Vorlesung 3

Themen dieser Vorlesung sind

- Schaltgeschwindigkeit des Inverters
- Kapazitäten
- AC-Stromverbrauch
- Taktbaum

Schaltgeschwindigkeit des Inverters

Versuchen wir die Geschwindigkeit eines Inverters zu berechnen:

Wir nehmen an, dass wir am Ausgang des Inverters eine kapazitive Last haben (Abbildung 1). Die Annahme ist auch, dass wir am Eingang einen unendlich schnellen Impuls haben – die Eingangsspannung ändert sich von GND auf VDD. In dem Fall wird NMOS momentan eingeschaltet und PMOS ausgeschaltet. (Abbildung 2)

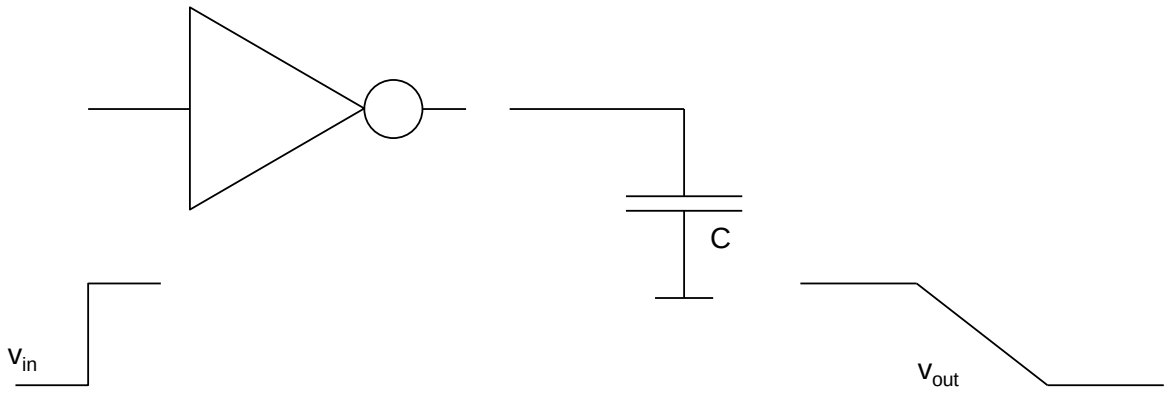

Abbildung 1

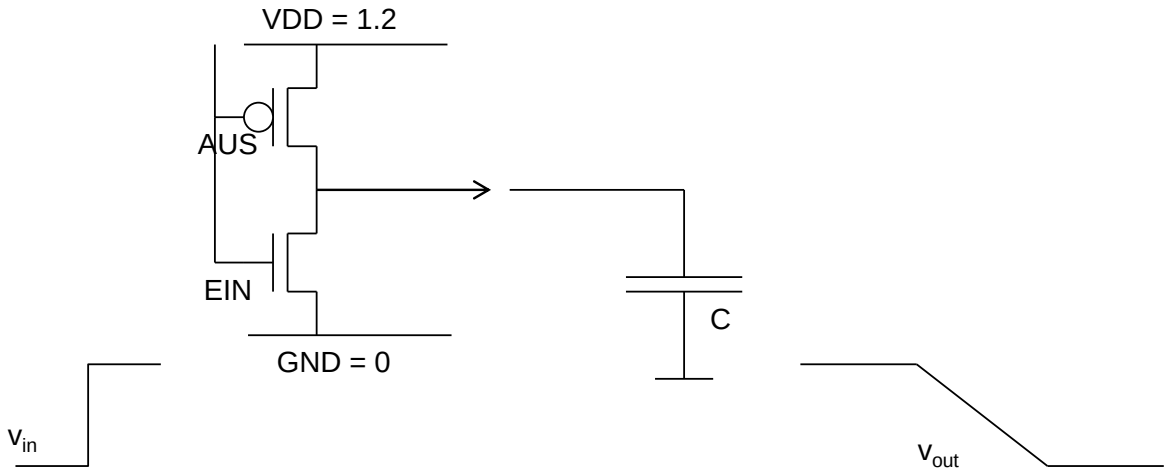

Abbildung 2

Die Ausgangsspannung war am Anfang VDD. Danach sinkt sie. Wir machen die Annahme dass die Schaltgeschwindigkeit etwa der Entladezeit entspricht.

Um die Entladezeit zu berechnen brauchen wir die I_{ds} -V_{ds} Kennlinie des NMOS Transistors (Abbildung 3).

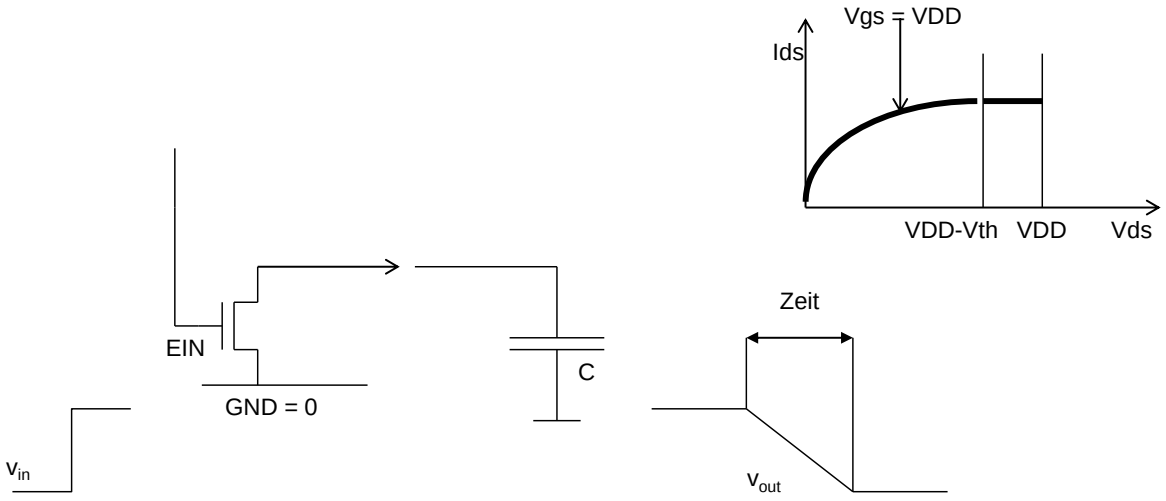

Abbildung 3

Im Ausgangsbereich zwischen VDD und VDD – V_{th} (Bereich 1) wird der Kondensator mit konstantem Strom entladen (Abbildung 4).

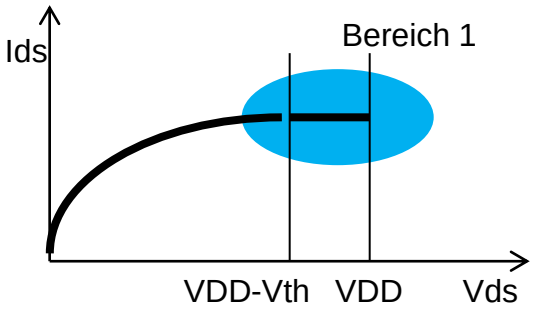

Abbildung 4

Im Bereich 2 hängt der Entladestrom vom V_{ds} ab (Abbildung 5).

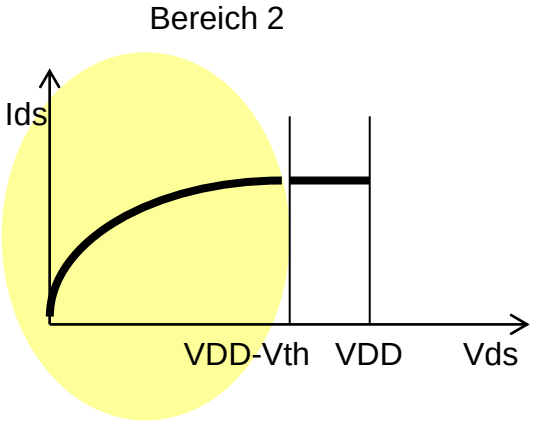

Abbildung 5

Wir werden den ersten Bereich aus folgenden Gründen vernachlässigen: Der Unterschied zwischen logisch 1 und 0 (logischer „Swing“) ist deutlich größer als V_{th}. Der Entladestrom im Bereich 1 ist am größten. Deshalb befindet sich das Potential V_{ds} viel kürzere Zeit im Bereich 1 als im Bereich 2. Die Entladezeit im Bereich 2 dominiert.

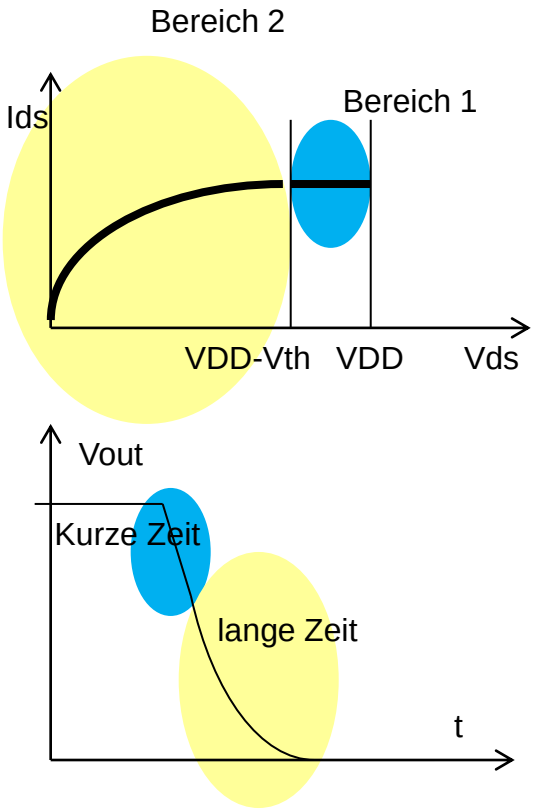

Abbildung 6

Das Entladeverhalten im Bereich 2 kann mit folgender Differentialgleichung hergeleitet werden:

$$C\frac{dU}{dt} = -I_{ds} = -k\left(V_{ds}V_{gst} - \frac{1}{2}V_{ds}^2\right) = -k(V_{gs}U - \frac{1}{2}U^2) \quad (1)$$

mit

$$V_{gst} = V_{gs} - V_{th} = VDD - V_{th}$$

Faktor k ist mit folgender Gleichung beschrieben:

$$k = \mu C'_{ox}\frac{W}{L} \quad (2)$$

μ ist Mobilität, C'_{ox} ist die Gate-Kapazität pro Fläche, W und L Breite und Länge des Kanals.

Normalerweise sind nichtlineare Gleichungen schwer zu lösen. Die Gleichung (1) kann analytisch gelöst werden.

Die Variablen U und t werden zuerst getrennt:

$$\frac{dU}{V_{gs}U - \frac{1}{2}U^2} = -\frac{k}{C}dt$$

Danach werden beide Seiten werden integriert – die Gleichung gilt für $U < V_{gst} = VDD - V_{th}$

Die Lösung ist:

$$U(t) = 2V_{gst}\frac{e^{-\frac{t}{\tau}}}{1+e^{-\frac{t}{\tau}}};\ \tau = \frac{C}{kV_{gst}} \qquad (3)$$

Beachten wir, dass sich im Bereich um $V_{ds} = 0$ der Transistor wie ein Widerstand R_{on} verhält

$$R_{on} = \frac{1}{kV_{gst}} \qquad (4)$$

Die Formel (3) kann wie folgend umgeschrieben werden:

$$U(t) = 2V_{gst}\frac{e^{-\frac{t}{R_{on}C}}}{1+e^{-\frac{t}{R_{on}C}}} \qquad (5)$$

Die Formel (5) hat eine ähnliche Zeitabhängigkeit, wie wenn eine Kapazität mit einem linearen Widerstand entladen wird:

$$U(t) = U(0)e^{-\frac{t}{R_{on}C}} \qquad (6)$$

Abbildung 7 zeigt den Unterschied zwischen den Funktionen (5) (Transistor) und (6) (Widerstand). Wenn die Kapazität durch Transistor entladen wird, dauert ist die Entladezeit von 100% auf 5% des Anfangswertes etwa $4 \times R_{on} C$ und wenn die Kapazität durch Widerstand entladen wird $3 \times R_{on} C$.

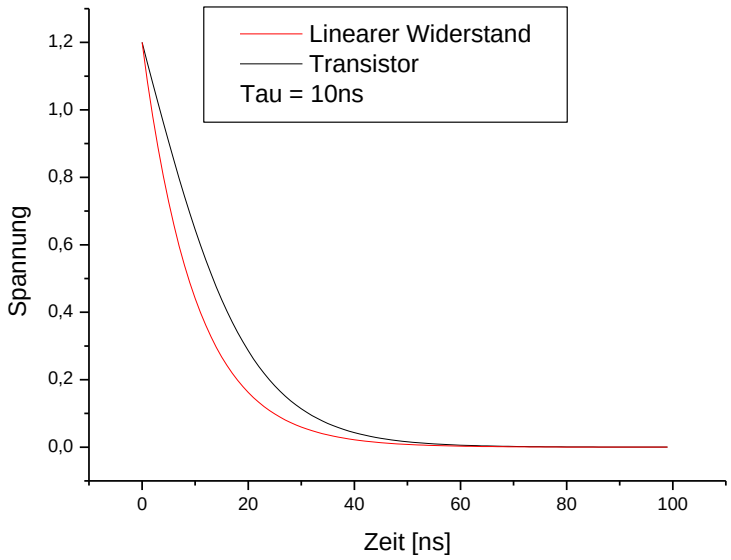

Abbildung 7

Wichtig ist das Folgende – die Geschwindigkeit des Inverters, also die Entladezeit hängt direkt von der Lastkapazität C ab und umgekehrt vom Faktor k. Wie erwähnt, hängt der Faktor k (2) hängt von der Mobilität der Ladungsträger µ, der Oxidkapazität C'_{ox} und vom Verhältnis W/L ab. Aus (2) und (3) leiten wir die Formel für die Entladezeit her:

$$\tau = \frac{C}{\mu C'_{ox} \frac{W}{L} V_{gst}} \quad (7)$$

Parameter μ, W/L, C'_{ox} und $V_{gst} = VDD - V_{th}$ sind die Parameter vom NMOS.

Im Fall, wenn der PMOS leitet gilt die gleiche Formel mit dem Unterschied, dass die Parameter μ, W/L, C'_{ox} und $V_{gst} = VDD - V_{th}$ die von PMOS sind.

Wenn fallende und steigende Flanken gleich sein sollen, müssen wir unterschiedliche Beweglichkeiten für Elektronen und Löcher mit verschiedenen W/L Faktoren kompensieren.

Deshalb sind die PMOS Transistoren normalerweise im Layout breiter.

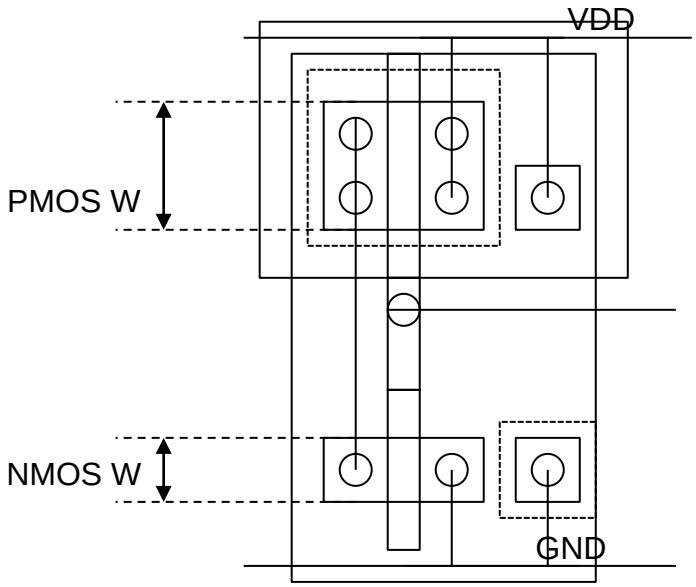

Abbildung 8

Die Kapazität C modelliert folgende Strukturen:

Ein Beitrag ist die Kapazität der digitalen Zelle, die an der Inverter angeschlossen ist (Abbildung 9).

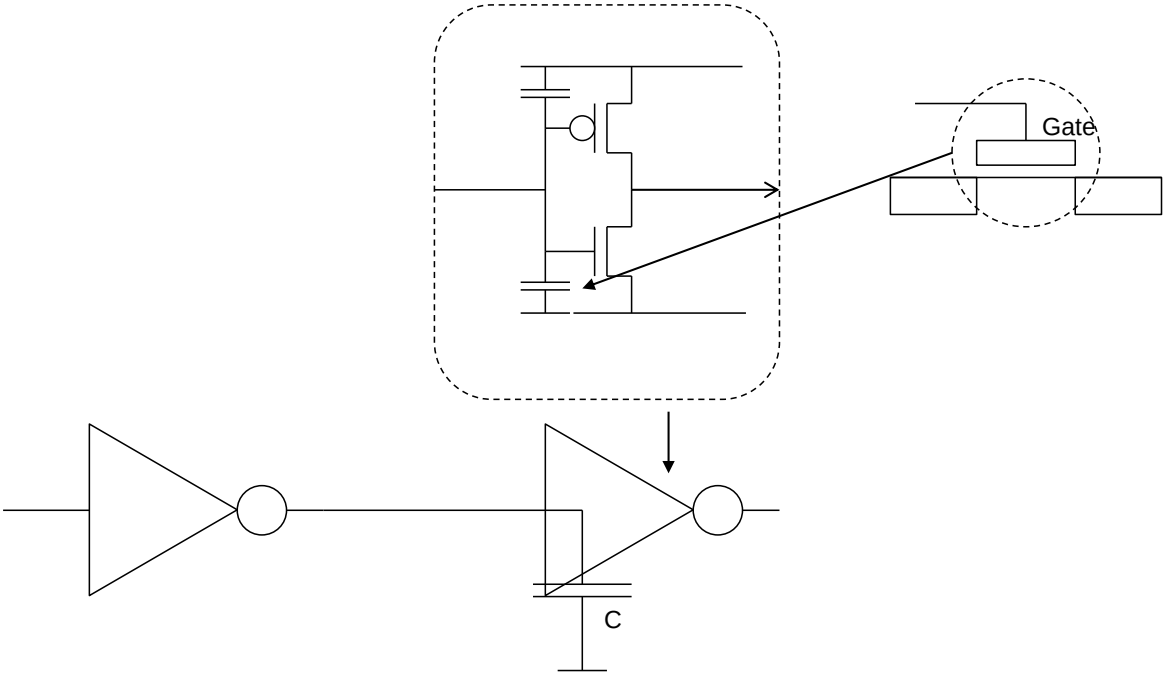

Abbildung 9

Zweiter Beitrag sind die die Kapazitäten von Metallleitungen (Abbildung 10).

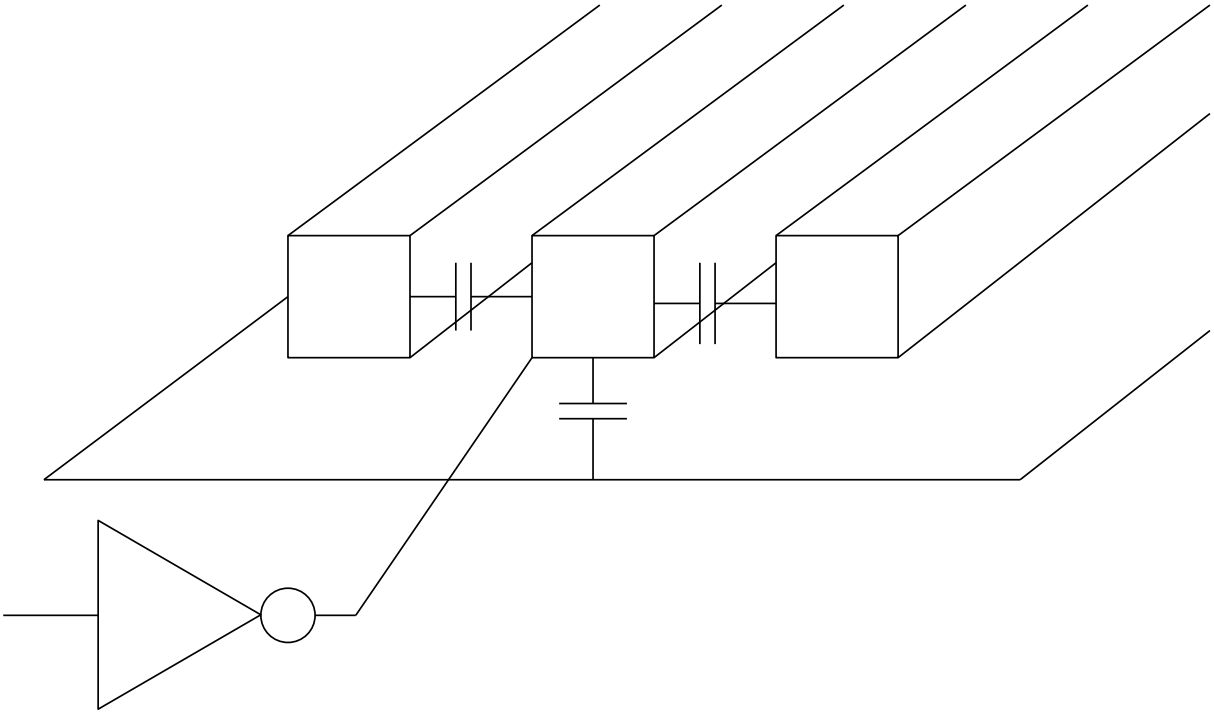

Abbildung 10

AC-Stromverbrauch

Wenn wir am Eingang eines CMOS Inverters logisch 1 oder 0 haten, ist entweder PMOS oder NMOS ausgeschaltet. Der Stromverbrauch ist dann (fast) null. (Wir vernachlässigen Drain-Source-Subthresholdstrom und Gate-Tunnelstrom)

Berechnen wir den mittleren AC-Stromverbrauch und den mittleren Leistungsverbrauch eines Inverters mit seiner kapazitiven Last. Wir stellen uns folgendes Modell vor (Abbildung 11).

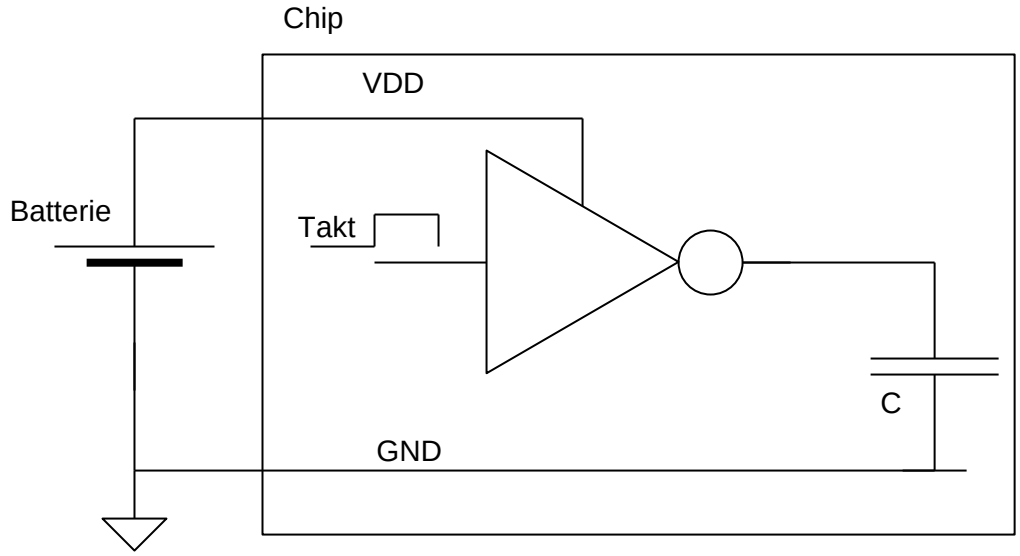

Abbildung 11

Der Inverter befindet sich auf einem Chip. VDD des Chips wird an eine externe Spannungsquelle angeschlossen, z.B. eine Batterie. Der Eingang des Inverters wird mit einem Taktsignal mit Frequenz f getaktet. Nehmen wir es an, dass der Ausgang des Inverters an eine Kapazität C angeschlossen ist. Die Kapazität modelliert die Eingangskapazität der nächsten CMOS Zelle und die Leitungkapazität. Wir nehmen momentan an, dass die andere Elektrode der Kapazität mit GND verbunden ist.

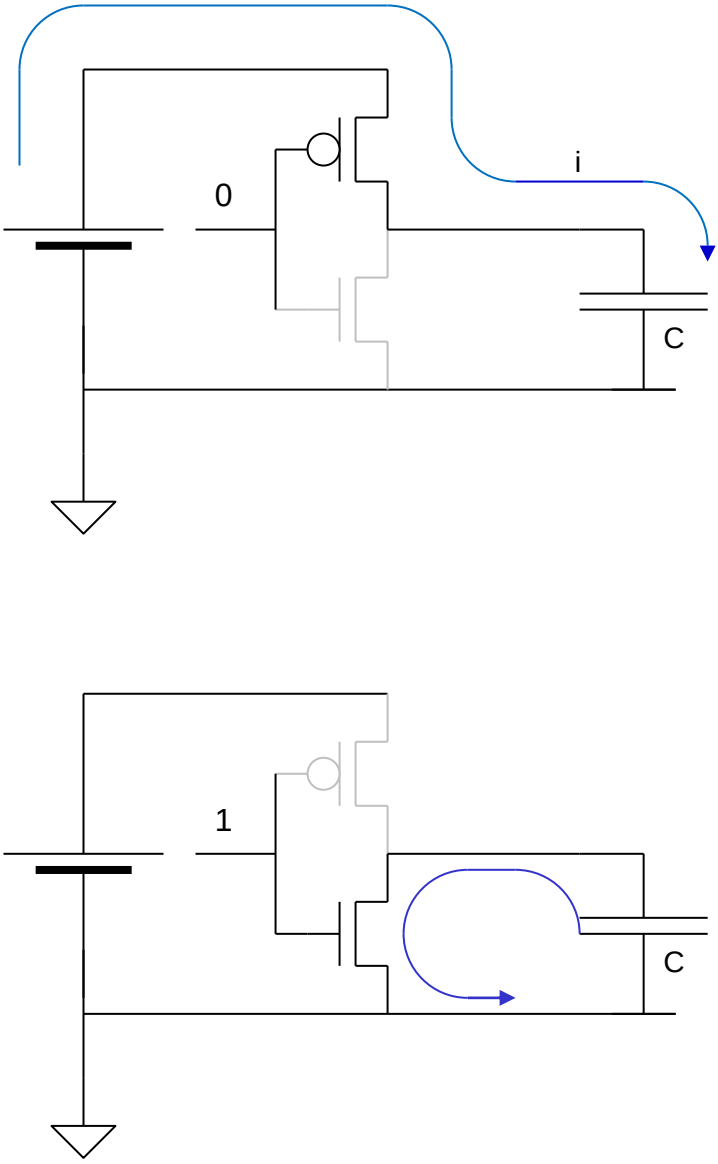

Abbildung 12

Wenn der Eingang von 1 auf 0 ändert (Abbildung 12, oben), wird der PMOS eingeschaltet. Die Ausgangskapazität muss von 0 auf VDD durch PMOS aufgeladen werden. Deswegen fließt eine Ladung:

$Q = VDD \times C$ durch die die Versorgungspannung, den PMOS Transistor und die Kapazität in die GND Leitung.

Diese Ladung muss die Spannungsquelle für die Versorgung des Chips bereitstellen.

Wenn der Eingang von 0 auf 1 ändert (Abbildung 12, unten), wird der NMOS eingeschaltet. Die Ausgangskapazität wird entladen. Es fließt dadurch ebenfalls eine Ladung VDD × C durch die Kapazität. In dem Fall ist der Stromkreis im Chip (der Stromkreis umfasst die Kapazität, NMOS und GND Leitung). Die Spannungsquelle für die Versorgungsspannung sieht die Ladung nicht.

Die Gesamtladung durch die externe Spannungsquelle in der Zeit T ist:

$$Q = N_{ck}\, VDD\, C = f \times T \times VDD \times C$$

Beachten wir, dass in der Formel N_{ck} die Zahl von Taktperioden ist, und nicht die Zahl von Taktflanken. Nur die Taktänderung 1 auf 0 führt zum Strom in der Batterie.

Der mittlere Strom ist:

$$< I > = \frac{Q}{T} = f \times VDD \times C$$

Wie groß ist der mittlere Leisugsverbrauch? Die Batterie erzeugt die folgende mittlere Leistung:

$$< P > = VDD < I > = f \times VDD^2 \times C \quad (8)$$

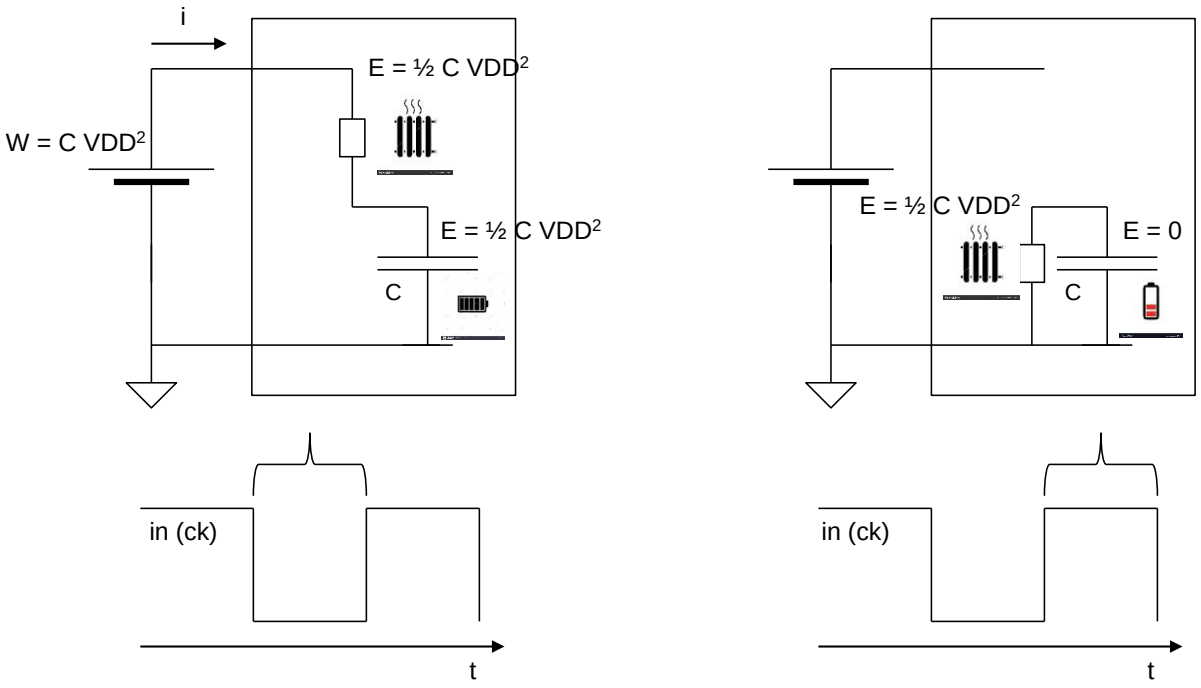

Abbildung 13

Wegen der Energieerhaltung muss diese Leistung auf dem Chip in Form von Wärme freigesetzt werden. Wo wird die Wärme erzeugt? Obwohl Kapazität C in Formel steckt, entsteht die Wärme nicht am Kondensator. Die Wärme entsteht an Widerständen der Transistoren

(Abbildung 13). Man kann die Energie die am Widerstand freigesetzt wird wenn C auf oder entladen wird herleiten. Es ist gegeben durch die Formel:

$$E = \int u(t)i(t)dt = R\int i^2(t)dt \qquad (9)$$

Es ist interessant, dass in dieser Formel R vorhanden ist. Wie passt dass zu Formel (8)?

Um (9) zu lösen, müssen eine Zeitfunktion für i(t) annnehmen. Die richtige Formel könnten man aus (5) herleiten (i = C dU/dt). Wir werden einfachkeitshalber annehmen:

$$i(t) = \frac{VDD}{R}e^{-\frac{t}{RC}} \qquad (10)$$

Wenn wir (10) in (9) einsetzten bekommen wir

$$E = R\int i^2(t)dt \; = R\frac{VDD^2}{R^2}\int e^{-\frac{2t}{RC}}dt = \frac{1}{2}C\,VDD^2 \qquad (11)$$

Beachten wir, dass die Energie in Transistoren sowohl beim Aufladen als auch beim Entladen in Wärme umgewandelt wird. Auch hier gibt es R in der endgültigen Formel nicht. Bei einem kleineren Widerstand ist zwar die Spannung in der Formel (9) kleiner aber der Strom ist um so viel größer. Der Faktor $VDD^2/2C$ entspricht der Energie des E-Feldes im geladenen Kondensator. Durch Entladen wird diese Energie in Wärme W umgewandelt. Aus diesem Grund ist würde eine nichtlineare Charakterstink des Widerstands das Ergebnis (11) nicht beeinflussen.

Man kann ebenfalls zeigen, dass genau die gleiche Energie am PMOS in Wärme umgewandelt wird, wenn Kondensator aufgeladen wird.

Die Gesamtenergie in Zeit T ist dann:

$$E_{ges} = N_{ck,flanken}\,\frac{1}{2}C\,VDD^2 = f\,T\,C\,VDD^2$$

Die mittlere Leistung ist:

$$< P > = \frac{E_{ges}}{T} = N_{ck,flanken}\,\frac{1}{2}C\,VDD^2 = f\,C\,VDD^2$$

Das ist das gleiche Ergebnis wie (8).

Wir können Leistung sparen, indem wir VDD verringern. Aus diesem Grund verwenden neuere Prozessgenerationen kleinere Versorgungsspannungen. Die Kapazitäten verringern sich nicht so stark. Die Gate-Kapazität / Fläche ist in einer kleineren Technologie linear größer, die Fläche quadratisch kleiner. Deshalb ist die Gate-Kapazität kleiner. Allerdings, die Kapazitäten von Metalllinien sind größer da die Abstände kleiner sind.

Takt-Baum

Oft werden Invertern als Treiber für die Takt-Leitung verwendet. Eine Taktleitung ist an viele Flip-Flops angeschlossen und hat deshalb eine große Kapazität (Abbildung 14).

In diesem Fall brauchen wir Inverter mit großem W/L Verhältnis.

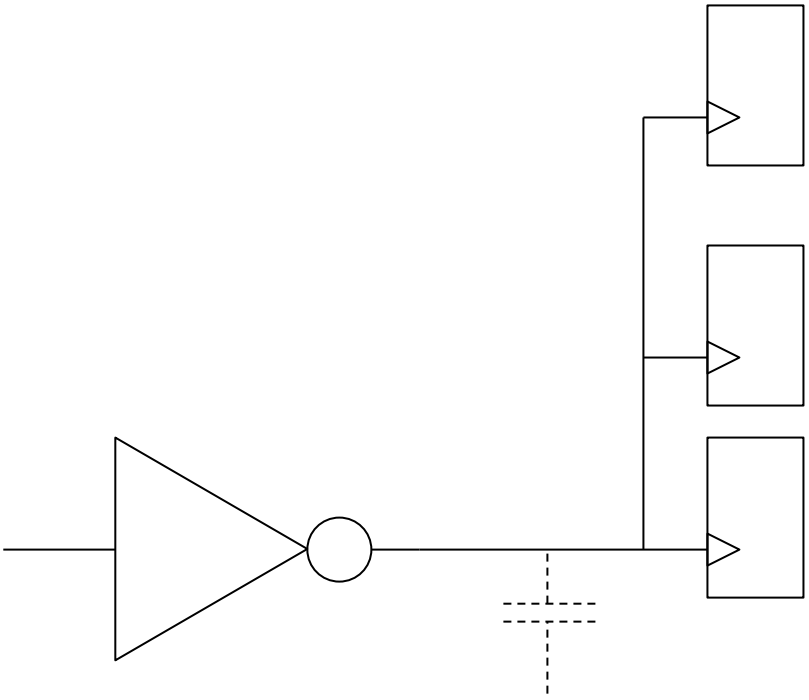

Abbildung 14

Deswegen gibt es in einer digitalen Bibliothek Invertern mit verschiedenen Stärken. Ein kleiner Inverter hat die Stärke 1 (oder 0) und wird als INV_1 (oder als INV_0) bezeichnet. Es gibt auch größere Invertern INV_2 … 4 … 8 (Abbildung 15).

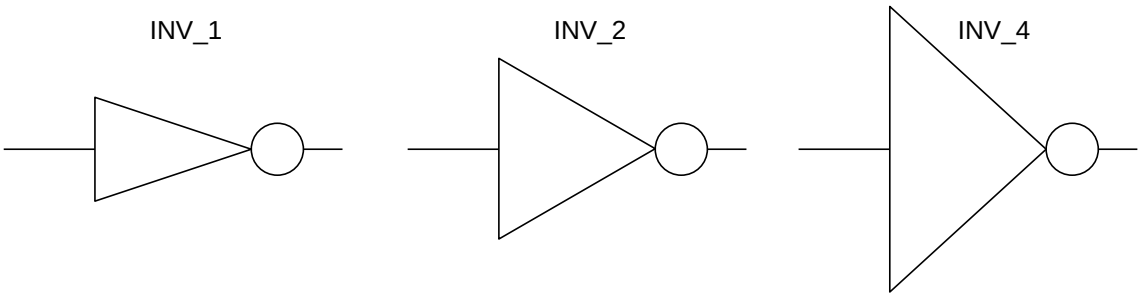

Abbildung 15

Oft entspricht ein INV_2n den zwei in parallel geschalteten INV_n. Das Layout ist normalerweise angepasst, so dass INV_2n nicht unbedingt im Layout zweimal größer ist, obwohl sein effektives W/L Verhältnis 2x größer ist (Abbildung 16).

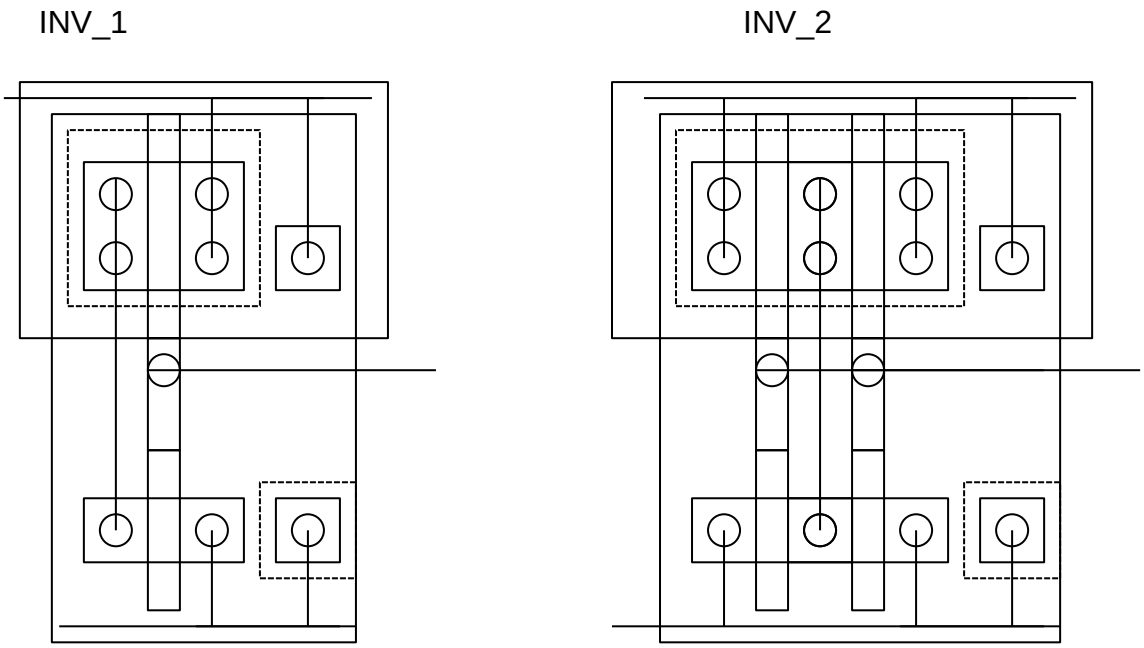

Abbildung 16

Die Wahl des Inverters wäre einfach, wenn die Invertern selbst keine Kapazitive Last für die vorherige Stufe erzeugen würden. Wir könnten dann immer den größten Inverter nehmen.

Woher kommt die Kapazität eines Inverters? Erinnern wir uns, dass die MOSFET Transistoren eine Oxidschicht zwischen dem Gate und dem Kanal haben. Das hieß, es gibt eine Gate-Kapazität mit der Größe

$$C_{ox} = WL\frac{\varepsilon_{SiO2}}{T_{ox}}$$

Eine detailliertere Analyse zeigt, dass die Gate-Kapazität auch vom Arbeitsbereich abhängig ist – sie ist 2/3 C_{ox} in Sättigung, volle C_{ox} im Linearbereich, usw.

Ein n-facher Inverter hat also eine n-fache Eingangskapazität.

Ich werde es mit der folgenden Optimierungsaufgabe illustrieren.

Wir haben ein Flip-Flop Ausgang, der einem Inverter mit Stärke 1 entspricht.

Wir möchten, dass das Flip-Flop ein Taktsignal generiert, das für 1100 weitere Flip-Flops verwendet wird. Wir haben die Invertern mit Stärken 1, 2 , 4 … zur Verfügung. Die Frage ist, wie die optimale Lösung im Sinne der Taktsignal-Verzögerung aussieht.

Wir können z.B. Flip-Flop-Ausgang direkt an die 1100 Flip-Flips anschließen, oder, eventuell, einen sehr starken (z.B. 128-fachen) Inverter dazwischenschalten, oder eine längere Kaskade von Invertern mit steigender Stärke benutzen.

Ich vernachlässige hier, dass ein Inverter den Takt invertiert.

Wenn wir also den Flip-Flop-Ausgang Q an 1100 FFs anschließen, können wir erwarten, dass die Takt-Anstiegszeit zu langsam ist. Die Taktfrequenz ist dann begrenzt.

Wenn wir einen großen Inverter an Q anschließen, stellt er selbst eine große kapazitive Last dar. Der Ausgang des Flip-Flops (realisiert mit einem Inverter) wird wegen der großen Lastkapazität stark verlangsamt.

Es ist vermutlich am besten, mehrere Invertern nacheinander zu schalten, die immer größer sind. Die Frage ist wie viele und wie sollen die Größenverhältnisse der Inverter sein?

Es ist interessant, dass man diese mathematische Optimierungsaufgabe analytisch lösen kann.

Das Ergebnis ist sehr einfach. Wir brauchen eine Kaskade von Invertern. Optimal wäre, dass der nächste Inverter immer um Faktor e = 2.718… größer ist als der vorherige. Um eine kapazitive Last zu „treiben“, die zB. 1100 INV_1 Invertern entspricht, beginnend von einem INV_1, brauchen wir ln(1100) = 7 Invertern:

Die Stärken sind: 1× (~e^0 ×), 2.7× (~e^1 ×), 7.9× (~e^2 ×), 20× (~e^3 ×), 55× (~e^4 ×), 148× (~e^5 ×), 402× (~e^6 ×).

Normalerweise wird statt e=2,718…, ein Verhältnis 2× oder 3× verwendet, da es im Layout einfacher zu realisieren ist (Abbildung 17).

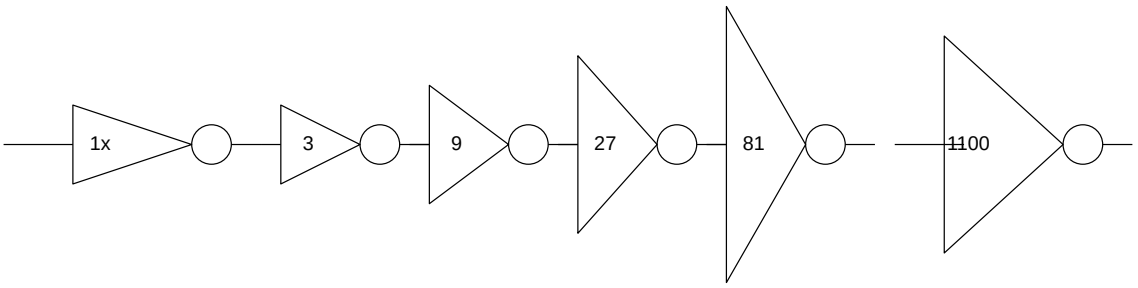

Abbildung 17

Herleitung (kompliziert – nur als Info)

Wir haben N Inverter in Serie. Inverter 0 modelliert den Ausgang des ersten Flipflops und Inverter N die Eingänge von allen Flipflops (in unserem Fall K = 1100) zusammen. Berechnen wir die optimalen Größen der Inverter-Transistoren und die optimale Zahl der Inverter in der Kette.

Die erste Gleichung stellt die Größe dar die wir minimieren möchten:

$$\tau_{gesamt} = \sum_{i=0}^{N-1} \tau_i \quad (12)$$

Zeit T_i ist die Zeitkonstante des Inverters i. Die Zeit T_i ist gegeben durch die Formel (7)

$$\tau_i = \frac{C_{i+1}}{\mu C'_{ox} \frac{W_i}{L_i} V_{gst}} \quad (13)$$

W_i und L_i sind die Dimensionen der Transistoren im Inverter i (Stufe i). C_{i+1} ist die kapazitive Last die Inverter i sieht. Wir nehmen hier an, dass C_{i+1} ungefähr die Gate-Kapazität vom Inverter i+1 ist. Es gilt:

$$C_{i+1} = W_{i+1} L_{i+1} C'_{ox}$$

(14)

Wenn wir (14) in (13) einsetzten, bekommen wir:

$$\tau_i = \alpha \frac{W_{i+1} L_{i+1}}{W_i / L_i} \quad (15)$$

Faktor α ist eine Konstante.

Falls wir annehmen, dass alle Transistorlängen gleich sind, bekommen wir

$$\tau_i = \alpha \frac{W_{i+1}}{W_i} \quad (16)$$

Für diese Optimierungsaufgabe brauchen wir eine Nebenbedingung für die Variablen T_i. Diese können wir folgenderweise herleiten:

Falls wir das Produkt von allen Zeitkonstanten berechnen, ergibt sich:

$$\prod_{i=1}^{N-1} \tau_i = \alpha^N \frac{W_N}{W_0} = \alpha^N K \quad (17)$$

Unsere Aufgabe ist es die optimalen Zeitkonstanten T_i und die Optimale Zahl von Invertern zu finden, welche zur kleinsten Gesamtzeit führen (12) und die Nebenbedingung (17) erfüllen.

Wir haben also eine Funktion $F(T_i)$ die minimiert werden soll, und eine Gleichung $f(T_i) = 0$ die eine Nebenbedingung für die Variablen T_i beschreibt.

$F(T_i)$ ist die Funktion (12):

$F(\tau_i) = \sum_{i=0}^{N-1} \tau_i$ (18)

Die Nebenbedingung ist mit der Gleichung (17) definiert:

$\prod_{i=1}^{N-1} \tau_i = \alpha^N K$ (19)

Oder

$f(\tau_i) = 0$ (20)

Wenn wir (19) und (29) vergleichen, bekommen wir:

$f(\tau_i) = \prod_{i=1}^{N-1} \tau_i - \alpha^N K = 0$ (21)

Wir werden die Methode der Lagrange-Multiplikatoren verwenden.

Wir definieren eine neue Funktion

$\Phi(\tau_i, \lambda) = F(\tau_i) - \lambda f(\tau_i)$ (22)

Die optimalen Werte für Ti und λ bekommt man als Minimum der Funktion Φ(Ti, λ).

Um dieses Minimum zu finden, müssen wir die partiellen Ableitungen

$\partial\Phi/\partial\tau_i$ und $\partial\Phi/\partial\lambda$ berechnen. Aus den Gleichungen

$\frac{\partial\Phi}{\partial\tau_i} = 0$ (23)

und

$\frac{\partial\Phi}{\partial\lambda} = 0$ (24)

Kann man die optimalen Werte berechnen.

Beachten wir, dass die letzte Gleichung (24) zur Nebenbedingung $f(T_i) = 0$ führt.

Berechnen wir (23):

$$\frac{\partial\Phi}{\partial\tau_i} = \frac{\partial}{\partial\tau_i}\sum_{i=0}^{N-1}\tau_i - \lambda\frac{\partial}{\partial\tau_i}\left(\prod_{i=1}^{N-1}\tau_i - \alpha^N K\right) = 1 - \frac{\lambda}{\tau_i}\prod_{i=1}^{N-1}\tau_i = 0$$

Daraus folgt:

$\tau_i = \lambda \prod_{i=1}^{N-1} \tau_i$ (25)

Wenn wir die Nebenbedingung (21) in (25) einsetzen, bekommen wir:

$\tau_i = \lambda \alpha^N K$ (26)

Aus der Gleichung (26) folgt, dass im Optimalfall alle Zeitkonstanten τ_i gleich sein müssen:

$\tau_i \equiv \tau$ (27)

Aus der Nebenbedingung (21), bekommen wir das Ergebnis:

$\prod_{i=1}^{N-1} \tau_i = \tau^N = \alpha^N K$ (28)

oder

$\tau = \alpha \times \sqrt[N]{K}$ (29)

Berechnen wir jetzt die optimale Zahl N.

Jetzt ist das Dimensionierungsproblem eindimensional, da die Gesamtzeit Funktion einer Variable N ist.

Wir haben:

$\tau_{gesamt} = \sum_{i=0}^{N-1} \tau_i = N\tau = N\alpha \times \sqrt[N]{K}$ (30)

Die Zeit τ_{gesamt} soll minimiert werden. Das ist äquivalent zu

$d\tau_{gesamt}/dN = 0$.

Berechnen wir $d\tau_{gesamt}/dN$.

$$\frac{d\tau_{gesamt}}{dN} = \frac{d}{dN}\left(N\alpha \sqrt[N]{K}\right) = \alpha\left(\frac{dN}{dN}\left(\sqrt[N]{K}\right) + N\frac{d}{dN}\left(\sqrt[N]{K}\right)\right)$$

Am schwierigsten ist der Term:

$\frac{d}{dN}\left(\sqrt[N]{K}\right)$

Wir rechnen ihn folgenderweise:

$$\sqrt[N]{K} = K^{1/N} = \exp\left(\ln\left(K^{\frac{1}{N}}\right)\right) = \exp\left(\frac{1}{N}\ln\left(K\right)\right)$$

Daraus folgt:

$$\frac{d}{dN}\left(\sqrt[N]{K}\right) =$$

$$\frac{d}{dN}\exp\left(\frac{1}{N}\ln(K)\right) = \exp\left(\frac{1}{N}\ln(K)\right)\frac{d}{dN}\left(\frac{1}{N}\ln(K)\right) = \sqrt[N]{K}\left(-\frac{1}{N^2}\right)\ln(K)$$

Wir bekommen:

$$\frac{d\tau_{gesamt}}{dN} = \alpha\left(\left(\sqrt[N]{K}\right) + N\sqrt[N]{K}\left(-\frac{1}{N^2}\right)\ln(K)\right)$$

Die Gleichung $d\tau_{gesamt}/dN = 0$ führt zu:

$$\left(\sqrt[N]{K}\right) = \sqrt[N]{K}\left(\frac{1}{N}\right)\ln(K)$$

oder

$$1 = \left(\frac{1}{N}\right)\ln(K)$$

oder

$1 = \left(\frac{1}{N}\right)\ln(K)$ oder

$N = \ln(K)$ (31)

Berechnen wir schließlich die Zeitkonstante τ:

Aus der Lagrange-Methode haben wir (29):

$\tau = \alpha \times \sqrt[N]{K}$ (32)

Wenn wir das Optimalwert für N (31) in (32) einsetzten, bekommen wir

$\tau = \alpha \times \sqrt[\ln(K)]{K}$ (33)

Vergleichen wir Ergebnis (33) mit der Formel für die Zeit τ_i (15):

$$\tau_i = \alpha\frac{W_{i+1}}{W_i}$$

Wir sehen, dass Faktor $\sqrt[\ln(K)]{K}$ das Verhältnis von Transistorbreiten W_{i+1}/W_i darstellt.

$\sqrt[\ln(K)]{K} = \frac{W_{i+1}}{W_i}$ (34)

Ausdruck $\sqrt[\ln(K)]{K}$ kann man folgenderweise vereinfachen:

$$\sqrt[\ln(K)]{K} = \exp\left(\ln\left(\sqrt[\ln(K)]{K}\right)\right) = \exp\left(\frac{1}{\ln(K)}\ln(K)\right) = e$$

Also, das Verhältnis von Transistorbreiten ist e und zwar für jede zwei nachfolgende Stufen. Nur in dem Fall sind die Zeitkonstanten gleich:

$$\tau_i = \tau = \alpha \times \sqrt[N]{K} = \alpha\, e$$