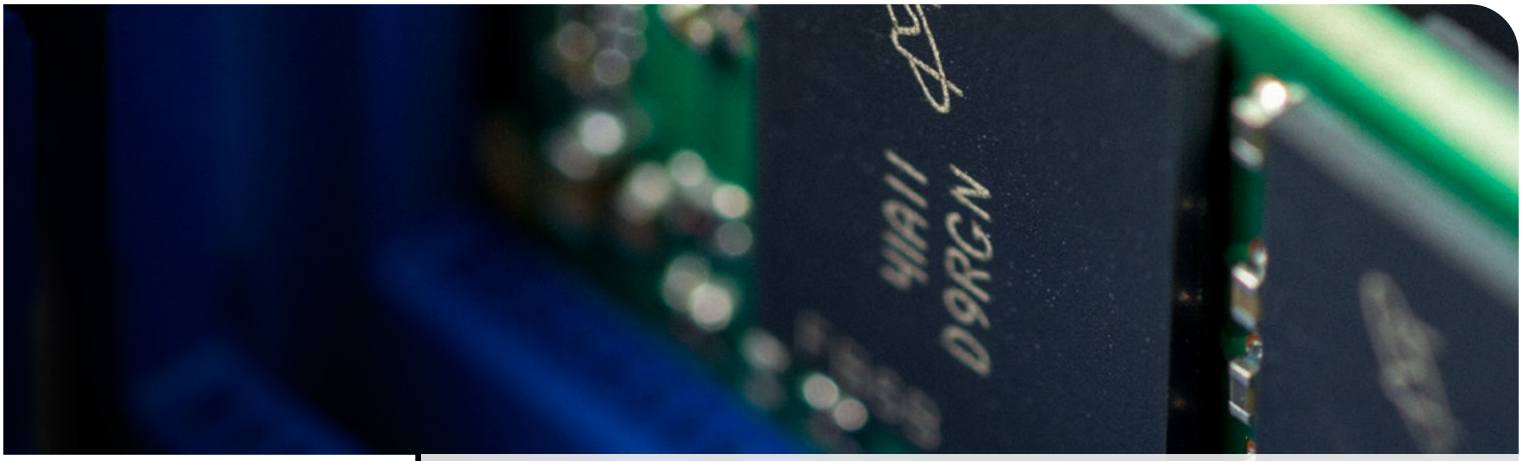


crucial by Micron

The memory & storage experts™

Vitesse ou Latence

Pourquoi la latence CAS n'est pas un bon indicateur des performances de la mémoire

Vitesse ou Latence

Bien que la vitesse et la latence soient toutes deux liées aux performances de la mémoire, ce lien n'est pas forcément évident. Pour la plupart des utilisateurs, la relation entre la vitesse et la latence est la suivante : lorsque la vitesse augmente, la latence fait de même. Cependant, ce n'est pas forcément le cas. En fait, cette vision est trompeuse et peut inciter les utilisateurs à réduire le niveau des performances de leur système. Nous allons donc découvrir les véritables liaisons entre la vitesse et la latence, pour mieux comprendre en quoi elles affectent les performances de la mémoire.

Défini par la vitesse

La vitesse est un concept facile à comprendre : il s'agit de la mesure de la rapidité de traitement des données par une barrette de RAM. La vitesse est mesurée en megatransfers par seconde (MT/s). Elle doit toujours être la plus élevée possible pour un coût réduit. Dans le secteur de la mémoire, chaque nouvelle avancée technologique va de pair avec une augmentation de la vitesse.

Défini par la latence

La latence est bien plus complexe à appréhender que la vitesse. En conséquence, elle est souvent mal comprise. Pour généraliser, on peut dire que la latence correspond au temps écoulé entre la saisie d'une commande et son exécution. Il s'agit de l'intervalle entre ces deux événements. D'un point de vue plus technique, la latence mesure le temps qu'il faut au contrôleur de mémoire pour ordonner à la RAM d'accéder à un emplacement précis, avant que les données de cet emplacement soient lues.

Comme la latence se limite à l'intervalle se déroulant entre la saisie d'une commande et son exécution, il est essentiel de bien comprendre ce qui se produit pendant ce temps. Une fois que le contrôleur de mémoire ordonne à la RAM d'accéder à un emplacement précis, les données doivent passer par un nombre défini de cycles d'horloge dans le CAS (Column Address Strobe, pour Temps d'accès à une colonne) pour atteindre l'emplacement désiré et « exécuter » la commande. En sachant cela, il faut donc tenir compte de deux variables pour déterminer la latence d'un module donné : **(1)** le nombre total de cycles d'horloge par lesquels les données doivent passer (mesuré en **CL**, pour latence CAS, ou Latence de temps d'accès à une colonne, sur les fiches techniques), et **(2)** la durée de chaque cycle d'horloge (mesurée en nanosecondes). Voici la formule exacte :

Formule de la latence

véritable latence$^{(ns)}$ = durée du cycle d'horloge$^{(ns)}$ x nombre de cycles d'horloge $^{(CL)}$

Le paradoxe de la latence

La latence est souvent mal comprise, car sur la plupart des prospectus publicitaires et des comparaisons de caractéristiques techniques, elle est indiquée en CL, une valeur correspondant uniquement à la moitié de l'équation déterminant la véritable latence. Comme les évaluations CL indiquent uniquement le nombre total de cycles d'horloge, elles ne tiennent pas compte de la durée de chaque cycle d'horloge et ne doivent donc pas être tenues pour seul indicateur des performances de latence.

Cela constitue le **paradoxe de la latence**. Veuillez consulter le *tableau 1*.

Tableau 1

VITESSE CONTRE LATENCE AU FUR ET À MESURE DES AMÉLIORATIONS TECHNOLOGIQUES
(NORMES DU SECTEUR)

TECHNOLOGIE	VITESSE DU MODULE (MT/s)	DURÉE DU CYCLE D'HORLOGE (ns)	LATENCE CAS (CL)	VÉRITABLE LATENCE (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

Au fur et à mesure des évolutions technologiques de la mémoire, la vitesse augmente et les durées de cycles d'horloge baissent, ce qui réduit la véritable latence, même si le nombre de cycles d'horloge augmente.

Nanosecondes :

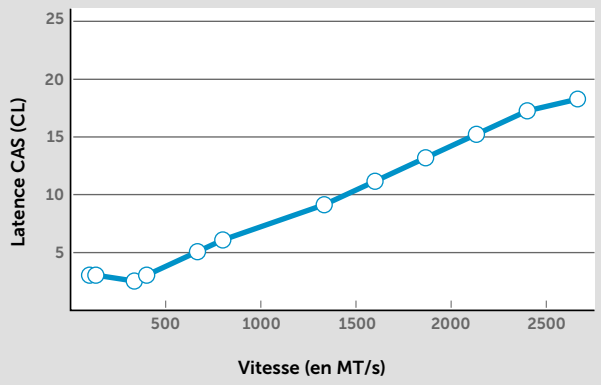
Une meilleure mesure des performances de latence

Comme la latence correspond au temps nécessaire pour que la mémoire exécute une commande saisie, il vaut mieux la mesure en nanosecondes qu'en CL (puisque ce dernier type de mesure se focalise sur le nombre de cycles d'horloge au lieu du temps nécessaire à leur réalisation). En mesurant la latence d'un module en nanosecondes, vous pouvez mieux juger si un module est, en fait, plus réactif qu'un autre. Pour calculer la véritable latence d'un module, multipliez la durée d'un cycle d'horloge par le nombre total de cycles d'horloge. Ces nombres figurent dans la documentation d'ingénierie officielle de la fiche technique d'un module.

Perception de la latence pour la plupart des utilisateurs

Lorsque la vitesse augmente, la latence fait de même

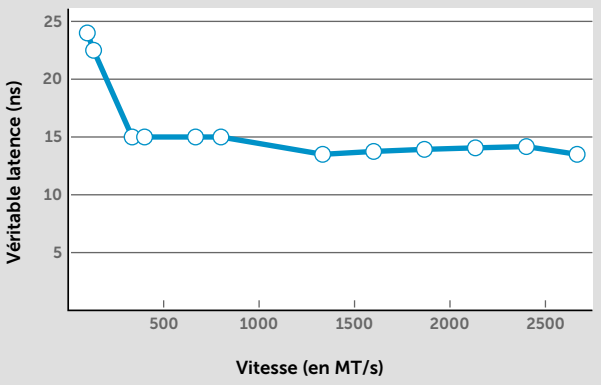
VITESSE CONTRE LATENCE CAS



Perception de la latence pour les ingénieurs de fabrication de semi-conducteurs

Lorsque la vitesse augmente, la latence diminue ou varie peu (environ 14 ns)

VITESSE CONTRE VÉRITABLE LATENCE



Lorsque la vitesse est comparée à la véritable latence, il est alors facile de s'apercevoir que, tandis que la technologie de la mémoire s'améliore, la latence n'augmente pas vraiment. De plus, comme la vitesse augmente et que la véritable latence reste globalement la même, vous pouvez améliorer vos performances en utilisant des mémoires plus récentes, plus rapides et plus économiques en énergie.

À présent, il nous faut préciser qu'en affirmant « La véritable latence reste globalement la même », nous voulons dire qu'entre la DDR3-1333 et la DDR4-2666 (la dernière évolution de la mémoire) la mesure de la véritable latence l'a d'abord située à 13,5 ns et après diverses évolutions, elle vient de revenir à ce niveau. Même si dans certains cas, la véritable latence a pu augmenter, il ne s'agit que d'une fraction de nanosecondes supplémentaire. Et pour ces mêmes cas, la vitesse a augmenté de plus de 1 300 MT/s, ce qui a contrebalancé cette augmentation.

Cependant, si vous vous souciez toujours de la règle générale stipulant que la latence peut toujours subir une augmentation, aussi faible qu'elle soit, voici une explication technique sur le fait que ceci soit une norme du secteur.

Pourquoi la vitesse et la latence sont liées

Pour garantir une réactivité rapide et cohérente des mémoires actuelles, les évaluations CL doivent généralement augmenter avec la fréquence pour maintenir un temps d'accès d'environ 14 ns.* Cela est important, car si les évaluations CL n'augmentent pas avec chaque cadence, alors : **(a)** les taux de données ne pourraient pas augmenter, **(b)** le volume de mémoire serait affecté par les vitesses les plus rapides et/ou les latences les plus basses, ou **(c)** la taille physique des modules de mémoire devrait nettement augmenter. Pour les utilisateurs finaux, une seule de ces trois conséquences entraînerait une augmentation significative de la mémoire. C'est pourquoi les normes JEDEC du secteur sont généralement définies par le marché pour permettre la production en masse de modules économiques qui améliorent les performances en situation réelle.

L'objectif

Les performances de mémoire reposent uniquement sur la relation entre la vitesse et la latence. Pour obtenir des performances optimales, installez autant de mémoire que possible, utilisez les technologies de mémoire les plus récentes et choisissez les modules aussi rapides que possible, compte tenu des besoins de vos applications et de votre budget. En règle générale, lorsque la vitesse augmente, la latence reste approximativement la même, ce qui veut dire qu'une augmentation de la vitesse vous permet d'améliorer votre niveau de performance. La véritable latence n'a pas forcément augmenté, contrairement à la latence CAS. Les évaluations CL ne sont donc pas un bon indicateur des véritables performances de latence.

*Notre objectif est toujours de réduire au maximum le temps d'accès. Les avancées technologiques et/ou de traitement de la mémoire modifient les temps d'accès actuels.

VERSION : 04/05/2015