Opinions Libres
le blog d'Olivier Ezratty

Le mystère MemComputing

La veille technologique a ceci de merveilleux que l'on tombe régulièrement à la fois sur de véritables bijoux comme le processeur de **Cerebras** que j'ai pu décrire en détail dans un **article précédent** et sur des prouesses qui génèrent un doute profond, et qui entraînent une indescriptible envie de faire un bon fact-checking scientifique.

En travaillant sur la mise à jour annuelle de mes gros ebooks sur l'informatique quantique et sur l'intelligence artificielle, à paraître respectivement en septembre et novembre 2019, je suis tombé sur la mystérieuse startup **MemComputing** (2016, USA). Elle développe un processeur de calcul qui pourrait accélérer significativement les calculs de problèmes complexes dont celui de l'entraînement de modèles de deep learning.

On peut la ranger dans une catégorie voisine de l'offre de **Fujitsu** et de son supercalculateur utilisant le recuit digital à température ambiante. Ce dernier propose une solution qui s'apparente à celle du Canadien D-Wave pour faire des calculs complexes, mais sans passer par du quantique. La technologie de Fujitsu dénommée "Digital Annealer" est développée sur silicium en CMOS et en partenariat avec l'Université de Toronto. Elle serait déjà proposée dans le cloud. Elle sert à résoudre des problèmes d'optimisation et notamment à réaliser du criblage de molécules dans les biotechs. Ce "Digital Annealer" est un chipset dédié comportant 1 024 blocs de mise à jour de bits intégrant de la mémoire pour stocker leurs poids et avec une précision de 16 bits, des blocs logiques pour réaliser des inversion de valeurs, et les circuits de contrôle associés. Voir **Fujitsu's CMOS Digital Annealer Produces Quantum Computer Speeds**, 2018.

L'Américain **MemComputing** met au point le **MemCPU Coprocessor**, une solution matérielle qui consiste à placer la mémoire près des unités de calcul dans des unités de traitement. Elle est décrite dans **Memcomputing: fusion of memory and computing** par Yi Li & Al, 2017 (3 pages) d'où provient ce schéma. Le rapprochement de la mémoire du calcul est une approche choisie par une bonne part des acteurs des composants de calcul neuromorphique qui cherchent à accélérer les traitements du deep learning, aussi bien dans les phases d'entraînement que d'inférence de ces réseaux de neurones complexes.

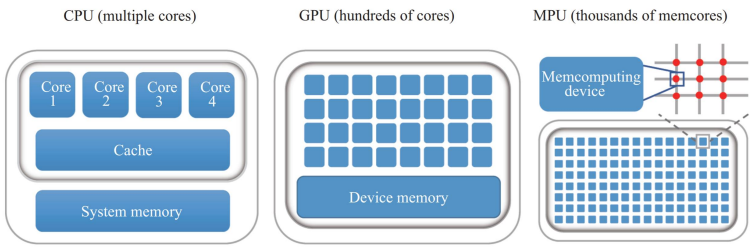
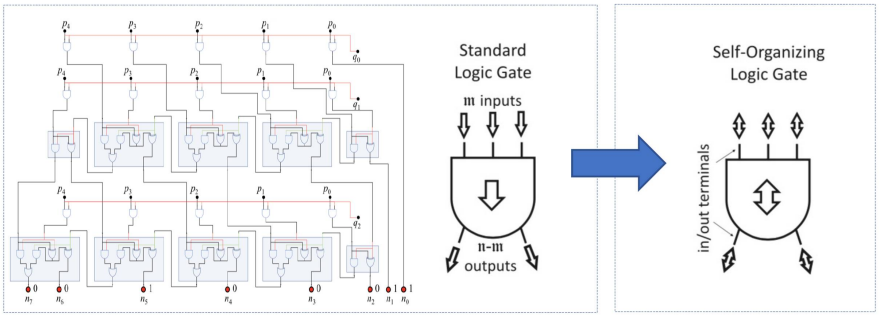


Figure 1 (Color online) Comparison of CPU, GPU, and MPU.

Mais ce n'est pas tout. Chez MemComputing, ces cellules de calcul auraient des entrées et sorties symétriques et interconnectées aux cellules avoisinantes. Elles gèrent uniquement des nombres entiers. Cette organisation des unités de calcul permettrait de trouver automatiquement un équilibre complexe d'un système paramétré.

C’est le principe des SOLGs (Self Organizing Logic Gates) du schéma *ci-dessous* qui provient de la **littérature** de MemComputing.



A traditional Boolean circuit using logic Gate for prime factorization
It multiplies two integers p and q to give pq = n = 35 = $(100011)_2$ (in the little-endian notation).

Moving from traditional gates to terminal-agnostic gates

Ce genre d’architecture permettrait d’après ses concepteurs de résoudre diverses classes de problèmes dits “NP-complets” et “NP-difficiles” en temps polynomiaux. En clair, cela servirait à transformer des problèmes complexes qui requièrent un temps exponentiel de calcul en fonction de la taille du problème à résoudre en solutions exécutables en un temps polynomial. C’est le Graal des mathématiques et du calcul ! C’est ce à quoi doivent servir les ordinateurs quantiques, une fois qu’ils seront au point à grande échelle, soit avec un grand nombre de qubits.

Les fondateurs de MemComputing annoncent des gains de performance significatifs qui seraient de quatre ordres de grandeur pour les applications de machine learning, donc une performance multipliée par 10 000 ! Voir **Memcomputing NP-complete problems in polynomial time using polynomial resources and collective states**, par Fabio Traversa, Massimilio Di Ventra & Al, 2014 (10 pages) et **Evidence of an exponential speed-up in the solution of hard optimization problems**, Fabio Traversa & Al, 2018. Enfin, voir cette **Conférence** de Massimiliano Di Ventra à Berkeley en 2016 (26 minutes).

La startup a été créée par le serial entrepreneur **John Beane** et deux chercheurs en physique, **Massimiliano Di Ventra** et **Fabio Traversa** qui ont une accumulé dense bibliographie sur le sujet du « memory computing » dont ils sont à l’origine. Voir **Universal Memcomputing Machines**, par Fabio Traversa et Max Di Ventra, 2014 (14 pages) et **Perspective: Memcomputing: Leveraging memory and physics to compute efficiently**, par Fabio Traversa et Massimilio Di Ventra 2018 (16 pages).

MemComputing annonce aussi résoudre des problèmes de planification et d’optimisation comme celui du voyageur de commerce, de combinatoire, de bioinformatique, d’entraînement de réseaux de neurones et même de factorisation de nombres entiers, à chaque fois, avec un gain exponentiel sur le temps de calcul. Tout ceci est documenté dans **Stress-testing memcomputing on hard combinatorial optimization problems** par Fabio Traversa, Max Di Ventra & Al, 2018 (6 pages), **Accelerating Deep Learning with Memcomputing** par Haik Manukian, Fabio Traversa et Massimiliano Di Ventra, 2018 (8 pages) et **Polynomial-time solution of prime factorization and NP-hard problems with digital memcomputing machines**, par Fabio Traversa et Max Di Ventra, 2017 (22 pages).

Pour l’instant, leur solution est émulée dans des ordinateurs classiques et fournie sous forme de SDK opéré dans le cloud qu’ils ont conçu en partenariat avec **Canvass Labs** (2017, USA). Leur composant électronique n’est pas encore fabriqué, même à l’état de prototype, et il n’est d’ailleurs pas évident de déterminer s’il est possible de le fabriquer.

Ils ont communiqué sur le traitement des problèmes de type **MIPLIB** (Mixed Integer Programming Library) qui sont considérés comme intractables (impossibles à résoudre de manière traditionnelle) avec un temps de réponse de 60 secondes sur un serveur tournant sous Linux. Le tout en battant au passage un ordinateur à recuit

quantique de D-Wave. La résolution de ce problème mathématique consiste à trouver par exemple une combinaison de nombres entiers donnés pouvant générer zéro une fois additionnés (le “Subset Sum problem”).

Traduisons tout cela en langage clair : cette startup prétend donc avoir trouvé une solution permettant de faire des calculs aussi rapidement que les ordinateurs quantiques qui ne sont pas encore au point et sur lesquels le monde entier a investi des milliards de dollars avec un espoir de réussir à obtenir des résultats probants dans une fenêtre de tir située entre 10 et 50 ans ! Et là, pas besoin de cryogénie à 15 mK et de frigo à 1M€. Pas de problème de décohérence quantique et d’erreurs à corriger avec de complexes systèmes de correction d’erreurs augmentant le nombre de qubits nécessaires de plusieurs ordres de grandeur. Cela semble donc énorme !

Bref, c’est tout bonnement révolutionnaire.

Ou pas.

Alors, il y a-t-il un ou plusieurs lézards ? S’il y en a, ils ne sont pas faciles à identifier pour le profane.

Les travaux de recherche des fondateurs sont publics depuis cinq ans. Il y a donc un peu de recul. Au-delà de la reprise par les médias-moutons de la communication officielle de la startup, il n’est pas évident de trouver de la matière en termes de fact-checking sur un tel projet. La barrière est placée assez haut : il faut connaître les théories de la complexité, l’électronique des SOLGs, les algorithmes et pouvoir déchiffrer les publications scientifiques des auteurs. Je vous avoue que cela dépasse mes limites en l’état.

J’ai commencé à douter de la capacité réelle à créer ces SOLGs dans des composants CMOS actuels, qui n’est pas démontrée. Ils n’ont pas encore créé de prototype ce qui fait douter un peu. Pourtant, le procédé employé par Fujitsu n’a pas l’air d’être très éloigné et ils arriveraient à en tirer quelque chose.

Autre source de doute : MemComputing arrive à obtenir un avantage d’échelle “quantique” avec l’émulation de son procédé sur des processeurs traditionnels. Cela a l’air bien séduisant mais, indirectement, revient à remettre en cause toutes les théories actuelles de la complexité. Cela impliquerait que P=NP au niveau des classes de complexité de problèmes, une égalité sujette à caution, le consensus des chercheurs en théorie de la complexité penchant plutôt pour une inégalité. Sachant que la question est posée depuis 1972 ! La majorité des chercheurs en complexité cherchent plutôt à démontrer que P est différent de NP. À savoir, que la classe des problèmes dont la vérification de la solution est réalisable en temps polynomial est différente de celle des problèmes qui ont une solution qui se calcule en temps polynomial. Bref, à en donner le tournis.

J’ai finalement trouvé quelques sources qui émettent des doutes sur le procédé de MemComputing :

- **Ken Steiglitz**, qui démontre rapidement un peu par l’absurde que cela n’est pas possible dans **A Note on `Memcomputing NP-complete problems…’ and (Strong) Church’s Thesis** 2015 (2 pages). Il fait référence à une tentative similaire datant de 1986 qui s’est soldée par un échec. Et il indique que si MemComputing tient ses promesses, soit P=NP, soit la thèse de Church-Turing est remise en cause : “*An (classical) analog device can be simulated by a Turing Machine using resources (including time) polynomial in the resources of the machine.”*. Okay.

- **Scott Aaronson** décrit les limites du modèle en 2017 dans **Memrefuting** . Ce n’est pas n’importe qui. C’est l’un des scientifiques les plus en vue de l’algorithmique quantique. En gros, l’approche de MemComputing ne scalerait pas à cause des phénomène de bruits induits dans les circuits. Les auteurs du procédé évoquent ce bruit mais pensent qu’ils peuvent passer outre : “*As anticipated, the machine we have built would ultimately be limited by unavoidable noise sources, thus requiring error-correcting codes. However we prove here that under the assumption of low noise, additive white noise does not affect the machine output”*

dans **Memcomputing NP-complete problems in polynomial time using polynomial resources and collective states**, 2015 (12 pages). Au demeurant, les ordinateurs quantiques dont la mise au point est très longue font aussi face à des problèmes de correction de bruit et d'erreurs.

- **Igor Markov** décrit d'autres lacunes du procédé de MemComputing dans **A review of « Memcomputing NP-complete problems in polynomial time using polynomial resources »** en 2015 (3 pages). En gros, en liaison avec des questions de ressources exponentielles cachées dans le procédé.

- **David Herrera Martí** qui fait de la recherche au Center for Excellence in Parallel Programming chez Atos Quantum à Grenoble pense que le procédé de MemComputing comprend plusieurs zones d'ombres. Notamment sur le fait que pour faire converger un système complexe vers un minimum global, il faut l'initialiser dans un état qui est proche de la solution à trouver, comme on le fait dans de nombreux algorithmes conçus pour les ordinateurs à recuit quantique de D-Wave. Et cette initialisation a un coût qui peut faire perdre une part de l'avantage obtenu dans le calcul. Les auteurs n'en parlent pas.

En gros, l'approche de MemComputing ne serait pas scalable ! Oubliez tout ce que vous venez de lire à leur sujet ! Ou alors, doutez des douteurs…

A-t-on pour autant affaire à des escrocs du calcul ou à des équivalents mathématiques de **Theranos**, cette startup américaine qui fait maintenant partie du groupe des escroqueries startupiennes à grande échelle, avec la mise en examen de sa fondatrice ? L'histoire est documentée par le journaliste John Carreyrou dans le livre "**Bad Blood**".

Ce n'est pas tout à fait la même situation. Tout d'abord, les chercheurs à l'origine de MemComputing sont plus sérieux et expérimentés que ne l'était Elisabeth Holmes lorsqu'elle avait créé Theranos après seulement une année d'études de licence en santé à Stanford. Ensuite, MemComputing n'a pas levé des montants significatifs. Ils ne sont même pas documentés, ce qui signifie qu'ils doivent être modestes. De nombreux travaux scientifiques ont été publiés, certains même dans des revues scientifiques à comité de relecture. J'ai même retrouvé une **trace** d'une intervention de Massimiliano Di Ventra au CEA en 2015.

Si cela se trouve, l'approche de MemComputing pourrait ne pas fonctionner mais tout de même inspirer la création d'autres architectures qui, elles, fonctionneraient.

On peut faire le lien avec les débats scientifico-sociétaux du moment comme celui qui a court autour du lien entre énergie nucléaire et mix énergétique décarbonné. Pour ou contre Greta, etc. Pour rentrer dans ces débats, il faut une certaine culture scientifique, hiérarchiser les problèmes, éviter le *whataboutism* et faire des comparaisons chiffrées objectives. S'il faut confronter les avis, cela serait plutôt entre experts scientifiques divergents qu'entre experts scientifiques et politiciens sans culture scientifique comme cela se produit souvent.

On doit toujours douter, mais il faut avoir un Pôle Nord. Le problème est qu'il n'y en a pas forcément qu'un seul. Donc, il faut débattre, mais posément et avec des raisonnements scientifiques et avec des faits, pas avec des émotions. C'est ça qui est le plus dur ! Heureusement cependant, le calcul exponentiel n'est pas encore devenu un sujet politique comme les questions environnementales.

Cet article a été publié le 9 septembre 2019 et édité en PDF le 17 mars 2024.
(cc) Olivier Ezratty – “Opinions Libres” – **https://www.oezratty.net**