Technologies des mémoires informatiques

Pierre Boulet

Professeur d'informatique à l'Université de Lille, CRIStAL

Le numérique est sur toutes les lèvres aujourd'hui. Son universalité (et son nom !) vient du codage sous forme de nombres de tout type d'information, y compris les programmes. Ces nombres sont représentés sous forme de suites de « bits » (chiffres binaires pouvant prendre les valeurs 0 ou 1). Ainsi, toute technologie matérielle qui a 2 états peut être utilisée pour stocker tout type d'information.
Au-delà du simple stockage de l'information, pour être utile, une technologie de mémoire doit supporter plusieurs opérations, dont la lecture, l'écriture, la modification, la recherche, ou encore l'effacement de l'information qu'elle contient. Selon le rôle qu'on donne à une mémoire, elle sera modifiable ou pas, volatile ou pas (une mémoire volatile perdant l'information qu'elle contient quand on coupe son alimentation électrique). On peut distinguer 2 grandes catégories de mémoires informatiques :
la *mémoire de travail*, aussi appelée mémoire vive, dans laquelle on manipule les programmes en cours d'exécution et les données en cours d'utilisation. Une telle mémoire doit être modifiable, et peut être volatile ou pas ;
la *mémoire de stockage*, aussi appelée mémoire de masse, qui permet la conservation d'information à longue échéance. Une telle mémoire peut être modifiable ou pas, mais doit être non volatile.
Dans la suite de cet article, nous passerons en revue les principales technologies matérielles des mémoires informatiques, puis le rôle du logiciel dans la fonction de mémorisation, et finirons par présenter 4 évolutions qu'on peut anticiper à court ou moyen terme.

Panorama historique des technologies matérielles des mémoires informatiques

Mémoires de travail

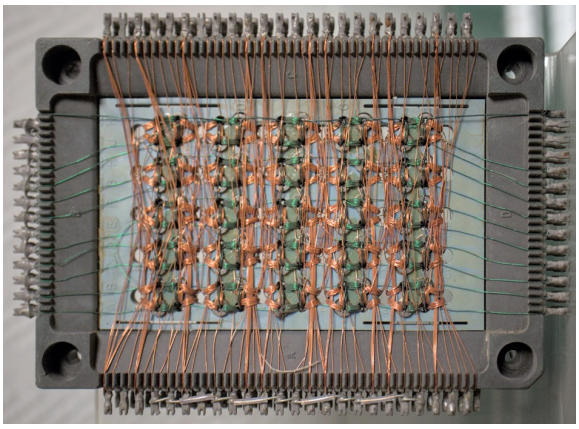
C'est vraiment l'apparition de la technologie des **tores magnétiques** qui a permis, par ses performances, de réaliser la vision de John Von Neumann de l'architecture (dite de Von Neumann) avec une mémoire de travail unifiée, séparée de l'unité de traitement et stockant les programmes et les données. Ces tores magnétiques se sont progressivement densifiés de 1955 à 1975. Ils ont été hégémoniques jusqu'en 1970, date de l'apparition des mémoires de travail sous forme de **semi-conducteurs** (en particulier la DRAM, *Dynamic Random Access Memory*) qui les ont complètement supplantés. Depuis 2008 et la découverte des memristors, on assiste à une diversification des technologies de semi-conducteurs utilisées pour faire des nouvelles mémoires non volatiles (ReRAM, PCM…), mais la bonne vieille DRAM apparue dans les années 1970 domine encore très largement le marché. En complément de la DRAM, les semi-conducteurs sont aussi mis à contribution dans les SRAM (*Static Random Access Memory*) qui permettent de faire des mémoires de travail plus rapides, mais moins denses et plus chères, comme les mémoires caches dont nous parlerons plus tard.

Mémoires de stockage

Les technologies de mémoire de stockage sont plus variées, mais peuvent être classées en 4 grandes familles de technologies qui utilisent des principes physiques différents : mécaniques, magnétiques, semi-conducteurs et optiques.

Les **cartes perforées** inventées par Jacquard en 1728 pour les machines à tisser ont été utilisées de 1945 à 1975 dans les ordinateurs, en particulier pour gérer les entrées-sorties, avant que nous ayons les claviers et les écrans.

Une autre technologie ancienne a été récupérée pour l'informatique : les **bandes magnétiques**. Inventées pour enregistrer et diffuser le son en 1928, elles sont utilisées depuis 1950 comme mémoire de stockage de grande densité et de longue durée de conservation. L'autre support utilisant des bits magnétiques est le **disque dur** apparu au milieu des années 1950, toujours utilisé aujourd'hui. Plus fragiles que les bandes magnétiques, les disques durs permettent l'accès direct à toutes les zones de stockage, là où l'accès est séquentiel dans une bande magnétique (il faut dérouler la bande jusqu'à l'endroit où se trouve l'information qui nous intéresse). Ces disques sont aussi plus chers que les bandes



Mémoire à tores magnétiques de 200 bits, musée de la DGDNum, Université de Lille, photo CC BY-SA Pierre Boulet

magnétiques car ils sont plus complexes à fabriquer. Enfin, une technologie simplifiée et amovible de stockage sur disque magnétique a été très largement utilisée entre 1971 et le début des années 2000 : la **disquette**.

Depuis 1988, les **semi-conducteurs** reviennent en force avec les technologies à base de mémoire flash, une technologie de stockage non volatile, plus rapide, mais plus chère et moins dense que les disques durs. On trouve des mémoires flash un peu partout : dans les disques SSD, les clés USB et les mémoires des téléphones.

Enfin, les technologies de stockage sur **support optique (CD-ROM, DVD-ROM...)** ont eu leur période d'utilisation en informatique entre 1980 et 2010 avant de ne subsister aujourd'hui que pour la diffusion de musique ou de vidéo (CD, DVD, Blu-Ray).

Rôle du logiciel

Comme nous venons de le voir, il y a finalement assez peu de technologies matérielles qui sont utilisées pour réaliser les mémoires utiles à l'informatique. Ceci vient du fait que l'on recherche des performances élevées en termes de

- rapidité, à la fois en temps de réponse et en débit, pour les opérations de lecture et d'écriture ;
- capacité ou densité (soit la quantité d'information qu'on peut y stocker) ;
- rémanence, robustesse et fiabilité (durée de conservation de l'information, nombre de cycles de lecture ou d'écriture avant détérioration) ;
- facilité d'accès à l'information (accès séquentiel, par adresse ou associatif) ;
- coût de fabrication et d'usage.

Aucune des technologies mentionnées précédemment ne permet de réaliser une grande mémoire rapide, robuste, non volatile et bon marché. Et pourtant, nous utilisons des ordinateurs au quotidien en ayant l'impression de travailler avec une telle mémoire de grande taille, rapide, très fiable, et à un coût sans cesse déclinant. Ceci est possible grâce au logiciel qui masque la variété des technologies utilisées dans les machines pour permettre aux programmeurs et aux utilisateurs de faire comme si chaque application ou processus avait à sa disposition une mémoire non seulement très grande et rapide, mais en plus isolée de celle des autres applications ou processus.

Les grandes fonctions du logiciel (principalement le système d'exploitation) concernant la mémoire sont :

- abstraire (ou virtualiser) le matériel, ce qui rend les programmes portables et masque la complexité matérielle ;
- permettre le partage des ressources matérielles, en isolant les processus et les utilisateurs ;
- organiser l'espace de stockage, en proposant des modes d'adressage, des systèmes de fichiers, des bases de données ;
- sécuriser l'information, en matière de confidentialité, intégrité et disponibilité.

Nous ne détaillerons pas ici tous ces mécanismes, mais attardons-nous tout de même sur celui des mémoires caches qui permet de réaliser une grande mémoire rapide avec une grande mémoire lente et une petite mémoire rapide. L'idée est simple : intercaler entre le processeur et la grande mémoire lente, une petite mémoire rapide qui va « cacher » l'information. Quand le processeur veut accéder à une cellule mémoire, il va d'abord regarder si le contenu de cette cellule est présente dans la **mémoire cache**. Si elle y est, l'accès est rapide, si elle n'y est pas, la mémoire cache va la chercher dans la grande mémoire lente et l'enregistre avant de la retransmettre au processeur. Ainsi, si on accède plusieurs fois à une même cellule mémoire, on travaille à la vitesse de la petite mémoire rapide. La raison pour laquelle cette astuce fonctionne si bien, est que les motifs d'accès à la mémoire ont en général deux bonnes propriétés : la **localité temporelle** et la **localité spatiale**. En effet, on a tendance à réutiliser la même adresse à des instants rapprochés d'une part, et, d'autre part, les programmes et les données sont le plus souvent stockés de manière contiguë en mémoire. Ainsi, si on accède à la grande mémoire lente par bloc, on précharge dans la petite mémoire rapide des zones de la mémoire qui ont de grandes chances d'être réutilisées par la suite. Ce principe fonctionne tellement bien qu'on a aujourd'hui des hiérarchies mémoires avec 3 niveaux de cache en plus de la mémoire interne au processeur (les registres) et de la mémoire de travail principale. Chaque niveau de cache a sa propre technologie, taille, taille de bloc, politique de remplacement (quand on veut ajouter une donnée dans un cache plein, il faut choisir laquelle on supprime), mode d'écriture, etc. Le diable (et les performances) est dans les détails !

Ces caches matériels sont gérés par des circuits matériels dédiés, mais les mêmes mécanismes existent dans le monde purement logiciel pour étendre la mémoire de travail avec la mémoire de stockage (mécanisme de mémoire virtuelle ou *swap*), ou pour accéder plus rapidement à des données distantes (web, bases de données, vidéo en flux, etc.).

Et demain ?

Voici une sélection personnelle de 4 technologies de stockage (3 matérielles et 1 logicielle) qui pourraient bien avoir un impact important dans un futur plus ou moins proche.

Architectures neuromorphiques

Ces architectures de traitement de l'information s'inspirent du fonctionnement du cerveau pour définir des circuits électroniques calculant sur des trains d'impulsions électriques avec des réseaux de neurones impulsionnels. Nous sommes donc ici dans le monde de l'informatique analogique, et non plus numérique. Les gains espérés de l'utilisation de telles architectures sont d'abord leur ultra faible consommation d'énergie, obtenue avec des technologies de fabrication existantes. Ces gains de plusieurs ordres de grandeurs de consommation viennent du fait qu'on ne sépare plus le stockage du calcul. Les deux sont localisés dans les synapses artificielles qui peuvent être réalisées avec des memristors, comme les nouvelles technologies de mémoire non volatile. En cassant l'architecture de Von Neumann, on doit cependant complètement repenser le traitement de l'information en le basant sur l'apprentissage, et non plus la programmation.

Stockage chimique ou sur ADN

L'idée est ici d'utiliser des polymères de grande stabilité pour stocker l'information. La disponibilité de séquenceurs d'ADN à haut débit pour la lecture et de techniques de synthèse chimique performantes rend envisageable de telles solutions pour de l'archivage à très grande durée de vie. Les points forts de ces technologies sont leur densité et leur durée de rétention de l'information extrêmes, et leur défaut majeur est que - le polymère encodant directement l'information- il n'y a pas de substrat : on doit donc encoder en même temps le contenu et le moyen de le retrouver, ce qui est particulièrement complexe. De ce fait, les vitesses de lecture et d'écriture sont encore très lentes.

Informatique quantique

L'informatique quantique repose sur un stockage de l'information dans des « qubits » qui peuvent porter chacun la superposition de 2 états. L'intérêt majeur de cette façon de stocker l'information est qu'elle permet la parallélisation massive des calculs, et donc permet d'accélérer dramatiquement certains calculs. Le revers de la médaille est que ces technologies souffrent de nombreux défis technologiques non résolus, dont la stabilité des qubits dans le temps, et la fabricabilité à grande échelle. Enfin, l'algorithmique quantique est complexe et ne permet pas aujourd'hui d'envisager un ordinateur quantique universel. Nous resterons, au moins pour quelques années, dans des cas d'usages spécifiques.

Registres décentralisés et technologies blockchains

Les registres décentralisés et les technologies blockchains sont une technologie logicielle apparues en 2007 avec le Bitcoin. Ils répondent à la problématique de créer un registre (espace de stockage où on ne peut qu'ajouter des informations comme un livre de comptes) infalsifiable, sans tiers de confiance. Ces technologies reposent massivement sur la cryptographie, les réseaux pair-à-pair, les algorithmes de consensus issus des recherches sur les systèmes distribués. Outre leur utilisation première pour réaliser des monnaies électroniques indépendantes des banques (centrales ou commerciales), elles ont des cas d'usage nombreux dès lors que des entités veulent échanger de l'information sans se faire confiance, ni faire confiance à un tiers. La confiance vient alors de la transparence de la technologie qui permet la vérification par toutes les parties que les échanges ou les transactions se passent comme prévu. Le logiciel permet ici de réaliser une mémoire gérée de manière décentralisée, infalsifiable, et de très grande durée de vie.

Pour résumer et conclure, j'aimerais insister sur l'universalité du codage binaire de l'information qui a rendu possible l'utilisation de technologies diverses pour stocker l'information. Ces technologies ayant des propriétés variées, les informaticiens ont développé des abstractions logicielles qui permettent de masquer la complexité de l'assemblage de technologies qui constituent les mémoires de nos ordinateurs. Il n'y a pas de stockage performant sans une combinaison de matériel et de logiciel.

Robot d'archivage à bandes magnétiques, DGDNum, Université de Lille, photo CC BY-SA

Pour aller plus loin

L'exposition numérique mémoire et stockage du *Computer History Museum* : https://www.computerhistory.org/revolution/memory-storage/8

Le modèle d'architecture de Von Neumann, par Sacha Krakowiak sur Interstices, 2011 : https://interstices.info/le-modele-darchitecture-de-von-neumann/

Et plus vite si affinités…, par Brice Goglin sur Interstices, 2011 : https://interstices.info/et-plus-vite-si-affinites/

Demain, un ordinateur inspiré de notre cerveau ?, par Hugo Leroux dans le journal du CNRS, 2018 : https://lejournal.cnrs.fr/articles/demain-un-ordinateur-inspire-de-notre-cerveau

La course aux bits quantiques, par Tristan Meunier sur Interstices, 2021 : https://interstices.info/la-course-aux-qubits/

Stocker les données : la piste prometteuse de l'ADN, par Dominique Lavenier, Marc Antonini, Anthony Genot & Yannick Rondelez sur Interstices, 2023 : https://interstices.info/stocker-les-donnees-la-piste-prometteuse-de-ladn/

Technologies blockchains en 20 min, par Pierre Boulet sur Hive, 2024 : https://peakd.com/hive-114606/@pboulet/technologies-blockchains-en-20-min