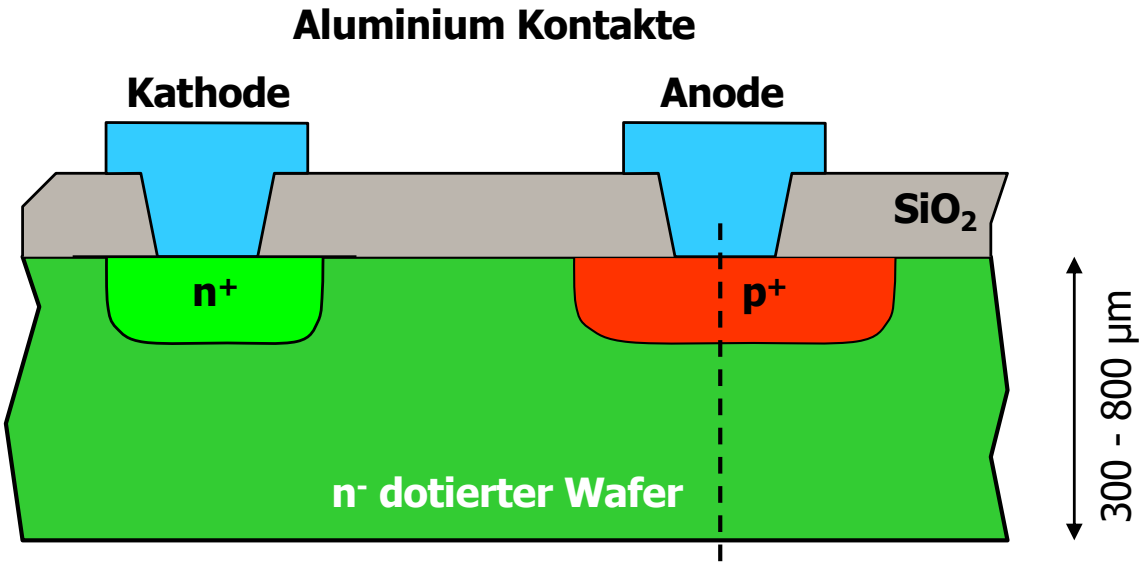
Layout von Bauelementen

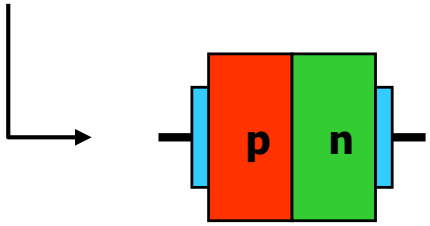
- MOS Transistoren
- Widerstände
- Kondensatoren
- (Spulen)
- Bipolare Transistoren in CMOS Prozessen
- Fuses (Sicherungen)
- Matching
- Verschiedenes: Elektromigration, Latchup..

Erinnerung: pn-Diode durch Implantation

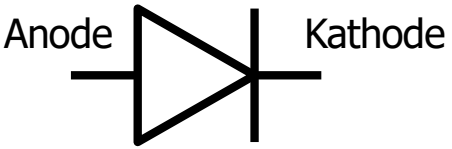
- In eine (z.B.) n- dotierte Si-Scheibe ('Wafer') werden an der Oberfläche stark dotierte Gebiete erzeugt
- JEDER pn-Übergang bildet eine Diode.
- Sie ist meist 'unerwünscht' – ein 'parasitäres' Element (insbesondere Drain, Source und Bulk des MOS)



Querschnitt eines pn-Übergangs auf einem Chip



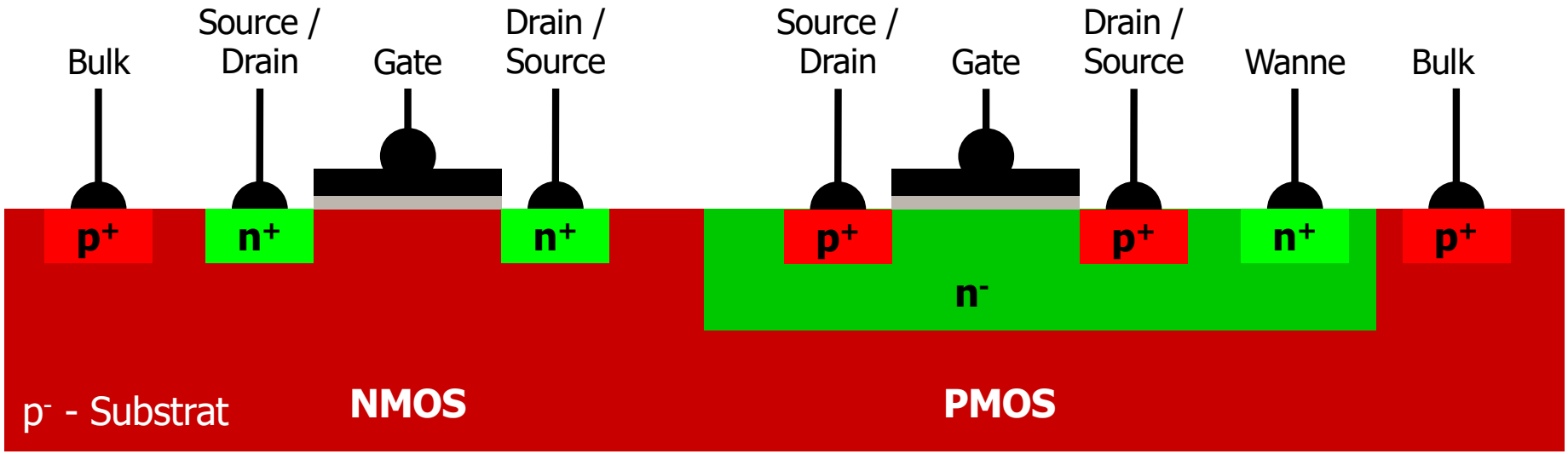
Prinzipieller Aufbau



Schaltsymbol

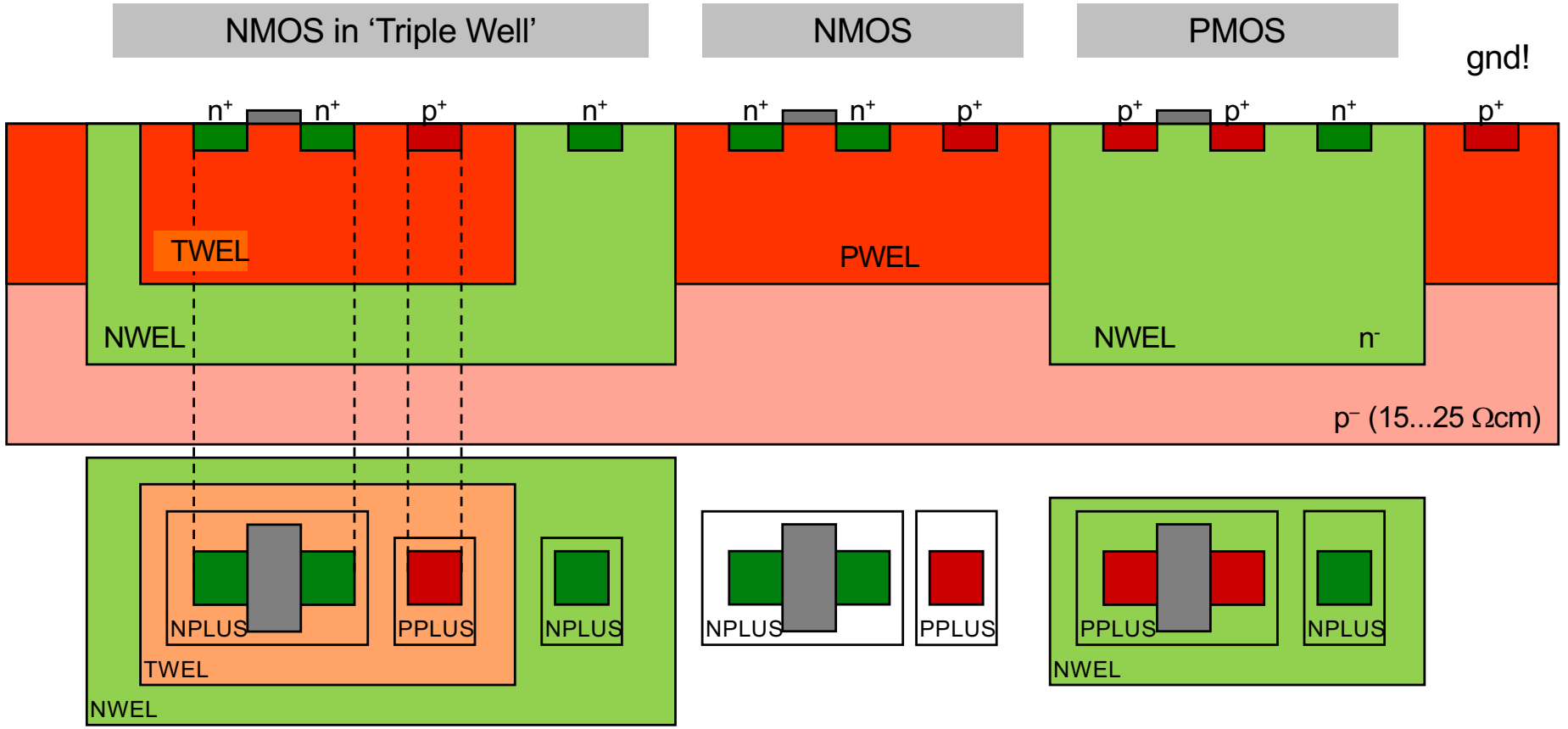
Erinnerung: NMOS / PMOS

- Bei einem p^--dotierten Wafer können die **NMOS** direkt im ‚Substrat' sitzen
- Der **PMOS** muss in n-dotiertem Silizium sitzen.
  Bei einem p^--dotieren Wafer benötigt man eine zusätzliche n^--dotierte **n-Wanne** (engl. **NWell**)



- Ein NMOS (in p-Substrat) **hat 4 Terminals**: G,S,D,B. Bulk (B) = Substrat liegt aber bei allen NMOS auf der gleichen GND! Source kann daher nur mit Bulk verbunden werden, wenn Source = GND.
- Ein PMOS (in N-Wanne in p-Substrat) **hat eigentlich 5 Anschlüsse**: G,S,D,B,SUB
  In den meisten Technologien ist in den PMOS Symbolen aber das ‚triviale' Sub-Terminal nicht enthalten.
- Drain und Source sind im Layout meist austauschbar. Die Namen kommen von den anliegenden Spannungen. Beim NMOS ist Source das Terminal mit der negativeren Spannung (also Drain)

NMOS in 'Triple Well'



- A **Triple Well NMOS** sits in a P-Well inside of an N-well (see left).
- Source can then take any potential. Sensitive devices can be shielded from the (noisy) substrate.
- Such a 'triple well' device **has 6 terminals** (G,D,S, TWELL, NWELL, SUB)
- In UMC, the 'Triple Well NMOS' is called BPW ('buried P-Well')
- In UMC, the BPW symbol has only 4 terminals (G,D,S,TWELL, no NWELL, no SUB)
  - Be careful to connect the correctly in the layout. This is compared by LVS!!!!!

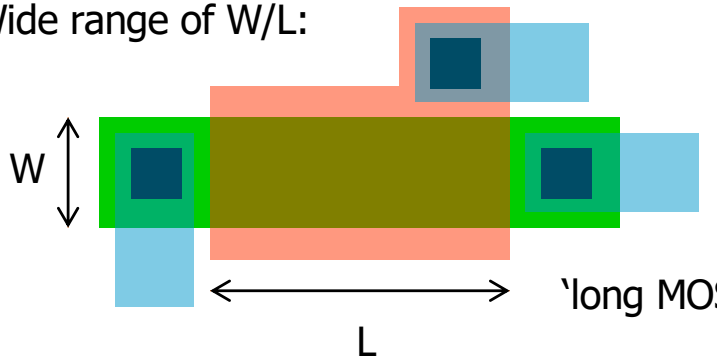
p-material

n-material

Transistoren

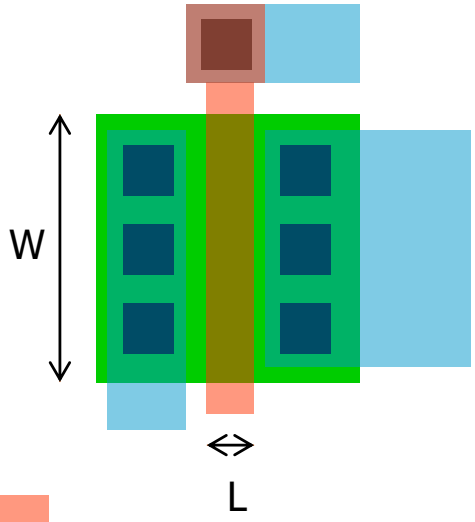
Transistors

- A transistor is formed by **Poly over Diffusion (=implant)**
- Wide range of W/L:

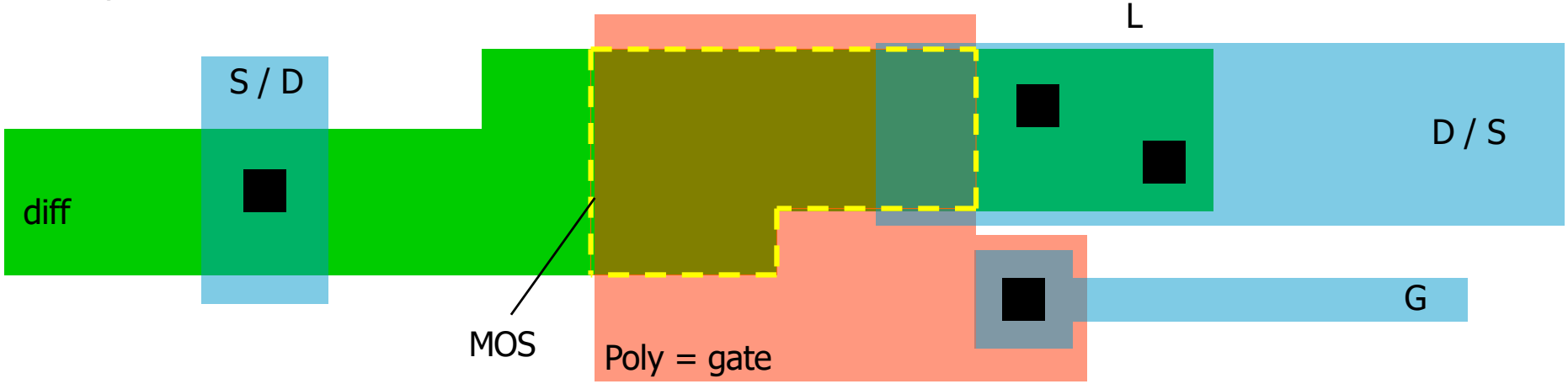


'long MOS'

'wide MOS'



- Bad Layout:

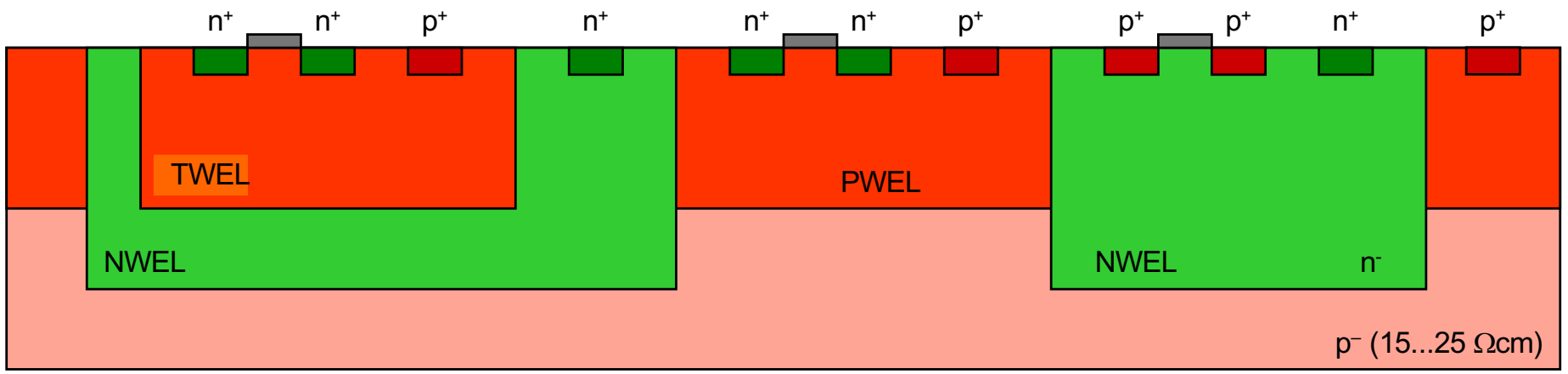


- Watch:
  - trace resistances of Drain / Source (implant) and Gate (poly)
  - unnecessary capacitances
  - In analogue layout: matching between parts -> identical layouts,...

Substrate and Well Contacts

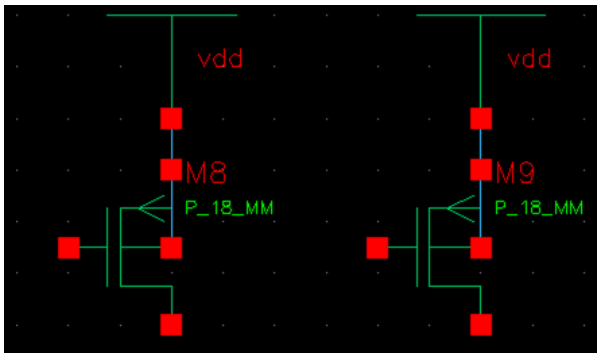
- The potential of each **WELL** must be defined. This is achieved with **well contacts**.
  For an NWELL, the contact is achieved by a n^+-implant.
  The NWELLs are normally connected to a positive supply (analogue supply vdda, digital supply vddd).
  In analogue design there are exceptions (to eliminate the substrate effect, e.g. in a source follower).
  - If a well is not connected to a 'supply' (as defined some rules file..), it is considered a **'hot NWELL'**, often with larger distance rules. This depends a lot on the design kit used..

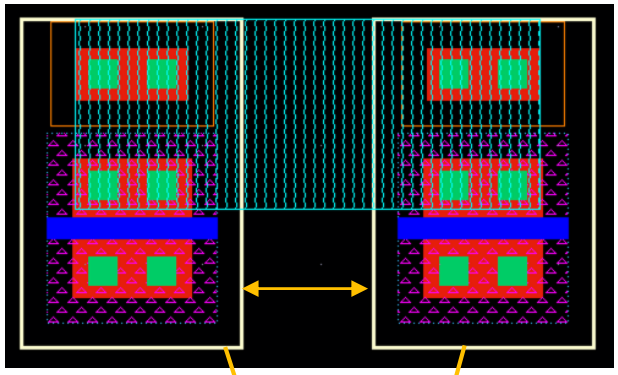
- The **substrate** (wafer) must also be connected using **substrate contacts**.
  In a p-substrate, these are p^+-implants.
  To avoid **Latchup** (see later), a sufficient density is required. The rules ask for a contact in the vicinity of every MOS.



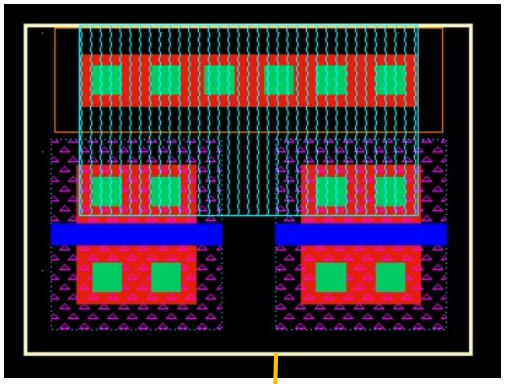
Sharing Wells

- WELLs of different MOS at the same potential can be shared to make the layout more compact:

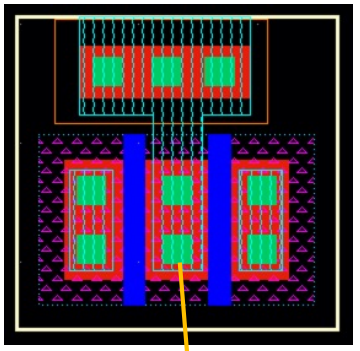




Two independent MOS in separate WELLs. Large WELL-WELL spacing

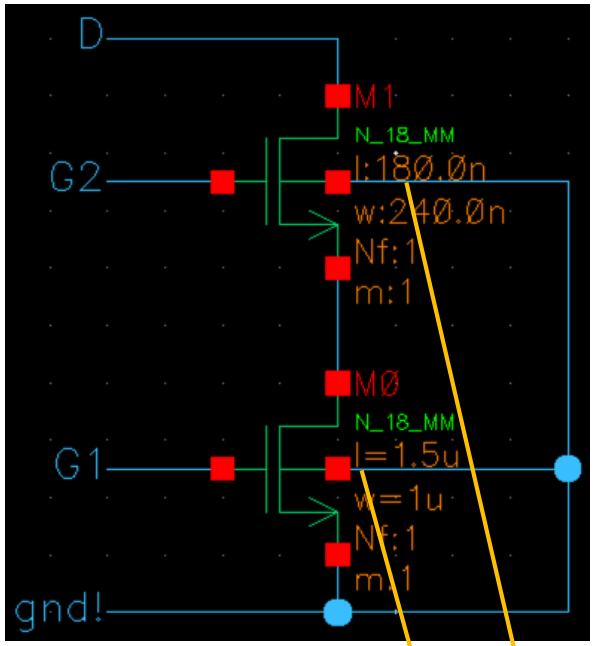


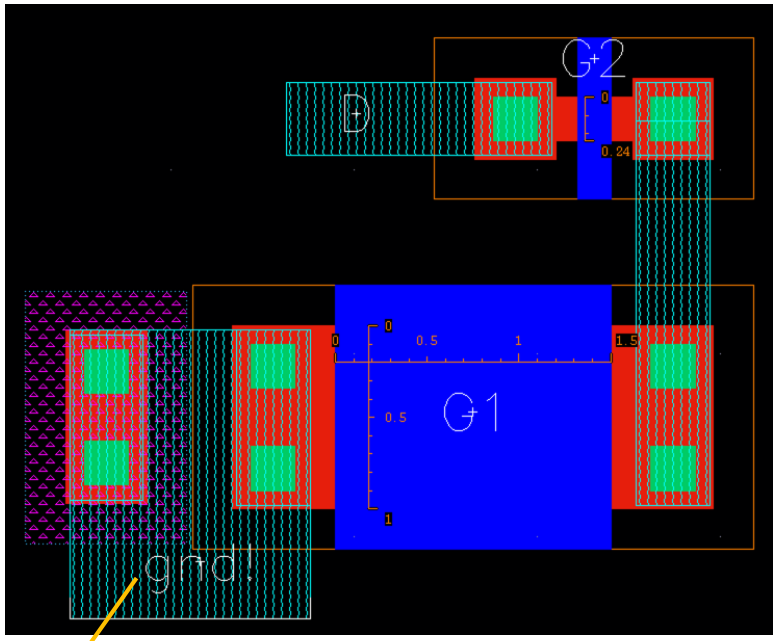
One 'merged' WELL for both MOS. Only one WELL contact.



Better MOS orientation allows to share source areas

Wells / Substrate Contacts: Layout Examples

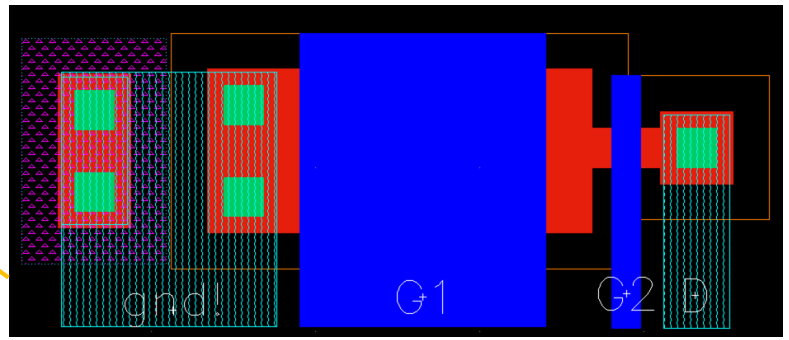


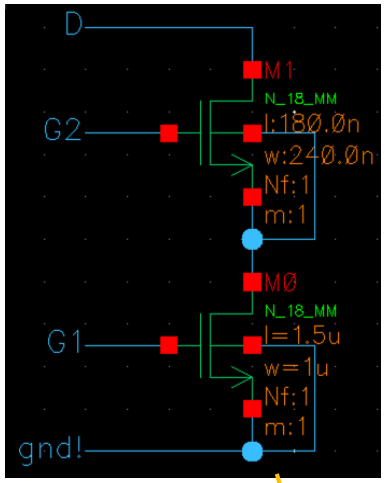


Both NMOS devices **MUST** have same bulk=substrate potential

One p+ contact in the vicinity is sufficient. (In UMC, a contact of type M1_PDIFF can be used.)

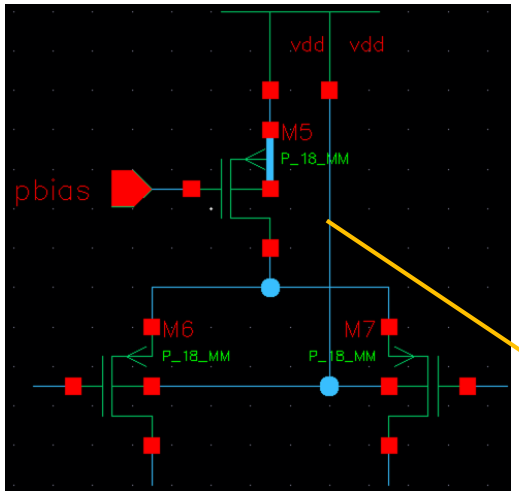
In this more compact layout version, drain of M0 and source of M1 are connected directly via the implant. No contacts / M1 are used.





This circuit is **NOT** possible using normal NMOS!

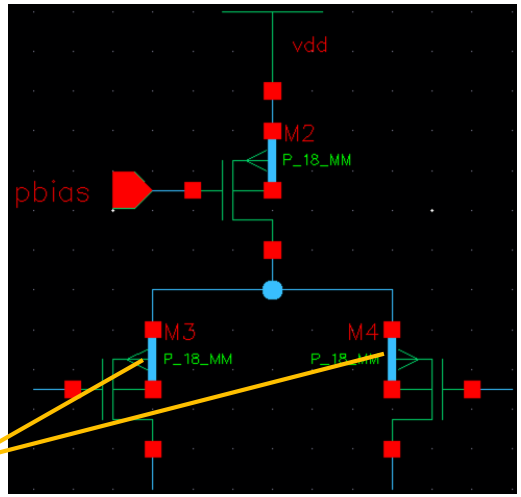
Wells / Substrate Contacts: Layout Examples

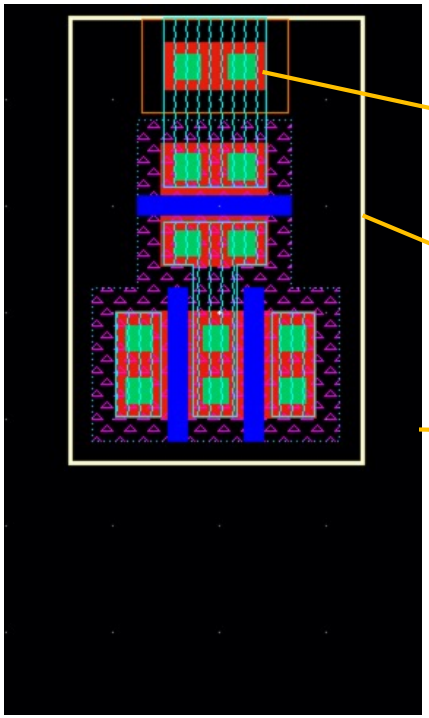


Circuit:
- PMOS current source
- differential PMOS pair

All NWELLs connected to vdd ('standard')

Substrate effect in diff-pair eliminated 😀 (source = bulk).





One n+ well contact For all 3 PMOS

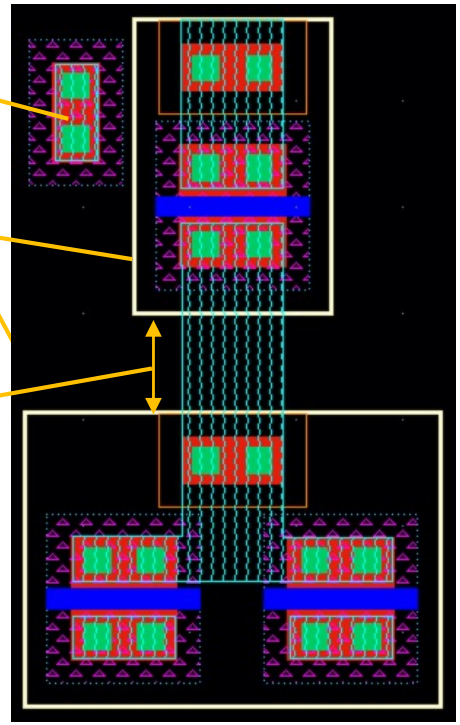
Shared NWELL

Smaller layout!

P+ substrate contact (somewhere)

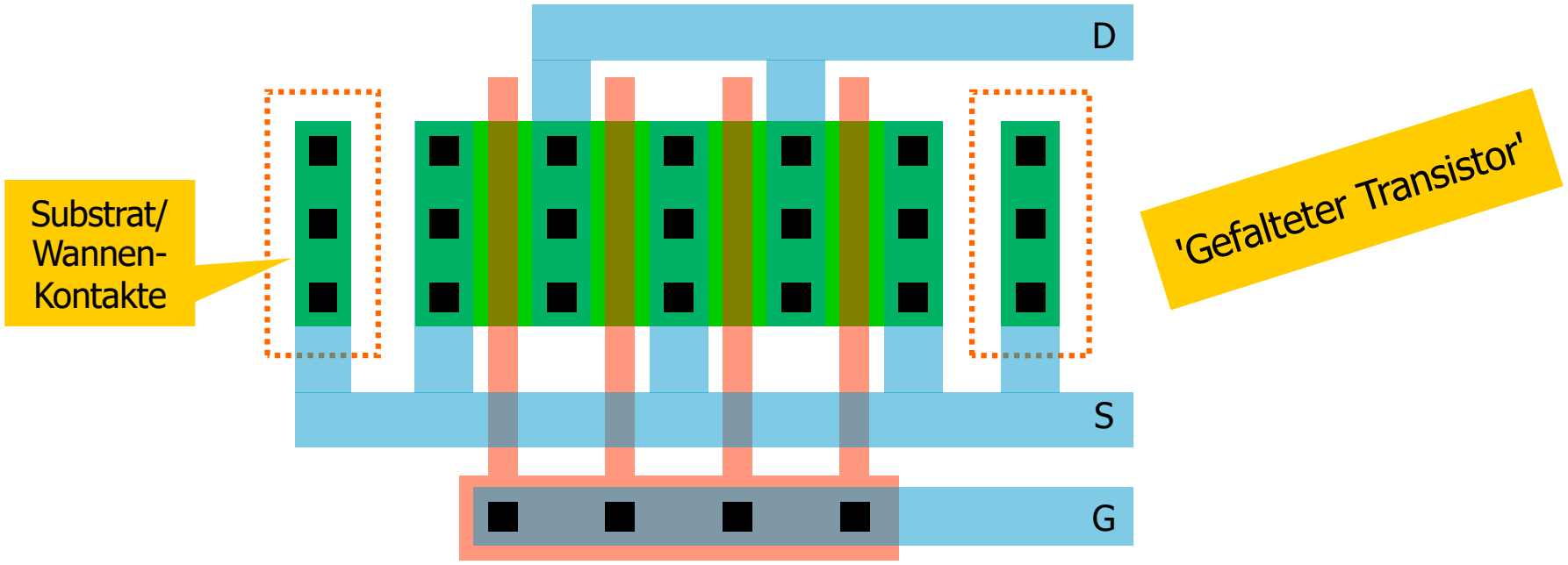
Two separate wells required !

Large well spacing!

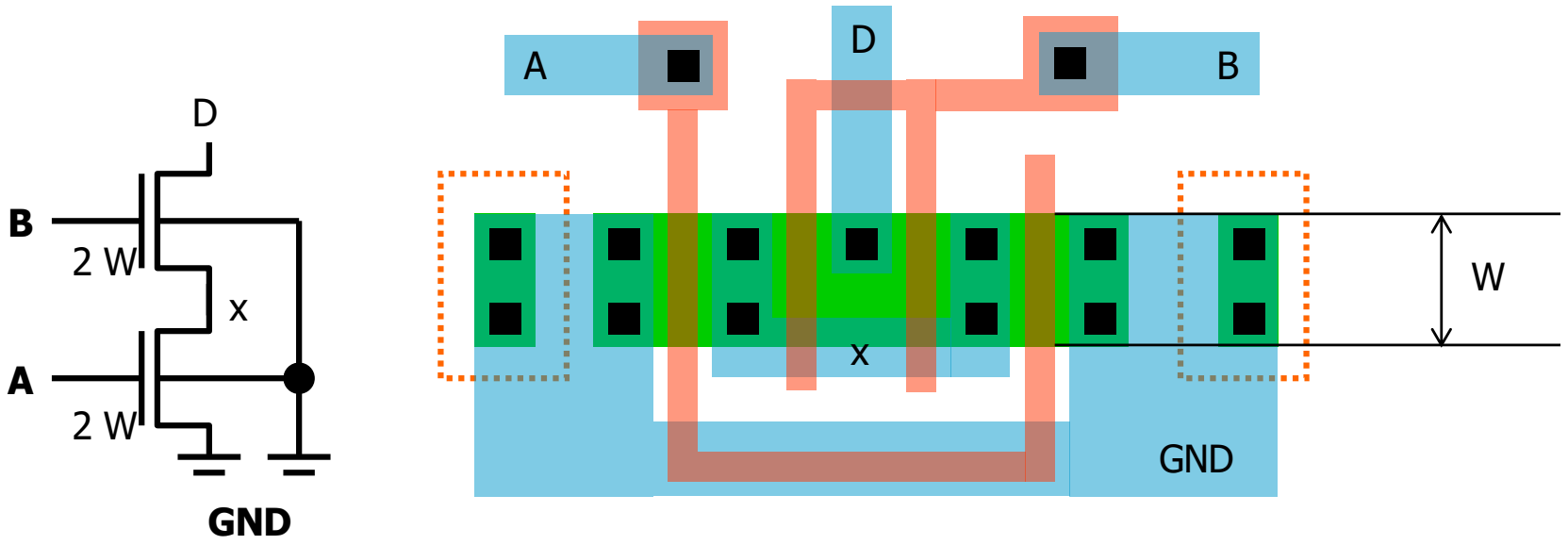


weite Transistoren (W groß)

- Bei großem W werden die Transistoren immer 'gefaltet':
  + **halbe Drain/Source-Kapazität bei gleichem W** !!! (gerade Anzahl Teiltransistoren ist besser !)
  + kleinerer Zuleitungswiderstand des Gates
- Es werden viele Drain/Source-Kontakte gezeichnet, um den Übergangswiderstand zu reduzieren



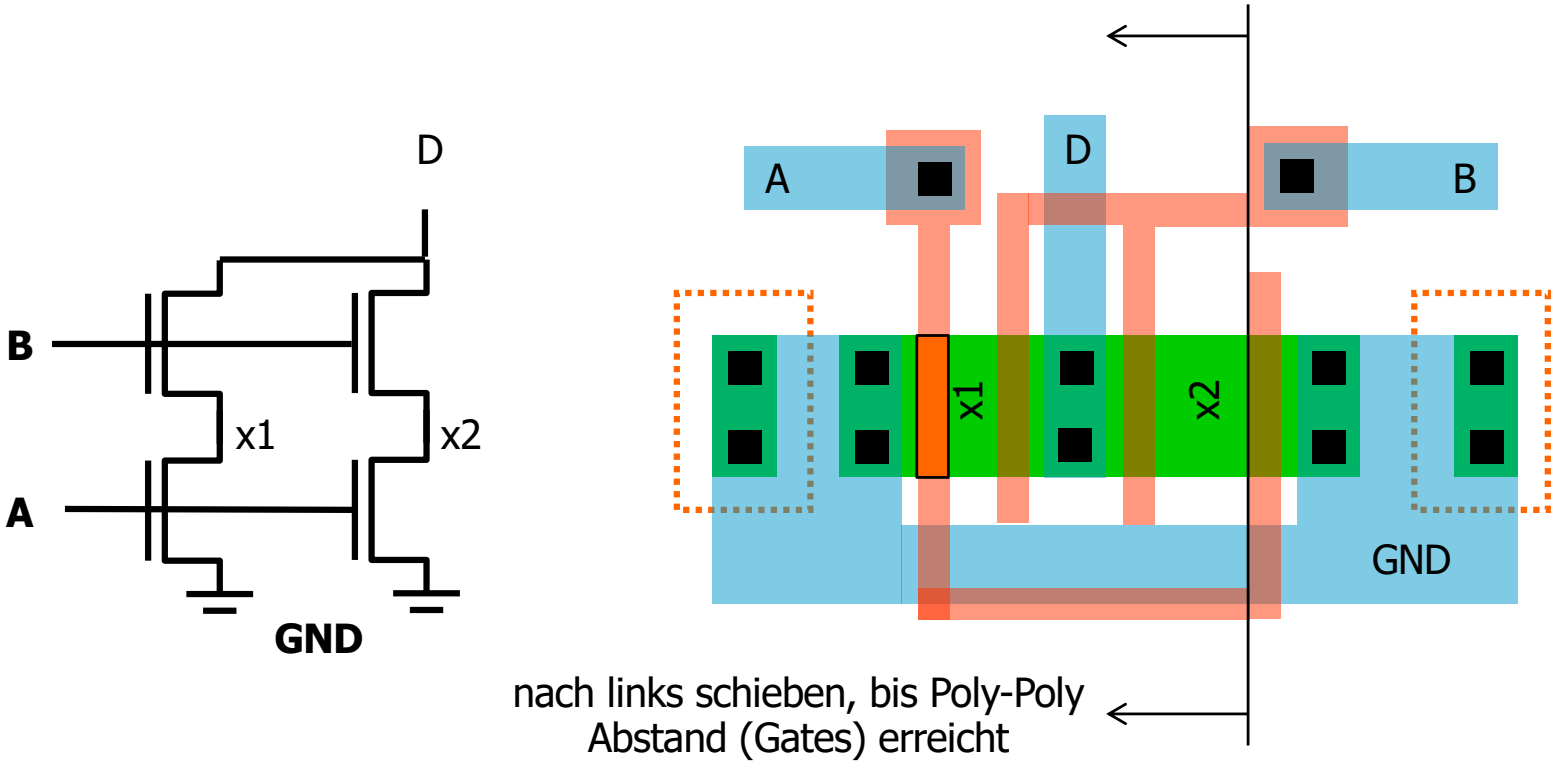
Beispiel: Zwei in Serie geschaltete weite MOS



- Hier: zwei NMOS in Serie wie z.B. beim NAND2 Gatter
- Substratkontakte hier sehr großzügig
- Es wäre **dumm**, den Ausgang außen und GND innen anzuschließen!

- Die Zwischenverbindung (x) kann man eigentlich weglassen, man benötigt dann auch keine Kontakte mehr → s. nächste Seite

Optimierung: Zwischenknoten nicht verbunden

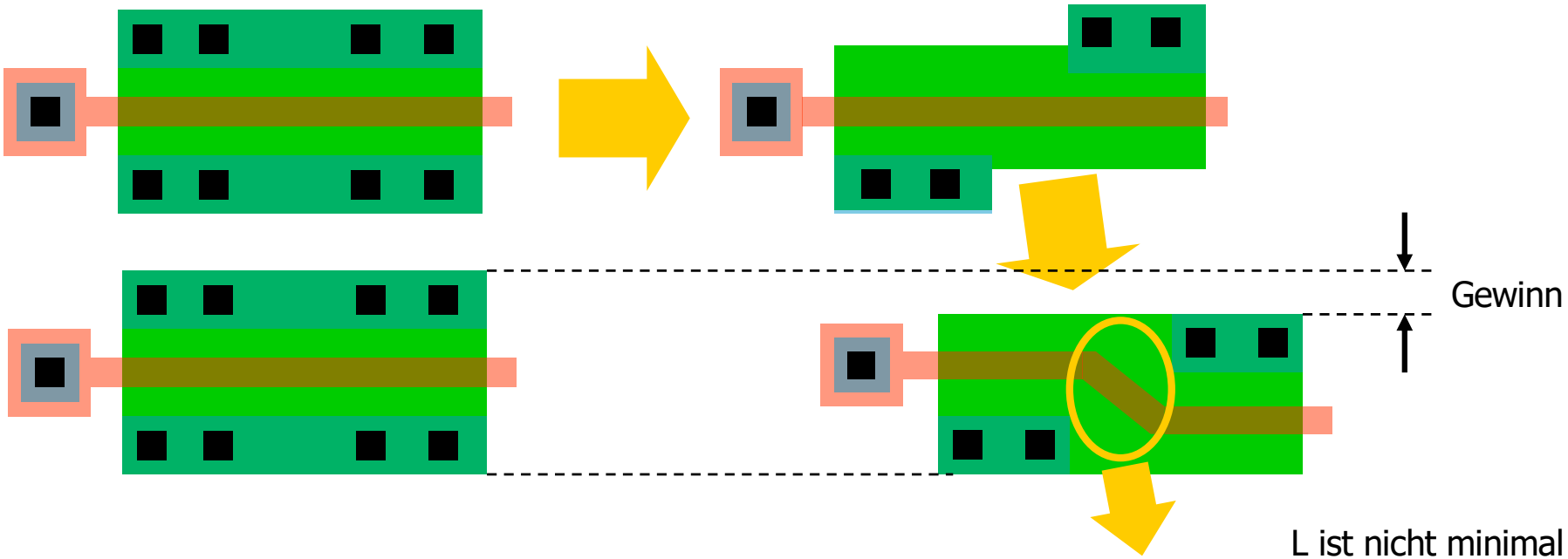


- Es gibt nun zwei unterschiedliche Knoten x1/x2. Diese müssen im Schematic getrennt sein!
- Die Kapazitäten dieser Knoten sind kleiner → gut!
- Da der ‚x'-Bügel nicht benötigt wird, kann die Masseleitung breiter sein

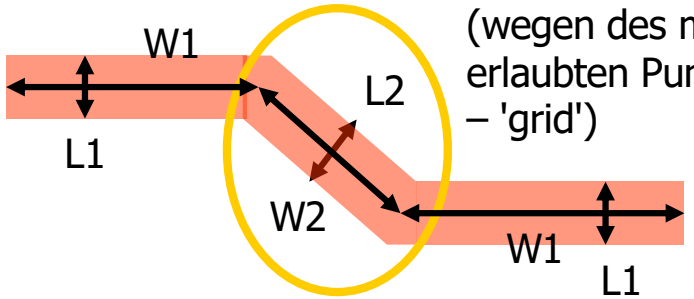
- Achtung: Im Schematic hat eine Instanz mit ‚Fingers' (m Parameter) nur 1 Pin, multiple Instanzen mehrere Pins.

'Bent Gate' Transistoren

- Das Layout wird manchmal kompakter, wenn das Gate einen Knick ('bend') macht.
- W/L ist nicht ganz eindeutig ⇒ Messungen sind erforderlich



– Gate-Geometrie schlecht definiert
– Widerstand in Drain / Source Diffusion (schmales Stück)
+ Kapazität Source / Drain kleiner
+ Layout insgesamt kleiner
+ Gut für ‚große' Transistoren in Buffern (s. nächste Folie)

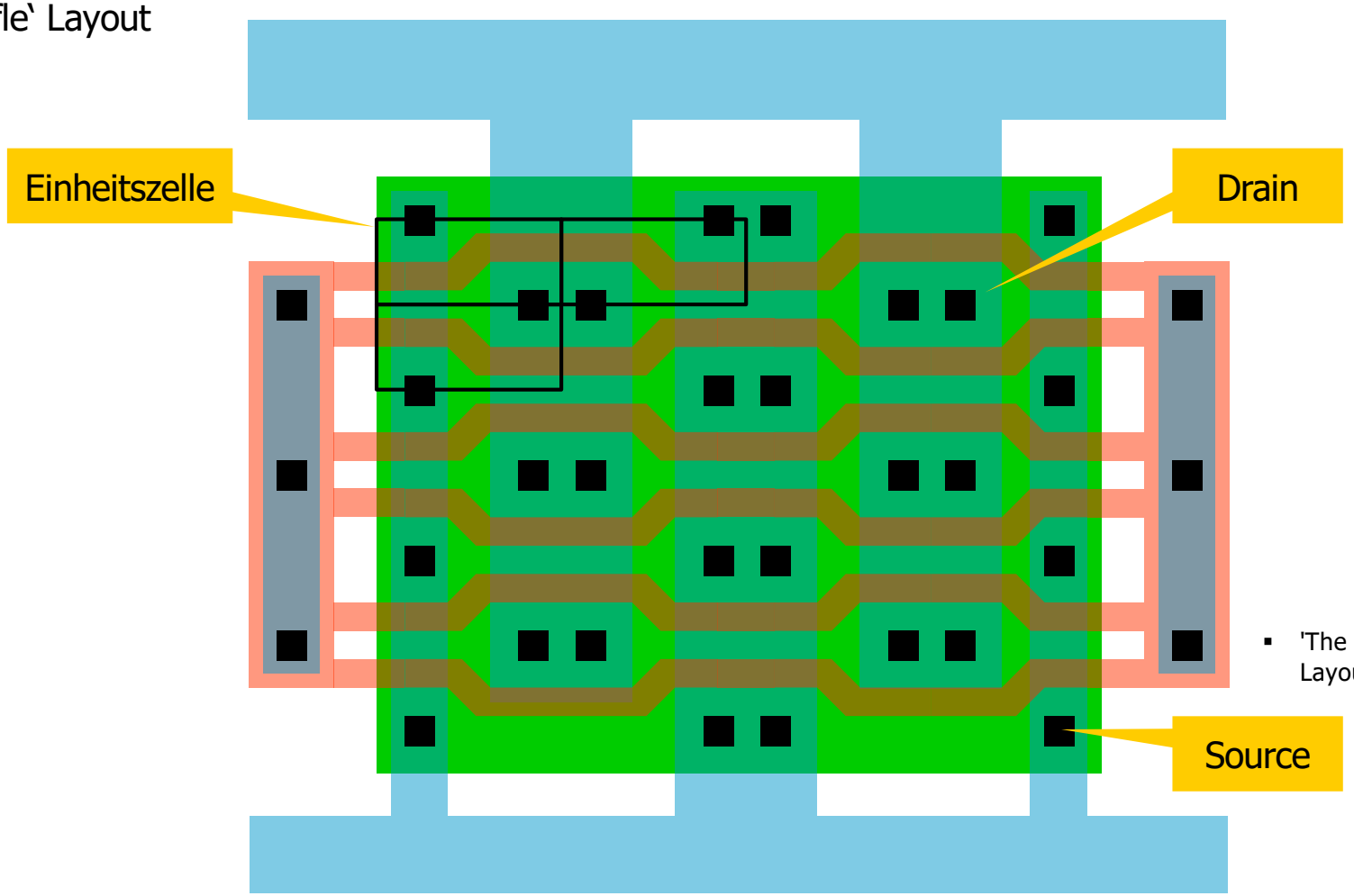


L ist nicht minimal (wegen des minimal erlaubten Punktrasters – 'grid')

Modellierung idealerweise als 2 MOS mit W1,L1 und einem mit W2,L2.

Große 'Bent Gate' Transistoren

- Geometrie für 'viel W/L pro Fläche' bei gleichzeitig kleinen Zuleitungswiderständen
- ‚Waffle' Layout

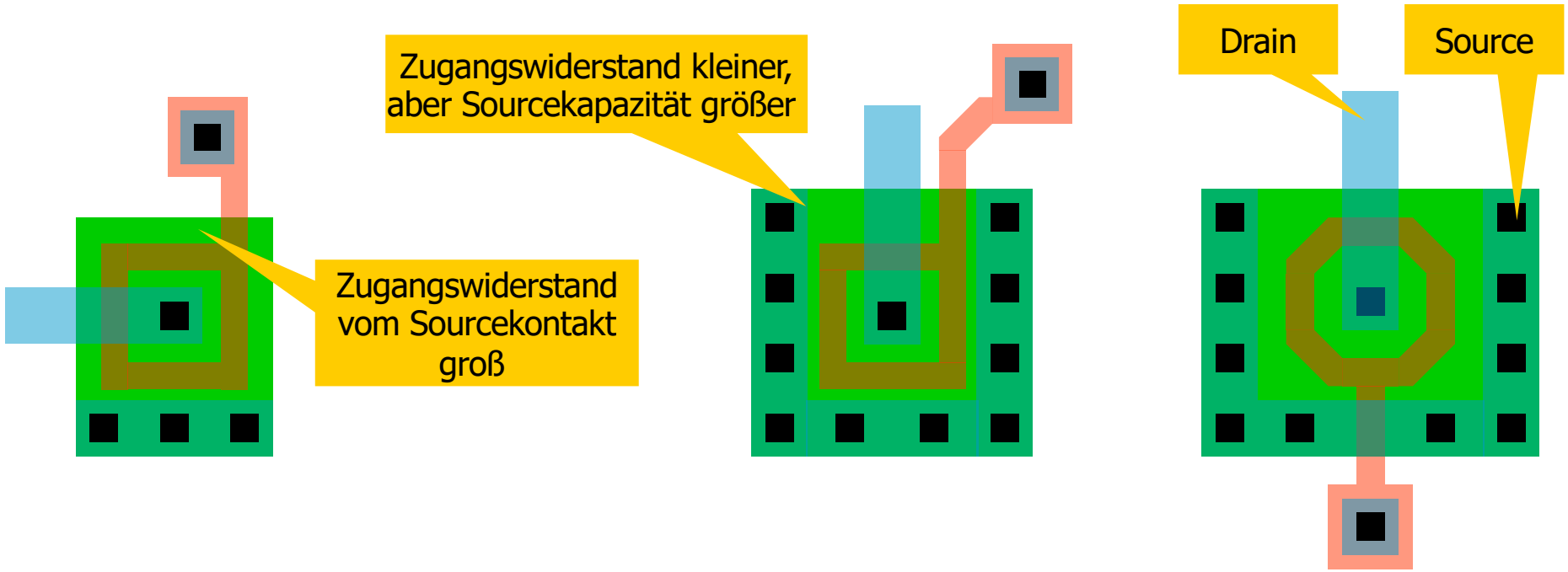


- 'The Art of Analog Layout', p. 417

- Achtung: Bei hohen (DC) Strömen kann sich die Struktur (zu) stark erhitzen. Daher evtl. doch größer machen!

'runde' Transistoren

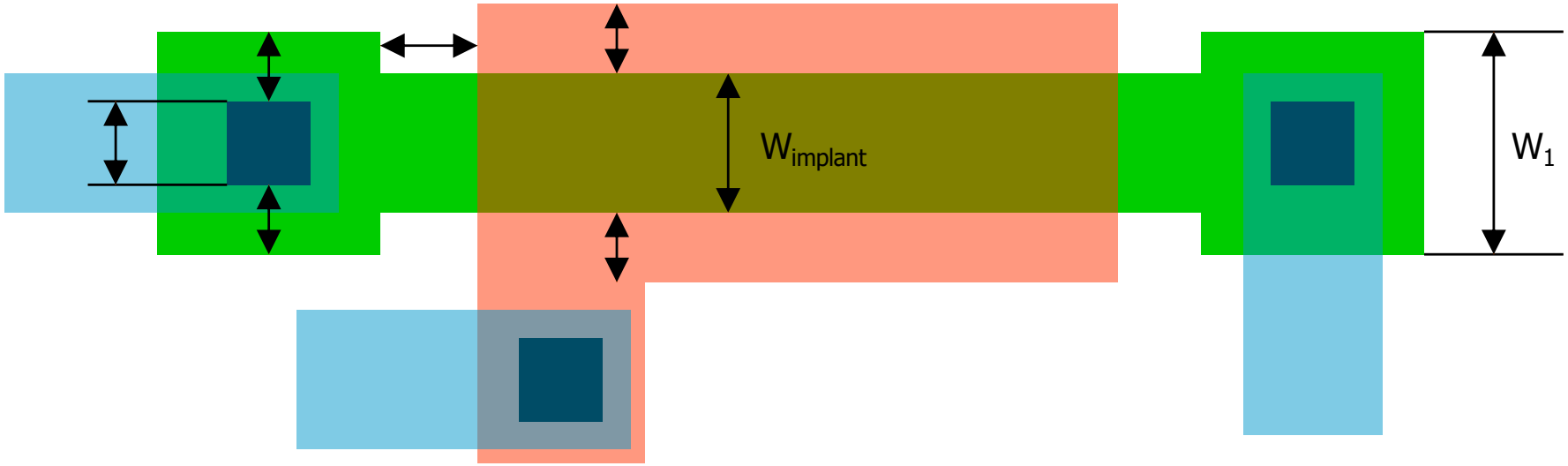
- Das kleinste (günstigste) Verhältnis von C_{GD} / g_m bekäme man mit **runden** Transistoren.
- Da diese meist die Design-Regeln verletzen (nur Vielfache von 45° erlaubt) zeichnet man **Rechtecke** oder **Achtecke**.
- Die Kapazitäten sind hier **SEHR** ungleich: C_S » C_D, C_{GS} » C_{GD}. NB: Source Cap oft unwichtig, da S=GND.
- Nützlich z.B. für Ausgangs-Pads oder Open-Drain-Netze



- Das 'effektive' W/L dieser Strukturen ist schwer zu ermitteln. Messungen an Teststrukturen sind nötig!
- Die Extraktionsprogramme scheitern hier oft völlig (unsinnige L-Werte). Ok für kleine L.

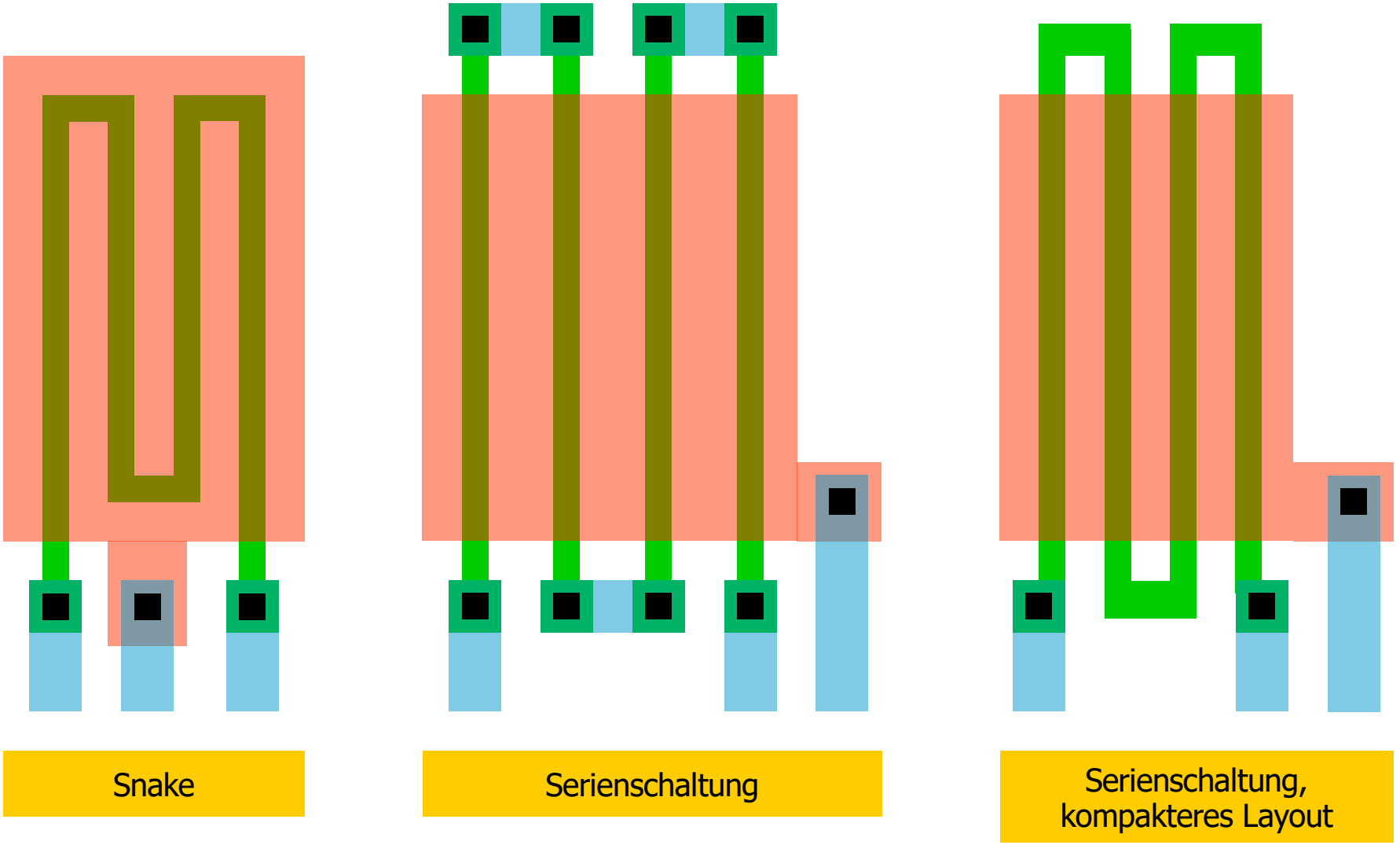
lange Transistoren (L groß)

- Minimales W der Implantation am Kontakt meist definiert durch $W_1 = W_{Kontakt} + 2 \times W_{Überlapp}$



Sehr lange Transistoren (L groß)

- Layout meist als 'Snake'. Besser definierte Länge bei Serienschaltung von mehreren linearen MOS.

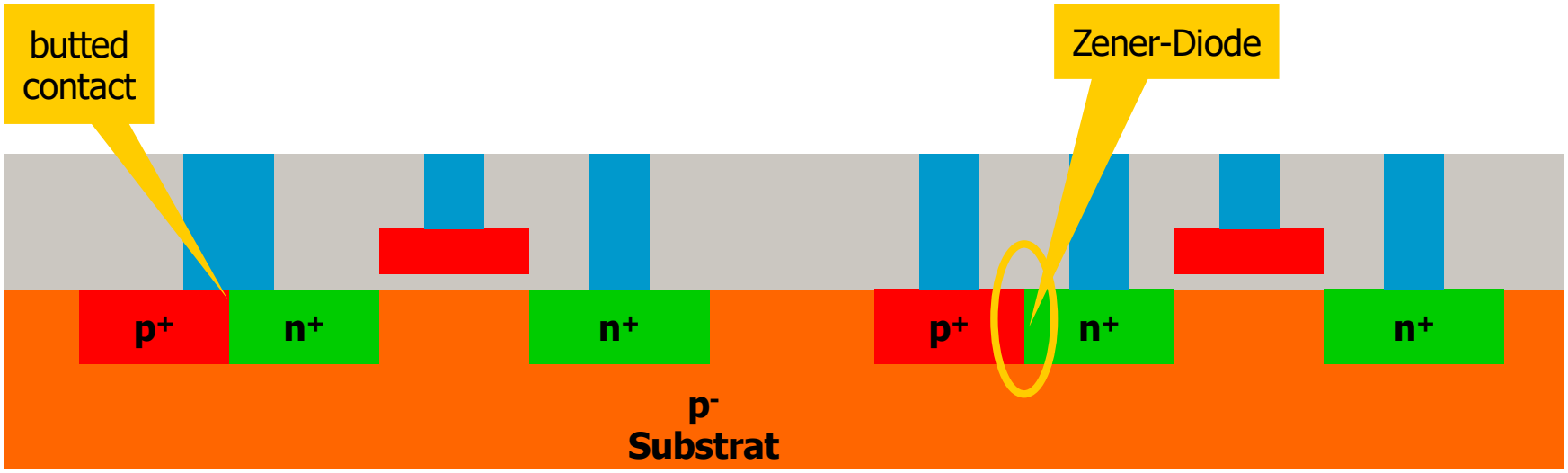


Butted Contacts

- Um kompakte Substratkontakte zu erlauben kann man einen gemeinsamen Kontakt über n^+ und p^+ legen.
- Man spricht dann von 'butted contacts'
- Sie sind in vielen Technologien nicht erlaubt

- Wenn sich n^+ und p^+ berühren, bildet sich u.U. eine Zenerdiode (hohe Dotierungen!). Solange beide Seiten auf gleichem Potential sind, ist das nicht schlimm, die Diode ist kurzgeschlossen.

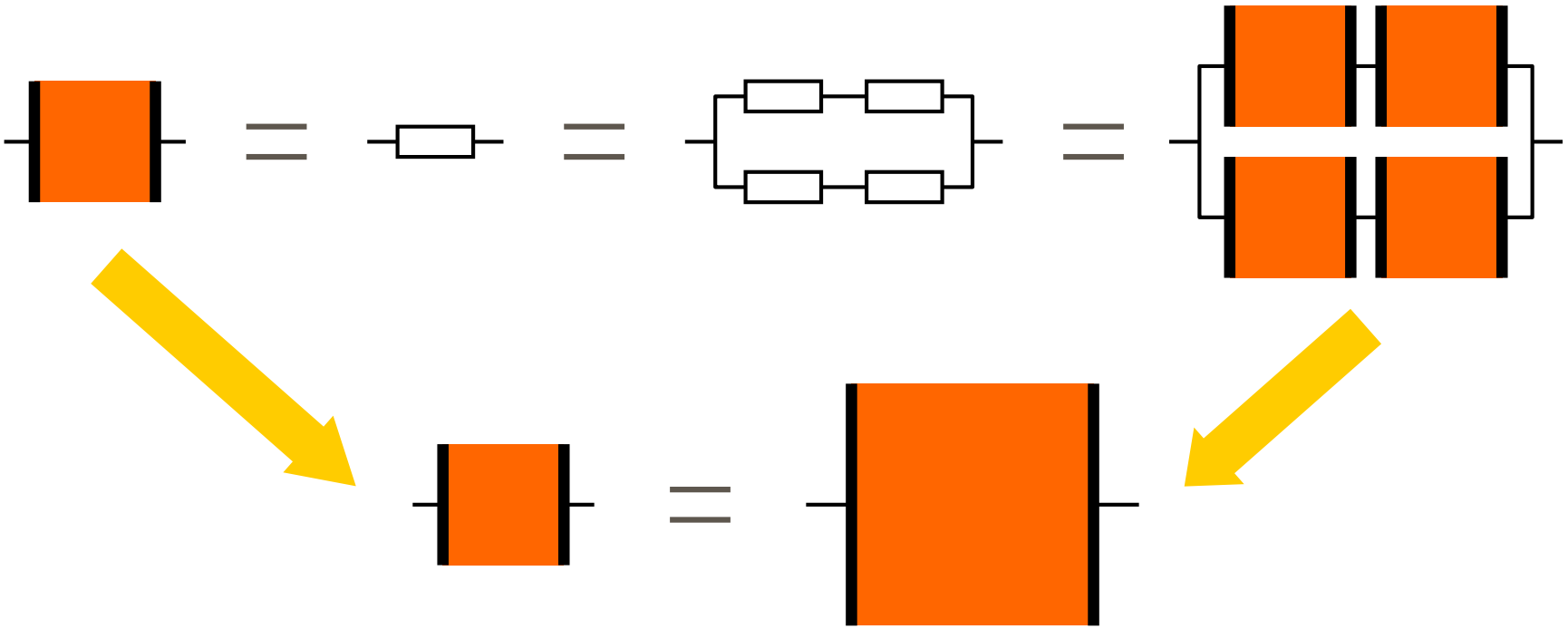
- Für 'analoge' Transistoren ist es generell nicht gut, wenn sich n^+ und p^+ berühren!



Widerstände

Widerstände

- Alle Lagen (NWELL, N+, P+, Poly, Metalle) können für Widerstände benutzt werden
- Wichtigster Parameter ist der **Flächenwiderstand $R_\square$ (sheet resistance)**
- Seine Einheit ist **'Ohm per square', $R_\square$**
- Eine Schicht der Dicke t mit dem spezifischen Widerstand ρ [Ωcm] hat den Flächenwiderstand $R_\square = \rho/t$
- Ein Quadrat hat immer den gleichen Widerstand, unabhängig von seiner Größe!



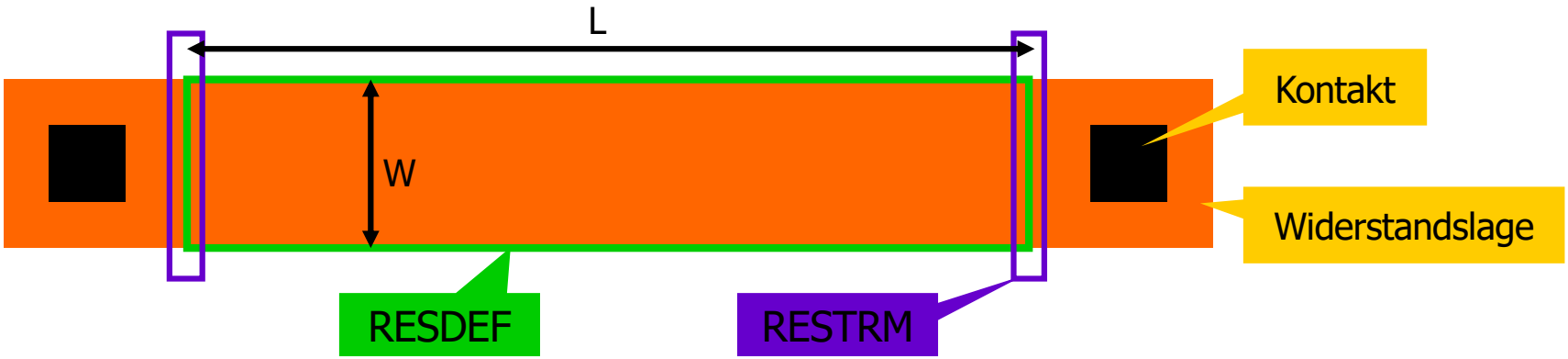
- Andere Betrachtung: Der Widerstand eines Quaders der Länge L, Breite W und Höhe t ist $R = \rho L/Wt$
  also ist der Widerstand eines Quadrats (W=L): $R_\blacksquare = \rho/t = R_\square$ unabhängig von der Größe

Rechteckige Widerstände

- Ein rechteckiger Widerstand hat den Wert

$$R = L/W \cdot R_{\square}$$

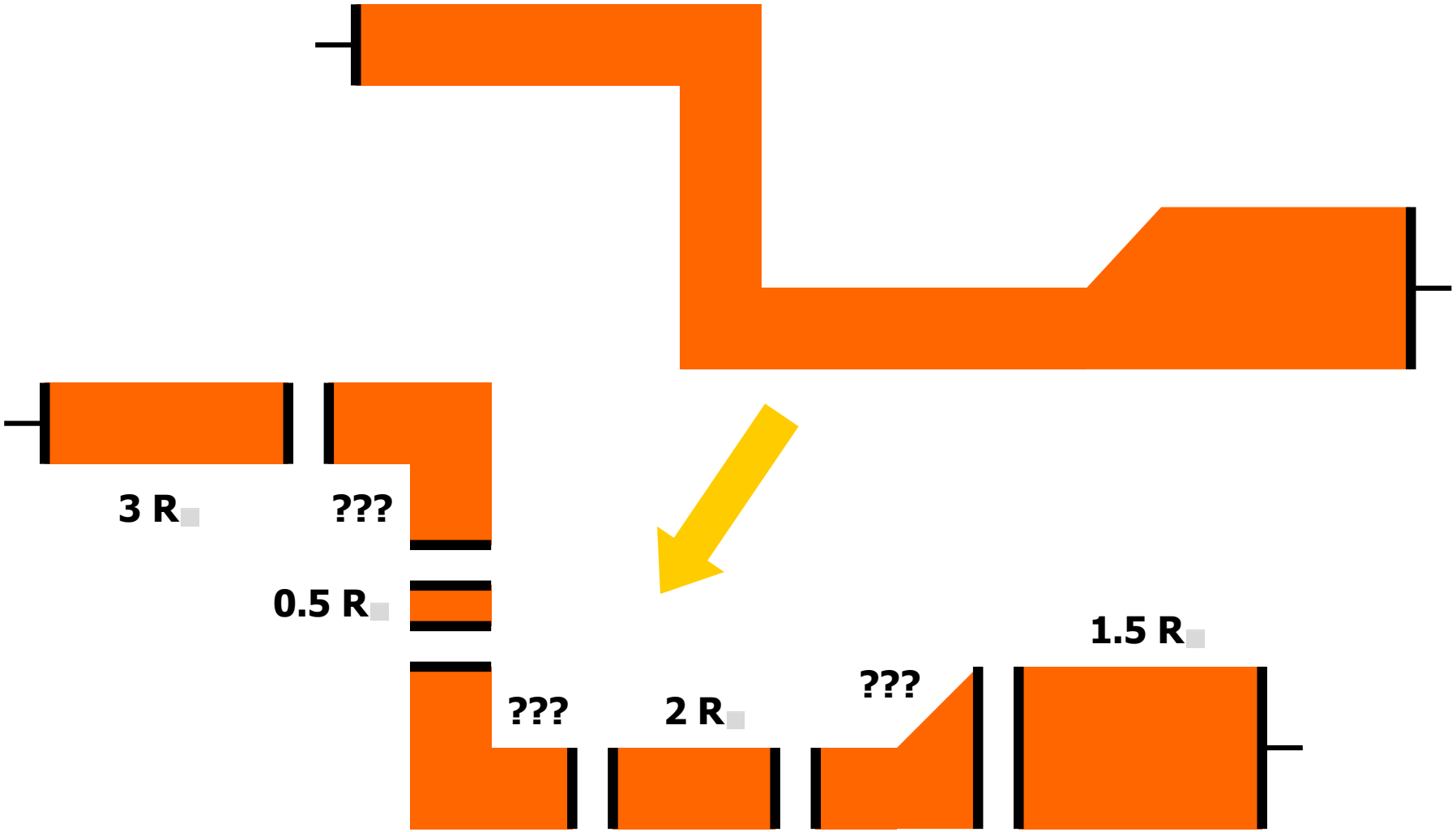
- Problem: Was sind W und L beim folgenden Layout ?



- Um der Extraktion zu 'helfen' werden der Widerstand und die Anschlüsse mit Sonderlagen markiert (in AMS Technologie: ‚RESDEF' und ‚RESTRM', in UMC: nur ‚PSYMBOL' anstelle ‚RESDEF')
- Details der Extraktion hängen vom Design-Kit der Technologie ab
- Strukturen ohne diese Lagen werden NICHT als Widerstände extrahiert!
- Der kurze Weg bis zu den Kontakten und die Kontakte selbst tragen auch einen kleinen Widerstand bei. Dieser ist meist vernachlässigbar
- Bei der Herstellung können laterale Dimensionen verkleinert oder vergrößert werden (z.B. durch Diffusion). Daher ist z.B. die wahre Breite $W = W_{drawn} - W_{offset}$. W_{offset} ist ein Parameter der Technologie.

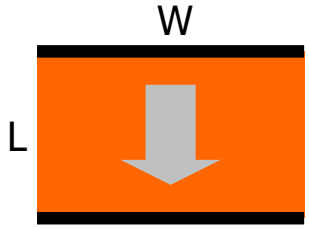
Widerstände mit allgemeiner Form

- Bei allgemeineren Formen muß (im Prinzip) die Poisson-Gleichung gelöst werden.
- **Näherungsweise** kann eine Struktur in einfache Elemente zerlegt werden:

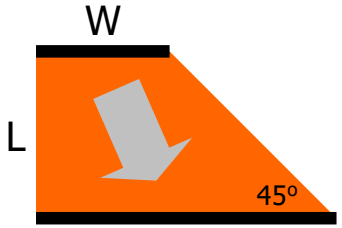


Kompliziertere Geometrien

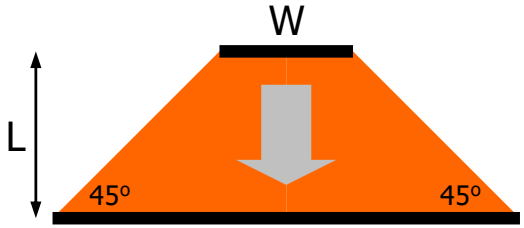
- Für einige einfache Formen gibt es Formeln:



$R_{\square} \cdot L/W$

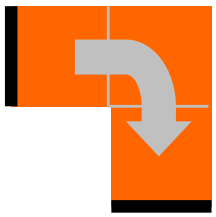


$R_{\square} \cdot 4L/(L+4W)$

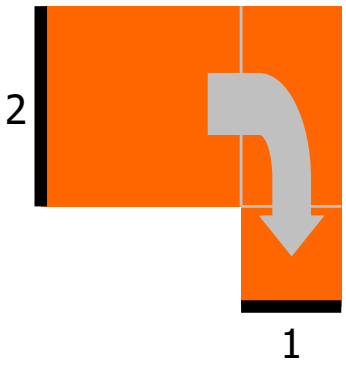


$R_{\square} \cdot 2L/(L+2W)$

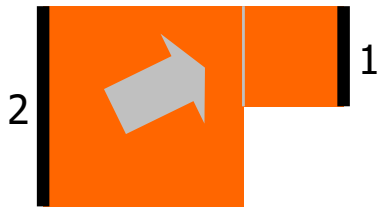
- Für elementare Ecken etc. gibt es Tabellen mit ungefähren Werten, z.B.:



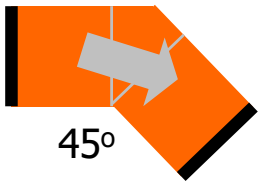
~2.5$R_{\square}$
(2.53...2.65 $R_{\square}$)
(my Tool: 2.52)



~2.55 $R_{\square}$



~2.25 $R_{\square}$
(my Tool: 2.29)



~2.2 $R_{\square}$
(auch
2.33$R_{\square}$)

~2.96 $R_{\square}$

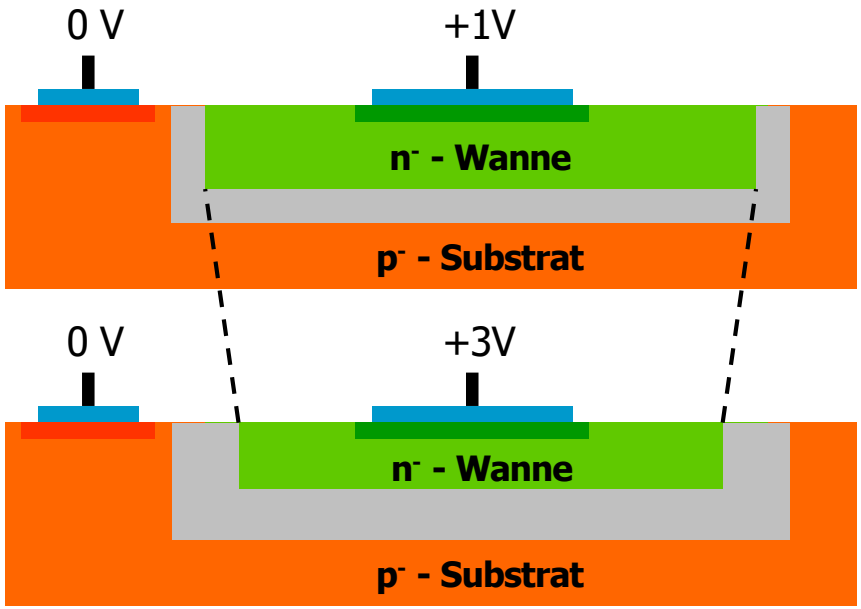
Große Widerstände

- Um hohe Widerstandswerte zu bekommen, muß man die Widerstände **lang** und **schmal** machen.
- Die Breite kann man durch ein **'dogbone'** (Hundeknochen) - Layout reduzierten.
  Achtung: Fehler in W durch W_{offset} wirken sich hier sehr stark aus !

- Lange Widerstände werden gefaltet (**'serpentine'**, **'snake', 'meander'**)

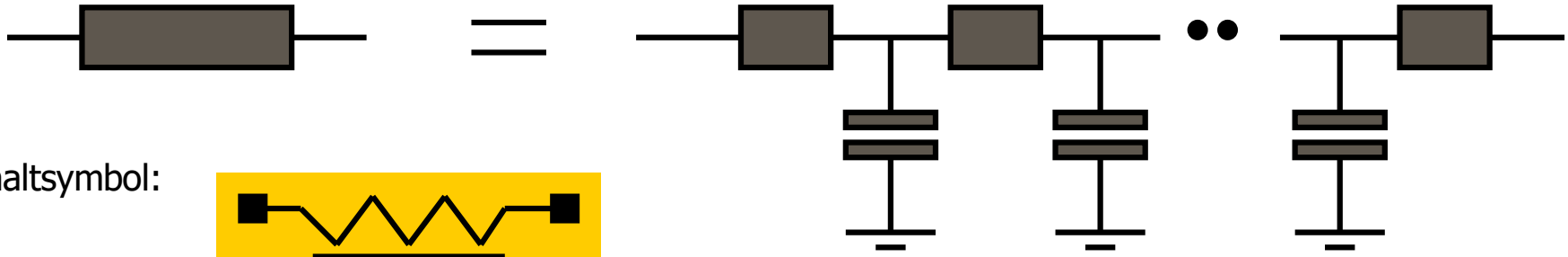
Typische Werte für $R_{\square}$

- NWELL 1000 Ω / □
- ACTIVE (N+ oder P+) 100 Ω / □
- Poly (mit Silicide) 10 Ω / □
- Metall 0.1 Ω / □

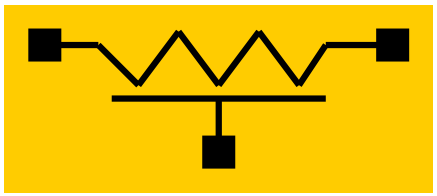


- Bei der Auswahl sind außerdem zu beachten:
  - **Toleranzen** (Min/Max-Werte der Technologie)
  - **Temperaturabhängigkeit** (z.B. bei NWELL)
  - **Spannungsabhängigkeiten** (Verarmungszonen!)

- Die Widerstände bilden **Kapazitäten**! (NWELL, ACTIVE). Modellierung daher oft als verteiltes RC-Netz:

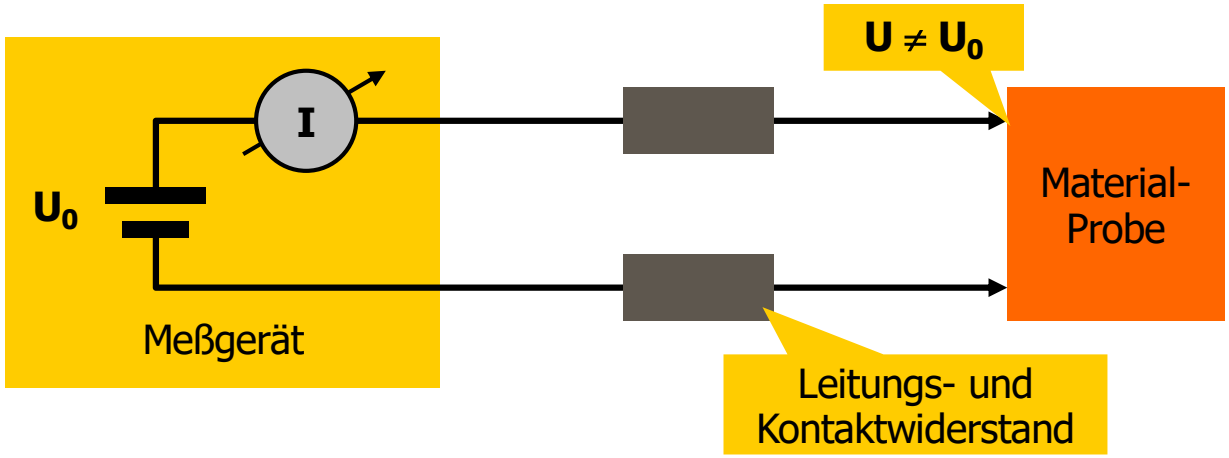


- Schaltsymbol:

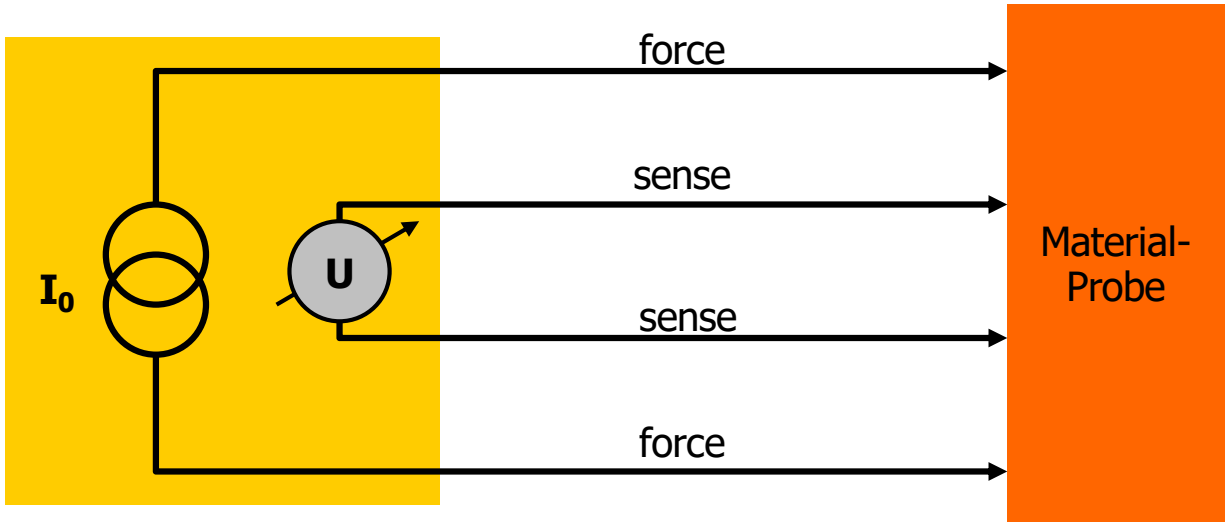


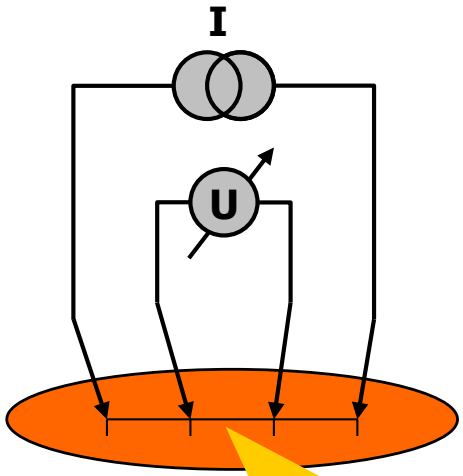
Messung von $R_{\square}$

- Problem bei Messung mit zwei Nadeln: Spannungsabfall in den Leitung und Kontaktwiderstand



- klassische Lösung: Vierpunktmessung





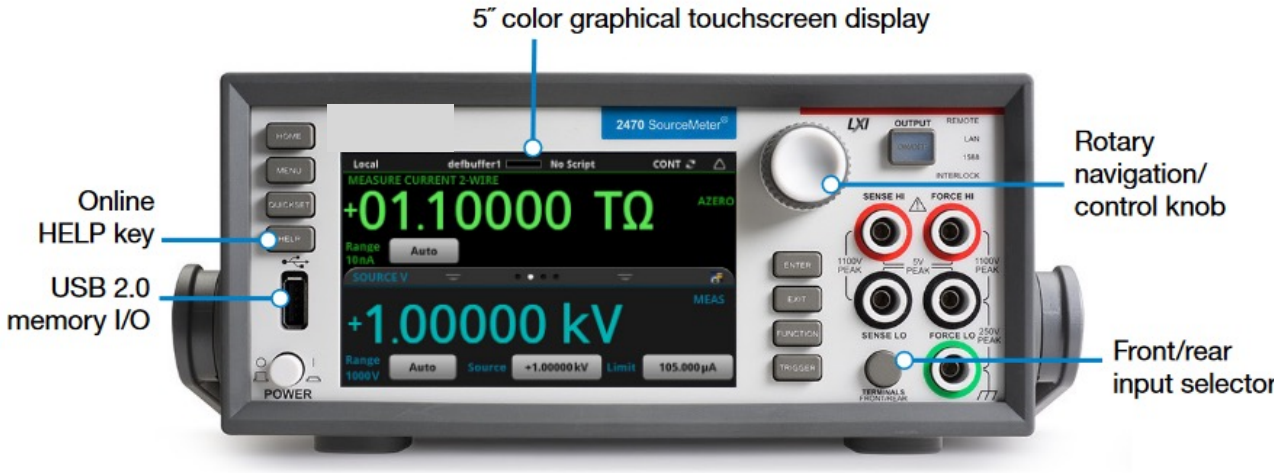
Wird bei 4 äquidistanten Meßspitzen der Widerstand R gemessen, so ist

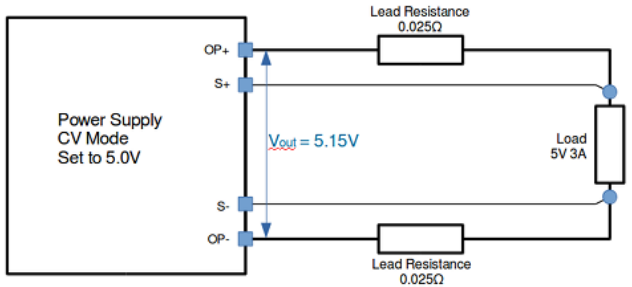
$R_{\square} = U/I \cdot \pi / \ln(2)$

(Die Herleitung erfordert nur elementare Elektrostatik!)

Meßgeräte mit Force / Sense

- Methode wird von Meßgeräten und Power supplies genutzt

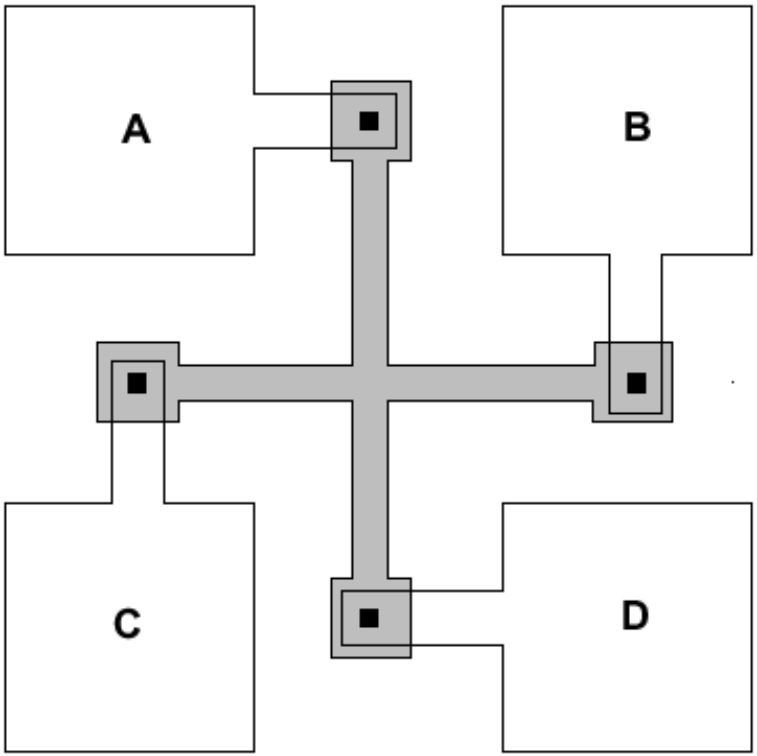




https://www.aimtti.com/remote-sense

Messung an Teststrukturen

- Mit speziellen Teststrukturen gibt es viele Möglichkeiten
- Man muß aber immer beachten, daß die Geometrie u.U. nicht bekannt ist (Überätzen etc.)
- Beispiel: Das 'Griechische Kreuz':



A.J.Walton: Microelectronic Test Structures

- Strom wird über AB aufgeprägt
- Spannung wird an CD gemessen
- $R_{\square} = U_{CD} / I_{AB} \times \pi / \ln(2)$
- Verbesserung: Mehrere Messungen mit vertauschten Anschlüssen und Mittelung
- Die 'Arme' müssen deutlich länger als das mittlere Quadrat sein
- Fehler dann bis 0.1%
- Der Meßstrom darf nicht zu hoch sein, damit sich die Struktur nicht zu stark erwärmt

Erwärmung

- Durch die in einem Bauteil (Widerstand, Transistor) anfallende Verlustleistung P_0 erwärmt sich das Silizium lokal. Durch den entstehenden Temperaturgradienten wird die Wärme nach außen geleitet. Nach längerem Betrieb stellt sich eine stabile Temperaturverteilung ein.
- Um die Erwärmung abzuschätzen, betrachten wir zunächst eine punktförmige Wärmequelle mit der Leistung P am Koordinatenursprung in einem unendlich ausgedehnten Silizium (nach allen Seiten)
- Der Wärmefluss $\vec{j}(\vec{x})$ wird durch den Gradienten in der Temperaturverteilung $T(\vec{x})$ angetrieben:

$$\vec{j} = -\sigma \cdot \vec{\nabla} T$$

(stationärer Fall). Dabei ist σ = 163 W/(mK) die Wärmeleitfähigkeit von Silizium.

- Solange keine Energie eingebracht wird (also **außerhalb** der punktförmigen Wärmequelle), muss die Divergenz von j zur Energieerhaltung verschwinden, also $\vec{\nabla}\vec{j} = 0$, so dass wir einfach

$$0 = \vec{\nabla}\vec{j} = -\sigma \cdot \Delta T \rightarrow \Delta T(\vec{x}) = 0$$

haben. (Nur an der Wärmequelle im Ursprung ist die Divergenz nicht Null.)

- Setzt man aus Symmetriegründen eine rein radiale Verteilung T(r) an, so lautet die obige Gleichung mit dem Laplace Operator in Kugelkoordinaten (nur der radiale Anteil ist relevant) einfach:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial T(r)}{\partial r}\right) = 0$$

Erwärmung

- Die Lösung dieser Gleichung ist einfach

$$r^2 \frac{\partial T(r)}{\partial r} = b \quad \rightarrow \quad \frac{\partial T(r)}{\partial r} = \frac{b}{r^2} \quad \rightarrow \quad T(r) = a - \frac{b}{r}$$

- Weit weg von der Hitzequelle ($r \longrightarrow \infty$) befinde sich das Silizium auf Umgebungstemperatur $\rightarrow a = T_0$.
- Der Skalierungsfaktor b wird von der eingespeisten Leistung P_0 bestimmt, die durch jede Kugelschale mit Radius r fließt:

$$P_0 = 4\pi r^2 \cdot j(r) = -4\pi r^2 \cdot \sigma \cdot \vec{\nabla} T(r) = -4\pi r^2 \cdot \sigma \cdot \frac{\partial}{\partial r} T(r) = -4\pi r^2 \cdot \sigma \cdot \frac{b}{r^2} = -4\pi\sigma b$$

- Hier haben wir wieder den Radialanteil des Gradienten in Kugelkoordinaten benutzt.
  Das Ergebnis ist unabhängig von r, wie es sein, muss, da der Gesamtfluss ja für alle r gleich ist.
- Mit dem so gefundenen $b=-P_0/4\pi\sigma$ ergibt sich:

Temperaturunterschied gegenüber T_0

$$\Delta T(r) = 2\frac{P_0}{4\pi\sigma}\frac{1}{r} \qquad \Delta T(r)[K] \approx \frac{P_0[mW]}{r[\mu m]}$$

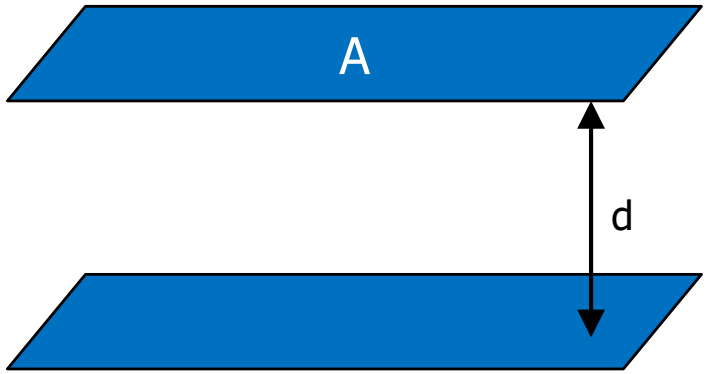
Hier muss noch die Wärmekapazität rein! 703 J/kg K

  Hierbei wurde die Leistung verdoppelt, da auf einem Chip nur ein Halbraum mit Silizium gefüllt ist.
- Man sieht, dass Leistungen im Bereich von mW noch keine hohen Temperaturen ergeben.

Kondensatoren

Idealer Plattenkondensator

- $C = \varepsilon\ \varepsilon_0\ A/d$



Kondensatoren

- Jedes Lagenpaar bildet Kapazitäten

- Man hat Anteile von **Fläche (Area)** und **Rand (Periphery, Fringe)**

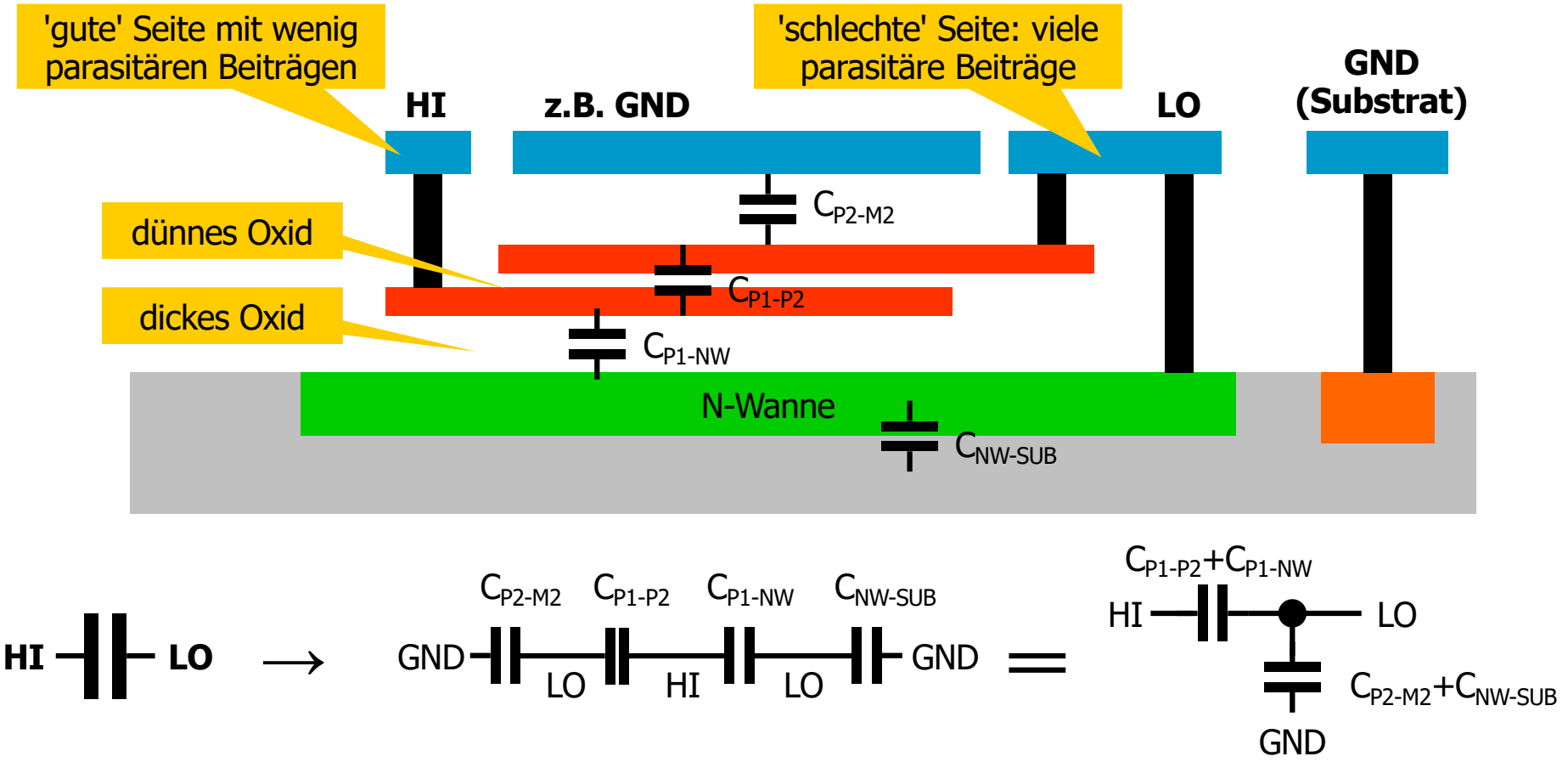
- **Hohe** Kapazitäten hat man durch das **Gate-Oxid**. **C ~ 5/10 fF/µm²** (in 350/180 nm)
  Diese Kapazität ist **spannungsabhängig** (s. VL Transistor: MOS Struktur)!
  Der Transistor muß 'on' sein ($V_{GS} > V_T$) oder in Akkumulation.
- **Lineare** Kapazitäten bekommt man immer parasitär als Poly-M1, M1-M2, ...
  - nur kleine Werte: **Area ~ 0.03 fF/µm², Peripherie ~ 0.04 fF/µm**
  - z.B. 1 µm breite Leiterbahn: (0.03 + 2 x 0.04) fF/µm

- Manche Technologien bieten einen dünnen Zwischenisolator für große, lineare Kapazitäten
  (z.B. **Poly1 – Poly2**, oder **MIM**-Cap 'Metal-Insulator-Metal' = **MMC** (Metal.Metal.Cap))
  - Dann **~ 1 fF/µm²** oder mehr

- Jede Lage trägt zu (unerwünschten) **parasitären Kondensatoren** bei!

Poly1-Poly2 Cap

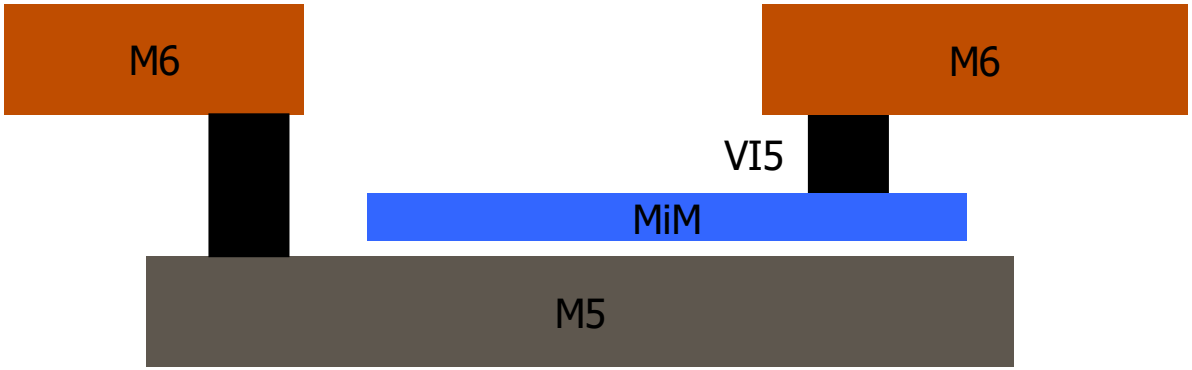
- Hier muß man beachten, daß es keine Kontakte von Poly1 nach Poly2 gibt (Herstellungstechnologie!)
- Zur Abschirmung zeichnet man oft eine NWANNE unter die Kapazität.
- Manchmal deckt man die Kapazität mit M1 ab, dann ist klar definiert, wohin die parasitären Beiträge gehen.
- Die parasitären Kapazitäten sind dann sehr unterschiedlich. Man muß den Kondensator 'richtig herum' in die Schaltung einbauen!



MiM Cap Example

- In this example, an extra metal layer 'MiM' with thin dielectric to M5 is available.
- In this case, MiM can only be on top of M5 and must be contacted from the top (vias M6)

- Cap is ~ 1fF / μm^2 (180 nm)



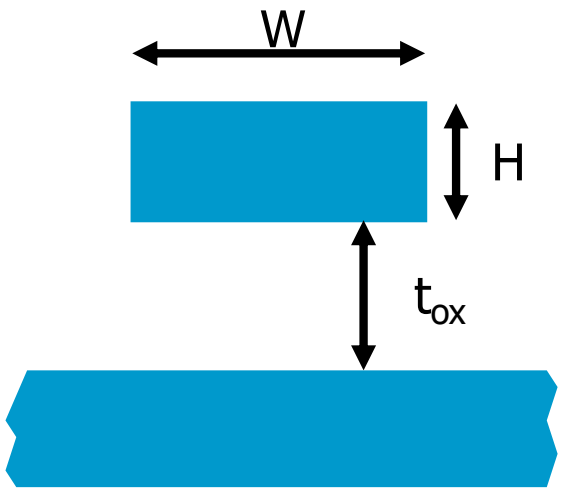
Kapazität von Leiterbahnen

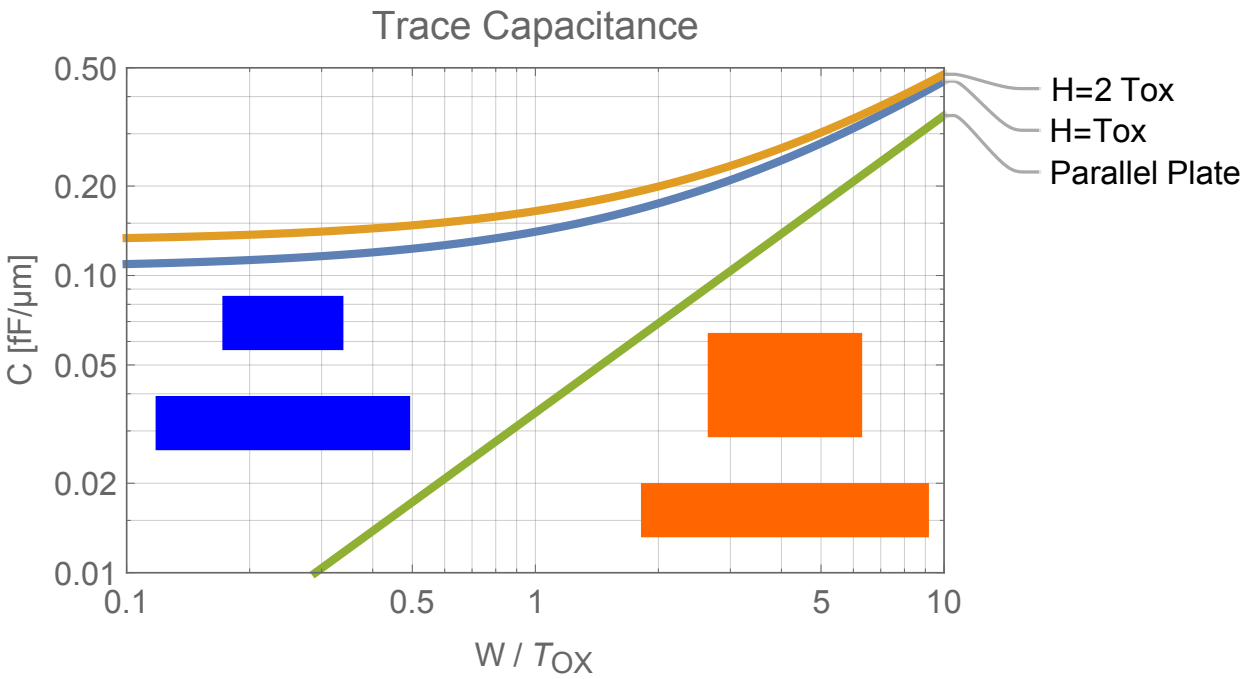
Für W/H > 0.5 gilt (~10% Genauigkeit):

$$\frac{C}{l} \approx \varepsilon\,\varepsilon_0 \left\{ w - \frac{h}{2} + \frac{2\pi}{\ln\left(1 + \frac{2}{h} + \frac{2}{h}\sqrt{1+h}\right)} \right\}$$

Yuan, Trick, 1982

$$w = \frac{W}{t_{ox}}, \qquad h = \frac{H}{t_{ox}}$$





Kapazität von Leiterbahnen

- Andere (empirische) Näherungsformeln:

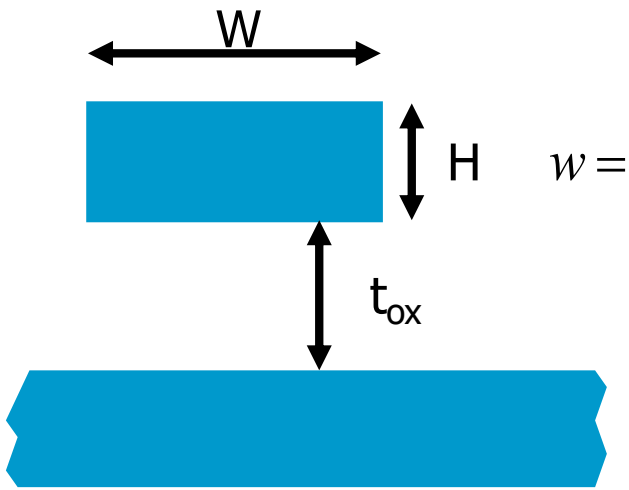
$$\frac{C}{l} \approx \varepsilon\,\varepsilon_0 \left\{1.15 \cdot w + 2.80 \cdot h^{0.222}\right\}$$

Sakurai, Tamaru, 1983

$$\frac{C}{l} \approx \varepsilon\,\varepsilon_0 \left\{w + 0.77 + 1.06 \cdot w^{0.25} + 1.06 \cdot h^{0.5}\right\}$$

Meijs, Kokkema, 1984

Sehr genau, einfache Exponenten



$$w = \frac{W}{t_{ox}}, \qquad h = \frac{H}{t_{ox}}$$

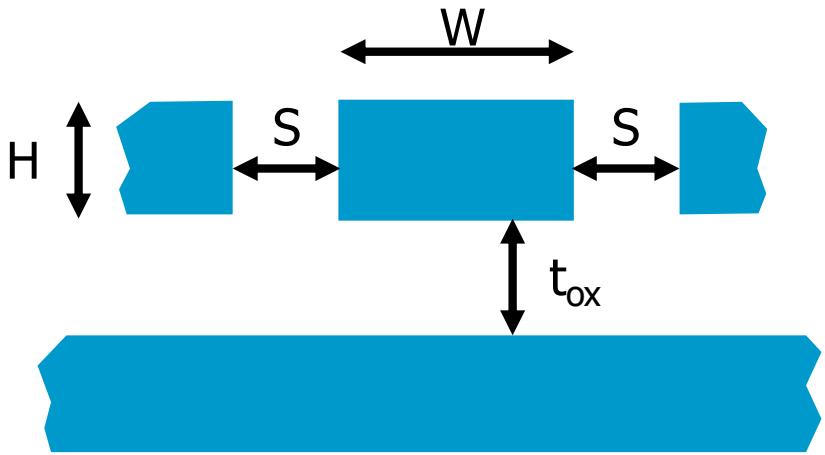
$$\varepsilon_{SiO_2}\ \varepsilon_0 \approx 3.9 \cdot 8.85 \times 10^{-12}\,\mathrm{F/m} = 34.52\,\mathrm{aF/\mu m}$$

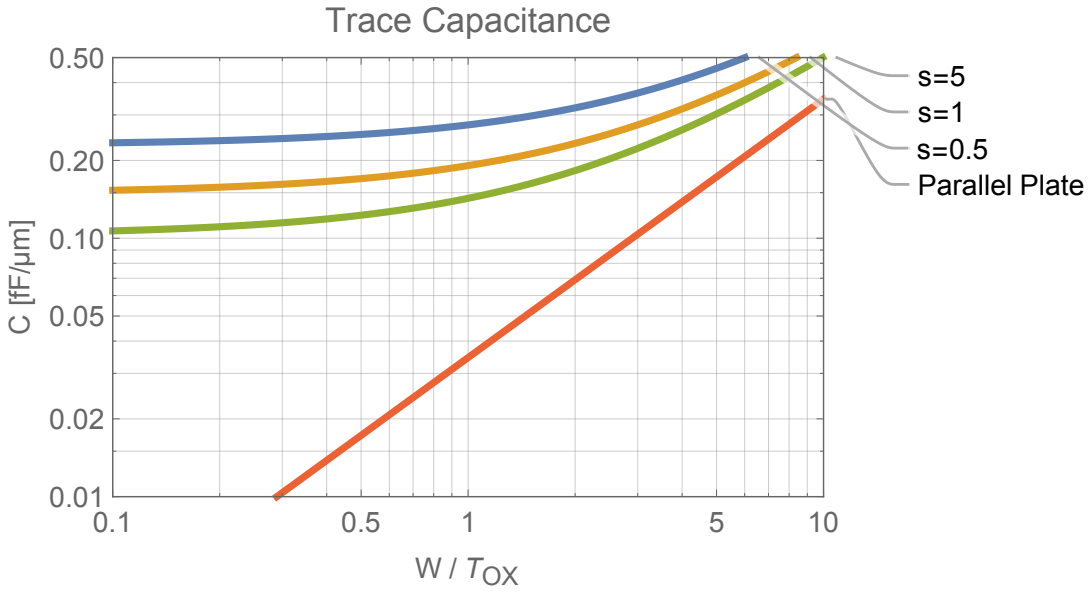
Kapazität von Leiterbahnen

Mit Nachbarn:

$$\frac{C}{l} \approx \varepsilon\,\varepsilon_0 \left\{ 1.15w + 2.8h^{0.222} + \frac{0.06w + 1.66h - 0.14h^{0.222}}{s^{1.34}} \right\}$$

$$w = \frac{W}{t_{ox}}, \qquad h = \frac{H}{t_{ox}}, \qquad s = \frac{S}{t_{ox}}$$

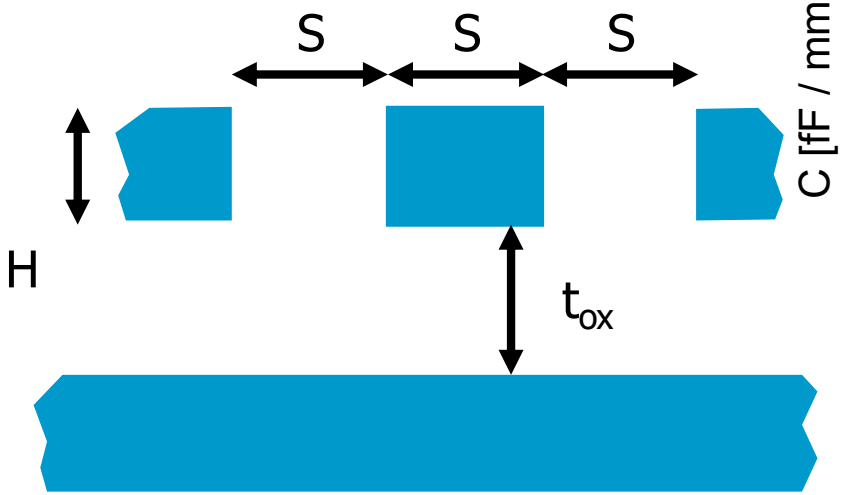


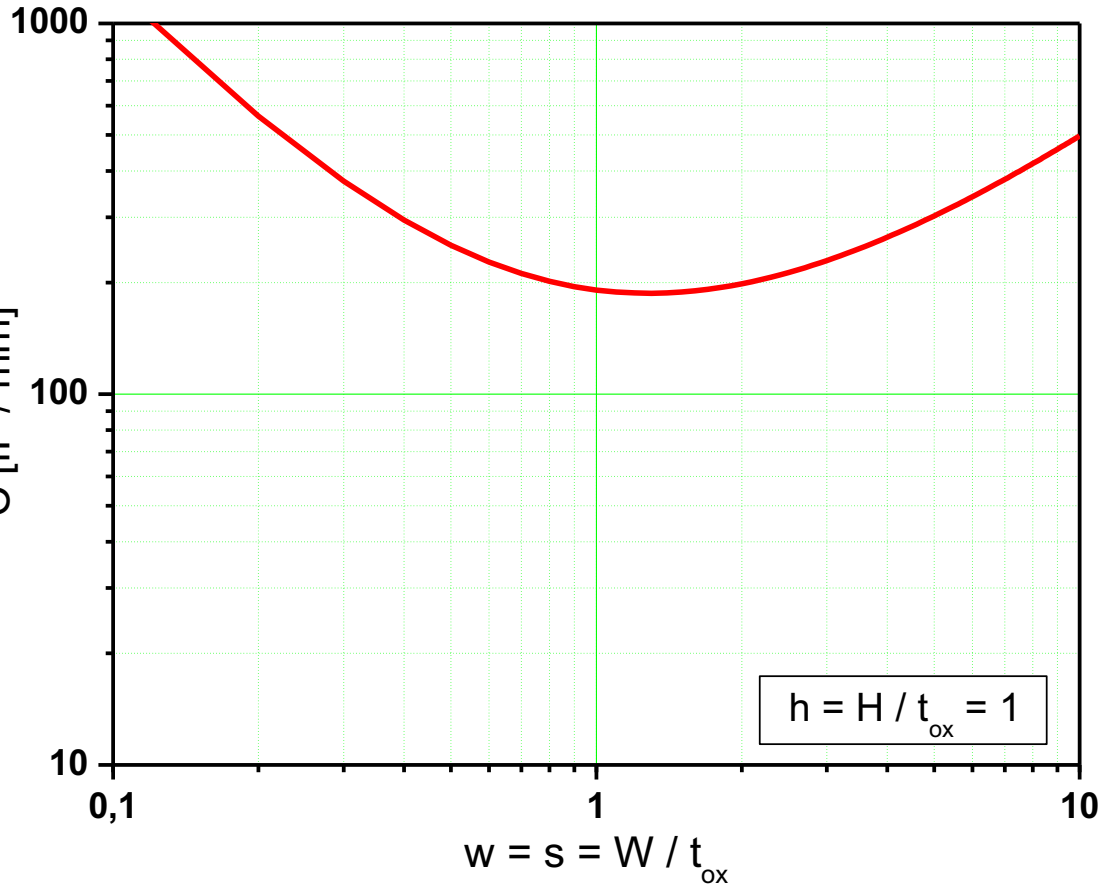


Kapazität von Leiterbahnen

Bei der Konfiguration Breite = Abstand gibt es ein minimales C!

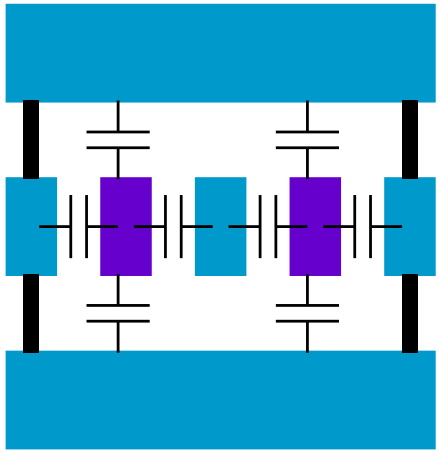
Spezialfall W = S



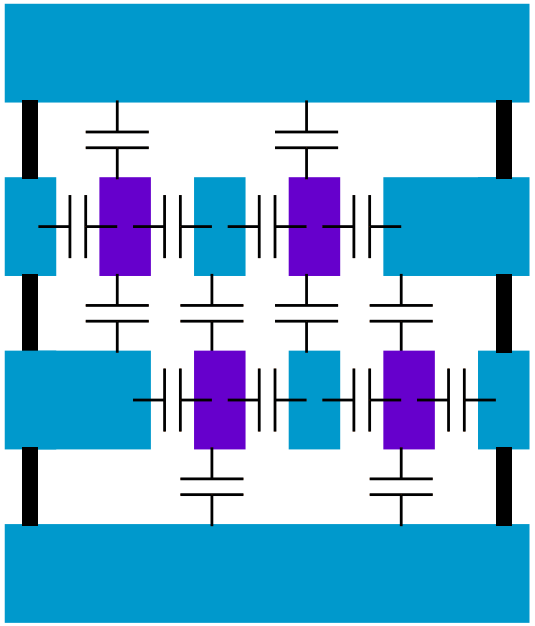


Fringe / Finger / MOM Capacitors

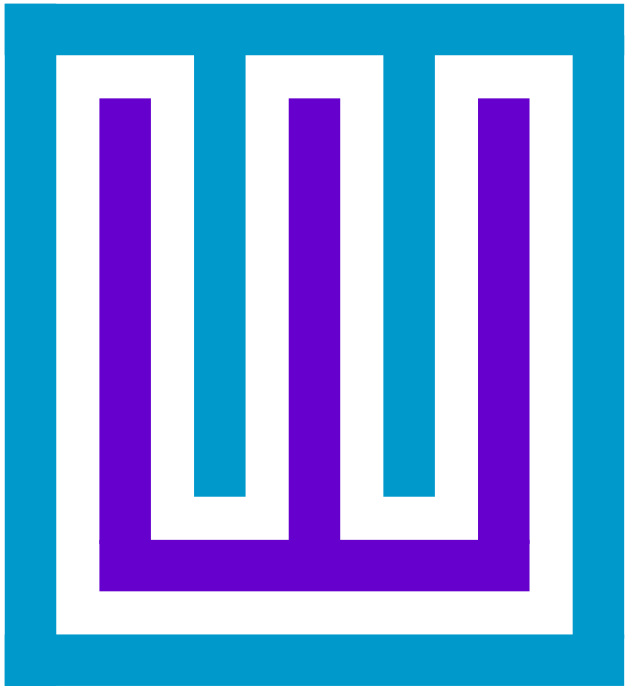
- Because modern metal structures are 'high' (in z) compared to minimal width (in x,y), the lateral 'fringe' cap can be higher that the area cap.
- → It possible and efficient to use lateral caps. These are often called 'Metal-Oxide-Metal' = MOM Caps



Side view



Side view



Top view (middle layer)

- Many layout variants are possible. Make sure to avoid stray contributions at borders!

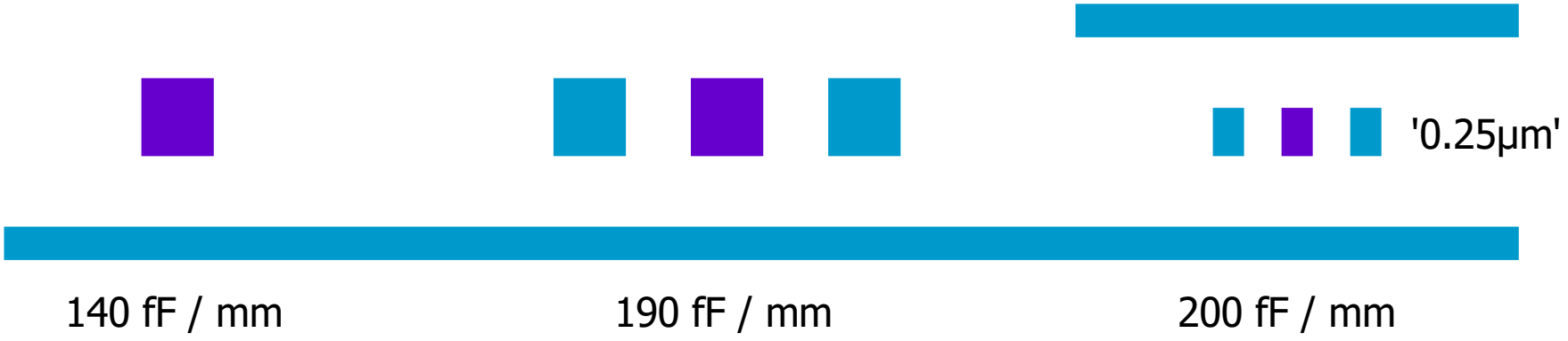
Streufelder - Zusammenfassung

- Plattenkondensator und Streufelder tragen zur Leitungskapazität bei
- **Bei schmalen Leitungen dominieren die Streufelder**
- Bei 0.25 µm Technologie und darunter dominieren oft die Cs zu den Nachbarn ('dicke Metalllagen')
- Nachbarleitungen und darüberliegende Metallebenen erhöhen die Kapazität
- Hier lohnt es sich daher, den Abstand zu erhöhen!
- Typische Werte (0.8µm Technologie): C_{area} = 35aF/µm^2, C_{finge} = 51aF/µm
- Wenn C genau definiert sein muß, sollten die Streufelder genau definiert sein:

- Schlecht:

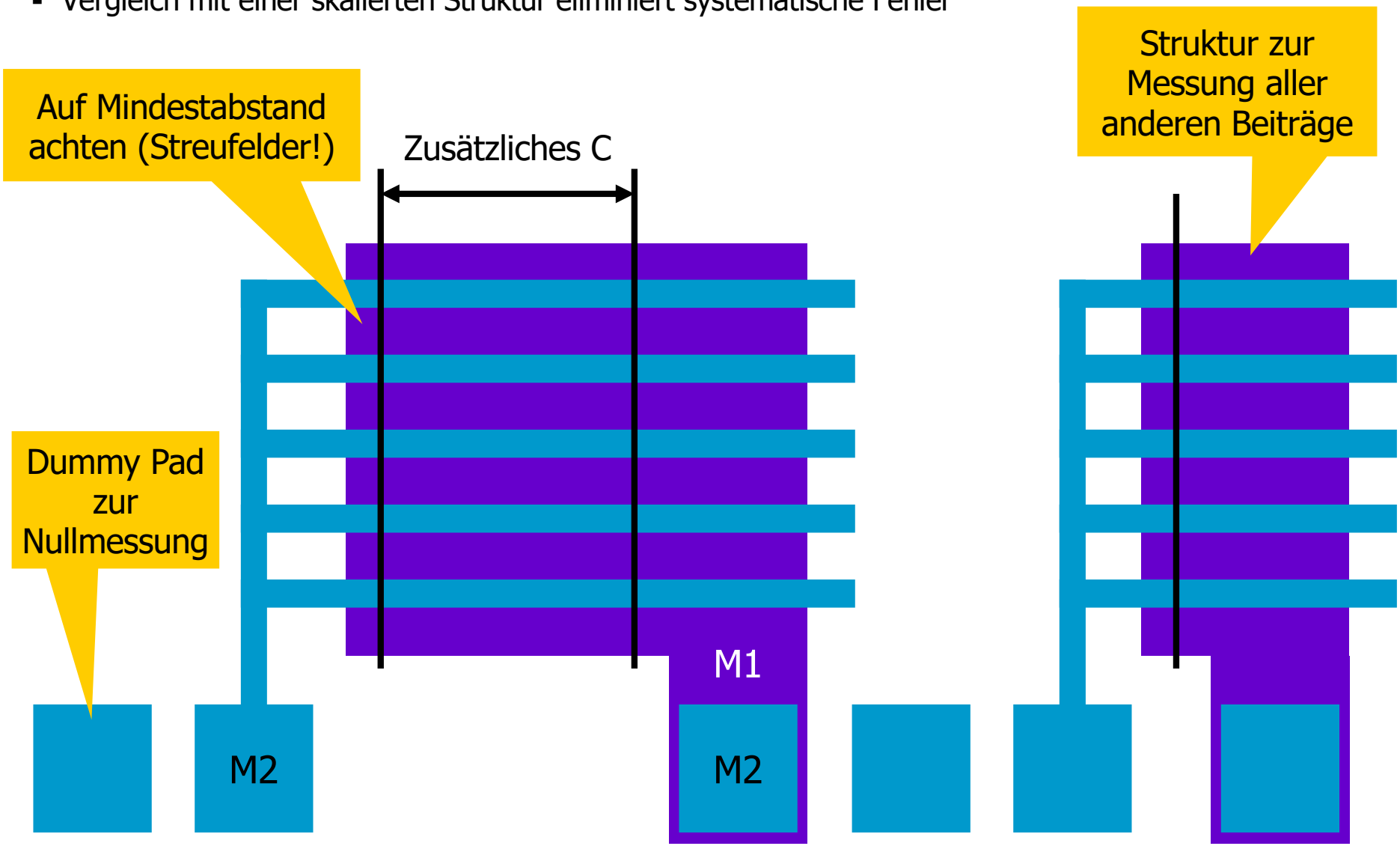
Besser:

- Typische Topologien:



Layout zur Kapazitätsmessung am Spitzenmeßplatz

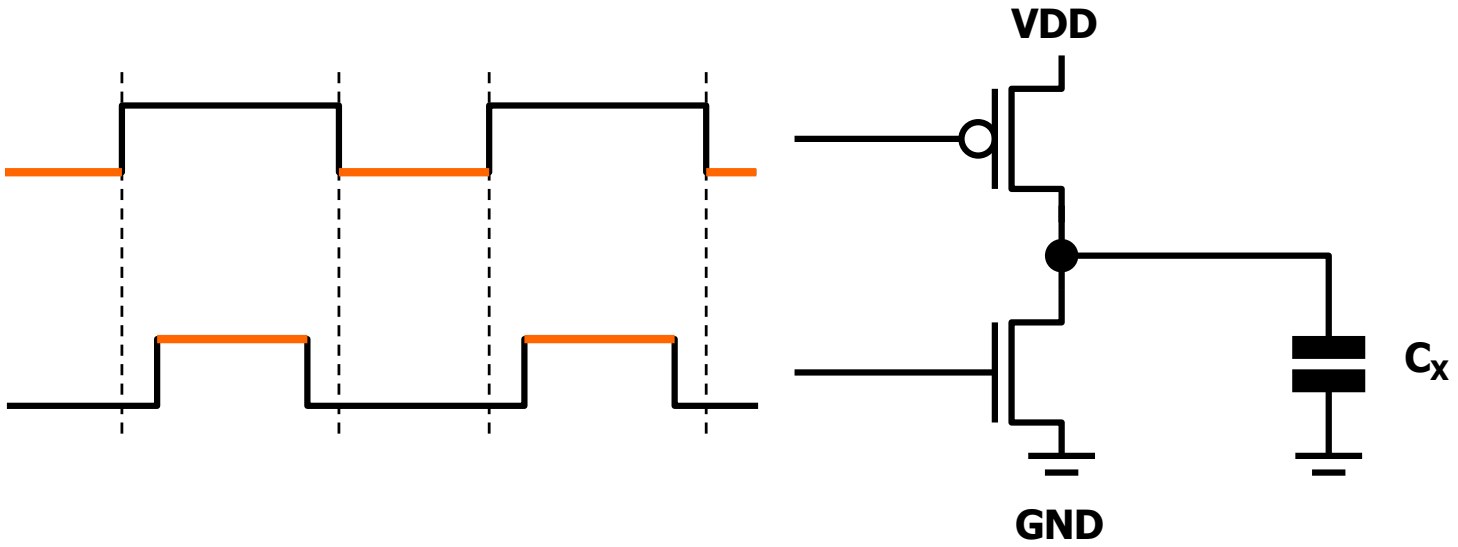
- Parallelschaltung vieler Elemente reduziert den Meßfehler
- Vergleich mit einer skalierten Struktur eliminiert systematische Fehler



Kapazitätsmessung mit Ladungspumpe

- Sehr einfache und genaue Methode:
- Die unbekannte Kapazität wird periodisch auf Spannung aufgeladen und entladen (nicht überlappende Takte PHI1 und PHI2)
- Der mittlere Strom wird gemessen.
  Pro Zyklus wird die Ladung VDD x C_X von VDD nach GND transportiert.

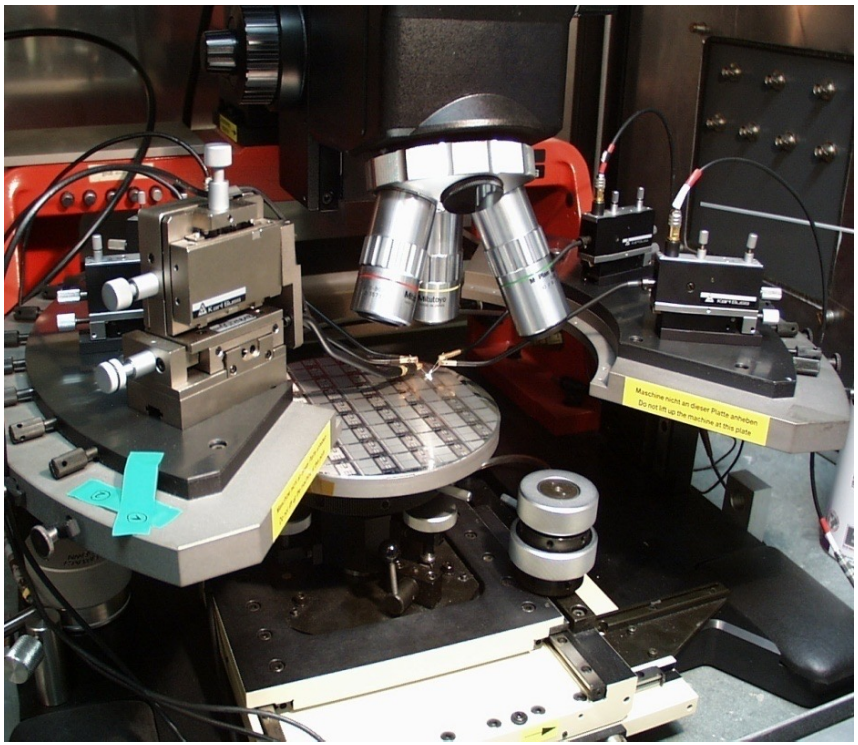
  Der mittlere Strom ist also **$I = f \times VDD \times C_X$**

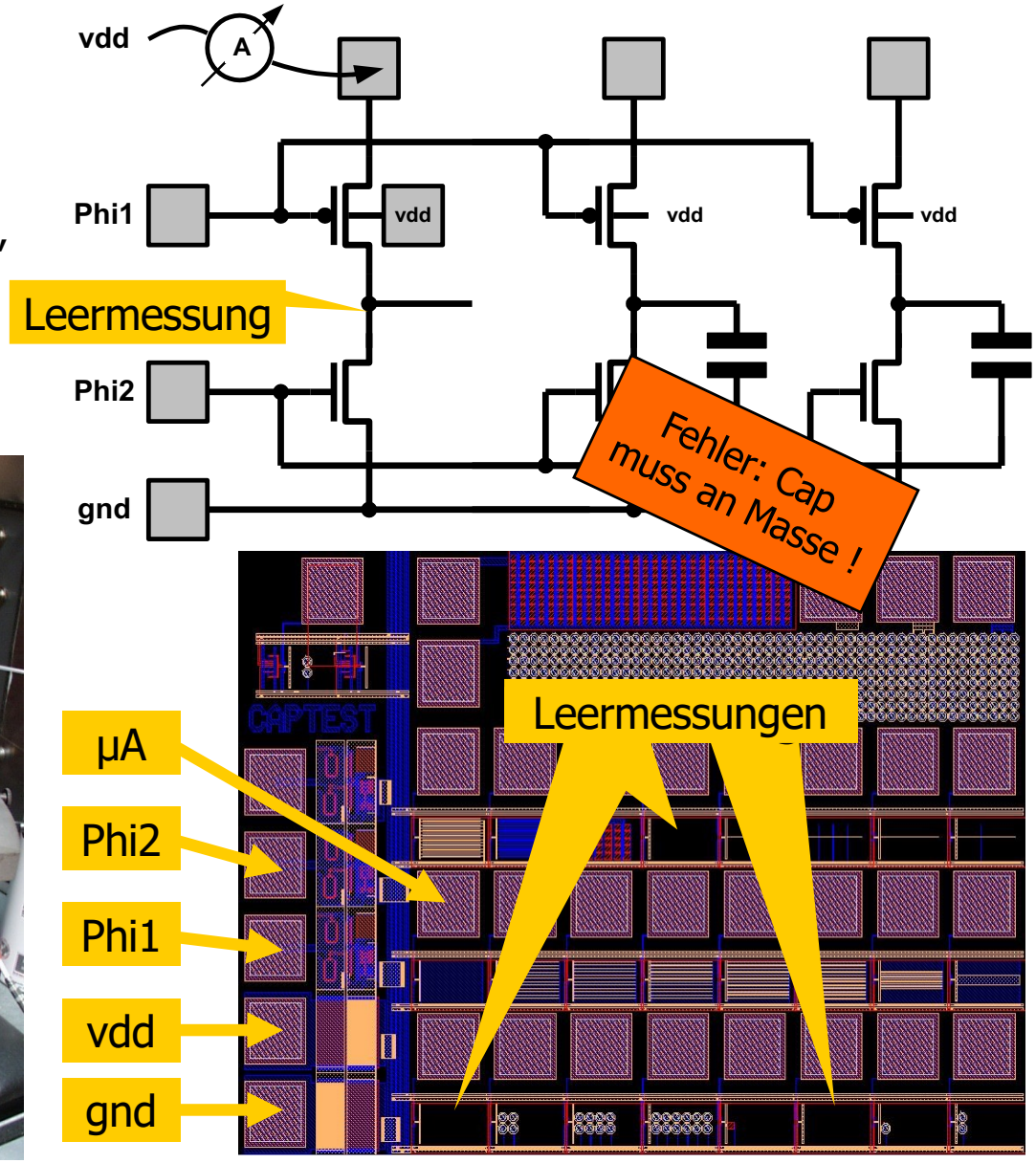


Referenz: J.C.Chen et al., An On-Chip Interconnect Capacitance Characterization Method with Sub-Femto-Farad Resolution, IEEE Trans. on Semiconductor Manufacturing, Vol. 11, No.2, May 1998.

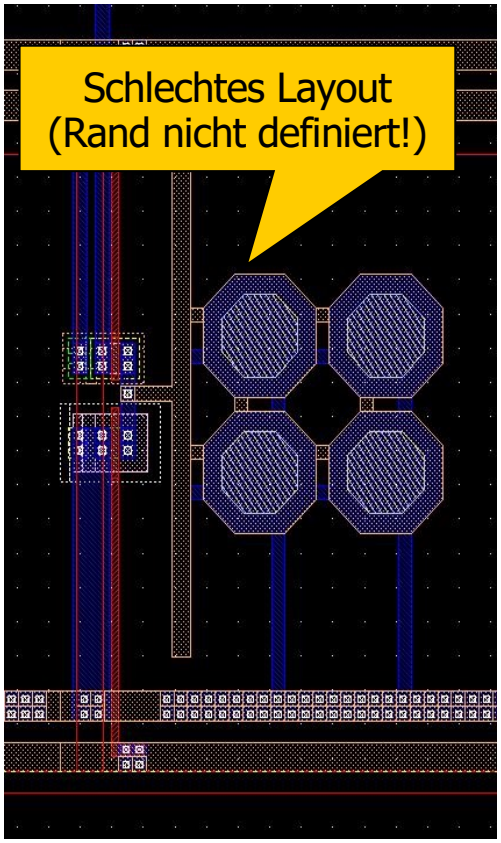
Messung auf Wafer am Spitzenmeßplatz

- Messung mit 5 Nadeln
- Nur eine Nadel wird bewegt
- Geräte:
  Pulsgenerator, Amperemeter
- Alle parasitären Elemente (Drain-Kapazitäten, Leitungen) werden durch Leermessungen ermittelt und abgezogen.

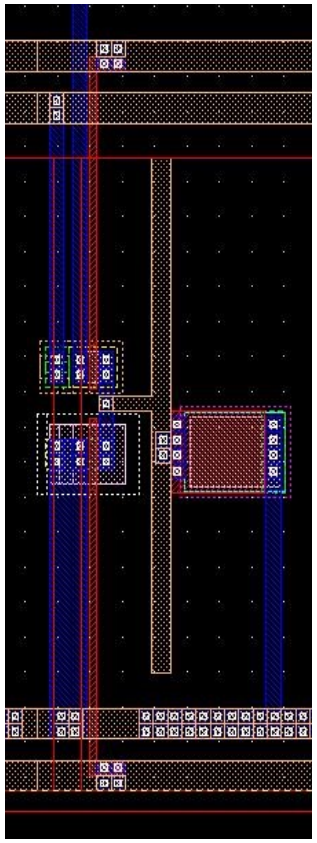




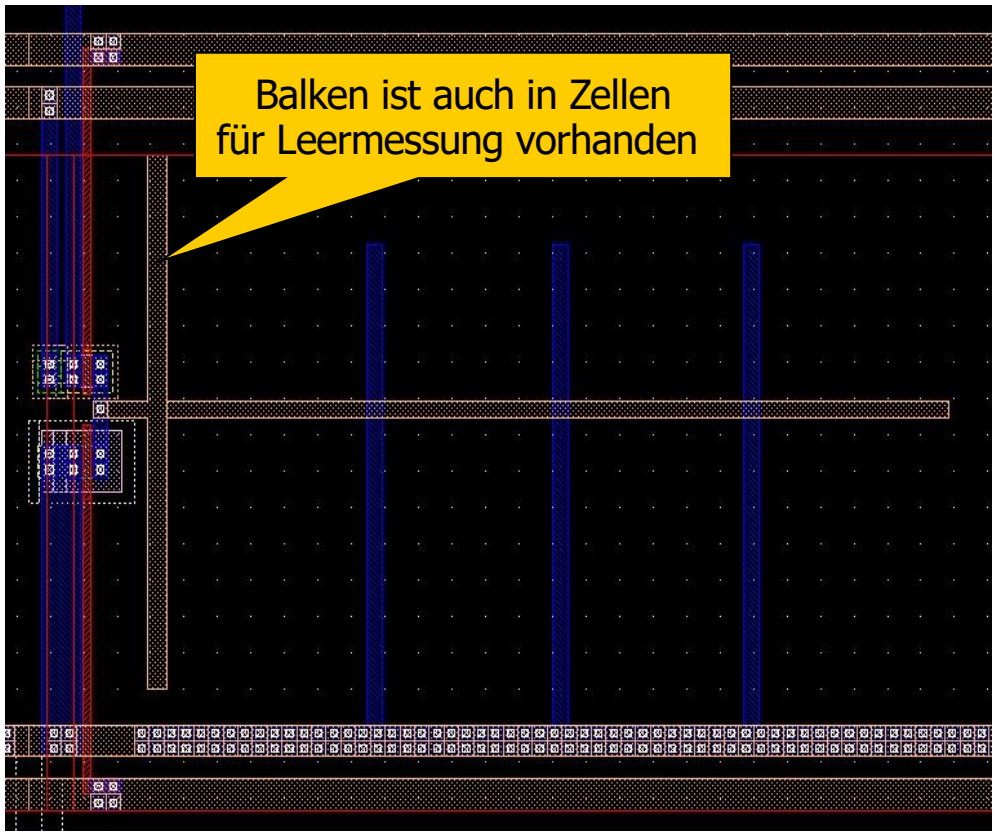
Layoutbeispiele



M1-M2 Kapazitäten



Poly1-Poly2
Kapazität



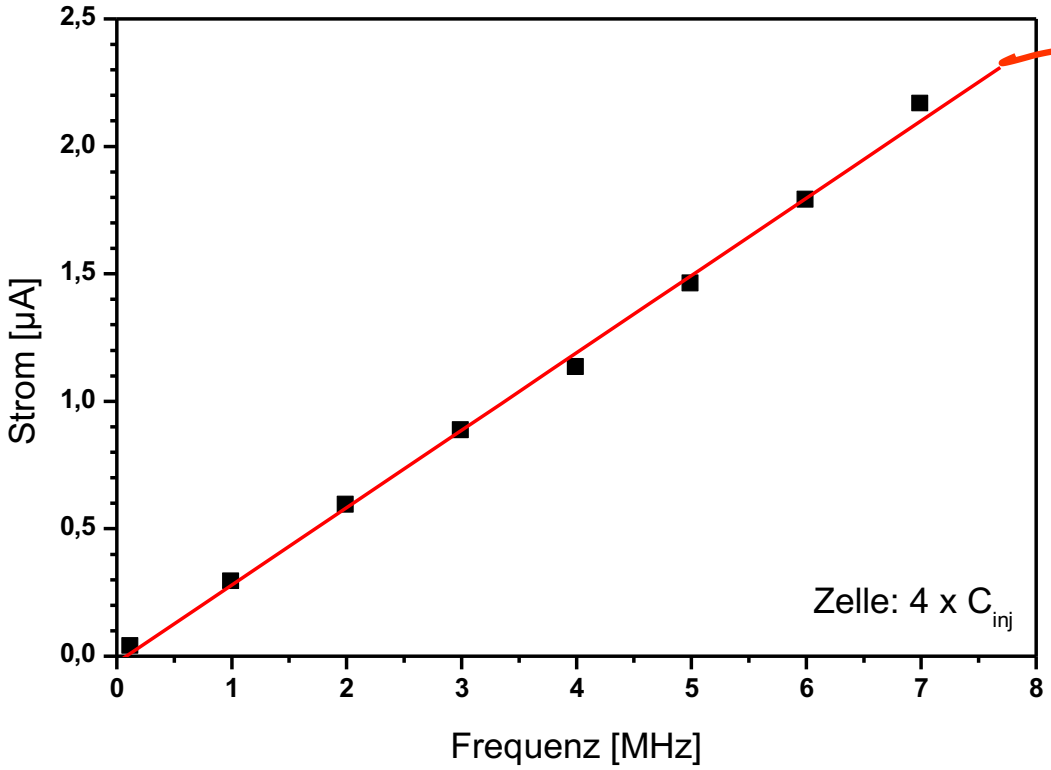
Drei M1 - M2 Kreuzungen
(auch Zellen mit 1 oder 2 Kreuzungen)

Messung: Strom = f(Frequenz)

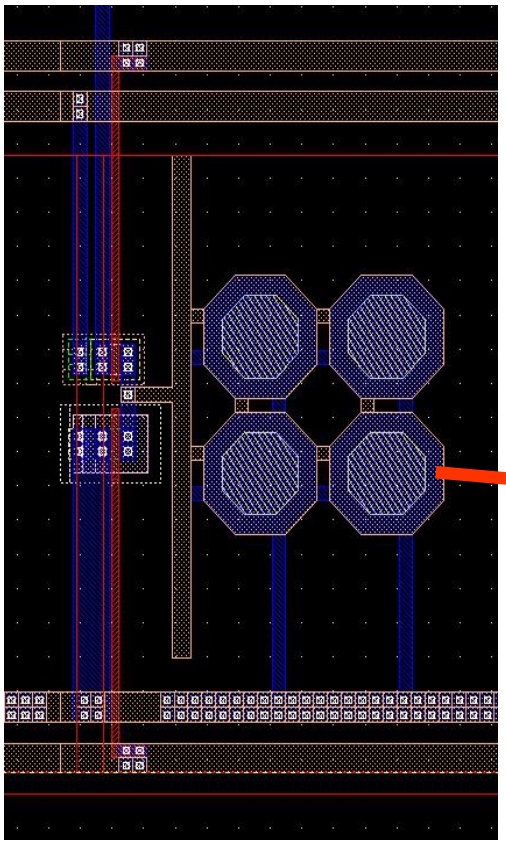
- Messung bei verschiedenen Frequenzen reduziert den Fehler:
- Zur Kontrolle: Messung bei verschiedenen Versorgungsspannungen vdd

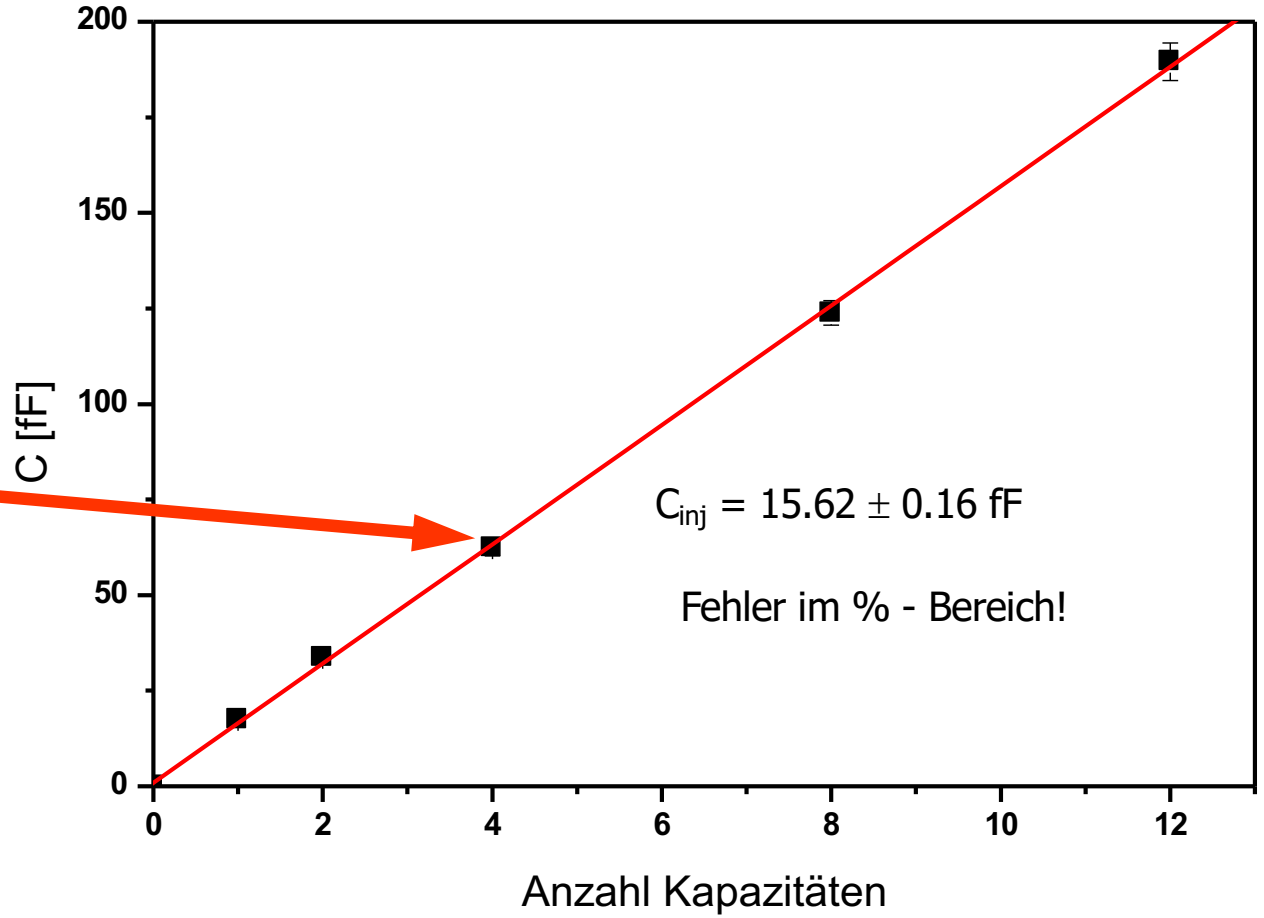
$$I = vdd \cdot C \cdot f \quad \Rightarrow \quad C = \frac{1}{vdd} \cdot \frac{I}{f}$$

- Die Messung aus leeren Zellen wird abgezogen (hier 38.9 $\pm$ 0,7 fF).
- Mit mehreren leeren Zellen kann der statistische Fehler abgeschätzt werden (Streuung der parasitären Kapazitäten)

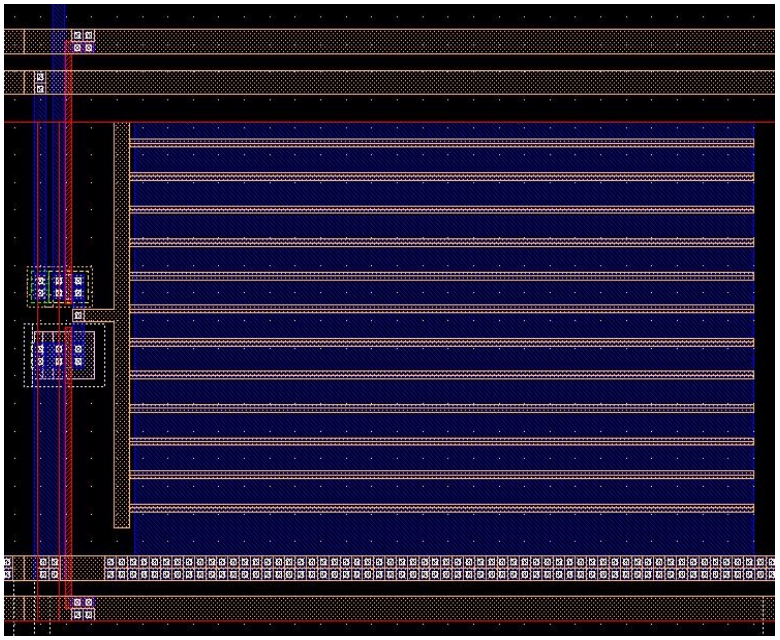


Beispiel für Ergebnisse: Injektionskapazitäten

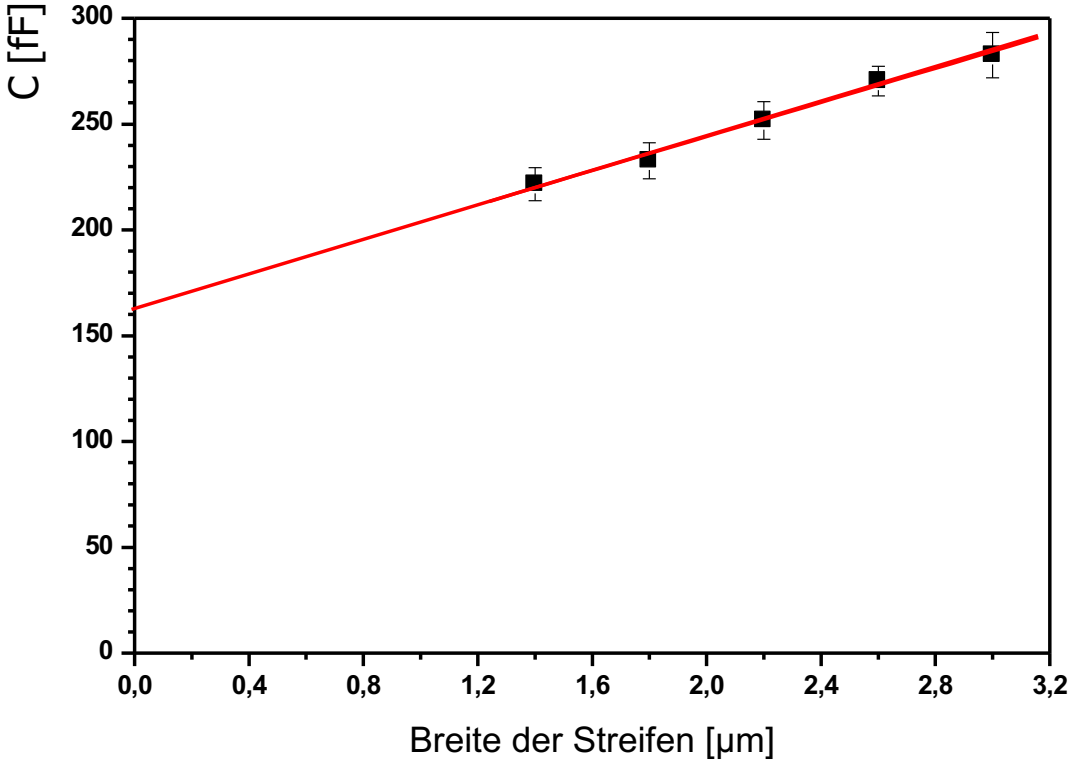




Messung Parallelplatten-C und Streufeld-C



12 Streifen mit unterschiedlicher Breite
(121.6 µm lang, 6.0 µm Raster)

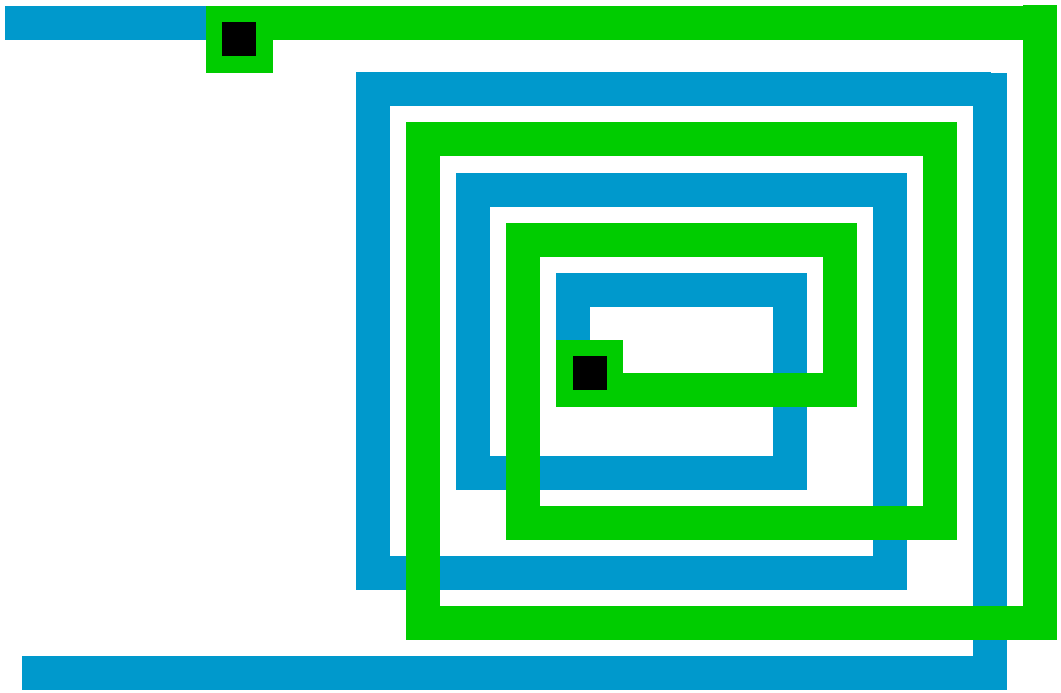


Steigung	⇒	C_{area} Rechnung:	=	**28.3 aF / µm² ± 8%** **28.8 aF / µm²**	(ε=3.9, t_{ox}=1.2µm)
Achsenabschnitt	⇒	C_{fringe} Rechnung:	=	**55.3 aF / µm ± 4%** **50.8 aF / µm**	(ε=3.9, t_{ox}=1.2µm, H=1.05µm)

Spulen

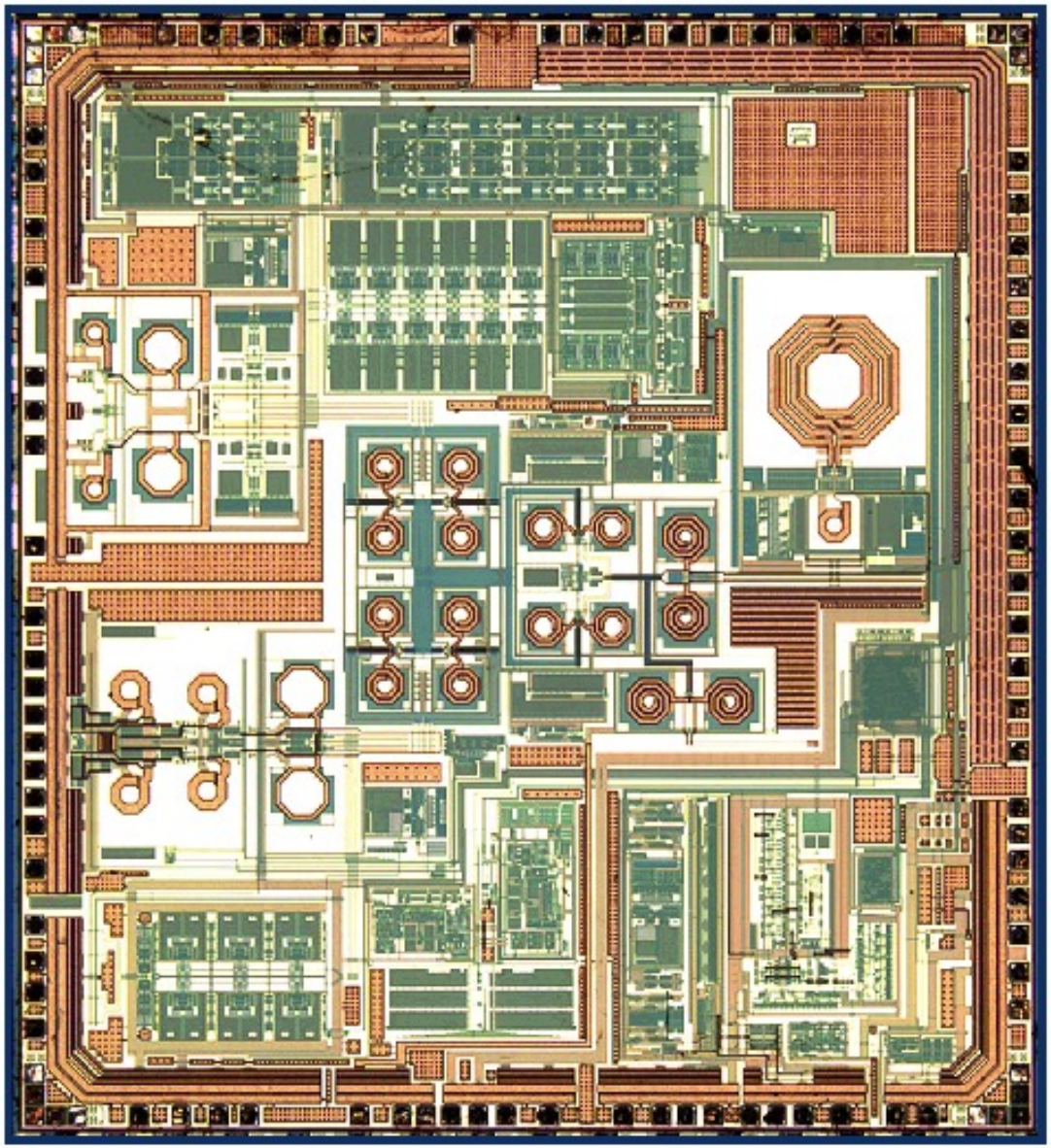
Spulen

- Sind möglich, seit viele Metalllagen verfügbar sind
- Problem: Qualität (Güte der Spule) wird durch parasitäre Kapazitäten verschlechtert
- Oft Wicklungen in mehreren Lagen
- Bevorzugt: Obere Lagen, da diese oft dicker sind und daher der Widerstand kleiner ist. Außerdem sind parasitäre Kapazitäten kleiner.
- Beispiel mit 2 Lagen:



Spulen

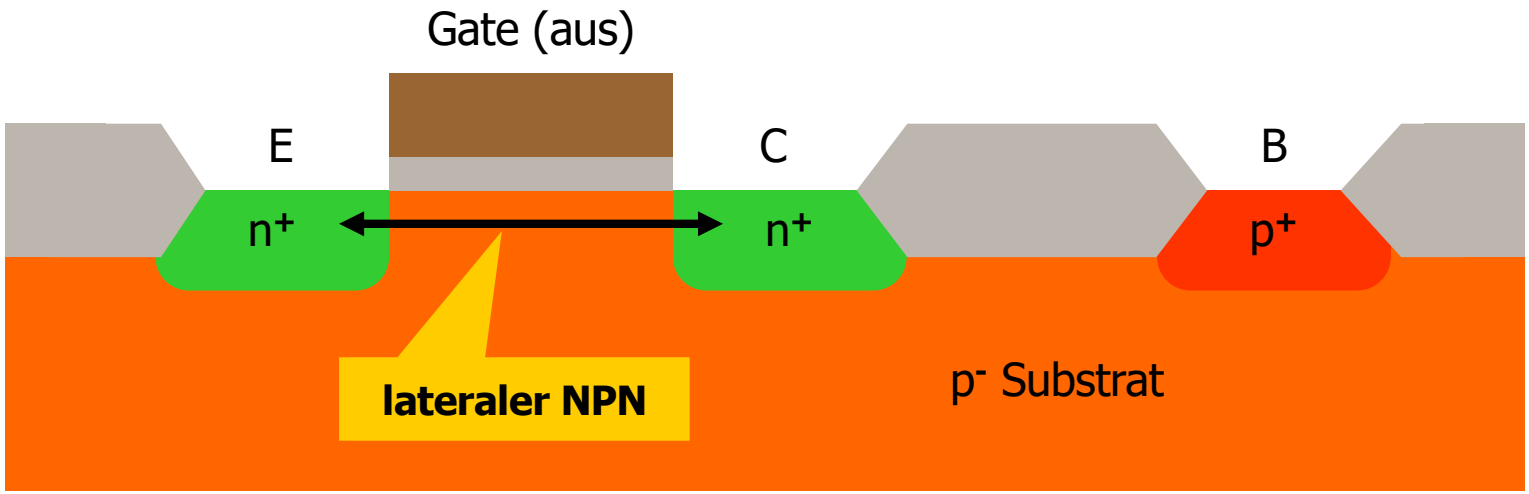
- Werden im HF Bereich (einige GHz) benötigt:
  - Telecom
  - WLAN, (Bluetooth, Zigbee)...
- Baugruppen:
  - LNAs (Low noise Amplifier)
  - VCO (Voltage Controlled Oscillator)
  - PLL (Phase Locked Loop)

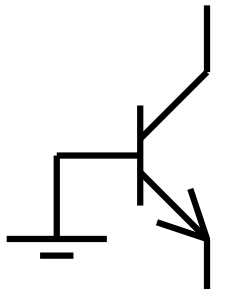


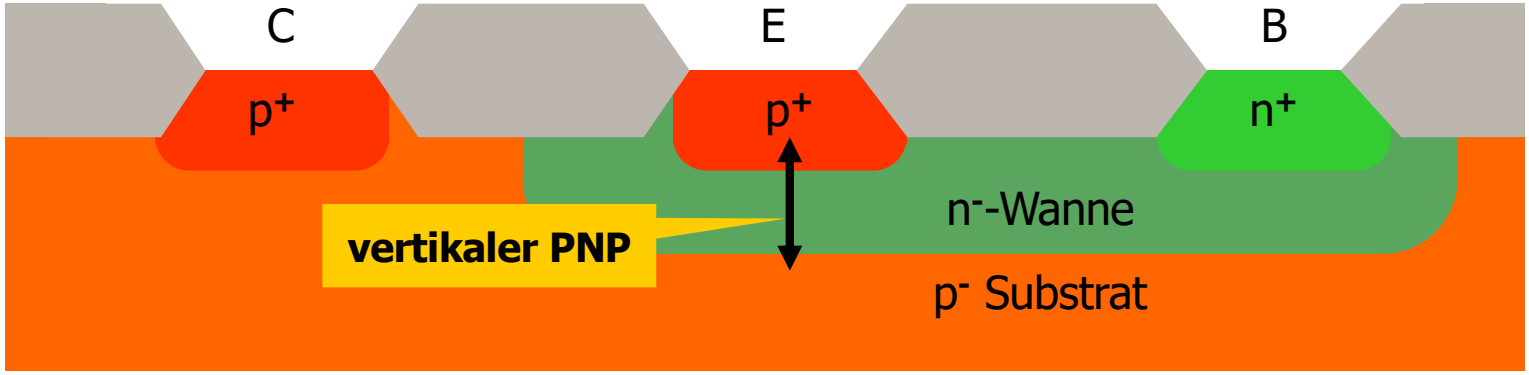
Broadcom BCM2050 WLAN Chip
Quelle Photo: TechInsight

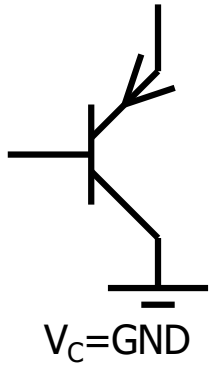
Bipolare Transistoren (in CMOS)

Parasitäre Bipolartransistoren in CMOS

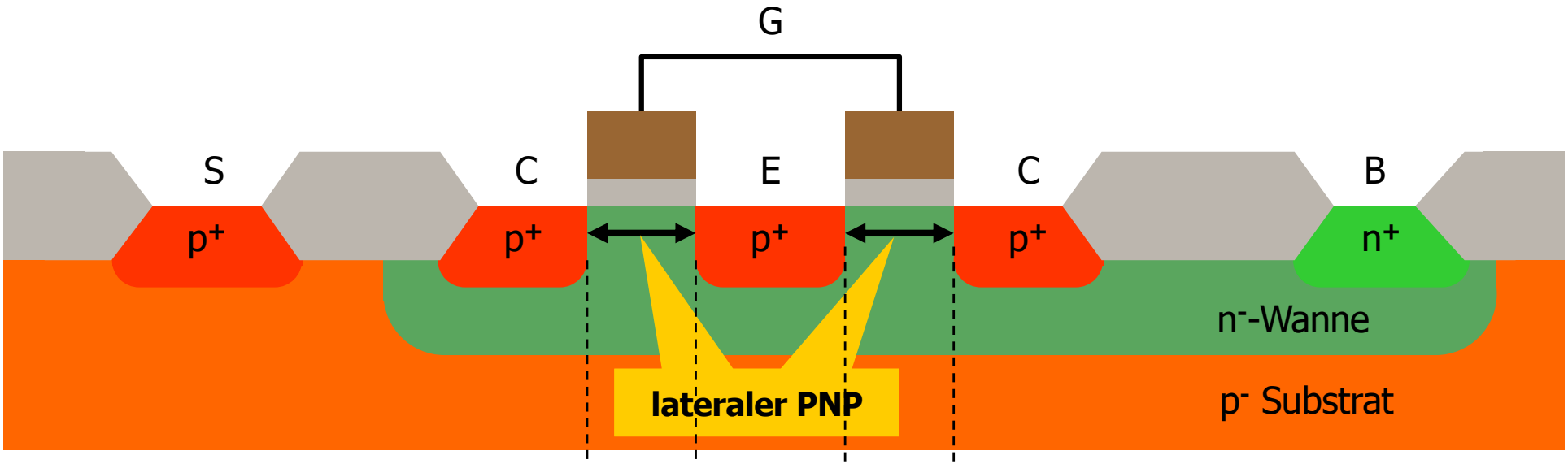




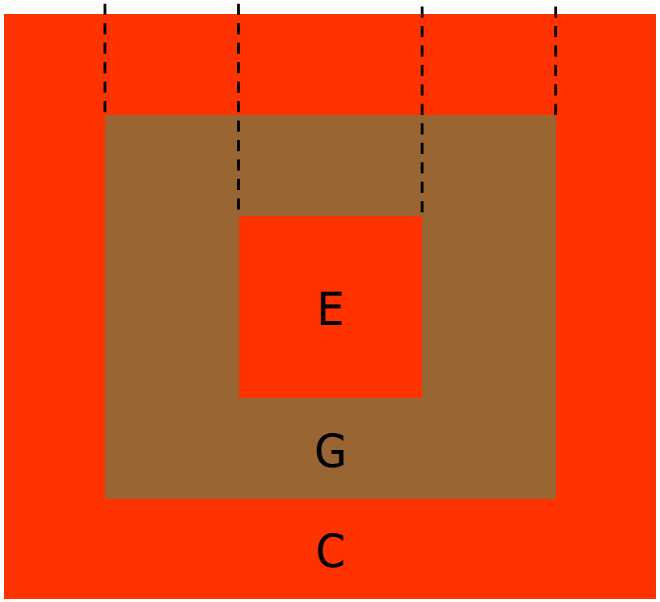


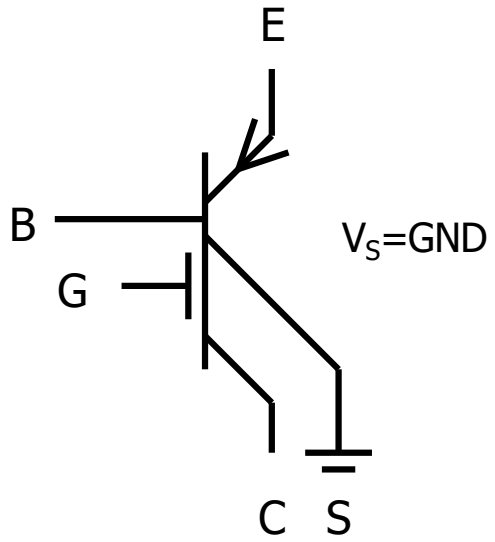


Lateraler pnp Bipolartransistor



- Analog zum lateralen npn, aber in einer N-Wanne
- Dadurch weniger Einschränkungen in den Spannungen (Basispotential kann frei gewählt werden)
- Es ist immer auch ein vertikaler pnp beteiligt
- Schaltsymbol daher kompliziert. (Gate hier explizit eingezeichnet, es wird meist ausgeschaltet)

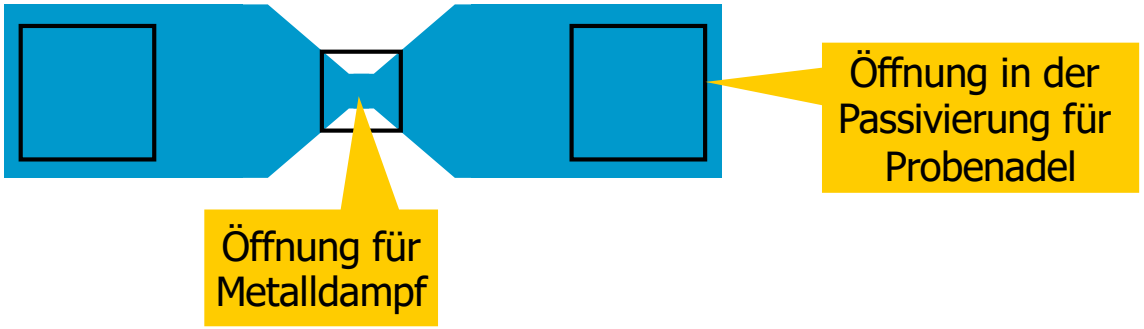




Fuses (Sicherungen)

Fuses ('Sicherungen') zum Trimmen

- Zum einmaligen, festen Programmieren von Chips (wenige Bits) oder zum Feinjustieren (Trimmen) werden manchmal Sicherungen benutzt, die beim Programmieren durchgebrannt werden können
- **Metal-fuses**: ein kurzes Stück Leiterbahn, das mit einem hohen Strom verdampft werden kann
  - Widerstand muß klein genug sein, damit der Strom bei wenigen Volt 'Brenn'-Spannung zur Verdampfung des Metalls ausreicht (benötigt werden einige 100mA für einige Millisekunden)
  - Eine Öffnung in der Passivierung (oberste Lage) läßt das verdampfende Metall entweichen, sonst besteht die Gefahr, daß eine unterbrochene Verbindung mit der Zeit wieder 'zuwächst'



- Manchmal auch **Poly-Fuses**: Weniger geeignet, da Widerstand höher und Schmelzen des Si schwieriger

- **Zener-Zapping**: Zener-Dioden können so überlastet werden, daß sich ein **Kurzschluß** bildet
  - Zerstörung des Kristallgitters!
  - Keine Öffnung in der Passivierung nötig, daher auch keine Kontaminierung des ICs möglich
  - Meist nur in speziellen Bipolartechnologien verfügbar

Matching

Matching

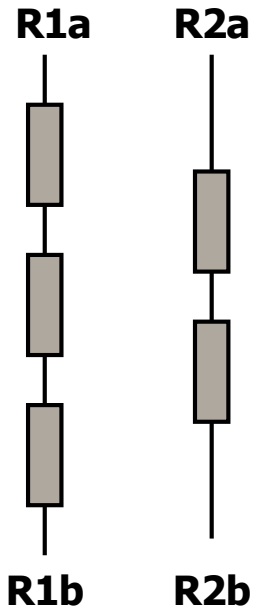
- Zwei identisch gezeichnete Bauteile (z.B. FETs) verhalten sich nicht identisch:
  - zufällige Variationen in der x- und y-Dimension und in LD, WD
  - zufällige Variationen in der Oxiddicke, der Dotierung, ...
  - unterschiedliches Ätzverhalten durch unterschiedliche Umgebung
  - unterschiedliche Temperaturen (Einfluß auf Halbleiterbauelemente, Thermoelektrischer Effekt: ca. 0.4mV/K)

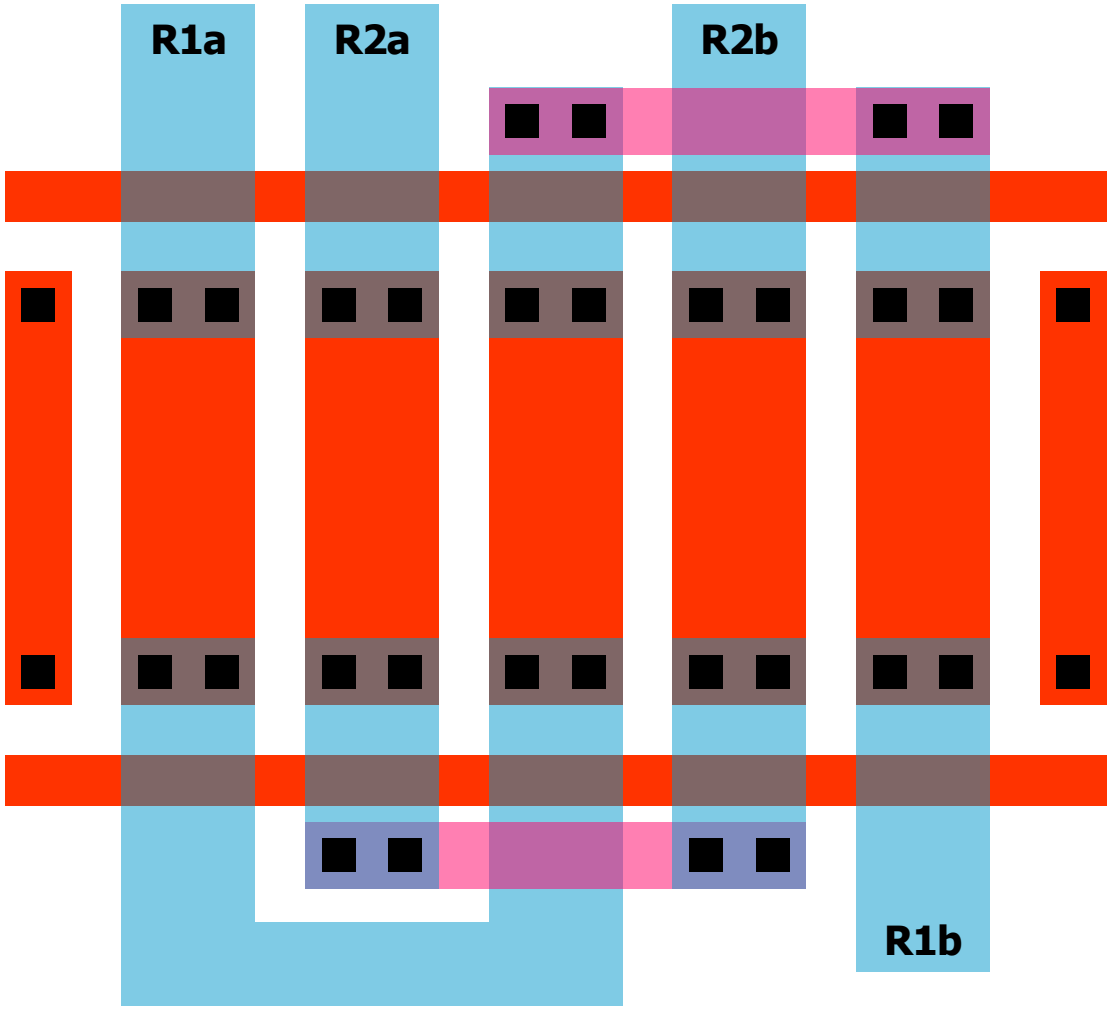
- Sollen sich unterschiedliche Bauteile identisch verhalten (gleicher Widerstand, Kapazität, Strom, Schwellenspannung,...) so muß man:
  - identische Layouts verwenden
  - die Bauteile so groß wie möglich machen (Problem: Kapazitäten, Fläche,..)
  - identische Umgebungen zeichnen ('Dummy Strukturen' = 'etch guards')
  - die Bauteile in der gleichen Orientierung zeichnen (wg. z.B. Piezoresistivität)
  - Bauteile gleichsinnig vom Strom durchlaufen lassen
  - Die Bauteile so nahe wie möglich anordnen, evtl. 'ineinander' legen (s. Beispiel)
  - 'common centroid' Geometrie verwenden, um Gradienten abzufangen (Schwerpunkte zusammengehöriger Bauteile fallen zusammen, s. Beispiel)
  - Temperaturgradienten vermeiden
  - Bauteile nicht an den Rand von Chips legen (Spannungen im Silizium)
  - (Keine Kontakte etc. auf Bauelemente legen, Strukturen in oberen Lagen gleich machen.)

- Sollen Bauteile feste Verhältnisse haben, so sollte man Vielfache von Einheitsbauelementen benutzen
  - Wo das nicht möglich ist, sollte das Verhältnis Rand / Fläche konstant gehalten werden.

Matching: Widerstände

- Beispiel: Poly-Widerstände im Verhältnis 2:3:
  - identische Einzelstrukturen, Dummy Strukturen, Common centroid
  - Stromrichtungen / Thermoelektrischer Effekt sind wegen der ungeraden Zahl Bauelementen nicht ganz kompensiert...
  - Annahme: $R_{M1} \ll R_{Poly}$

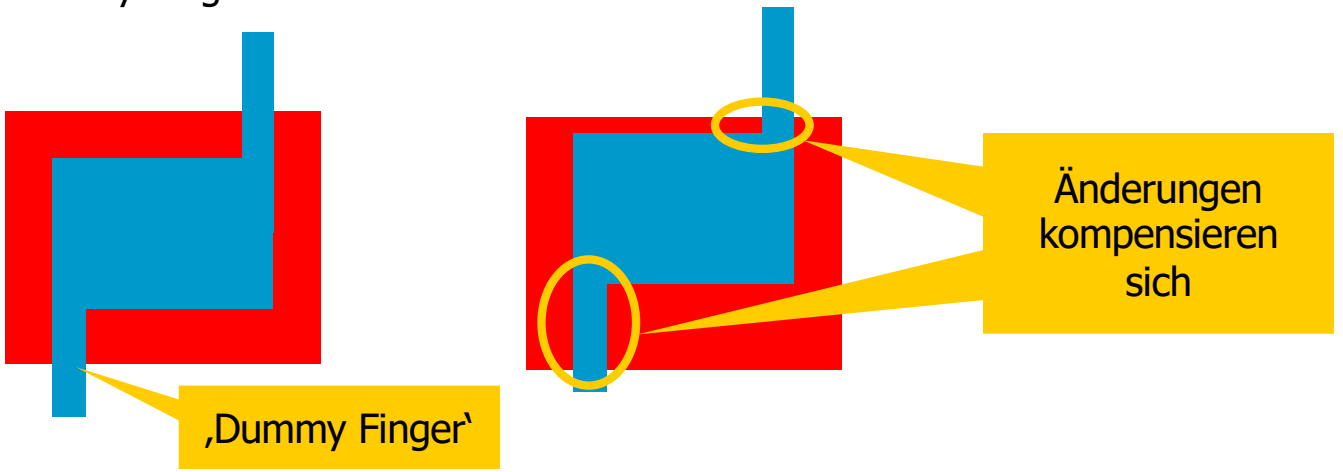




Kondensatoren – Matching 1

- Problem: Durch Verschiebung der 2 Plattenlagen kann es zu Toleranzen kommen.

- Lösung: ‚Dummy-Finger‘

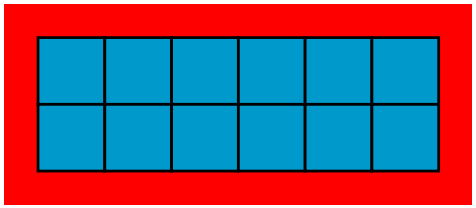


- Zum besseren Matching umgibt man das Layout wieder mit Dummy-Strukturen.

Kondensatoren – Matching 2

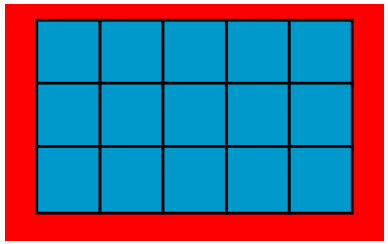
- Verhältnis Fläche/Rand (Area/Periphery) gleich halten:
  - Die Kapazität bei Fläche A und Rand P ist: C = A · Carea + P · Cfringe = **A · (Carea + Cfringe · P / A)**
  - Ein **beliebiges Vielfaches** k · C erhält man daher, wenn A → k · A unter **Beibehaltung** von P/A gesetzt wird
  - Die einfache lineare Skalierung der Struktur um √k funktioniert nicht, da dann A → k · A, aber P → 4 · √k · P wird.

Gegeben:
‚Einheitskapazität'



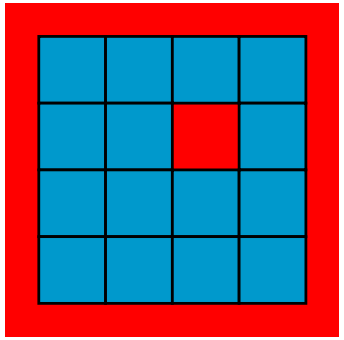
A = 12 Flächeneinheiten (FE)
P/A = 16LE / 12FE = **4/3** LE/FE

Gesucht z.B.:
Layout f. 25% mehr Cap.
i.e. 12 x 5/4 = 15 FE
mit GLEICHEM P/A



A = 15 FE
P/A = 16/15 = **4/5** LE/FE ☹

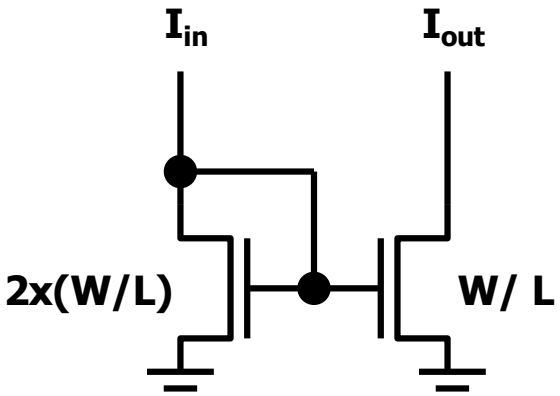
Lösung:
Kapazität mit Loch

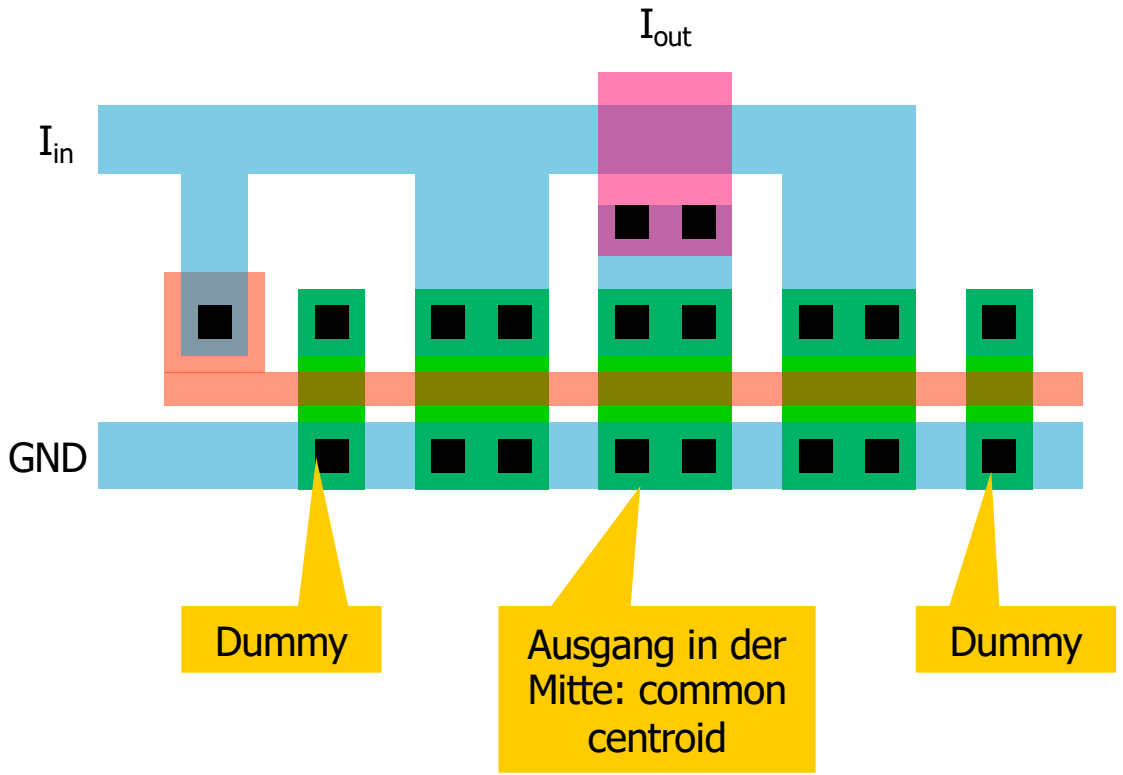


A = 15 Flächeneinheiten
P/A = 20LE / 15FE = **4/3** LE/FE

Matching: Transistoren

- Wenn exakte Vielfache benötigt werden: Transistoren duplizieren, NICHT Länge oder Breite verändern!
- Beispiel: Präziser Stromspiegel 2:1:
  - keine exotischen Formen
  - gleiche Transistorgeometrie
  - gleiche Stromrichtung
  - gleiche Umgebung (Dummy Strukturen)
  - common centroid zur Elimination von Gradienten

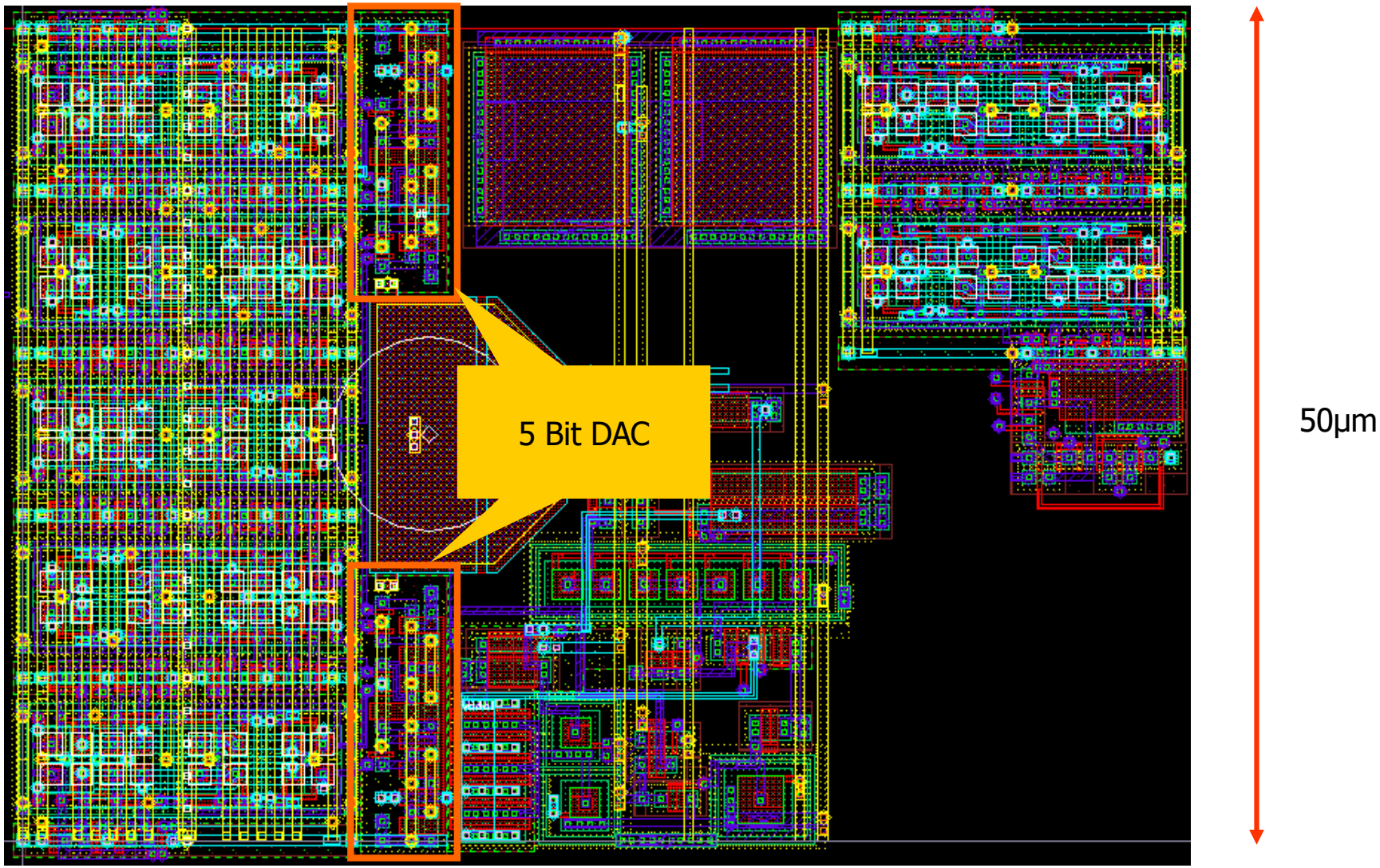




- Präzision ist nicht immer erforderlich !!!

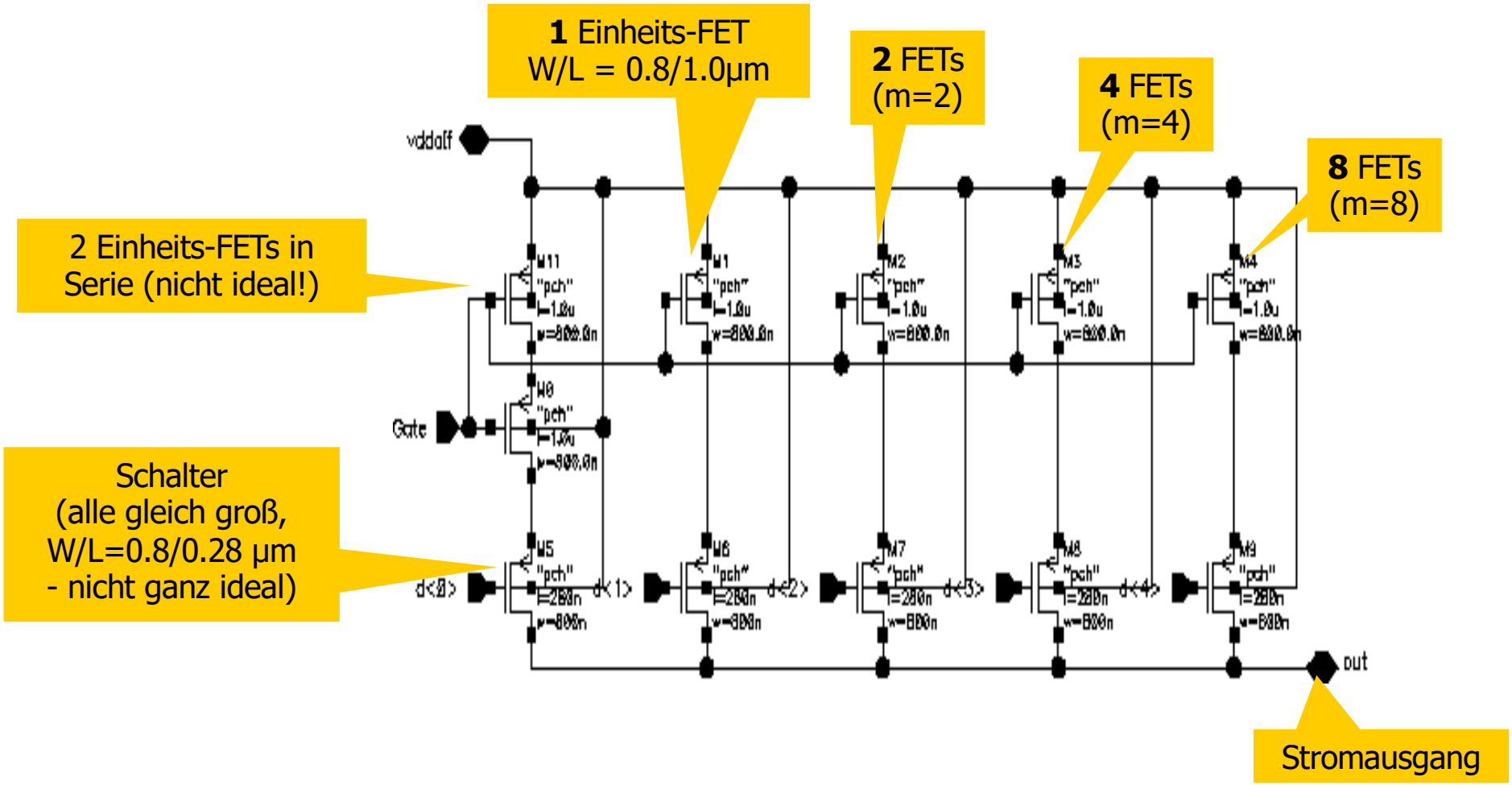
Beispiel aus der Praxis: 5 Bit DAC

- 5 Bit DAC in einem Pixelchip in 0.25µm Technologie. 2 x 2880 Stück auf einem Chip



Schaltung

- Sehr einfach: skalierte Stromquellen werden zu- und abgeschaltet

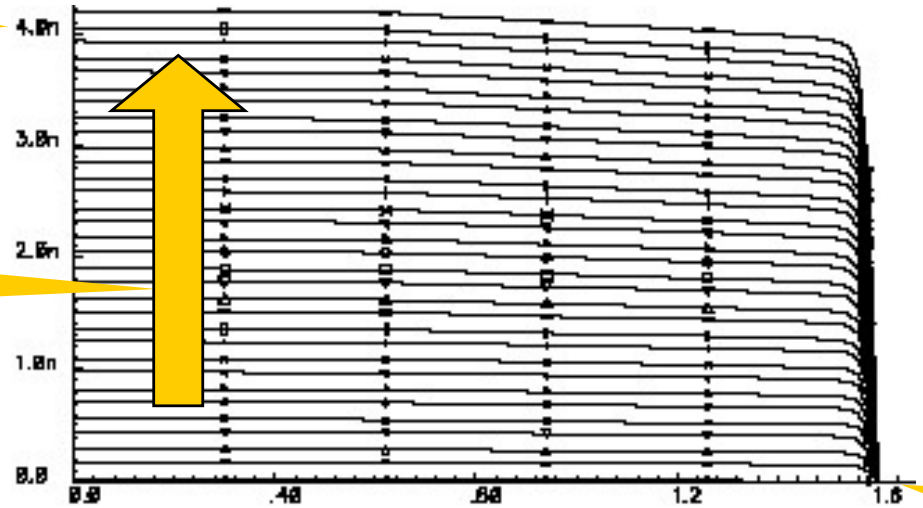


Simulation ist sehr linear

Ausgangskennlinie (VDD=1.6V)

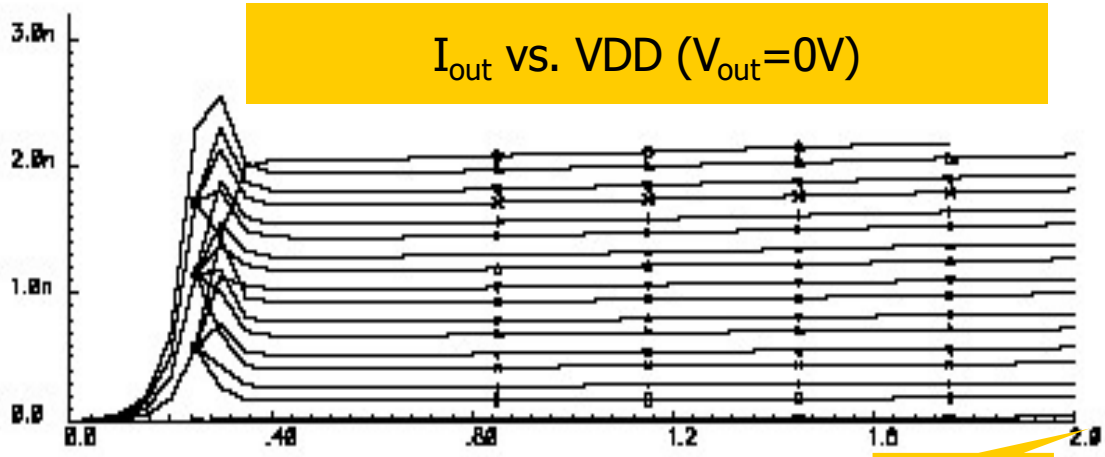
I_{out} (0..4nA)

DAC=0...31



V_{out} (0..1.6V)

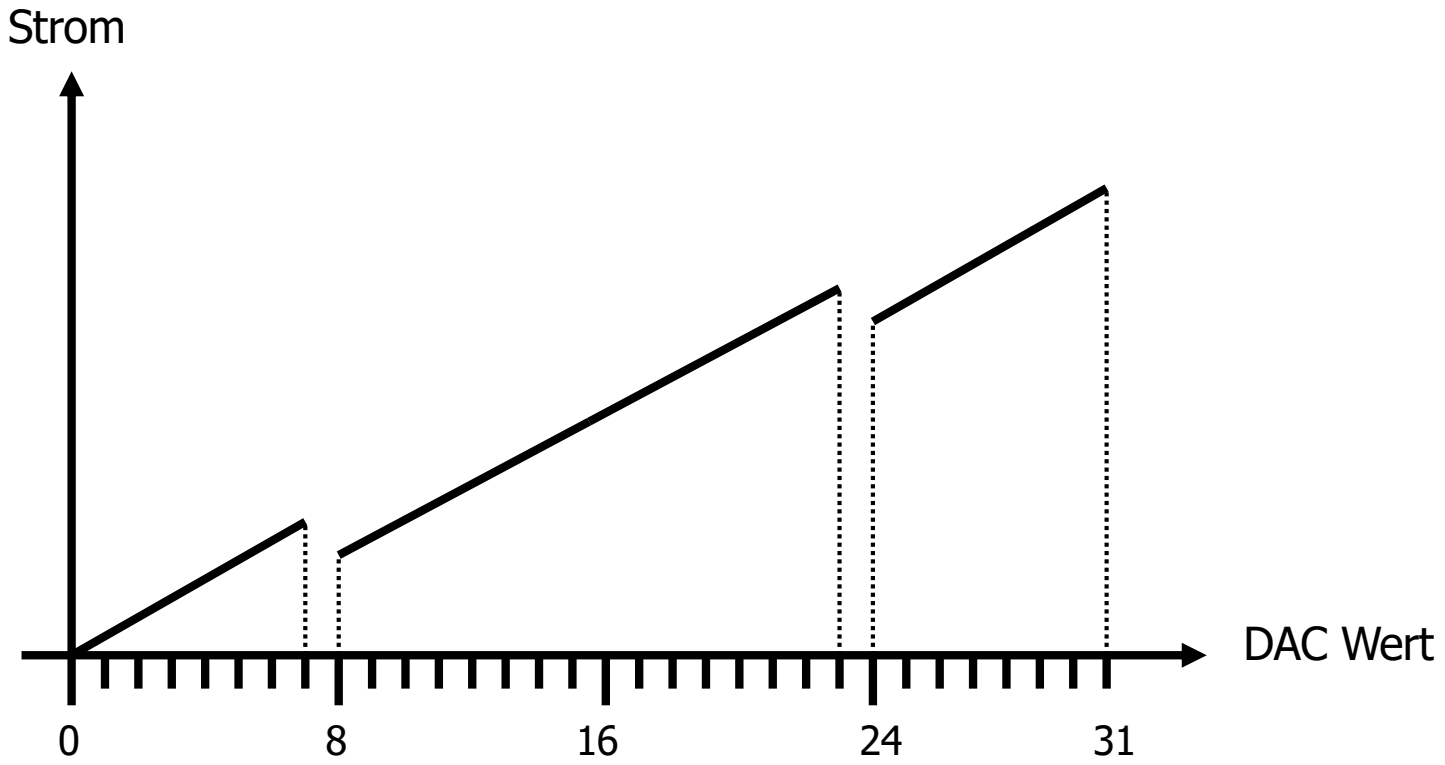
I_{out} vs. VDD (V_{out}=0V)



VDD

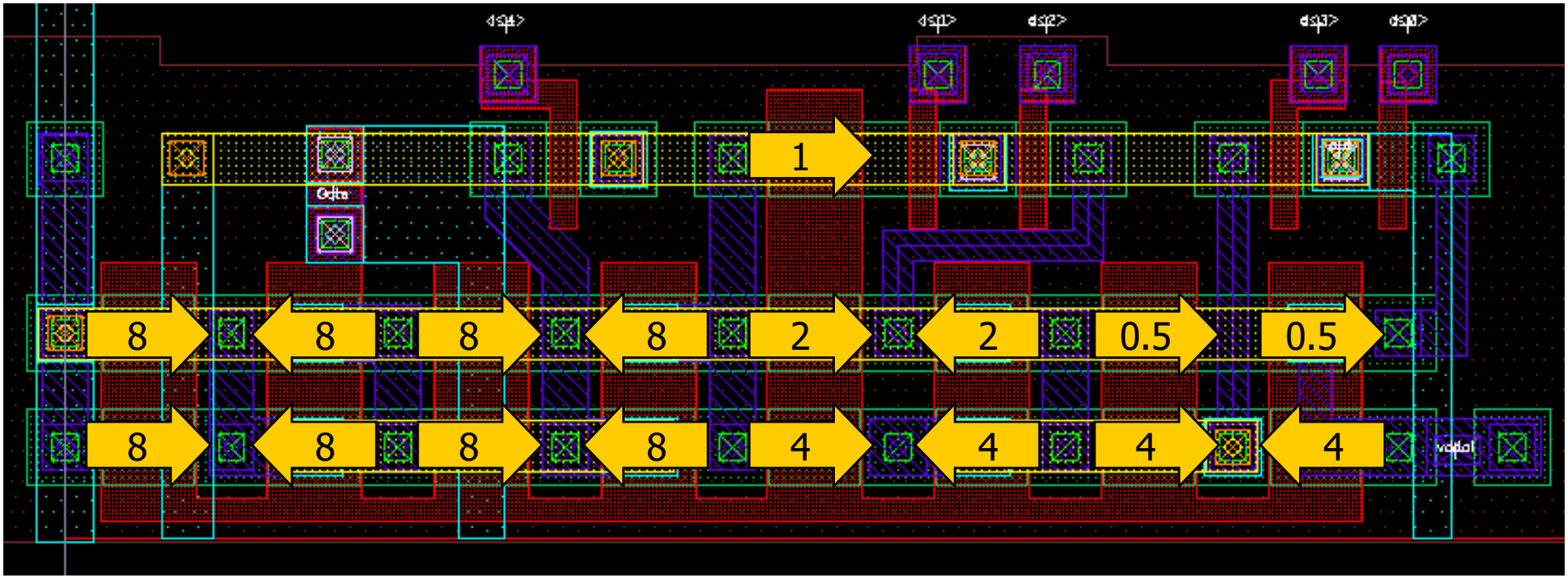
Messung, schematisch

- Große Stufen bei 7⇒8 and 23 ⇒24.
- Kein Problem bei 15 ⇒16 (das ist normalerweise der kritische Punkt!)



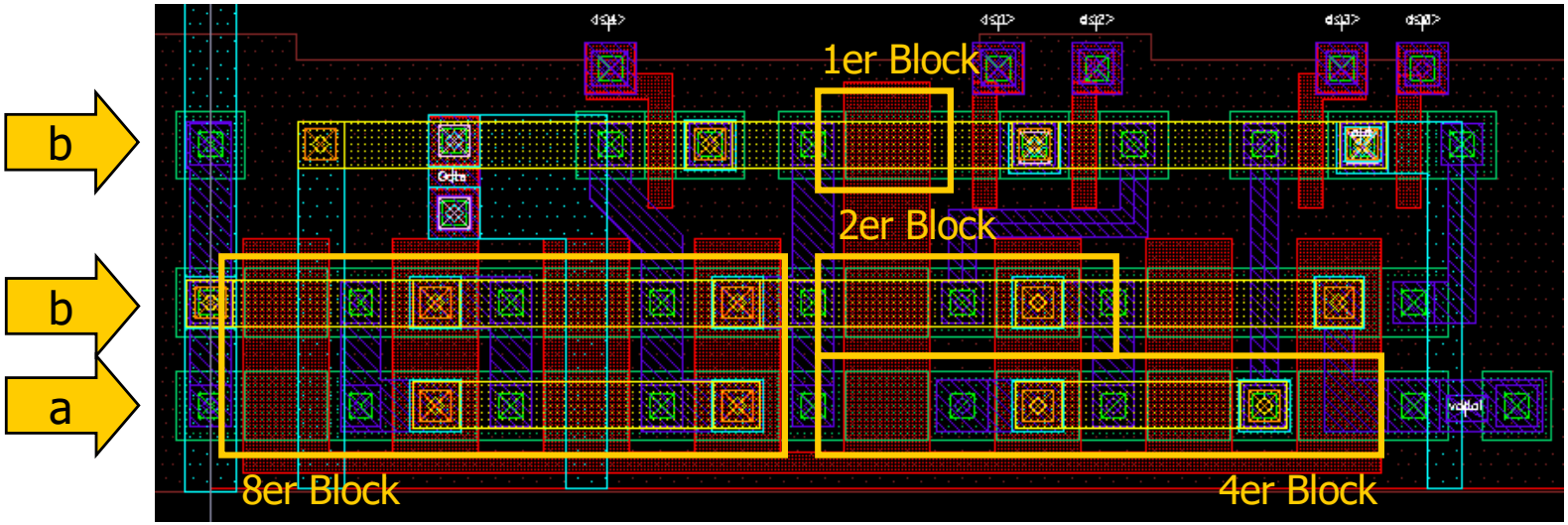
5 bit DAC: Layout

- Im Layout wurde versucht die Stromrichtung in den einzelnen FETs gleichmäßig zu verteilen

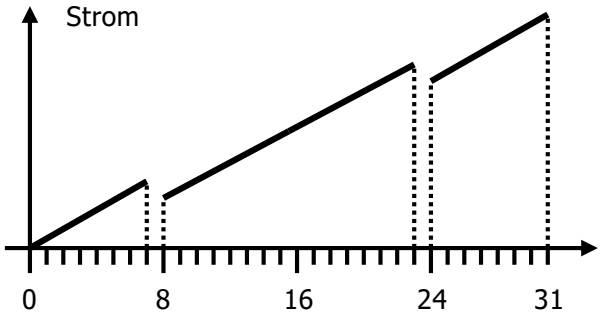


Erklärung

- Erklärungsversuch: Obere und untere Transistorreihe haben unterschiedlichen Strom (Evtl. wegen der Poly1 Leitung unterhalb der unteren Transistorreihe)



- Test: Annahme: FETs in unterer Reihe erzeugen den Strom a, FETs in oberer Reihe den Strom b. (a>b)
  - Schritte < 7: je 0.5b
  - Schritt bei 7 ⇒ 8: 3.5b ⇒ 4a Fehler: 4*(a-b)
  - Schritt bei 15 ⇒16: 3.5b+4a ⇒ 4b+4a kein Fehler
  - Schritt bei 23 ⇒24: 7.5b+4a ⇒ 8b+4a Fehler: 4*(a-b)



Verschiedenes

Spannungsabfälle, Elektromigration

Spannungsabfälle

- Hohe Ströme führen durch den Leitungswiderstand zu Spannungsabfällen ('IR-Drops').
  Hiervon sind insbesondere Versorgungsspannungen (statisch) und Clock-Treiber betroffen (dynamisch)
  - Breite Leitungen benutzen
  - Metalllagen mit niedrigem Widerstand benutzen
  - Mehrere Lagen benutzen
  - Kontakte vervielfältigen

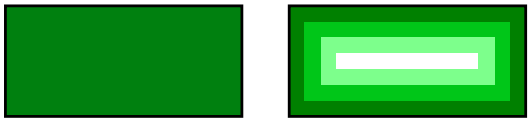
Elektromigration

- bei hohen DC Strömen wandern die Metallatome und erodieren mit der Zeit das Metall
- Durch diese Elektromigration können Leitungen unterbrochen werden oder (durch laterale Ablagerung und die Bildung von nadelförmigen 'Whiskers') Kurzschlüsse entstehen
- Dies kann langfristig zum Ausfall des Chips führen.
- Die MTF (Mean Time to Failure) ist proportional zu $1/I^2$ und hängt exponentiell von der Temperatur ab. Durch Test bei hoher Temperatur kann man daher die MTF ermitteln ('accelerated aging').
- Die Stromdichte muß überall unter einem (hoffentlich vom Hersteller vorgegebenen) Limit bleiben
  Literaturwert: max. 5×10^5 A/cm^2 für Cu-dotiertes Aluminium bei 85^{o}C (0.5-4% Cu erhöht Al-Haltbarkeit)
  Faustregel: max. 1.5mA pro µm Bahnbreite
- Leiterbahnen, die über Oxid-Stufen laufen, sind anfälliger (da dort meist etwas dünner)

Skineffekt

- Bei niedrigen Signalfrequenzen fließt der Strom im gesamten Volumen eines Leiters
- Bei **sehr hohen Frequenzen** (GHz) werden die Ladungsträger durch das von ihnen selbst erzeugte Magnetfeld aus dem Inneren des Leiters verdrängt.
- Stromfluss findet nur noch in einer dünnen ‚Haut' - Schicht (skin) an der Oberfläche statt.

- Die Anzahl Ladungsträger nimmt exponentiell mit der Eindringtiefe ab.
- In der **Skin Tiefe** δ (‚skin depth') ist die Stromdichte auf 1/e abgefallen.



niedrige Frequenz hohe Frequenz

$$\delta = \mathbf{sqrt\ (1\ /\ \pi\ f\ \mu\ \sigma)} = \mathbf{sqrt\ (\rho\ /\ \pi\ f\ \mu)}$$

mit

f = Signalfrequenz,
μ = (magnetische) Permeabilität = $4\pi\ 10^{-7}$ H/m,
σ = Leitfähigkeit,
$\rho = 1/\sigma$ = spez. Widerstand

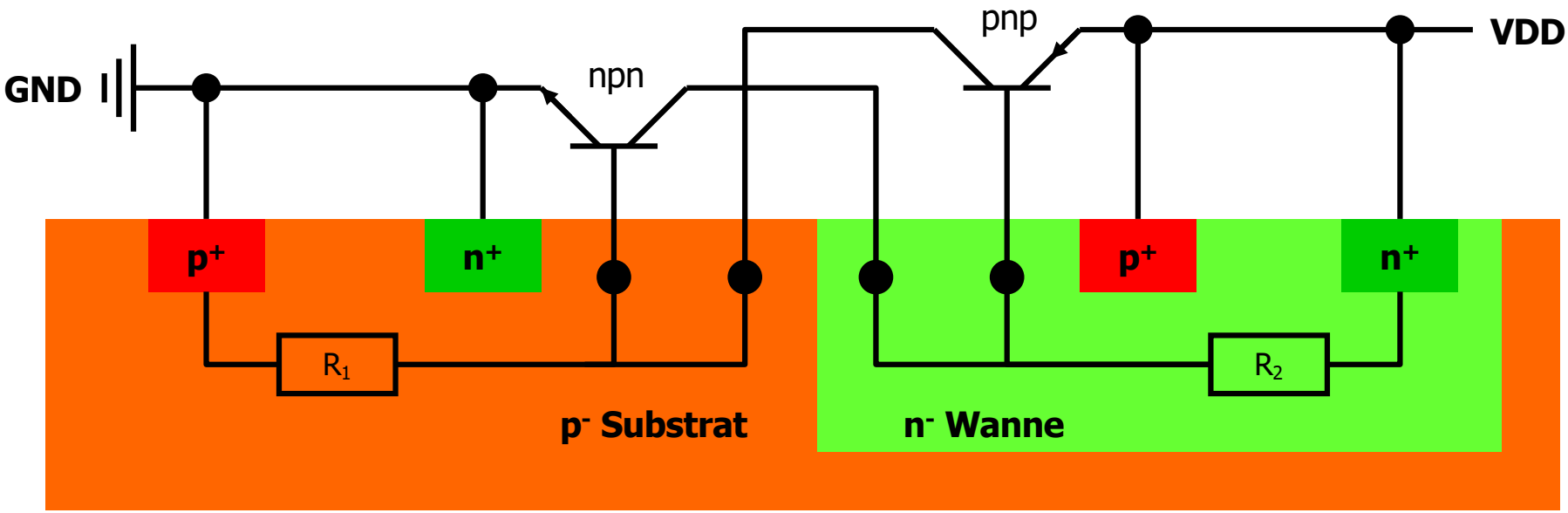
- Beispiele:
  - Für Aluminium (ρ=2.65μΩcm) bei 1 GHz ist δ**=2.6 μm**
  - Für Kupfer (ρ=1.67μΩcm) bei 10 GHz ist δ**=0.7 μm**

- Ergebnis: Bei Leiterbahndicken von μm spielt der Skineffekt in Chips (noch) **keine große Rolle**

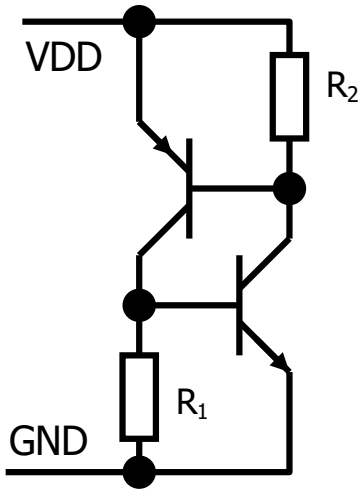
- Herleitung z.B.: http://scienceworld.wolfram.com/physics/SkinDepth.html

Latchup

- Durch die verschieden implantierten Bereiche gibt es parasitäre Strukturen aus npn und pnp Transistoren

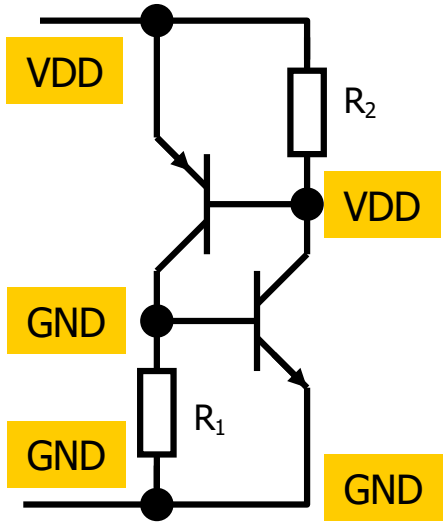


Äquivalentes Schaltbild:

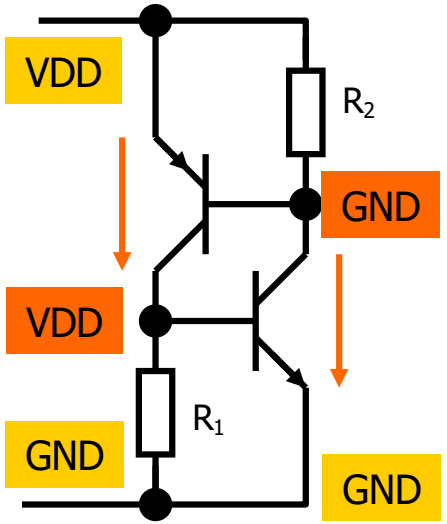


Latchup

- Die äquivalente Schaltung kann 'gezündet' werden, wenn eine BE-Diode in Vorwärtsrichtung leitet
- Dies kann durch kurze Spannungsspitzen passieren
- Die npnp-Struktur schaltet dann ein (Thyristor, SCR=Silicon Controlled Rectfier) und BLEIBT angeschaltet
- Der hohe Querstrom zwischen VDD und GND kann die Schaltung zerstören



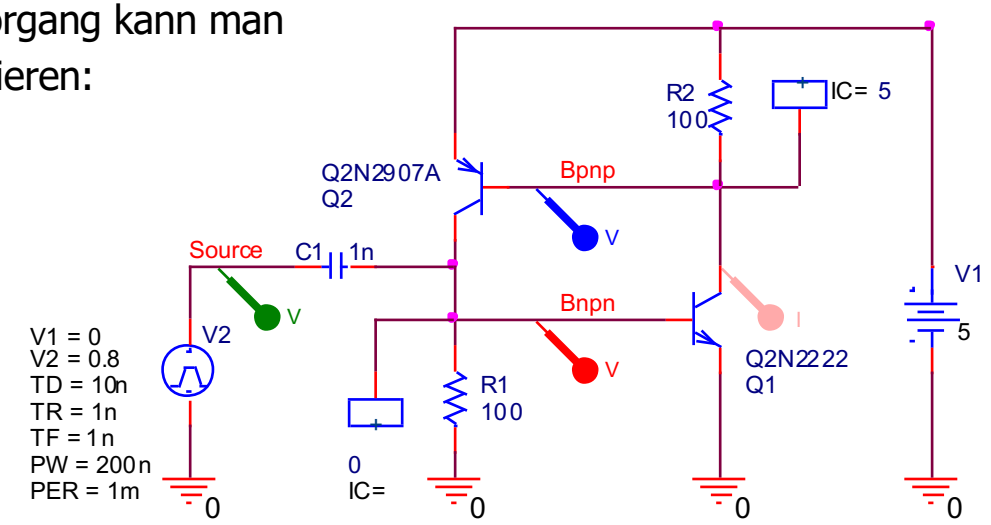
Normalzustand:
Beide Transistoren sind OFF

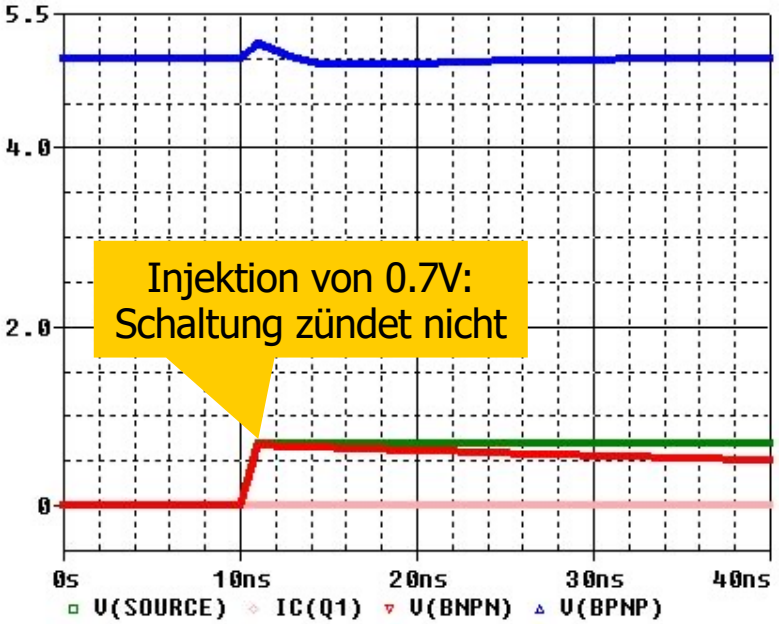


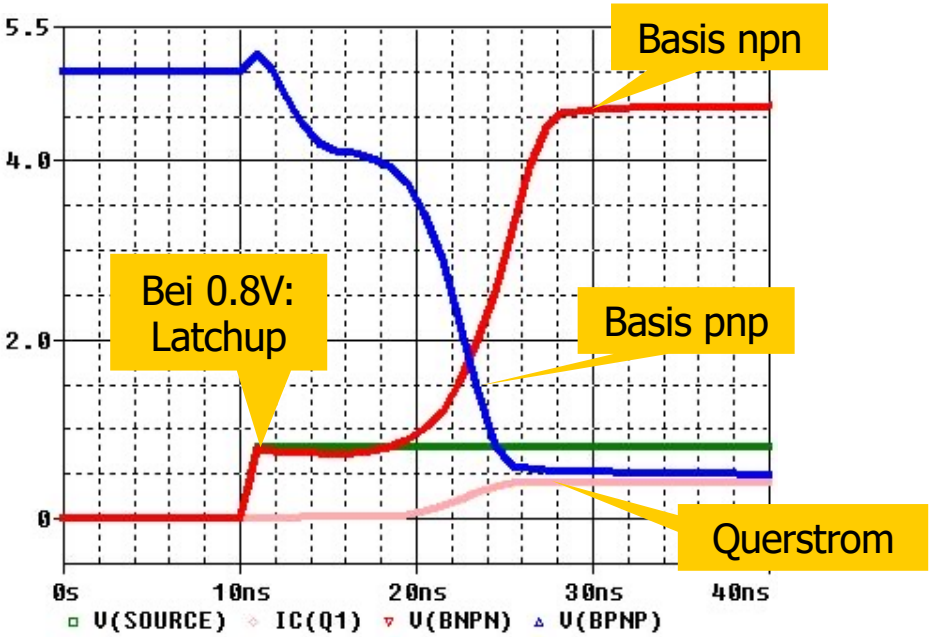
Gezündeter Zustand (vereinfacht):
Beide Transistoren sind ON

Latchup Simulation mit SPICE

- Den beschriebenen Vorgang kann man leicht mit SPICE illustrieren:







Latchup Vermeidung

- Die Schleifenverstärkung muß klein sein. Dies kann durch geeignete Technologieschritte erreicht werden

- Die Dioden dürfen nicht in Leitung kommen:
  - Widerstände müssen klein sein
  - **Substrat/Wannenkontakte nahe bei den Transistoren**
  - Niederohmige Anbindung der Kontakte an Versorgungen und Transistoren (Metall!)
  - **Jede Wanne muß Kontakte bekommen!**
  - Abstand der Transistoren vergrößern
  - Guard-Ringe benutzen

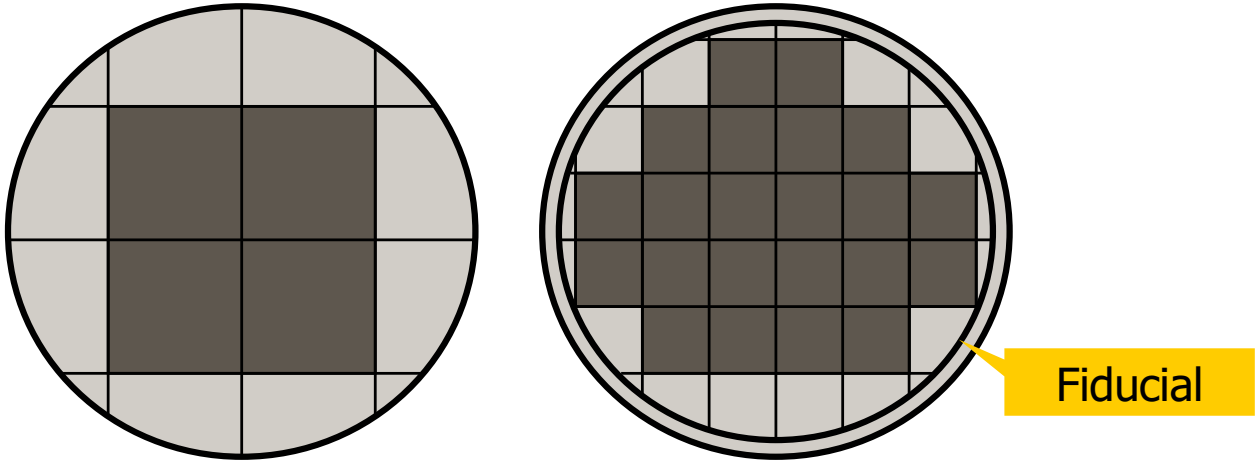
- Besondere Vorsicht ist bei Schaltung mit hohen Transientenströmen (Buffer, IO Pads) geboten

Ausbeute (Yield)

Ausbeute (Yield)

$$\text{Yield} = \frac{\text{Anzahl } \textbf{gute} \text{ Chips ('}\textbf{die}\text{') auf dem Wafer}}{\text{Anzahl alle Chips auf dem Wafer}}$$

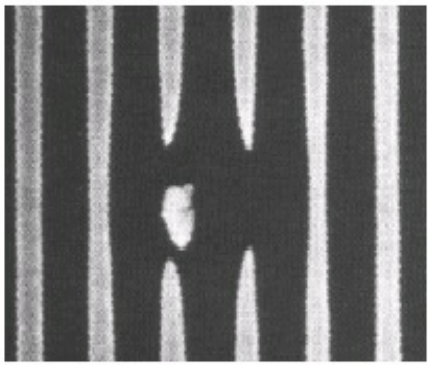
- Ein gewisser am Rand ist unbrauchbar (Prozessierung dort schlecht, Verletzungen durch Anfassen des Wafers,…). Es werden daher nur die Chips innerhalb des brauchbaren Bereichs ('fiducial area,) gezählt.

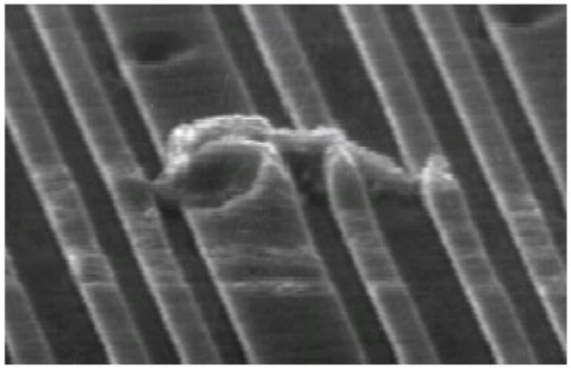


- Partikel größer als ~ min. Strukturbreite/3 können (müssen aber nicht, p=0.2-0.4) zu Totalausfall führen
- Beispiel 0.5µm Prozeß, 1cm^2 Chip: Fußballfeld – Stecknadelkopf
- Nur 4-5 Ebenen sind kritisch (Poly, Implantationen, unteres Metall), dann werden die Strukturen gröber.
- Eine Gesamtausbeute (Wafer $\Rightarrow$ fertig verpackte Bauteile) von 70% ist gut.

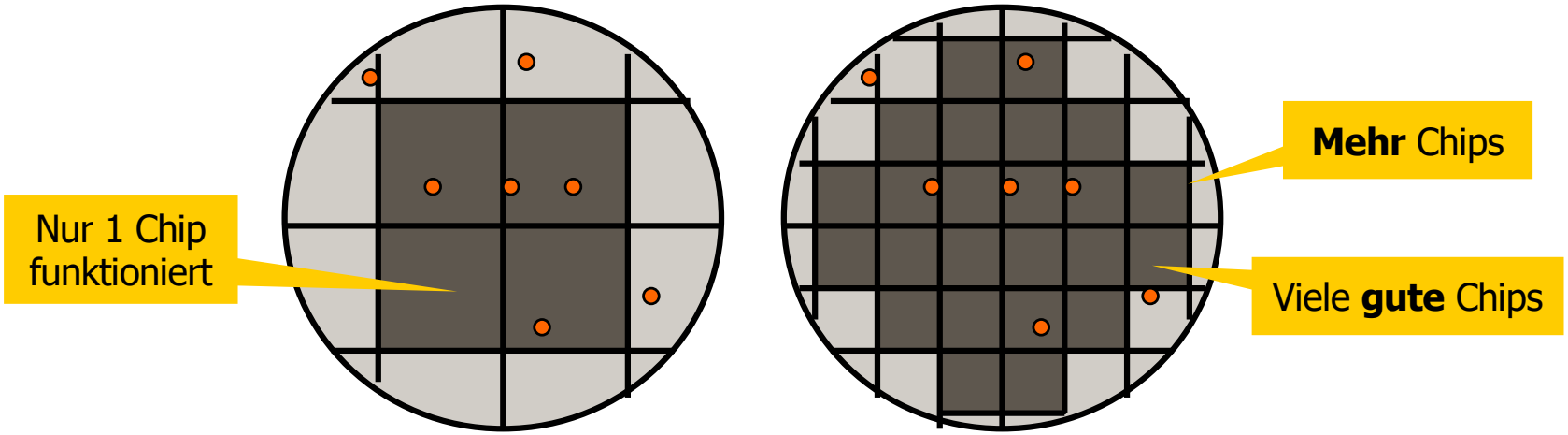
Defekte

- Einfachste Defekte: open / short





- Die Wahrscheinlichkeit für den Ausfall eines Chips ist um so höher, je größer der Chip ist.
- Gleichzeitig gibt es bei großen Chips sehr viel weniger Chips auf dem Wafer



⇒ **Die Ausbeute nimmt mit der Größe der Chips sehr stark ab!**