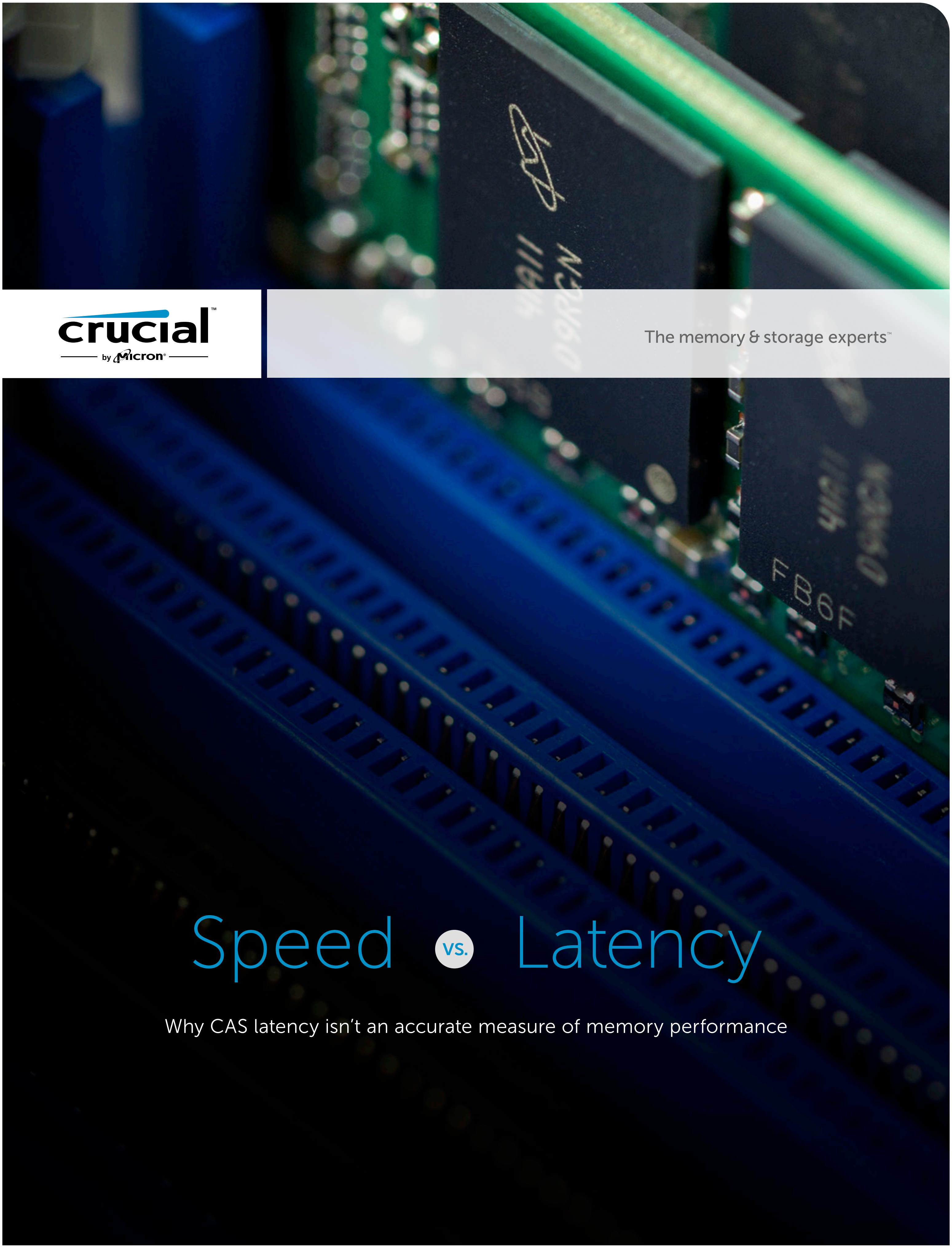


crucial™ by Micron®

The memory & storage experts™

Speed vs. Latency

Why CAS latency isn't an accurate measure of memory performance

Speed vs. Latency

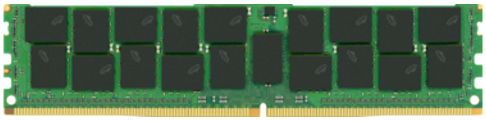
While speed and latency are both related to memory performance, they're not necessarily related in the way that you might think. Most people understand the speed/latency relationship in that as speeds increase, so do latencies. However, this isn't necessarily the case. In fact, it's highly misleading and can lead users to settle for lower levels of performance. Here's how speed and latency are related—and what it means in terms of your memory's performance.

Speed defined

Speed is easy to understand—it's a measure of how fast a stick of RAM is able to process data. Speed is measured in megatransfers per second (MT/s) and you want as much of it as is possible and/or cost-effective. In the history of the memory industry, speeds have always increased with each new memory technology.

Latency defined

Compared to speed, latency is much more complex—and often misunderstood. At a basic level, latency refers to the time delay between when a command is entered and executed. It's the gap between the two. At a precise technical level, latency measures the time it takes for the memory controller to tell the RAM to access a particular location, and when the data in that location is actually read.



Because latency is all about the gap between when a command is entered and executed, it's critical to understand what happens during this gap. After the memory controller tells the RAM to access a particular location, the data must go through a set number of clock cycles in the Column Address Strobe in order to get to its desired location and "complete" the command. With this in mind, there are two variables when it comes to determining a given module's latency: **(1)** the total number of clock cycles the data must go through (measured in CAS Latency, or **CL** on data sheets), and **(2)** the duration of each clock cycle (measured in nanoseconds). Here's the exact formula:

Latency Formula

$$\text{true latency}^{(\text{ns})} = \text{clock cycle time}^{(\text{ns})} \times \text{number of clock cycles}^{(\text{CL})}$$

The Latency Paradox

Latency is often misunderstood because on most product flyers and spec comparisons, it's noted in CL, which is only half of the latency equation. Since CL ratings only indicate the total number of clock cycles, they don't have anything to do with the duration of each clock cycle, and thus, they shouldn't be extrapolated as the sole indicator of latency performance.

This presents us with the **Latency Paradox**. Take a look at *Figure 1*.

Figure 1

SPEED VS. LATENCY AS MEMORY TECHNOLOGY HAS MATURED
(INDUSTRY STANDARDS)

TECHNOLOGY	MODULE SPEED (MT/s)	CLOCK CYCLE TIME (ns)	CAS LATENCY (CL)	TRUE LATENCY (ns)
SDR	10E	8.00	3	24.00
SDR	133	7.50	3	22.50
DDR	335	6.00	2.5	15.00
DDR	40B	5.00	3	15.00
DDR2	667	3.00	5	15.00
DDR2	800	2.50	6	15.00
DDR3	1333	1.50	9	13.50
DDR3	160B	1.25	11	13.75
DDR4	1866	1.07	13	13.93
DDR4	2133	0.94	15	14.06
DDR4	2400	0.83	17	14.17
DDR4	2666	0.75	18	13.50

In the history of memory technology, as speeds have increased, clock cycle times have actually decreased, resulting in lower true latencies as technology has matured, even though there are more clock cycles to complete.

Nanoseconds:

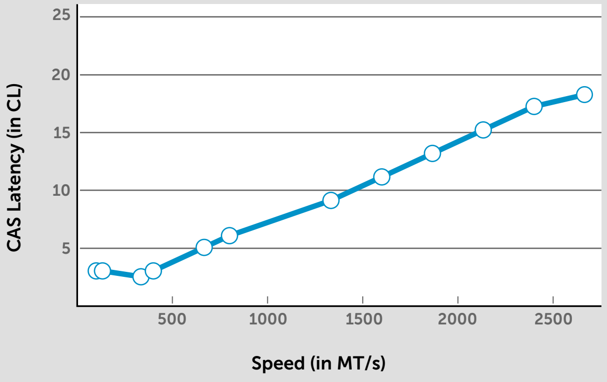
A better measure of latency performance

Since latency is all about how long it takes the memory to execute a command once it's been entered, it's best measured in pure nanoseconds rather than in CL (which is all about the number of clock cycles, rather than how long they take to complete). By looking at a module's latency in terms of nanoseconds, you can better judge if one module is, in fact, more responsive than another. To calculate a module's true latency, multiply clock cycle duration by the total number of clock cycles. These numbers will be noted in official engineering documentation on a module's data sheet.

How most people view latency

As speeds increase, latencies increase

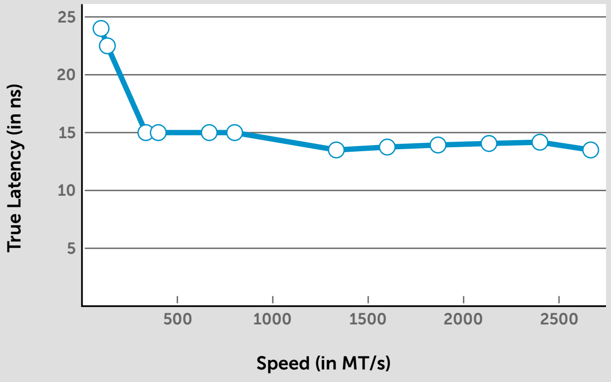
SPEED VS. CAS LATENCY



How semiconductor engineers view latency

As speeds increase, latencies decrease and/or remain roughly the same (approximately 14ns)

SPEED VS. TRUE LATENCY



When speed is compared to true latency, it's easy to see that as memory technology has improved, latencies haven't really increased. What's more, since speeds are increasing and true latencies are remaining roughly the same, you're able to achieve a higher level of performance using newer, faster, and more energy efficient memory.

At this point in the discussion, we need to note that when we say "true latencies are remaining roughly the same," we mean that from DDR3-1333 to DDR4-2666 (the span of modern memory), true latencies started at 13.5ns and returned to 13.5ns. While there are several instances in this range where true latencies increased, the gains have been by fractions of a nanosecond. In this same span, speeds have increased by over 1,300 MT/s, effectively offsetting any trace latency gains.

However, if you're still concerned about the general principle—that latencies can increase, even if by the smallest of amounts—here's a technical explanation as to why this is the industry standard.

Why speed and latency are related

In order to ensure consistently fast responsiveness in modern memory, CL ratings typically have to increase alongside frequency in order to maintain an approximate 14ns access time.* This is important because if CL ratings didn't increase with each cadence, then **(a)** data rates wouldn't be able to increase, **(b)** memory yields at the fastest speeds/lowest latencies would be affected, or **(c)** the physical size of the memory modules would have to increase substantially. For end users, any one of these three things would make memory significantly more expensive. For this reason, JEDEC industry standards typically are set by the market to enable the mass production of cost-effective modules that deliver better real-world performance gains.

The bottom line

Memory performance is all about the relationship between speed and latency. For optimal performance, install as much memory as possible, use the latest memory technology, and choose modules with as much speed as is cost-effective and/or relevant for the applications you're using. In general, as speeds have increased, true latencies have remained approximately the same, meaning faster speeds enable you to achieve a higher level of performance. True latencies haven't necessarily increased, just CAS latencies. And CL ratings are an inaccurate, and often misleading, indicator of true latency performance.

*We always aim for the lowest possible access time. Current access times are subject to change as memory technology and/or process advancements mature.

REVISION: *05/01/2015*