

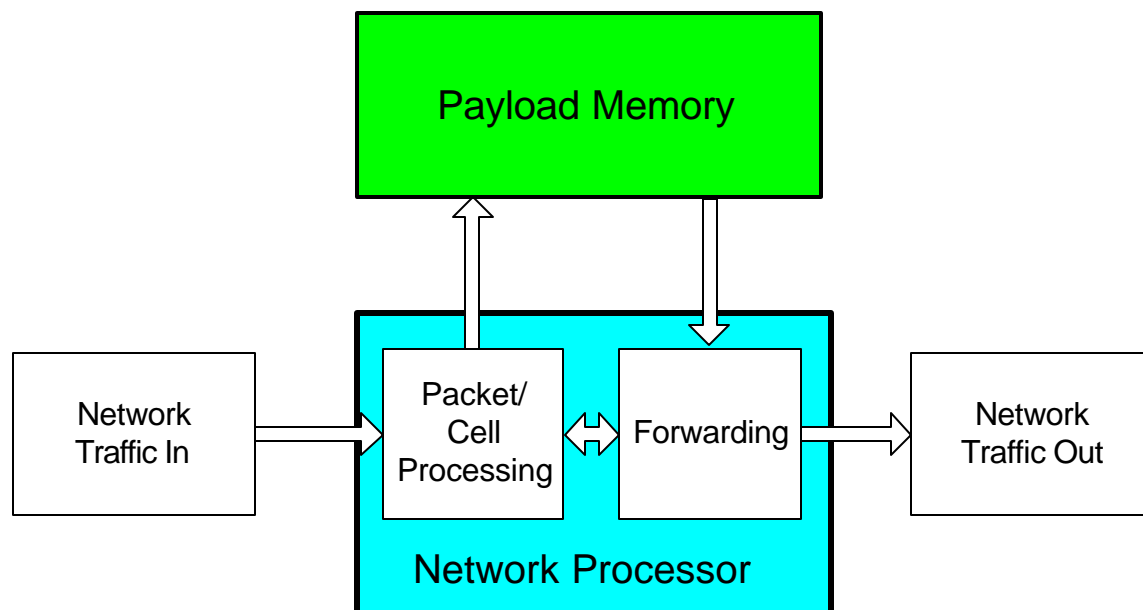
# Memory Subsystem: A Question of Cost vs. Speed

## Overview

A next-generation network processor (NP) serving the 10 Gb/s market (OC192c) must provide traffic engineering and queue management functions in addition to basic NP functions. While doing so, it must sustain line-level performance regardless of traffic patterns that may occur or network services implemented. It must be able to receive packets from line or fabric, process the packets by applying any required combination of network services, and forward the processed packets to either line or fabric.

When performing traffic engineering and queuing functions, an NP must examine every incoming packet to see what must be done before forwarding it. Because of this, incoming traffic must be stored temporarily while the NP processes it. Once the NP determines what to do (classification, protocol conversion, Qos/CoS etc.), it processes the packet and forwards it to its next destination. For temporary packet storage an NP uses a companion payload memory.

The payload memory, however, is a potential bottleneck in the network processor environment because of the speed at which traffic must be stored and forwarded. Unlike a PC memory where read operations outnumber write operations by about 3:1, payload memory read and write operations are equal—for every packet received and stored, there is a corresponding read and forward. Both store and forward functions must be done at speeds that maintain OC192c line rates.



*Incoming network traffic is temporarily stored in Payload Memory while it is processed by the Network Processor. Once a packet is processed, the NP's Forwarding engine sends it on its way.*

Implementing payload memory using high-speed SRAM would eliminate the potential bottleneck, but the number of chips required, board real estate taken, power consumed, and cost per chip eliminate SRAM as a viable alternative.

Several high-speed double data rate (DDR) DRAM families have the basic bandwidth to handle store and forward functions. They require fewer chips than SRAM, take up a fraction of the board space, consume less power, and cost thousands of dollars less than SRAM. However, every

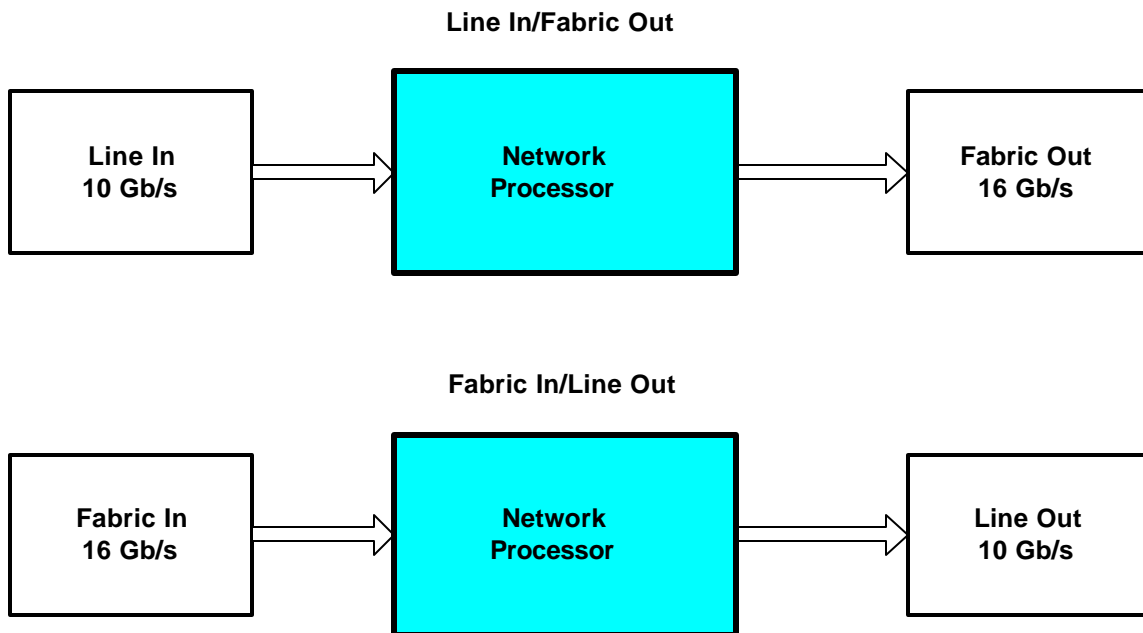
DRAM family suffers from latency delays. These delays can cause serious problems when processing strings of small packets that must be stored into random locations in payload memory.

This paper examines key performance factors associated with using DRAM subsystems to support NPs. It provides an analysis of the effective bandwidth of various specialty and commodity memory subsystems options, and identifies specific bottlenecks that affect overall system performance. It includes measured results for all options, with emphasis on a low-cost DDR SDRAM solution developed by Bay Microsystems.

The Bay Microsystems DDR SDRAM solution implements a patented memory management system and associated algorithms that Bay developed for NPU/TM applications. The Bay solution allows DDR DRAM devices to handle store-and-forward functions at full 10 Gb/s line speeds.

### The situation

OC192c network operations demand a *sustained* throughput as high as 26 Gb/s. Some NPs specify peak rates which meet this performance level under ideal conditions, but that is not good enough for true OC192c networks. The only meaningful metric with which to judge NP performance is the throughput that it can sustain regardless of network traffic patterns and regardless of the number and type of network services being provided. Bay refers to this as effective bandwidth.



*When operating in an OC192c network environment, a network processor must handle traffic at 10 Gb/s on the Line side, and at 16 Gb/s (line speed plus 0.6x overspeed) on the Fabric side. Total throughput (in and out) is up to 26 Gb/s.*

An NP must handle Line traffic (in or out) at 10 Gb/s and Fabric traffic (in or out) at an overspeed rate that may reach 1.6 times the Line rate. The total sustained throughput, therefore, which the NP and its associated payload memory must provide is at least:

$$10 \text{ Gb/s} + 16 \text{ Gb/s} = 26 \text{ Gb/s}$$

Because of the real-time nature of a network environment, payload memory must possess certain characteristics that differ from conventional computer memory. Issues that designers must consider include:

- **Memory speed.** To sustain 26 Gb/s throughput, payload memory must have the effective bandwidth to support this throughput.
- **Power consumption.** Payload memory will be large (typically in the 512 MB per channel range). But network equipment is sensitive to power consumption. Payload memory, therefore, must minimize chip count to minimize power consumption.
- **Board real estate.** This issue is closely related to power consumption. If payload memory is populated with a large number of chips, it will take up a lot of space on a board, and will typically consume a lot of power. Fewer chips take up less space and usually consume less power.
- **Cost.** Depending on the types of DRAM memory devices used, the cost for a single 512 MB channel of payload memory on a single line card can range from about \$32 (depending on current market rates) to more than \$200. (SRAM would be about \$4000.)

The clear choice is between specialty DRAM families (FCRAM, RLDRAM, RDRAM) and the commodity memory, DDR SDRAM. The majority of system design managers will choose the lower cost option, if it also addresses the speed, power, and real estate issues.

### **DDR SDRAM vs. specialty DRAM families**

The following table compares four DDR DRAM families that are candidates for payload memory applications. The table compares clock rate, bus width, raw bandwidth, multi-channel effective bandwidth, unit cost, and total cost to implement a single 512 MB payload memory channel.

*Raw bandwidth* is calculated as clock frequency times double the bus width. *Theoretical usable bandwidth* is the bandwidth that can be realized when all overhead issues are considered. This is a worst-case number for a single DRAM channel system; *it is also the worst case for a multi-channel system that does not implement sophisticated memory management algorithms*. For SDRAM, usable bandwidth varies from about 33 to 66 percent of raw bandwidth depending on page-miss and bank-select overhead (7.01 Gbps to 14.02 Gbps). For FCRAM it is about 66 percent. *Effective multi-channel bandwidth* factors the overhead associated with the choosing of the correct channel to store and forward the data when using multiple channels to yield higher total system bandwidth. These numbers reflect the data measured by Bay Microsystems when applying our memory management algorithms.

Memory Type	Clock (MHz)	Bus Width (bits)	Raw Bandwidth (Gb/s)	Theoretical Usable Bandwidth	Effective multi-channel Bandwidth	Ratio of Total Cost/ Channel
DDR SDRAM	166	64	21.5	7.01	8.41	1x
DDR FCRAM	166	64	21.5	14.02	11.22	6x-8x
DDR RLDRAM	200	64	25.6	12.03	9.62	6x-8x
DDR RDRAM	400	16	12.8	5.25	not measured	1.5x-2x

Of the specialty memory available today FCRAM is the current performance leader. For this reason Bay's products were designed to support FCRAM memories. But to provide our customers a lower cost option we needed to determine if DDR-SDRAM could be supported as well. When comparing DDR-SDRAM and FCRAM, two things stand out: FCRAM's "theoretical" usable

bandwidth is a fixed number varying between twice that of SDRAM to parity. However, at current market prices, FCRAM costs from six to eight times more than SDRAM. The market still considers FCRAM a specialty memory while SDRAM is more commonly thought of as a commodity memory. Depending on FCRAM's higher "effective" bandwidth, it may allow slightly smaller memory bus widths to be used, but the price difference should still favor DDR-SDRAM.

The question is, can DDR-SDRAM be configured to support the 26 Gb/s or more of throughput that OC192c requires? The answer is "yes," and that is the subject of the remainder of this paper.

### The SDRAM bottleneck

The most significant DDR-SDRAM bottleneck involves reading packets from payload memory in order to forward them to their next destination. When storing packets in payload memory, the NP can select the most efficient storage location. However, packets are not necessarily read in the order that they are stored. Packets are read in the order that they are to be forwarded, and that is usually from random locations.

When reading from DDR-SDRAM, three overhead factors must be considered:

1. **Refresh.** Any DRAM device must be refreshed periodically. This is a small penalty and is insignificant compared to other latency issues.
2. **Page-miss overhead.** DDR-SDRAM is organized in pages and banks. When a memory operation addresses a location that is not in the page currently being addressed, the DDR-SDRAM needs 4 to 6 cycles to switch to the new page.
3. **Read Latency.** This is the time from the issuing of a read command to the availability of data. For DDR-SDRAM this is 3 or 4 clock cycles.

The worst-case situation involves consecutive memory reads from random locations where each access moves a small amount of data. In this case, the combined delays can consume a high percentage of the total bandwidth of the memory. Assuming each access reads eight bytes (two per clock cycle), a single transfer, in memory cycles, looks like this:

PAGE MISS = 6	READ LATENCY = 3 OR 4	DATA = 4
---------------	-----------------------	----------

*When a packet contains a relatively small amount of data, memory overhead eats up a high percentage of memory bandwidth.*

For DDR SDRAM to be used in a payload memory application, it is necessary to find a way to overcome the page miss and latency overhead.

### Bus width selection

The first step in optimizing payload memory is to decide on memory bus width. Storing a byte at a time eats up memory bandwidth very quickly. However, a very wide bus, for example 256 bits, reduces memory efficiency because of wasted memory space when partially filled words are stored. Payload memory will write and read a certain amount of data on every access. If the bus width is 64 bits, payload memories will read/write 8 bytes on every clock. If the bus width is 256 bits, they will read/write 32 bytes. This is true even if there is only one valid data byte on the bus.

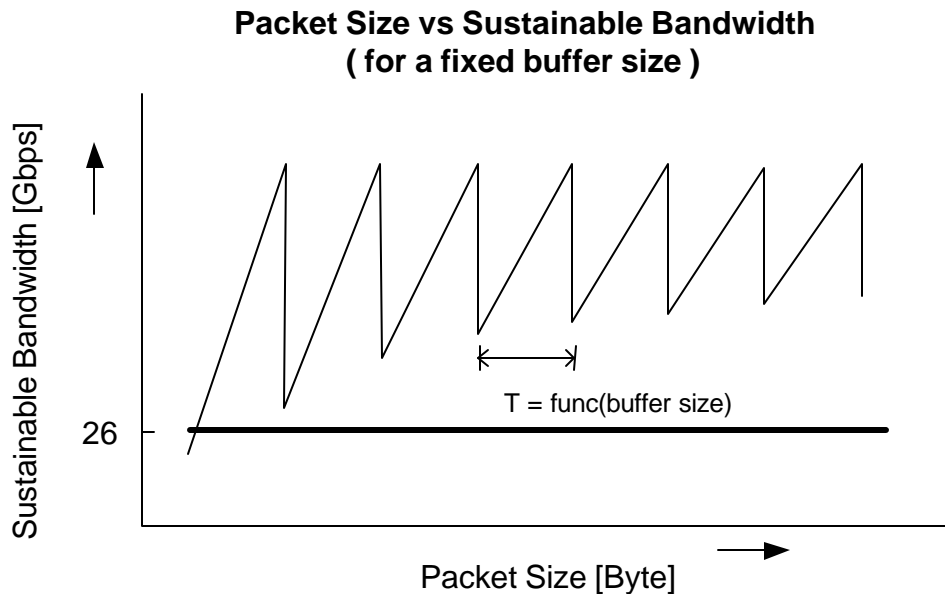
Bay Microsystems engineers evaluated bus width and determined that a 64-bit bus achieves the best balance of memory bandwidth and memory efficiency. A 64-bit bus achieves 33 percent

worst-case bandwidth utilization while a 128-bit bus achieves 20 percent and a 256-bit bus achieves only 11 percent.

Using a 64-bit bus in a system operating at 166 MHz provides a raw bandwidth of 21.5 Gb/s when using DDR SDRAM

### **Buffer size selection**

Another consideration in designing a payload memory is the size of the buffer used to store the data. The following diagram illustrates how bandwidth relates to buffer size. Memory bandwidth increases as the size of the packet increases. For example, for a 256-byte buffer, the maximum efficiency is achieved when a packet exactly fills the buffer (the high peaks in the diagram). Lowest efficiency happens when the packet contains one byte more than the buffer size (the low points in the diagram). Each low point is a case where packet size is one byte more than the buffer size.



*This diagram plots packet size and sustainable bandwidth for a given buffer size. Highest efficiency (highest bandwidth) occurs when a packet size exactly matches buffer size (high peaks of the sawtooth diagram). Lowest efficiency (lowest bandwidth) occurs when the packet size is one byte more than the buffer size (low peaks of the sawtooth).*

In order to guarantee line-rate for any traffic patterns, all the low points of the sawtooth must be equal to or higher than the 26Gbps requirement of the NPU.

### **Turning “theory” into practice**

By looking at the “Theoretical Usable Bandwidth per channel” numbers it may be assumed that by multiplying this by, let’s say 3 (channels), will yield three times the bandwidth. This would only be true if the data was non-bursty, uniformly sized, and was forwarded in exactly the same order as stored. That of course, is not the case for true payload traffic in practice.

The three-channel configuration may appear simple in concept, but is actually a very complex structure. In practice, the payload memory interface could overload one channel while leaving the others idle, thus giving up the bandwidth that the channels offer. To achieve maximum usable throughput, a scheme must be developed to make the bursty, non-uniform, random data appear as non-bursty, uniform data to the memories.

Bay Microsystems developed a patented memory management system and associated algorithms that maximize bandwidth among all channels.

### ***The end result***

By carefully taking all the factors described above into consideration, Bay Microsystems has developed a patented memory management system and associated algorithms. The effectiveness of Bay's solution allows for the use of commodity DDR SDRAM memory subsystems while enabling as much as 32Gbps of store and forward capacity regardless of traffic patterns. This memory subsystem is used throughout Bay Microsystems' Internetworking Processor Family. The initial instantiation of this memory subsystem is within Bay's OC129c/10G NPU/TM, Montego. It allows network system designers to choose either low-cost DDR SDRAM or low-latency FCRAM for their OC192c/10G applications.