Synchronous Cache Popularity Increases

Pentium Chip Sets, Faster Clock Speeds Spur Interest in New SRAMs

by Linley Gwennap

Specialty SRAMs have seen a surge in popularity recently, driven by their increased use in high-end PC designs. Synchronous caches, supported by Intel's Neptune and Triton chip sets among others, can raise system performance by 5–10% with an increase of just $30–$50 in manufacturing cost. With this performance boost, PC vendors can differentiate their high-end systems. In addition, several new RISC processors require synchronous caches. We expect that, as the price of these new parts drops, the number of synchronous cache designs will surpass asynchronous caches within two years.

With more vendors selling synchronous SRAMs, several variations are now available. The terminology describing these parts is even more varied: registered, latched, burst, pipelined, flow-through, etc. This article unravels this knot of confusion, looking at the different types of parts and their typical applications.

Faster CPUs Force SRAM Design Changes

For years, processors have used standard asynchronous SRAMs for secondary caches. With both CPUs and SRAMs riding the same IC process curve, memory access times kept pace with processor cycle times, allowing the same basic design to scale in speed. When cache speeds approached 100 MHz, however, designers discovered that one factor in the equation was not scaling: transmission time between the CPU and the cache RAMs.

At 50 MHz, for example, a 15-ns SRAM allows a couple of nanoseconds for the address to flow from the CPU to the SRAM and a similar time for the data to return to the processor chip. But at 100 MHz, the same 5 ns of wire delay leaves just 5 ns for the memory access. Thus, SRAM speed must triple to account for a doubling of processor speed. Current manufacturing technology does not support standard SRAMs at this speed.

This situation has led to the development of a variety of alternative SRAM designs. Known broadly as synchronous SRAMs, these parts use a clock signal to latch the inputs and sometimes the data output. This structure spreads the cache access across two or three cycles while maintaining a bandwidth of one access per cycle. Extending the number of cycles for each access allows plenty of time for both the SRAM access and wire delays, even at high clock speeds.

Pipelined Versus Flow-Through

Figure 1 compares the design of the two basic variations of synchronous SRAMs with a standard asynchronous memory. The flow-through design (b) is identical to an asynchronous part (a) except for registers on the input signals. In this design, the address is sent from the CPU to the SRAM during cycle 1 and stored in the register on the rising clock edge.

During the next cycle, the memory array is accessed and the data flows back to the CPU, as the timing diagram shows. The input register allows the CPU to send the next address even as the array access is in progress, keeping the bandwidth to one access per cycle. The terms registered and "standard" synchronous SRAM are sometimes applied to this design.

The alternative, called pipelined, adds a register to the data output as well, as Figure 1(c) shows. As in the previous design, the memory array is accessed in cycle 2, but in this case, the output data is stored in a register. The CPU reads this register in cycle 3 to obtain the data. This design allows the array access to consume nearly an entire cycle. Although the data is available one cycle later, it is avail-

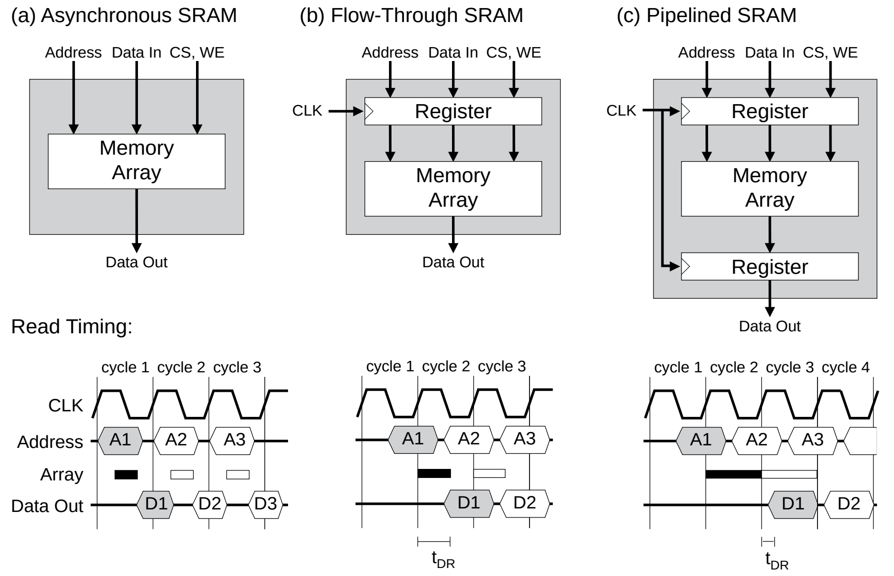

Figure 1. All synchronous SRAMs have registers on the inputs, but only pipelined versions have a registered output. The extra register pushes data availability into the third cycle but greatly reduces the data-output delay (t_{DR}).

able much earlier in the cycle than with the flow-through design. As with the flow-through design, accesses can be overlapped to maintain a bandwidth of one per cycle.

Both designs have similar write timing. On a write, the input data, like the address and control signals, is stored in a register. The memory array is written in the second cycle.

Pipelined Parts Are Less Expensive

With pipelined SRAM, signals can take up to a full cycle to move between the CPU and the cache, and vice versa. Alternatively, the CPU can begin processing the data during the same cycle it is received from the cache; in other designs, the data is not available until near the end of the cycle. For most desktop applications, the extra cycle of latency is not a significant performance issue.

The incentive to use pipelined parts is that they are less expensive than flow-through parts. A 66-MHz (15-ns) flow-through design, for example, might support a data-out time of 9 ns to allow time for the data to flow back to the CPU. A pipelined 66-MHz part, on the other hand, could get away with a 15-ns array. Thus, the flow-through part must use a more aggressive manufacturing process, and thus carry a higher price, to operate at the same cycle time.

Pipelined SRAMs have about the same manufacturing cost as an asynchronous part of the same speed; the die area of the pipelined part is only slightly larger, to include the registers, and the core must be slightly faster. Today, these parts are selling for a significant premium due to high demand and a relatively low supply, but this premium should diminish over time.

Because of the more aggressive manufacturing processes required to build their faster arrays, flow-through parts have a significantly higher manufacturing cost, although the ratio is not as high as the 2× price premium that these parts currently carry. Again, this premium will diminish over time, but flow-through parts will always be more expensive than pipelined devices that support the same cycle time.

Many of the pipelined SRAMs offered today are "generic" synchronous SRAMs that can operate in either flow-through or pipelined mode. Although these parts offer both the vendor and the system designer flexibility, they must meet the stricter timing for flow-through parts, and thus they do not gain the cost advantages of a pipelined-only part.

Burst Parts Ideal for Cache Refills

Both types of synchronous parts can sustain one access per cycle: the CPU can send a series of addresses and receive a series of data words, offset by one or two cycles. Traditionally, the CPU must generate a series of sequential addresses to refill a complete line in the on-chip cache. Burst SRAMs are synchronous parts that automatically increment the address to deliver a full cache line. The CPU simply sends the first address and receives four data words on consecutive cycles. Figure 2 shows a burst read transaction for a flow-through part.

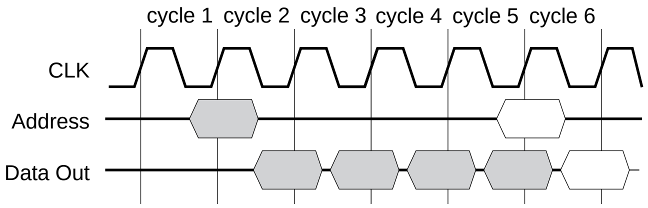

Figure 2. A burst SRAM can return several data words from a single address by automatically incrementing the address. This example shows a flow-through SRAM.

The 486, for example, uses a 16-byte line size for its internal cache. With its 32-bit bus, four read transactions supply enough data to refill one cache line. Pentium uses a 32-byte cache line but increases the bus width to 64 bits, so it still takes four cycles to read one cache line.

To improve performance, many processors read the "critical" word—the one requested by the stalled instruction—first, followed by the other words in the cache line. Thus, the SRAM's address logic must generate different sequences of addresses depending on the low-order bits of the first address, as Table 1 shows.

This situation is further complicated because Intel processors use a patented burst-address order that differs from the linear address order used by most other processors. Some burst SRAMs implement one order or the other, while other chips can be configured for either. SRAMs are frequently touted as "optimized for Pentium" or "PowerPC-specific" parts; these labels simply indicate which burst order the parts implement.

Improving Pentium System Performance

Most Pentium systems currently use an asynchronous cache. With a 60- or 66-MHz bus, 15-ns parts take two bus cycles to deliver data; an extra cycle is required at the beginning of each access for the tag lookup. This response time can be written as 3-2-2-2 (three cycles for the first access, two for subsequent accesses). Eight 32K×8 (256-Kbit) parts, at about $3.50 each, store the data for a 256K cache in a typical design. The total cost of this cache is about $30.

Starting Address	Linear Order	Intel Order
0	0, 1, 2, 3	0, 1, 2, 3
1	1, 2, 3, 0	1, 0, 3, 2
2	2, 3, 0, 1	2, 3, 0, 1
3	3, 0, 1, 2	3, 2, 1, 0

Table 1. Burst SRAMs typically support either a linear address sequence—used by PowerPC, 68040, and other processors—or the Intel scheme implemented in its 486, Pentium, and 960 chips.

Vendor	Phone	Synch. Type	Burst Mode	Speeds
Alliance	408.383.4900	pipelined	either	50–66 MHz
Cypress	408.943.2600	flow-thru	either	50–60 MHz
Hitachi	415.589.8300	either	Intel	50–66 MHz
IBM Micro	800.426.0181	either	either	50–100 MHz
IC Works	408.922.0202	either	Intel	50–66 MHz
IDT	800.345.7015	either	either	50–66 MHz
ISSI	408.733.4774	pipelined	either	50–66 MHz
Micron	208.368.3950	either	either	50–100 MHz
Motorola	512.933.7726	either	either	50–100 MHz
NEC	800.366.9782	either	either	50–150 MHz
Paradigm	408.954.0500	flow-thru	either	50–66 MHz
Samsung	408.954.7000	either	either	50–100 MHz
SGS-Thomson	617.259.0300	either	either	33–50 MHz
Sharp	800.642.0261	either	either	50–66 MHz
Sony	408.432.0190	pipelined	either	50–100 MHz
Toshiba	800.879.4963	pipelined	either	50–66 MHz

Table 2. Many vendors offer synchronous SRAMs. Most vendors target Pentium systems with cache speeds up to 66 MHz, but a few sell to RISC systems that require much faster parts.

Premium systems often use a synchronous cache for higher performance. Chip sets from Intel and other vendors support either flow-through or pipelined burst SRAMs as well as standard parts. With a synchronous cache, these chip sets can deliver 3-1-1-1 performance. Accesses that hit in the same SRAM row as the previous access can proceed at a 1-1-1-1 rate.

Currently, such designs typically use 32K×8 flow-through parts, which sell for about twice the price of asynchronous chips. The next generation of designs is moving to 32K×32 (1-Mbit) parts that allow just two SRAMs to create a 256K cache. These chips, available from NEC and Sony among others, typically use a 3.3-V supply but offer 5-V–tolerant I/O. This arrangement permits the parts to function in older 60- and 66-MHz Pentium systems but is optimized for faster 3.3-V Pentium chips. At about $30, these larger SRAMs carry the same price per bit as their smaller predecessors, but these prices should fall as volume increases.

Table 2 lists several vendors that offer synchronous SRAMs. Most offer parts with a top speed of 66 MHz; these are suitable for all Pentium systems as well as for most PowerPC systems. The faster parts are used primarily by high-end RISC processors.

RISC Designers Seek Higher Clock Rates

For example, HP processors up to the PA-7200 *(see **0905MSB.PDF**)* have relied on asynchronous SRAMs in their single-cycle caches. HP uses a technique called "wave pipelining" to launch a new address before the data from the previous read is completed. This technique spreads the cache access across one-and-a-half cycles and allows HP to use asynchronous parts with an access time near the CPU cycle time. For example, a 100-MHz PA-RISC processor operates with 8- and 9-ns SRAMs.

With the PA-8000, however, HP is targeting a 200-MHz clock speed, which would require 3- or 4-ns parts using wave pipelining. Such parts are not available today and will be very expensive even when the PA-8000 ships next year. Instead, the new design uses synchronous caches operating at the CPU clock speed. This change increases latency but supports the same bandwidth as a true single-cycle design at a lower cost.

Other next-generation RISC processors—including the PowerPC 620, MIPS R10000, and UltraSparc—need synchronous parts for their external caches; the Alpha 21164 allows either synchronous or asynchronous parts. UltraSparc requires the external cache to operate at the CPU clock speed, which Sun expects to reach 167 MHz when the first systems ship this summer. The other chips allow the external cache to run at fractions of the CPU frequency, reducing implementation cost.

At speeds of 100 MHz or above, most SRAMs are moving away from the 3.3-V I/O used by Pentium-class parts. Recently, the JEDEC group approved a standard called HSTL (high-speed transceiver logic) that uses reduced signal levels—0.55 V for zero and 0.95 V for one—to speed transmission times. It is not yet clear whether fast synchronous SRAMs will use GTL *(see **070301.PDF**)*, HSTL, or low-voltage TTL (LVTTL) interfaces.

Faster parts are also leading the way to BGA (ball-grid array) packaging. Most parts today use PLCC or TQFP packages, but BGAs provide superior electrical characteristics, improving transmission times at high frequencies. Motorola, one of the leading vendors of high-speed SRAMs, is a BGA proponent; other SRAM vendors are also beginning to offer this packaging option for fast parts and for those with larger pinouts.

Growth Opportunities for SRAM Vendors

According to market-watcher In-Stat (Scottsdale, Ariz.), worldwide consumption of synchronous SRAMs will grow from fewer than 10 million units in 1994 to more than 70 million units in 1997. The latter figure is less than 10% of the overall SRAM market; the bulk of these parts will be used in PC and workstation caches. At 2–4 SRAMs per system, this figure translates to about 25 million synchronous caches, supporting vendor claims that these designs will be more common than asynchronous caches by 1997.

Growth in synchronous SRAM sales today is limited by the high price premium of these parts. This situation has created a chicken-and-egg problem: SRAM vendors need higher volumes to cut their prices, but system designers want to see lower prices before using the new parts. We expect that a few aggressive vendors, such as Samsung and IDT, will ramp up volumes and cut prices, jump-starting the cycle. These prices will lead to increased use of synchronous caches in PCs and, ultimately, better performance for end users. ♦