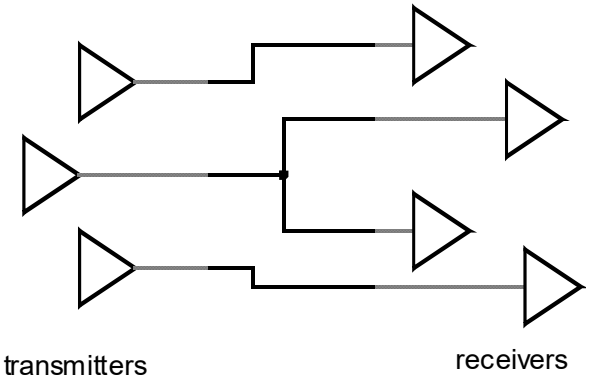
Diritti d'autore

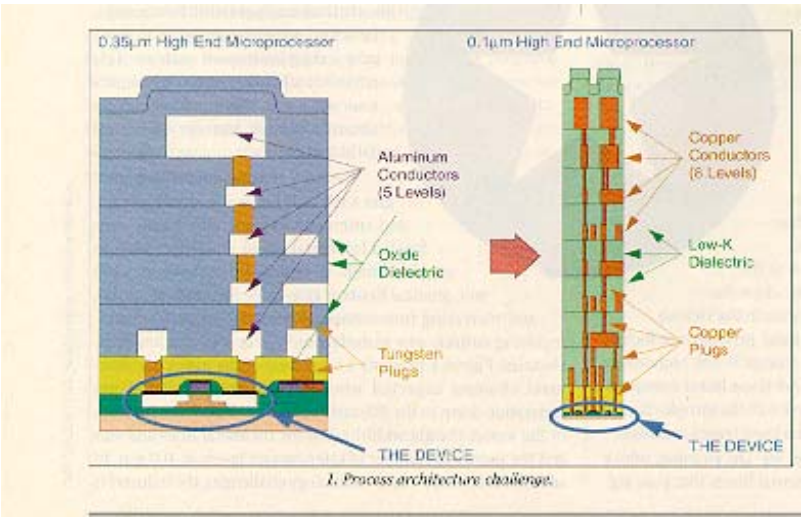
INTERCONNESSIONI

Interconnessioni



schematics

physical

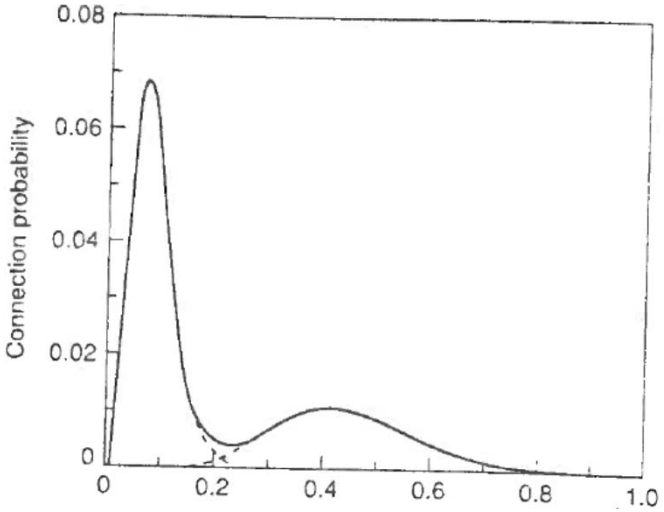


cross-section

- Sono linee di materiale conduttore (metallo, polisilicio fortemente drogato con comportamento quasi metallico) che collegano tra loro diversi nodi del circuito e diverse porte logiche.
- In un tipico SoC si sviluppano per molti chilometri e sono distribuite su molti livelli (piani) che nel loro complesso formano il cosiddetto Back End Of Line (BEOL)
- Il numero di livelli è andato storicamente crescendo con l'evolvere delle tecnologie.
- L'occupazione di area delle interconnessioni contribuisce in modo fondamentale all'area complessiva del chip.
- I consumo di potenza per il pilotaggio delle interconnessioni rappresenta spesso oltre l'80% del consumo totale

Interconnessioni locali e globali

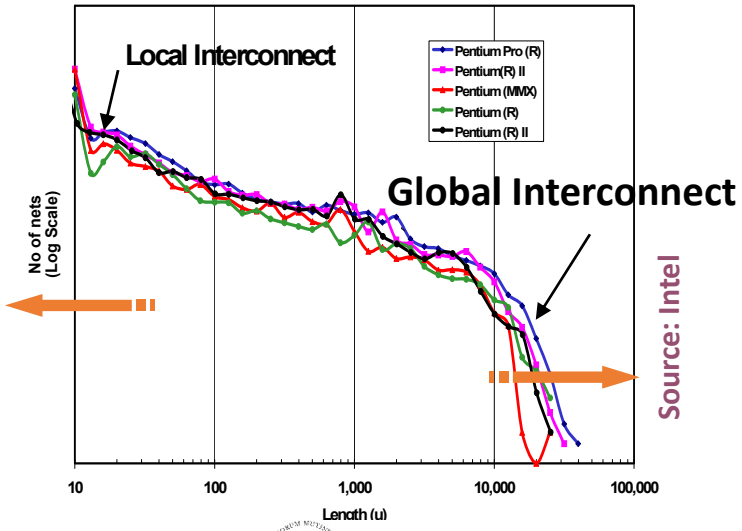
Distribuzione della lunghezza delle linee di interconnessione in un chip (idealized, lin.scale - top; realistic, log scale - bottom)



Normalized wire length $L_{en}/sqrt(A_{chip})$

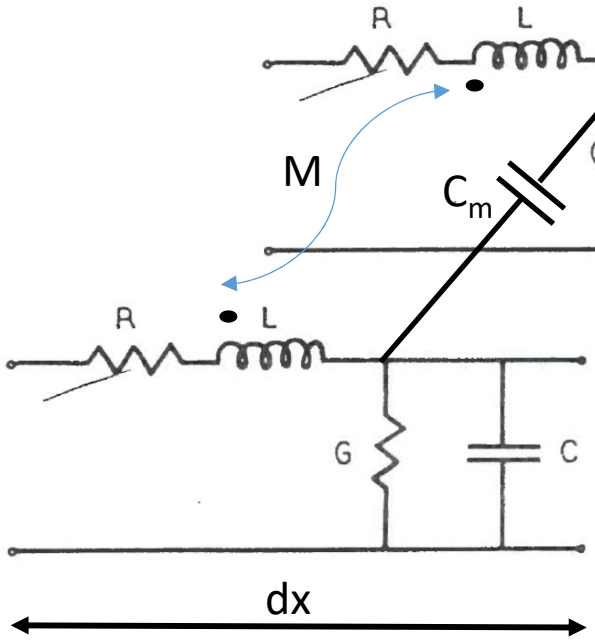
Interc.locali: tante ma di piccola lunghezza (connessioni tra porte logiche e celle standard)

Interc.globali: poche ma di elevata lunghezza (bus di comunicazione, distribuzione del clock, VDD, GND). Len è circa proporzionale alla radice quadrata dell'area del chip



Interconnessioni

Ciascun tratto di lunghezza infinitesima *dx* di due linee accoppiate può essere schematizzato attraverso un circuito equivalente come quello in figura. Il modello si presta ad essere esteso facilmente al caso di più linee.



R=r*dx [Ohm]
G=g*dx [Siemens]
C=c*dx [Farad]
L=l*dx [Henry]
M=m*dx [Henry]
C_m=c_m*dx [Farad]

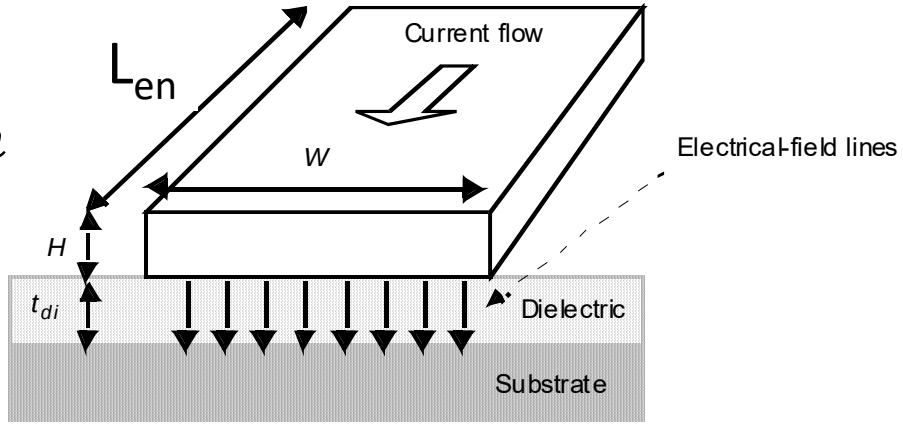
r = resistance per unit length [Ohm/m] , c = capacitance per unit length [Farad/m]
g = conductance per unit length [Siemens/m] ,
l = inductance per unit length [Henry/m] , m=mutual inductance per unit length [Henry/m]
c_m=coupling capacitance per unit length [Farad/m]

Interconnessioni: calcolo parametri

$$C_W = \frac{\varepsilon_{di}}{t_{di}} W L_{en} = \frac{\varepsilon_0 \varepsilon_r}{t_{di}} W L_{en}$$

$$R_W = \frac{\rho}{H\, W} L_{en}$$

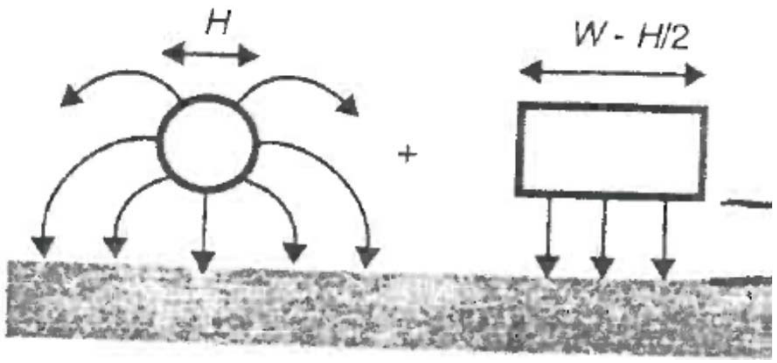
$$C_W R_W = \frac{\rho}{H} \frac{\varepsilon_{di}}{t_{di}} L_{en}^2$$

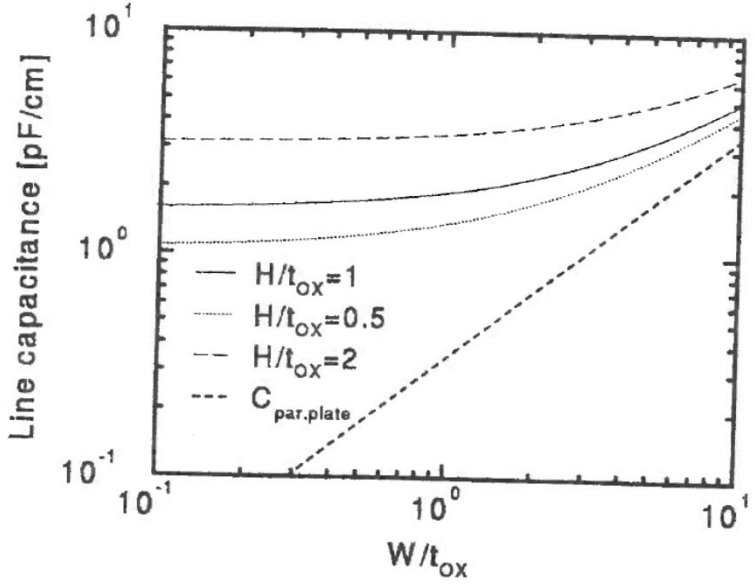


- Requisiti: bassa rho e bassa epsilon (low-k materials o aria)
- Il ritardo introdotto dall'interconnessione tende a rimanere constante con lo scaling mentre quello di propagazione dei gate tende a calare → problema

Material	ε_r
Free space	1
Aerogels	~1.5
Polyimides (organic)	3-4
Silicon dioxide	3.9
Glass-epoxy (PC board)	5
Silicon Nitride (Si_3N_4)	7.5
Alumina (package)	9.5
Silicon	11.7

Interconnessioni: capacità di fringing





$$C_W = C_{pp} + C_{fr} = \frac{\varepsilon_{di}}{t_{di}} W L_{en} + 2\pi \frac{\varepsilon_{di}}{\ln(1 + 4t_{di}/H)} L_{en}$$

Contributo facce piane parallele

Contributo di fringing

Interconnessioni: capacità totale linee debolmente accoppiate

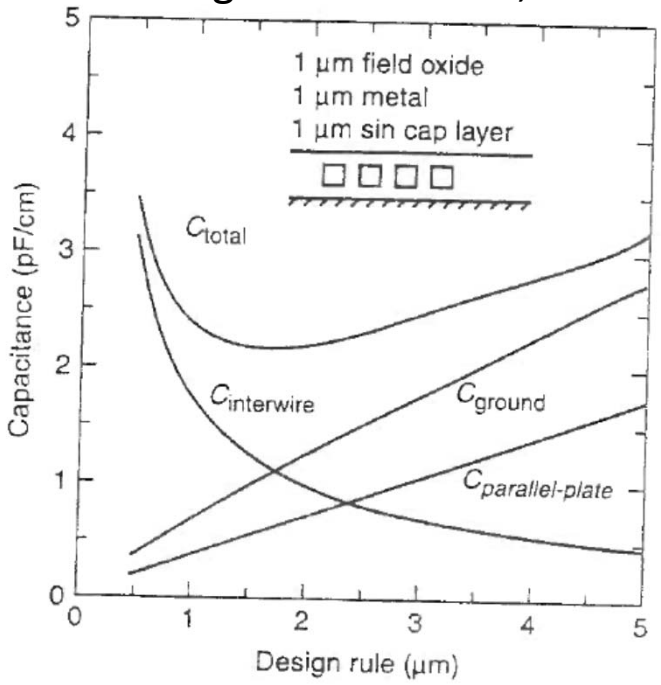
$$C_W = \frac{\varepsilon_{di}}{t_{di}} W L_{en} + 2\pi \frac{\varepsilon_{di}}{\ln(1 + 4t_{di}/H)} L_{en} + \frac{\varepsilon_{di}}{S} H L_{en}$$

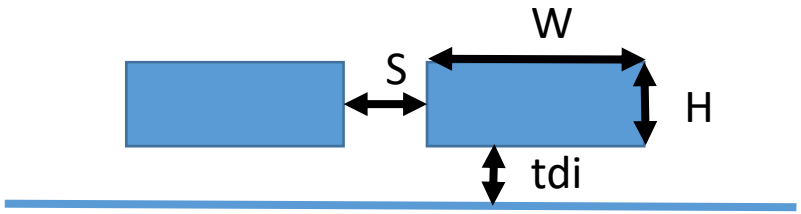
Contributo facce piane parallele

Contributo di fringing

Contributo accoppiamento laterale

Scaling ad H e tdi fissi, W=S



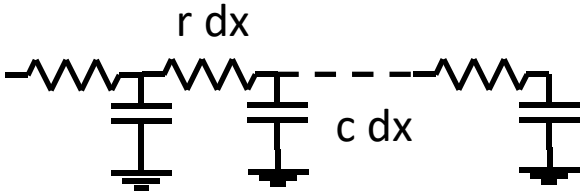


Semplificazione del modello

- Solitamente G è trascurabile.
- L è anche trascurabile a patto che la lunghezza d'onda dei segnali sulla linea λ >> lunghezza complessiva della linea L_{en}. Questo corrisponde a ritenere la cadute sulle induttanze trascurabili rispetto alle cadute di tensione sulle resistenze

$$V_L = l\,dx\,\frac{\partial I}{\partial t} \approx l\,dx\,\frac{I_0}{t_r} \ll r\,dx\,I_0 \text{ che implica } l \ll r\tau$$

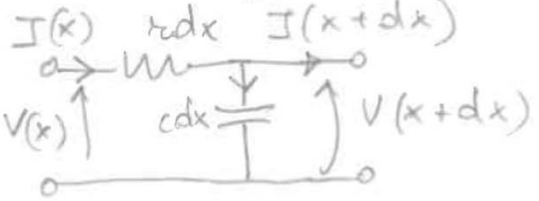
- Questa condizione è normalmente verificata per interconnessioni interne ai chip dove R è elevata in quanto W e H hanno dimensioni molto piccole.
- Per una linea isolata M e C_m sono nulle
- Sotto queste ipotesi ogni tratto di linea può essere rappresentato come un tratto di resistenza in serie seguito da una capacità verso massa.



$$c = \frac{C_W}{L_{en}} \qquad r = \frac{R_W}{L_{en}}$$

Modello a parametri distribuiti

- Se la resistenza della linea è significativa non è possibile schematizzare tutte le capacità con un unica capacità concentrata C_L. Occorre ricorrere ad un modello distribuito.



$$I(x+dx) - I(x) = -c\frac{\partial V}{\partial t}dx$$

$$V(x+dx) - V(x) = -r\,I\,dx$$

$$\frac{\partial V}{\partial x} = -rI \qquad \frac{\partial I}{\partial x} = -c\frac{\partial V}{\partial t}$$

$$\frac{\partial^2 v}{dx^2} = rc\frac{\partial v}{dt}$$

Si tratta della ben nota equazione di diffusione la cui soluzione per una linea infinita a partire da condizioni iniziali tutte nulle è

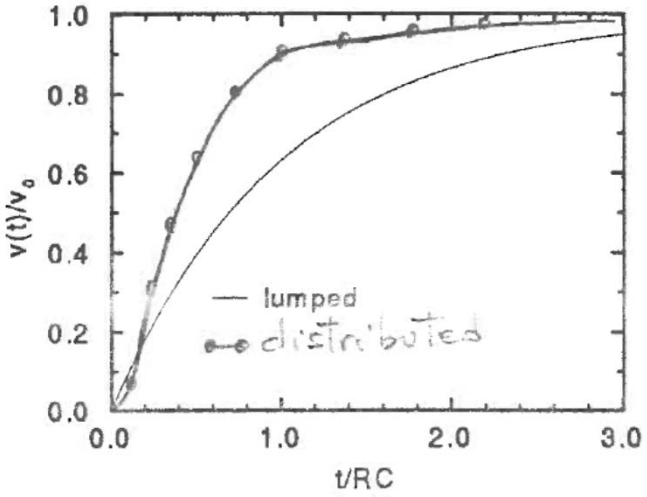
$$V(L_{en}, t) = V_0 erfc\left(\sqrt{\frac{R_W C_W}{4t}}\right)$$

Se invece approssimiamo la linea con una sola resistenza e capacità abbiamo

$$V(L_{en}, t) = V_0\left(1 - \exp\left(\frac{t}{R_W C_W}\right)\right)$$

Modello a parametri distribuiti

- Confronto tra soluzione a parametri distribuiti e a parametri concentrati

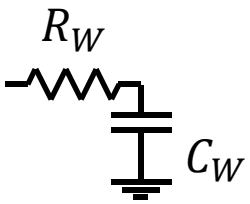


Range %	time (distr.) [RC]	time (lumped) [RC]
0-90	1.0	2.3
10-90	0.9	2.2
0-63	0.5	1.0
0-50	0.4	0.7
0-10	0.1	0.1

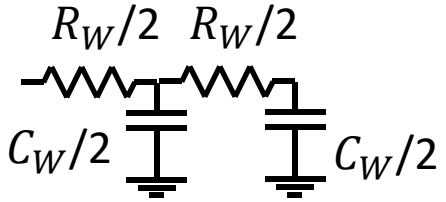
- Il comportamento di una linea distribuita implica una risposta più rapida che un modello a parametri concentrati a parità di resistenza e capacità totali della linea
- L'implementazione di modelli distribuiti è tipicamente piuttosto onerosa dal punti di vista computazionale. E' opportuno trovare soluzioni approssimate

Approssimazioni a parametri concentrati

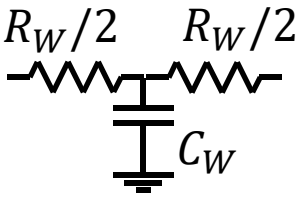
- Approssimazioni della linea con un numero finito di celle di tipo L, π o T costituite da elementi concentrati
- Una cella ad L



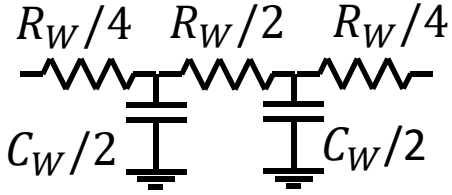
Due celle a L



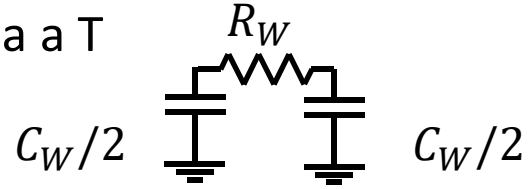
- Una cella a π



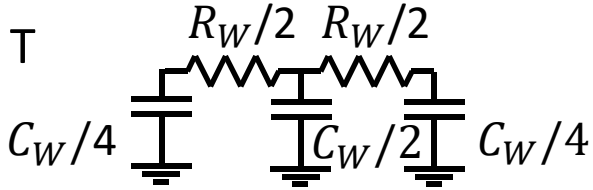
Due celle a π



- Una cella a T



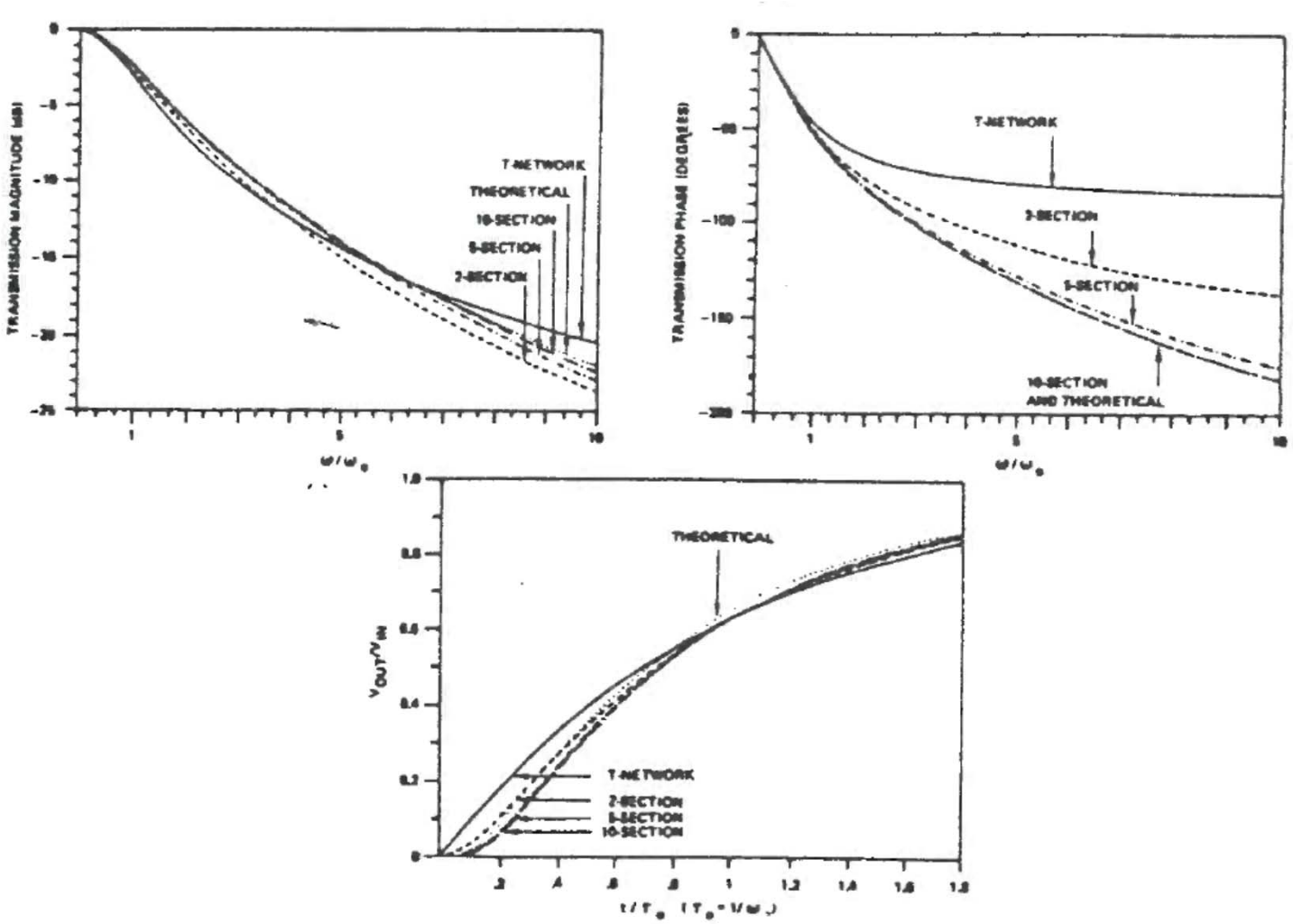
Due celle a T



- Si possono utilizzare più celle uguali per una singola interconnessione a patto di scegliere coerentemente i valori delle resistenze e capacità
- Per segnali digitali in cui conta solo l'ampiezza della tensione l'accuratezza del modello converge rapidamente al caso distribuito (3 celle → 3% errore)

Modello «distribuito» vs. «concentrato»

- Accuratezza del modello a più celle di parametri concentrati



Tempo di propagazione del segnale in interconnessioni

- Il modello a parametri concentrati della linea comprende numerose capacità. → non esiste un'unica costante di tempo. La risposta temporale è la sovrapposizione di esponenziali, ciascuno con la sua costante di tempo.
- Ai fini del calcolo del ritardo possiamo ipotizzare che esista una costante di tempo dominante (quella di durata maggiore)
- L'espressione della constate di tempo dominante è data da

$$\tau_D = \sum_{k=1}^{N} R_{sk}\, Ck = \sum_{k=1}^{N} R_k\, Cke$$

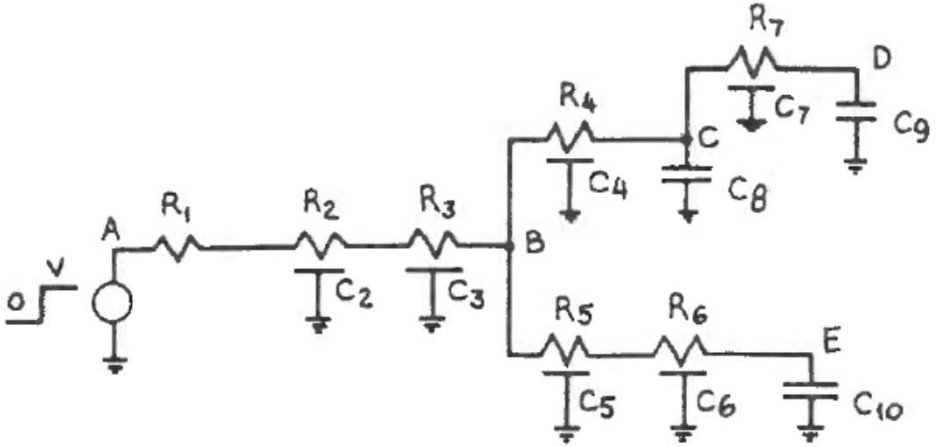
dove R_{sk} è la resistenza totale tra la sorgente del segnale e il nodo k, C_{ke} è la capacità totale tra il nodo k e l'ultimo nodo (N) della linea

- Nel caso di una linea uniforme di celle ad L abbiamo R_k=r Δx e C_k=c Δx

$$\tau_D = \sum_{k=1}^{N} kR_k\, Ck = rc\Delta x^2 \sum_{k=1}^{N} k = rc\Delta x^2 \frac{N(N+1)}{2} \rightarrow rc\frac{L_{en}^2}{2}$$

coerentemente con l'espressione del ritaro 0-63% fornita nella tabella precedente

Tempo di propagazione del segnale in interconnessioni



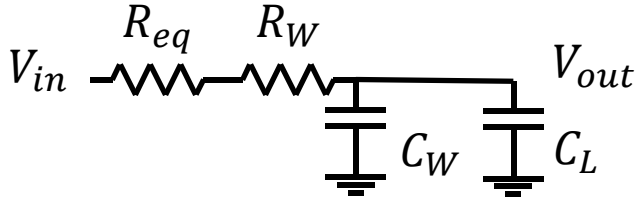
- **Alberi RC:** Il contributo di ciascun ramo deve essere sommato per trovare il ritardo dell'albero.

$$\tau_{A-D} = \tau_{A-B} + \tau_{B-D}$$
$$\tau_{A-E} = \tau_{A-B} + \tau_{B-E}$$

dove τ_{A-B}, τ_{B-D} e τ_{B-E} sono i ritardi di Elmore dei rami $A-B$, $B-D$ e $B-E$, rispettivamente.

Tempo di salita di linea e driver

- Nel caso in cui la linea sia pilotata da una porta logica driver il segnale al suo ingresso in generale non è assimilabile ad un gradino di tensione.
- Al fine di stimare il l'effetto del driver sul ritardo del segnale fino al carico consideriamo un semplice modello RC ad L per la linea e schematizziamo il driver tramite la sua resistenza equivalente Req.



- La costante di tempo del circuito vale: $\boldsymbol{\tau} = (R_{eq} + R_W)(C_W + C_L)$
- Il tempo di salita/discesa al 90% vale:

$$t = 2.3(R_{eq} + R_W)(C_W + C_L) = 2.3\left(R_{eq}C_W + R_{eq}C_L + R_W C_W + R_W C_L\right)$$

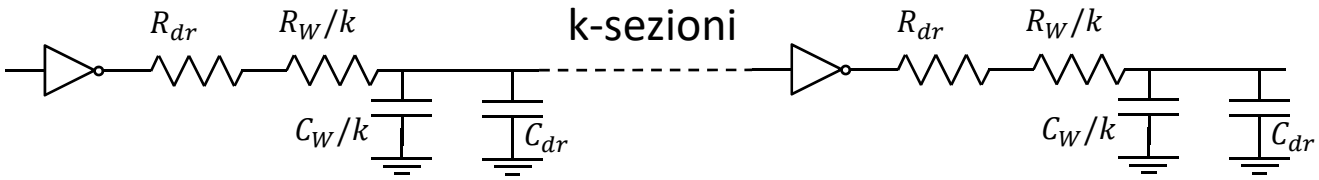
- Riconosciamo che il termine $2.3R_W C_W$ nella rappresenta il ritardo intrinseco della linea. Se calcolato accuratamente con il modello distribuito esso sarebbe pari a solo $1.0R_W C_W$. Decidiamo allora di corregere empiricamente l'espressione del ritardo per evitare di sovrastimarlo eccessivamente

$$t = 2.3\left(R_{eq}C_W + R_{eq}C_L + R_W C_L\right) + R_W C_W$$

- Analogamente per transistori al 50% usando coefficienti 1.0 e 0.5 come da tabella

Ottimizzazione del ritardo

- Il ritardo intrinseco di una interconnessione può essere molto elevato. La capacità dell'interconnessione può essere eccessiva per il driver.
- Per migliorare la situazione inseriamo ripetitori (invertitori o buffer non invertenti) lungo la linea e ne ottimizziamo il numero

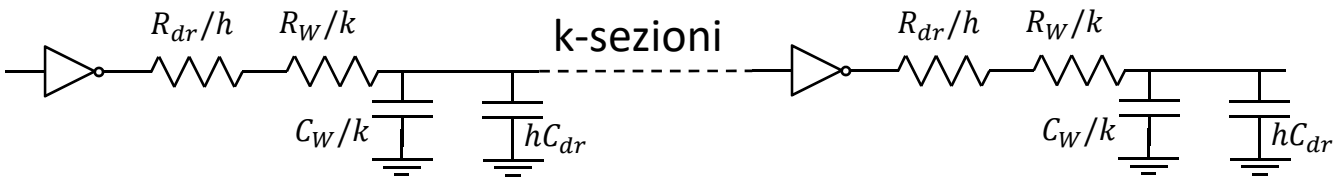


- Utilizzo di *k r*ipetitori identici di area minima, resistenza R_{dr} e capacità di ingresso C_{dr}. Si ha

$$t_{0-90} = k[2.3\left(\frac{R_{dr}C_W}{k} + R_{dr}C_{dr} + \frac{R_W C_{dr}}{k}\right) + \frac{R_W C_W}{k^2}]$$

- Imponendo dt/dk=0 (minimo) $k = \sqrt{R_W C_W / 2.3 R_{dr} C_{dr}}$
- Il ritardo ottimo vale $t_{0-90} = \left(\sqrt{2.3 R_{dr} C_W} + \sqrt{2.3 R_W C_{dr}}\right)^2$
- Analogamente per transistori al 50% usando coefficienti 1.0 e 0.5 come da tabella

Ottimizzazione del ritardo



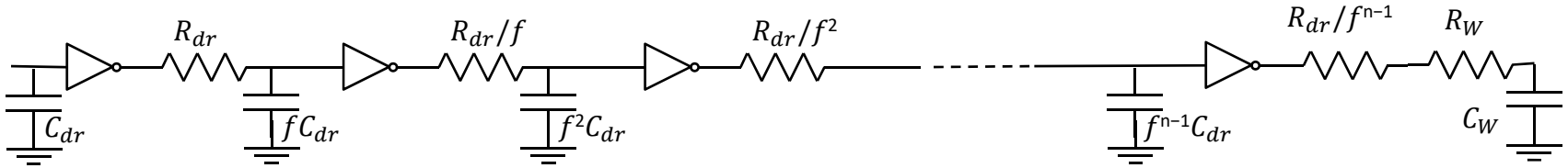
- **Utilizzo di *k r*ipetitori identici ottimizzati** con fattore di forma h. Si ha

$$t_{0-90} = k[2.3\left(\frac{R_{dr}C_W}{hk} + R_{dr}C_{dr} + \frac{R_W h C_{dr}}{k}\right) + \frac{R_W C_W}{k^2}]$$

- Imponendo dt/dk=0 e dt/dh=h0 (minimo) abbiamo

$k = \sqrt{R_W C_W / 2.3 R_{dr} C_{dr}}$ $h = \sqrt{R_{dr} C_W / R_W C_{dr}}$

- Il ritardo ottimo vale $t_{0-90} = 7.6\sqrt{R_{dr}C_{dr}R_W C_W}$ è inferiore al caso precedente



- **Utilizzo di buffer a cascata** di «n» invertitori di dimensioni crescenti («f»)
- Imponendo dt/dn=0 e dt/df=0 (minimo) abbiamo

$f = \mathrm{e} \approx 2.7$ $n = \ln(C_W/C_{dr})$

- Il ritardo ottimo vale $t_{0-90} = 2.3\,\mathrm{e}\,R_{dr}C_{dr} + R_W C_W$

Ottimizzazione del ritardo