

# T1 Architecture des processeurs

## Introduction

Nicolas Gac

Maître de conférences à l'université Paris Saclay  
L2S (Univ. Paris Saclay/CNRS/CentraleSupélec)

*\* slides de cours rédigés par Alain Mérigot (université-paris-saclay)*

université  
PARIS-SACLAY

*M2 SETI 2022-23*

## 3 ECTS

CM (15h) / TD (4h) / TP (9h)

## Intervenants

- Omar Hammami (ENSTA) *omar.hammami@ensta-paris.fr*
- Nicolas Gac (UPSay - L2S) *nicolas.gac@universite-paris-saclay.fr*

## Partie 1 - Cours (3h)

amphi S.A4/S.104

- C1 (N.Gac) **Introduction** Amphi S.A4 - jeudi 29/09 8h45/12h
- C2 (N.Gac) **Jeux d'instructions** S.104 - jeudi 13/10 8h45/12h
- C3 (N.Gac) **Pipeline** S.104 - vendredi 14/10 8h45/12h

## Partie 2 - Cours (1h30) puis TD (2h)

S.015 + S.201 (G1) et S.211 (G2)

- C4 (O.Hammami) **Processeurs superscalaires et VLIW** S.015 - lundi 17/10 8h45/10h15  
TD1 (G1 O.Hammami) 10h30/12h30  
TD1 (G2 N. Gac) jeudi 20/10 8h45/10h45
- C5 (O.Hammami) **Hiérarchie mémoire** S.015 - lundi 24/10 8h45/10h15  
TD2 (G1 O.Hammami/G2 N. Gac) 10h30/12h30
- C6 (O.Hammami) **Optimisations logicielles** S.015 - lundi 31/10 8h45/10h45  
Séminaire S.015 - 11h/12h

## Partie 3 - TP (3h)

S.201 (G1) et S.210 (G2)

- G1 TP1 (O.Hammami) **Analyse de performances** lundi 31/10 13h45/17h
- TP2 (O.Hammami) **Simulateur de processeur MIPS** lundi 07/11 8h45/12h
- TP3 (N.Gac) **Simulation de cache de données et TLB** mardi 15/11 13h45/17h
- G2 TP1 (N.Gac) **Analyse de performances** jeudi 10/11 8h45/12h
- TP2 (N.Gac) **Simulateur de processeur MIPS** mardi 15/11 8h45/12h
- TP3 (N.Gac) **Simulation de cache de données et TLB** jeudi 17/11 8h45/12h

## Compte rendu TD/TP

TD à rendre pour le 7 novembre	10 %
TP à rendre pour le 5 décembre	20 %
Dépôt en ligne sur e-campus	

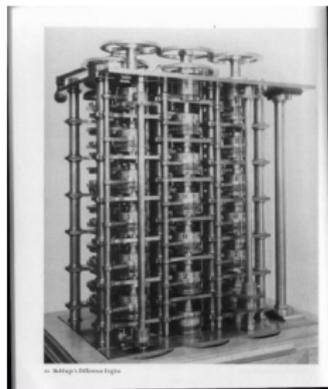
## Examen sur table (avec document)

Jeudi 1er décembre 9h/11h	70 %
---------------------------	------

# Brève histoire des ordinateurs

## Les pionniers

- Blaise PASCAL (1623–1662)  
1642 *Pascaline*  
Première calculatrice mécanique
- Gottfried LEIBNITZ (1646–1716)  
1670 Ajout des opérations de multiplication  
et division
- Charles BABBAGE (1791–1871)  
et Ada LOVELACE (1815–1852)  
1850 La machine analytique :  
Première machine programmable



## Premiers calculateurs électroniques 1940-1950

- 1943 : la *bombe* Alan TURING (1912-1954) & Gordon WELCHMAN (1906-1985) (cryptographie) [électromécanique]
- 1944 : Colossus Max NEWMAN (1897-1984) & Tommy FLOWERS (1905-1998) (cryptographie) [tubes]
- 1944 : Mark I (Harvard) Howard AIKEN (1900-1973) [électromécanique] 1 addition/s
- 1946 ENIAC John MAUCHLY (1907-1980) et John-Presper ECKERT (1919-1995) (Un. Penn.)  
Entièrement électronique [tubes]  
1 multiplication 2.2 s
- 1947 EDVAC MAUCHLY, ECKERT et John VON NEUMANN (1903-1957) (U. Penn.)  
Premier ordinateur moderne  
Calcul en binaire  
Programme enregistré en mémoire (*machine de von Neumann*)
- 1949 Manchester Mark-I Max NEWMAN (Un. Manchester)  
(premier ordinateur commercialisé Ferranti Mark-I)

### Caractéristiques de l'EDVAC

1000 mots de 44 bits (5ko)

6000 tubes

56kW,

1.16 kHz

8 tonnes, 45m<sup>2</sup>

MTBF<sup>1</sup>8h



<sup>1</sup>Mean Time Between Failure

1947 Invention du transistor (John BARDEEN (1908–1991) et William SHOCKLEY (1910–1989)) (Bell Labs)

Apparition des premiers ordinateurs commerciaux (IBM, Digital, CDC, Honeywell, etc)

Premiers systèmes d'exploitation  
Multiprogrammation

Premiers compilateurs (Fortran, John BACCHUS 1959)

Premiers multiprocesseurs



Mini ordinateur PDP-8 (Digital Equipment)

# Brève histoire des ordinateurs

(cont.)

## Apparition des circuits intégrés 1970–1980

1970 4004

premier microprocesseur :  
processeur 4 bits,  
2300 transistors

1972 8008

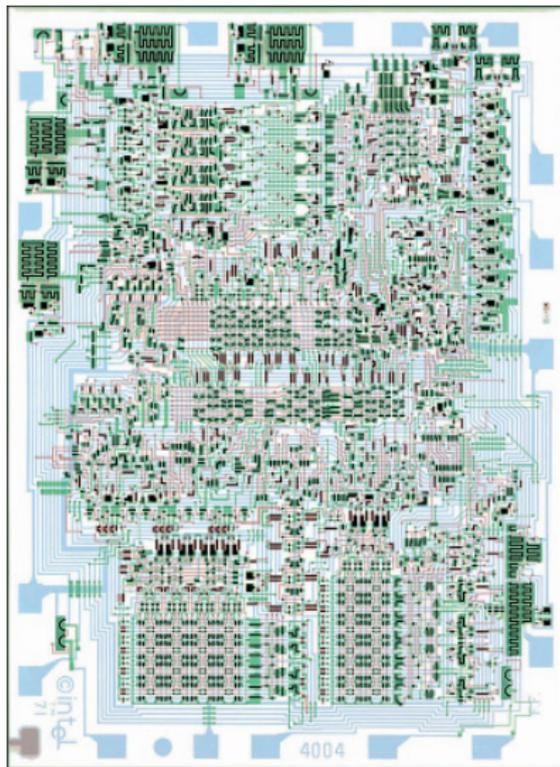
premier processeur 8 bits  
3500 trans. 300 kHz  
premier microordinateur Micral (R2E)  
2ko RAM 1973

1974 8080

processeur 8 bits 4500 transistors,  
2MHz

1978 8086

premier processeur 16 bits  
29ktrans. Utilisé dans les premiers PC



Le processeur 4004 d'Intel

1980-1990 Processeurs pipeline, processeurs de calcul flottant, caches, ...

1990-2000 Parallélisme d'instruction (superscalaire et VLIW), exécution spéculative, parallélisme de données (SIMD), ...

2000-2010 Parallélisme de thread, multicoeurs

2010+ *system-on-chip* (SoC)

## Ordinateurs personnels

- Ordinateurs de bureau
- Ordinateurs portables

## Serveurs de calcul ou de stockage

- *Supercomputers*
- *Data centers (cloud computing)*

## Systèmes embarqués

*Embedded systems* (informatique enfouie)

## Les systèmes embarqués



	Ordinateur de bureau	Serveur	Système embarqué
Prix	100–1000\$/proc.	200–2000\$/proc.	0.1–200\$/proc.
Ventes mondiales (2013)	300 M	10 M	15 G
Critères	Rapport prix/performance. Graphique	Performances, débit, disponibilité, évolutivité	Prix, consommation, performance pour l'application

## Haut de gamme

Processeurs des PC et serveurs

## Spécialisé

Haut de gamme des générations précédentes.

Ex : MIPS, PowerPC

## Embarqué

Faible consommation, temps réel.

## Contraintes

- Prix
- Performance
- Encombrement
- Consommation
- Temps réel

# Evolution des performances des processeurs

Liés à de nombreux facteurs

- Amélioration de la technologie
  - augmentation du nombre de transistors/circuit
  - réduction du temps de traversée
  - diminution de la consommation d'un transistor
- Améliorations architecturales
- Améliorations logicielles (compilateurs)
- Demandes du marché et contraintes économiques

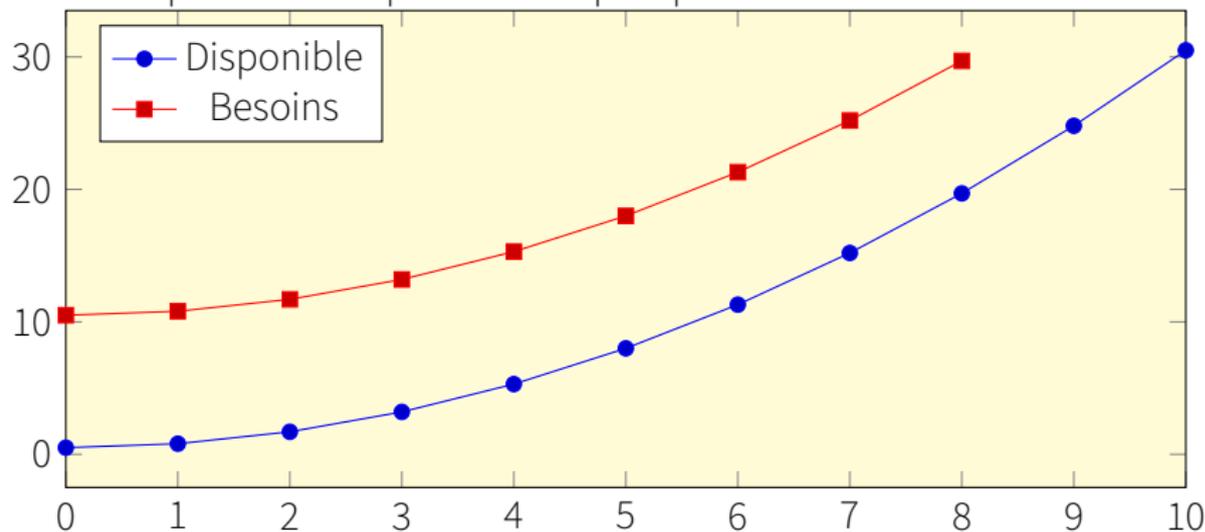
# Aspect économiques et applicatifs

## Des besoins croissants

Modèle économique piloté par la technologie (Intel)

De nombreuses applications nécessitent toujours plus de puissance de calcul (serveurs, ordinateurs personnels, embarqué haut de gamme).

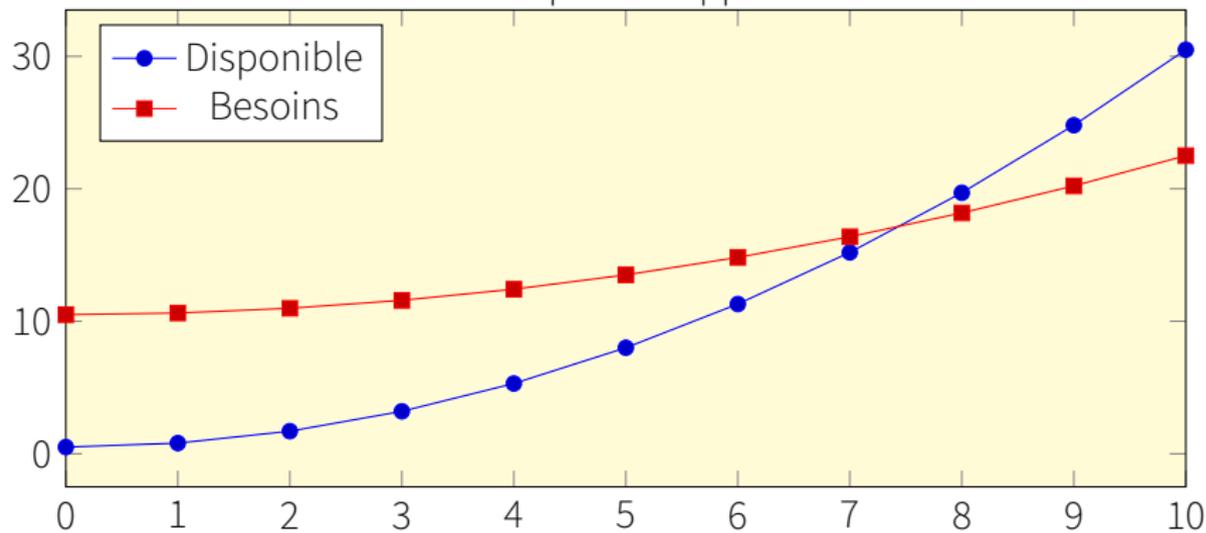
Il suffit de produire des processeurs plus puissants et le marché suivra.



Modèle économique piloté par les applications

Cas où les performances sont supérieures aux besoins (comme un certain nombre d'applications embarquées).

Il faut créer de nouveaux besoins par des applications.

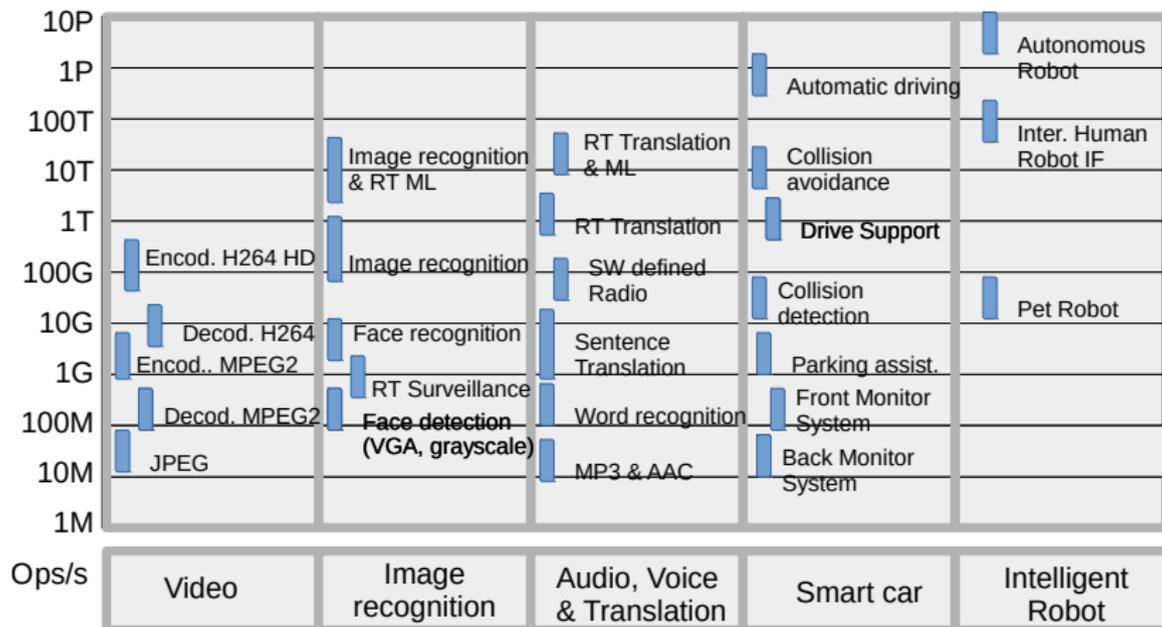


Recherche de la *killer application*

# Aspect économiques et applicatifs

(cont.)

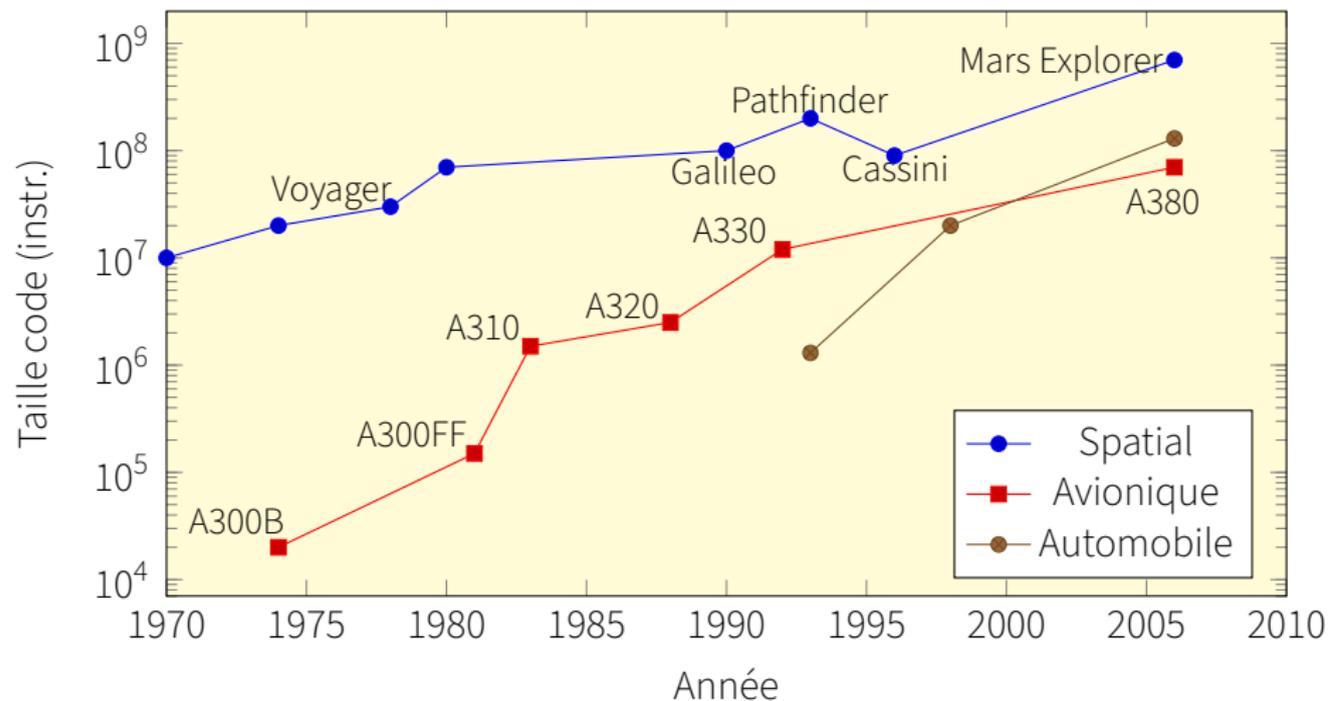
## Prévisions de charge sur les applications embarquées



# Aspect économiques et applicatifs

(cont.)

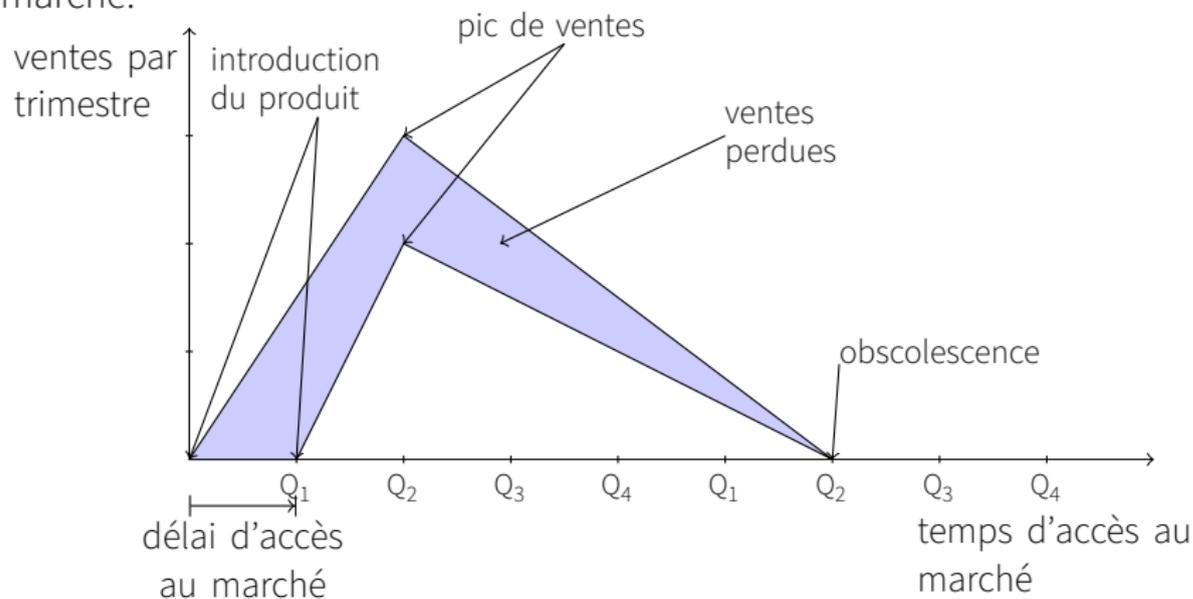
Des logiciels embarqués de plus en plus complexes



*firefox=10M, linux kernel+drivers=15M, windows 10 60M, google 2G*

## Temps d'accès au marché

Obsolescence très rapide du matériel (18 mois) : il faut être le premier sur le marché.



Un délai d'introduction d'un produit entraîne une perte de ventes.  
*time-to-market*

# Les grandes tendances de la technologie

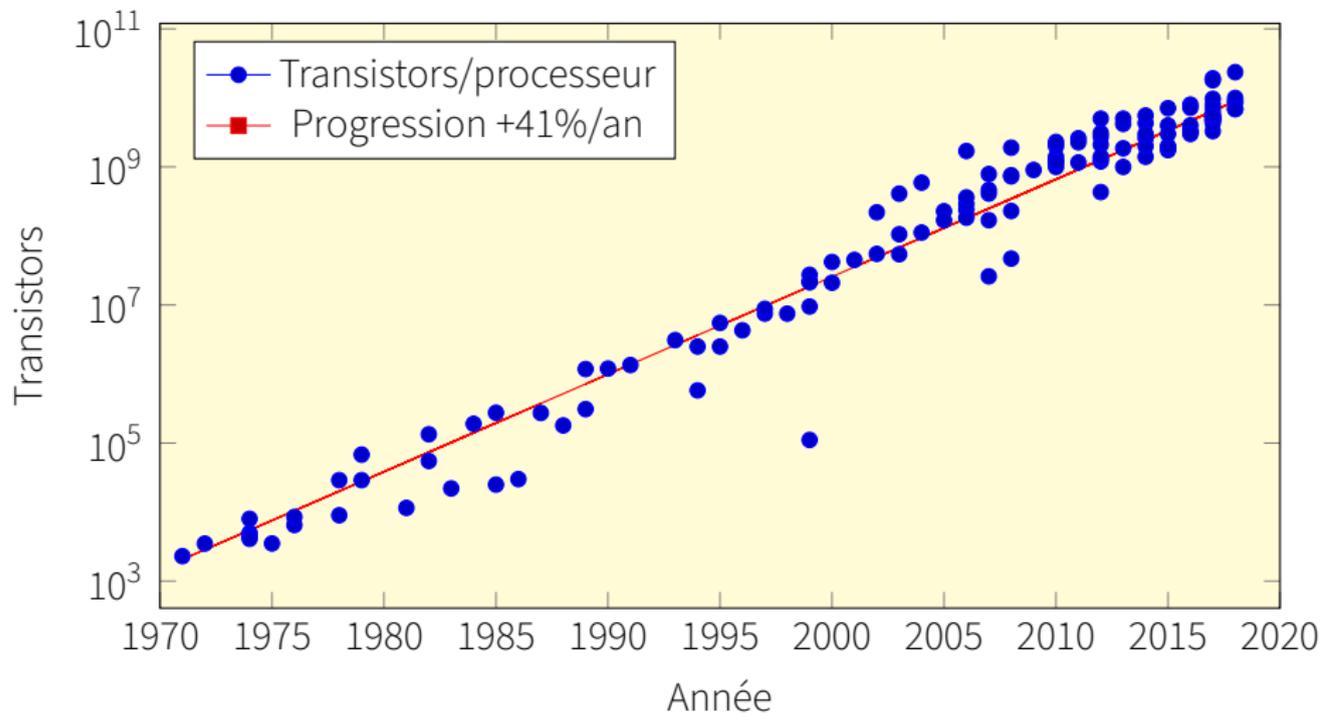
## Loi de Moore

- Dimensions technologiques réduites de 16%/an
  - Densité de transistors X 1.35/an
- Augmentation de la taille des circuits de 5 à 10%/an
- Le nombre de transistors/circuit croit de 40 à 45%/par an ( $\times 2$  tous les 22 à 24 mois)  
**Loi de Moore** (Gordon MOORE, né en 1929) [1965/1975]

Pour les mémoires l'augmentation est similaire, mais se réduit :

- $\times 4$  tous les 3 ans dans les années 90
- $\times 2$  tous les ans dans les années 2000
- $\times 2$  tous les 2 à 3 ans actuellement

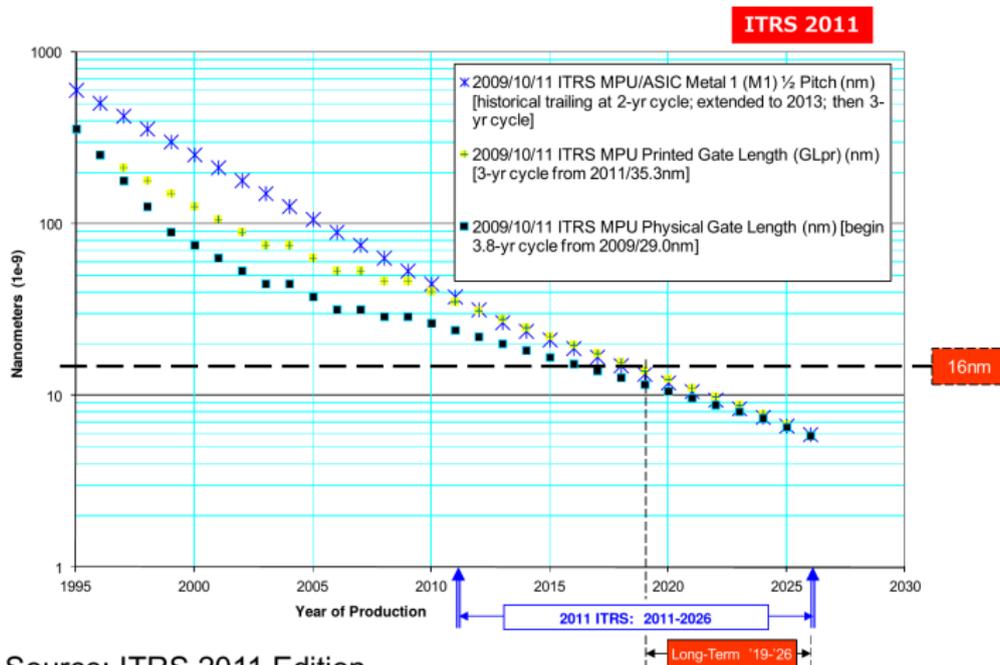
Nombre de transistors/processeur



# Les grandes tendances de la technologie

(cont.)

## Loi de Moore



Dimensionnement des transistors

MAIS....

Maille de silicium : 0,543nm

Actuellement (2021) gravure 10nm à 5nm : un canal fait de 20 atomes (10nm) à 10 atomes (5nm) !

Nombreux problèmes de gravure, fiabilité, dispersion lors du dopage ( $10^{-2}$  –  $10^{-5}$ ), effet tunnel, etc

Limites physiques de l'effet transistor en dessous de 2–3 nm.

**Fin de la loi de Moore dans quelques années ( $\approx 2025$ )<sup>2</sup>.**

(sauf rupture technologique)

---

<sup>2</sup>Intel a annoncé 18 Ångströms fin 2025.

Loi de Rock (Arthur Rock né en 1926) (ou seconde loi de Moore)

« *Le coût d'une usine de fabrication de circuits semiconducteurs<sup>3</sup> double tous les 4 ans.* »

Augmentation exponentielle des coûts de fabrication

Actuellement  $\approx$  15G\$/usine

---

<sup>3</sup> parfois appelée « fonderie de silicium »

L'évolution de performance des processeurs est la combinaison du progrès technologique et des améliorations architecturales.

$$t_e = n_i \times \overline{\text{CPI}} \times T_c$$

(formule de Hennessy-Patterson)<sup>4</sup>

$t_e$  temps d'exécution d'un programme

$n_i$  nombre d'instructions exécutées

$\overline{\text{CPI}}$  nombre moyen de cycles par instruction

$T_c$  temps de cycle

A la fin du XXe siècle amélioration globale des performances de près de 50%/an.

MAIS...

<sup>4</sup> John HENESSY, né en 1952 et David PATTERSON, né en 1947.



Jusqu'en 2000, augmentation systématique des performances de 50%/an

Il suffisait d'attendre les améliorations technologiques pour avoir un programme plus rapide.

***The free lunch is over*** (Herb SUTTER 2005)

## Performances des processeurs

L'amélioration technologique permettait l'amélioration des performances :

- en augmentant la fréquence des processeurs grâce à des transistors plus petits
- en améliorant les architectures des processeurs grâce à un nombre plus important de transistors :
  - pipeline,
  - exécution dans le désordre, parallélisme d'instruction (superscalaires, VLIW),
  - exécution spéculative,
  - élargissement des chemins de données : 8 bits, puis 16, 32, 64 (et même 128 à 512 bits en parallélisme de données),
  - parallélisme de données (SIMD)
  - *multithreading*
  - ajout de caches L1, puis L2, et L3 dans les processeurs
  - etc

### Evolution des CPU Intel

Proc.	ann.	Taille/fréq./CPI	innovations
4004	1971	2300tr. / $\approx 10\text{CPI}^5$	0.75MHz / Premier $\mu$ processeur 4 bits
8080	1974	4500tr. / 4–11CPI	2MHz / Premier $\mu$ processeur 8 bits
8086	1978	29000tr. 2–20CPI	5MHz / Premier $\mu$ processeur 16 bits (code partiellement compatible), mémoire segmentée
80286	1982	135ktr. 2–20CPI	6MHz / unité de gestion mémoire (MMU), bus adresses/données non multiplexés, multiplieur entier câblé
80386	1985	275 ktr. 3–10CPI	12MHz / Processeur 32 bits (bus, registres, ALU), cache L1 externe, mémoire paginée (jusqu'à 4GB)

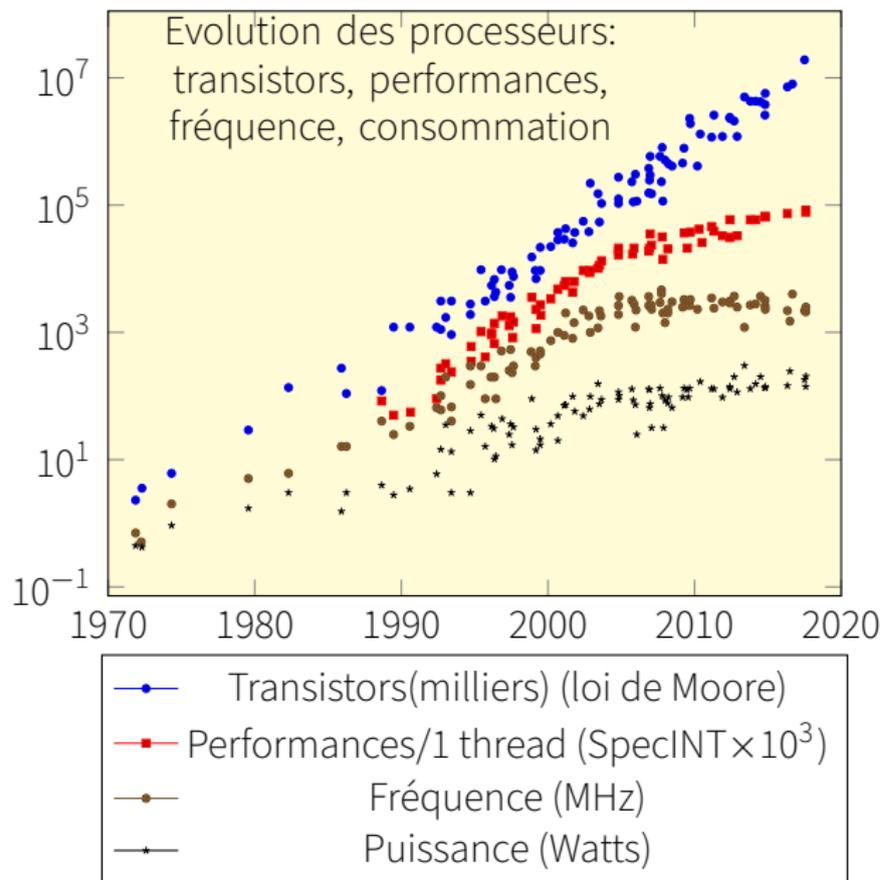
<sup>5</sup>CPI : cycles par instruction

# Les grandes tendances de la technologie

(cont.)

Proc.	ann.	Taille/fréq./ipc	innovations
80486	1989	1Mtr. /50MHz / 40MIPS <sup>6</sup>	Pipeline, cache L1 intégré (unifié), opérateurs flottants intégrés
Pentium (P5)	1993	3Mtr. / 60MHz / 100 MIPS	Superscalaire (2 instr.), bus 64 bits, cache L1 séparé I/D, prédiction branchement
Pentium MMX	1995	4.5Mtr. / 150MHz	Instructions “multimedia” MMX
Pentium Pro (P6)	1996	5.5Mtr. / 250MHz / 180 MIPS	Exécution dans le désordre, exécution spéculative, coeur RISC
Pentium III	1999	29 Mtr. / 700MHz /400 MIPS	Ext. SSE ( <i>Coppermine</i> 1999), cache L2 intégré
Pentium 4	2000	40–200 Mtr. / 1.5–3GHz	Pipeline profond, cache $\mu$ ops, SSE2 ( <i>willamette</i> 2000), SSE3, <i>hyperthreading</i> ( <i>Prescott</i> 2004)
Pentium core	2008	1 Gtr.+ / 1.5–3GHz	Multicoeur, L3 intégré, x86-64, AVX ( <i>Nehalem</i> 2008, <i>Sandy Bridge</i> 2011, <i>skylake</i> 2015, <i>coffee-lake</i> 2018, <i>icelake</i> 2019, <i>tigerlake</i> 2020)

<sup>6</sup>MIPS : *Million instructions per second*



Depuis  $\approx$  2005,  
stagnation des  
performances des  
processeurs  
individuels

- performance/cy-  
cle (un seul  
thread)
- fréquence  
d'horloge
- puissance  
consommée

Il s'avère que :

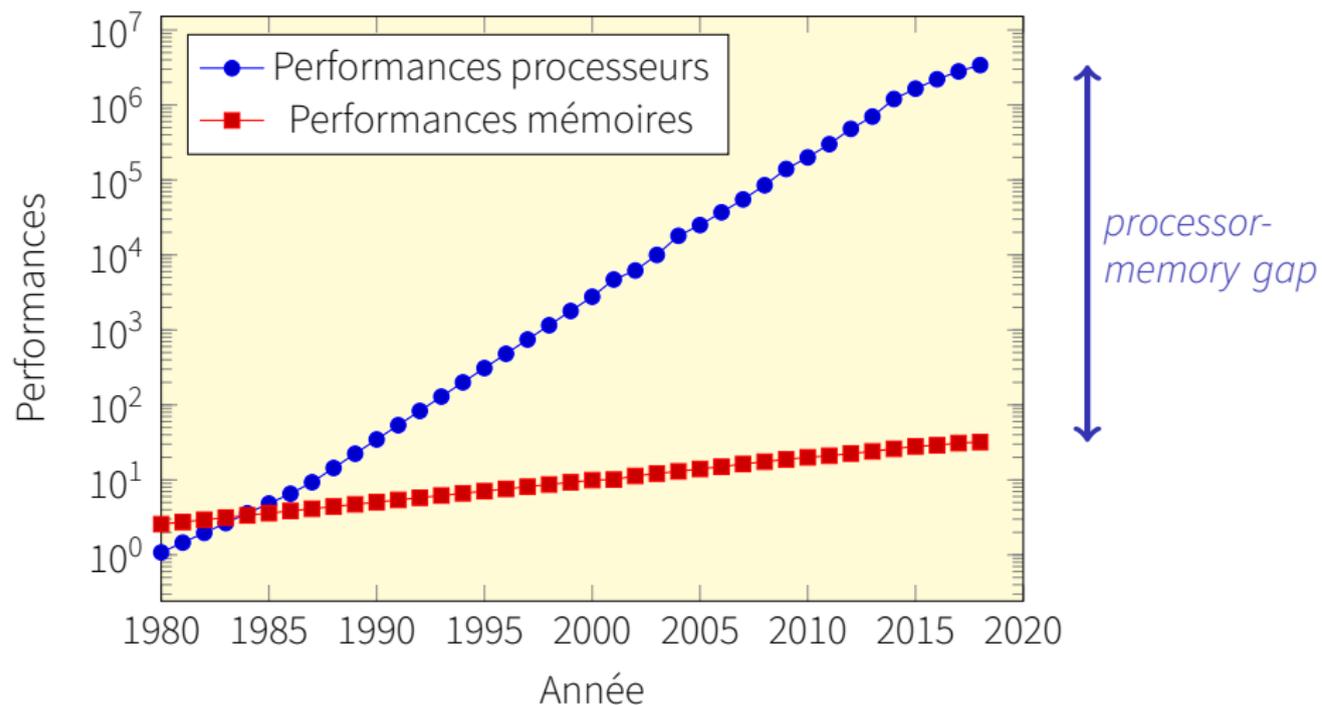
- l'augmentation de fréquence se heurte au problème de la consommation. Doubler la fréquence revient à quadrupler la consommation (à cause de l'augmentation de tension qui accroît le courant de fuite des transistors). (**power wall**)
- il n'y a plus beaucoup de progrès architecturaux exploitables sur un processeur (**ILP wall**)<sup>7</sup>
- la différence entre performances des mémoires et des processeurs ne fait que s'accroître (**memory wall**)

---

<sup>7</sup>ILP : *Instruction level parallelism*

# Les murs de l'architecture : *Memory wall*

Evolution technologique comparée des processeurs et des mémoires



Lors de l'amélioration de la technologie, on va chercher à **augmenter** la taille des mémoires.

Cette augmentation de taille compense en grande partie l'accélération obtenue par le progrès technologique.

L'accès mémoire nécessite de 50 à 200 cy sur un processeur actuel...

On peut faire :

- des mémoires de grande taille lentes
- des mémoires de petite taille rapides

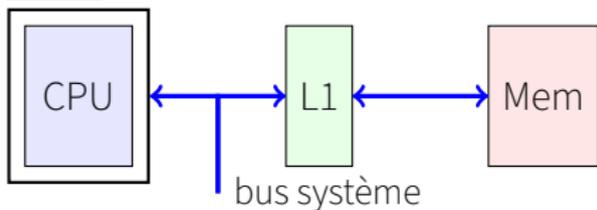
→ mémoire cache

Complexité croissante de la hiérarchie mémoire : caches L1, L2, L3, mémoire principale

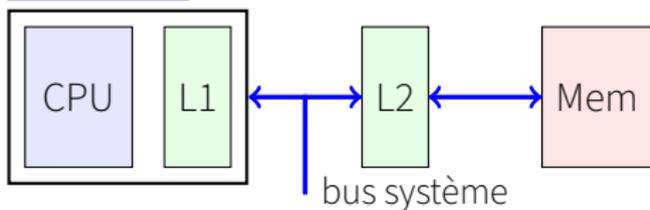
# Les murs de l'architecture : *Memory wall*

(cont.)

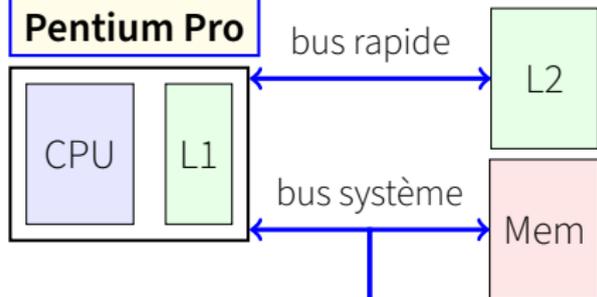
386



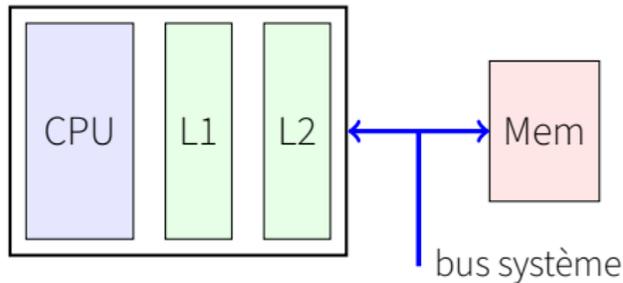
Pentium



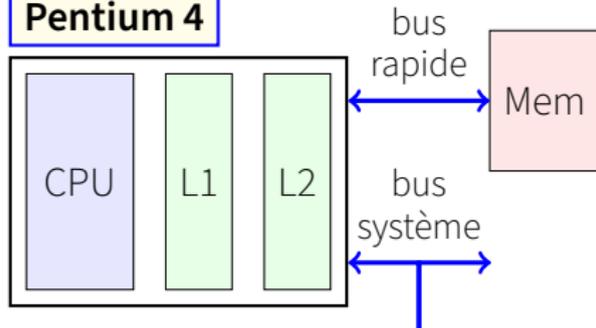
Pentium Pro



Pentium III

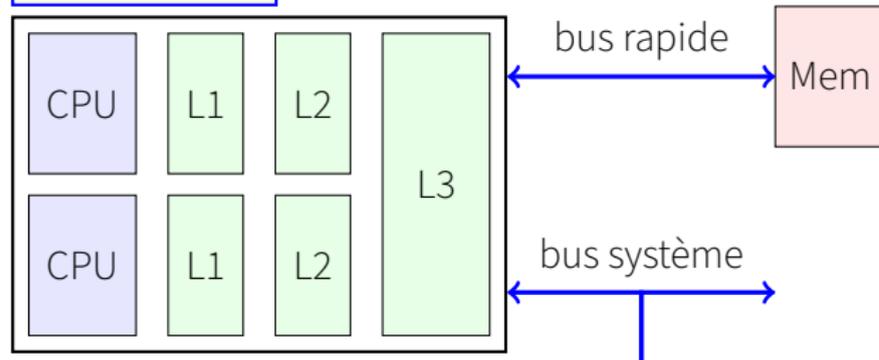


Pentium 4



La hiérarchie mémoire est un point très critique dans les processeurs actuels

## Pentium core



# Les murs de l'architecture : *Power wall*

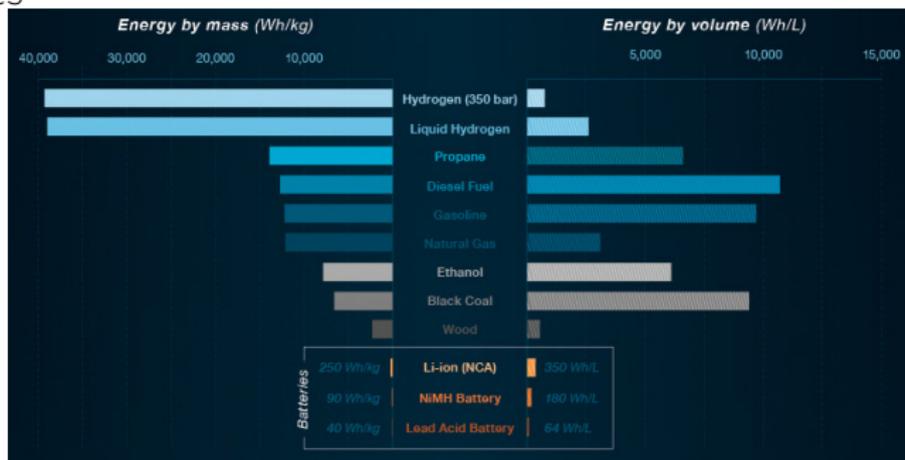
La puissance consommée est un problème majeur dans plusieurs applications :

- équipements nomades
- *data-centers* (4% de l'énergie électrique produite aux USA en 2013, 20% de croissance/an)
- objets connectés

Impact

- technologique
- microarchitecture/circuit logique
- architecture
- algorithmique
- système

## Les batteries



1950 : 40 Wh/kg (Plomb acide)

1978(concept)/1992(commercialisation) : 90 Wh/kg, (Lithium-Ion)

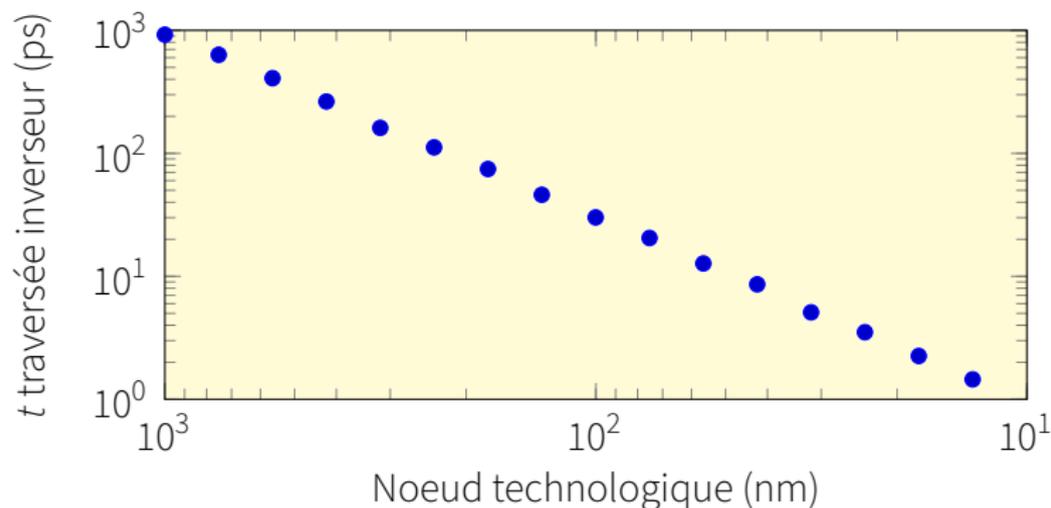
2000 : 140 à 160 Wh/kg . (Lithium-Polymère)

2002 : 190 à 250 Wh/kg (Lithium-Ion-NiCoAl)

2013 (lithium-air) : 1700 - 2400 Wh/kg (expérimental)

Essence : 14000 Wh/kg

L'évolution de la technologie entraîne une amélioration des temps de traversée des portes.



Réduction du temps de traversée des portes, mais l'augmentation de fréquence de fonctionnement est limitée par la consommation (et les problèmes de synchronisation à très haute fréquence).

## Puissance consommée dans un circuit synchrone

$$P \approx \frac{C \times V^2}{T}$$

$C$  capacité totale à commuter,  $V$  tension d'alimentation,  $T$  période d'horloge.

Amélioration de la technologie:  $\rightarrow$  Toutes les dimensions  $\times \alpha$  ( $\alpha < 1$ )  
(réduction des dimensions)

### Impact sur la capacité $C$

- $C = n_{\text{portes}} \times C_{\text{porte}}$
- $n_{\text{portes}} = \frac{\text{surf. circuit}}{\text{surf. porte}}$  (surf. circuit  $\approx$  cst, surf. porte  $\times \alpha^2$ )  
 $\rightarrow n_{\text{portes}} / \alpha^2$  (nombre de portes  $\nearrow$ )
- $C_{\text{porte}} \approx S/e$  (transistor ou interconnexions  $\equiv$  capacité plane)  
 $C_{\text{porte}} \rightarrow C_{\text{porte}} \times \alpha^2 / \alpha = C_{\text{porte}} \times \alpha$  (capacité par porte  $\searrow$ )
- $C \rightarrow C' = C / \alpha$  (capacité totale  $\nearrow$ )

augmentation de la capacité totale

## Impact sur la période d'horloge

Liée au plus grand temps de traversée de portes entre deux fronts d'horloge.  
(chemin critique)

Pour une architecture donnée,  $T$  est :

- $1/\alpha$  courant dans les transistors ( $\nearrow$ )
- $\alpha$  capacité d'une porte ou interconnexion ( $\searrow$ )
- $\alpha$  délais RC de propagation ( $\searrow$ )

$T \rightarrow T' = T \times \alpha$  (période horloge  $\nearrow$ )

D'où la puissance  $P' = C' \times V^2 / T' = P / \alpha^2$

Augmentation quadratique de la puissance avec  $\alpha$  !

Technologie améliorée d'un facteur 2  $\implies$  Puissance multipliée par 4 !!!!

(à  $V$  constant)

Actuellement, un CPU d'un ordinateur de bureau consomme 40 W. Un GPU peut atteindre 200 W.

Une solution consiste à réduire la tension d'alimentation (et donc les performances)

Règle de dimensionnement de Dennard (*Dennard scaling* 1974)

- Tension et courant doivent être proportionnels aux dimensions d'un transistor
- Ainsi dimensions  $\searrow \implies V$  et  $I \searrow$
- Garanti puissance/unité de surface constante

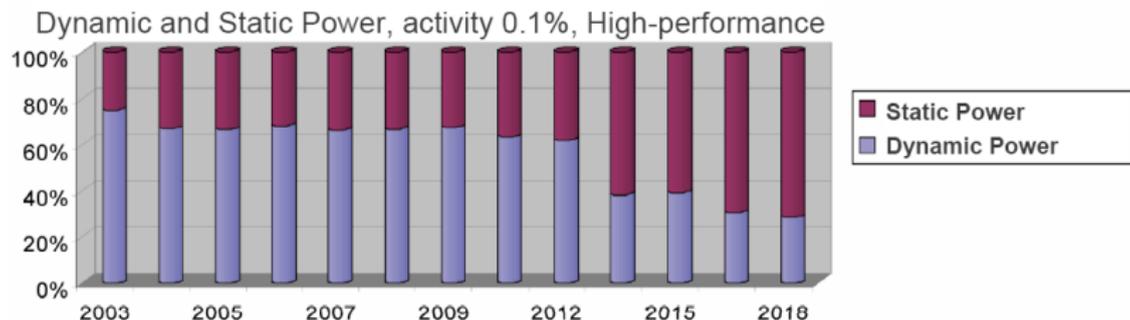
Partiellement appliquée pour limiter la puissance consommée  
Mais réduit les performances  $f \approx V^2$

De plus la loi de Dennard ignore :

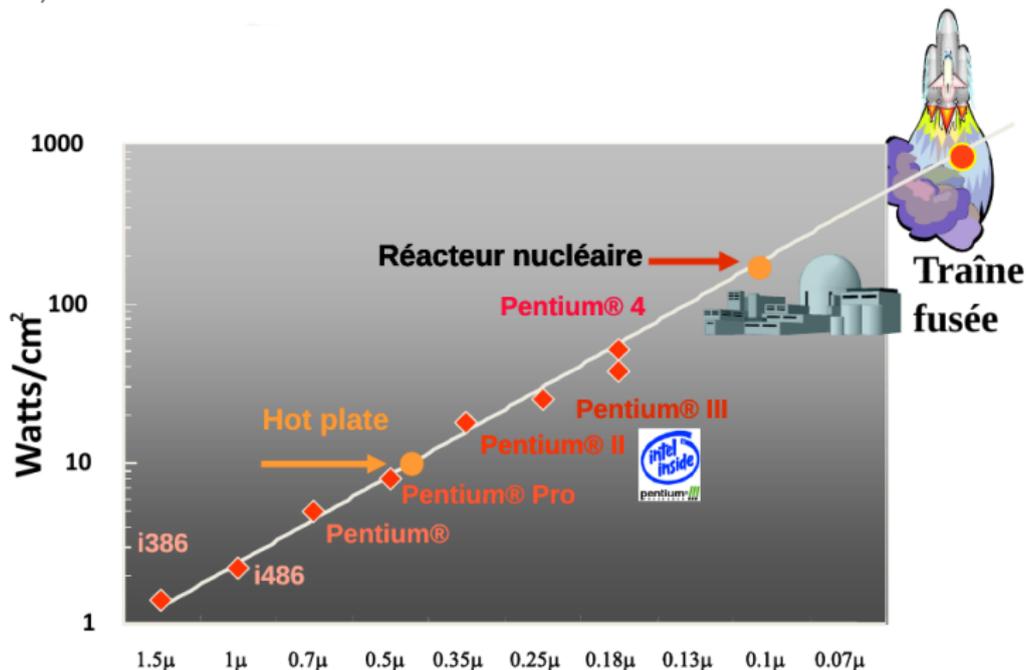
- l'existence d'une tension de seuil (atteinte)  
Alimentation actuelle  $\approx 0.3 - 0.7V$   
Nombreux problème pour réduction supplémentaire
- les courants de fuite

Les courants de fuite. *subthreshold leakage*

Avec la réduction des dimensions technologiques, il y a un courant de fuite même pour un transistor qui ne conduit pas.



Forte augmentation de la densité de puissance par unité de surface  
Problème majeur (évacuation de la chaleur, possibilité de destruction des transistors)



A terme la dissipation risque d'être le facteur limitant de l'intégration; « *dark silicon*<sup>8</sup> »

Fin de la loi de Dennard vers 2006

Fort ralentissement de la loi de Koomey (*la quantité de calcul par Joule double tous les deux ans*)

⇒ Forte augmentation de la puissance à dissiper.

Possibilité d'intégrer un grand nombre de processeurs par circuits, *mais* on ne pourra en activer qu'une partie pour limiter la dissipation thermique

Pourrait atteindre  $\approx 80\%$  en 3nm.

Très forte limite des performances.

Peut pousser à l'utilisation de processeurs très spécialisés...

---

<sup>8</sup>Dark Silicon and the End of Multicore Scaling, Hadi ESMAEILZADEH et al., ISCA 2011

Une des meilleures solutions pour réduire la consommation est le *parallélisme*.

Supposons que l'on ait certaines performances sur un processeur.

Utilisation de deux processeurs pour avoir les **mêmes** performances,  
→ réduction fréquence par deux  
(et donc réduction tension d'alimentation, consommation).

A performances données, la puissance **diminue** *quadratiquement* avec le parallélisme.

L'utilisation de processeurs parallèles permet également de lutter contre le mur du parallélisme d'instruction (**ILP wall**).

Présent dans la totalité des architectures actuelles.

# Grandes tendances architecturales

- Multicoeurs (parallélisme symétrique).  
Quasi généralisé dans les processeurs pour PC et les systèmes embarqués
- Parallélisme au sein des processeurs : parallélisme d'instruction, parallélisme SIMD (traitement parallèle de données)
- Utilisation de processeurs graphiques à parallélisme massif pour le calcul (GP-GPU)
- Systèmes hétérogènes. Réunion sur un même circuit de processeurs, GPU, systèmes parallèles spécialisés, DSP, etc.
- Intégration de processeurs et de matériel programmable (FPGA)

***Welcome to the parallel jungle!*** (Herb SUTTER 2012)

La difficulté essentielle dans le parallélisme est son exploitation efficace.

*Loi d'Amdahl* (Gene AMDAHL 1922–2015)

Soit un programme ayant un temps d'exécution  $t = t_s + t_a$

$t_s = t(1 - \tau)$  temps intrinsèquement séquentiel (non parallélisable)

$t_a = t\tau$  temps de la partie du programme accélérable (parallélisable)

$\tau$  fraction parallélisable

Si on accélère  $t_a$  par un facteur  $A > 1$ , l'accélération finale est:

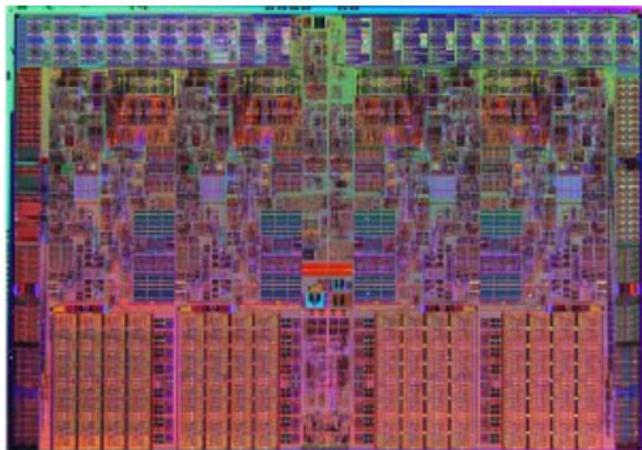
$$A' = \frac{t}{t'} = \frac{t_s + t_a}{t_s + \frac{t_a}{A}}$$

$$A' = A \times \frac{1}{\tau + A(1-\tau)}$$

Exemple : Si on accélère par  $A = 10$  une fraction  $\tau = 0.5$  d'un programme, le gain final est de  $A' = \frac{1}{0.5+0.05} = 1.82$

Parallélisme dans les processeurs pour PC.

- *hyperthreading*
- multicoeurs/*manycores*



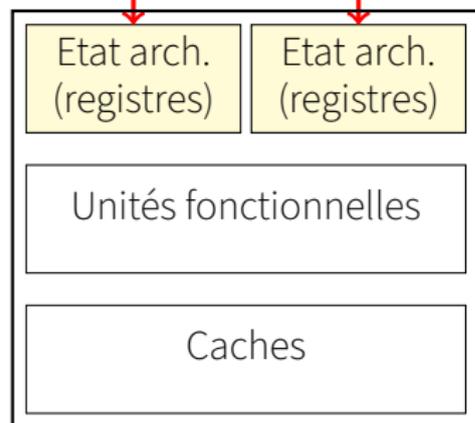
Processeur core i7 d'intel (4 coeurs).

## *multithreading*

*multithreads* et multiprocesseurs

Processeur *multithread*

proc. log. 1    proc. log. 2

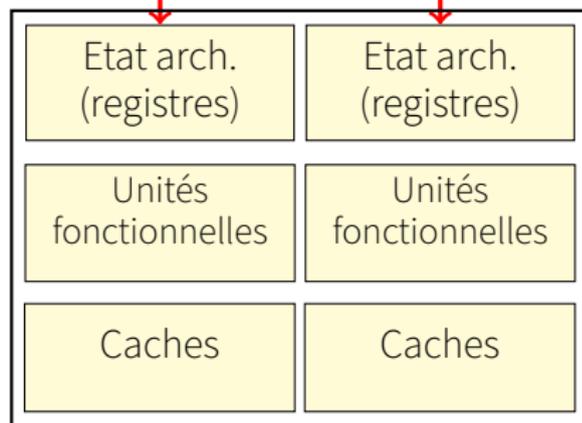


Mémoire principale

2 proc. logiques / proc. physique

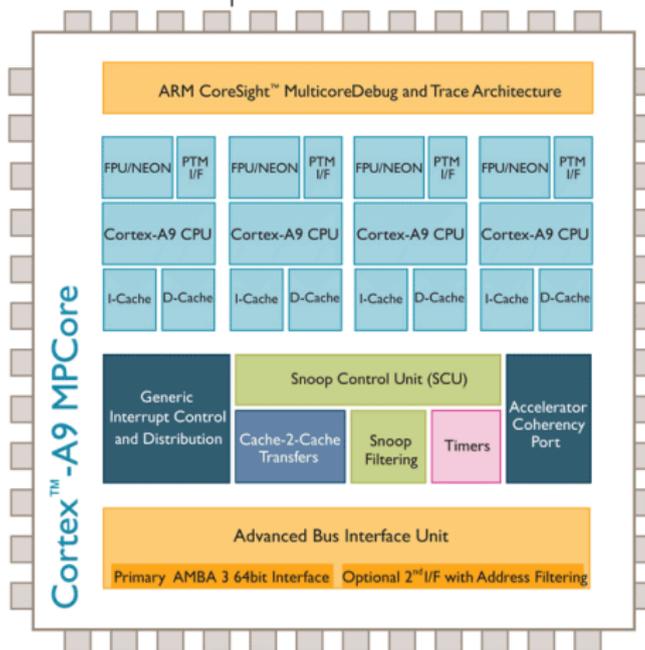
Processeur multicœurs

proc. phys. 1    proc. phys. 2



Mémoire principale

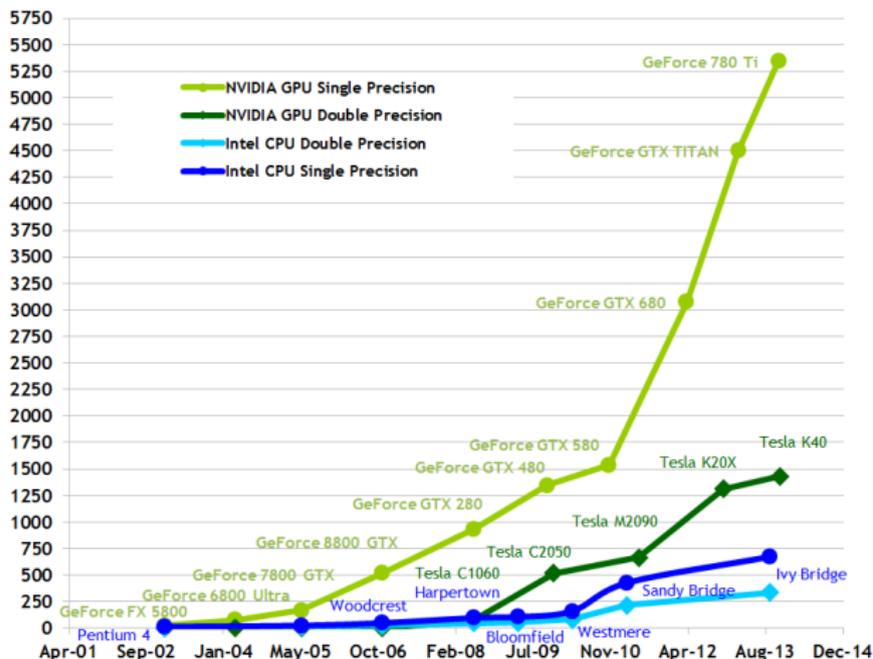
### Parallélisme en systèmes embarqués

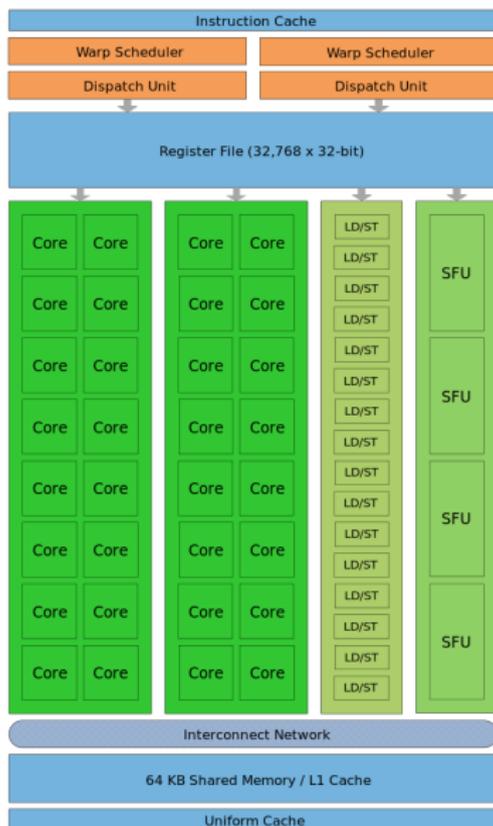


Mais le parallélisme peut être beaucoup plus important.  
MPPA Kalray 256 processeurs.

Les processeurs graphiques (GPU) donnent une puissance de calcul considérable pour des problèmes réguliers.

Theoretical GFLOP/s





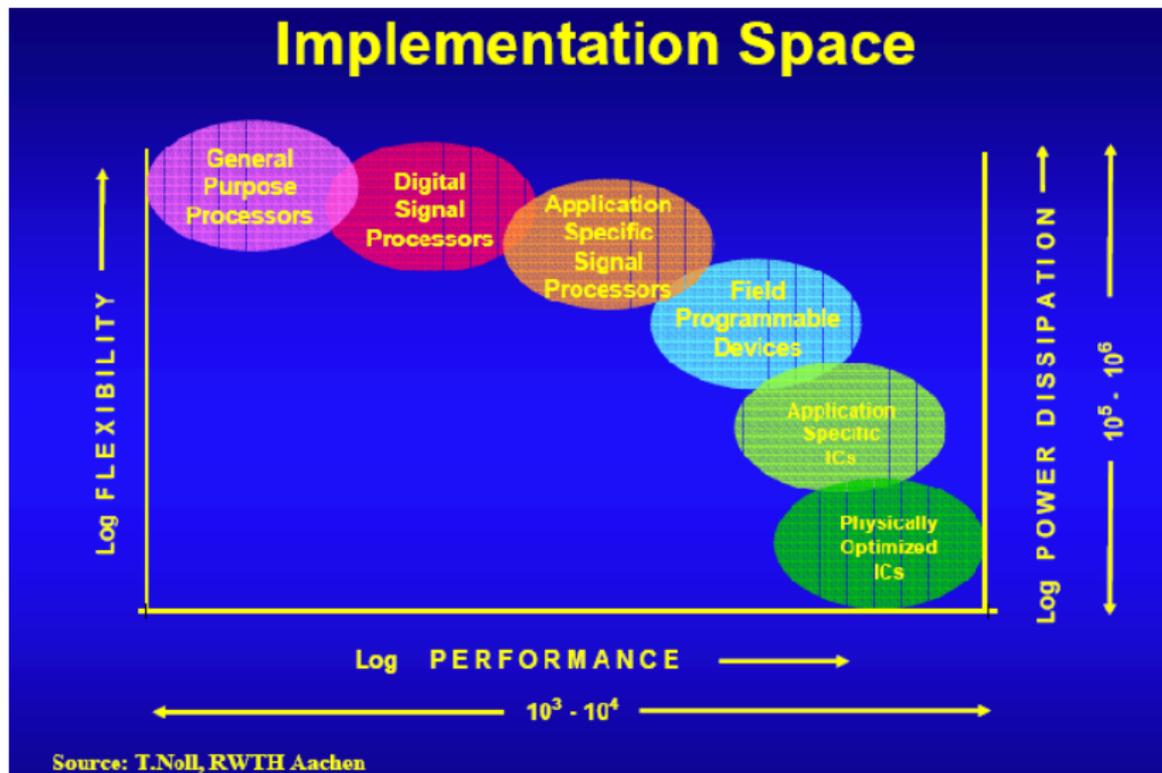
Microarchitecture de la série Fermi de NVidia

Existent en version embarquée avec une faible consommation ( $\approx 10W$ ) (Tegra).

Ces versions intègrent des processeurs *arm* et un GPU de taille réduite (256-512 coeurs)

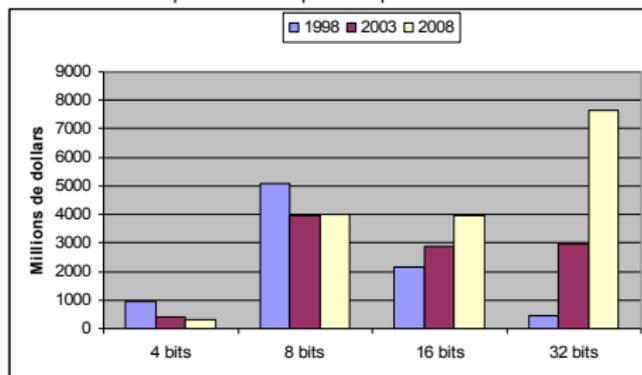
# Le spectre d'implémentation

ASIC	Processeur	Reconfigurable
Circuit intégré dédié à une application	Utilisation de processeurs généraux	FPGA
Hautes performances	Programmable	Bon compromis
Consommation réduite	Flexible	Permet de mélanger sur un circuit matériel et processeurs
Long et coûteux à développer	Non optimisé pour une application	

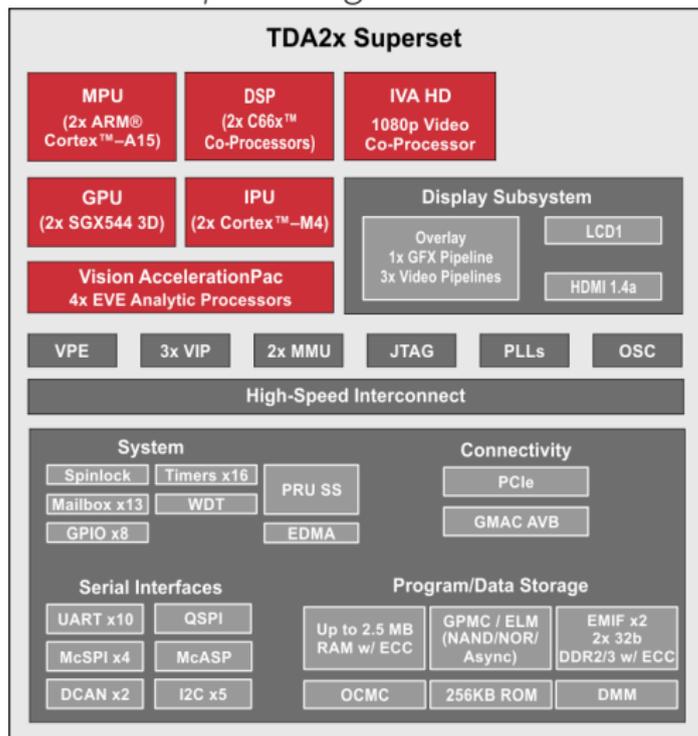


Microcontrôleurs

Intègrent des processeurs de plus en plus performants + RAM/ROM + E/S



*systems on chip* hétérogènes



INFO-001

TDA2X (TI)

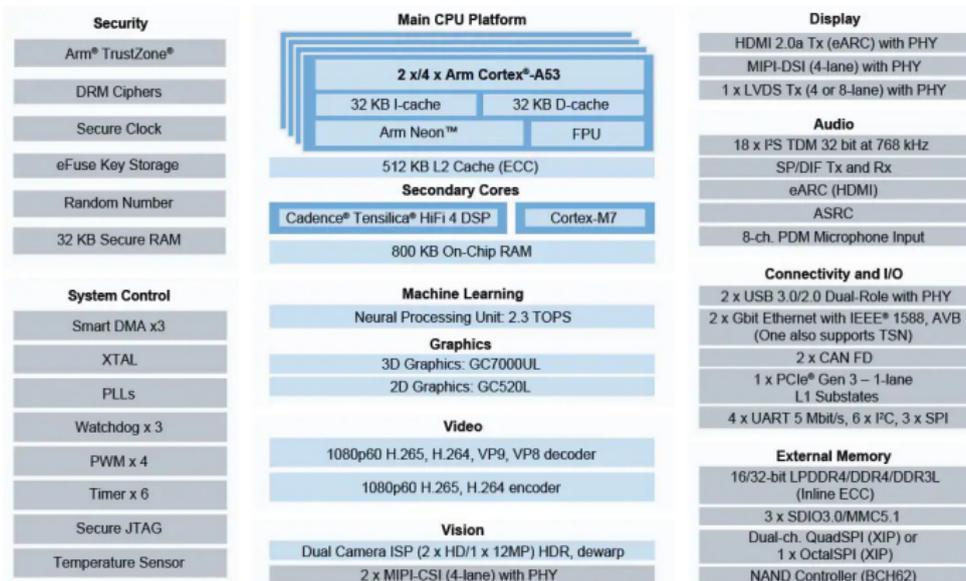
Circuit dédié aux applications « véhicules intelligents »

- 2x ARM Cortex-A15 + 2x ARM Cortex-M4
- 2x DSP C66x
- 2x GPU
- 4x accélérateurs pour traitement d'images

# Le spectre d'implémentation

systèmes hétérogènes

(cont.)



## IMX8M+ de freescale (nxp)

- 4 ARM Cortex A-53
- 1 ARM Cortex M7

- Processeur neuronal NPU 2.3TOPS
- GPU 16GFLOPS
- 2 processeurs d'images 375MPixels/s
- 4 DSP Tensilica HiFi

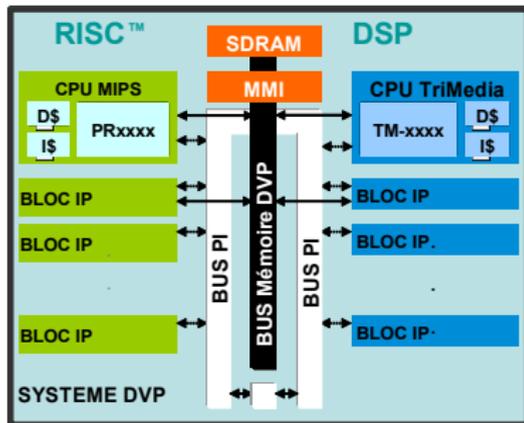
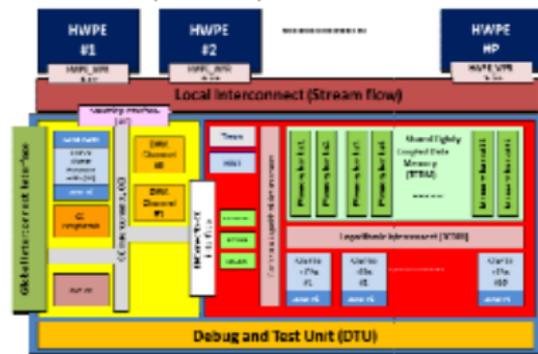
# Le spectre d'implémentation

systèmes hétérogènes

(cont.)

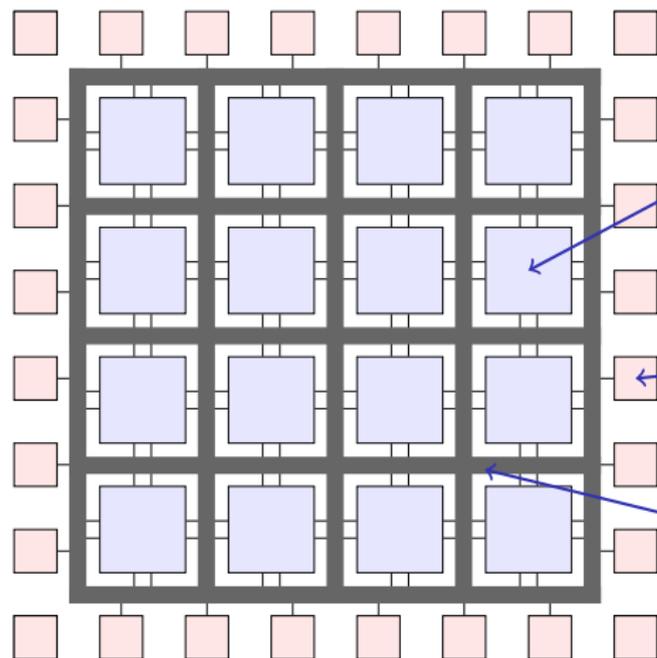
*systems on chip* intégrant des opérateurs matériels spécifiques

La “plateforme 2012” de ST-micro



Le système nexperia (Phillips)

Les circuits programmables FPGA (xilinx, altera, lattice, ...)



blocs logiques pour mettre en oeuvre la logique combinatoire et séquentielle (CLB : *configurable logic blocs*)

blocs logiques spéciaux en périphérie du circuit pour les connexions avec l'extérieur

canaux de routage pour connecter les CLB entre eux et aux entrées-sorties

Les famille de FPGA les plus récentes (Virtex (Xilinx), Stratix (Altera/Intel)):

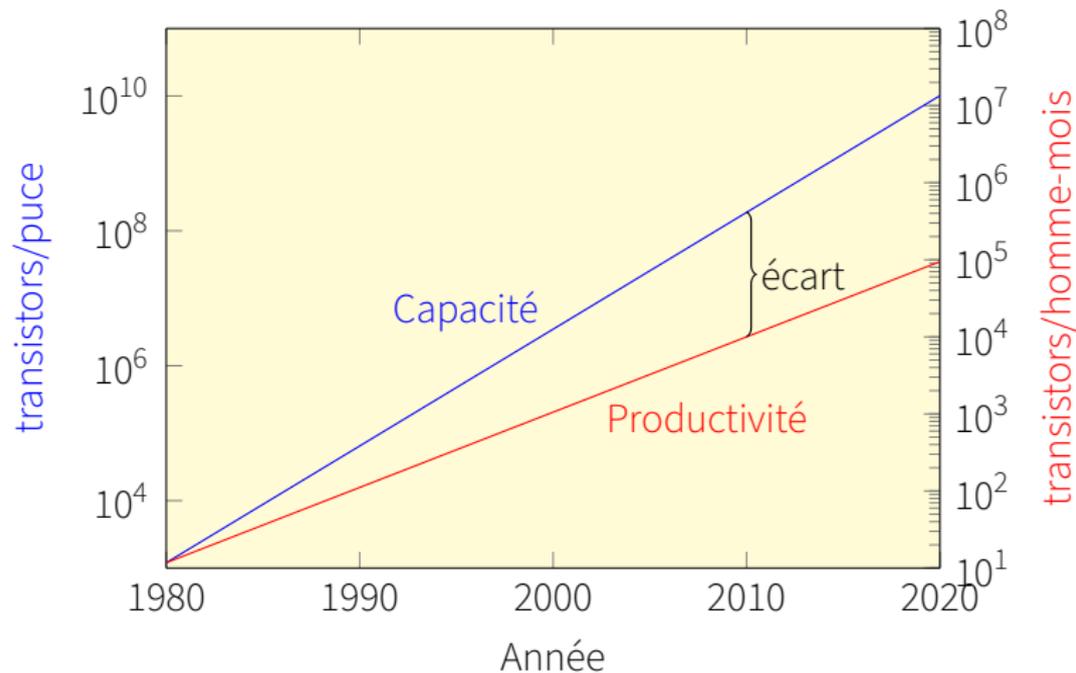
- intègrent des blocs logiques programmables sophistiqués
- comprennent des blocs de mémoire (jusqu'à 500Mo)
- ont des possibilités de calcul élevées (blocs DSP multiplieur-accumulateurs flottants).  
Puissance de calcul crête  $\approx 10\text{--}20$  Tflops
- mise en oeuvre avec des langages de haut niveau (open-CL)

Il peuvent :

- mettre en oeuvre des processeurs *softcore* personnalisables (*customisables*) (nios, microblaze).
- intégrer un ou plusieurs processeurs *hardcore* (Arm)

# Productivity-Design gap

La capacité potentielle des circuits croît plus vite que la productivité des concepteurs



# Etat de l'art (2019)

Transistors  $\times 10^9$

---

CPU : AMD Epyc Rome 64 coeurs	32	7 nm
GPU : Nvidia A100 (Ampere) (6912 CUDA Cores)	54	7nm
FPGA : Xilinx Virtex ultrascale+	35	16nm
Mémoire (DDR5) : 512Gb	600	10nm

Performances crête maximales théoriques (TFlops)

---

CPU ( <i>skylake</i> ) (DP avec AVX-512 et $2 \times$ FMA) (/coeur)	0.1
CPU Arm Cortex A72 (DP avec Neon) (/coeur)	0.006
GPU (7000 SM) (DP)	18
avec <i>tensor cores</i> (SP) (Volta, Turing, Ampere)	150
FPGA (SP)	20

Débit mémoire maximum (Go/s)

---

CPU	100
GPU	600

# Les règles d'or de l'architecture

- Prendre en compte la technologie
- Accélérer le cas le plus fréquent
- Utiliser le parallélisme
- Utiliser le pipeline
- Utiliser la prédiction
- Hiérarchiser la mémoire

Jeux d'instructions

Processeurs pipeline, dépendances, réordonnancement, prédiction de branchement

Parallélisme d'instructions : processeurs superscalaires et VLIW

Hierarchie mémoire, caches, mémoire virtuelle

Optimisations logicielles