy networks experience congestion. It comes and goes and usually isn't persistent. It is evidence of the liveness of the network and its heartbeat. Many protocols, such as TCP, rely on congestion to achieve their goal of reliable transmission. What happens when a network device experiences congestion? Where is the extra data saved? For this, we have the buffer. Every network device consists of a buffer - memory that saves the data that cannot be transmitted immediately. Historically, the belief was that this buffer should hold at least all in-flight data between the source node to the target node, known also as Bandwidth-Delay-Product. Additionally, the interface to this buffer had to accommodate the entire bandwidth of the device, letting all incoming traffic be absorbed by the buffer.

These two requirements could coexist in early switches and routers, where external memories were used to meet the buffer size requirement stemming from the network Round Trip Time (RTT), and the memory bandwidth was sufficient to meet the device requirements.

However, as we can see in Figure 1, the bandwidth growth of external memories couldn't keep up with the bandwidth and radix growth of ethernet switches. For some time, this was solved by brute force: just increasing the number of memory devices. However, by mid-2010 the widening gap between switch bandwidth and memory bandwidth was becoming an unsolvable engineering problem. The only reasonable choice was to compromise on one of the requirements and create two different product lines: low bandwidth "routers" with large external-memory buffers and high bandwidth "switches" with small, internal buffers.

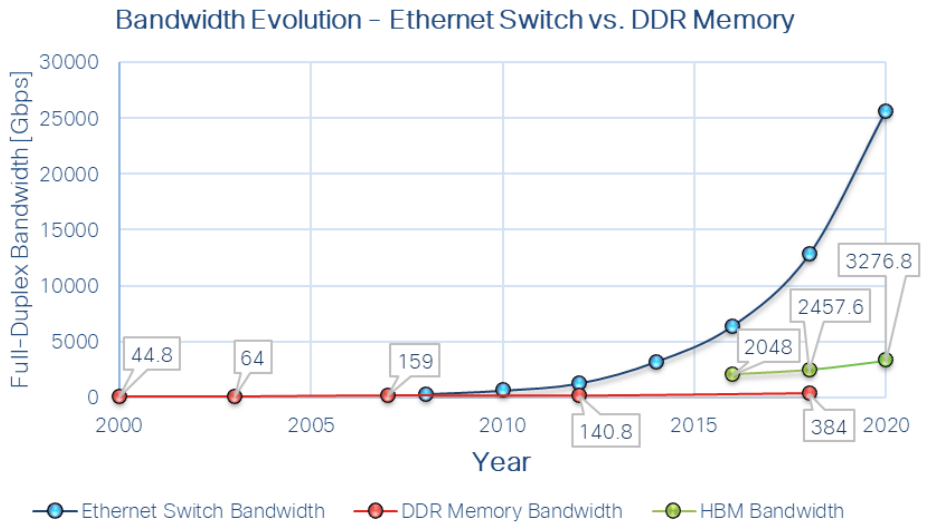

Figure 1. Bandwidth growth in DDR memories and ethernet switches

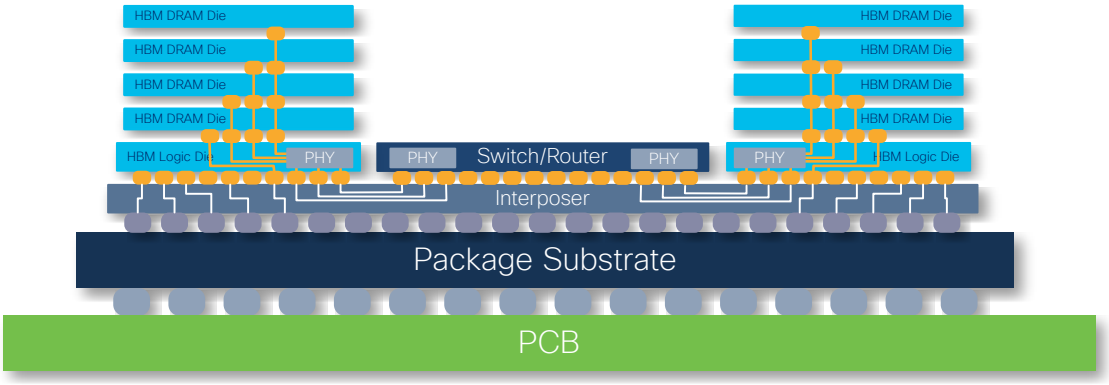

Figure 2. HBM - In-package memory

In 2013 a new technology arrived to help. High Bandwidth Memory (HBM) is a game-changing technology that allows high bandwidth switches to be equipped with deep buffering traditionally only available on much lower bandwidth systems. The HBM physical interface uses 2.5D packaging technology to allow the interconnect between the main ASIC die and the memory without having to route very large numbers of signals across a printed circuit board, saving critical routing resources and power. A typical deployment is shown in Figure 2. These HBM devices offer capacities that are more than 100x greater than the internal on-die SRAM. However, even with HBM devices, the memory bandwidth is still lower than the switch bandwidth, and as a result the access to the HBM must be carefully managed.

The Cisco Silicon One architecture employs a hybrid-buffer scheme that benefits from both worlds – internal memory bandwidth and external memory size. With efficient usage and smart management of the HBM interface, it helps enable the unification of both high-bandwidth switching and routing in a single device, as demonstrated by our Q200 – a 12.8Tbps router with 8GB of buffer in HBM.

Hybrid buffer architecture

Traditional architectures use external memories as the device's main memory by immediately buffering all incoming traffic in memory. All incoming packets, regardless of output port congestion state, are written to this memory. When a port is ready for transmission the packets are read from the memory and sent to the output port. To compensate for the read latency from the external memory and avoid underrun the device incorporates a small internal memory as a prefetch buffer. The flow of the packet data is shown in Figure 3.

Figure 1 shows the increasing gap between the total switch bandwidth and HBM bandwidth. Therefore, as not all incoming traffic can be written and read from the external memory, the traditional approach is not viable anymore and a new approach is needed.

Importantly, buffers in routers and switches aren't designed to absorb constant oversubscription, and in fact any buffering scheme under persistent oversubscription will be forced to drop packets. The buffering allows network equipment to ride through bursts of oversubscribed traffic without dropping packets

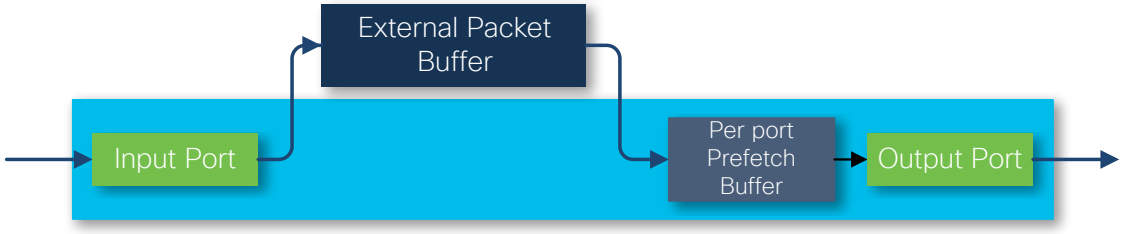

Figure 3. Traditional architecture with an external buffer

(or intelligently determining which packets to drop first). Analysis of network behavior shows that most of the time many of the flows don't experience significant congestion – or said another way: significant congestion events occur on only a small subset of the flows. This means that most of the time there's no oversubscription and no congestion, so a small "transit" memory is sufficient to accommodate the device processing time. The deep buffer is therefore only required for short periods when meaningful congestion occurs. In this case, if we could identify the flows that contribute to the oversubscription, or the aggressors, we could have buffered only these flows into the external memory. When the transient oversubscription is finished, the flows return from the external memory to use only the internal memory. This approach, adopted by Cisco Silicon One, is called hybrid memory architecture and is illustrated in Figure 4.

The hybrid memory architecture treats the interface towards the external memory and the external queues as resources and manages them accordingly. A careful and efficient management scheme should answer these questions:

- What are the data flows for packets that do not require external buffering?
- How are those flows changed when external buffering is required?
- What are the criteria to evict a queue to the external buffer?
- How is a queue returned from the external buffer to use only the internal buffer?

Cisco Silicon One implements a sophisticated management scheme that handles the HBM interface in a very efficient way by dynamically moving queues in and out of the external buffer. With that, it hides the limitations of the HBM technology and provides the same level of performance as the traditional buffering architectures for all real-life scenarios, even for high bandwidth devices like the 12.8Tbps Q200.

The management algorithms, as well as the congestion management functions, are an integral part of the Cisco Silicon One architecture and don't require software intervention. We provide a set of pre-configured parameters suitable for most applications. However, customers can use the provided APIs to tune and adjust these parameters to better match their application needs.

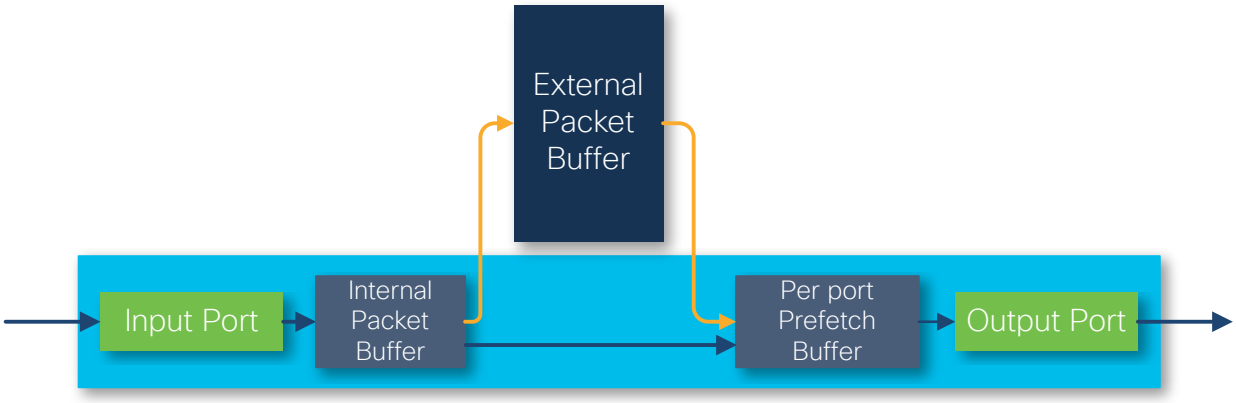

Figure 4. Hybrid memory architecture

Learn more

For additional information, please visit the Cisco Silicon One page.

Conclusion

New silicon technologies now enable the development of faster switches and routers with a higher radix, increasing the gap between system bandwidth and the underlying memory technologies. The traditional solutions for building routing silicon are impractical and unsustainable and contribute to the splitting between “high bandwidth switches with internal buffers” and “low bandwidth routers with external buffers”. Cisco Silicon One implements a hybrid memory architecture that benefits from both approaches and allows high bandwidth devices with deep buffers.

Hybrid memory architecture requires careful management of the HBM interface as well as the queuing resources. Cisco Silicon One’s hybrid memory architecture integrates sophisticated management algorithms that consist of unique connectivity between the internal and external buffers, and mechanisms that dynamically move queues between HBM and the internal buffer. The combination of these elements provides the performance expected from a high bandwidth switch with the buffering capabilities of an external buffered router.