



HAL
open science

Conception pour la faible consommation

Alain Guyot

► **To cite this version:**

Alain Guyot. Conception pour la faible consommation. DEA. Ce cours est enseigné de DEA de Microélectronique., 2006. cel-00092966

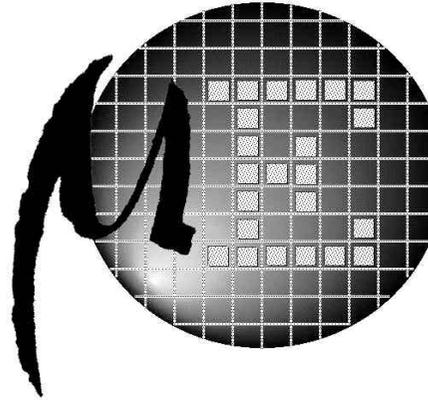
HAL Id: cel-00092966

<https://cel.hal.science/cel-00092966v1>

Submitted on 12 Sep 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Conception pour la Faible Consommation

Alain GUYOT
Sélim ABOU-SAMRA

Email: Alain.Guyot@imag.fr

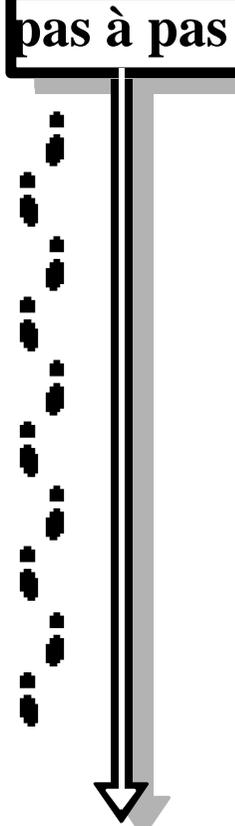
Tel: 04 76 57 46 16

<http://tima-cmp.imag.fr/~guyot/Cours/Basseconso/>



Plan du cours

pas à pas

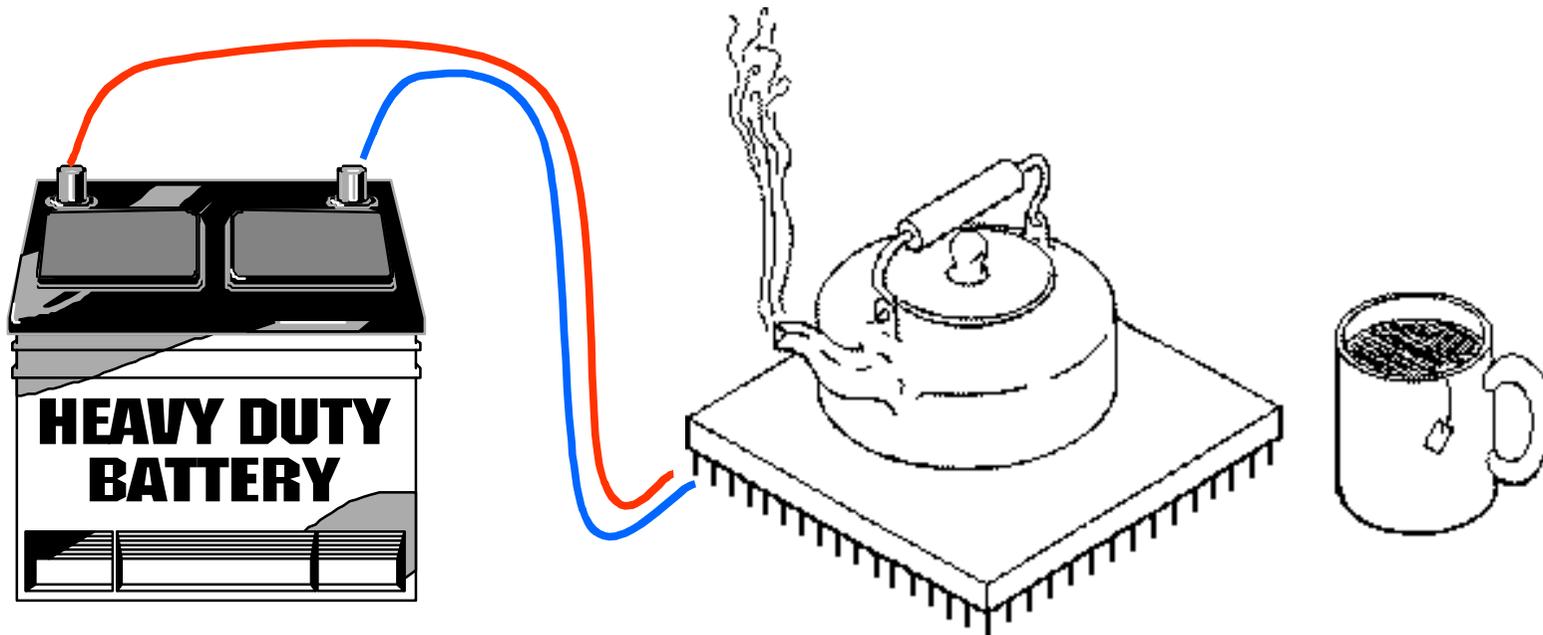


- mécanismes de dissipation de puissance
 - contributions à la dissipation
 - effet de l'alimentation V_{dd}
 - effet du seuil V_t
 - effet des fuites
- comment réduire la dissipation
 - jouer sur les capacités
 - jouer sur la tension d'alimentation
 - jouer sur la fréquence/ le délai
 - jouer sur l'activité
- mesure statistique de l'activité moyenne
- activité redondante
- adaptation dynamique des paramètres
- comment prédire la consommation
- codages pour réduire la consommation
- systèmes à récupération d'énergie
- technologie basse consommation
- conclusion



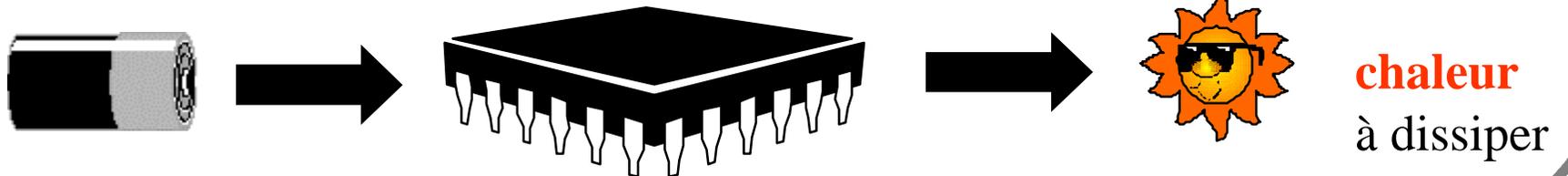
Introduction

- Critères traditionnels de la conception
 - Système sur une puce (complexité)
 - Délai (performance)
 - Surface de Silicium (coût)
 - Testabilité (coût)
- Soucis plus récent: faible consommation



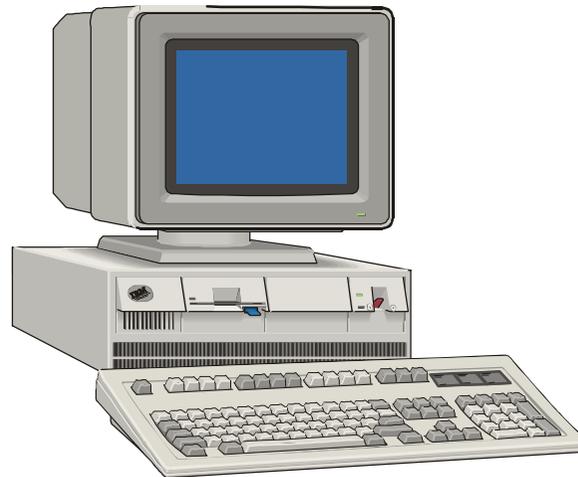
Introduction 2

- Pourquoi la faible consommation ? :
 - les circuits tombent en panne quand ils surchauffent (fiabilité)
 - les systèmes de refroidissement (ventilateur, radiateur) sont coûteux
 - les applications portables se développent (taille, poids et durée de la batterie)
 - les ordinateurs consomment 10% de l'énergie électrique totale (écologie)
 - un boîtier 2-watt (céramique) est 4 fois plus cher qu'un boîtier 1-watt (plastique)
- Pourquoi la dissipation augmente ?
 - plus de transistors/mm²
 - fréquence d'horloge plus élevée
 - capacités parasites par unité de surface plus élevés
 - courants de fuite plus élevés



Traiter de l'information dissipe de l'énergie

	Acide-plomb	NiCd	NiMH	Li-ion	Li-métal
Tension	2V	1,2V	1,2V	3,6V	3V
Wh/Kg	35	50	90	105	140
Wh/l	75	150	230	300	300
atodécharge	5%/mois	20%/mois	30%/mois	8%/mois	2%/mois
effet mémoire	grand	grand	faible	sans	sans
cycles	500	900	1000	1200	1200



Où chercher à réduire la dissipation

Systeme

Contrôle d'activité, Partitionnement, "Power down",
"Clock gating"

Algorithme

Complexité, Concurrence, Localité,
Régularité, Représentation des données

Architecture

Diminution de la tension V_{dd} , Parallélisme,
Corrélation spatio/temporelle de signaux

Circuit

Ajustement de la taille des transistor,
Optimisation logique et réordonancement,
Faible activité, logique à faible excursion,
Commutation adiabatique

Technologie

Réduction dimensions, réduction seuil V_t ,
seuil auto-ajustable, multi-seuil, SOI



Pourquoi délai et dissipation

Les circuits dissipent et ont du délai pour les mêmes raisons:

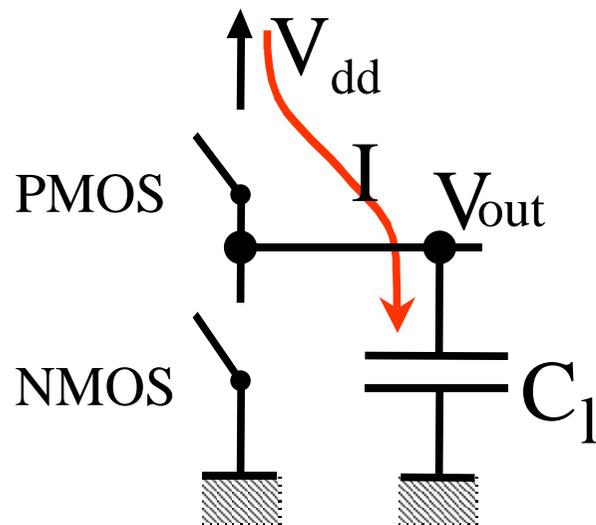
- les variables booléennes sont représentées par des tensions ("0" \equiv 0V, "1" \equiv V_{dd})
 - les circuits ont des capacités (parasites ou utiles)
- \Rightarrow il faut donc des courants pour charger ou décharger les capacités.



Energie dans les capacités parasites

effet du chargement d'une capacité C_1 :

porte CMOS

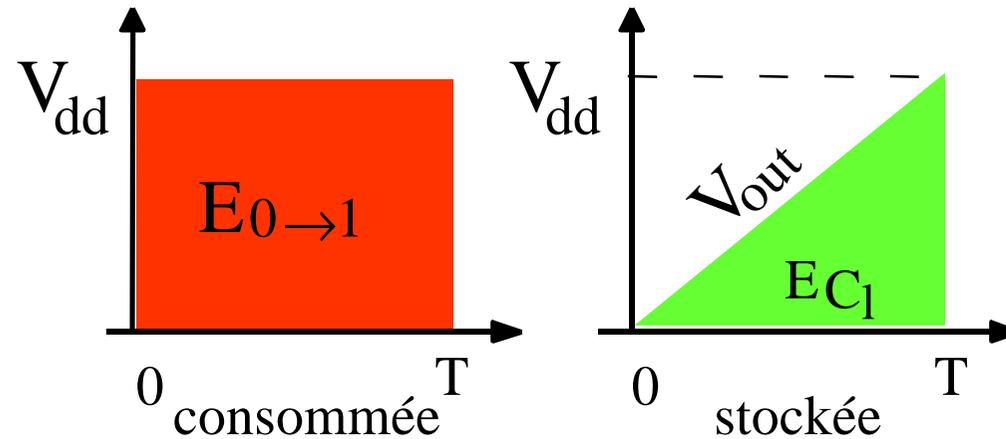


énergie totale tirée de V_{dd}

$$E_{0 \rightarrow 1} = \int V_{out} * I * dt = \int_0^{C_1 * V_{dd}} dq = C_1 * V_{dd}^2$$



Energies stockée et dissipée



énergie stockée dans C_1

$$E_C = \int V_{out} * I * dt = \int V_{out} * \frac{dq}{dt} * dt = \int V_{out} * dq$$

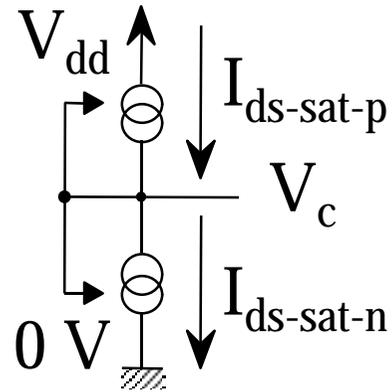
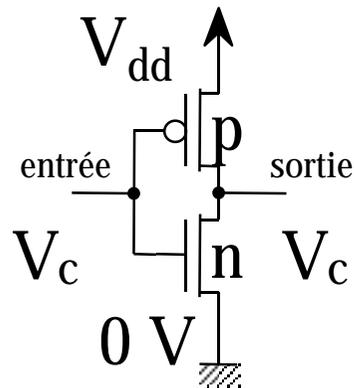
$$E_C = \int_0^{V_{dd}} V_{out} * C_1 * dV_{out} = \frac{1}{2} * C_1 * V_{dd}^2$$

énergie dissipée en chaleur $E_W = E_T - E_S = \frac{1}{2} * C_1 * V_{dd}^2$



Seuil de commutation de l'Inverseur

porte CMOS



Analyse en continue

$$\frac{I_{ds-sat-n}}{I_{ds-sat-p}} = 1$$

$$K = \frac{\mu\epsilon}{2t_{ox}}$$

$$\alpha = \sqrt{\frac{K_n}{K_p} * \frac{W_n/L_n}{W_p/L_p}}$$

$$\text{Seuil de commutation } V_c = \frac{V_{dd} - \alpha V_{tn} - |V_{tp}|}{1 + \alpha}$$



Courant de court circuit de l'Inverseur

$$\alpha = \sqrt{\frac{K_n}{K_p} * \frac{W_n}{W_p} * \frac{L_p}{L_n}}$$

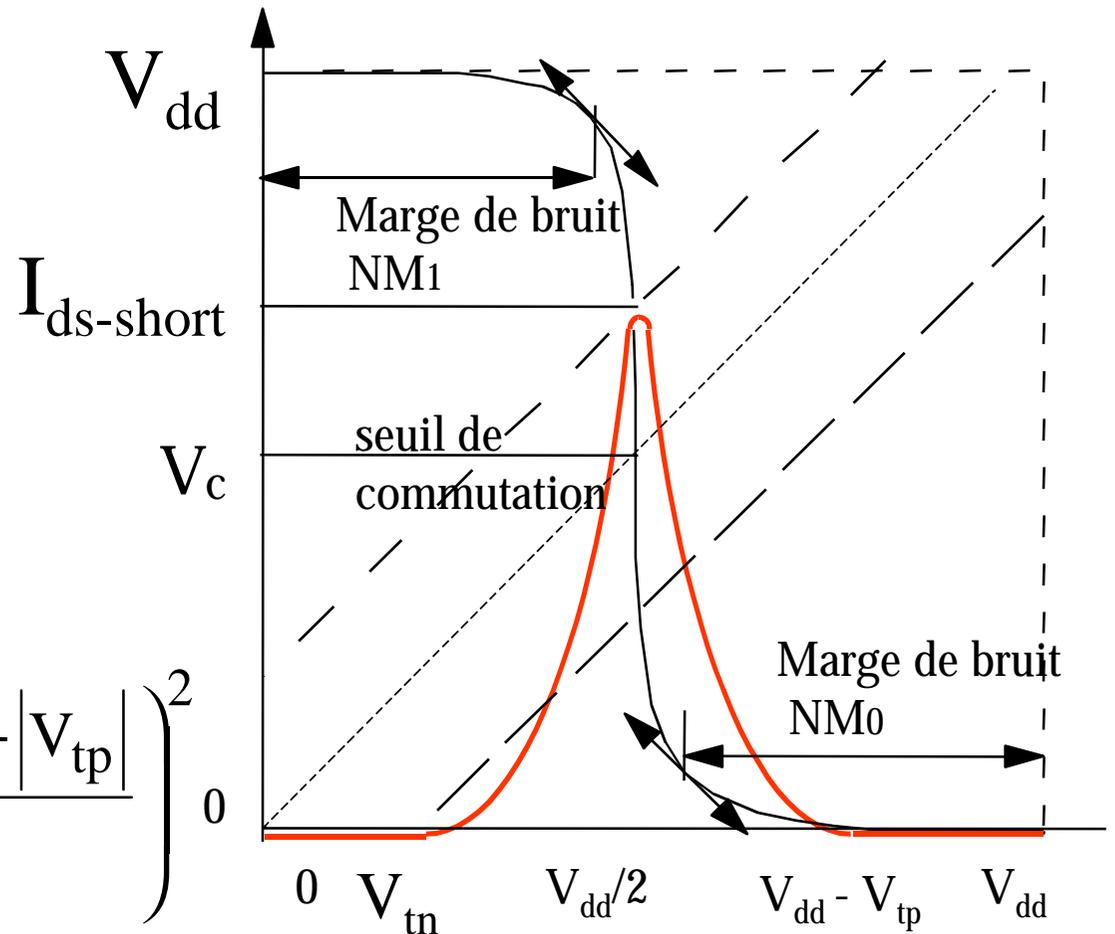
$$\frac{I_{ds-sat-n}}{I_{ds-sat-p}} = 1$$



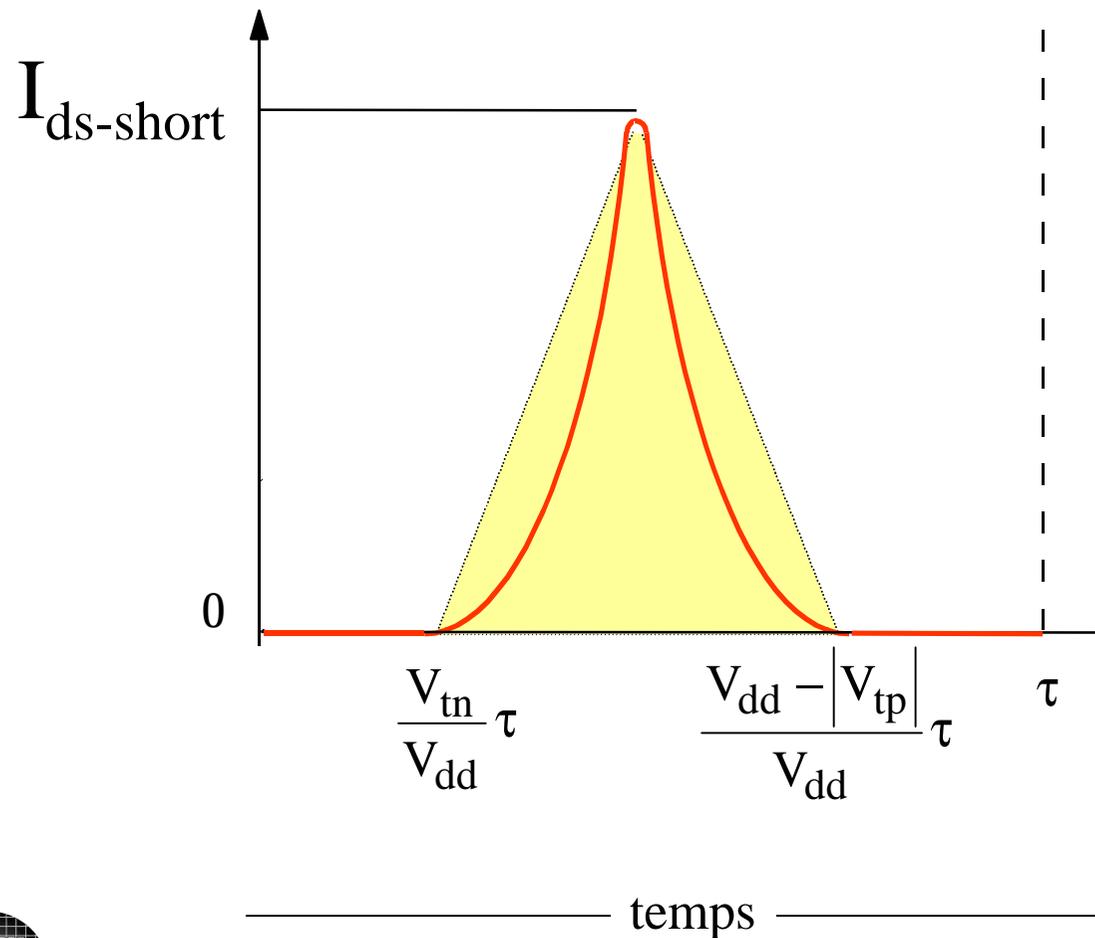
$$I_{ds} = K \frac{W}{L} (V_{gs} - V_t)^2$$

$$I_{ds-short} = K_n \frac{W_n}{L_n} \left(\frac{V_{dd} - V_{tn} - |V_{tp}|}{1 + \dots} \right)^2$$

analyse en continue



Energie de court circuit de l'Inverseur



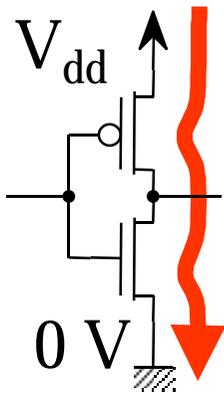
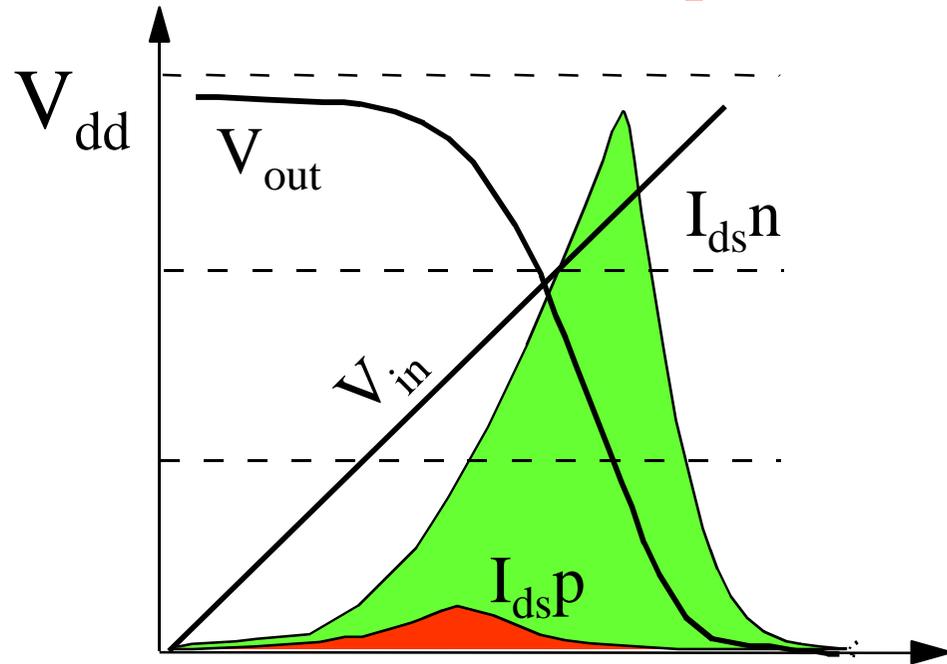
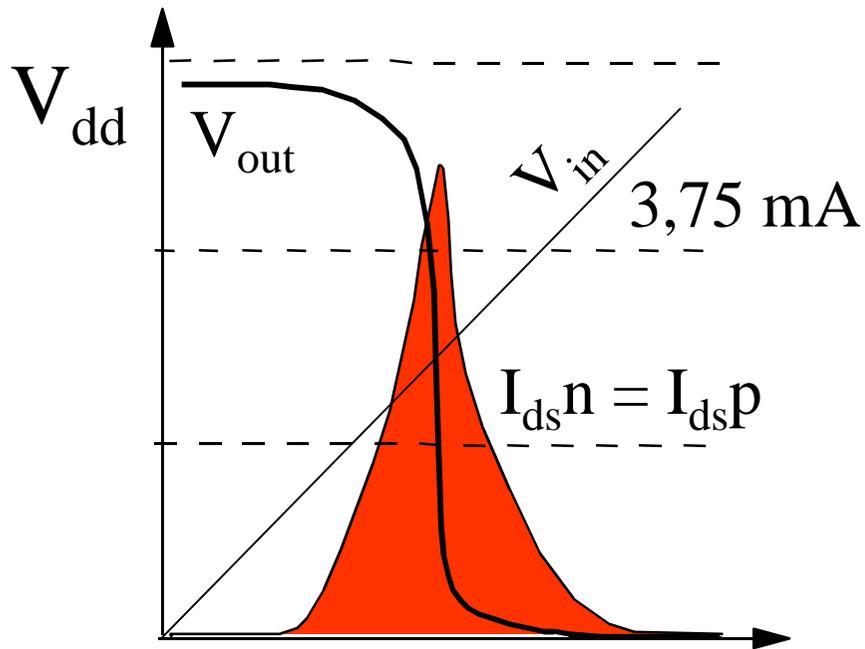
Une bonne approximation de l'énergie de court-circuit est donnée par la surface du triangle.

$$E_{cc} \approx \frac{\tau * K * \frac{W}{L}}{2} (V_{dd} - 2 * V_t)^3$$

- τ est le temps de monté
- K est la technologie
- W/L la taille

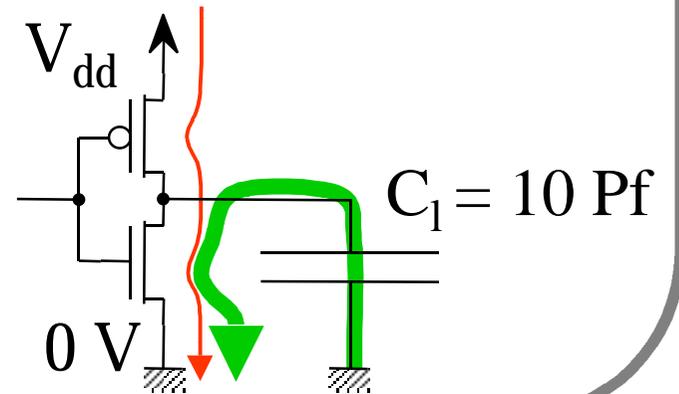


Effet de la charge de sortie C_1



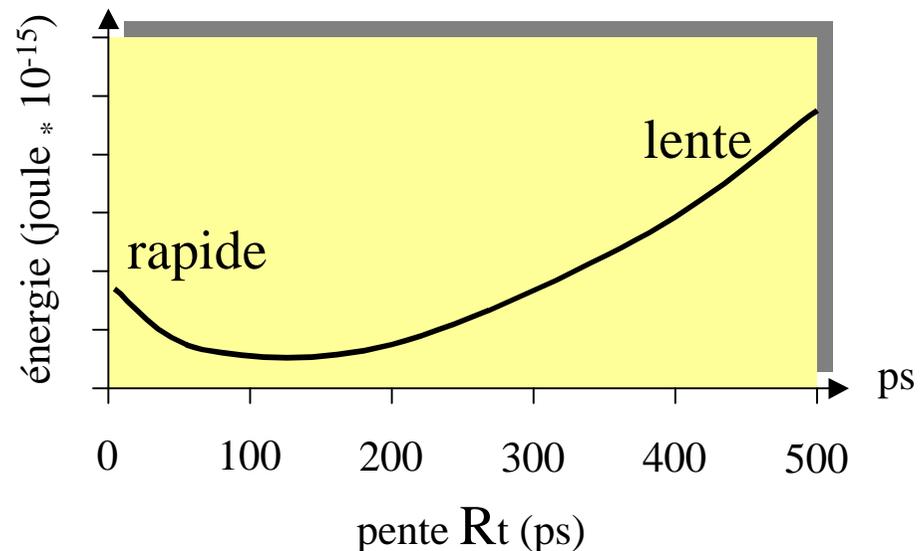
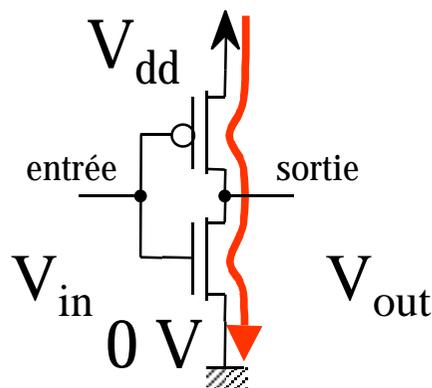
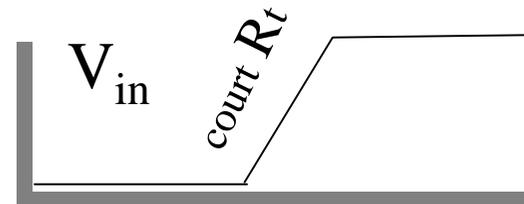
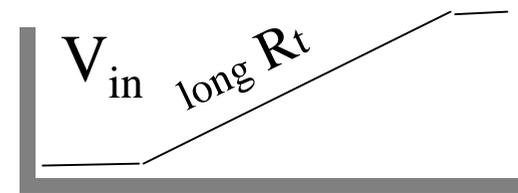
Surface rouge :
énergie perdue (court circuit)

Surface verte :
énergie stockée (capacité)



Effet de la pente sur le court circuit

- entrée lente (faible pente)
 - durée du court-circuit longue
 - transistors petits (faible capacité)
- entrée rapide (forte pente)
 - fort courant de sortie
 - transistors doivent être gros



Composantes de la dissipation en CMOS

puissance = énergie/temps

puissance = fréquence * activité * énergie par transition logique + dissipation statique

$$\begin{aligned}
 & \text{fréquence (MHz)} \\
 & \text{activité } (0 \leq A \leq 1) \\
 \text{puissance} &= F * A * \left(\underbrace{(C_1 * V_{dd}^2)}_{\substack{\text{énergie dans capacités} \\ \text{(puissance dynamique)}}} + \underbrace{(I_{ds\text{-short}} * R_t * V_{dd})}_{\substack{\text{énergie de court-circuit} \\ \text{(puissance dynamique)}}} \right) + \underbrace{(I_{diodes} + I_{ds\text{-leak}}) * V_{dd}}_{\substack{\text{Fuite des diodes} \\ \text{courant sous seuil} \\ \text{(puissance statique)}}}
 \end{aligned}$$



Quelques ordres de grandeur

Approximations pour du CMOS 0,7 μm alimenté sous 5 V

- charge C_1 typique \approx 0,1 pF
- une charge \approx 8 million d'électrons
- énergie de charge pour une transition \approx 1 picojoule
- énergie de court-circuit pour une transition \approx 0,1 picojoule
- courant de fuite \approx 0,2 nA par porte

Quelques comparaisons d'énergie

- transition d'un neurone \approx 10^{-13} joule
- transition d'une porte CMOS \approx 10^{-12} joule
- transition d'une entrée/sortie \approx 10^{-10} joule
- instruction de microprocesseur \approx 10^{-7} joule
- pile de type AA (NiCd) \approx 10^3 joule
- canette de bière \approx 10^6 joule



Moyens de réduire la dissipation

$$\text{dissipation} = F * A * \left(C_1 * V_{dd}^2 + I_{ds\text{-short}} * R_t * V_{dd} \right) + \underbrace{(I_{diodes} + I_{ds\text{-leak}})}_{\text{Faible contribution}} * V_{dd}$$

The diagram shows arrows pointing from the variables in the equation to a list of strategies:

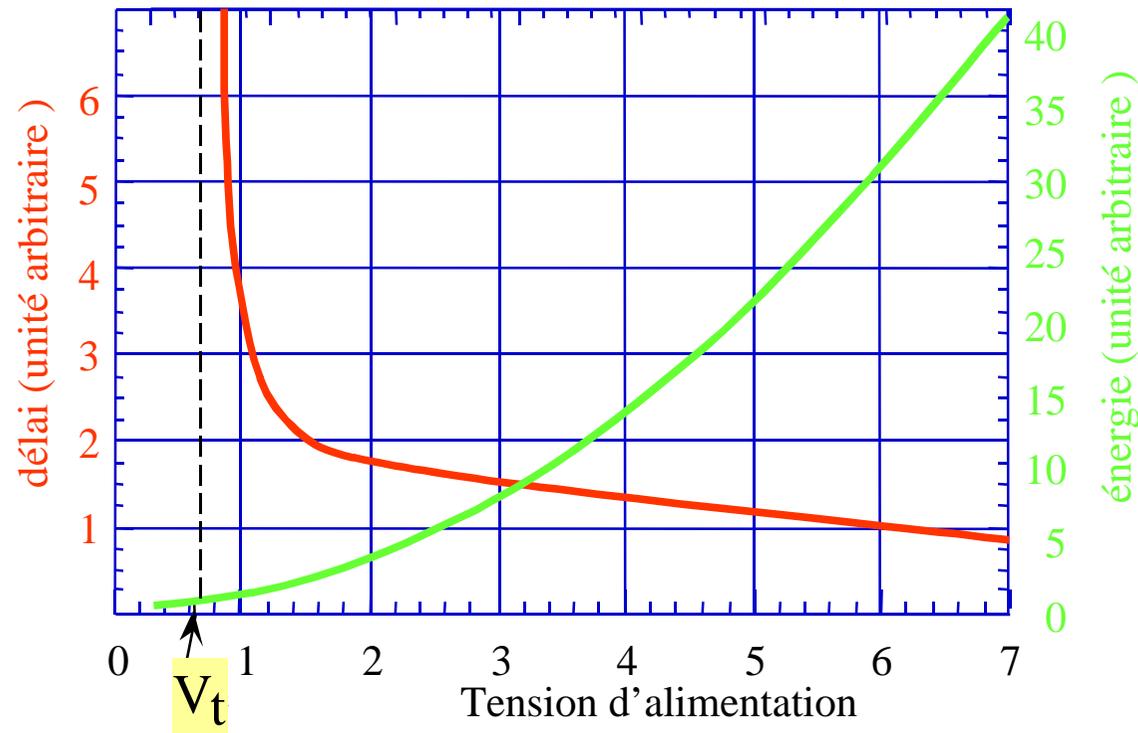
- réduire la fréquence est contre-productif
- réduire l'activité redondante (excellent)
- réduire la capacité parasite C_1
- réduire le V_{dd} (excellent, inconvénients)
 - réduit la dissipation quadratiquement
 - augmente linéairement le délai
- augmenter V_{tn} & V_{tp} augmente le délai
- pas de pente lente (R_t)



Réduction de l'alimentation V_{dd}

$$\text{délai} \approx \frac{\text{charge} * \text{escursion}}{\text{courant}} \approx \frac{C_1}{K(W/L)} \frac{V_{dd}}{(V_{dd} - V_t)^2}$$

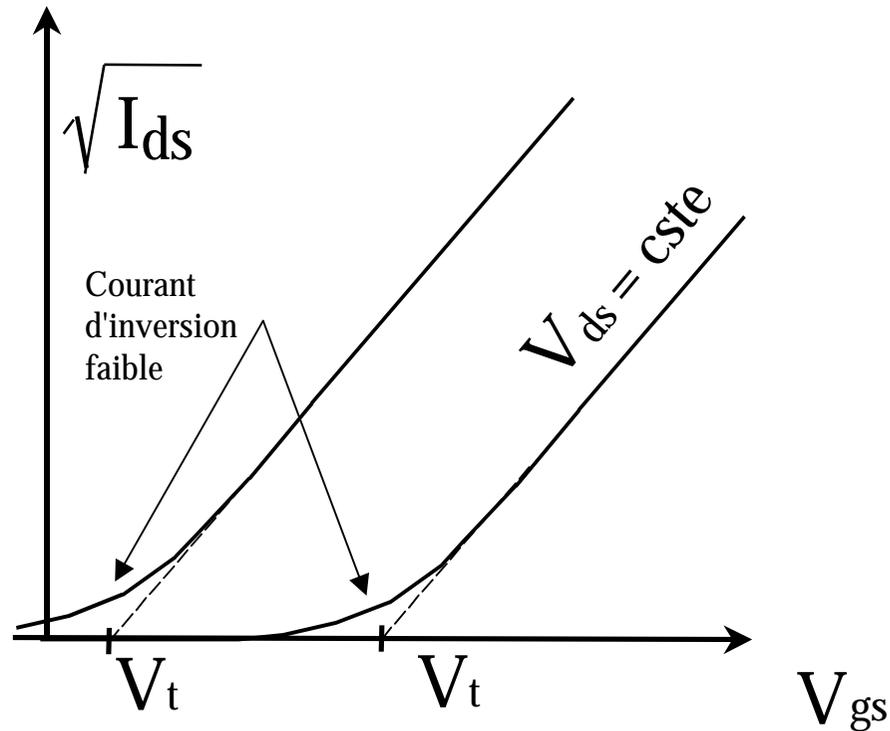
Le délai augmente comme $\frac{1}{V_{dd}}$



Réduire V_t proportionnellement à V_{dd} ($V_{dd} \approx 3 * V_t$)



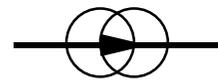
Réduction du seuil V_t



$$G_m = \frac{\partial I_{ds}}{\partial V_{gs}}$$

Transconductance
ou gain du transistor
(petit signal)

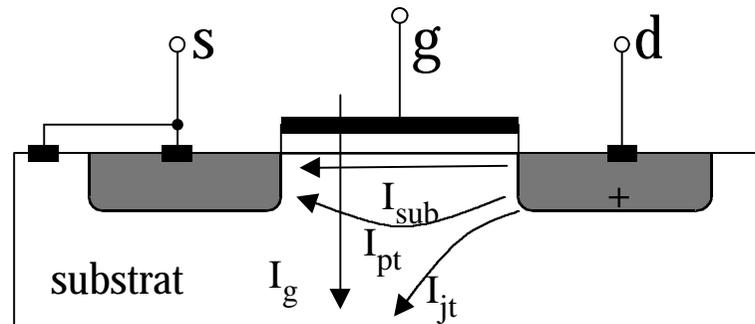
$$V_{gs} > V_t \Rightarrow I_{ds}$$


$$I_{ds} = K \frac{W}{L} (V_{gs} - V_t)^2$$



Courants de fuite

Remarque: Les courants de fuite sont la seule contribution lorsque le circuit est inactif (mode veille)



$$I_{sub} = K \frac{W}{L} V_t^2 \exp\left(\frac{V_{gs} - V_t + \eta V_{ds}}{\eta V_t}\right)$$



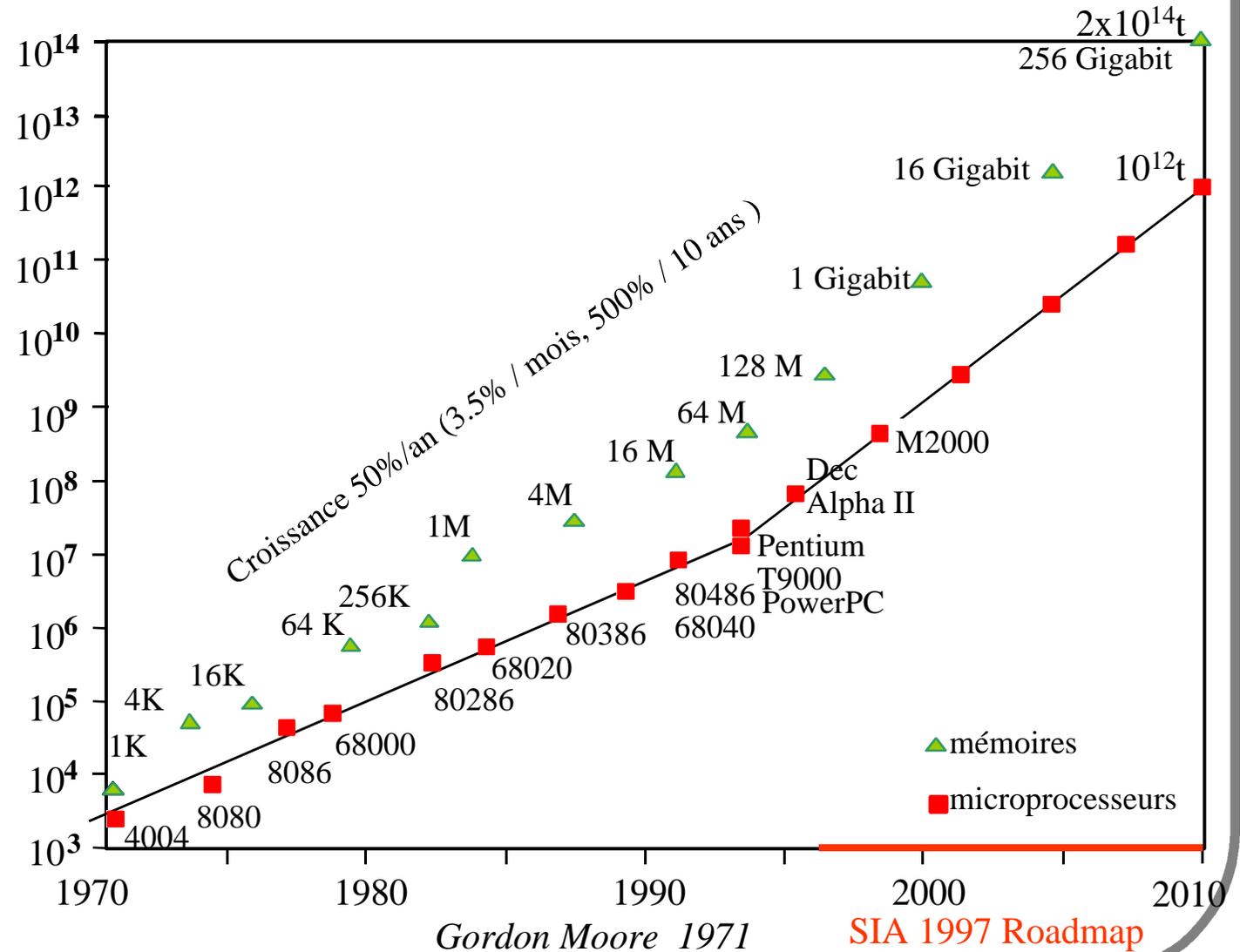
Evolution des technologies et dissipation

L'évolution du nombre de transistors par puce est dû au maigrissement des technologies.

Maigrir la technologie par un facteur $1/K$ avec:

- V_{dd} constant
- champs constant
- longueur seulement

Donne différents résultats



Evolution des technologies et dissipation

Paramètre	Modèle de réduction		
	champs constant	tension constante	longueur canal
Longueur (L)	1/K	1/K	1/K
Largeur (W)	1/K	1/K	1
Tension (V_{dd})	1/K	1	1
Epaisseur d'Oxyde (t_{ox})	1/K	1/K	1
Courant ($I=(W.V^2)/L$)	1/K	K	K
Transconductance (g_m)	1	K	K
Profondeur de jonction (X_j)	1/K	1/K	1
Dopage du substrat (N_A)	K	K	1
Champs électrique dans oxyde	1	K	1
Zone de déplétion (d)	1/K	1/K	1
Capacité parasite ($C = V.L/ t_{ox}$)	1/K	1/K	1/K
Délai ($V.C/I$)	1/K	$1/K^2$	$1/K^2$
Effet sur la puissance dissipée			
Puissance statique (P_s)	$1/K^2$	K	K
Puissance dynamique (P_d)	$1/K^2$	K	K
Energie Par Opération (EPO)	$1/K^3$	1/K	1/K
Surface (W.L)	$1/K^2$	$1/K^2$	1
Densité de puissance (V.I/A)	1	K^3	K^2
Densité de courant	K	K^3	K^2



Evolution des technologies et dissipation

	Modèle de réduction		
	champs constant	tension constante	longueur canal
Puissance statique (Ps)	$1/K^2$	K	K
Puissance dynamique (Pd)	$1/K^2$	K	K
Energie par Opération (EPO)	$1/K^3$	$1/K$	$1/K$
Surface A (W.L)	$1/K^2$	$1/K^2$	1
Densité de puissance (V.I/A)	1	K^3	K^2
Densité de courant	K	K^3	K^2

Réduction de dimension à tension V_{dd} constante
préférée pour le délai ($1/K^2$)
augmente la dissipation
augmente champs électrique (destruction)

Réduction de dimension à champs constant
réduction du délai $1/K$
la puissance réduite $1/K$ également



Réduction de la dissipation

$$\text{dissipation dynamique} = F * V_{dd} * V_{\text{escur}} * \left(\begin{array}{c} \text{nombre} \\ \text{de nœuds} \\ \sum_{i=1} A_i * C_{1i} \end{array} \right)$$

- Puissance de traitement = nombre d'opérations par seconde
- Parallélisme = nombre d'opérations par cycle
- Puissance de traitement = fréquence * parallélisme
- Fréquence = 1/délai

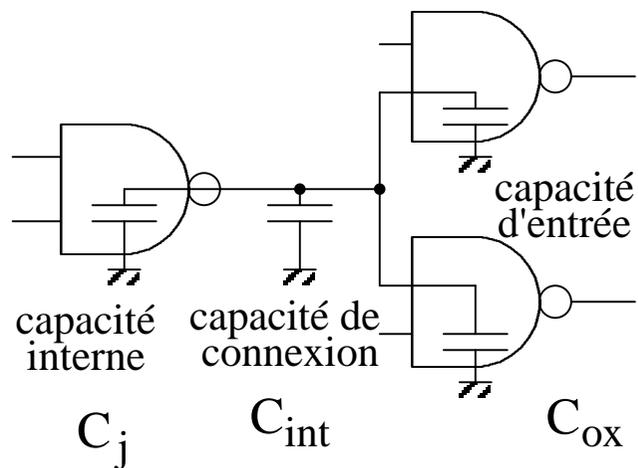
A puissance de traitement constante on peut jouer avec

- la fréquence F
- le parallélisme de traitement
- la tension V_{dd} ou l'excursion logique V_{escur} ou les deux
- le nombre de nœuds
- l'activité moyenne de chaque nœud A_i
- la capacité parasite de chaque nœud C_{1i}



Ajustement de la taille des transistors

- La dissipation minimale n'est pas donnée par la taille minimale des transistors



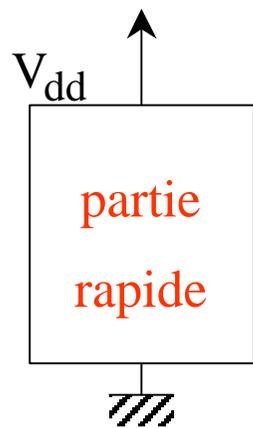
Tensions d'alimentation multiples

Idée: alimenter avec une tension d'alimentation plus basse les parties qui peuvent être lentes

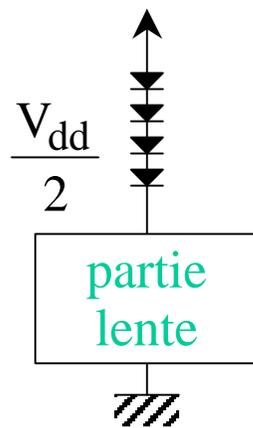
Problèmes: générer la tension d'alimentation

interfacer les parties alimentées différemment (l'alimentation est la référence logique)

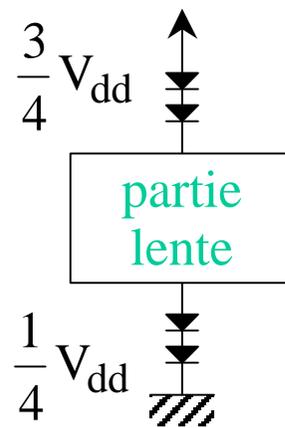
On peut réduire soit l'excursion logique V_{escur} seule, soit la tension V_{dd} et l'excursion V_{escur}



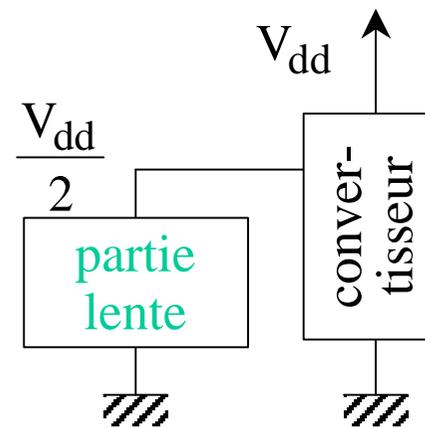
partie à
délai critique



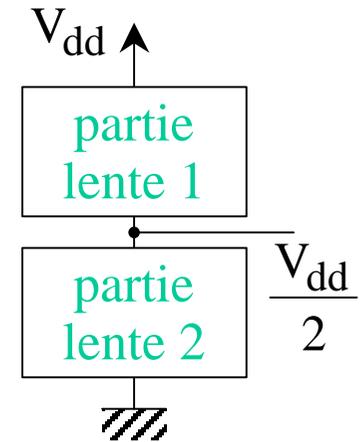
limiteur
d'excursion
logique



même seuil
que partie
rapide



rendement
du
convertisseur
continu/continu



nécessite
un
régulateur



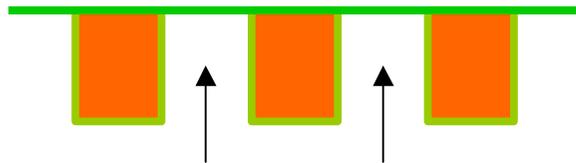
Réduction des capacités d'interconnexions

Idée: conductivité plus grande, section plus petites, permittivité plus faible

Aluminium

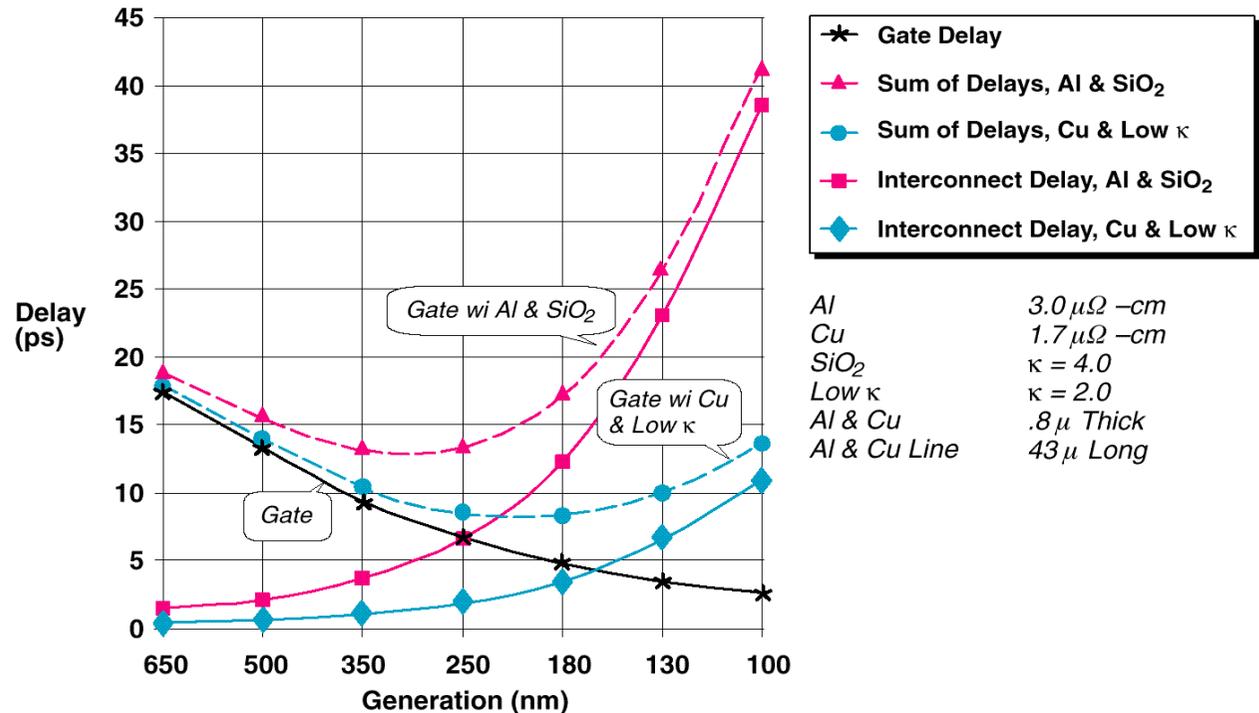


Cuivre



Diélectrique à faible permittivité

SPEED / PERFORMANCE ISSUE *The Technical Problem*



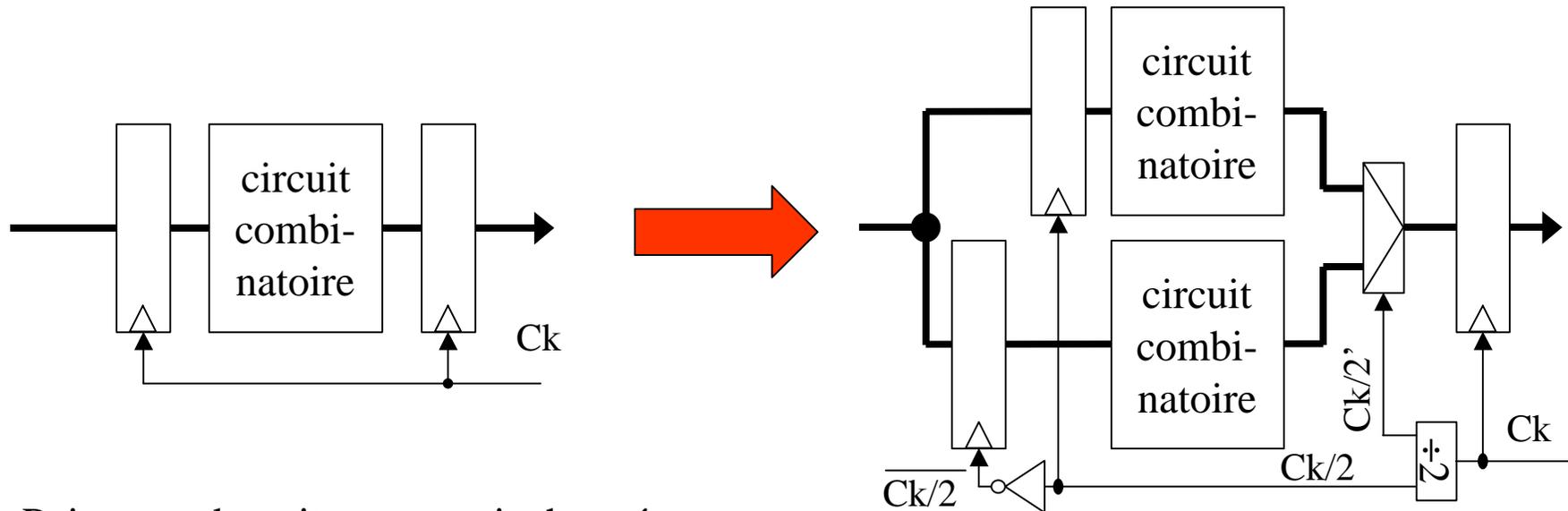
Al	3.0 μΩ-cm
Cu	1.7 μΩ-cm
SiO ₂	κ = 4.0
Low κ	κ = 2.0
Al & Cu	.8 μ Thick
Al & Cu Line	43 μ Long

Tiré de SIA 1997 Roadmap

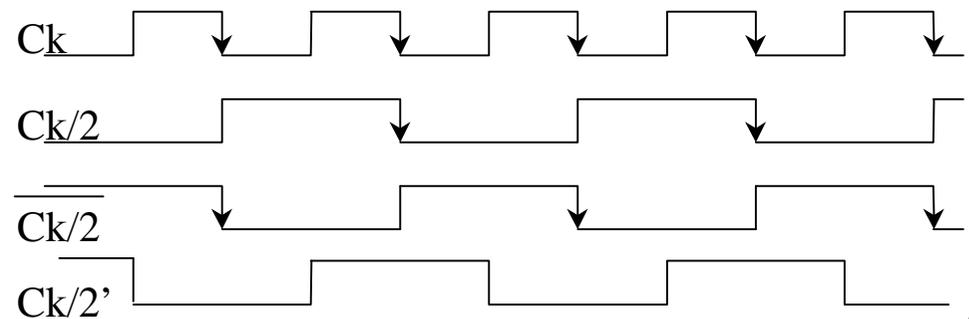


Parallélisme

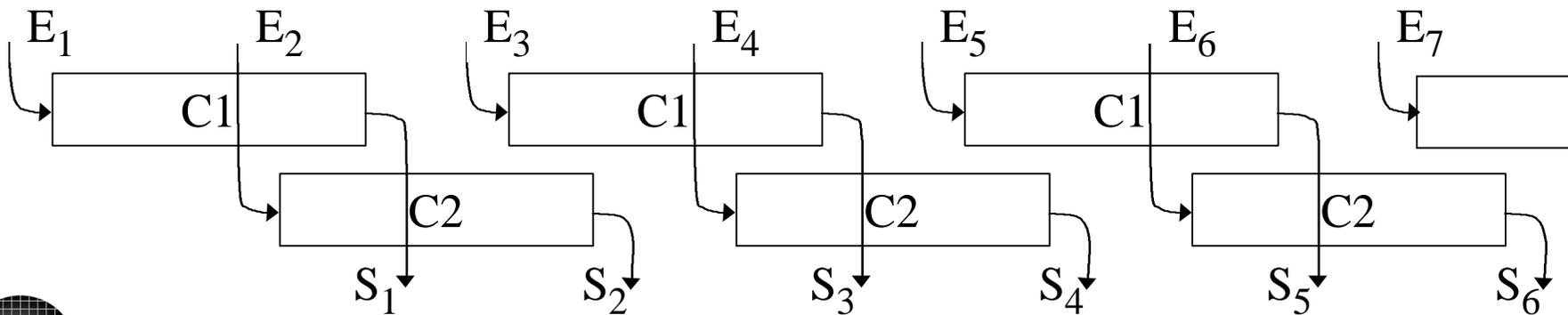
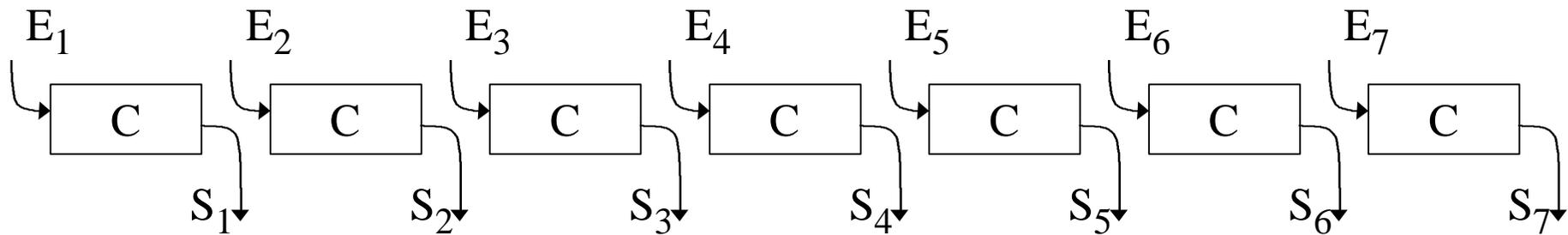
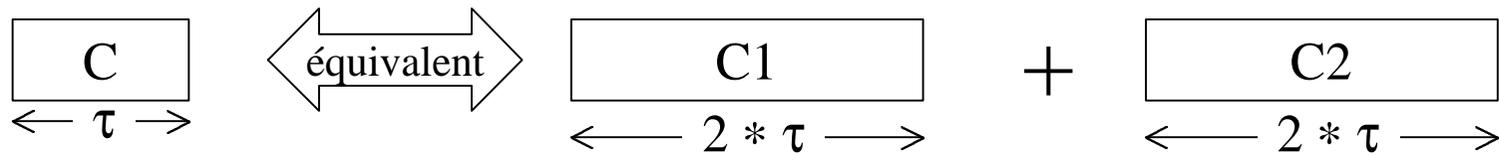
Idée: diviser par 2 la fréquence F de Ck . Doubler le circuit combinatoire conserve le débit



Puissance de traitement	inchangée
Complexité	\approx doublée
Latence	doublée
Délai	doublé
Fréquence	divisée par 2
Dissipation	?

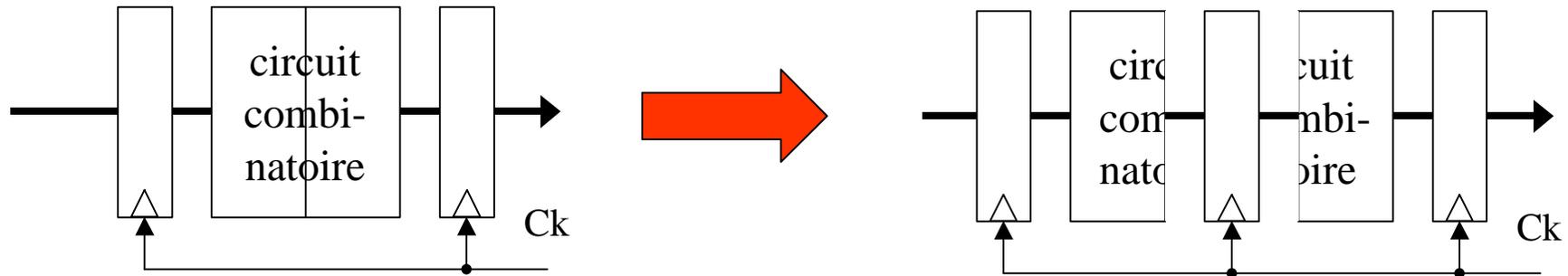


Chronogramme de blocs parallèles



Pipelining

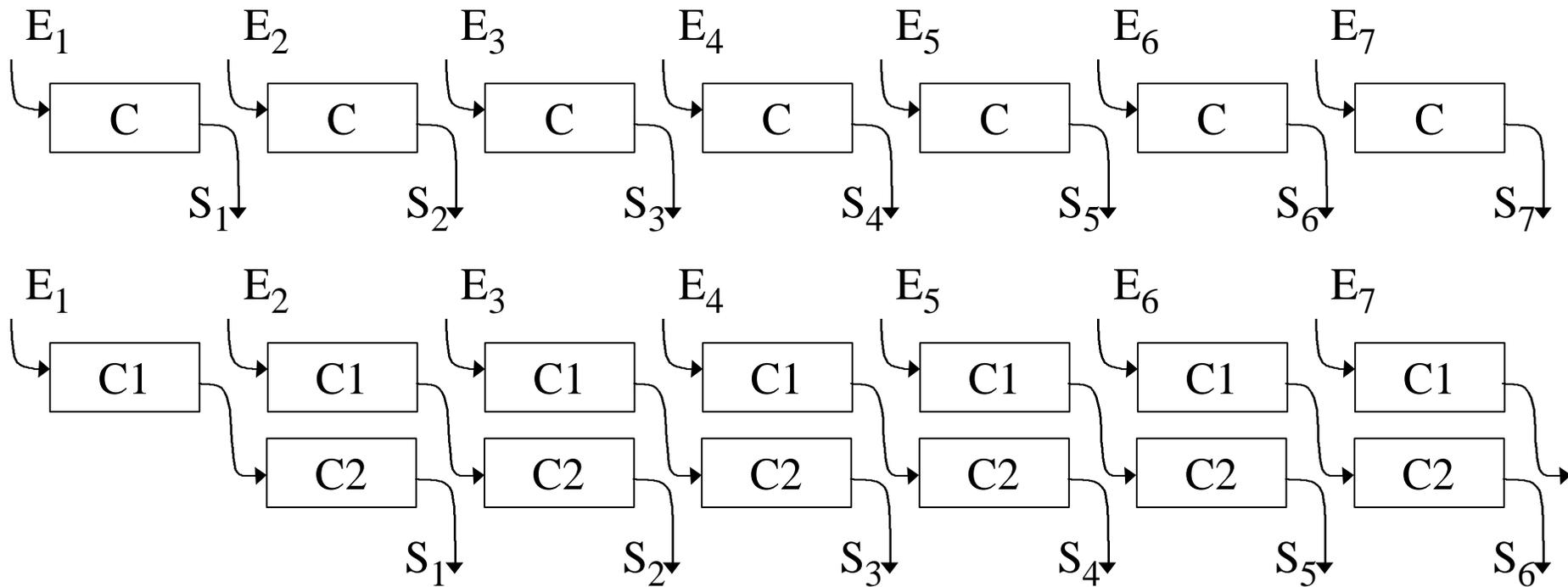
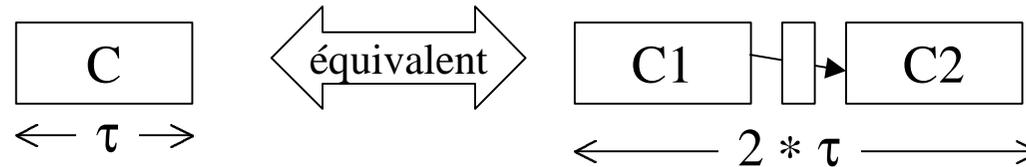
Idée: multiplier par 2 le délai. Un pipeline à 2 étages conserve le débit



Puissance de traitement	inchangée
Complexité	≈ inchangée
Latence	doublée
Délai	doublé
Fréquence	inchangée
Dissipation	?



Chronogramme de blocs "Pipelínés"



Une idée plus formelle de l'activité

Soit un circuit de 10 millions de portes (10^7 nœuds) fonctionnant à 10^8 Hz.
Il dissiperait en 1 seconde $10^7 * 10^8 * 10^{-12}$ joules soit 10^3 Watts si tous les nœuds commutent à chaque cycle de l'horloge.

Comment prédire l'activité d'un nœud ?

Comment réduire l'activité d'un nœud ?

$$\text{dissipation dynamique} = \frac{1}{2} * F * V_{dd}^2 * \left(\begin{array}{l} \text{nombre} \\ \text{de nœuds} \\ \sum_{i=1} A_i * C_{1i} \end{array} \right)$$



Probabilité des nœuds

On appelle nœud la variable booléenne connectant une sortie de porte à une entrée d'une autre porte.

On note P_a la probabilité que le nœud a ait la valeur 1. $0 \leq P_a \leq 1$, $a \in \{0,1\}$

La probabilité s'appelle également espérance mathématique.

Si on note "0" par "faux" et "1" par "vrai", P_a est tout simplement la probabilité de a.

La probabilité que le nœud a ait la valeur 0 est $P_{\bar{a}} = 1 - P_a$.

Les deux valeurs des nœuds ne sont généralement pas équiprobables, en général $P_a \neq 1/2$.

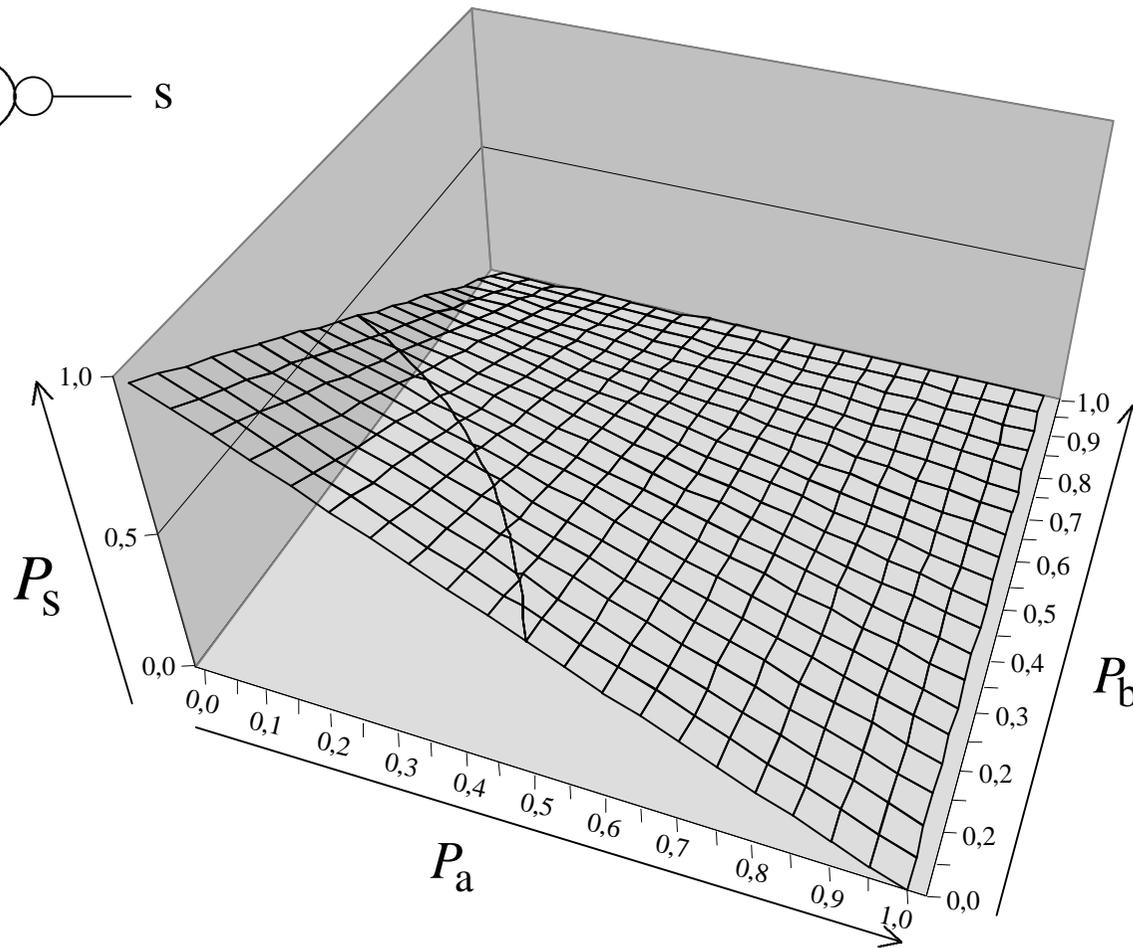
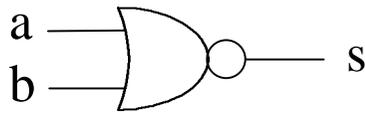
Il suffit d'observer la table de vérité d'un NOR à 2 entrées. Si les valeurs des entrées sont équiprobables et indépendante, alors la sortie s a 3 chances sur 4 de valoir 0 et seulement 1 sur 4 de valoir 1. Le déséquilibre est encore plus grand avec davantage d'entrées.

L'analyse statique permet l'évaluation des probabilités beaucoup plus rapidement que des statistiques

sur un très grand nombre n de cycles: $P_a = \frac{1}{n} * \sum_{i=1}^n a$



Probabilité de sortie d'une porte NOR



On peut calculer la probabilité de la sortie s à partir des probabilités des deux entrées

$$P_s = (1 - P_a) * (1 - P_b)$$



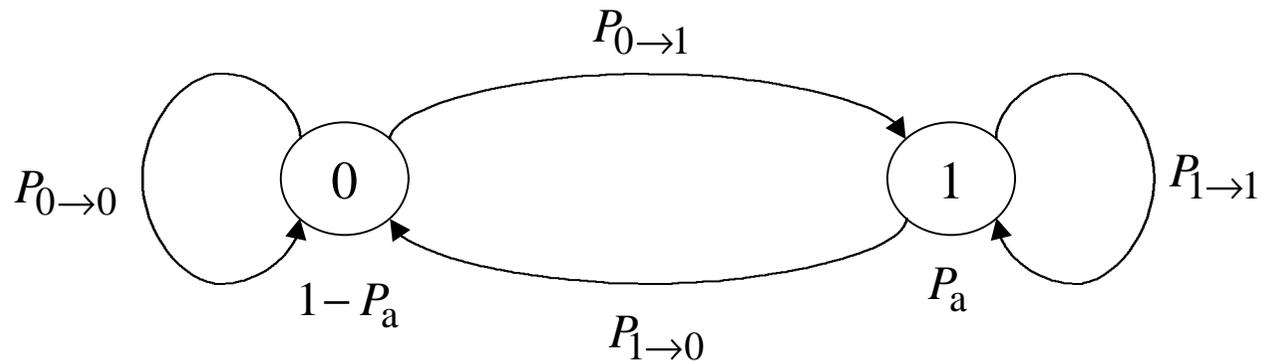
Activité moyenne des nœuds

On appelle nœud la variable booléenne connectant une sortie de porte à une entrée d'une autre porte.

On note P_a la probabilité que le nœud a ait la valeur 1, $0 \leq P_a \leq 1$.

La probabilité de transition d'un nœud est donnée directement par la probabilité de la valeur de ce nœud pourvu que les valeurs successives soient non corrélées.

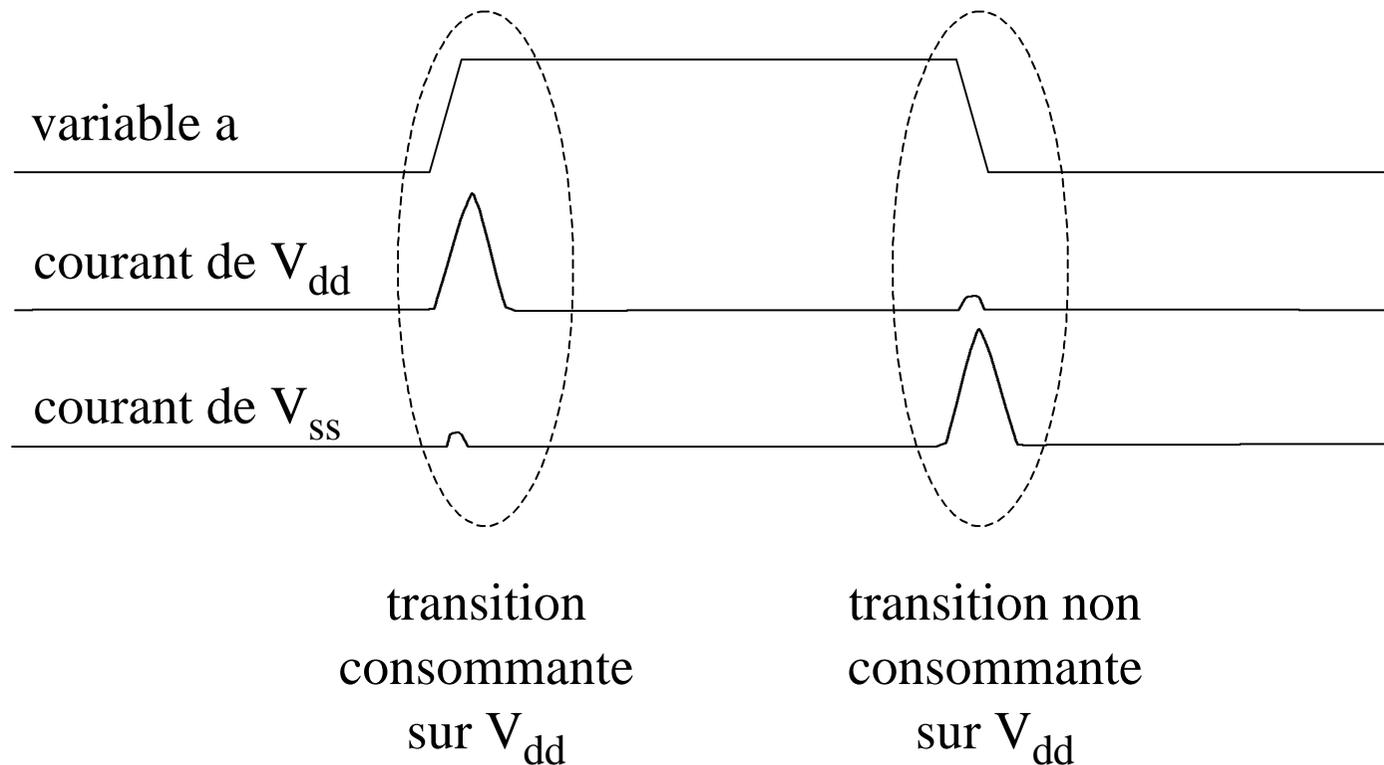
$$P_{0 \rightarrow 1} = (1 - P_a) * P_a \text{ et } P_{1 \rightarrow 0} = P_a * (1 - P_a).$$



probabilité de valeur et de transition du nœud a



Transition consommante

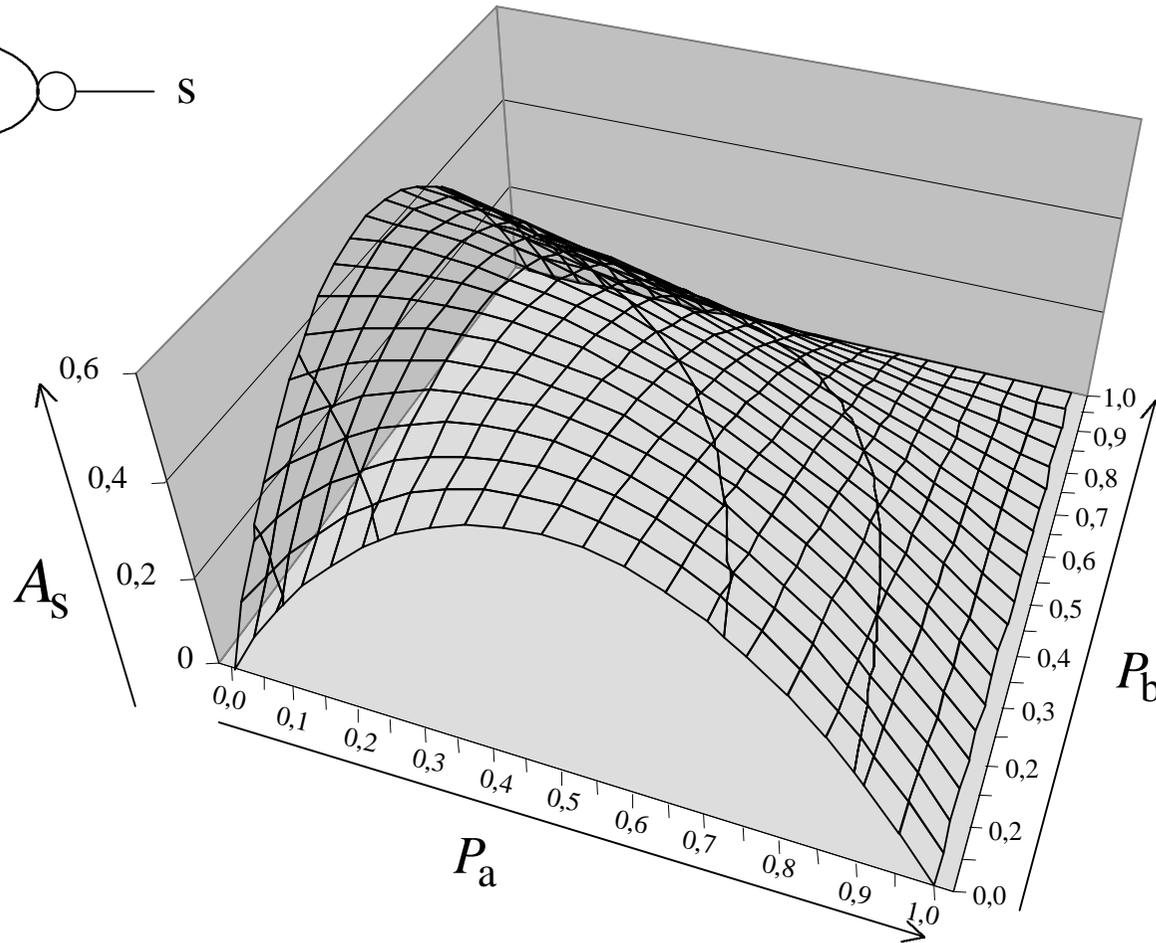
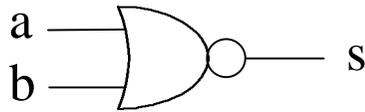


Par convention on appelle activité tout changement c'est à dire toutes transitions (montantes et descendantes). On pondère la consommation.

$$A = P_{0 \rightarrow 1} + P_{1 \rightarrow 0}$$



Activité d'une porte NOR



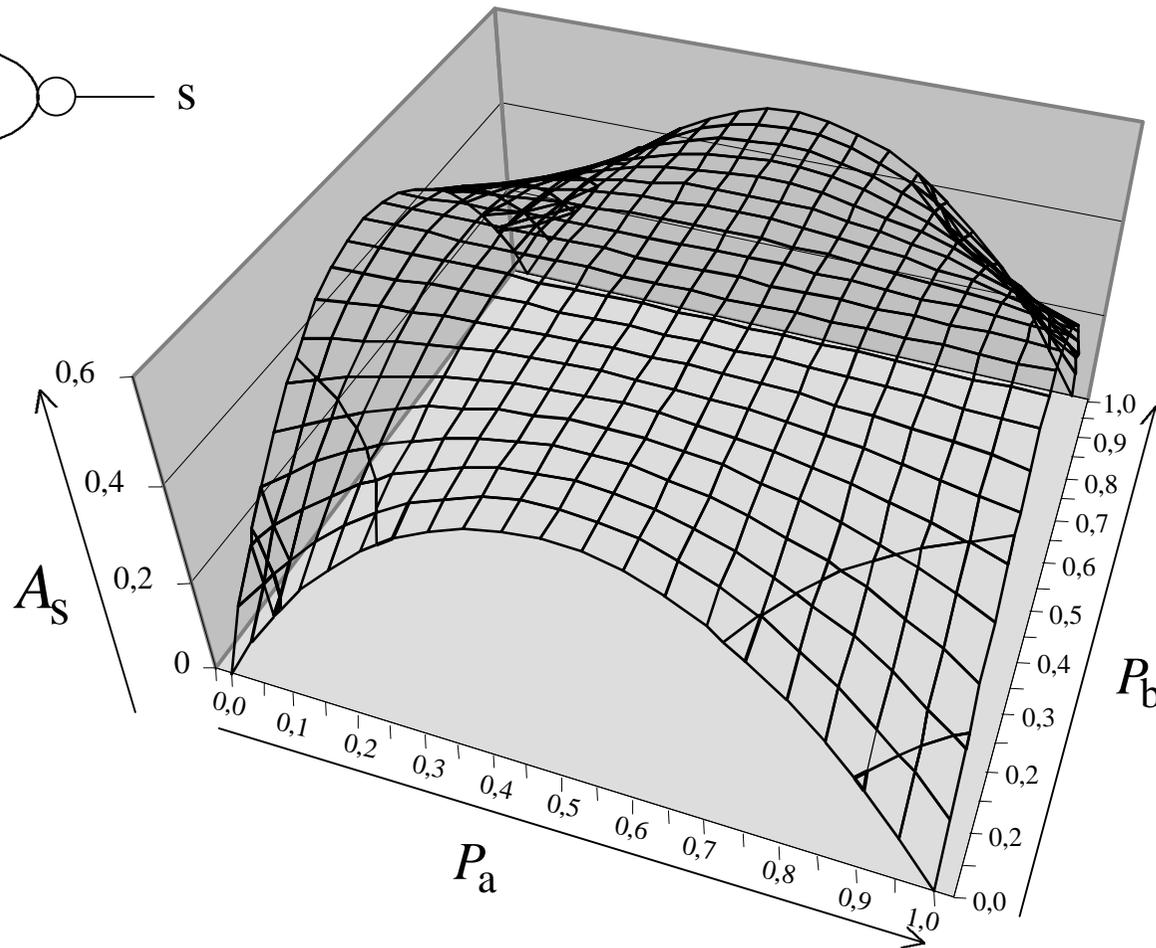
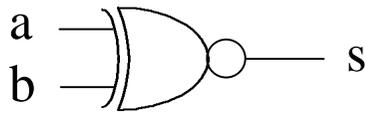
On peut calculer l'activité avec la probabilité mais pas l'inverse

$$A_s = 2 * (1 - P_s) * P_s$$

$$A_s = 2 * ((1 - (1 - P_a) * (1 - P_b))) * (1 - P_a) * (1 - P_b))$$



Activité d'une porte XOR



$$A_s = 2 * ((1 - (P_a + P_b - 2 * P_a * P_b)) * (P_a + P_b - 2 * P_a * P_b))$$



Limites du modèle

Les hypothèses d'indépendance temporelle et d'indépendance spatiale de valeurs de signaux, qui permettent de déduire l'activité de la probabilité, ne sont pas toujours observées:

Corrélation temporelle

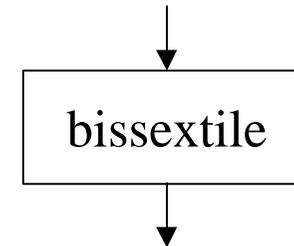
$$P_s = \frac{1}{4} - \frac{1}{100} + \frac{1}{400} \approx 0,25$$

Corrélation spatiale

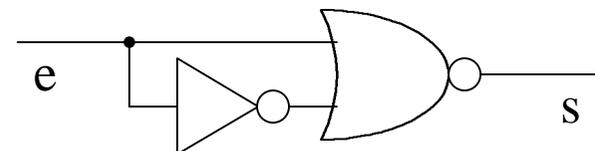
$$P_e = 0,5 \quad A_e = 0,5$$

$$P_s = 0 \quad A_s = 0$$

4 chiffres



$s = 1 \Leftrightarrow$ année bissexile



Reconvergence de chemins

Il faut propager simultanément l'activité et la probabilité des nœuds



Propagation des probabilités

inverseur $P_{\bar{e}} = 1 - P_e$

et logique $P_{a \wedge b} = P_a * P_b$

ou logique $P_{a \vee b} = P_a + P_b - (P_a * P_b)$

soit $y = F(x_1, x_2, \dots, x_n)$ une porte à n entrées.

$$P_y = P_{x_i} * PF(x_1, x_2, \dots, 1 \dots x_n) + (1 - P_{x_i}) * PF(x_1, x_2, \dots, 0 \dots x_n)$$

correlation $(x, y) = \frac{P_{x \wedge y}}{P_x * P_y}$ probabilité que x et y soient vrais simultanément

et logique $P_s = \prod_{x \in \text{entrées}} P_x * \prod_{y > x} \text{correlation}(x, y)$

ou logique $P_s = 1 - \left(\prod_{x \in \text{entrées}} (1 - P_x) * \prod_{y > x} \text{correlation}(\bar{x}, \bar{y}) \right)$



Propagation des activités

Soit $y = F(x_1, x_2, \dots, x_n)$ une porte à n entrées.

On nomme dérivée partielle booléenne de y par rapport à x_i :

$$\frac{\partial y}{\partial x_i} = F(x_1, x_2, \dots, 0 \dots x_n) \oplus F(x_1, x_2, \dots, 1 \dots x_n)$$

$P\left(\frac{\partial y}{\partial x_i}\right)$ est la probabilité que y change lorsque x_i change.

$$L'activité A_y = \sum_{i=1}^n P\left(\frac{\partial y}{\partial x_i}\right) * A_{x_i}$$

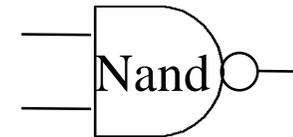
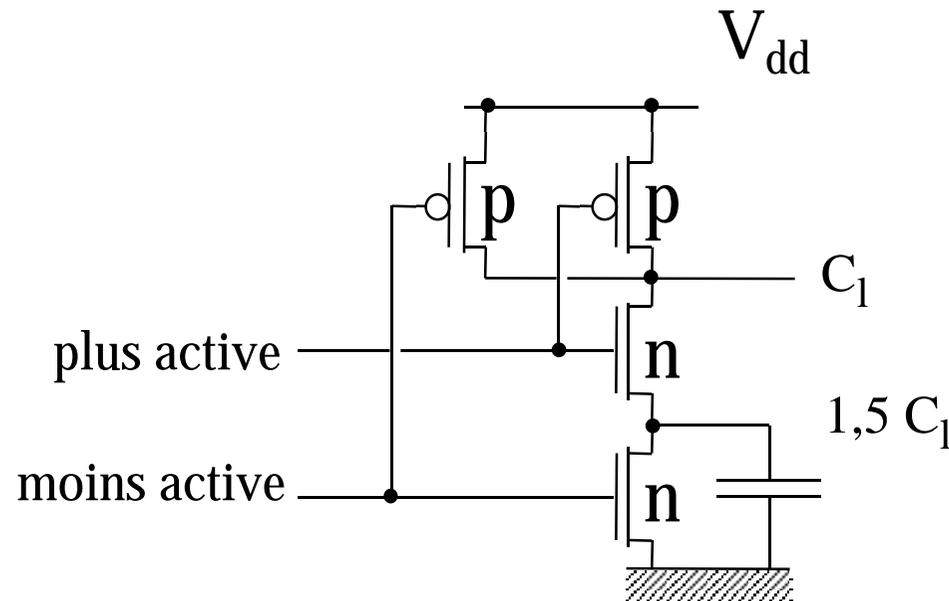
inverseur $A_{\bar{e}} = A_e$

et logique $A_{a \wedge b} = P_a * A_b + P_b * A_a$

ou logique $A_{a \vee b} = (1 - P_a) * A_b + (1 - P_b) * A_a$



Réordonnancement des entrées



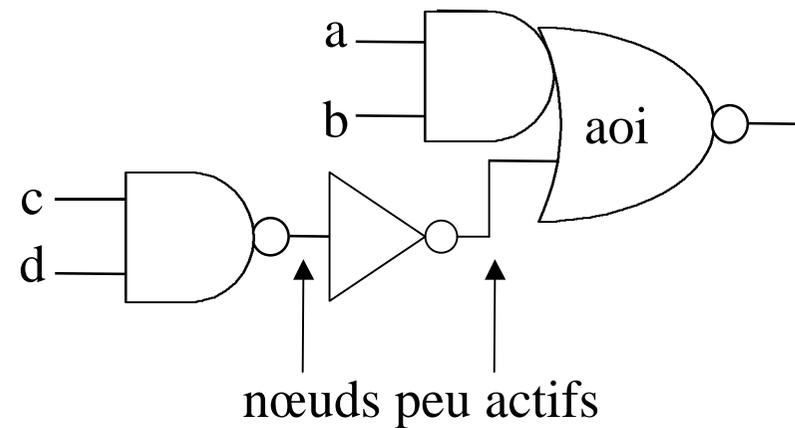
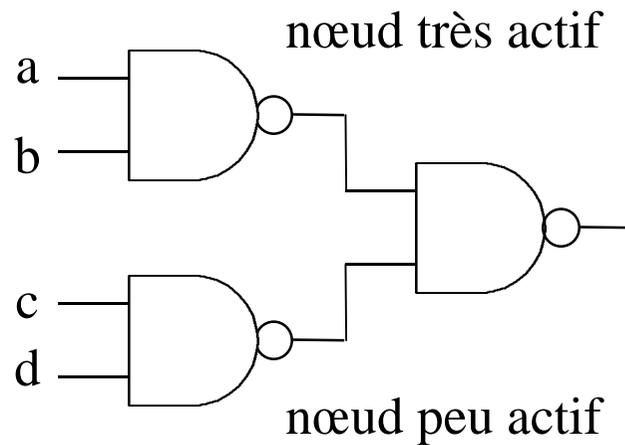
Demande de calculer
l'activité en tout nœud

minimise
$$\left(\begin{array}{c} \text{nombre} \\ \text{de nœuds} \\ \sum_{i=1} A_i * C_{1i} \end{array} \right)$$

Cet ordre peut être en conflit avec l'optimisation du délai.



Synthèse pour la faible consommation



Exemple "PowerCompiler™" de Synopsys



Redondance pour réduire l'activité

Comment pouvons nous diminuer la redondance de l'information pour diminuer l'activité et la puissance dissipée

Nous voulons calculer $S = X + Y$, toutes les valeurs de X et Y sont équiprobables et indépendantes.

Définissons $o_i = x_i \vee y_i$ et $a_i = x_i \wedge y_i$ Les valeurs de o_i et a_i ne sont plus indépendantes puisque $a_i = 1 \Rightarrow b_i = 1$:

x_i	y_i	$o_i = x \vee y$	$a_i = x \wedge y$	$s = x + y$	$o_i + a_i$
0	0	0	0	0	0
0	1	1	0	1	1
1	0	1	0	1	1
1	1	1	1	2	2

même somme



Recomptons les transitions

Distance de Hamming entre la nouvelle et l'ancienne valeur ($x_i y_i$)

$x_i y_i$	00	01	10	11
00	0	1	1	2
01	1	0	2	1
10	1	2	0	1
11	2	1	1	0

Toutes les occurrences sont équiprobable

Nombre moyen de transitions = $16/16 = 1$



Bilan du nombre moyen de transitions

Distance de Hamming entre la nouvelle et l'ancienne valeur de $(o_i a_i)$

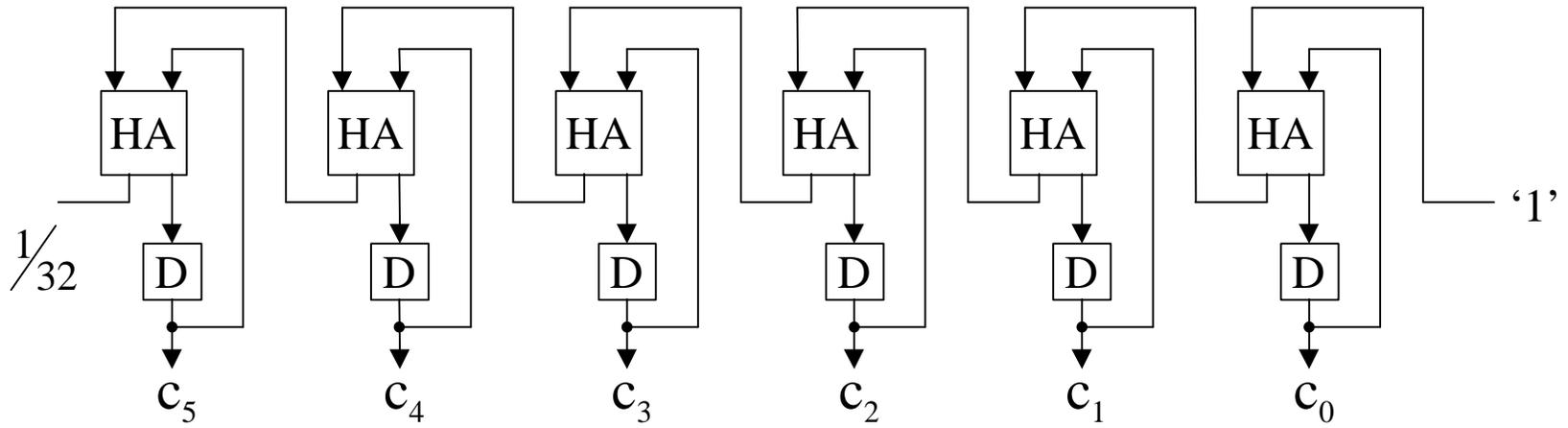
$o_i a_i$	00	10	10	11
00	0	1	1	2
10	1	0	0	1
10	1	0	1	1
11	2	1	1	0

Le nombre moyen de transitions est réduit à 12/16.

Le gain en activité est 25% au coût de 2 portes logiques (or, and)



Compteur binaire



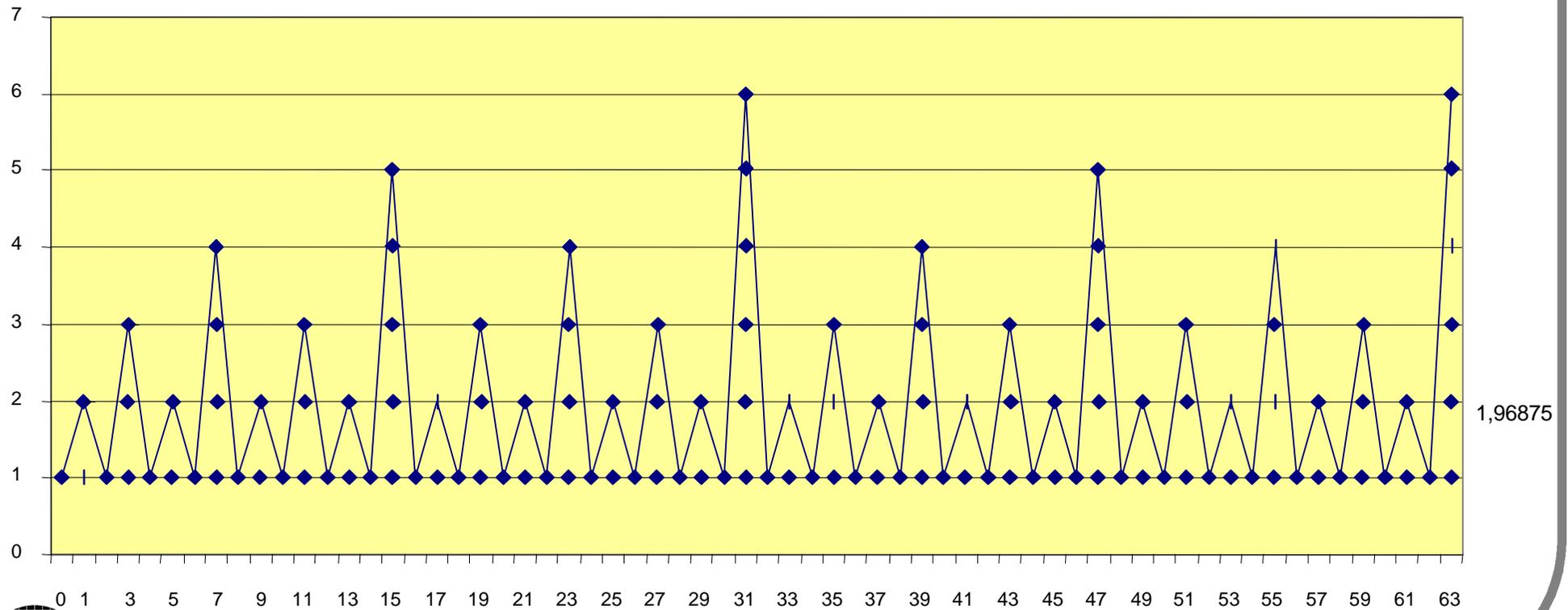
probabilité	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
activité	$\frac{1}{32}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{1}$

l'activité moyenne en sortie d'un compteur binaire est $\sum_{i=0}^{n-1} \frac{1}{2^i} \approx 2$



Activité d'un compteur

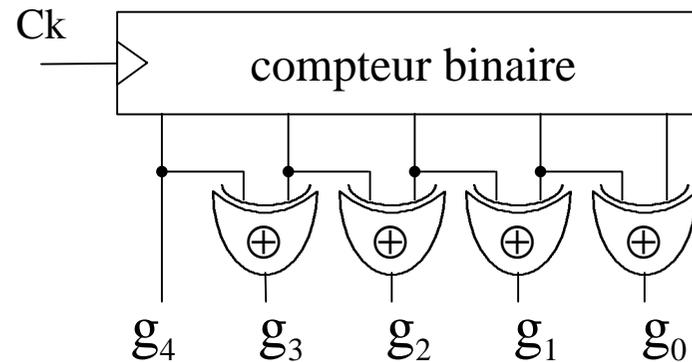
Distance de Hamming des valeurs consécutives d'un compteur binaire



Compteur de Gray

	binaire	gray
0	0 0 0 0 0	0 0 0 0 0
1	0 0 0 0 1	0 0 0 0 1
2	0 0 0 1 0	0 0 0 1 1
3	0 0 0 1 1	0 0 0 1 0
4	0 0 1 0 0	0 0 1 1 0
5	0 0 1 0 1	0 0 1 1 1
6	0 0 1 1 0	0 0 1 0 1
7	0 0 1 1 1	0 0 1 0 0
8	0 1 0 0 0	0 1 1 0 0
9	0 1 0 0 1	0 1 1 0 1
10	0 1 0 1 0	0 1 1 1 1
11	0 1 0 1 1	0 1 1 1 0
12	0 1 1 0 0	0 1 0 1 0
13	0 1 1 0 1	0 1 0 1 1
14	0 1 1 1 0	0 1 0 0 1
15	0 1 1 1 1	0 1 0 0 0
16	1 0 0 0 0	1 1 0 0 0
17	1 0 0 0 1	1 1 0 0 1
18	1 0 0 1 0	1 1 0 1 1
19	1 0 0 1 1	1 1 0 1 0
20	1 0 1 0 0	1 1 1 1 0
21	1 0 1 0 1	1 1 1 1 1
22	1 0 1 1 0	1 1 1 0 1
23	1 0 1 1 1	1 1 1 0 0
24	1 1 0 0 0	1 0 1 0 0
25	1 1 0 0 1	1 0 1 0 1
26	1 1 0 1 0	1 0 1 1 1
27	1 1 0 1 1	1 0 1 1 0
29	1 1 1 0 0	1 0 0 1 0
29	1 1 1 0 1	1 0 0 1 1
30	1 1 1 1 0	1 0 0 0 1
31	1 1 1 1 1	1 0 0 0 0

- Code de Gray: 2 valeurs ayant une distance arithmétique de 1 ont une distance de Hamming de 1.
- un compteur en Code Gray est plus complexe qu'un compteur en binaire pur.
- Le code de Gray s'appelle également code réfléchi.

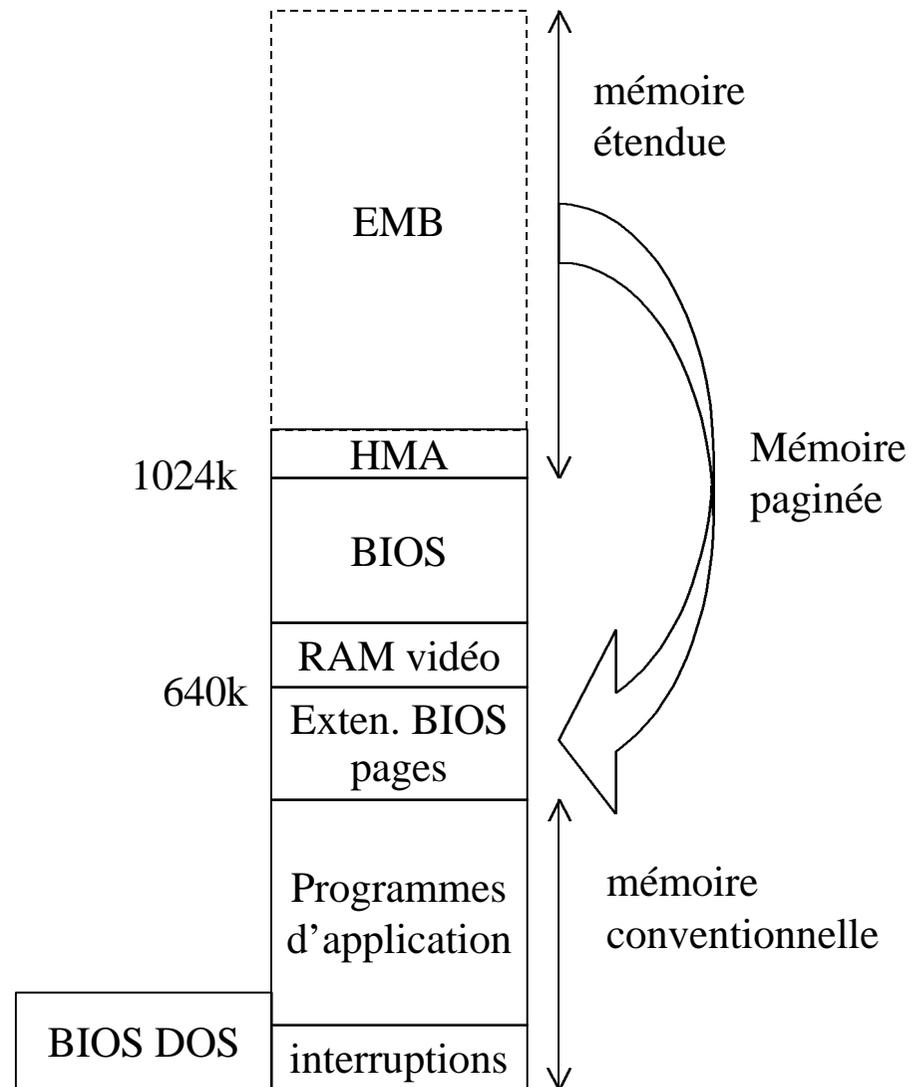


$$g_i = b_{i+1} \oplus b_i$$

activité minimum



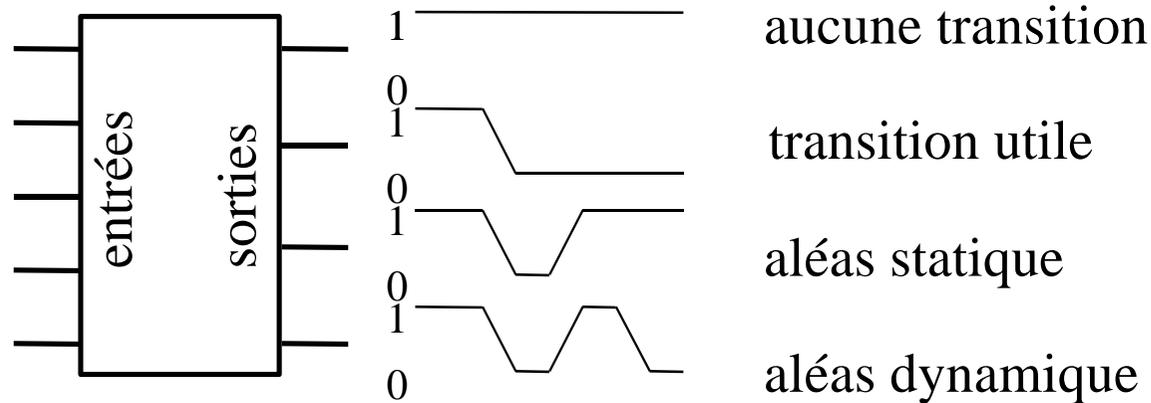
Compteur semi-Gray



En binaire, l'examen d'un petit nombre de bits permet de décider quelle zone adresser



Taxonomie des Transition



les transitions des sorties sont causées par les transitions des entrée
les aléas résultent de différences de délais

activité redondante $\begin{cases} \rightarrow \text{transition complète (0} \Leftrightarrow V_{dd}) \\ \rightarrow \text{transition incomplète} \\ \text{(long } R_t \Rightarrow \text{ courant de court circuit)} \end{cases}$



Transition utile ou redondante

- Transition utile
 - Prédite dans la description fonctionnelle
 - Arriverait même avec un délai nul
 - Désirable
- Transition redondante
 - Non prédite dans la description fonctionnelle
 - Arrive parce que les délais sont non nuls
 - Non désirable
 - Peut être une important contribution à la puissance totale
- Exemple du multiplieur de Braun 16x16-bits (multiplieur très populaire)
 - Utile 35 %
 - Redondant 65 %



Activité moyenne

Définition:

$$\text{Activité moyenne} = \frac{\text{Nombre total de transition générées par toutes les combinaisons de paires de vecteurs d'entrée}}{\text{Nombre total de paires de vecteurs d'entrée}}$$



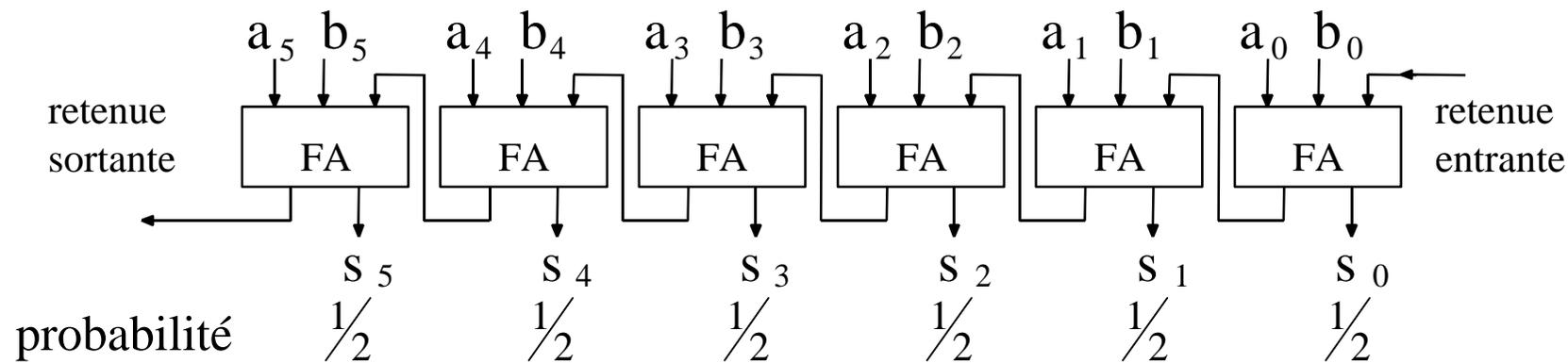
l'activité peut être divisée en deux composantes:

$$A = A_{\text{utile}} + A_{\text{redondante}}$$



Activité de l'additionneur à propagation

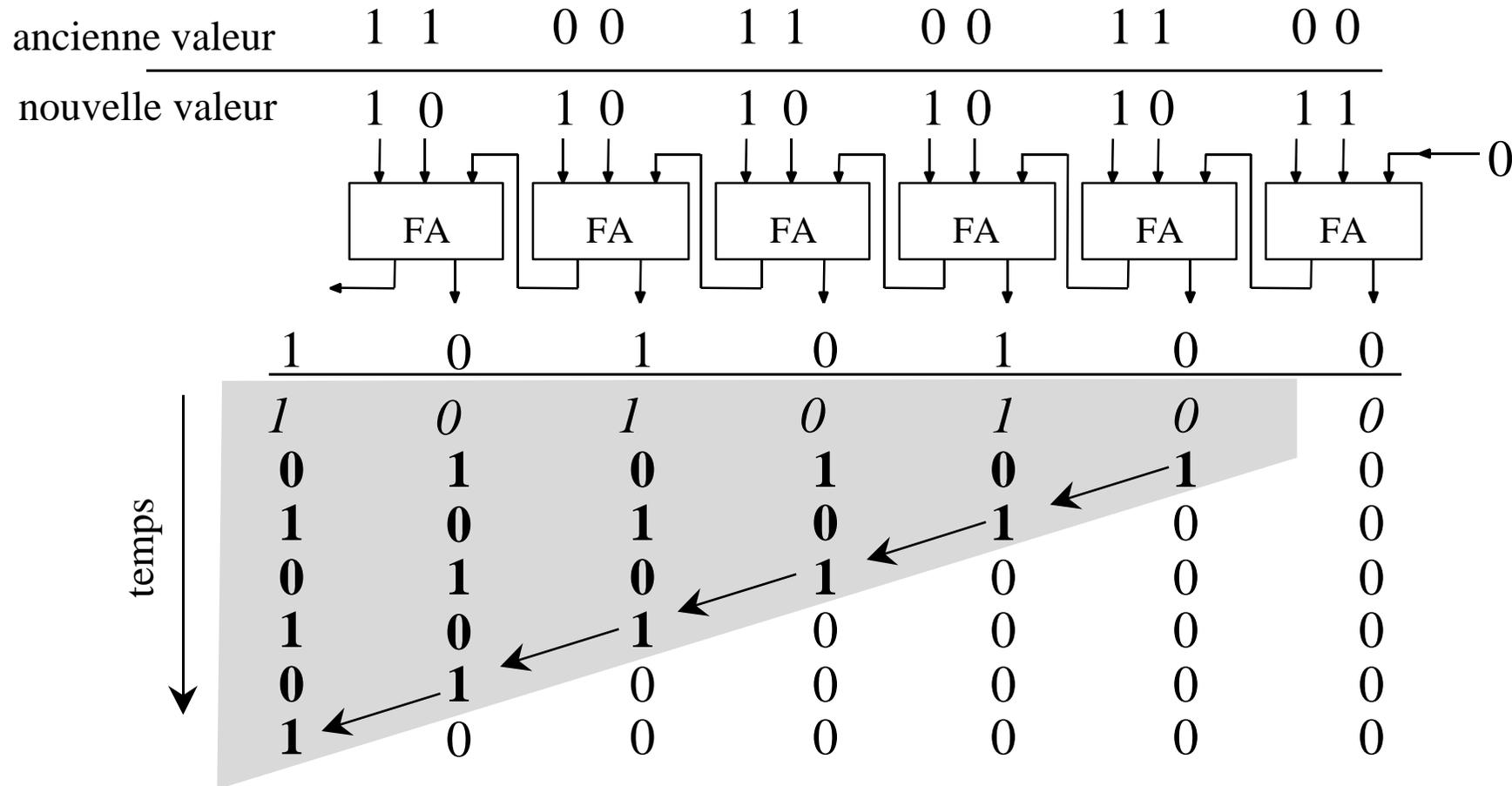
$$A = \sum_{i=0}^5 a_i 2^i \quad B = \sum_{i=0}^5 b_i 2^i \quad S = \sum_{i=0}^5 s_i 2^i \quad a_i, b_i, s_i \in \{0, 1\}$$



- Si toutes les valeurs sont indépendante et équiprobables, l'activité moyenne de la sortie est $n/2$ si on ne tient pas compte des délais (modèle à délai nul)
- La sortie poids fort dépend de toutes les entrées avec des délais différents.
L'activité réelle est plus importante que prédite par la probabilité de la valeur des nœuds.



Activité pire cas d'un additionneur



$$\text{activité pire cas} \approx \frac{n^2}{2}$$



Modèle analytique de l'activité du RCA₍₁₎

- l'activité est proportionnelle au carré de la longueur du chemin de la retenue
- compter le nombre moyen de chaînes de longueur k dans un mot de longueur n.
- Soit T(k,n) le nombre de chaînes de longueur k dans tous les mots de longueur n

$$\frac{\underline{11\dots10}}{k+1} \quad \frac{\underline{011\dots01}}{n-(k+1)}$$

$$2 * 2^{n-k-1}$$

$$\frac{\underline{01\dots10}}{k+2} \quad \frac{\underline{011\dots01}}{n-(k+2)}$$

$$(n-k-1) * 2^{n-k-2}$$

$$T(k,n) = 2^{n-k} \left(1 + \frac{n-k-1}{4} \right) \text{ pour } 0 < k < n$$

et

$$T(0,n) = 0 \text{ et } T(n,n) = 1$$



Modèle analytique de l'activité du RCA₍₂₎

- ◆ Somme sur tous les cas possibles :

$$\text{Activité moyenne} = A(n) = \frac{1}{2^n} \sum_{k=0}^n T(k, n) \frac{k^2}{2}$$

On obtient :

$$A(n) = \frac{3n-4}{4} - \frac{3n^2}{2^{n+3}} \xrightarrow{n \rightarrow \infty} \frac{3n}{4}$$

- ◇ L'activité utile d'un additionneur de n bits est $n/2$



Modèle analytique de l'activité du RCA₍₃₎

◆ Longueur moyenne du chemin de propagation :

$$\frac{1}{2^n} \sum_{k=0}^n k * T(k, n) = \frac{n}{2} - \frac{3n}{2^{n+2}} \xrightarrow{n \rightarrow \infty} \frac{n}{2} \quad (\text{résultat connu})$$

◆ Proportion d'activité redondante dans activité totale :

$$h = \frac{A_{\text{redondant}}}{A} = \frac{A - A_{\text{utile}}}{A} = \frac{n - 4}{3n - 4} = 33\% \text{ pour les grands } n$$

◆ Erreur:

# de bits	Activité (%)	Délai (%)	η (%)
8	9.3750	2.34	8.2759
16	0.15	0.02	0.04
32	8.94e-06	5.59e-07	4.26e-06
64	8.33e-15	2.60e-16	2.58e-06



Réduction de l'activité parasite

- L'activité redondante représente typiquement de 10% à 40% de l'activité.
- Les délais des chemins convergent vers une même porte doivent être équilibrés.
- Insérer des portes dans les chemins plus court ne change pas le délai du chemin critique.
- Par contre cela réduit l'activité redondante mais augmente le nombre de nœuds.

⇒ Il y a compromis et optimisation

Remarque 1: il n'y a pas d'activité redondante en sortie des registres donc le pipeline réduit l'activité redondante.

Remarque 2: il n'y a pas d'activité redondante dans les circuits "self timed" qui ont cependant plus de nœuds ⇒ il y a compromis .



Comparaison d'activité d'additionneurs

Type additionneur	# de Δ -cells	délai (Δ -cell)	Max. fan-out	activité utile
propagation (RCA)	$n - 1$	$n - 1$	2	$n / 2$
2-level carry select	$\lceil 2n - \sqrt{2n} \rceil$	$\lceil \sqrt{2n} \rceil$	$\lceil \sqrt{2n} \rceil$	$\lceil 2n - \sqrt{2n} \rceil / 2$
3-level carry select	$5/2 n - 3 \log_2(n)$	$\lceil \sqrt[3]{6n} \rceil$	$\lceil \sqrt[3]{6n} \rceil$	N.A.
Brent-Kung	$\lceil 2n - \log_2(n) \rceil$	$\lceil 2 \log_2(n) - 2 \rceil$	$\lceil 2 \log_2(n) - 2 \rceil$	N.A.
Sklansky	$\lceil n / 2 \log_2(n) \rceil$	$\lceil \log_2(n) \rceil$	$n / 2$	$\approx \lceil n / 4 \log_2(n) \rceil$
Kogge et Stone	$\lceil n (\log_2(n) - 1) \rceil$	$\lceil \log_2(n) \rceil$	2	$\approx \lceil n / 2 \log_2(n) \rceil$
Han et Carlson	$\lceil n / 2 \log_2(n) \rceil$	$\lceil \log_2(n) \rceil + 1$	2	$\approx \lceil n / 4 \log_2(n) \rceil$

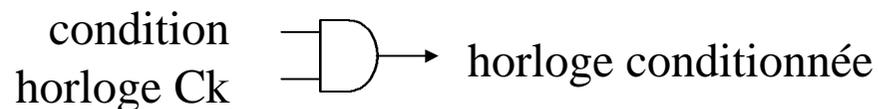


Adaptation dynamique

Remarque: Les circuits sont dimensionnés pour le pire cas et non le cas moyen.

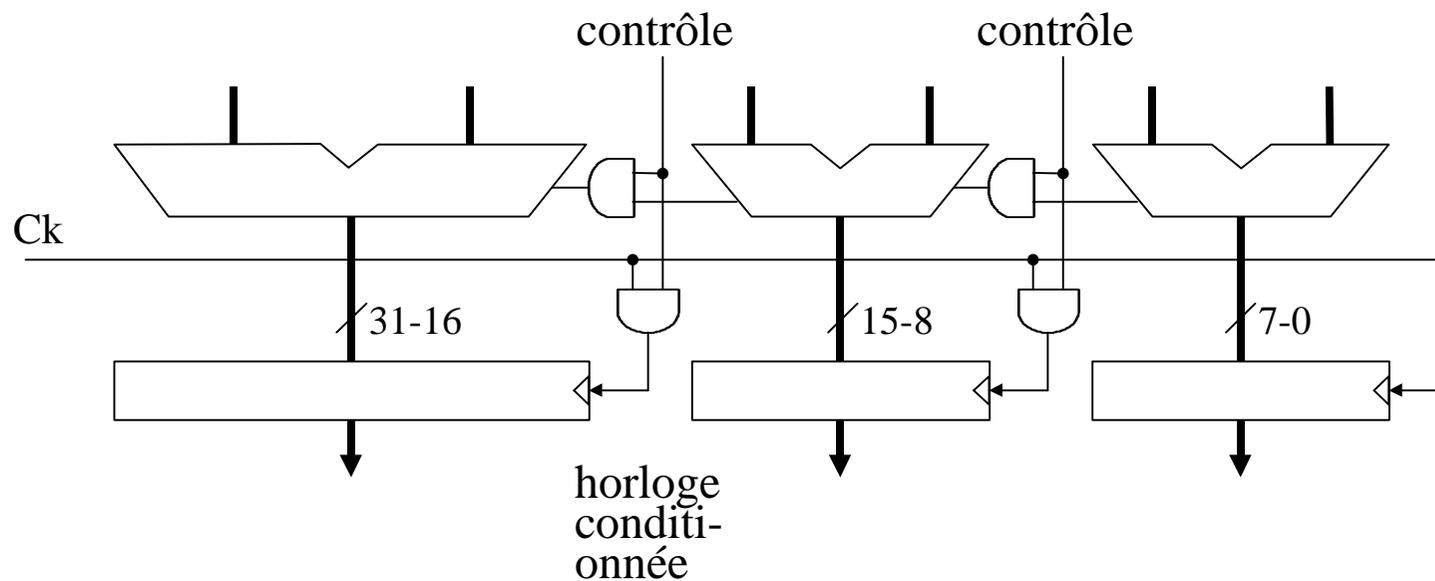
- le délai des circuits synchrones est calculé pour le pire cas, pour le cas moyen beaucoup plus fréquent le circuit est plus rapide que nécessaire
- la dynamique des nombres est calculée pour le pire cas, pour le cas moyen beaucoup plus fréquent les nombres ont plus de bits que nécessaire
- les opérations arithmétiques sont effectuées sur tous les bits des nombres, inutilement pour les valeurs les plus courantes (comparaison)

Méthode: ne pas transmettre l'horloge C_k aux parties de circuit dont l'activité n'est pas nécessaire ("gated clock" ou "horloge conditionnée")



Ajustement du chemin de données

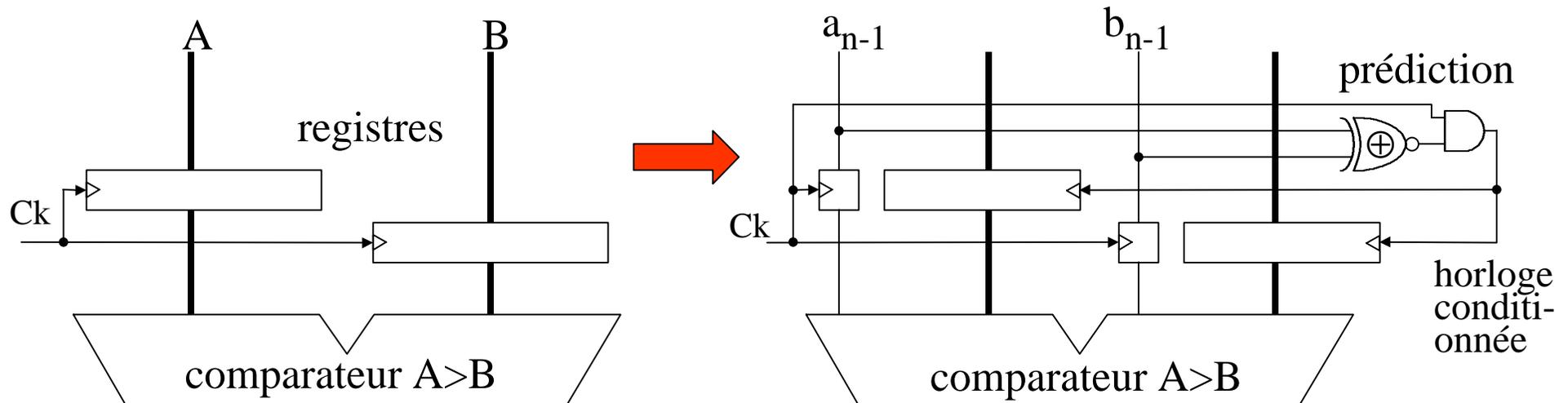
Idée: ajuster le nombre de bits du chemin de données à la précision demandée pour supprimer l'activité non nécessaire



Prédiction: exemple la comparaison

Idée: pour une comparaison, si les bits poids fort a_{n-1} et b_{n-1} des deux nombres sont différents, il est inutile d'examiner les autres bits, ce qui permet de réduire l'activité.

$$A = \sum_{i=0}^{n-1} a_i 2^i \quad B = \sum_{i=0}^{n-1} b_i 2^i \quad \text{on compare les nombres A et B}$$



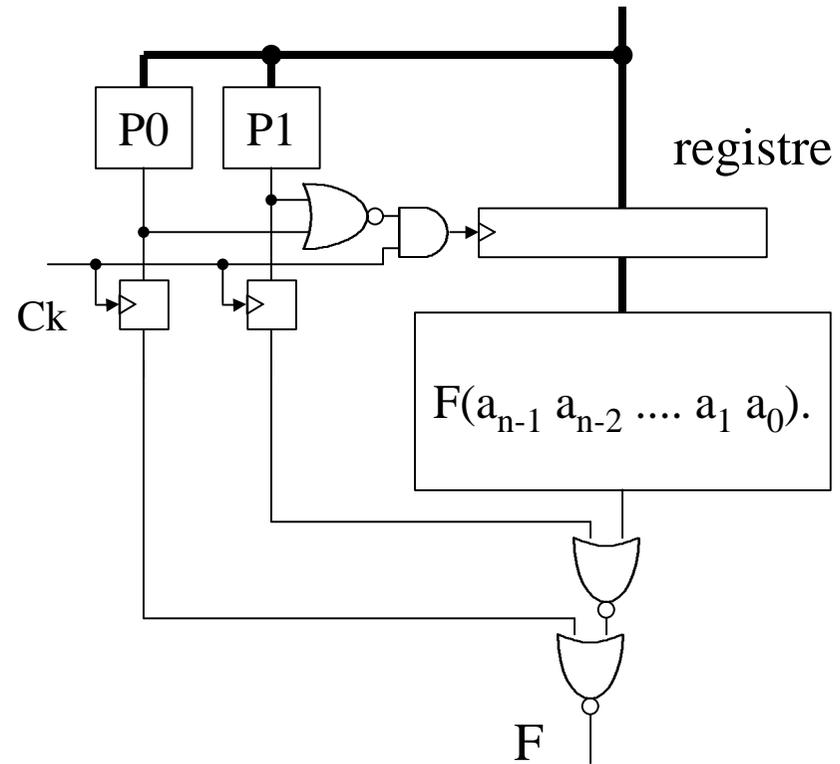
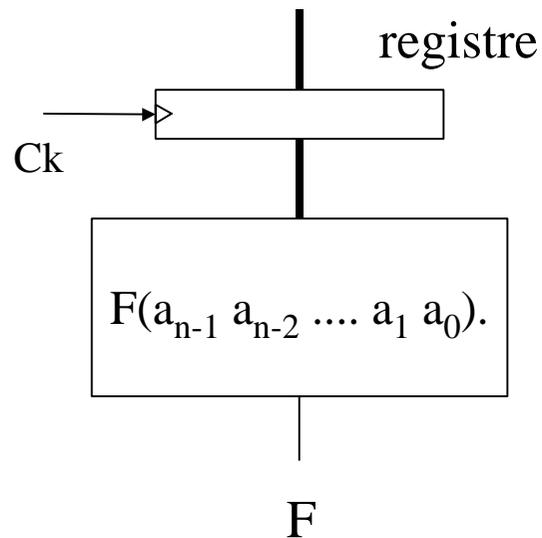
La prédiction est souvent utilisée pour les automates à basse consommation, elle est appelée également précalcul.



Généralisation de la prédiction

Idée: on veut calculer $F(a_{n-1} a_{n-2} \dots a_1 a_0)$.

On cherche 2 fonctions $P0(a_{n-1} a_{n-2} \dots a_1 a_0)$ et $P1(a_{n-1} a_{n-2} \dots a_1 a_0)$ telles que $P0 = 1 \Rightarrow F = 0$ et $P1 = 1 \Rightarrow F = 1$. Si ni $P0$ ni $P1$ ne sont vrais, il faut calculer F . $P0$ et $P1$ sont appelées les "fonctions de prédiction" de F .

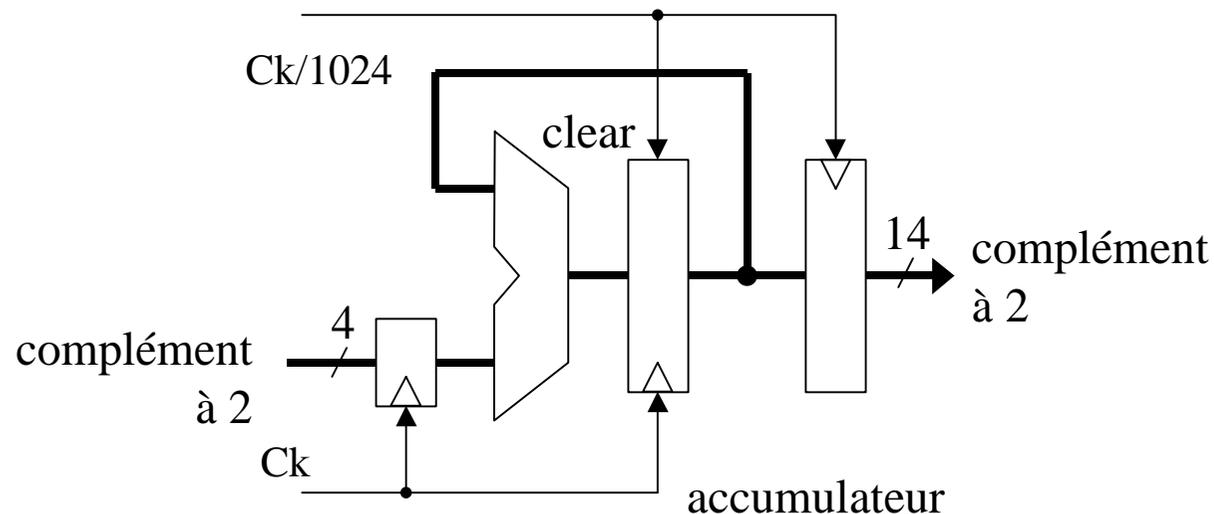


Accumulation de données

Exemple: circuit accumulant 1024 échantillons de 4 bits signés à 64 MHz.

L'accumulation nécessite $4 + \log_2(1024) = 14$ bits.

Chaque fois que l'accumulateur change de signe, 10 bits au moins commutent.

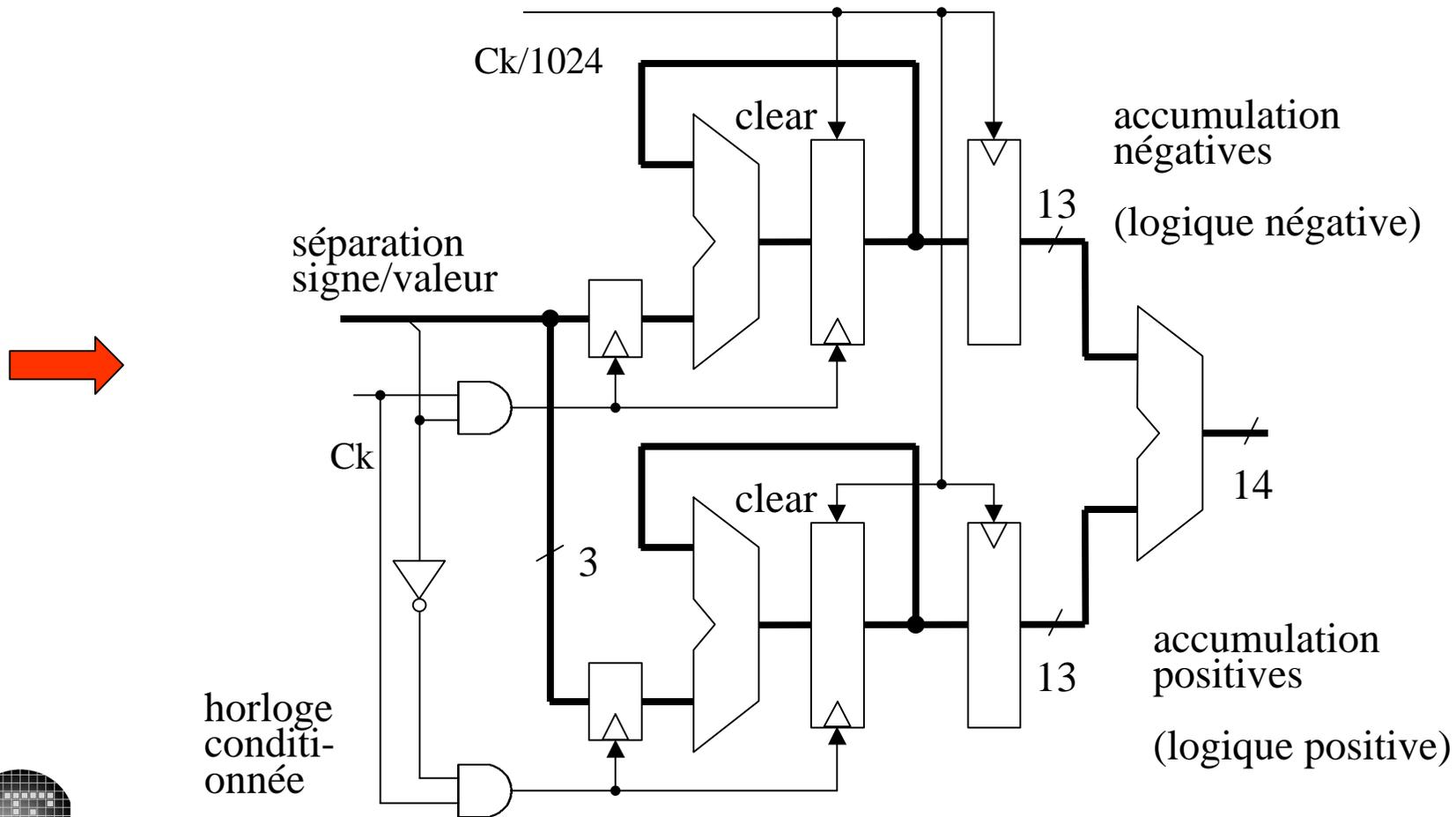


activité très grande si convergence bidirectionnelle vers 0 due au changement de signe de l'accumulateur et à sa propagation sur beaucoup de positions.



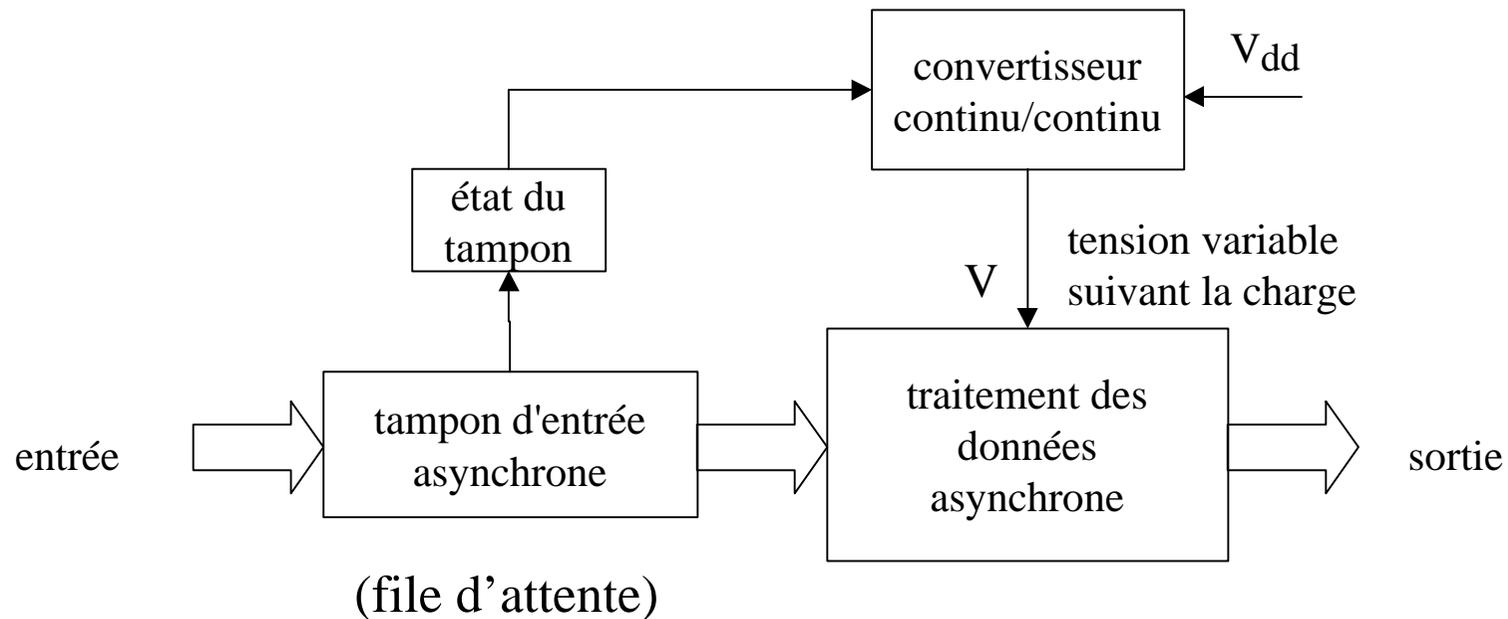
Accumulation de données

Idée: accumuler séparément les échantillons ≥ 0 et les échantillons < 0
 \Rightarrow il n'y a plus de changement de signe.



Contrôle dynamique de tension

Idée: si le tampon d'entrée est plein, augmenter V pour accélérer la cadence
si le tampon d'entrée est vide, diminuer V pour moins consommer

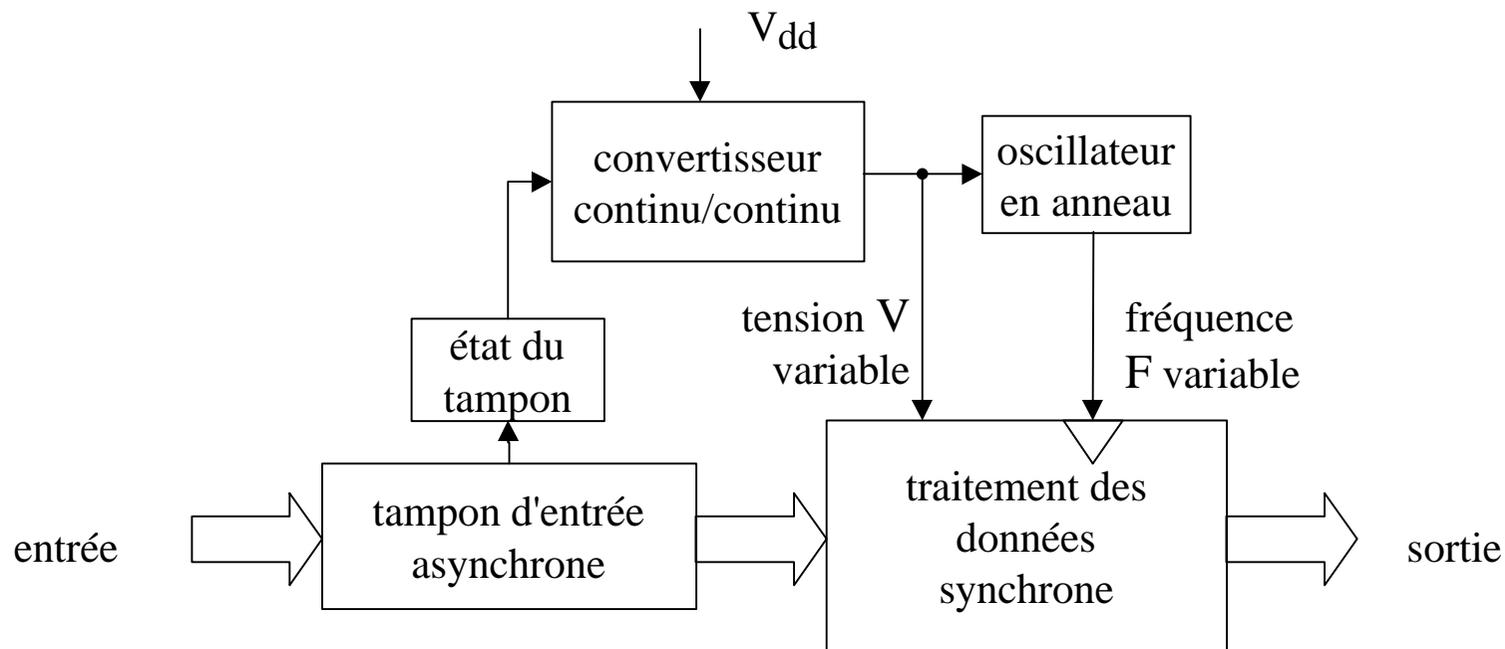


Inconvénient: forte activité due au codage double rail
rendement du convertisseur continu/continu



Contrôle de tension synchrone

Idée: ajuster la fréquence d'horloge F et la tension V en fonction de la charge de sorte que: délai de traitement $\leq 1/F$



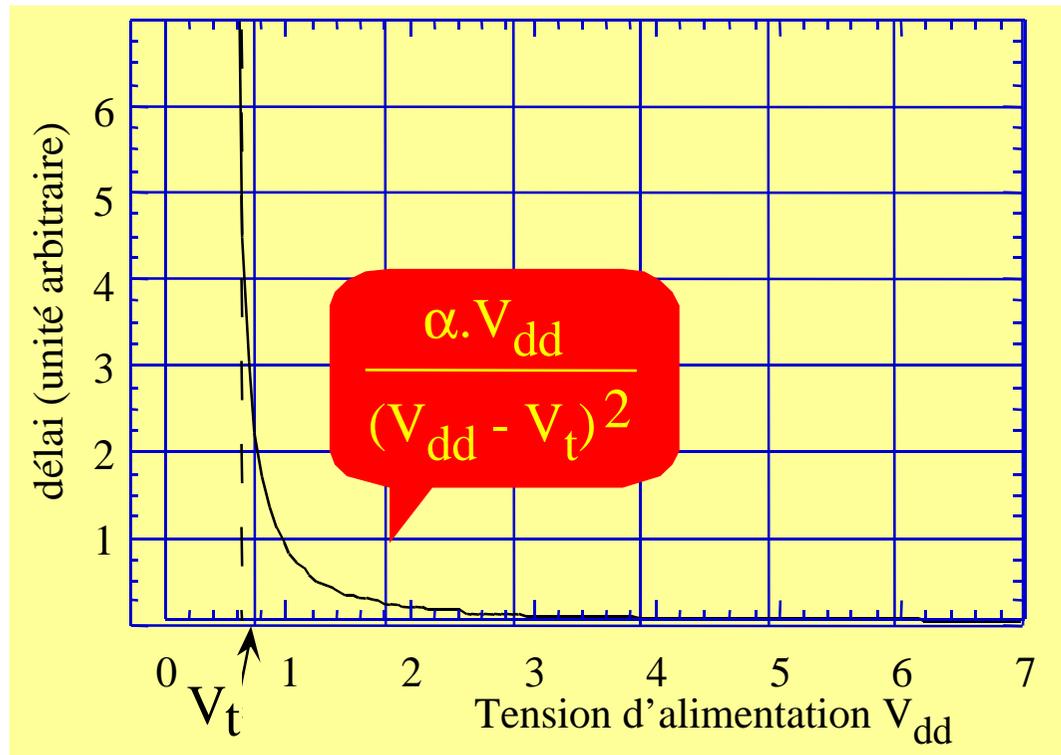
problème: maintenir la période $1/F$ proche du délai



Puissance de traitement

Pas de minimum \Rightarrow Pas de compromis

$$\text{délai} \sim \alpha \cdot V_{dd} / (V_{dd} - V_t)^2$$



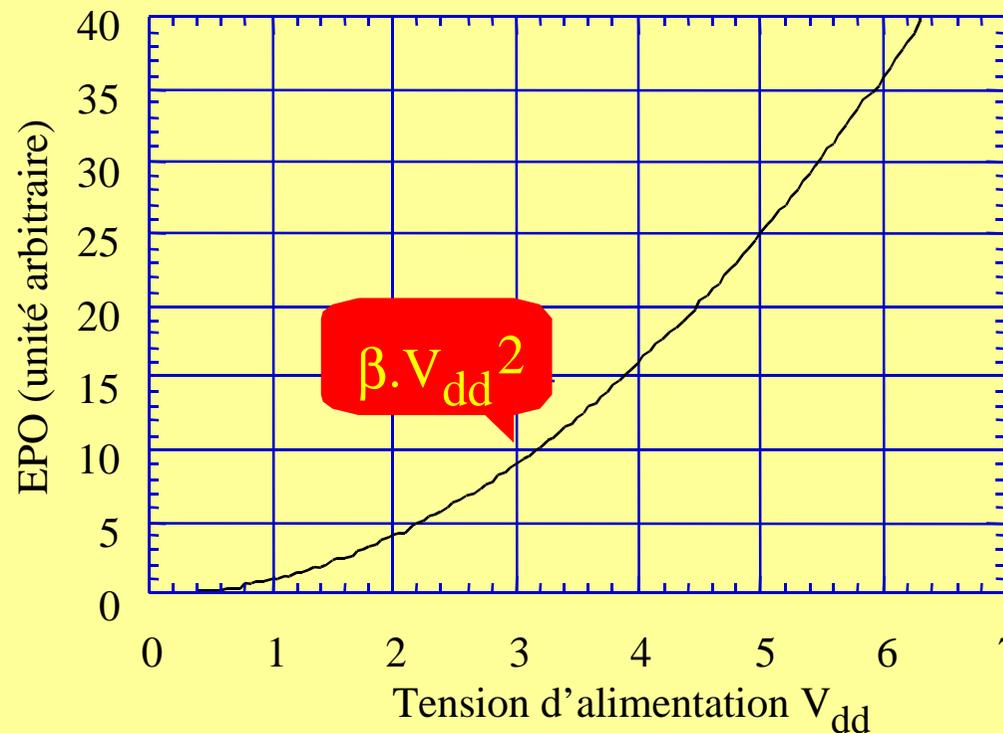
Durée de vie de la batterie

Energie par Opération vs. Tension d'alimentation V_{dd} :
pas de minimum \Rightarrow pas de compromis

$$P = \beta \cdot V_{dd}^2 \cdot f = \beta \cdot V_{dd}^2 / D; \text{ donc, } P \times D = PDP = \beta \cdot V_{dd}^2 = EPO$$

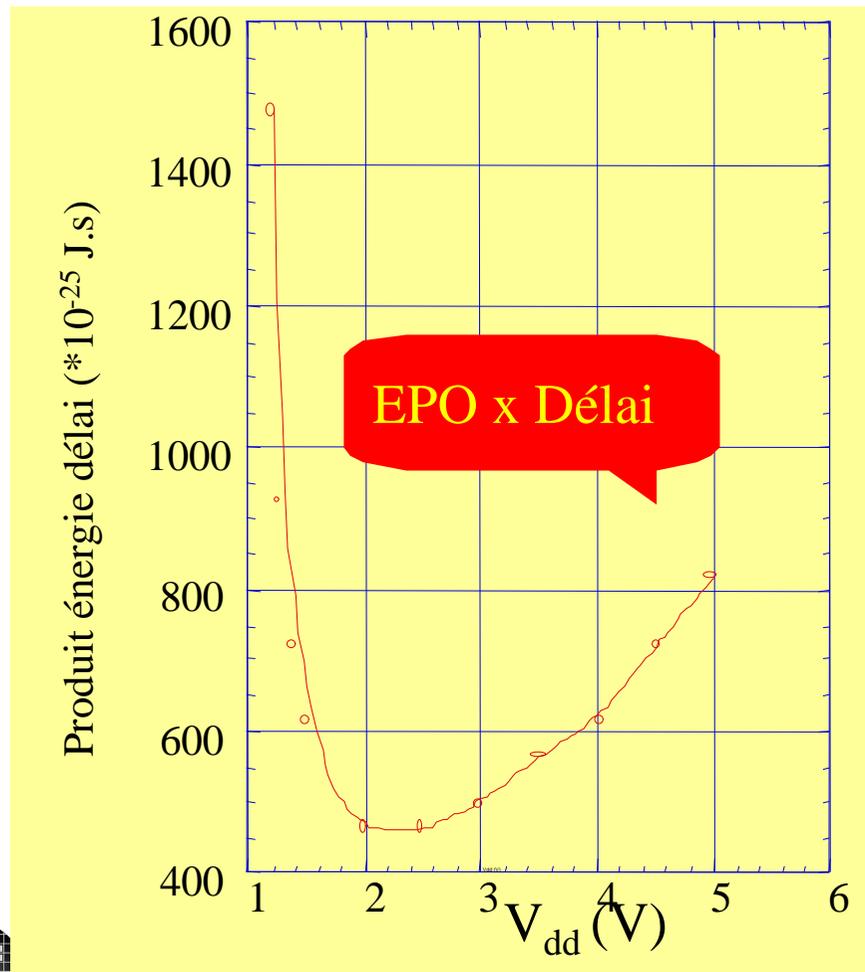


≈ 1 watt*heure



Effacité

EDP = Produit énergie x délai



EDP = EPO x délai \Rightarrow Compromis !!

$$\frac{\partial \text{EDP}}{\partial V_{\text{dd}}} = 0$$

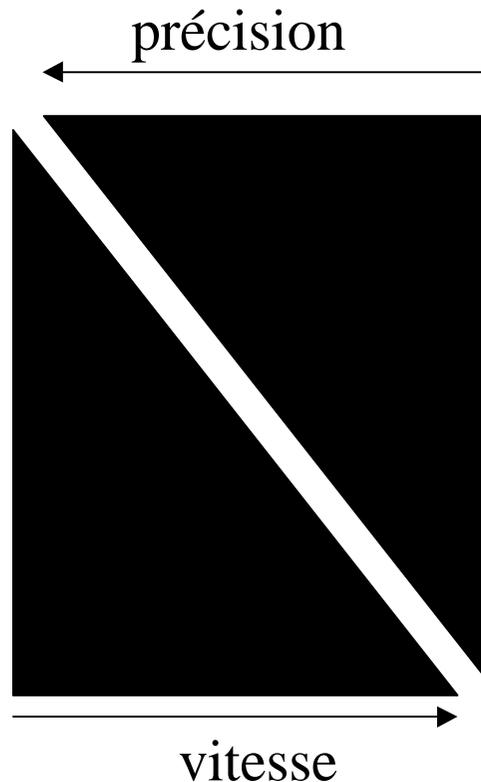
Meilleur EDP

$$V_{\text{dd}} \approx 3V_t$$

Chaîne d'inverseurs ECPD07



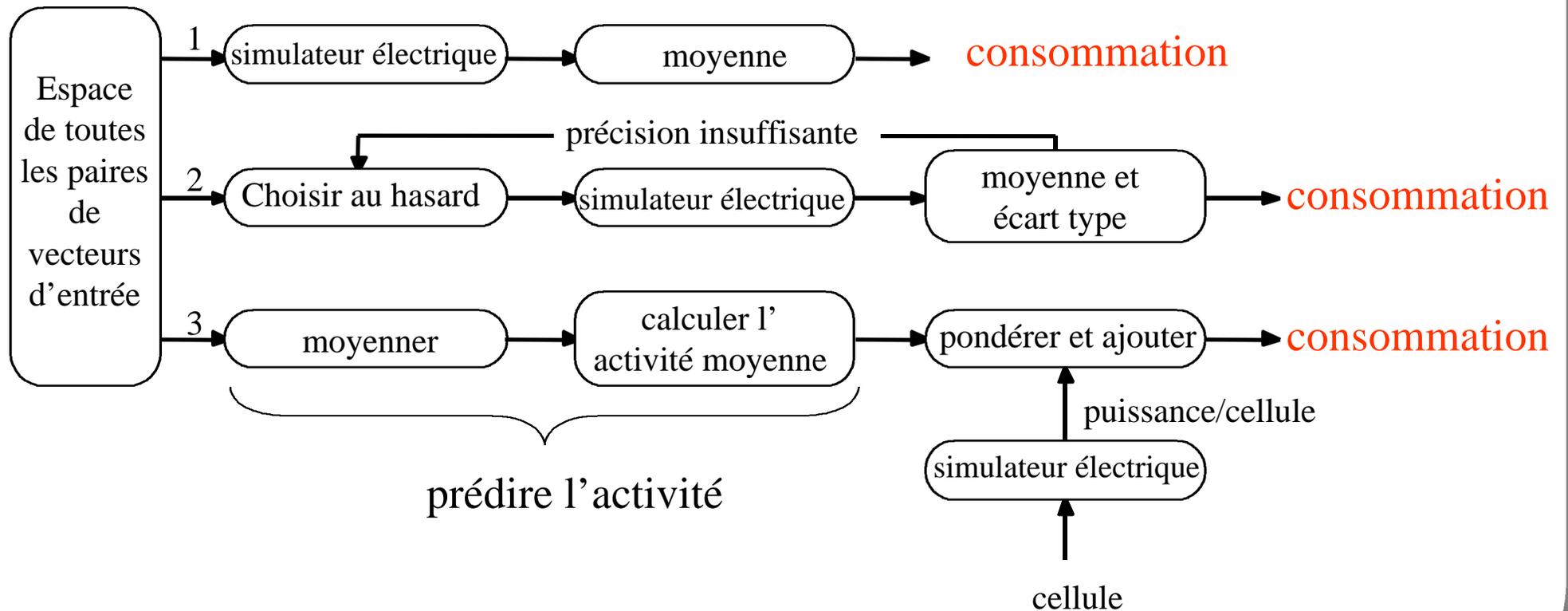
Outils de mesure disponibles



- 1- “Spice” ou “Eldo”
- 2- “PowerMill” (Synopsys, anciennement EPIC)
- 3- Simulateur mixte logique/électrique (e.g. HEAT)
- 4- “DesignPower™” (Synopsys)
“QuickPower™” (Mentor Graphics)
- 5- Simulateurs logiques (VERILOG, BDD, ...)



Trois principes pour prédire la consommation



1- Simulation Exhaustive

1^{ère} approche

- Appliquer toutes les paires différentes de vecteurs d'entrée
- Faire une simulation logique
- Mesurer la puissance $\int I_t * V_{dd} * dt$
- Pondérer par la probabilité de la transition d'entrée
- Applicable seulement à des petits circuits
- Nombre de vecteurs = $2^{\text{nombre d'entrées}}$



2- Simulation Statistique

2^{ème} approche

- Appliquer des vecteurs d'entrée choisis au hasard
- Simuler au niveau électrique
- Mesurer la puissance
- Mettre à jour l'écart (degré de confiance)
- Itérer jusqu'à ce qu'un écart prédéfini soit atteint



Problème de précision: dans les opérateurs arithmétiques, des entrées peu probables causent une dissipation relativement très élevée



3- Modélisation Analytique

3^{ème} approche

- Déduire un modèle analytique de l'activité
- Simulation exhaustive de chaque modèle de cellule (pas les instances)
- Multiplier l'activité de la cellule par l'énergie de cette activité

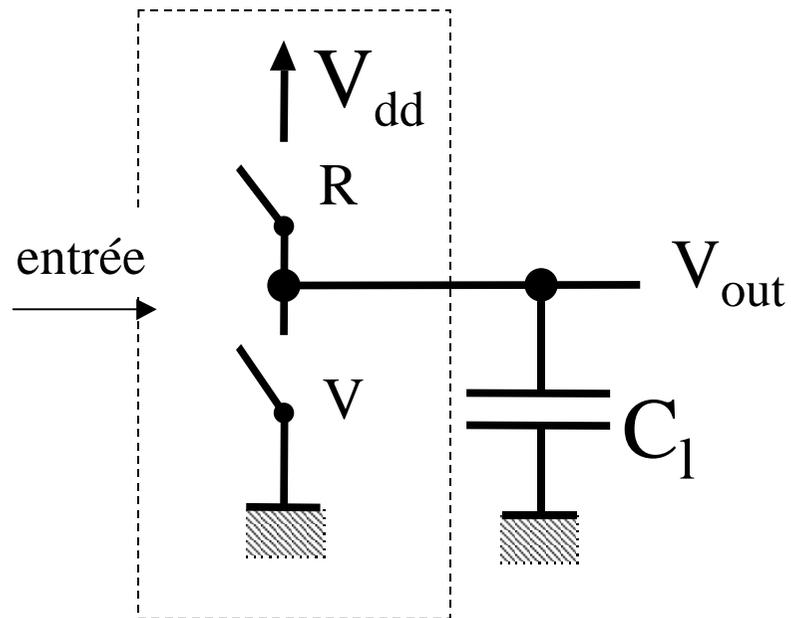


C'est l'approche préférée pour les opérateurs arithmétiques

Modèle → estimateur → outil de synthèses



Transmission sur des bus à forte capacité

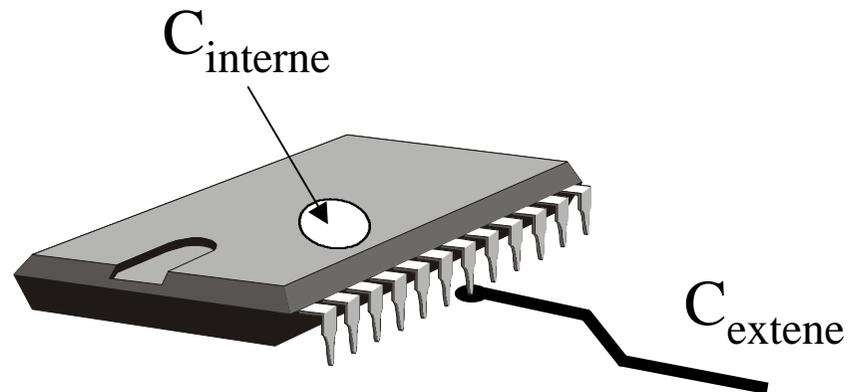


- La capacité parasite des plots d'entrées/sorties est de plusieurs ordres de grandeur supérieure aux capacités internes
- Les entrées/sortie consomment environ la moitié du budget dissipation d'un circuit.
- Diminuer l'activité des entrées/sorties, même au coût d'un codeur augmentant l'activité interne peut être payant.
- Un cache dans le circuit réduit l'activité externe.
- Des modèles ou des mesures statistiques d'activité sont nécessaires.

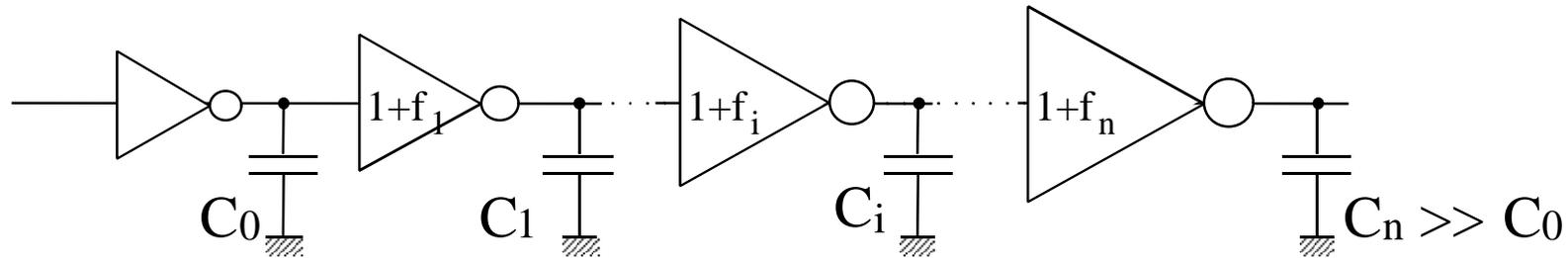


Transmission sur des bus à forte capacité

La capacité parasite d'interconnexion entre boîtiers est de deux ordres de grandeur supérieure à la capacité parasite interne.



Minimisation du délai d'une chaîne



Soit $f_i = \frac{C_{i+1}}{C_i}$. Le délai du $i^{\text{ème}}$ inverseur est $T_i \approx (1 + f_i)T_{\text{mod}}$ (délai d'un inverseur modèle).

On a $\prod_{i=1}^n f_i = \frac{C_n}{C_0}$; on veut minimiser le délai total $\sum_{i=1}^n T_i$ proportionnel à $\sum_{i=1}^n (1 + f_i)$.

Le minimum est obtenu lorsque les $f_i \forall i$ sont égaux. Alors $f_i = \sqrt[n]{\frac{C_n}{C_0}}$ soit $n = \frac{\log\left(\frac{C_n}{C_0}\right)}{\log(f_i)}$.

Alors le délai total = $n(1 + f_i) = \log\left(\frac{C_n}{C_0}\right) \frac{1 + f_i}{\log(f_i)}$.

Le minimum de $\frac{1 + f_i}{\log(f_i)}$ est obtenu pour $f_i \approx 3,5$ donc $n = \left\lceil 1,838 * \log\left(\frac{C_n}{C_0}\right) \right\rceil$.



Minimisation de la dissipation d'une chaîne

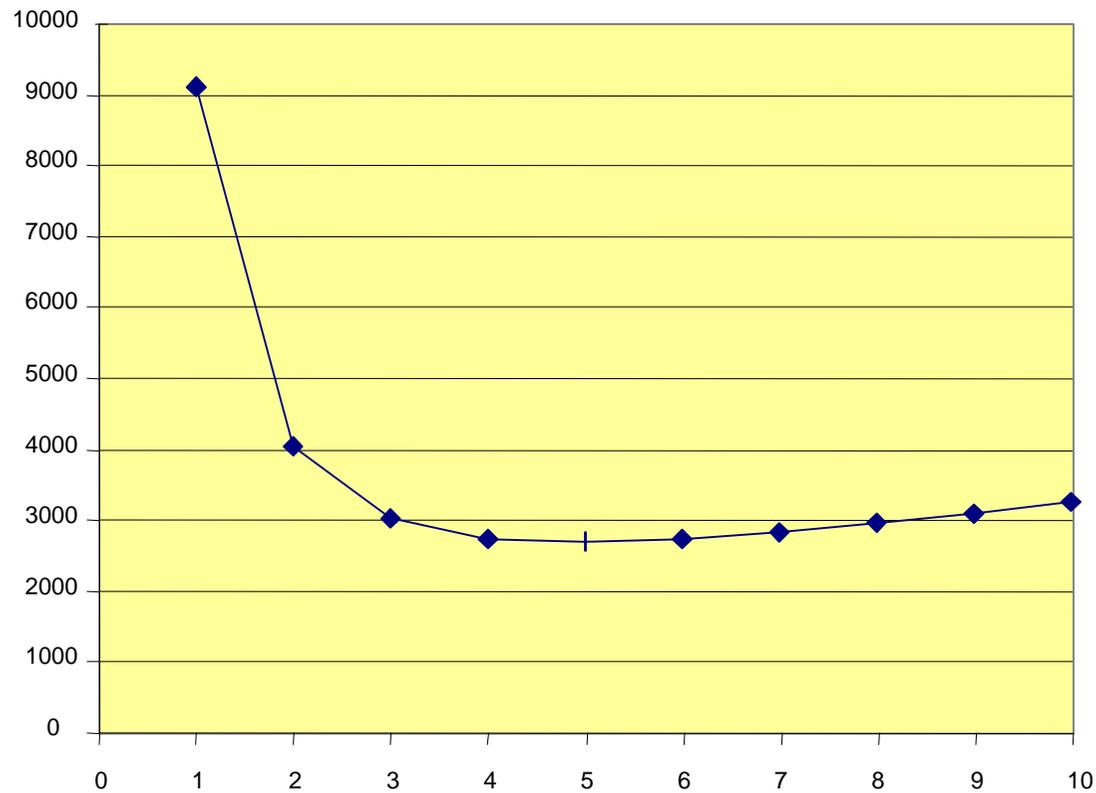
- On se fixe C_0 , C_n et un délai T_{total} (supérieur au délai minimum), et on veut minimiser l'énergie dissipée tout en respectant le délai.
- On ne peut pas jouer sur le V_{dd} car on veut maintenir l'excursion, ni sur le débit.
- Reste à minimiser la capacité parasite $\sum_{i=1}^{n-1} C_i$.
- Il faut trouver le plus petit n tel que $n \left(1 + \sqrt[n]{\frac{C_n}{C_0}} \right) T_{\text{mod}} \leq T_{\text{total}}$.

Application numérique: $C_0 = 5\text{fF}$, $C_n = 10\text{pF}$, $T_{\text{total}} = 3\text{ns}$, $T_{\text{mod}} = 100\text{ps}$.



Minimisation de la dissipation d'une chaîne

Application numérique: $C_0 = 5\text{fF}$, $C_n = 10\text{pF}$, $T_{\text{total}} = 3\text{ns}$, $T_{\text{mod}} = 100\text{ps}$.

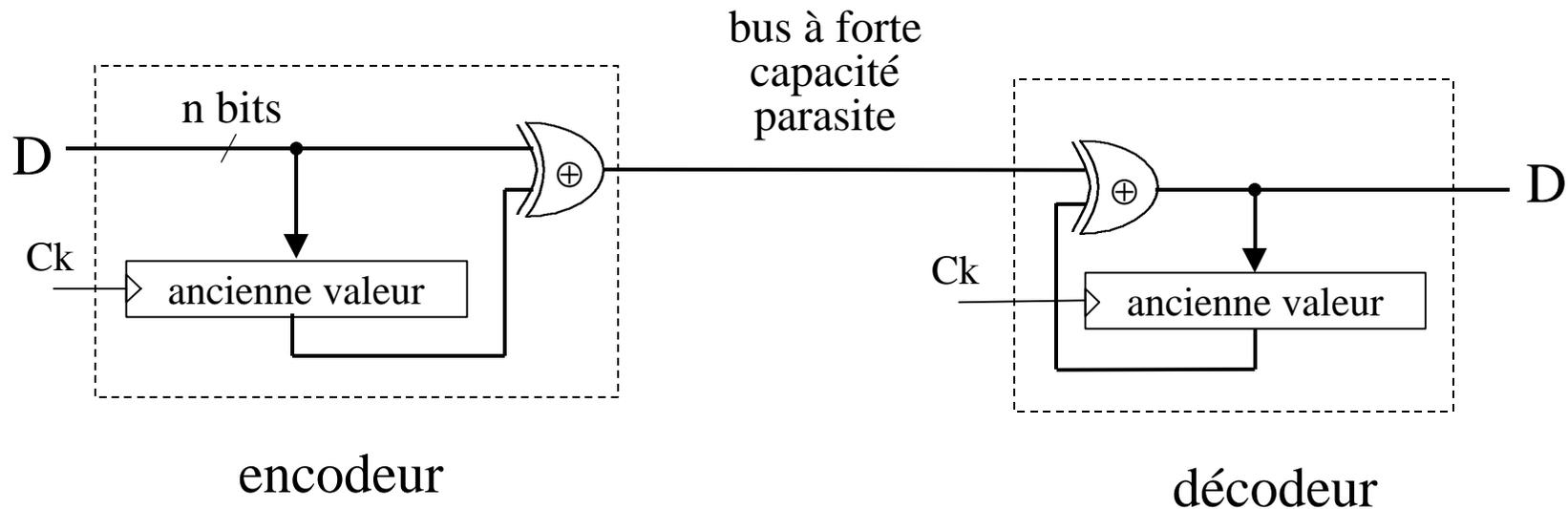


évolution de $n * (1 + \sqrt[3]{2000}) * 100\text{ps}$



Transmission des changements

Idée: ne transmettre que les changements de valeur de bits.
L'activité est réduite s'il y a une forte corrélation temporelle des valeurs (exemple images).

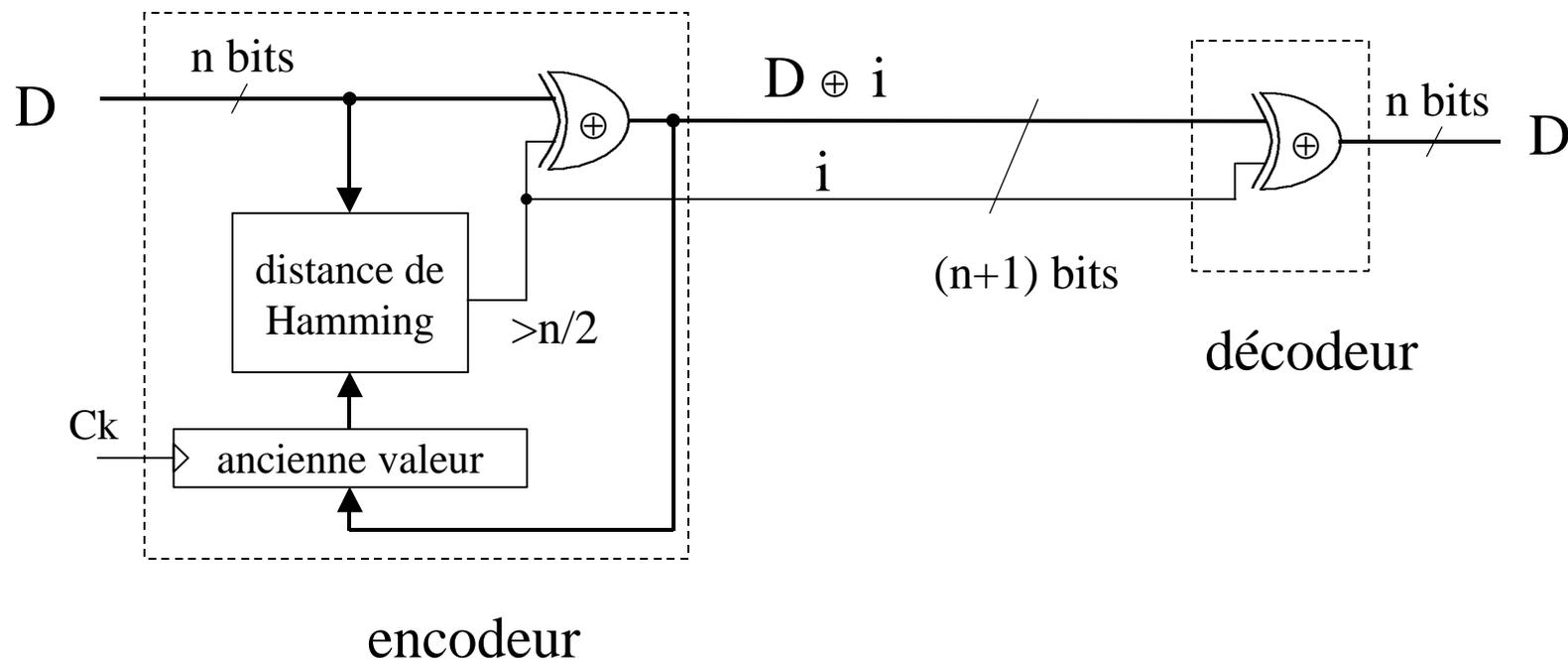


"ancienne valeur" est initialisé à 0 des deux côtés.

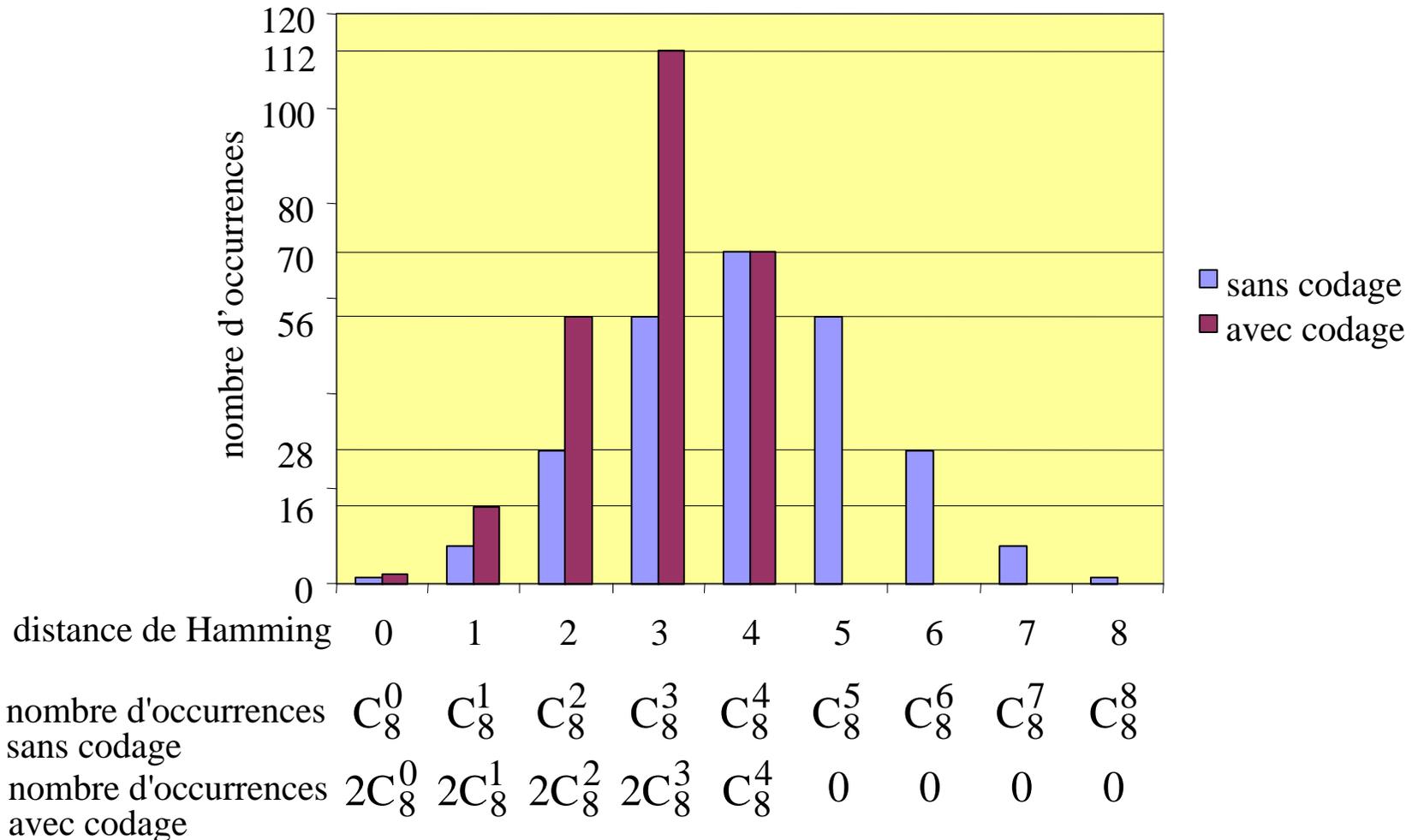


Codage de bus inversé

Idée: choisir entre la valeur D et la valeur $\neg D$ celle qui est la plus proche de la valeur transmise au cycle précédent et la transmettre



Répartition d'activité de bus de 8 bits



activité moyenne sans codage = 4

avec codage $(744+93)/256 = 3,2695$

gain 18%



Réduction d'activité de bus de n bits

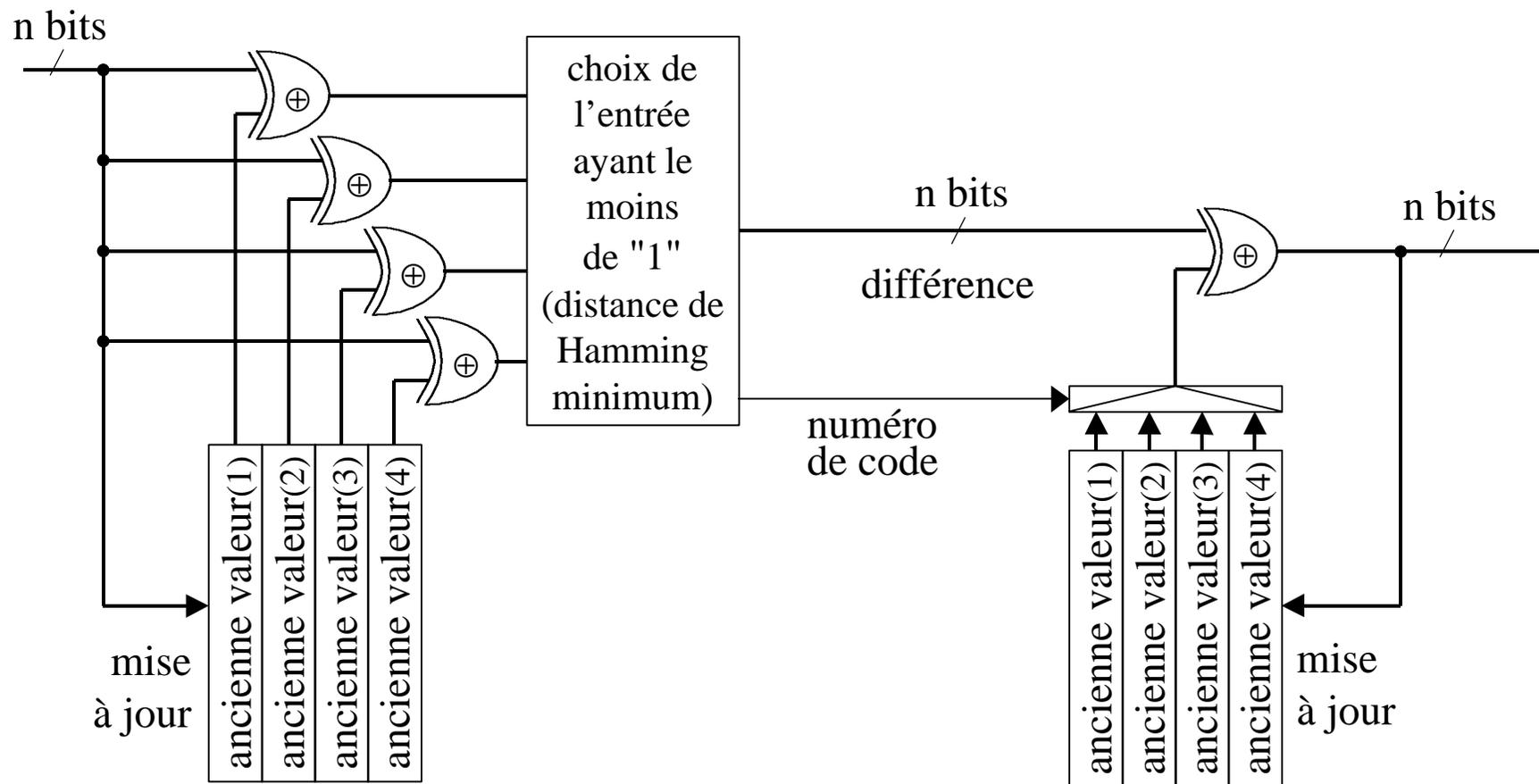
$$\text{activité sans codage} = \sum_{i=0}^{n-1} i * C_n^i = \sum_{i=0}^{\frac{n-1}{2}} (i + n - i) * C_n^i = n * \frac{2^n}{2}$$

$$\text{activité avec codage} = \sum_{i=0}^{\frac{n-1}{2}} (i + i) * C_n^i = 2 \sum_{i=0}^{\frac{n-1}{2}} i * C_n^i$$



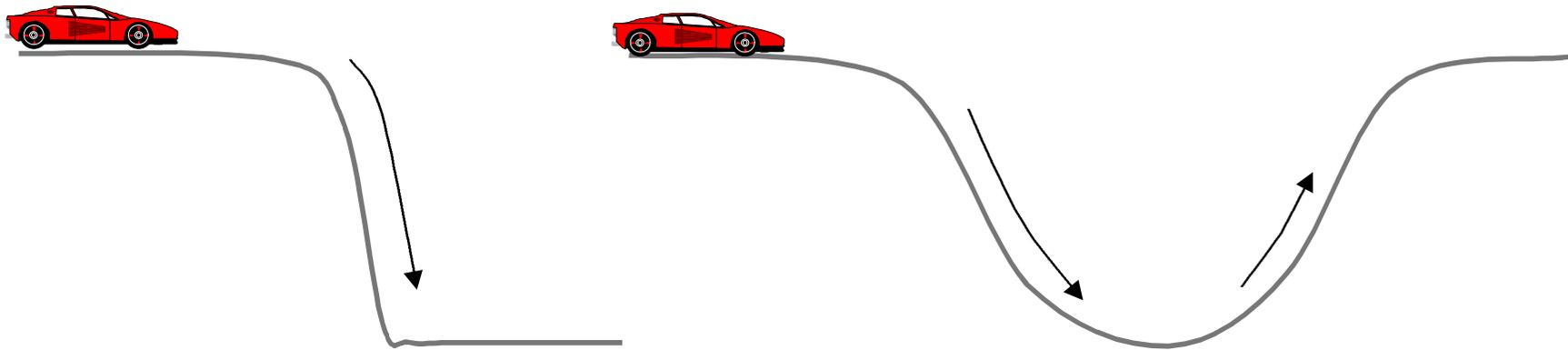
Utilisation d'un "livre de codes"

Idée: choisir parmi plusieurs "anciennes valeurs" celle qui est la plus proche.
C'est le "code" dont on ne transmet que le "numéro de code".



Récupération d'énergie

Idée: récupérer de l'énergie à la descente pour aider à la remontée au lieu de transformer toute cette énergie en chaleur

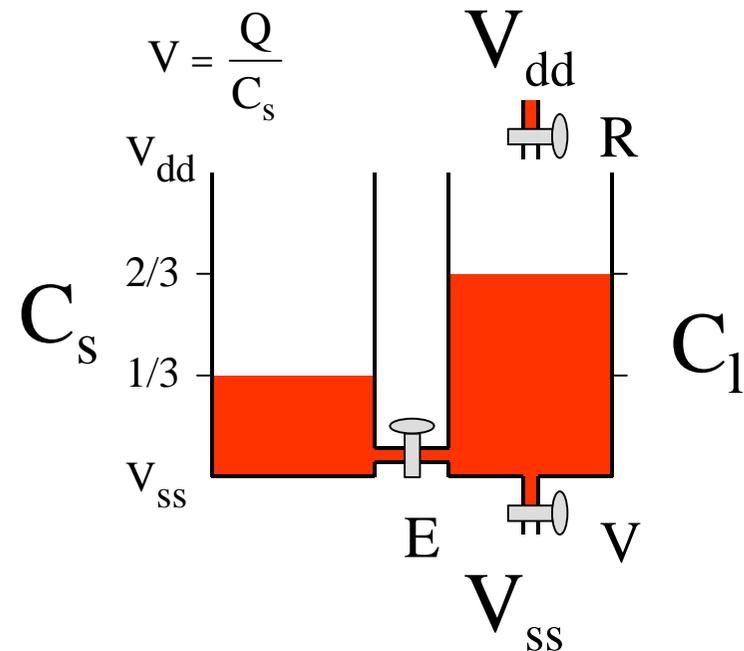
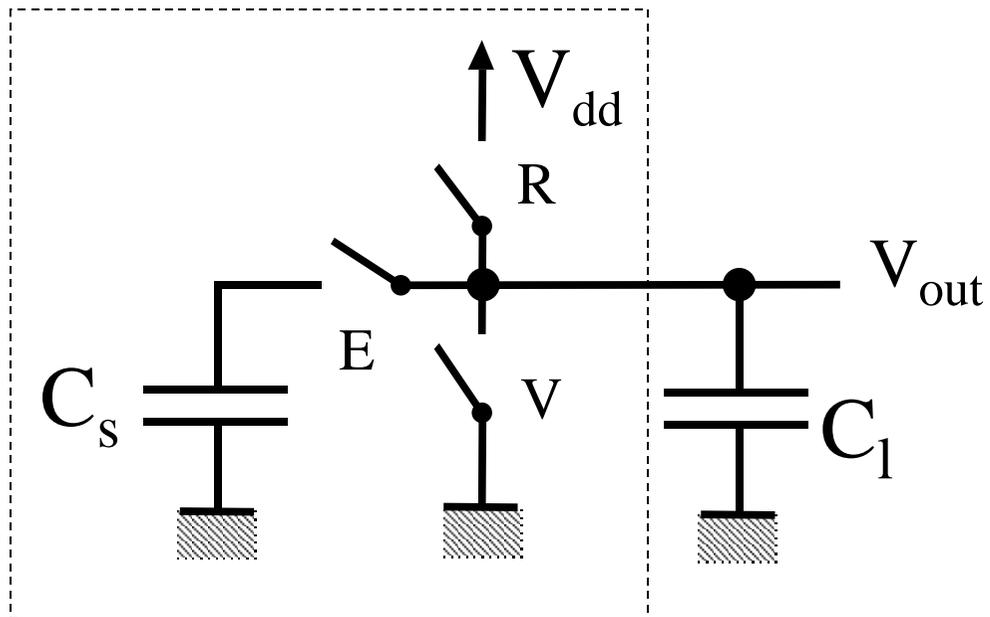


Remarque: on récupère d'autant plus qu'on va plus lentement

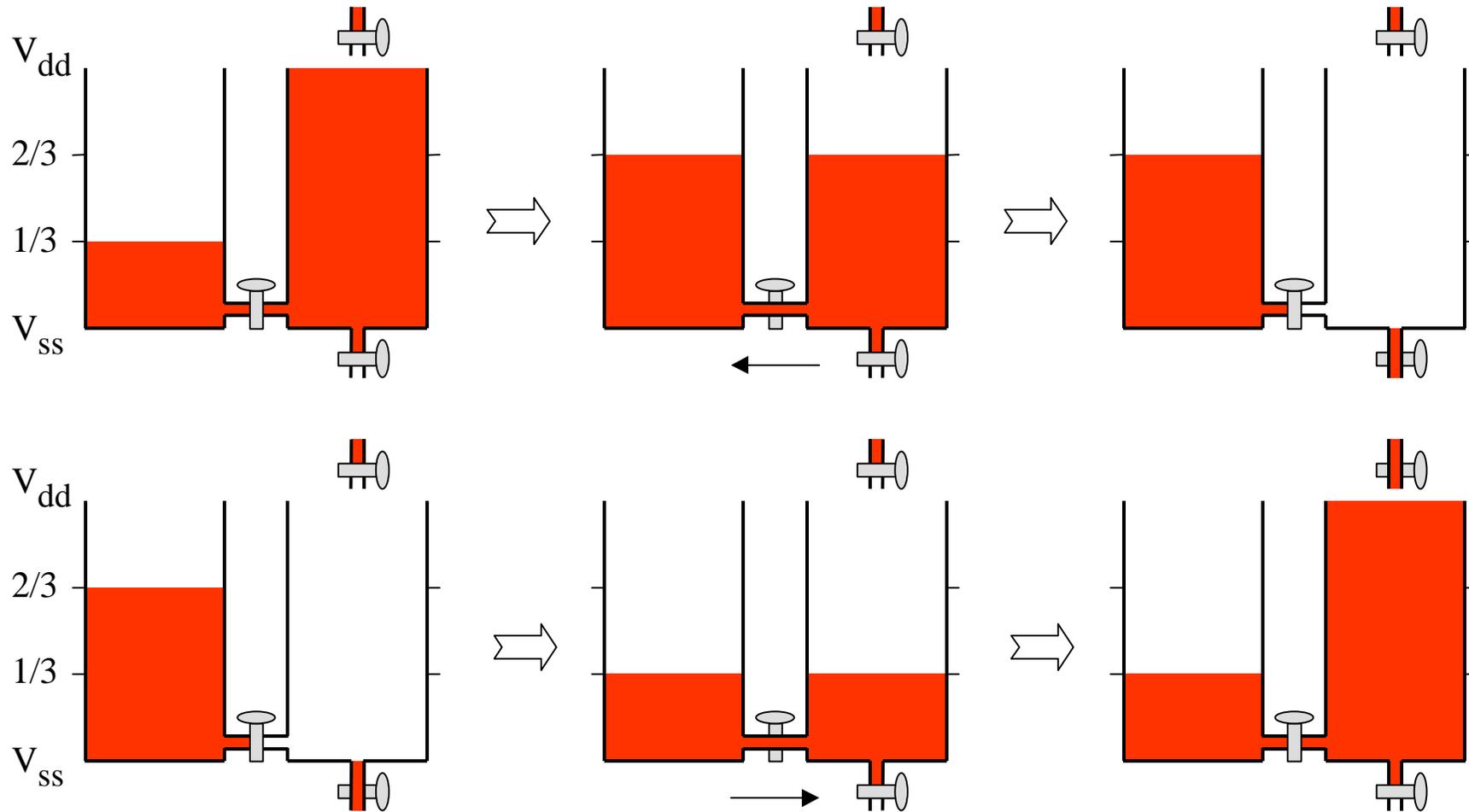


Circuit de récupération d'énergie

Idée: on ajoute une capacité de récupération C_s dans laquelle on va puiser ou stocker de l'énergie

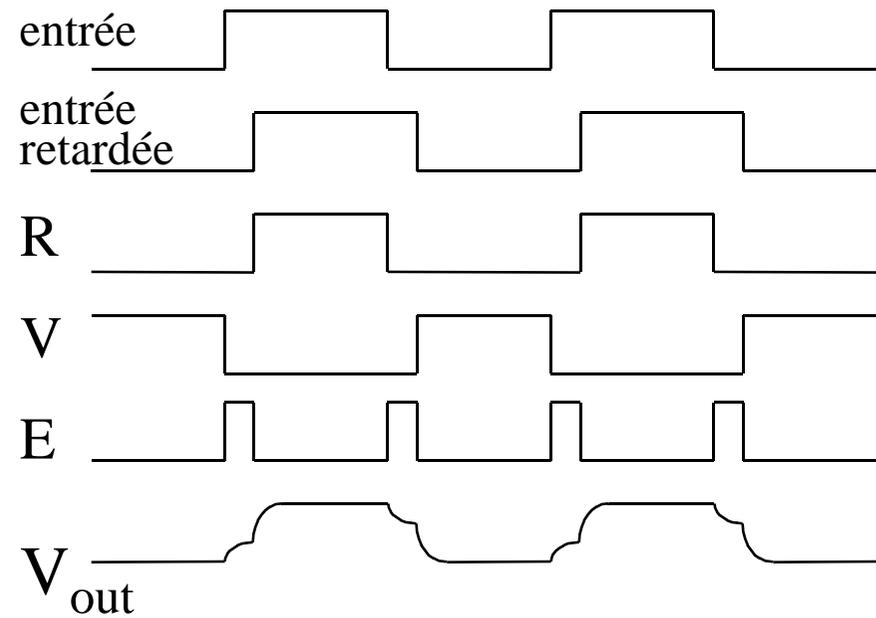
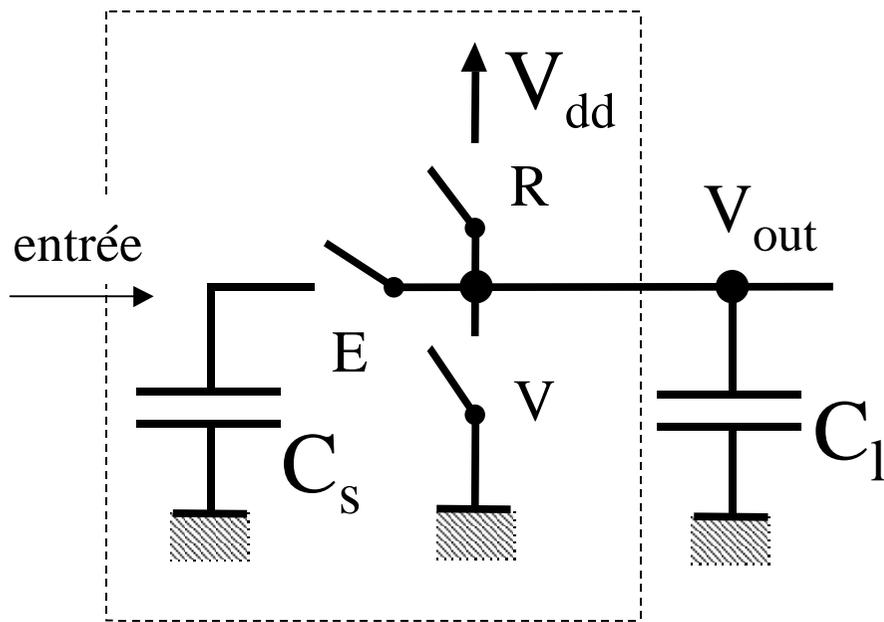


Cycle de récupération d'énergie



Contrôle du plot à récupération d'énergie

Chronogramme des signaux de contrôle



Technologie CMOS SOI

80 à
300 nm

couche mince de silicium
oxyde de silicium
substrat (silicium monocristallin)

Avantages

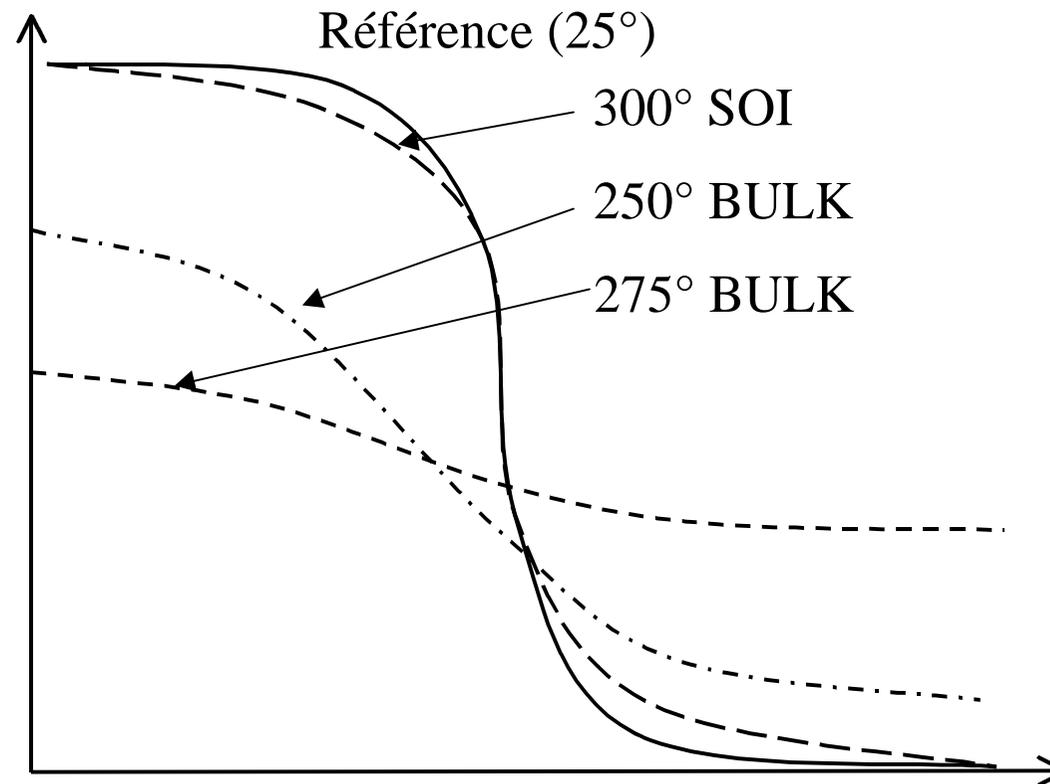
- pas de courant de fuite de jonction
- capacités parasites divisées par 2 à 3
- résistance rayonnement ionisant
- pas de "latch-up »
- résistance à la température (300°)

Inconvénients

- coût de la tranche
- règles de dessin à changer



Influence de la température SOI/BULK



Caractéristiques de transfert d'un inverseur



Conclusion

- Les moyens de réduire la dissipation de puissance sont nombreux
- Pour être utilisée, une stratégie doit
 - avoir peu d'impact sur le flot de conception traditionnel
 - permettre une réduction substantielle
- Qu'est-ce qui sera utilisé ?
 - réduction de V_{dd}
 - contrôle/réduction de l'activité
 - horloge conditionnée "gated clock"
 - représentation des nombres
 - cellule standard basse consommation (surtout latches)
 - seuil multiple V_t
 - transition à faible excursion logique
 - plots de sortie à basse consommation



Références

A. Chandarakasan and R. W. Brodersen, "Low power Digital CMOS Design", Kluwer Academic Publishers, Boston, 1995, ISBN : 0-7923-9576-X

A. Amara, Proc. of the "Journées d'études Faible Tension Faible Consommation" (FTFC'97), Paris, Nov. 1997

S. Cristoloveanu and J. Boussey, Proc. of the "MIGAS summer school on Low Power, Low Voltage Integrated Circuits : Technology and Design", Microelectronics Engineering, Vol. 39, Elsevier, Dec. 1997

J. M. Rabaey and M. Pedram, Low Power Design Methodologies, Kluwer Academic Publishers, Boston, 1996, ISBN : 0-7923-9630-8

W. Nebel and J. Mermet eds, Low Power Design in Deep Submicron Electronics, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1996, ISBN : 0-7923-4569-X



Table des matières

1- Conception pour la Faible Consommation	23- Evolution des technologies et dissipation	48- Activité d'un compteur	72- Efficacité
2- Plan du cours	24- Réduction de la dissipation	49- Compteur de Gray	73- Outils de mesure disponibles
3- Introduction	25- Ajustement de la taille des transistors	50- Compteur semi-Gray	74- Trois principes pour prédire la consommation
4- Introduction 2	26- Tensions d'alimentation multiples	51- Taxonomie des Transition	75- Simulation Exhaustive
5- Traiter de l'information dissipe de l'énergie	27- Réduction des capacités d'interconnexions	52- Transition utile ou redondante	76- Simulation Statistique
6- Où chercher à réduire la dissipation	28- Parallélisme	53- Activité moyenne	77- Modélisation Analytique
7- Pourquoi délai et dissipation	29- Chronogramme de blocs parallèles	54- Activité de l'additionneur à propagation	78- Transmission sur des bus à forte capacité
8- Energie dans les capacités parasites	30- Pipelining	55- Activité pire cas d'un additionneur	79- Transmission sur des bus à forte capacité
9- Energies stockée et dissipée	31- Chronogramme de blocs "Pipeliné"	56- Modèle analytique de l'activité du RCA(1)	80- Minimisation du délai d'une chaîne
10- Seuil de commutation de l'Inverseur	32- Une idée plus précise de l'activité	57- Modèle analytique de l'activité du RCA(2)	81- Minimisation de la dissipation d'une chaîne
11- Courant de court circuit de l'Inverseur	33- Probabilité des nœuds	58- Modèle analytique de l'activité du RCA(3)	82- Minimisation de la dissipation d'une chaîne
12- Energie de court circuit de l'Inverseur	34- Probabilité de sortie d'une porte NOR	59- Réduction de l'activité parasite	83- Transmission des changements
13- Effet de la charge de sortie C_1	35- Activité moyenne des nœuds	60- Comparaison d'activité d'additionneurs	84- Codage de bus inversé
14- Effet de la pente sur le court circuit	36- Transition consommante	61- Adaptation dynamique	85- Répartition d'activité de bus de 8 bits
15- Composantes de la dissipation en CMOS	37- Activité d'une porte NOR	62- Ajustement dynamique de la précision	86- Réduction d'activité de bus de n bits
16- Quelques ordres de grandeur	38- Activité d'une porte XOR	63- Ajustement du chemin de données	87- Utilisation d'un "livre de codes"
17- Moyens de réduire la dissipation	39- Limites du modèle	64- Prédiction: exemple la comparaison	88- Récupération d'énergie
18- Réduction de l'alimentation V_{dd}	40- Propagation des probabilités	65- Généralisation de la prédiction	89- Circuit de récupération d'énergie
19- Réduction du seuil V_t	41- Propagation des activités	66- Accumulation de données	90- Cycle de récupération d'énergie
20- Courants de fuite	42- Réordonnancement des entrées	67- Accumulation de données	91- Contrôle du plot à récupération d'énergie
21- Evolution des technologies et dissipation	43- Synthèse pour la consommation	68- Contrôle dynamique de tension	92- Technologie CMOS SOI
22- Evolution des technologies et dissipation	44- Redondance pour réduire l'activité	69- Contrôle de tension synchrone	93- Influence de la température SOI/BULK
	45- Reconstituons les transitions	70- Puissance de traitement	94- Conclusion
	46- Bilan du nombre moyen de transitions	71- Durée de vie de la batterie	
	47- Compteur binaire		

