
crucial
by Micron
The memory & storage experts™
Geschwindigkeit vs. Latenzzeit
Darum ist die CAS-Latenz keine zuverlässige Messgröße für die Speicherleistung

Geschwindigkeit vs. Latenzzeit

Auch wenn die Geschwindigkeit und die Latenzzeit mit der Speicherleistung in Verbindung stehen, ist diese Verbindung möglicherweise anders geartet, als Sie vielleicht glauben. Die meisten Menschen verstehen die Beziehung zwischen Geschwindigkeit und Latenzzeit so: Wenn sich die Geschwindigkeit erhöht, verlängert sich auch die Latenzzeit. Das ist jedoch nicht notwendigerweise der Fall. Tatsächlich ist es sogar sehr irreführend und kann die Benutzer dazu verleiten, sich mit einer geringeren Leistung zufriedenzugeben. Sie erfahren nun, wie Geschwindigkeit und Latenzzeit zusammenhängen und wie sich das auf Ihre Speicherleistung auswirkt.

Definition von Geschwindigkeit

Die Geschwindigkeit ist leicht zu verstehen. Sie gibt an, wie schnell ein RAM-Stick Daten verarbeiten kann. Die Geschwindigkeit wird in Megatransfer pro Sekunde (MT/s) gemessen. Natürlich möchte man eine möglichst hohe und/oder kostengünstige Geschwindigkeit erreichen. Im Laufe der Geschichte der Speicherbranche hat sich die Geschwindigkeit mit jeder neuen Speichertechnologie erhöht.

Definition von Latenzzeit

Die Latenzzeit ist wesentlich komplexer als die Geschwindigkeit und wird oft missverstanden. Auf einer grundsätzlichen Ebene bezeichnet „Latenzzeit" die Zeitverzögerung zwischen der Eingabe und der Ausführung eines Befehls. Sie ist der Abstand zwischen diesen beiden Punkten. Auf einer präzisen, technischen Ebene bezeichnet „Latenzzeit" die Zeit, die der Speichercontroller benötigt, dem RAM zu befehlen, auf einen bestimmten Speicherort zuzugreifen, bis zu dem Zeitpunkt, zu dem die Daten an diesem Speicherort tatsächlich gelesen werden.

Da es bei der Latenzzeit vor allem um den zeitlichen Abstand zwischen der Eingabe eines Befehls und seiner Ausführung geht, ist es wichtig zu verstehen, was während dieser Zeit geschieht. Nachdem der Speichercontroller dem RAM den Befehl gegeben hat, auf einen bestimmten Speicherort zuzugreifen, durchlaufen die Daten eine bestimmte Anzahl von Taktzyklen im Column Address Strobe, um an den gewünschten Speicherort zu gelangen und den Befehl „vollständig" auszuführen. Vor diesem Hintergrund gibt es zwei Variablen, wenn es darum geht, die Latenzzeit eines bestimmten Moduls zu ermitteln: **(1)** die Gesamtanzahl der Taktzyklen, die die Daten durchlaufen müssen (in Datenblättern angegeben als CAS-Latenz bzw. **CL**) und **(2)** die Dauer der einzelnen Taktzyklen (angegeben in Nanosekunden). Die genaue Formel sieht so aus:

Latenzformel

$$\text{Wahre Latenzzeit}^{(ns)} = \text{Taktzykluszeit}^{(ns)} \times \text{Anzahl der Taktzyklen}^{(CL)}$$

Das Latenz-Paradox

Die Latenzzeit wird häufig missverstanden, da sie in vielen Produktbroschüren und Vergleichen technischer Daten in CL angegeben wird. Das ist jedoch lediglich die Hälfte der Latenzformel. Da CL-Angaben nur die Gesamtzahl der Taktzyklen anzeigen, machen sie keinerlei Angaben zur Dauer der einzelnen Taktzyklen und sollten daher nicht als einziger Indikator für die Latenzleistung extrapoliert werden.

Das stellt uns vor das **Latenz-Paradox**. Schauen Sie sich *Abbildung 1 an.*

Abbildung 1

GESCHWINDIGKEIT IM VERGLEICH ZUR LATENZZEIT IM LAUFE DER ENTWICKLUNG DER SPEICHERTECHNOLOGIE
(BRANCHENSTANDARDS)

TECHNOLOGIE	MODULGESCHWIN-DIGKEIT (MT/s)	TAKTZYKLUSZEIT (ns)	CAS-LATENZ (CL)	WAHRE LATENZ (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

Im Laufe der Entwicklung der Speichertechnologie hat sich die Geschwindigkeit erhöht, während sich die Taktzykluszeiten sogar verkürzt haben. Das hat im Zuge der Entwicklung der Technologie zu kürzeren wahren Latenzzeiten geführt, auch wenn immer mehr Taktzyklen durchlaufen werden.

Nanosekunden:

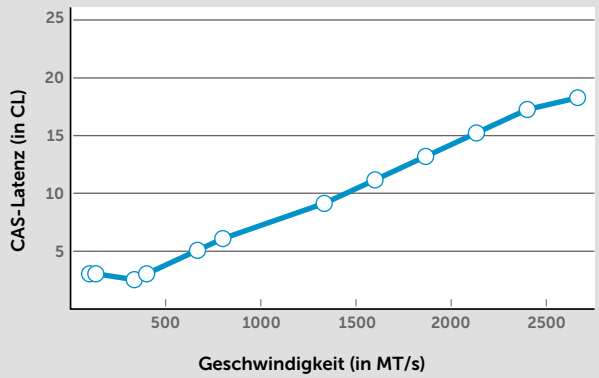
Eine bessere Messgröße für die Latenzleistung

Da es bei der Latenzzeit darum geht, wie lange es dauert, bis der Speicher einen eingegebenen Befehl ausführt, sollte sie am besten in reinen Nanosekunden an anstatt in CL (also in der Anzahl der Taktzyklen anstelle der Dauer ihrer Ausführung) gemessen werden. Wenn Sie sich die Latenzzeit eines Moduls in Nanosekunden anschauen, können Sie besser beurteilen, ob ein Modul tatsächlich schneller reagiert als ein anderes. Um die wahre Latenzzeit eines Moduls zu berechnen, multiplizieren Sie die Dauer der Taktzyklen mit der Gesamtanzahl der Taktzyklen. Diese Zahlen werden in die offizielle technische Dokumentation im Datenblatt eines Moduls aufgenommen.

So sehen die meisten Menschen die Latenzzeit

Mit der Geschwindigkeit erhöht sich auch die Latenzzeit.

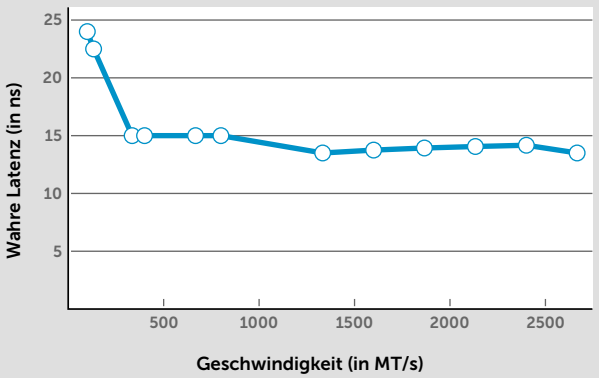
GESCHWINDIGKEIT IM VERGLEICH ZUR CAS-LATENZ



So sehen Halbleiteringenieure die Latenzzeit

Mit zunehmender Geschwindigkeit verkürzen sich die Latenzzeiten und/oder bleiben in etwa gleich (etwa 14 ns).

GESCHWINDIGKEIT IM VERGLEICH ZUR WAHREN LATENZZEIT



Vergleicht man die Geschwindigkeit mit der wahren Latenzzeit, sieht man leicht, dass sich die Latenzzeiten während der Weiterentwicklung der Speichertechnologie nicht wirklich verlängert haben. Und da die Geschwindigkeit sich erhöht, während die wahren Latenzzeiten in etwa gleich bleiben, können Sie mit neuerem, schnellerem und energieeffizienterem Speicher außerdem mehr Leistung erzielen.

An diesem Punkt in der Diskussion müssen wir feststellen, dass wir, wenn wir sagen „die wahren Latenzzeiten bleiben in etwa gleich" meinen, dass die wahren Latenzzeiten vom DDR3-1333 bis zum DDR4-2666 (die Spanne des modernen Speichers) bei 13,5 ns begannen und dann wieder zu 13,5 ns zurückkehrten. Auch wenn es innerhalb dieser Spanne mehrmals vorgekommen ist, dass die Latenzzeiten sich verlängert haben, handelte es sich bei dieser Zunahme um Bruchteile von Nanosekunden. Innerhalb derselben Spanne sind die Geschwindigkeiten um mehr als 1.300 MT/s gestiegen und haben damit effektiv die geringfügigen Zunahmen bei der Latenzzeit wieder aufgehoben.

Wenn Sie sich jedoch noch immer über das allgemeine Prinzip Gedanken machen – dass Latenzzeiten sich verlängern, wenn auch nur geringfügig – ist dies die technische Erklärung, warum das Branchenstandard ist.

Darum hängen Geschwindigkeit und Latenzzeit zusammen

Um gleichbleibend schnelle Reaktionszeiten zu gewährleisten, müssen CL-Angaben typischerweise zusammen mit der Frequenz ansteigen, damit eine durchschnittliche Zugriffszeit von etwa 14 ns erhalten bleibt.* Das ist wichtig, denn wenn CL-Angaben nicht mit jeder Kadenz ansteigen würden, würde **(a)** die Datenmenge nicht zunehmen, **(b)** die Speicherleistung mit den höchsten Geschwindigkeiten/der kürzesten Latenzzeit beeinträchtigt werden oder **(c)** die physische Größe der Speichermodule deutlich zunehmen. Diese drei Dinge würden den Speicher für die Endbenutzer erheblich verteuern. Daher werden JEDEC-Branchenstandards normalerweise vom Markt festgelegt, um die Massenproduktion kostengünstiger Speichermodule für höhere Leistungssteigerungen im realen Betrieb zu ermöglichen.

Das Fazit

Bei der Speicherleistung kommt es auf das Verhältnis zwischen Geschwindigkeit und Latenzzeit an. Installieren Sie so viel Speicher wie möglich, verwenden Sie die neueste Speichertechnologie und wählen Sie Module mit einer so hohen Geschwindigkeit, wie sie kostengünstig und/oder für die von Ihnen verwendeten Anwendungen relevant ist, um eine optimale Leistung zu erzielen. Im Allgemeinen ist die wahre Latenzzeit in etwa gleich geblieben, während sich die Geschwindigkeit erhöht hat. Das bedeutet, dass Sie dank höherer Geschwindigkeiten eine bessere Leistung erreichen können. Die wahren Latenzzeiten sind nicht notwendigerweise länger geworden, sondern lediglich die CAS-Latenzen. Außerdem sind CL-Angaben ein ungenauer und häufig irreführender Indikator für die wahre Latenzleistung.

*Wir streben stets eine möglichst kurze Zugriffszeit an. Die derzeitigen Zugriffszeiten werden sich verändern, wenn die Speichertechnologie und/oder die Prozesse weiterentwickelt werden.

ÜBERARBEITETE VERSION:
28/04/2015