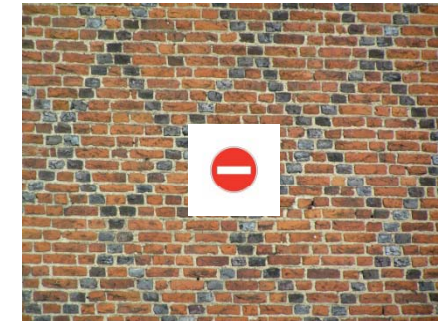
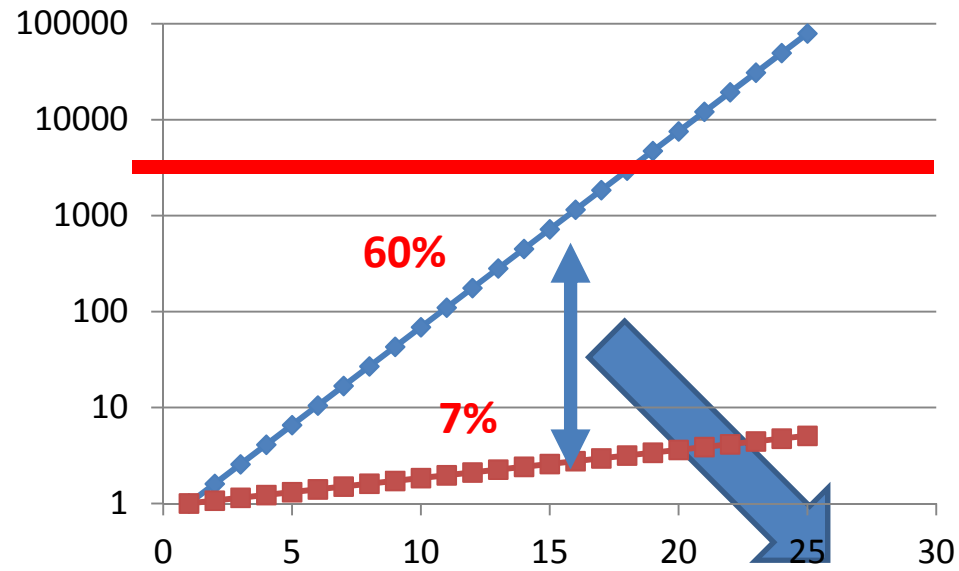


Exponentielles et Murs

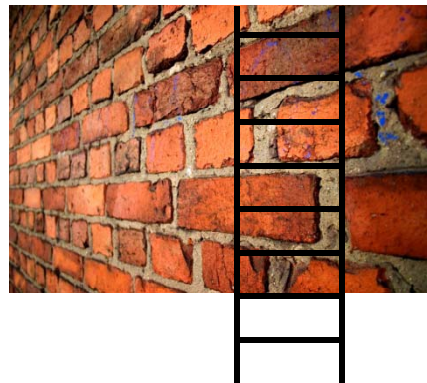
Daniel Etiemble

LRI – Université Paris Sud

Exponentielles et murs



Ne pas franchir



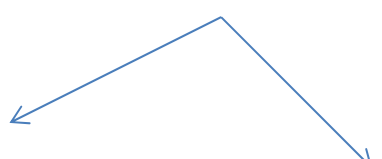
A franchir !

Des équations fondamentales

- Temps d'exécution d'un programme

$$T_{ex} = \frac{NI}{IPC * F}$$

- Puissance dissipée CMOS

$$P_d = P_{stat} + \alpha * \sum C_i * V_{dd}^2 * F$$


= 0 (autrefois !)

≠ 0

Plan (en vrac)

Exponentielles

- Technologie
- Moore
- Performances
- Puissance dissipée
- Coûts de fabrication
-

Murs

- Chaleur
- IPC
- Mémoire
- Conception
- Programmation parallèle ?
- ...

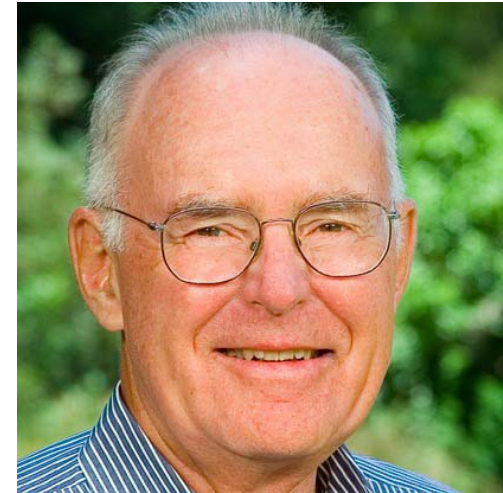
Relations entre exponentielles et murs

Les exponentielles

- En profiter
 - Utilisation dans les applications
 - Puissance de calcul, capacité mémoire, etc.
 - Exemple : la traduction automatique
- Les combattre
 - Continuer à réaliser la progression exponentielle
 - Performances : microarchitecture et architecture
 - Traitement : explosion de la taille des données...
- Les deux à la fois
 - Exemple : calcul scientifique

La loi de Moore

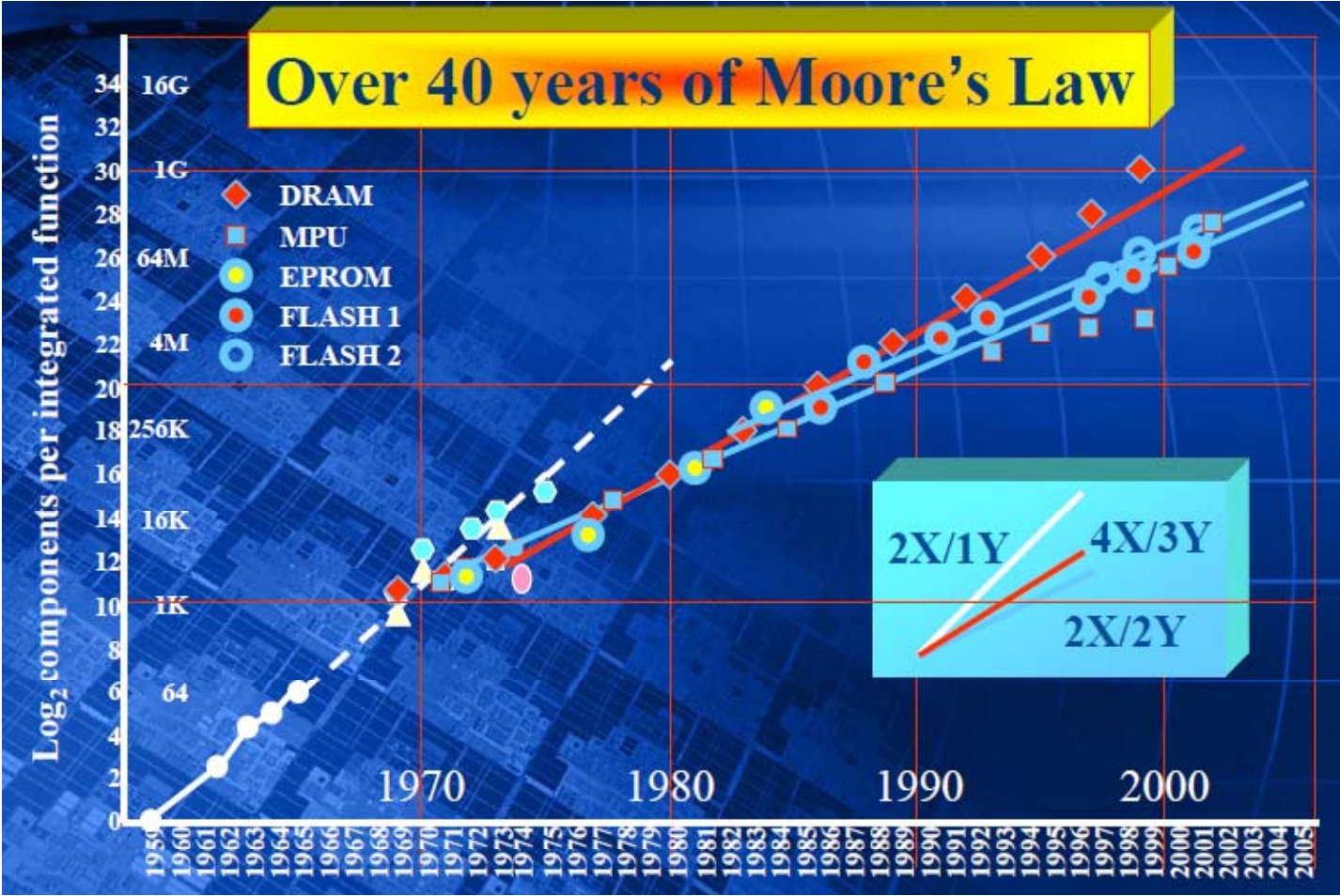
- Version correcte
 - Le nombre de transistors par puce double tous les ans / **18 mois** / 2 ans
- Versions « incorrectes »
 - Doublement tous les deux ans
 - De la performance
 - Du nombre de cœurs par puce
 - Etc



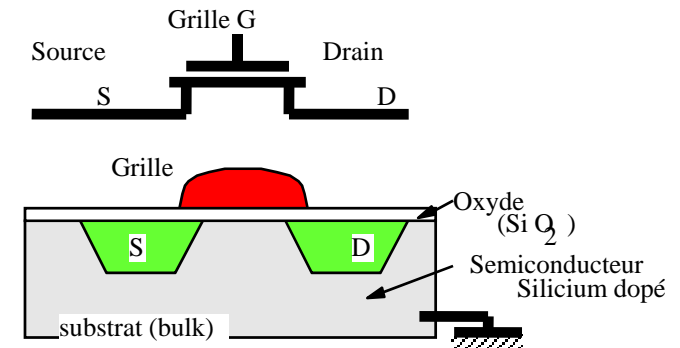
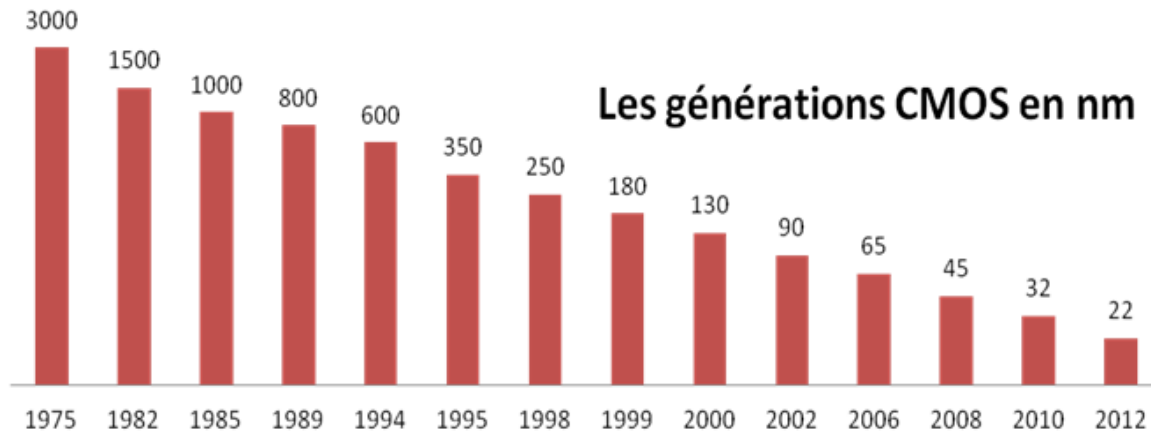
La loi de Moore

IEEE March 11, 2014

P.Gargini



Les « nœuds » technologiques



- **Une fin ?**

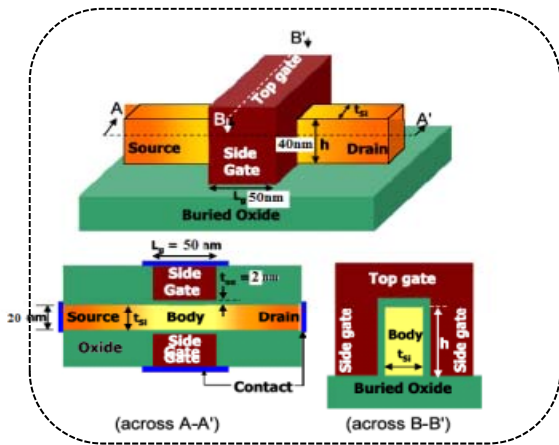
- Il est difficile de faire des prévisions (surtout pour le futur)
 - Voir les prévisions passées et la réalité
- Vers les limites physiques
- Quand ?

Remarques sur les prévisions

- Paolo Gardini (Intel Fellow – ITRS Chairman)
 - <http://www.ewh.ieee.org/r6/scv/eds/slides/2014-Mar-11-Paolo.pdf>
- Lesson 2
 - « Predictors of Engineering Limits have Always been Proven Wrong by The Right Improvements »
- Lesson 3
 - « It Would be Wrong to Believe that the Right Fundamental Limits Don't Exist »

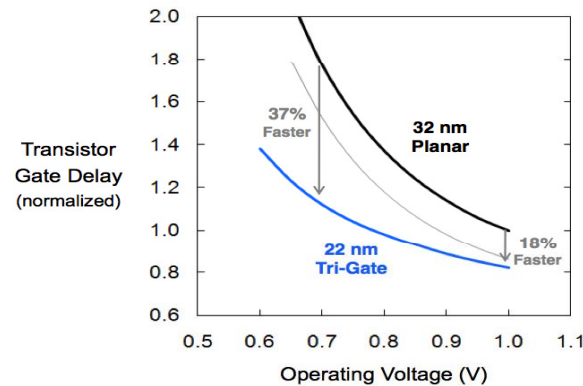
« Predictors of Engineering Limits have Always been Proven Wrong by The Right Improvements »

- Un exemple
 - Transistor 3D et Technologie FinFET

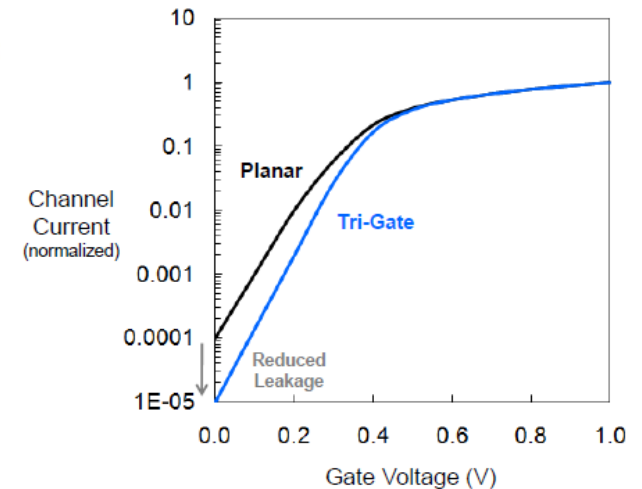


Vitesse

Transistor Gate Delay

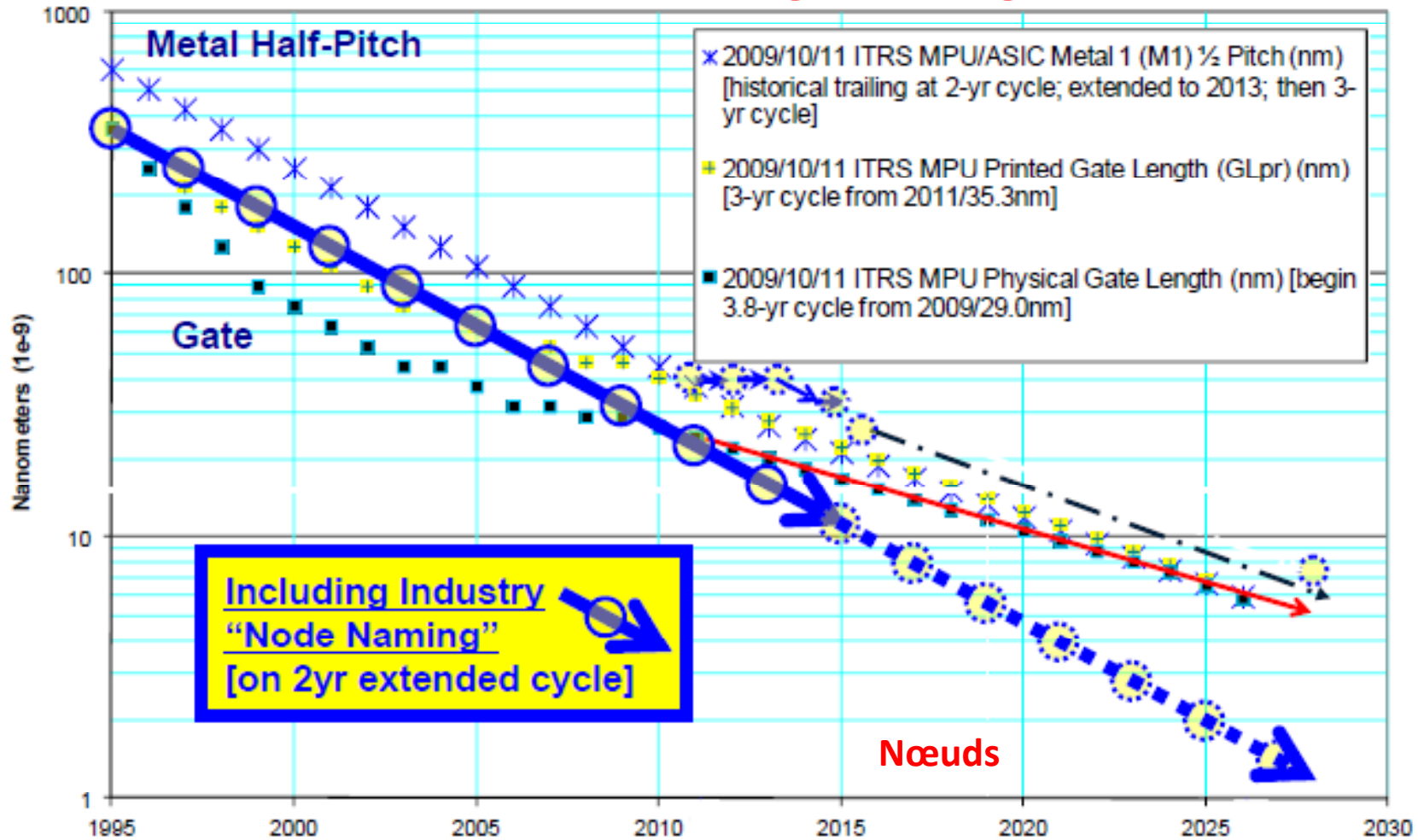


Courant de fuite



« It Would be Wrong to Believe that the Right Fundamental Limits Don't Exist »

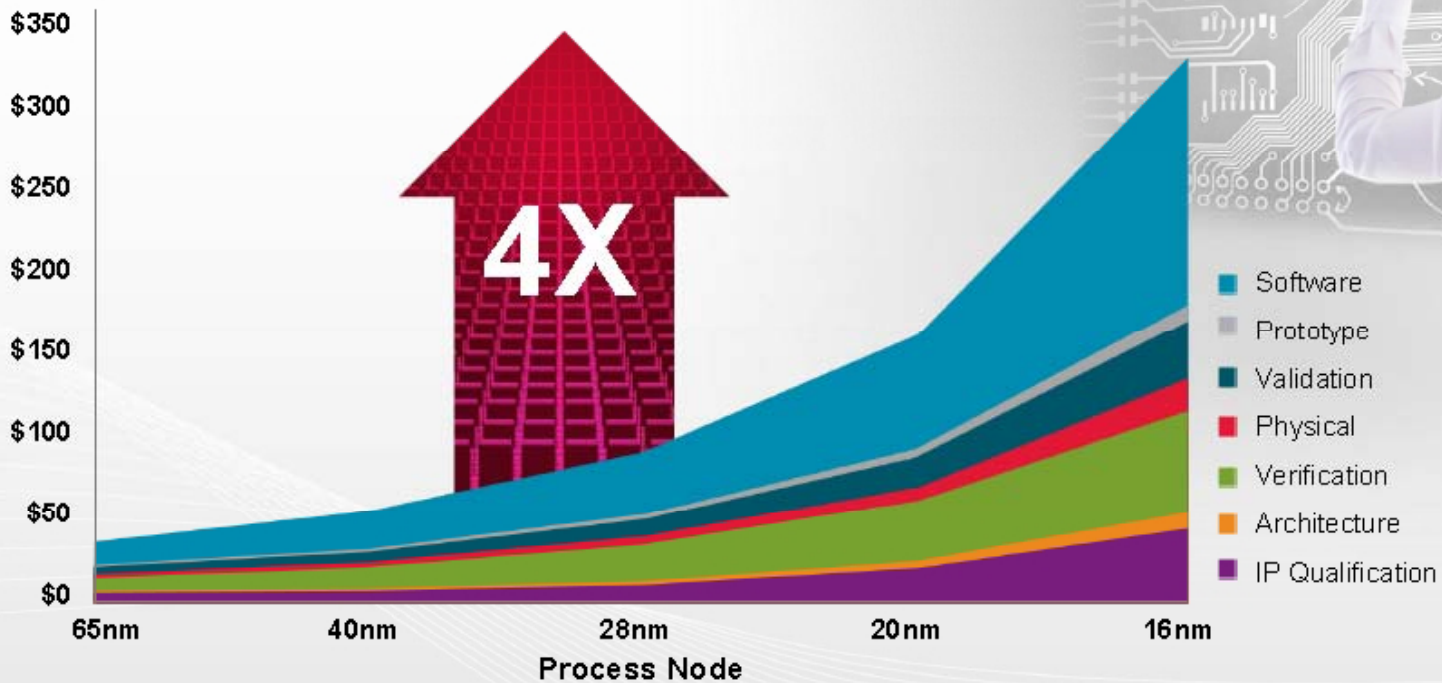
Nœuds et longueur de grille



Coûts de fabrication

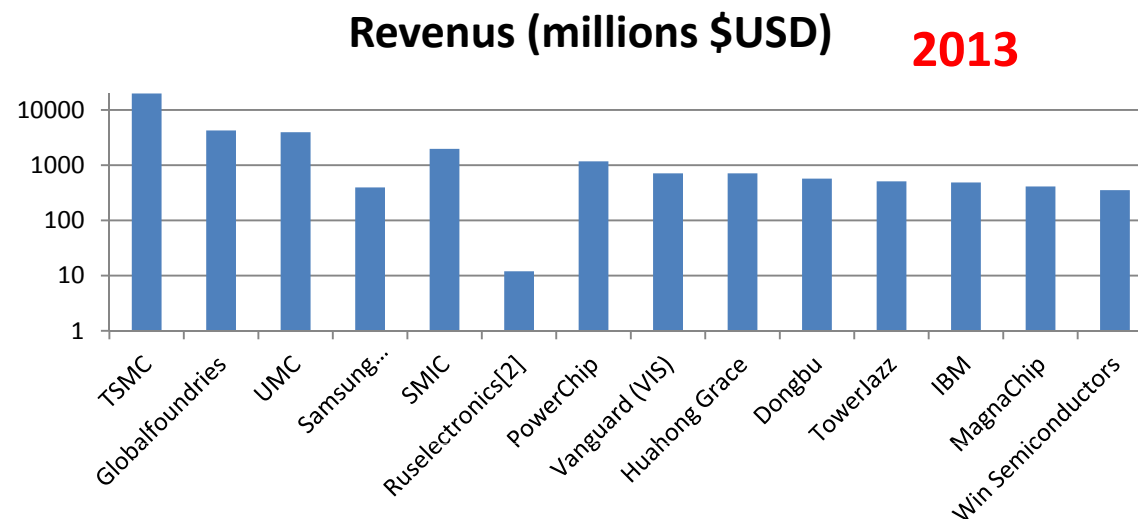
DRAMATIC RISE IN DESIGN COSTS

Design Cost (\$M)



Fonderies pour les autres (14 13)

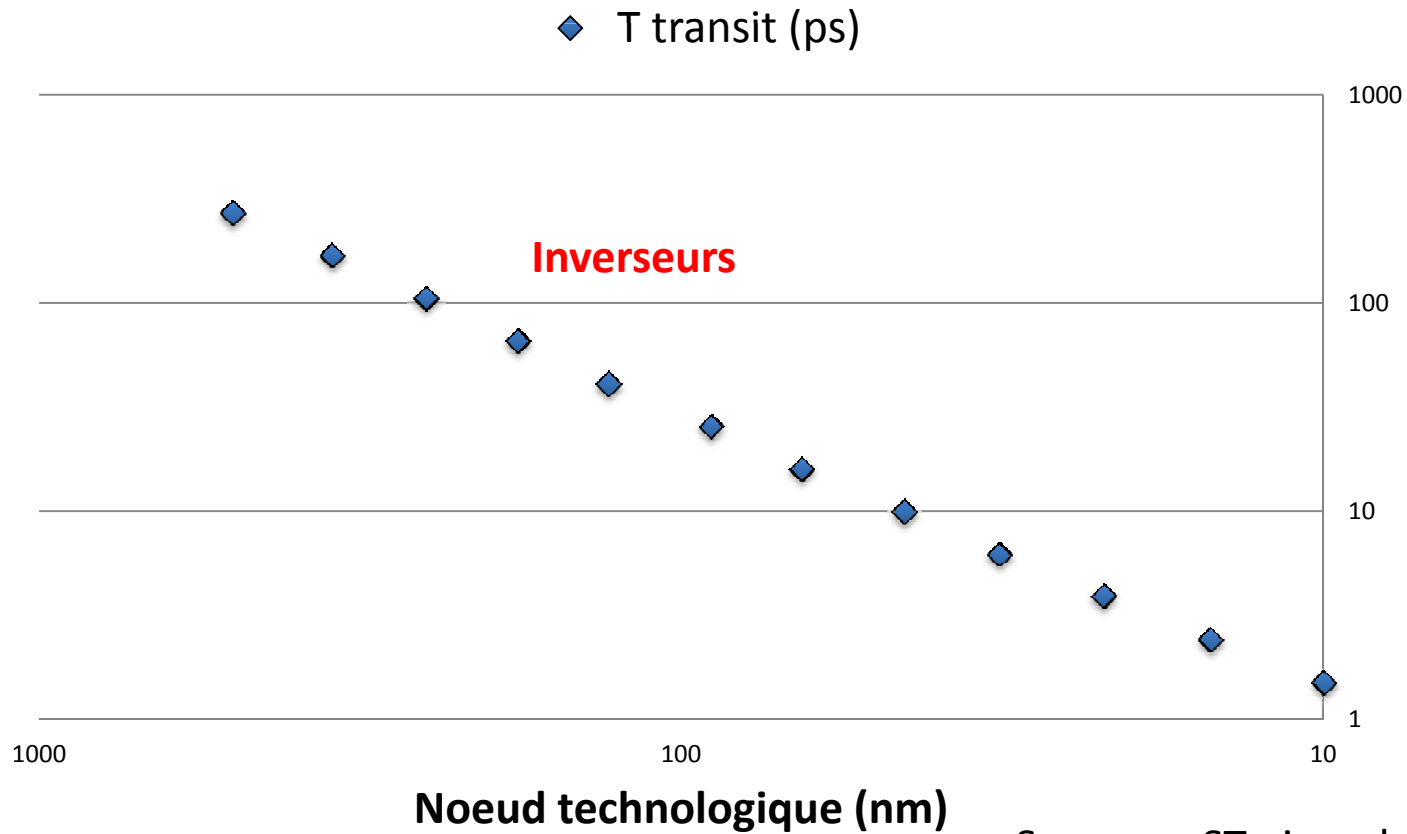
Croissance exponentielle des coûts de fonderie (*loi de Rock*)



GLOBALFOUNDRIES TO ACQUIRE IBM'S MICROELECTRONICS BUSINESS

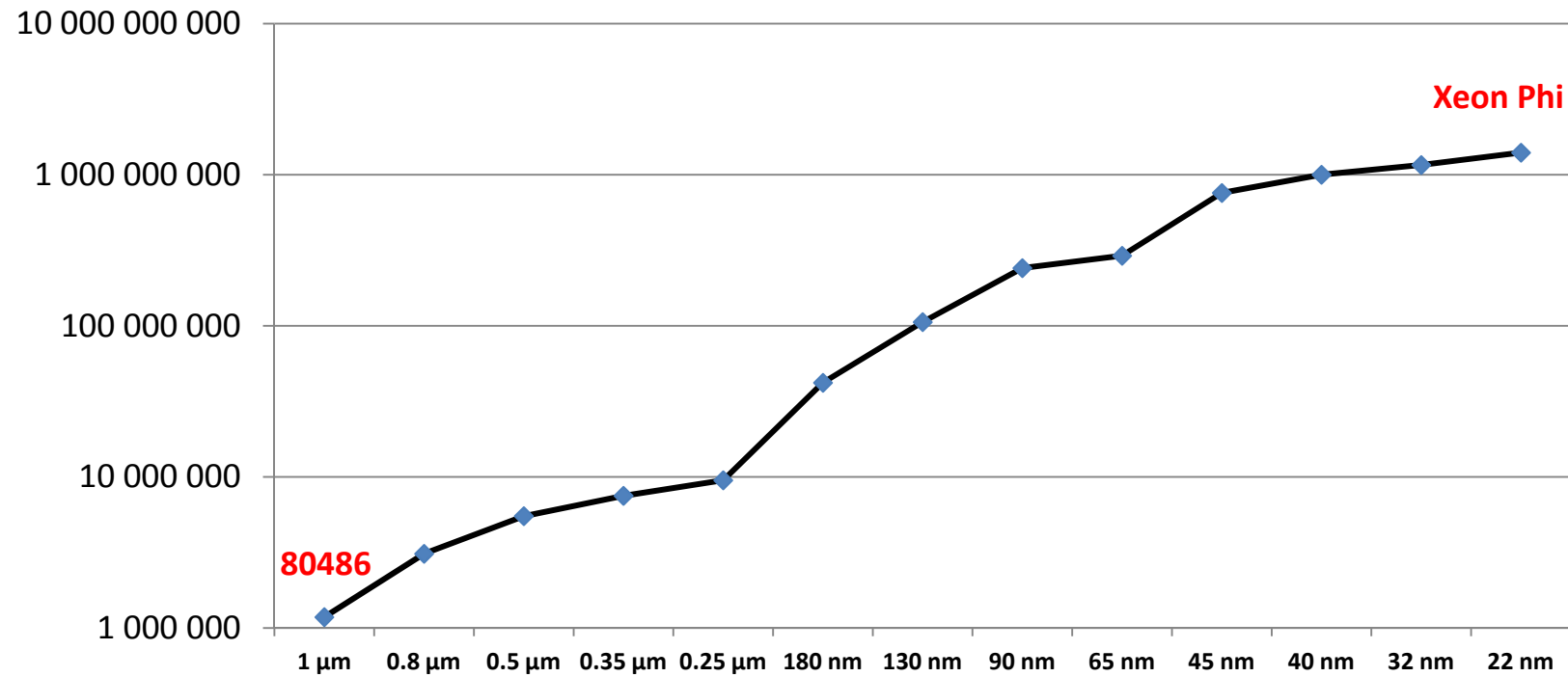
Acquisition Enables GLOBALFOUNDRIES to Become a World Leader in Semiconductor Foundry Technology;

Vitesse des portes



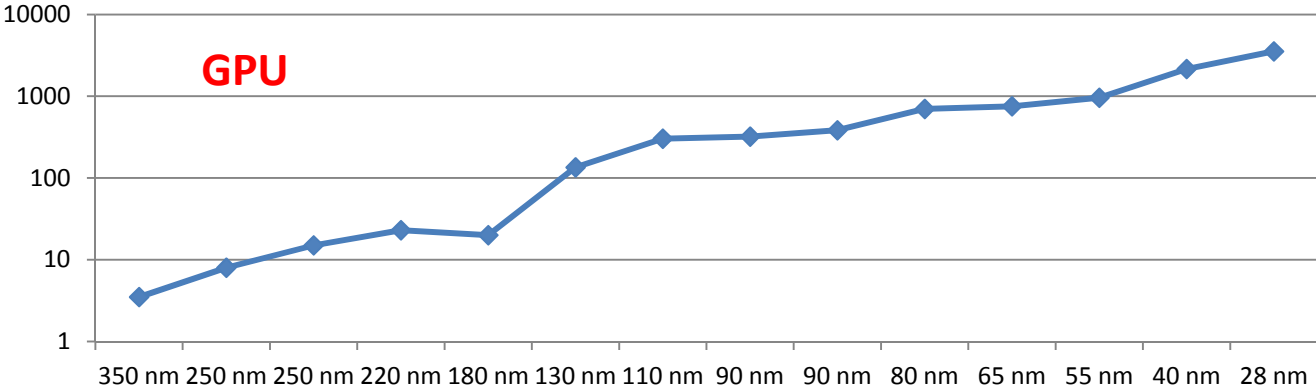
Sources : STmicroelectronics

Nombre de transistors (microprocesseurs)

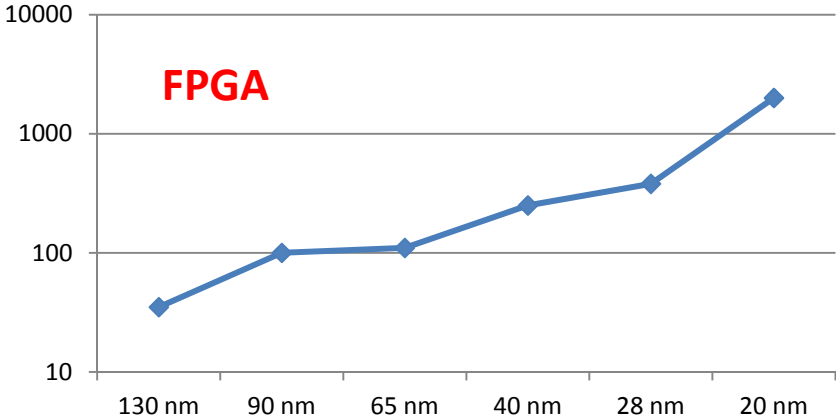


Nombre de transistors

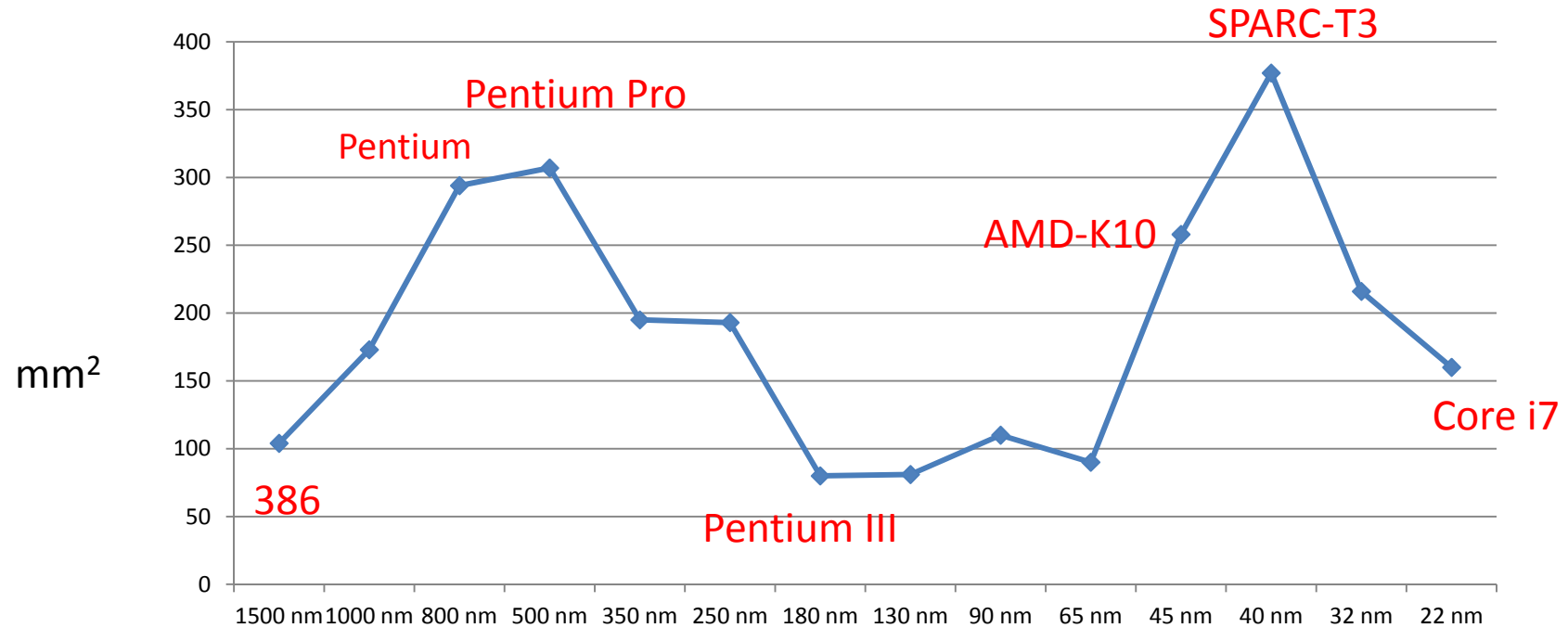
Millions de transistors



Millions de transistors



Surface de puce



Surfaces entre 1 et 7 cm²

Neon Nehalem – 45 nm – 685 mm²

Power8 – 22 nm – 650 mm²

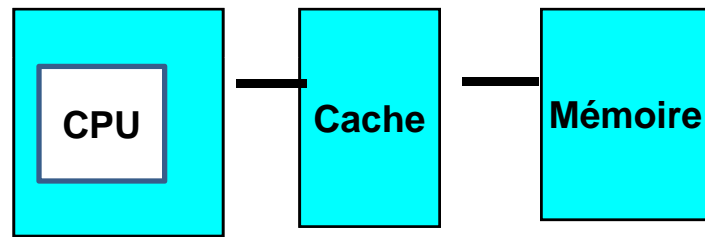
Une ou deux puces

- Opérateurs flottants
 - Coprocesseurs (ex : x87)...
 - Intégrés
- Caches
 - Externes
 - Internes (L1 puis L1-L2 puis...)
- CPU et GPU
 - 2 puces distinctes
 - GPU intégré (APU)
- Multi-cœurs
 - 1 puce
 - Cluster de multi-cœurs

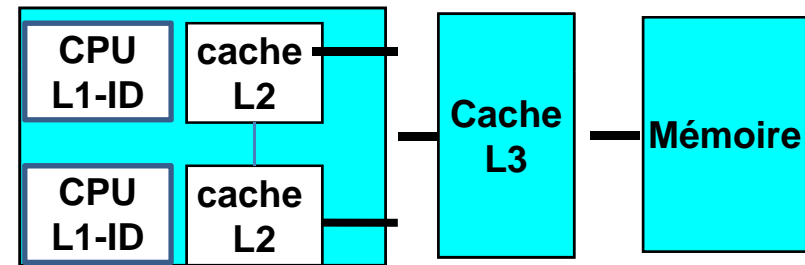
**Du coprocesseur à
l'intégration
dans la même puce**

**De puces distinctes
à une seule puce**

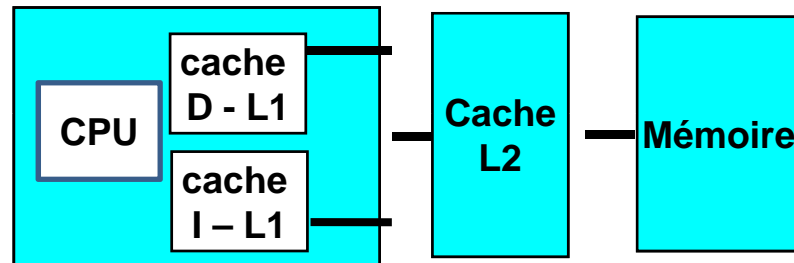
Evolution des caches



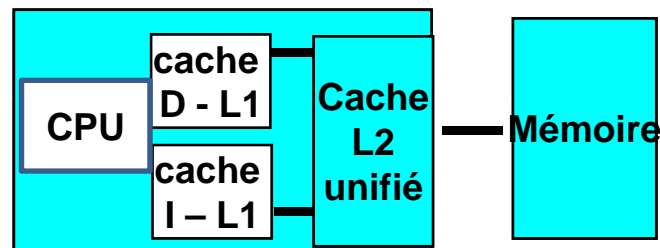
386



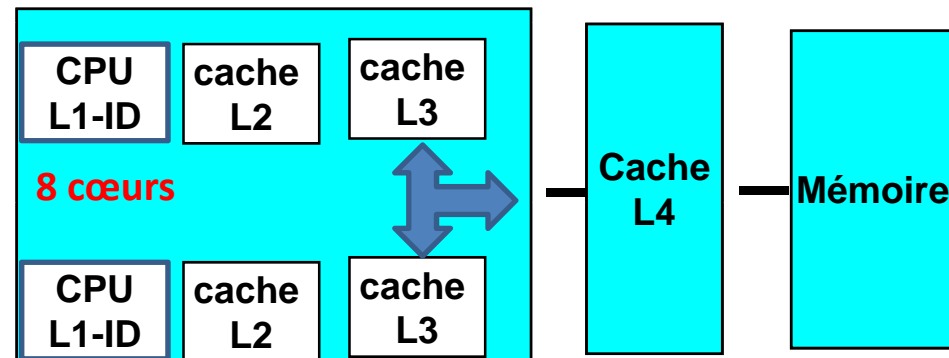
Power6



Pentium



Pentium 4

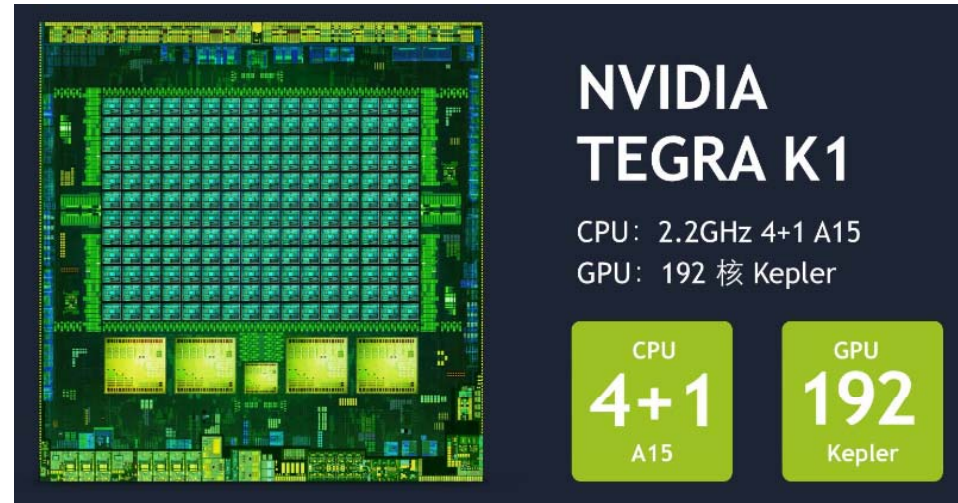


8 cœurs

Power 8

Un ou plusieurs circuits intégrés

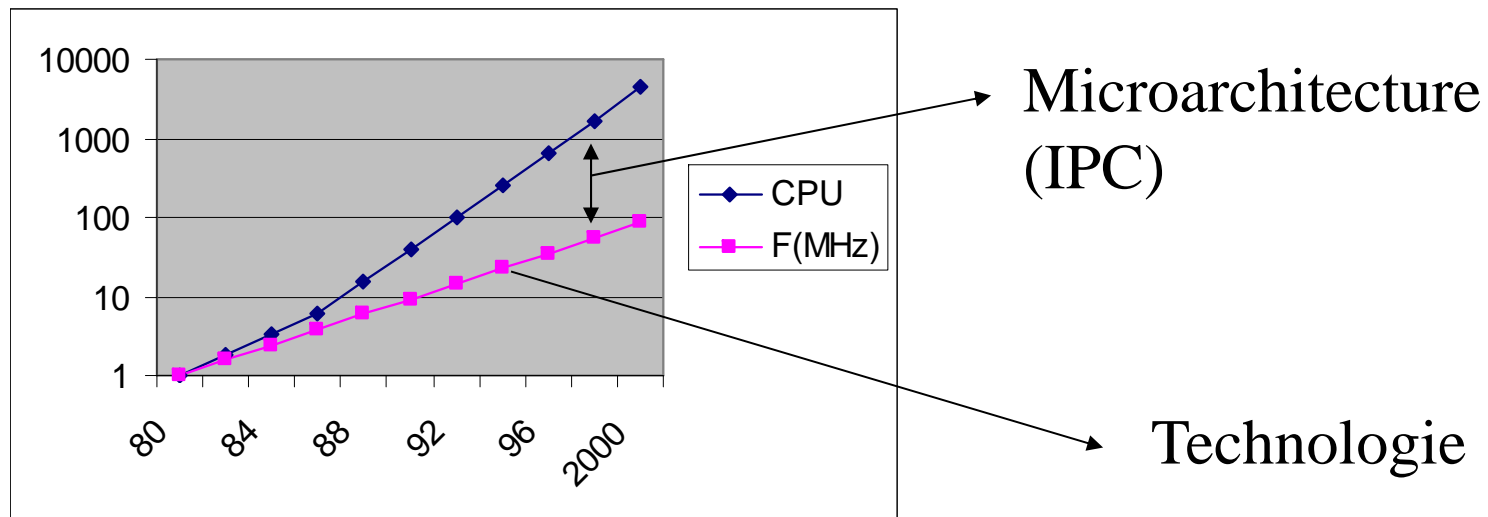
- Exemple : GPU ou APU
 - APU = CPU + GPU dans la même puce.



	Tesla K40 + CPU	Nvidia Tegra K1
Single Precision Peak	4.2 TeraFlops	326 GFlops
Single Precision SGEMM	3.8 TeraFlops	290 GFlops
Memory	12GB @ 288GB/s	2GB @ 14.9GB/s
Power (CPU + GPU)	~ 385Watt	<11Watts
Performance Per Watt	10SP GFlops Per Watt	26SP GFlops Per Watt

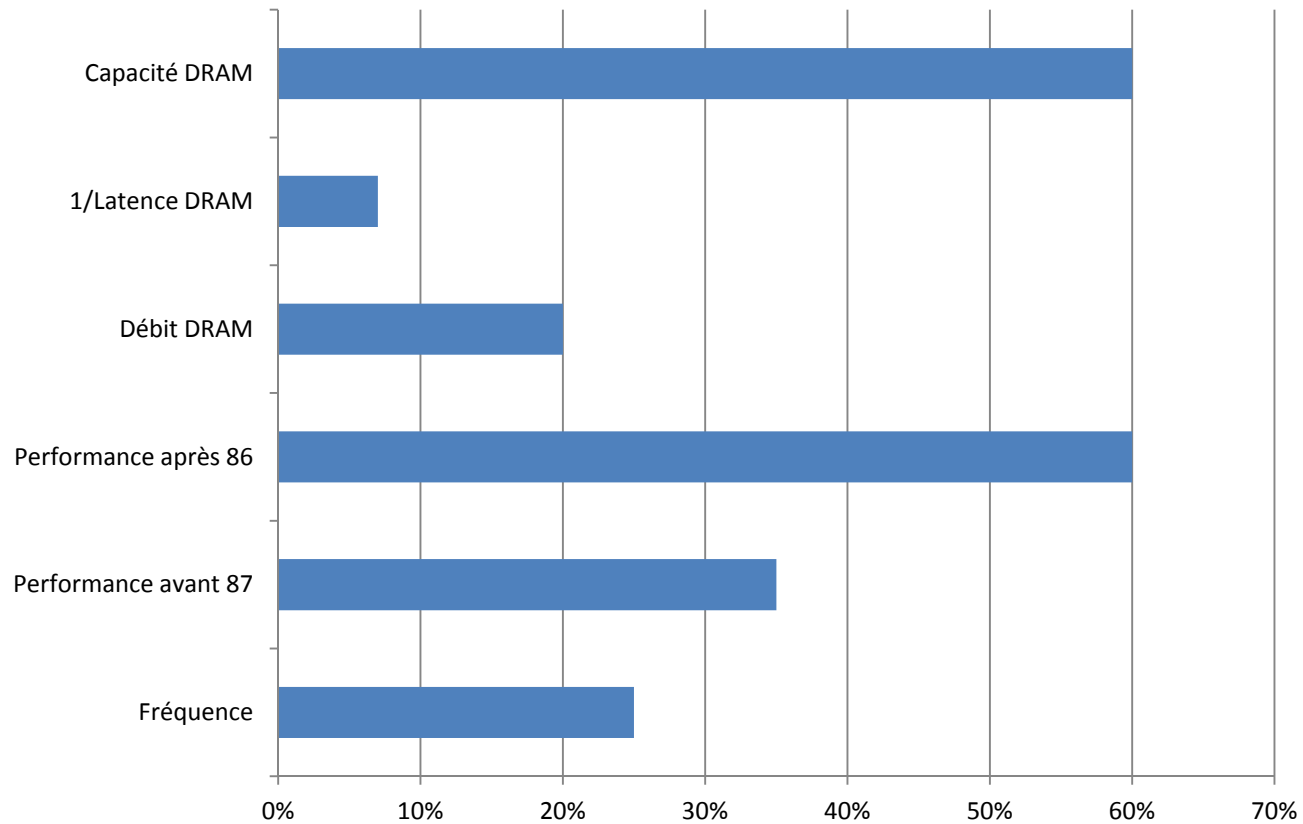
Performance microprocesseurs (1980-2005)

$$T_{\text{exécution}} = \text{NI} * \text{CPI} * T_c = \frac{\text{NI}}{\text{IPC} * F}$$



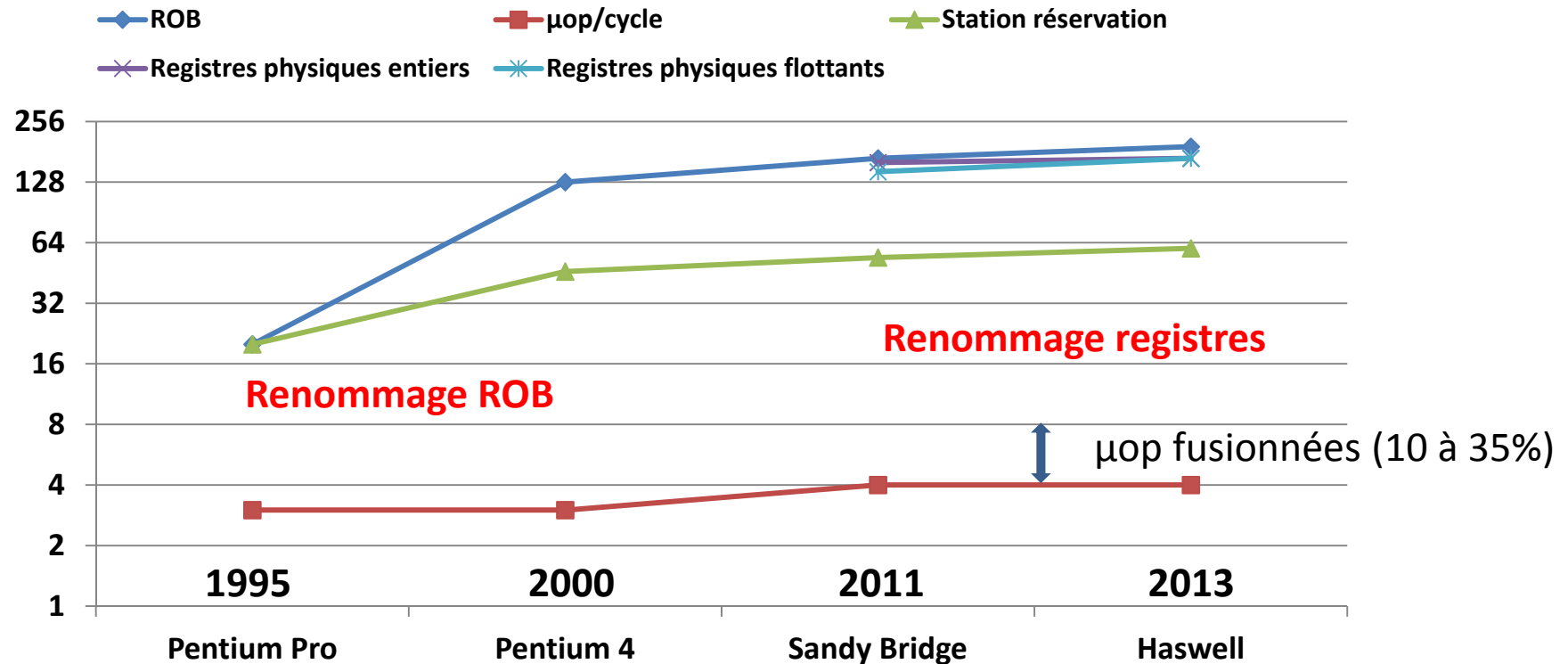
Exponentielles...(plus exponentielles que d'autres)

- Evolution / an



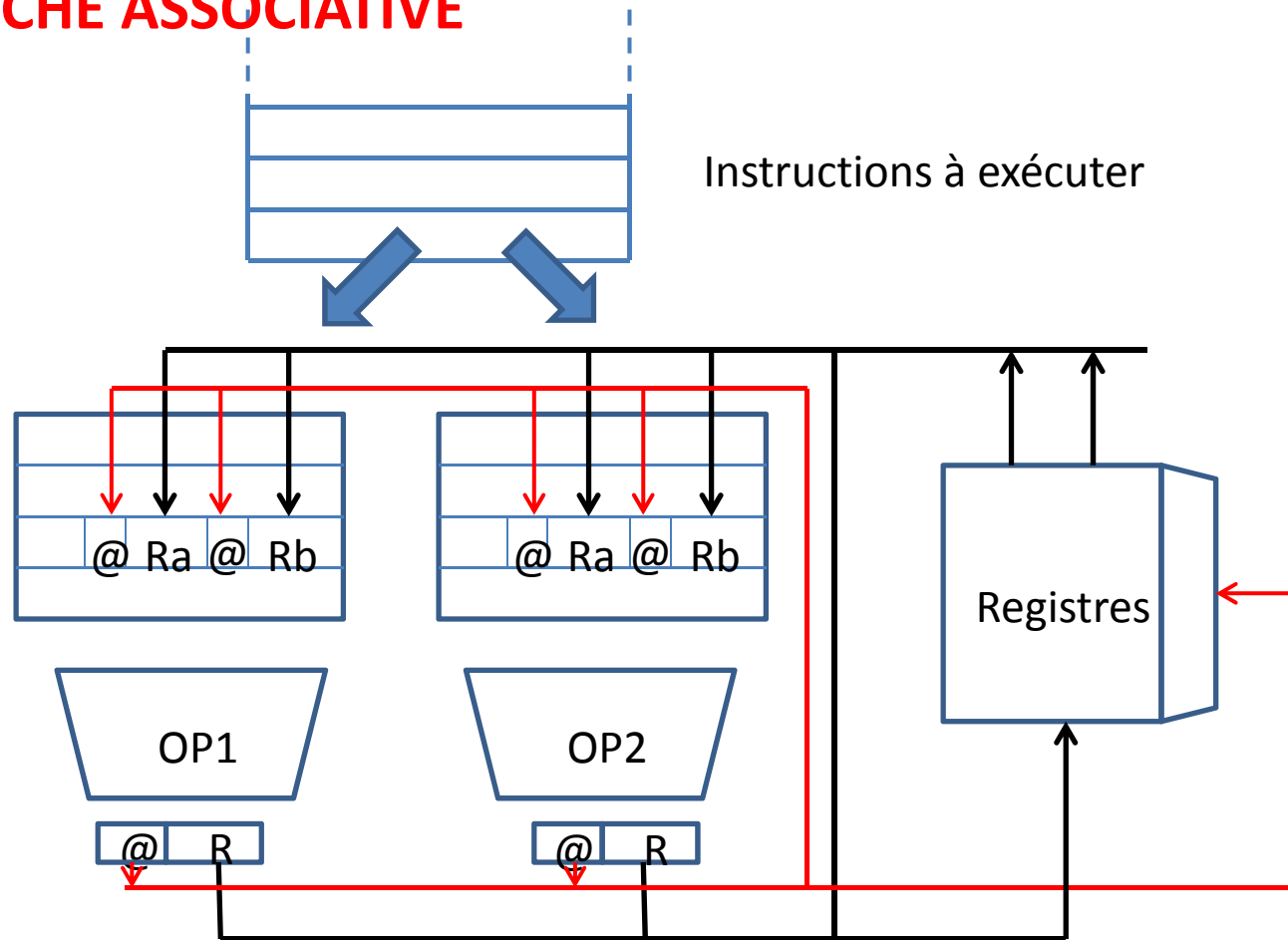
Mur de l'IPC

- Parallélisme d'instructions
 - Nombre d'instructions exécutables/cycle
- Evolution très limitée de 1990 à 2015



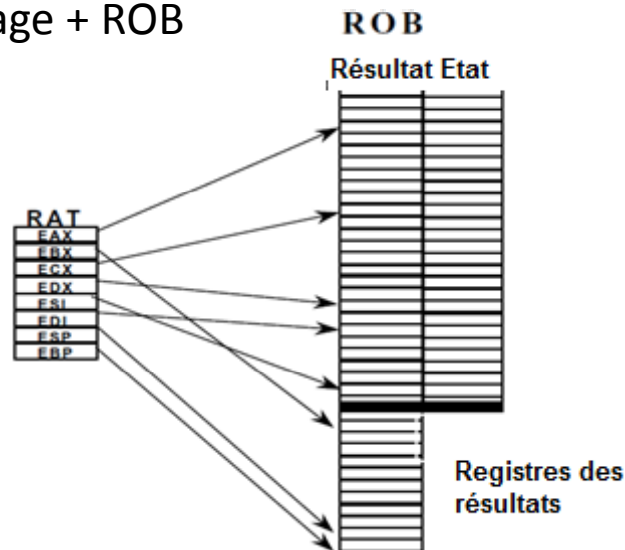
Algorithme de Tomasulo

RECHERCHE ASSOCIATIVE

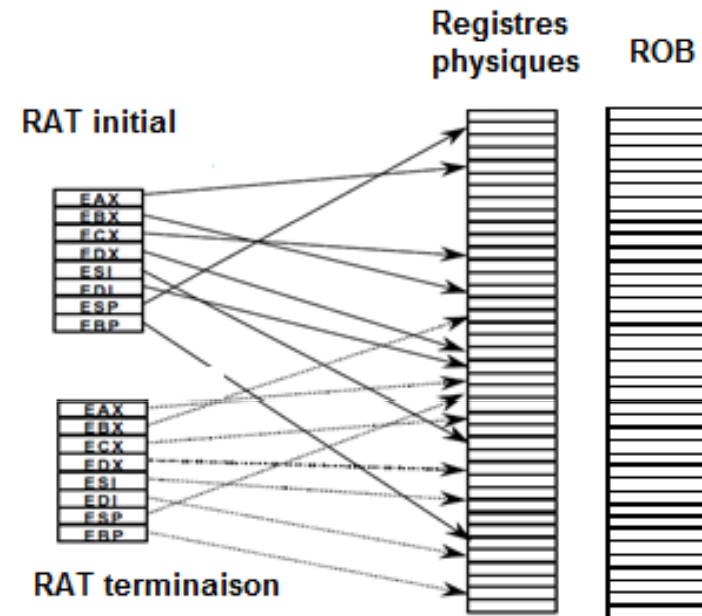


o-o-o : simplifier les recherches associatives

Renommage + ROB



Pentium Pro, PIII, Nehalem



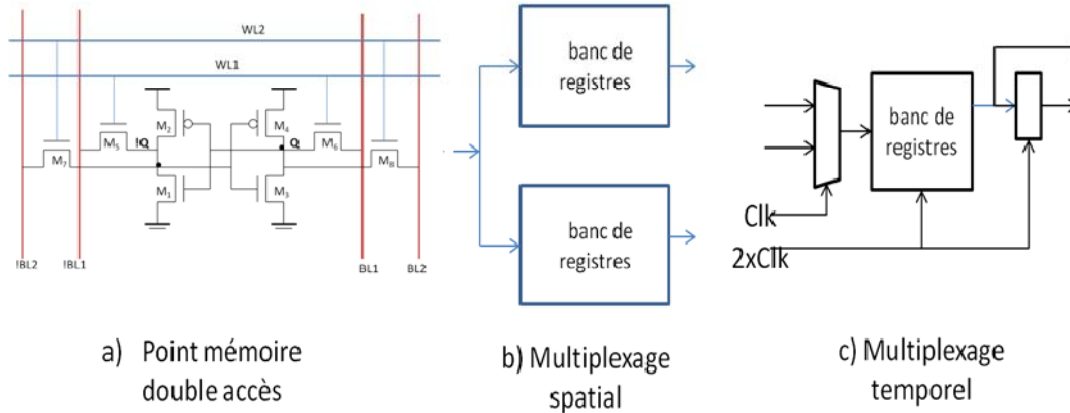
Netburst (P4)... Sandy Bridge, Haswell

Recherches associatives complexes

- Bibliothèques versus « full custom » (cas ARM)

Banc de registres multi-ports

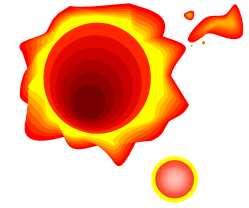
- Registres double accès



- Registres N ports

- Temps accès proportionnel à N
- Surface proportionnelle à N^2

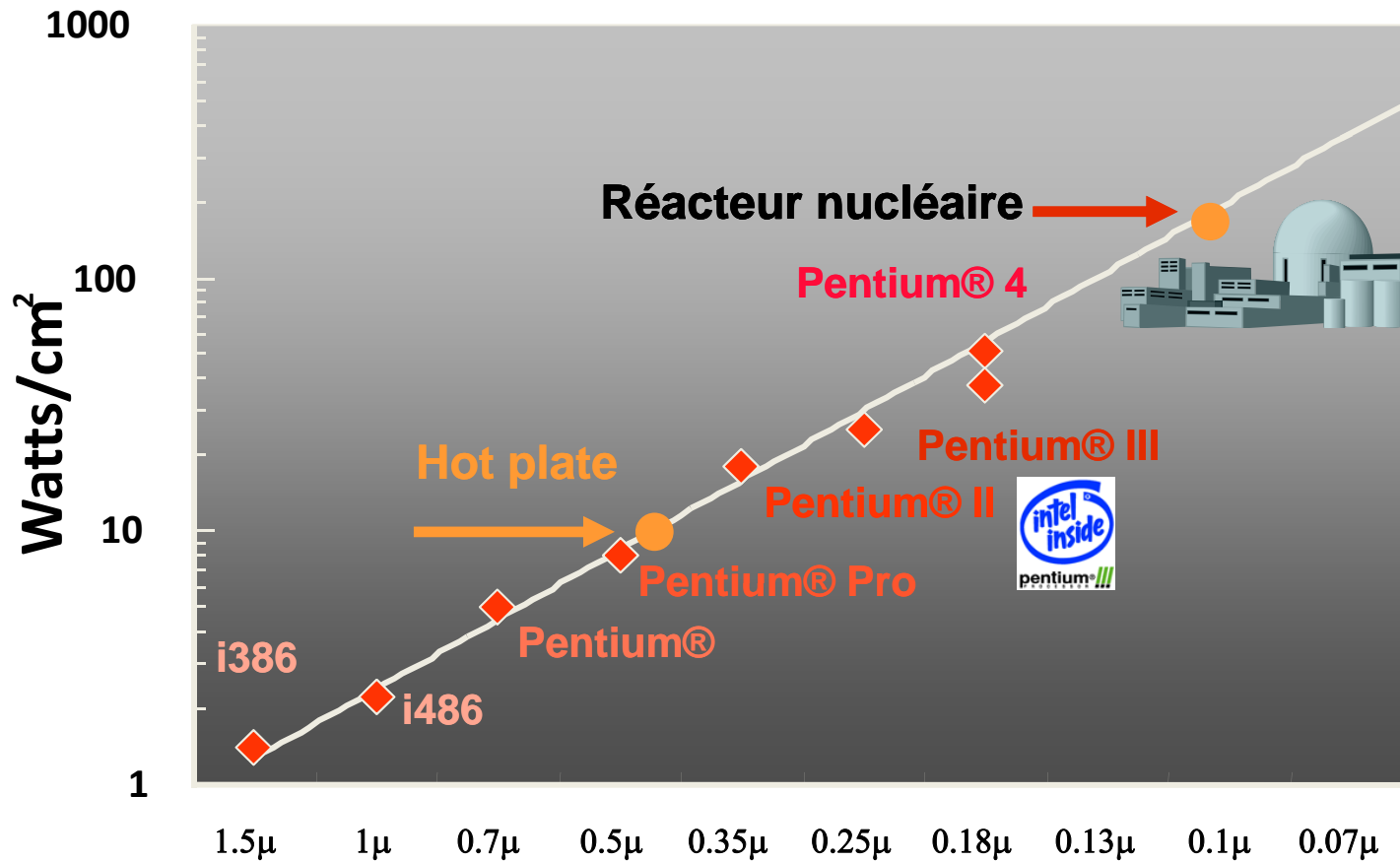
Le mur de la chaleur



*Surface
du soleil*

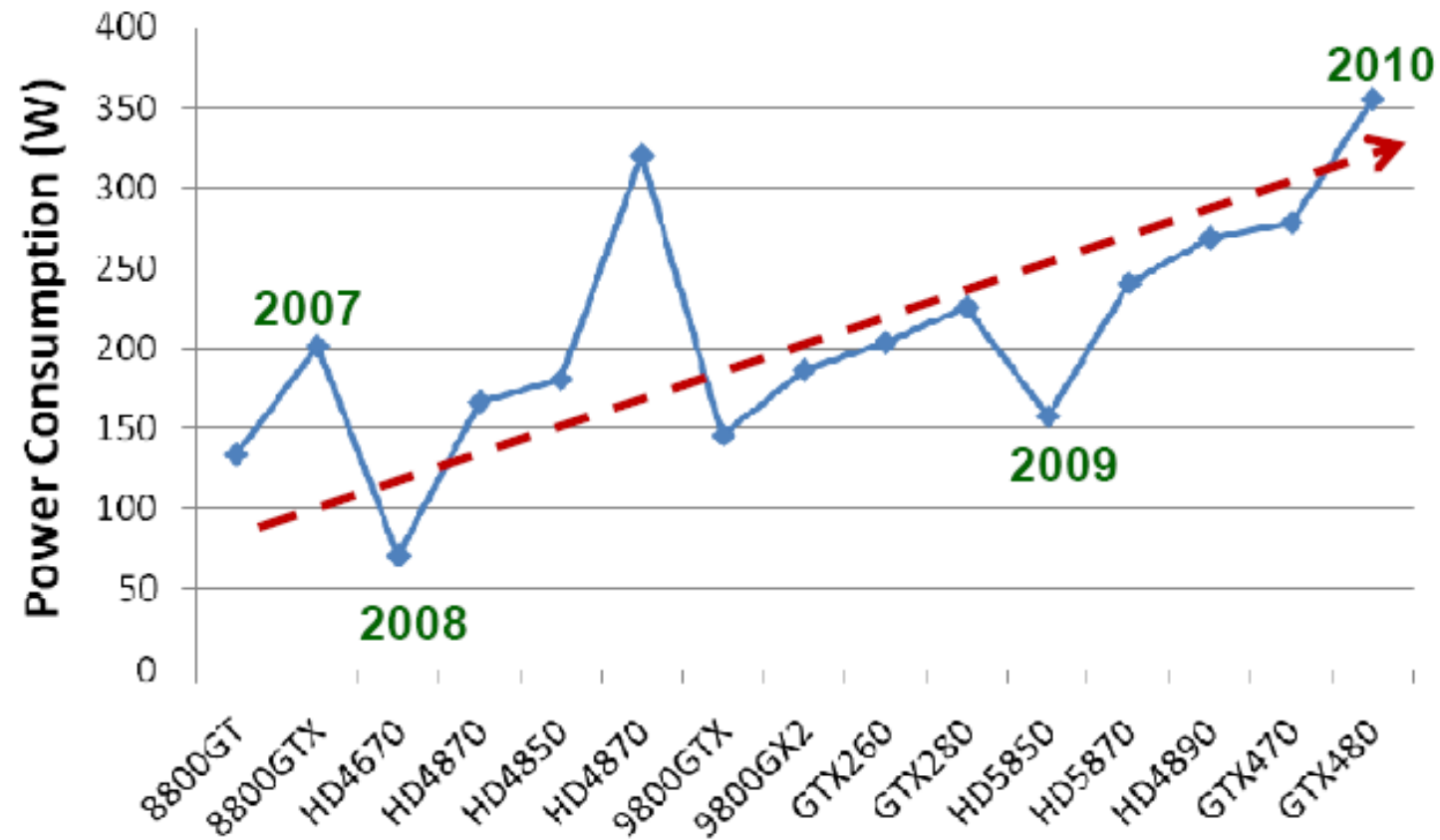


**Traîne
fusée**



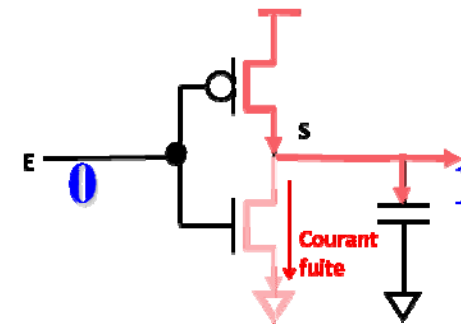
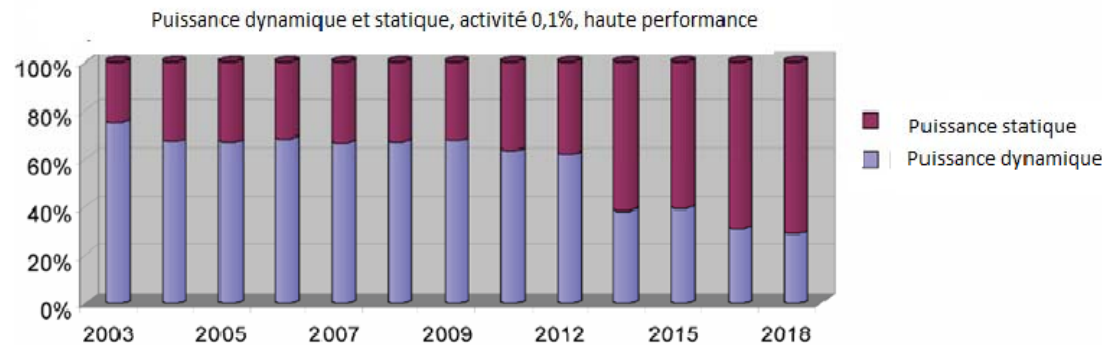
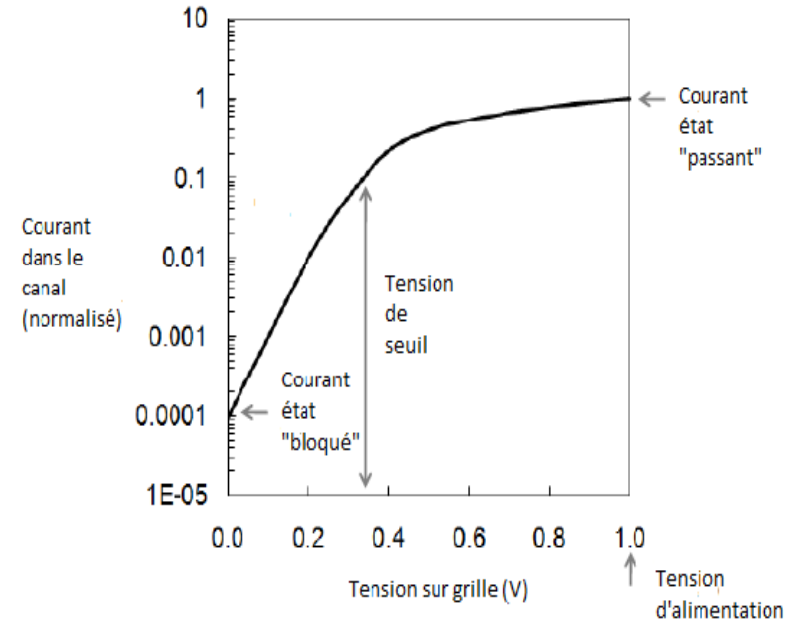
* “New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies” – Fred Pollack, Intel Corp. Micro32 conference key note - 1999.

Puissance dissipée - GPU



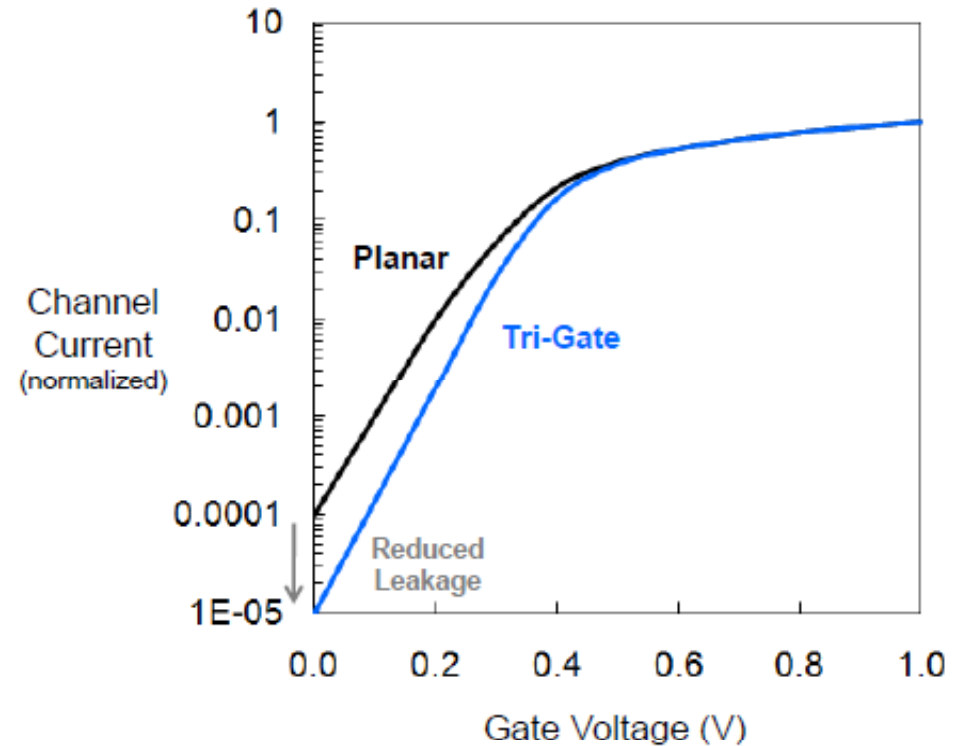
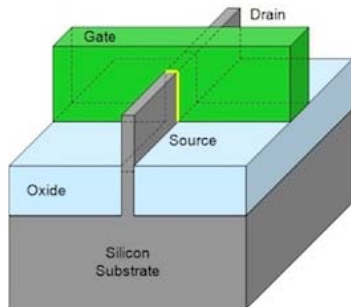
Puissance dissipée en CMOS

- $P_d = P_{d_{statique}} + P_{d_{dynamique}}$
- Puissance statique
 - Courants de fuite
- Puissance dynamique
 - $P_{dyn} = \alpha \cdot \sum C_i \cdot V_{dd}^2 \cdot f$
 - Fréquence
 - Tension d'alimentation



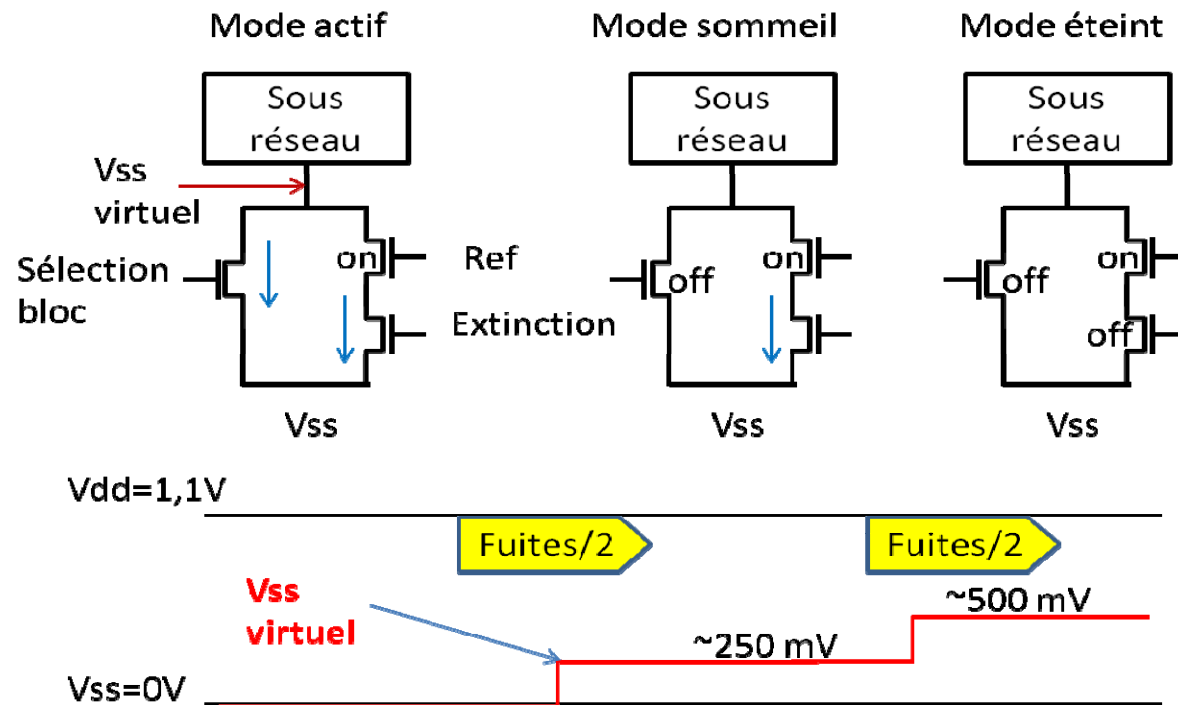
Puissance statique : technologie

- Technologie
 - Ex : Technologie Sol (IBM) réduit les capacités
 - Ex : transistor 3D (Intel)



Puissance statique : circuiterie

- Circuiterie
 - Ex : Masses virtuelles (caches du Xeon)



Puissance dynamique

- Circuiterie
 - « clock gating » : n'activer que les parties utiles
 - Plusieurs tensions d'alimentation
 - Transistors rapides, moyens et lents (tension de seuil V_t)
- **Architecture**
 - **Fréquence d'horloge (F)**
 - **Complexité du circuit**
 - **Processeur**
 - **Caches**

N'activer que les parties utiles

Plusieurs modes de fonctionnement

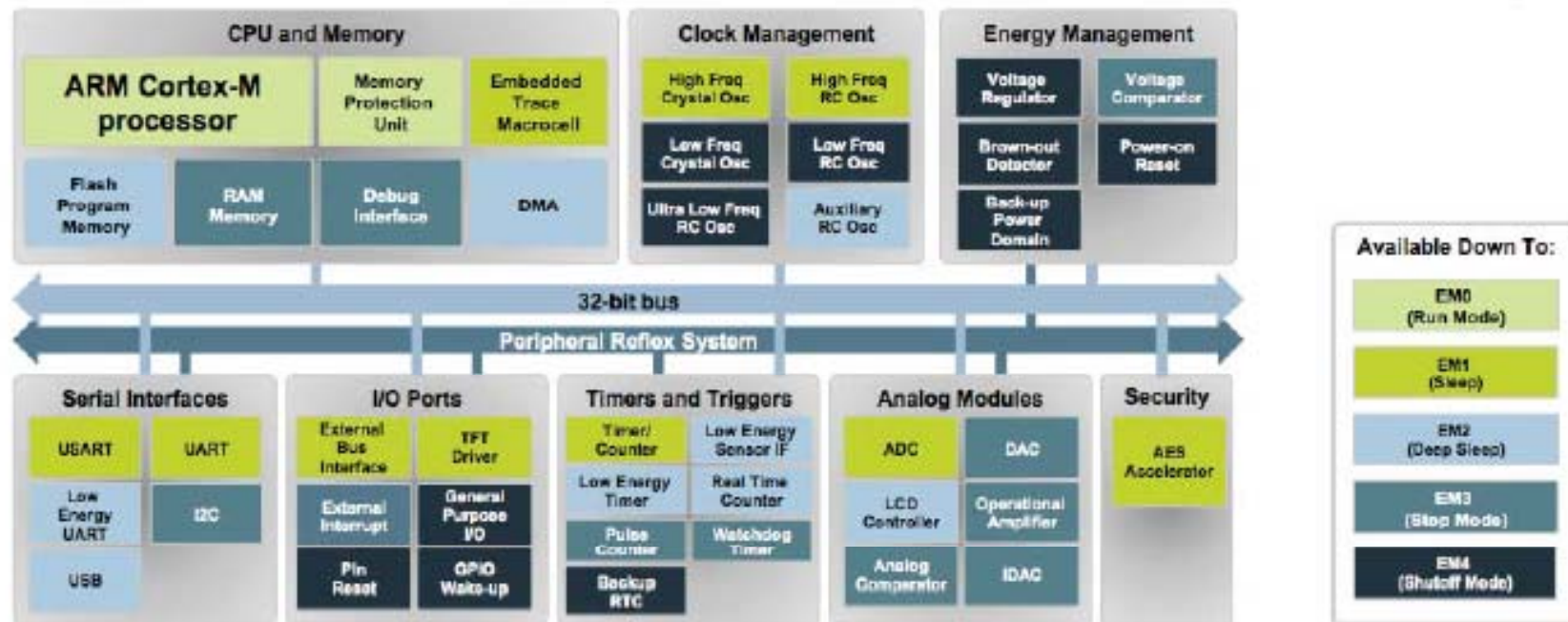
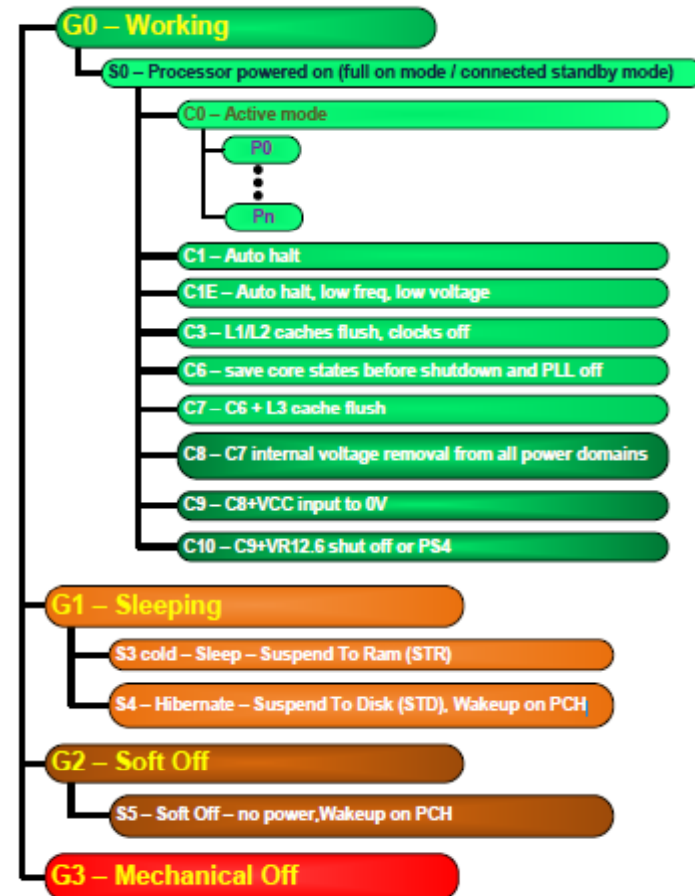


Figure 2. EFM32 Gecko block diagram and the five energy modes.

5^{ème} génération Intel Core

Core M
Pentium
Celeron

LES DIFFERENTS ETATS



* Note: Power states availability may vary between the different SKUs

Mur de la chaleur

- Arrêt de l'augmentation de F
 - Fin du « free lunch »
 - Haut et milieu de gamme
 - Tournant vers les multi-cœurs
 - Bas de gamme
 - Simplification des architectures
- Utilisation du parallélisme
 - Parallélisme de données dans les monoprocesseurs (SIMD et SIMT)
 - Multi-cœurs

Parallélisme de données : réduire NI

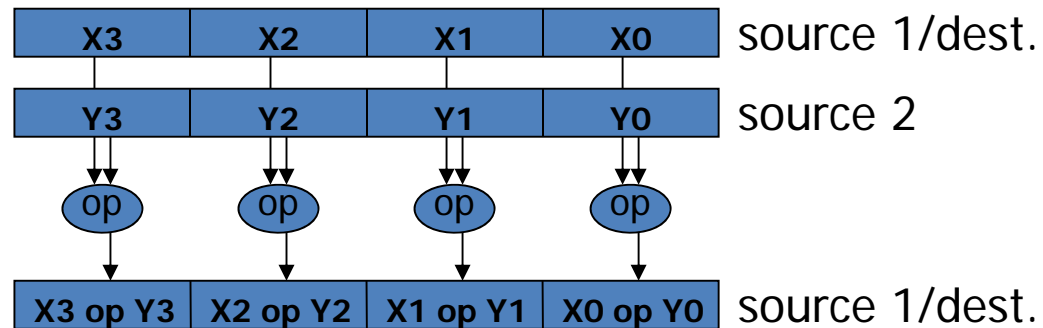
$$T_{ex} = NI * CPI * T_c$$

CPU

SIMD

1 instruction à
plusieurs données

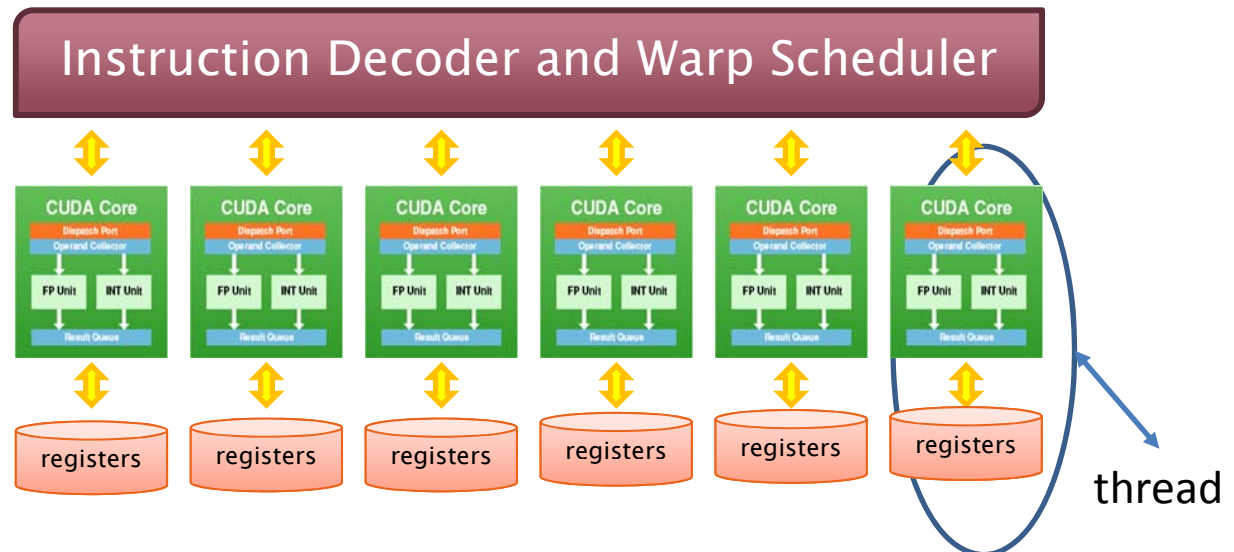
SSE2/3/4 – Neon - AltiVec



GPU

SIMT

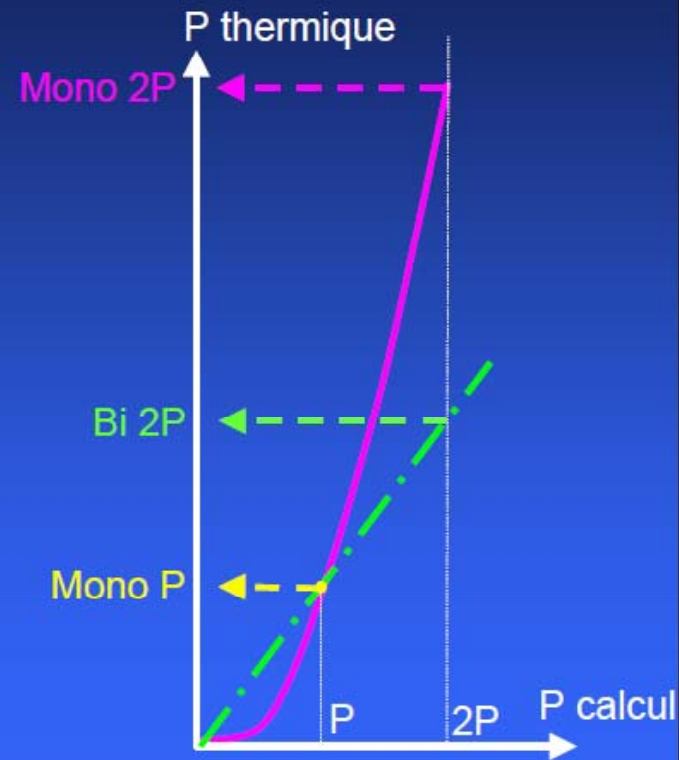
1 instruction pour
plusieurs threads



Multiprocesseurs et puissance (d'après F. Anceau)

Intérêt d'un multiprocesseur

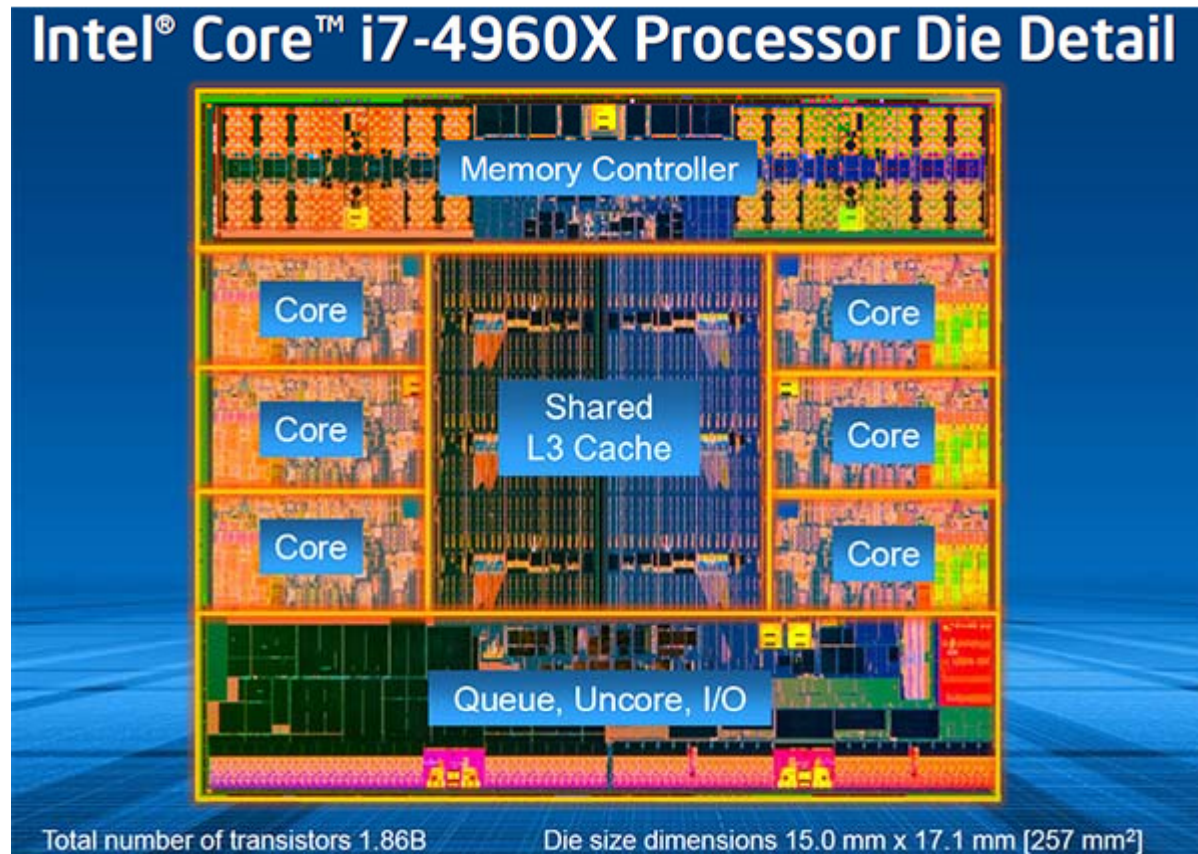
- La puissance thermique dégagée par un processeur dépend **exponentiellement** de sa puissance de calcul (à technologie constante!).
- Celle d'un multiprocesseur varie linéairement.
- Il est thermiquement plus économique de gagner de la puissance de calcul par des configurations **multiprocesseurs**



Multi-cœurs haut de gamme

Ivy-bridge

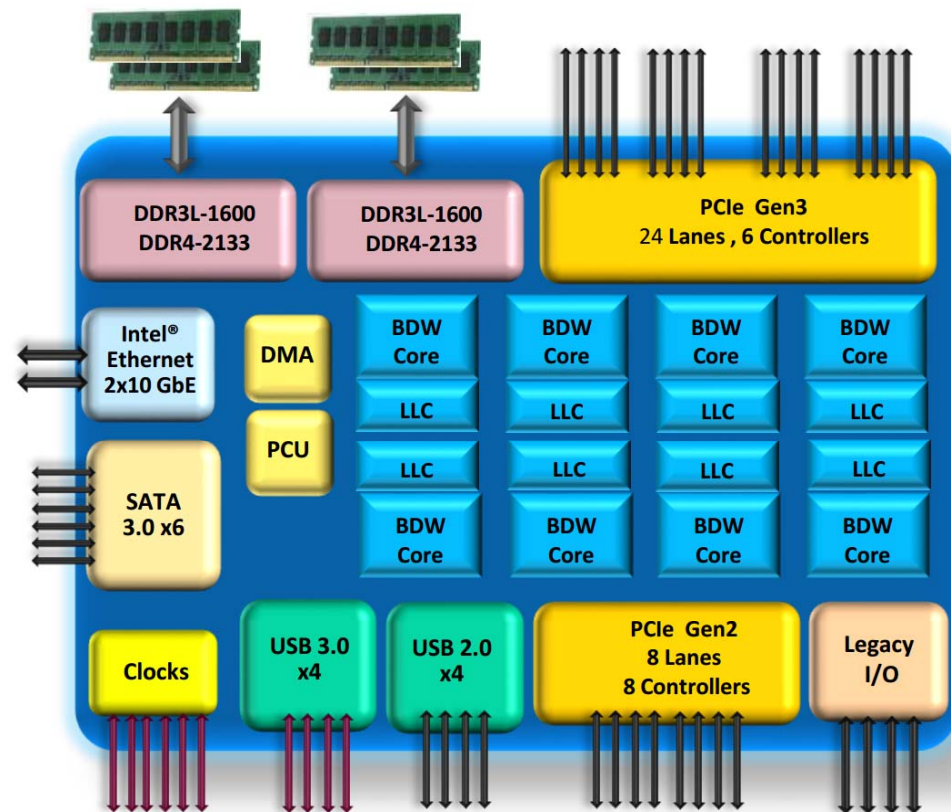
2013
22 nm
1,86 10⁹ T
2,57 cm²



Et le « *petit* » dernier ...

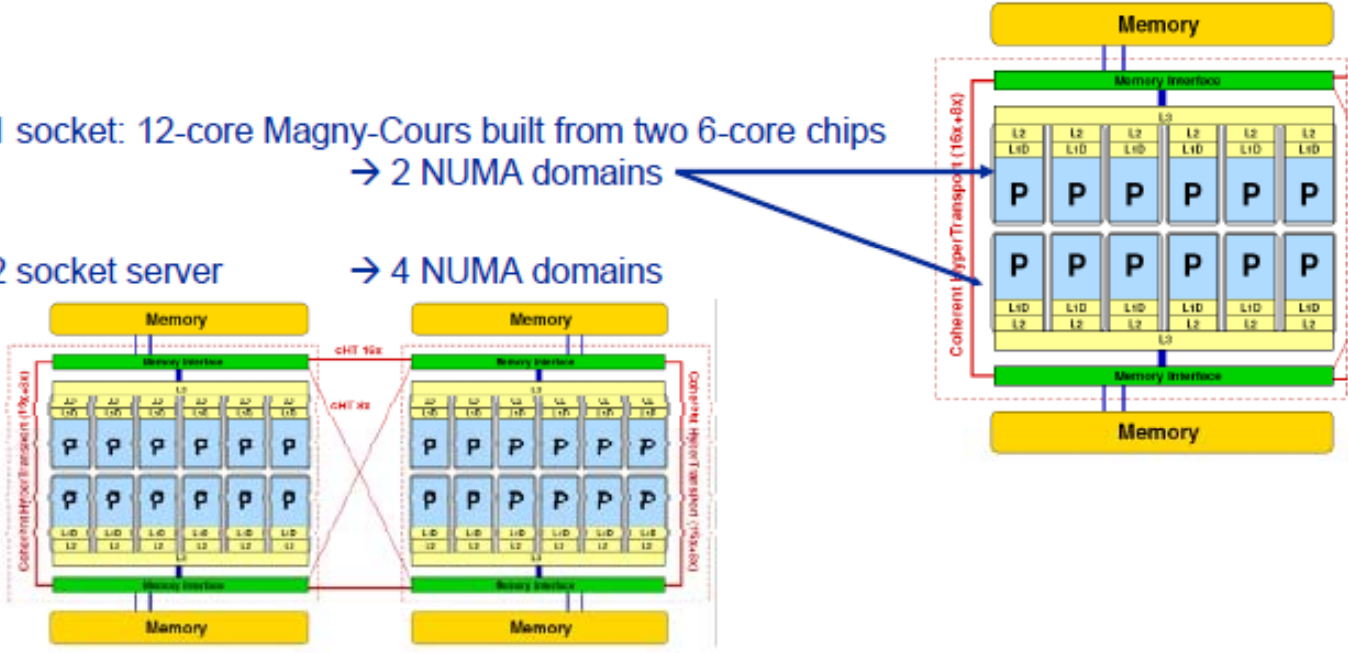
Intel® Xeon® Processor D - SoC Architecture

- 14 nm
- D1540
 - 8 cœurs
 - 15 Mo caches L3
 - 2 GHz
 - **45 W**



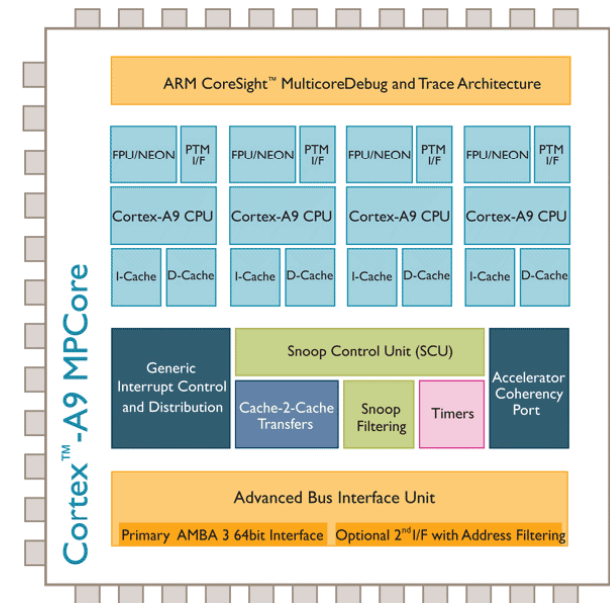
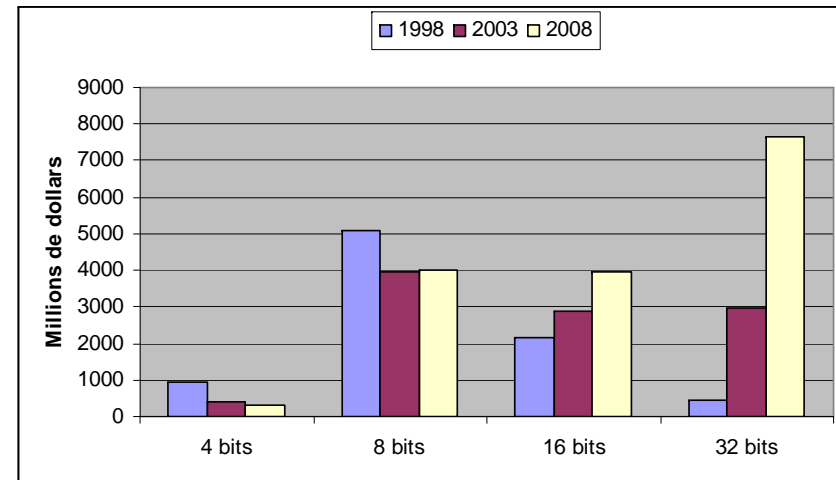
Cc-NUMA (exemple AMD)

- 1 socket: 12-core Magny-Cours built from two 6-core chips
→ 2 NUMA domains
- 2 socket server → 4 NUMA domains



Embarqué : des processeurs très différents

- Microcontrôleurs 8 bits
 - 8051 !
 - Des processeurs 32 bits
 - Des multi-cœurs
 - Des SoC
-
- Contraintes
 - Taille du code
 - Consommation
 - Surface de puce



Intel Atom x3 c3440

- CPU 4 cœurs
- Modem 4G
- Vidéo

- WiFi
- Bluetooth
- Radio FM



Kalray MPPA-256

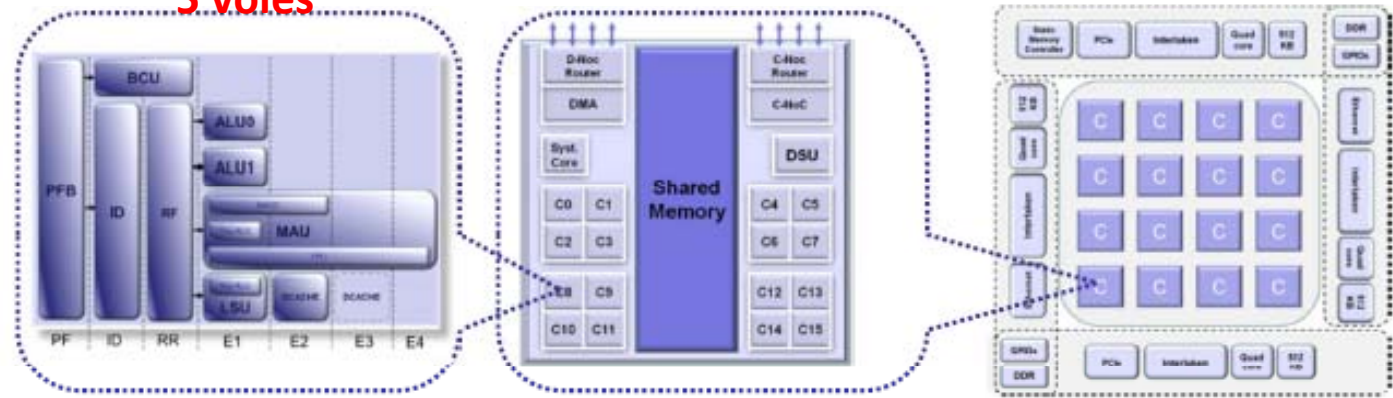
400 MHz

VLIW Core

5 voies

Compute Cluster

Manycore Processor



Instruction Level Parallelism

Thread Level Parallelism

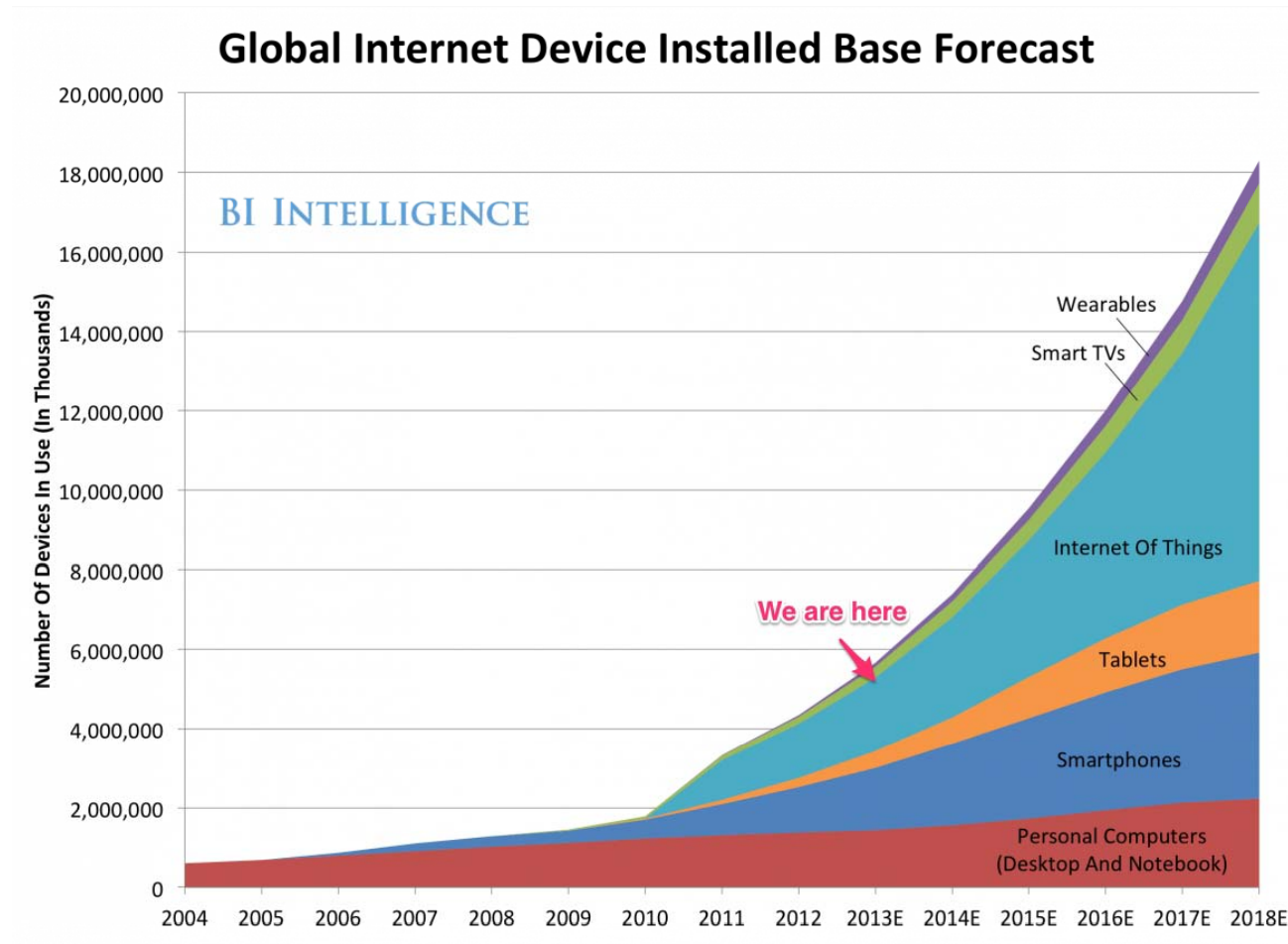
Process Level Parallelism

28 nm	Cores	GFLOPS (SP)	Active Power	Real Time	DDR	Ethernet
Kalray MPPA	288 K1	230	10W	Yes	2 DDR3 1600	8 10G

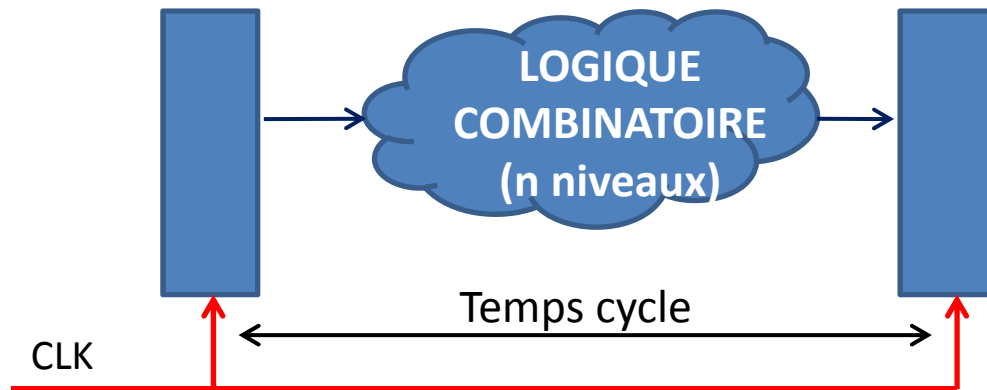
Monte-Carlo Option Pricing

Accelerator	Time (s)	Performance	Energy (J)
i7-3820	13.86	0.17	1802.2
Tesla C2075	2.37	1.00	531.7
MPPA-256	5.75	0.41	86.3

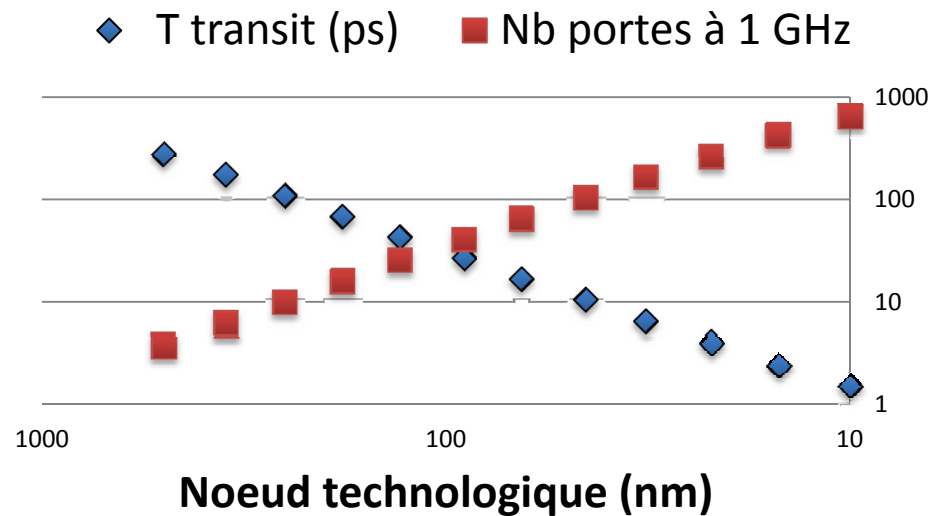
Modification des « moteurs »



La logique synchrone



- Temps de cycle fixé
 - Mur de la chaleur
- Avec les nœuds technologiques successifs



- **Diminution possible du nombre d'étages de pipelines**

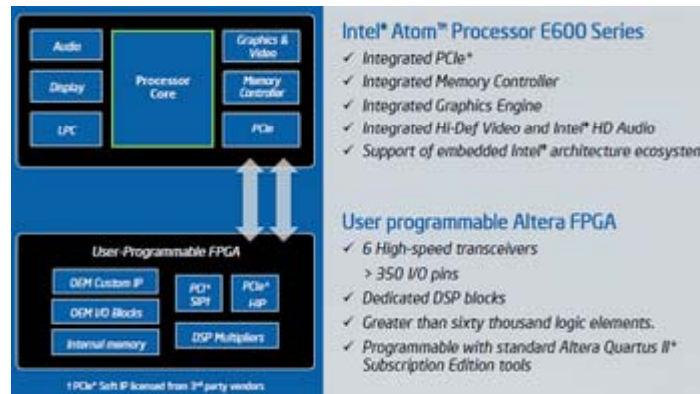
Des pipelines plus courts ?

- Moins d'étages de pipeline
 - Structures matérielles plus simples
 - Prédiction de branchement moins importante
 - **Applicables à tous les types de processeurs**
- Processeurs sans pipelines d'exécution ??????
 - Suppression des aléas (structurel, données, branchement)
 - Moins de registres
 - Contrôle plus simple
 - Consommation réduite
 - Simplification de la hiérarchie mémoire



Multi-cœurs et FPGA

- Réduction du différentiel de fréquences entre processeurs et FPGA
 - Altera : 1 GHz en 14 nm Intel



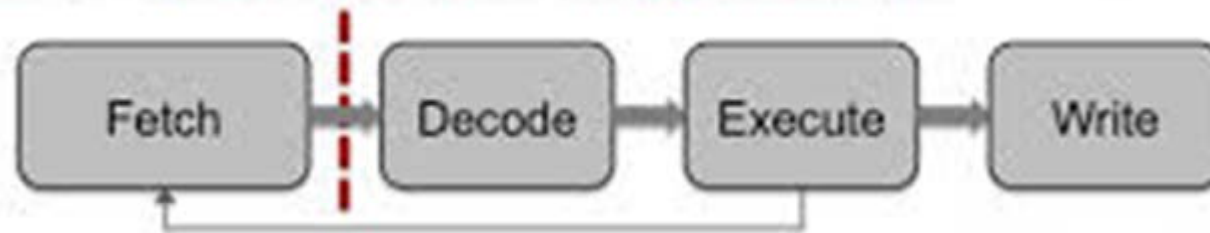
Intel s'offre les puces reprogrammables d'Altera pour 16,7 milliards de dollars

CPU BA20 (IP Cast)

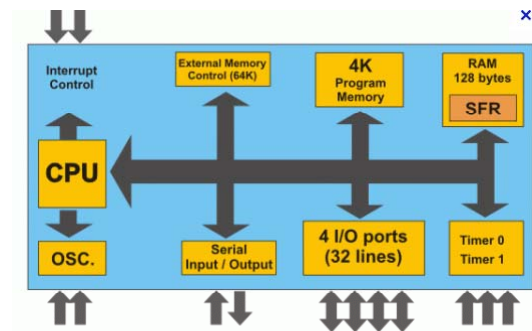
Re-invent the basic architecture

- What about those "old fashioned" non-pipelined CPUs?

BA20 PipelineZero Approach



VERSION MODERNE DU 8051

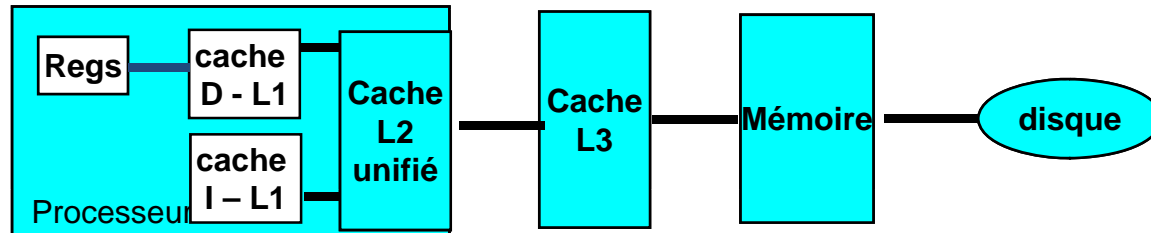


Fréquence – Etages de pipeline

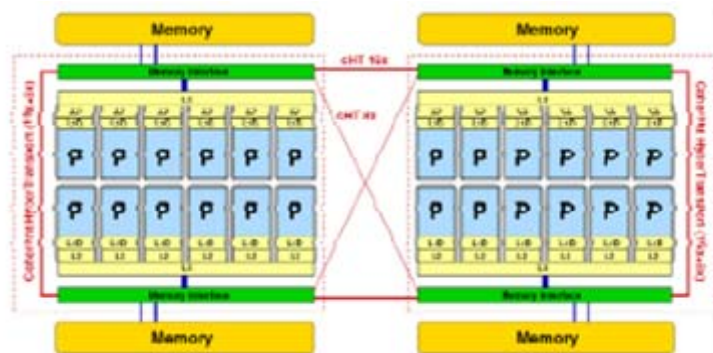
Comparaison de différents processeurs scalaires
(jeu d'instructions BA2)

Caractéristiques	BA20	BA21	BA22 (CE/AP)	BA25
NB étages	1	2	5	7/12
CoreMarks/MHz	3,41	2,77	2,93	2,51
FMAX @TMS65LP	75 MHz	150 MHz	400 MHz	800 MHz
CoreMarks	256	415	1175	2000
Portes équivalentes	> 10 K	> 10 K	>25 K >35 K	> 150 K
Caches	-	-	L0 /L0	L0 et L1
MMU	-	-	-/L0	L0 et L1

Mur mémoire : NUMAS



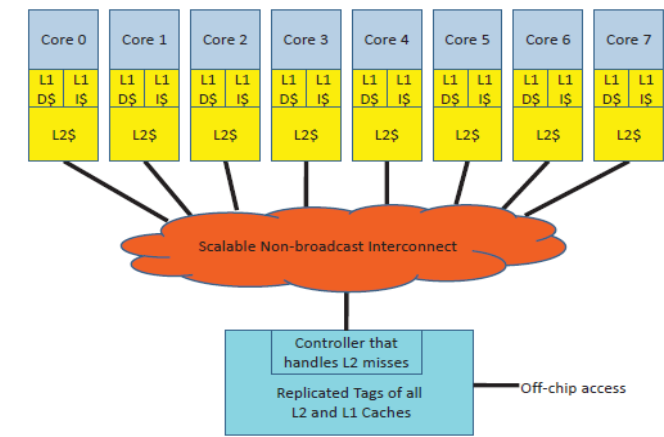
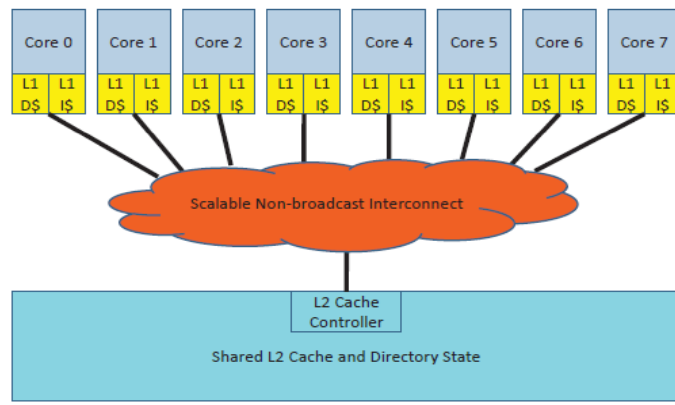
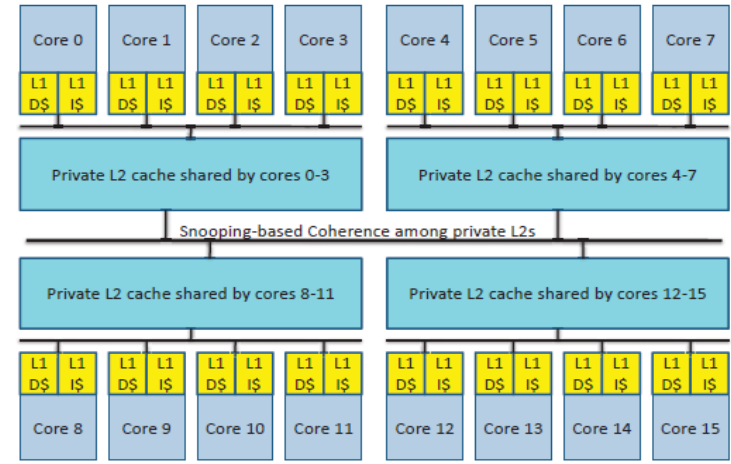
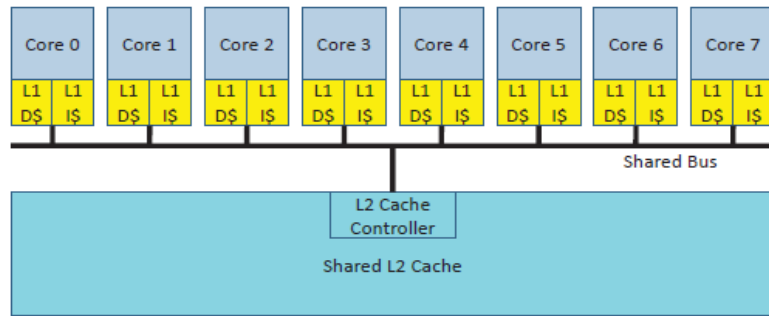
Hiérarchie de caches



Clusters

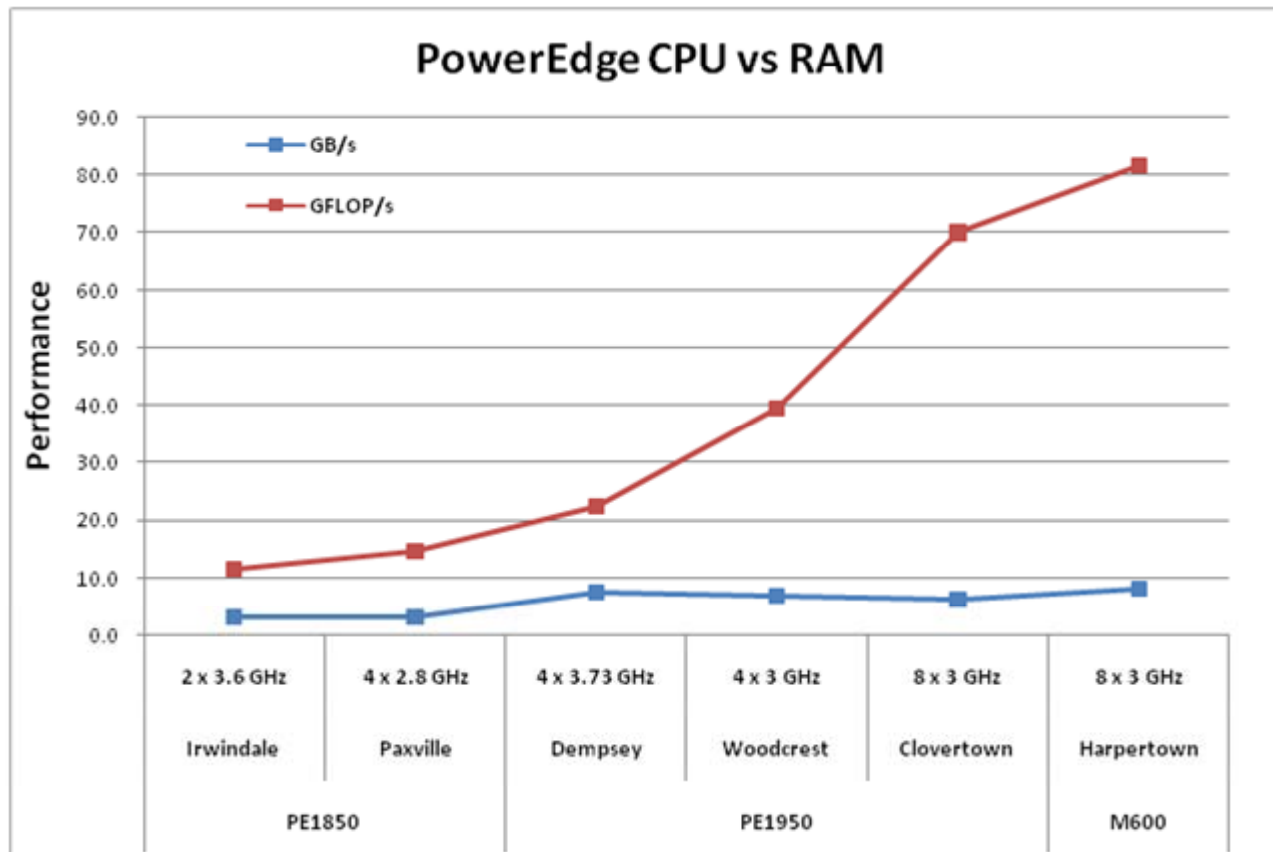
**STRUCTURATION DES DONNEES EN FONCTION DE LA HIERARCHIE MEMOIRE
MODELES DE PROGRAMMATION : Mémoire partagée / Mémoire distribuée**

Caches privés ou partagés



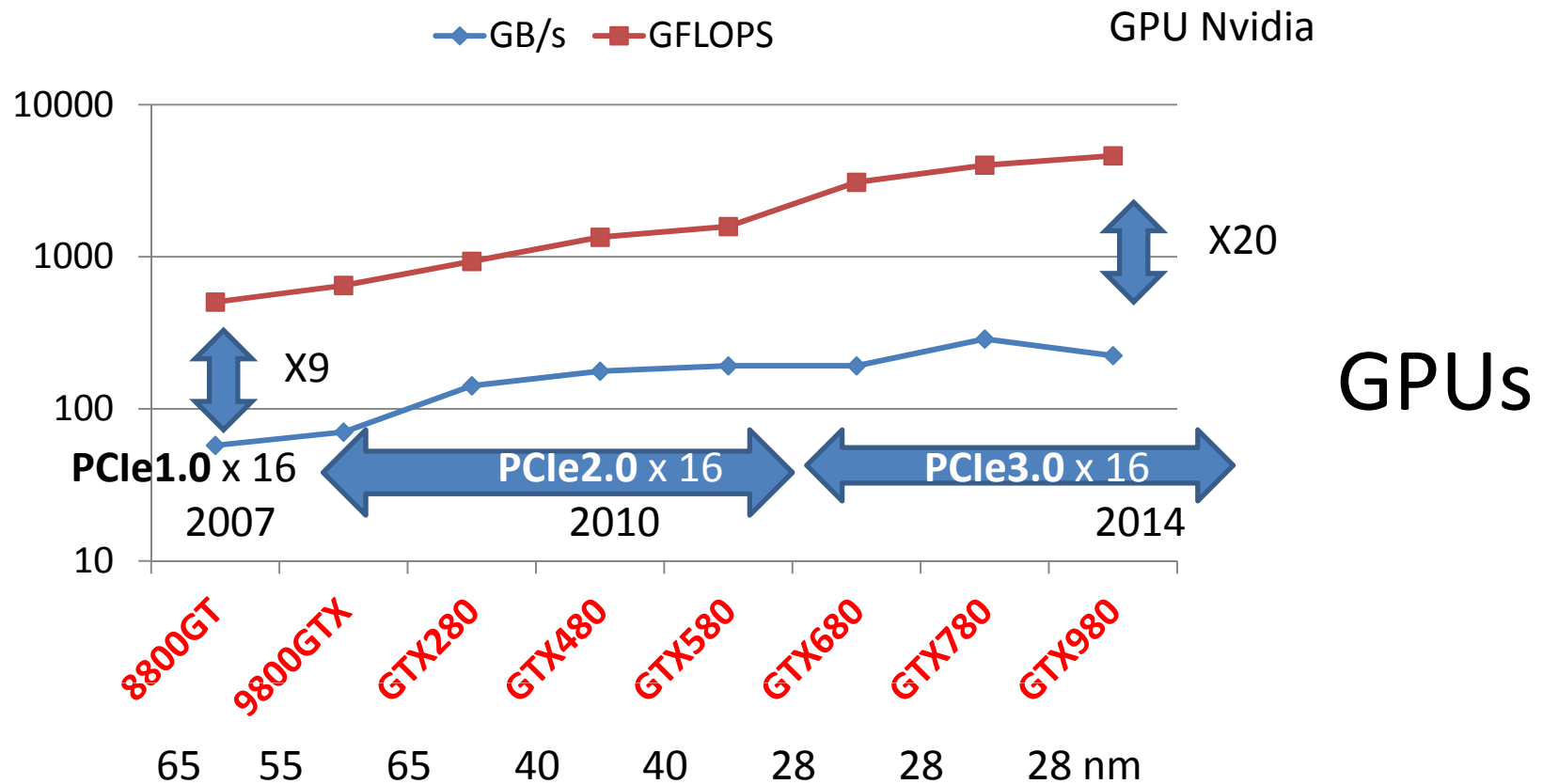
Complexification des problèmes de cohérence

Mur mémoire : débits

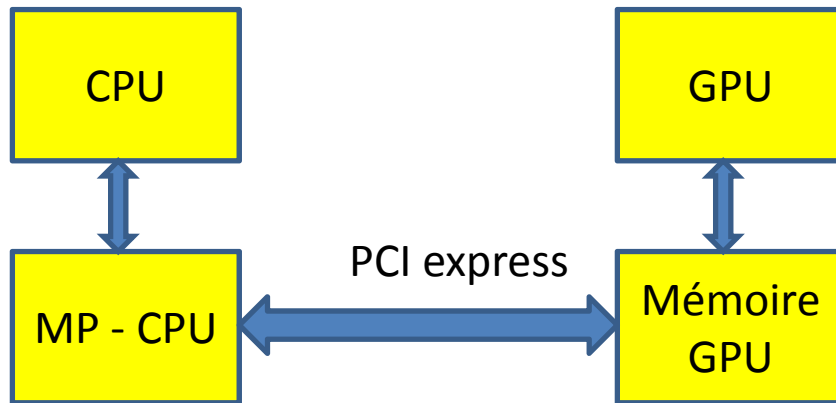


**SERVEURS
DELL**

Mur mémoire : débits



Mur mémoire : CPU et GPU



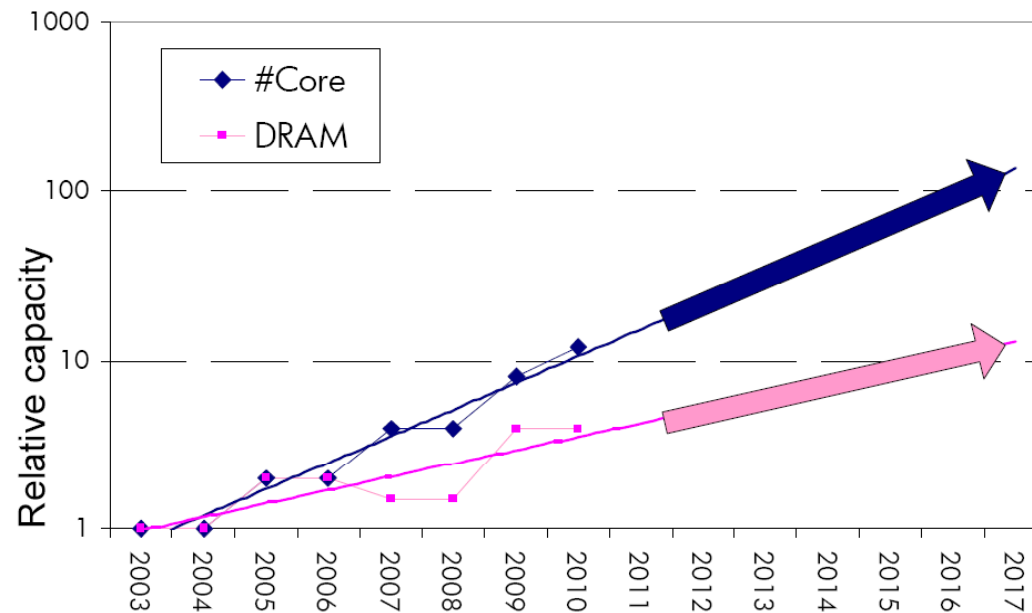
Eviter PCIe
- Puces soudées
- APU

- Débit maximum PCIe
 - PCIe 1.0 x 16 : 4 Go/s
 - PCIe 2.0 x 16 : 8 Go/s
 - PCIe 3.0 x 16 : 16 Go/s
 - PCIe 4.0 (2016)
- Contrôleur mémoire « pinned » du CPU

Example: The Memory Capacity Gap

Core count doubling ~ every 2 years

DRAM DIMM capacity doubling ~ every 3 years



Source: Lim et al., ISCA 2009.

- *Memory capacity per core* expected to drop by 30% every two years
- Trends worse for *memory bandwidth per core!*

Hétérogénéité : CPU + accélérateurs

- Superordinateurs
 - Multi-cœurs et GPUs
 - Multi-cœurs et Xeon-Phi
 - Multi-cœurs et FPGA
- MPSoC
 - CPU / Multi-cœurs/DSP/ FPGA

Et la programmation ?

- Clusters de multi-cœurs
 - Mémoire partagée (OpenMP – Pthreads)
 - Mémoire distribuée (MPI)
- + GPU
 - Cuda / OpenCL
- + DSP + FPGA + ...

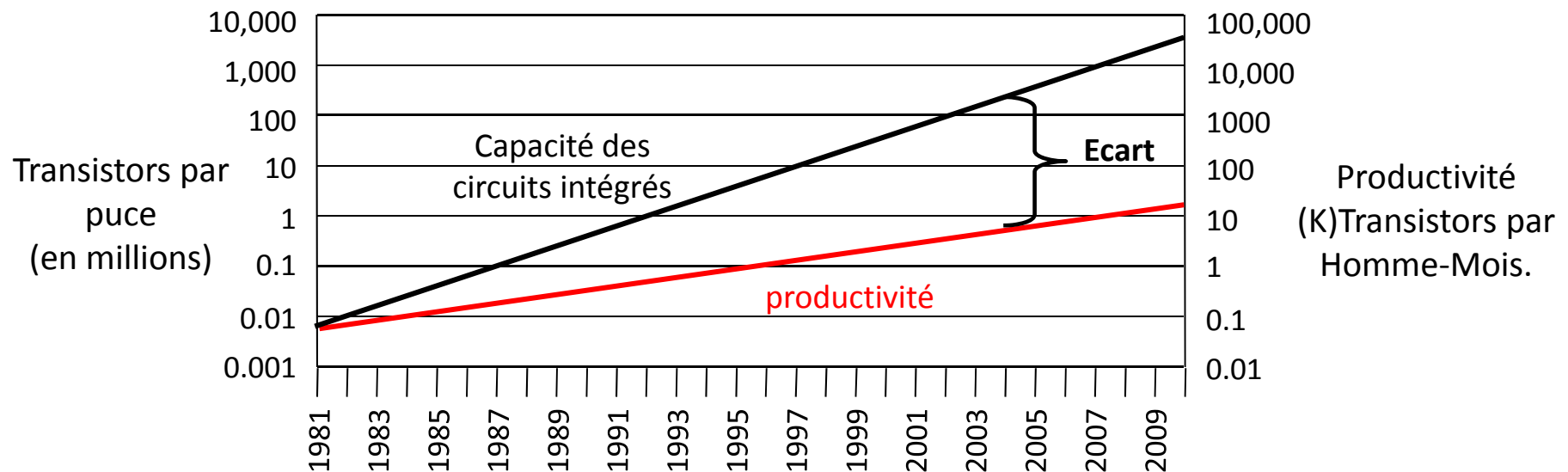
D. A. Paterson (IEEE Spectrum – Juin 2010)

- ***“The Trouble With Multicore - Chipmakers are busy designing microprocessors that most programmers can't handle”***

Un autre différentiel

- Évolution comparée du temps de conception d'un circuit et du nombre de portes disponibles

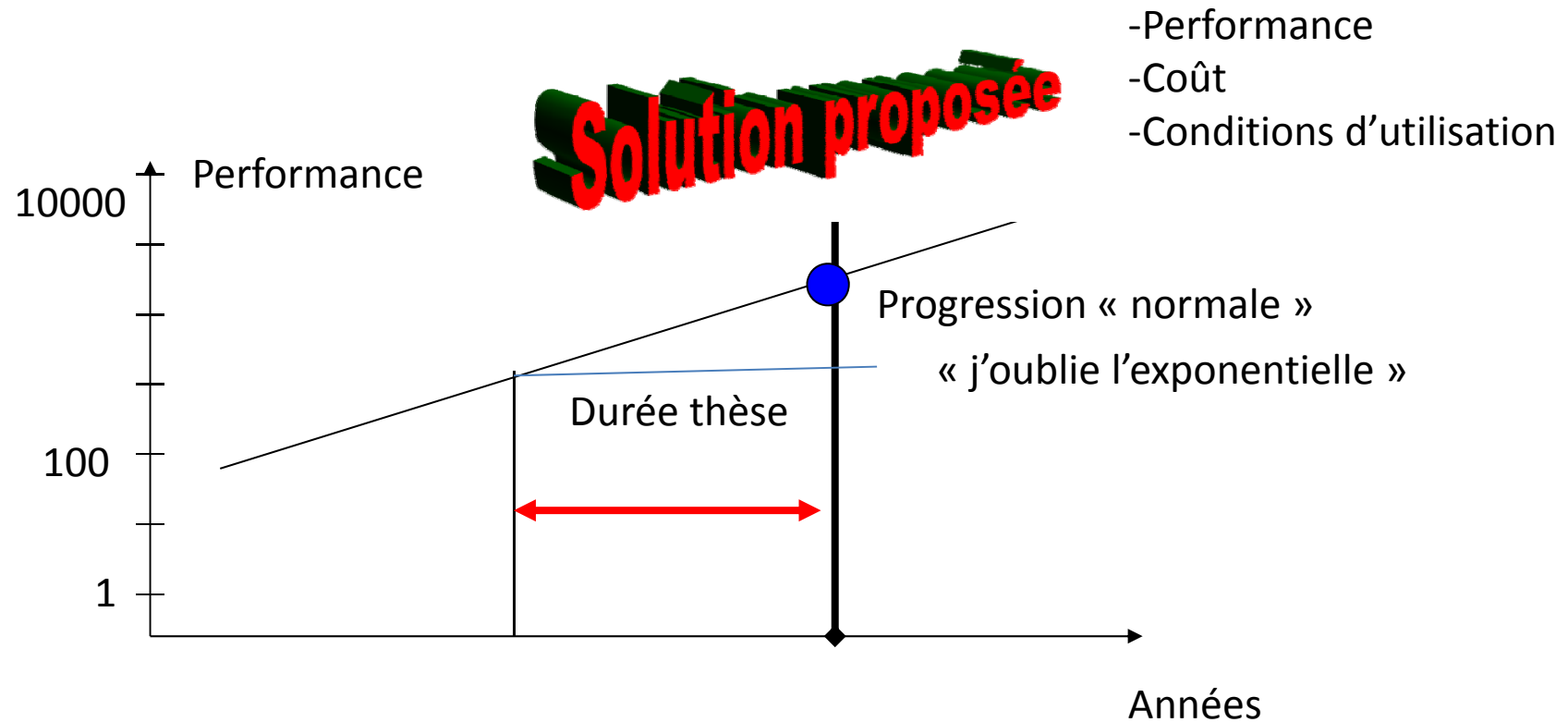
Écart de productivité de conception



Conclusions (1)

- Exponentielles : propriété n°1 de l'architecture des processeurs / ordinateurs
- Les exponentielles provoquent des murs, infranchissables ou à contourner
- Complexification
 - Hétérogénéité matérielle
 - Les accélérateurs disparaissent (on-chip) et réapparaissent plus nombreux (GPU, FPGA...)
 - Complexification de la hiérarchie mémoire
 - Hétérogénéité logicielle
 - Modèles de programmation différents

Conclusion (2) : exponentielle et... thèse



- Le référentiel (état de l'art) évolue exponentiellement.

Recherche en archi : masochisme ?

- Pourquoi avoir choisi info/archi ?
 - Référentiel à évolution rapide ou **exponentielle**
 - **Architecture des ordinateurs**
 - **Données massives**
- Et non un domaine à référentiel fixe ou à évolution très lente ?
 - Science de la terre
 - Physique - Chimie
 - Biologie (végétale – animale)
 - Recherche médicale
 - Ou... Mathématiques...

Dernier théorème de Fermat

- Formulé par Fermat vers 1630

$$x^n + y^n = z^n$$

n'a pas de solutions entières non nulles quand $n > 2$.

- Démontré par Andrew Wiles en 1995

– Plus de 350 ans plus tard

- ***Problèmes et... conditions de recherche sont immuables dans le temps***



Questions ?

- En évitant, dans la mesure du possible, les questions sur le futur ...! *{joke}*

Exécution o-o-o = recherches associatives

Exemple : « ROB » dans Metaflow (DRISS)

	Op source 1			Op source 2			Destination			
N°	V	N° reg	ID	V	N° reg	ID	D ?	N° reg	Résultat	
ID1	V	R2		V	R1		D ?	R1		DIV R1,R2,R1
ID2	V	R7			R1	ID1	D ?	R3		ADD R3,R7,R1
ID3	V	R8		V	R2		D ?	R1		SUB R1,R8,R2
ID4		R3	ID2		R1	ID3	D ?	R4		MUL R4,R3,R1
ID5	V	R6		V	R5		D ?	R4		XOR R4,R6,R5