Opinions Libres
le blog d'Olivier Ezratty

La dérive des exponentielles 1/3

Les prospectivistes qui sont nombreux dans les industries high-tech s'appuient souvent sur la notion de progrès exponentiel pour justifier leurs oracles. C'est notamment le cas de Ray Kurzweil, dans son fameux "**The Singularity is Near**" paru en 2005. Pour lui, le progrès scientifique et technique n'est pas linéaire mais exponentiel. On départ, quand les valeurs sont faibles, on ne le distingue pas du bruit ambiant. Puis à force de multiplications, il explose et transforme toutes les industries. Avec au centre de tout, le numérique.

Ces thèses du progrès exponentiel s'appuient sur la Loi de Moore et ses dérivés (Metcalfe, etc). Il y a en fait trois lois de Moore :

- La **première** selon laquelle la densité des semiconducteurs double tous les ans à coût constant.
- La **seconde**, la plus souvent évoquée, selon laquelle le nombre de transistors dans les microprocesseurs sur puce de silicium double tous les deux ans, et pas 18 mois comme souvent répété.
- La **troisième** qui est son application à n'importe quel domaine avec une grandeur technique ou économique (coût, densité, rapidité, stockage, etc) qui double ou est divisée par période comprise généralement entre 18 et 24 mois.

Certains n'hésitent pas à extrapoler ces lois à d'autres secteurs d'activités et technologies alors qu'il ne s'agit que de conjectures dont la validité est tout à fait empirique. D'autres emploient les courbes exponentielles qui les arrangent car on peut jouer avec tout un tas de paramètres pour faire passer le message. Comme les courbes ne sont pas parfaites, certains analystes inventent des explications empiriques expliquant la non linéarité des progressions (*ci-dessous*) par les phases de croissance et de maturité des technologies. Il faudrait dans ce cas faire la distinction entre les technologies disponibles dans des produits courants et celles qui sortent des laboratoires de recherche. Le passage à l'industrialisation a souvent été un obstacle dans la généralisation de nouvelles technologies. Cela a notamment été longtemps le cas des écrans de TV en technologie OLED.

Les lois exponentielles s'appliquent à tout un tas de domaines, comme le montre le tableau suivant qui liste le nombre de mois entre chaque facteur multiplicateur x2, une fois sur deux dans la dimension économique plus que dans la dimension technique. Et le tableau date un peu.

Fiber optic throughput	wavelengths per fiber	9
Optical network	$/bit	9
Wireless	bits per second	10
Communication	bits per dollar	12
Magnetic areal storage	gigabit/in2	12
digital cameras	pixels per dollar	12
Microprocessor	$ per cycle	13
Supercomputer power	flops	14
RAM	MiB/$	16
RAM	bits per dollar	18
DNA sequencing	$ per base pair	18
Transistor	$ per transistor	18
PCU Power consumption	watts/cm2	18
Pixels	per array	19
Harddrive Storage	Gigabyte per $	20
Chip	MIPS	21
DNA sequencing	$ per base pair	22
Trunk line data speeds	bits/sec	22
Microprocessor	transistors per	24
Chip processor	MHz/$	27
Bandwidth	kilobits per second per $	30
Microprocessor	hrtz	36

Cela aboutit à une vision très linéaire et prédictible de l'innovation technologique qui peut en fait aussi bien ralentir qu'accélérer brutalement. Certaines filières peuvent ainsi se satisfaire du respect de cette loi alors que le progrès pourrait être plus rapide. D'autres se font discrètes quand les progrès ne suivent plus cette loi. D'où l'intérêt d'y voir un peu plus près.

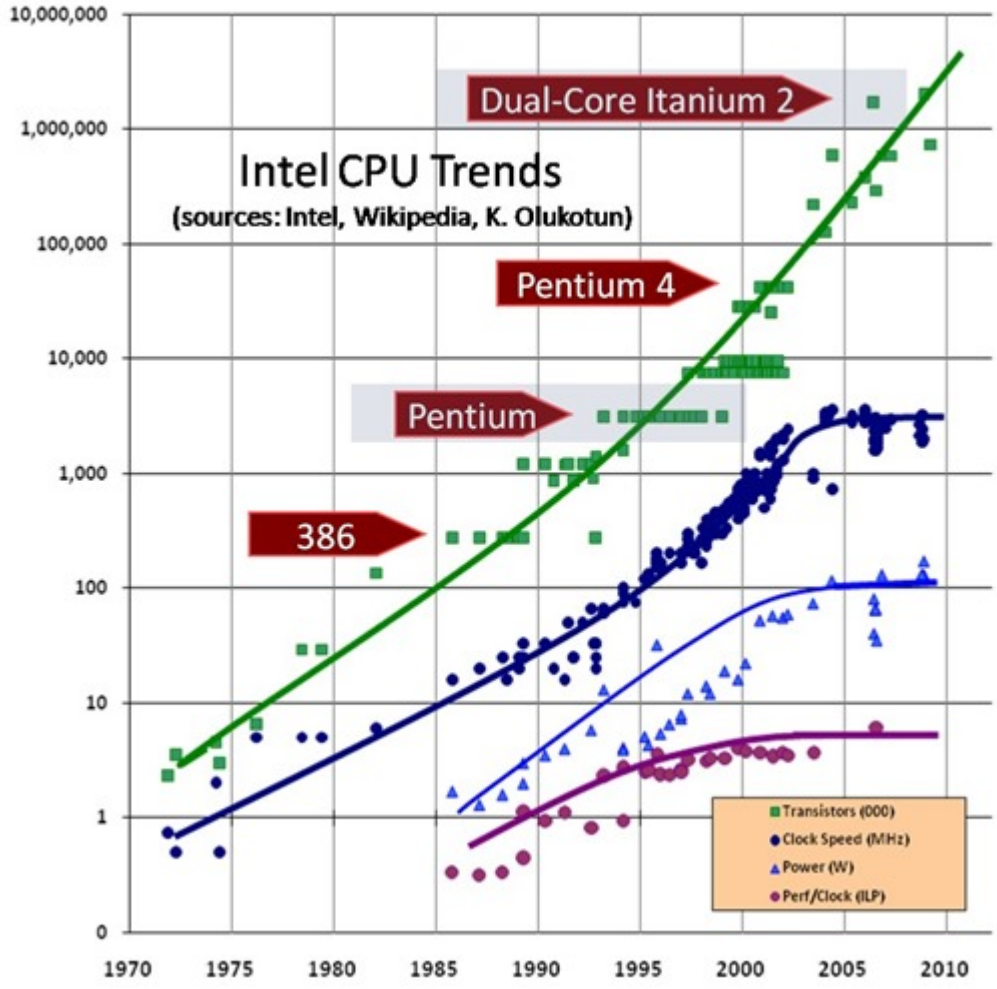
Je vais prendre ici trois exemples pour examiner ce phénomène : les **microprocesseurs**, le **solaire photovoltaïque** et le **séquençage de l'ADN**. Dans les trois cas, on peut constater que cette loi du progrès exponentiel connait des "hoquets" ou des dérives. Ces variations peuvent aussi bien ralentir les progrès que les accélérer au-delà de l'exponentielle. Elles donnent notamment lieu à des non linéarités liées aux variétés des techniques employées pour faire progresser l'état de l'art.

Dans le cours de cet argumentaire, je vais utiliser des courbes exponentielles récupérées à droite et à gauche. La plupart ne sont pas régulièrement actualisées, ce qui peut cacher parfois d'autres phénomènes de divergence dans les courbes. Mais elles font bien passer le message !

Je terminerai avec quelques autres exemples, moins scientifiques, tels que ceux qui concernent les prévisions de chiffre d'affaire des startups dans leurs business plans et la confusion qui existe entre deux courbes qui démarrent de la même manières : les exponentielles et les gaussiennes.

Semiconducteurs

Les progrès dans ce secteur sont continus depuis des décennies. Ils vont se poursuivre pendant au moins quelques décennies à venir, mais de manière probablement plus chaotique que jusqu'à présent. Les progrès dans les semiconducteurs et les microprocesseurs, qui sont les plus complexes d'entre eux, sont variables. Cela dépend de la grandeur que l'on mesure. La loi de Moore continue d'être vérifiée au niveau de la densité de transistors et du nombre de transistors par puce. Elle a cependant tendance à plafonner dans d'autres dimensions comme la vitesse d'horloge des processeurs.



Plusieurs facteurs conditionnent tous ces progrès :

Le **niveau d'intégration,** qui change d'échelle environ tous les trois ans. Les microprocesseurs les plus intégrés en ce début 2015 sont fabriqués en technologie 14 nm, chez Intel avec sa série Core M et ses récents Atom x3, x5 et x7. Dans la mobilité, le Samsung Galaxy S6 annoncé au MWC de Barcelone **serait basé** sur un chipset Exynos 7420 construit en technologie FinFET 14 nm. Le processeur A8 de l'iPhone 6 lancé en septembre 2014 est produit chez Samsung en technologie 20 nm. Hors Intel, les générations de processeurs mobiles de 2013/2014 étaient en technologie 28 nm. La majorité des processeurs mobiles seront en 14 ou 16 nm d'ici la fin 2015 ou début 2016. On les retrouvera dans les smartphones et tablettes "mainstream" en 2016 et pas simplement dans le haut de gamme. A la clé : plus d'autonomie et de puissance.

Ce processus d'intégration prend du temps. Il demande notamment des systèmes de gravure à ultraviolet en phase liquide capables de limiter les effets de la diffraction de la lumière dans le dessin des transistors sur le silicium. Les systèmes dits EUV (Extreme Ultra Violet) doivent prendre le relais pour permettre d'affiner encore plus la gravure. Ils sont complexes à mettre au point et le leader du marché, le hollandais ASML, a plusieurs années de retard dans sa roadmap. Le processus d'amélioration du niveau d'intégration de la fabrication des microprocesseurs est très lent et itératif, comme décrit dans ma **série d'articles** sur la visite de l'unité de fabrication Crolles 300 de STMicroelectronics. Comme l'EUV n'est pas encore au point, les fondeurs utilisent le procédé du multi-patterning consistant à graver en plusieurs fois des sillons non adjacents dans le silicium. Limitant les risques de recouvrement entre sillons, cela reste cependant un pis-aller.

Depuis le 28 nm, Intel, TSMC et Samsung utilisent des transistors montés "verticalement" (les "FinFET") et avec des portes en matériaux alternatifs comme le Silicium-Germanium. A ce niveau d'intégration, la

mécanique quantique, les phénomènes de fuite de courant et les effets thermiques peuvent jouer des tours à la loi de Moore et bloquer la course à l'intégration. Quid des atomes ? On n'a pas encore atteint leur taille puisque 1 nm = 4 atomes de silicium cristallin. Mais on s'en rapproche ! Les roadmaps des fondeurs prévoient de descendre jusqu'à 5 nm.

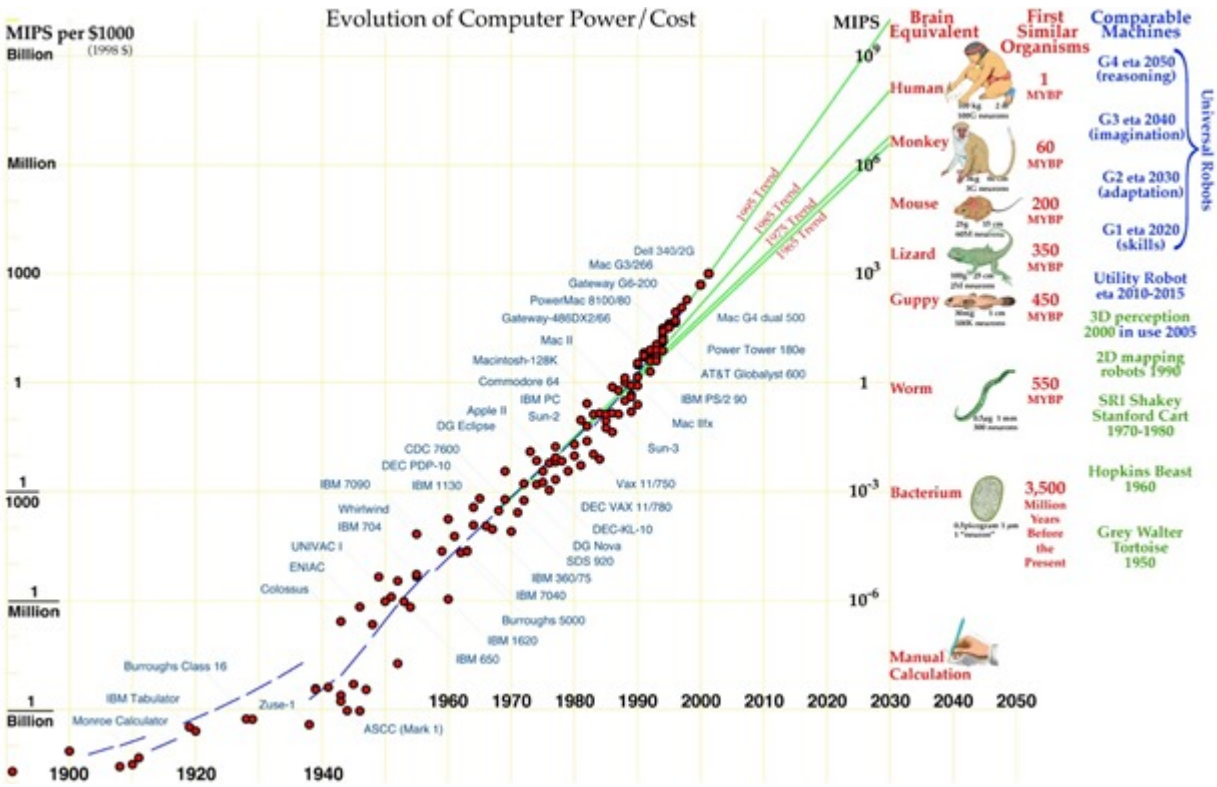
La **vitesse de commutation des transistors** qui dépend notamment de la nature du diélectrique qui forme la grille d'un transistor. Elle commande le passage du courant entre la source et le drain. C'était historiquement de l'oxyde de silicium. On utilise aujourd'hui de l'oxyde de hafnium ou de zirconium. Dans un futur indéterminé, on pourra passer au graphène, des cristaux savamment arrangés d'atomes de carbone pur. Il présentent la particularité d'avoir des électrons très mobiles. Ceux-ci permettront d'augmenter la fréquence de commutation des transistors et de passer de 3 GHz à plus de 50 GHz, voire même d'atteindre 400 GHz. Le jour où l'on saura faire cela, on gagnera quelques années dans l'application de la loi de Moore du côté non pas de la densité mais de la puissance de calcul. Et on redémarrera un autre cycle, celui de la miniaturisation. L'autre axe de progrès réside dans l'optronique qui est aussi capable de commuter des circuits à 50 GHz et à faire transiter de très gros volumes d'information. L'optronique est pour l'instant en retard d'environ deux décennies au niveau de l'intégration par rapport aux transistors CMOS. Nous sommes ici aux tous débuts d'une **autre exponentielle** !

La **consommation électrique** qui peut baisser avec les deux techniques précédentes, éventuellement accompagnées du procédé du SOI (Silicon on Insulator) lancé par le français SOITEC et utilisé par STMicroelectronics et sous licence par Samsung. C'est un sujet critique où la loi de Moore n'a pas l'air de trop s'appliquer. On le voit à la difficulté qu'a Intel à baisser la consommation de ses processeurs. Il est difficile aujourd'hui de trouver un laptop avec plus de 10 heures d'autonomie, même avec un Core M ou un Atom x3/5/7 gravés en 14 nm. Baisse de consommation des processeurs aidant, le problème se déplace progressivement vers d'autres composants avec, en premier, les écrans LCD dont le rétroéclairage, même en LED, consomme encore pas mal d'énergie.

Le **multicœur** reste le moyen le plus utilisé d'augmenter la puissance des microprocesseurs. Ce n'est plus une question de gravure mais de logique. On répartit les traitements sur plusieurs cœurs ou bien génériques (dans le CPU) ou spécifiques (GPU pour le graphisme, codecs vidéo et audio, fonctions réseau ou sécurité). Le facteur limitant est lié aux logiciels qu'il est difficile de répartir automatiquement, surtout sur les cœurs du CPU. C'est ce qui explique la relative futilité de ces chipsets à CPU à huit cœurs que l'on trouve maintenant dans les smartphones alors que l'on n'est pas en mesure de répartir une application mobile courante sur plus de deux cœurs avec les outils de développement actuels. Et comme on n'utilise généralement qu'une application à la fois sur mobile, un bon nombre de cœurs restent inutilisés.

Reste à savoir si tous ces progrès technologiques sont perçus par les utilisateurs. Est-ce qu'un micro-ordinateur d'aujourd'hui est significativement plus rapide qu'un micro-ordinateur d'il y a 10 ou 20 ans ? On manque de repères pour s'en rendre compte. On peut juste noter que l'on gère plus de photos et de vidéos qu'avant avec et que tout est plus facile d'accès via les réseaux haut-débit. Même remarque sur les smartphones : le processeur d'un iPhone 6 est plus de cinquantaine fois plus rapide que celui des premiers iPhone. Le perçoit-on pour autant ? Pas forcément car si le matériel évolue bien, le logiciel a du mal à suivre. Ou tout du moins, il prend ses aises avec l'abondance de ressources matérielles. Les langages compilés ont été progressivement remplacés par des langages interprétés notamment dans l'Internet (PHP, JavaScript). Les micro-ordinateurs, les mobiles et les tuyaux se remplissent comme une baignoire à raz-bord quelle que soit la capacité disponible.

Une exponentielle est souvent présentée sous forme de droite. Pour ce faire, on utiliser une échelle logarithmique dans les ordonnées (Y), celle des abscisses restant linéaire pour les années (X). C'est le cas de la courbe ci-dessous qui fait un parallèle entre la puissance des processeurs au travers des âges et celle du cerveau humain. Ce genre de courbe n'a pas beaucoup de sens car on compare des puissances brutes et pas des architectures.



Un cerveau ne fonctionne pas du tout comme un ordinateur actuel. Comparer les deux revient à mettre en regard la capacité de vol d'un porte-avions et des avions qu'il transporte. Même si on voit bien des porte-avions voler dans The Avengers ! Un cerveau est un ordinateur ultra-lent tournant à environ 100 Hz mais massivement parallèle et hyperconnecté, le tout avec 100 milliards de neurones et 1000 fois plus de synapses les reliant entre eux. Le tout est associé à une floraison de capteurs à faire pâlir n'importe quel objet connecté. Ils nous fournissent non pas cinq sens, **mais 18**. Si on ne prend que l'œil humain, il est connecté au cerveau par le nerf optique avec un million de connexions parallèles. Un capteur CMOS de smartphone est connecté au CPU via trois fils, avec une sérialisation totale de l'image (comme dans le bus **SCCB** chez Omnivision). Le cerveau parallélise donc massivement les liaisons avec les capteurs. Les capteurs du futur auront peut-être des architectures de bus complètement différentes des architectures sérialisées actuelles !

Bref, comparer la puissance d'un ordinateur avec celle du cerveau n'a pas beaucoup de sens car on ne compare pas des architectures équivalentes. La question est de savoir comment l'architecture du cerveau pourra être plus ou moins imitée dans des ordinateurs.

Aujourd'hui, on sait reproduire des fonctions isolées du cerveau, une par une et de manière limitée. On bute sur l'intégration et sur la gestion des ambiguïtés, un blocage classique dans la reconnaissance de la parole ou l'interprétation des images. Même si la dernière fonction de détection de sport joué dans les vidéos lancée par Facebook va en surprendre quelques-uns. Ce n'est pas de l'intelligence. Ce sont des morceaux d'algorithmes en pièces détachées que les logiciels ne savent pas encore bien intégrer.

Dans le même temps, les ordinateurs savent faire un tas de choses qui sont totalement inaccessibles au cerveau humain. Comme le calcul mental qui a ses limites chez nous et n'en a quasiment aucune dans l'ordinateur. Le calcul est-il une forme d'intelligence ? Qu'est-ce que l'intelligence d'ailleurs ? Question aussi bien technique que philosophique ! Si l'on ne rentre pas dans des considérations religieuses sur l'origine de l'âme et de la conscience, et en adoptant une vision "mécanique" et "chimique" de l'intelligence, on peut cependant anticiper que l'homme sera un jour capable de la reproduire avec des machines.

Les progrès "Kurzweiliens" qui verront les ordinateurs se rapprocher du cerveau humain devront cependant

faire appel à des architectures nouvelles et à des combinaisons d'architectures complémentaires. Il y aura notamment les **réseaux synaptiques** pour imiter les connexions multiples des synapses dans le cerveau et les **ordinateurs quantiques** pour dépasser de très loin le cerveau humain dans la recherche de patterns par la force brute (décryptage, statistiques, moteurs de recherche, reconstitution de puzzles complexes comme dans le séquençage de l'ADN). On sortira alors de l'exponentielle dans laquelle on est un peu coincés, celle des microprocesseurs à transistors CMOS.

En résumé sur les semiconducteurs :

- Les ralentisseurs de l'exponentielle sont les effets de la mécanique quantique, les effets thermiques et de fuite en dessous de 5 nm pour les technologies CMOS actuelles.
- Les accélérateurs de l'exponentielle sont les grilles de transistor en graphène pour augmenter radicalement la "clock" des processeurs, les calculateurs utilisant l'optronique pour faire de même, les architectures massivement parallèles, les ordinateurs synaptiques, les réseaux neuronaux et les ordinateurs quantiques.

Dans la **seconde partie de cet article**, nous irons voir ce qu'il en est du côté du solaire photovoltaïque. Son coût a baissé très rapidement mais pas forcément du fait de progrès technologiques.

Cet article a été publié le 6 avril 2015 et édité en PDF le 17 mars 2024.