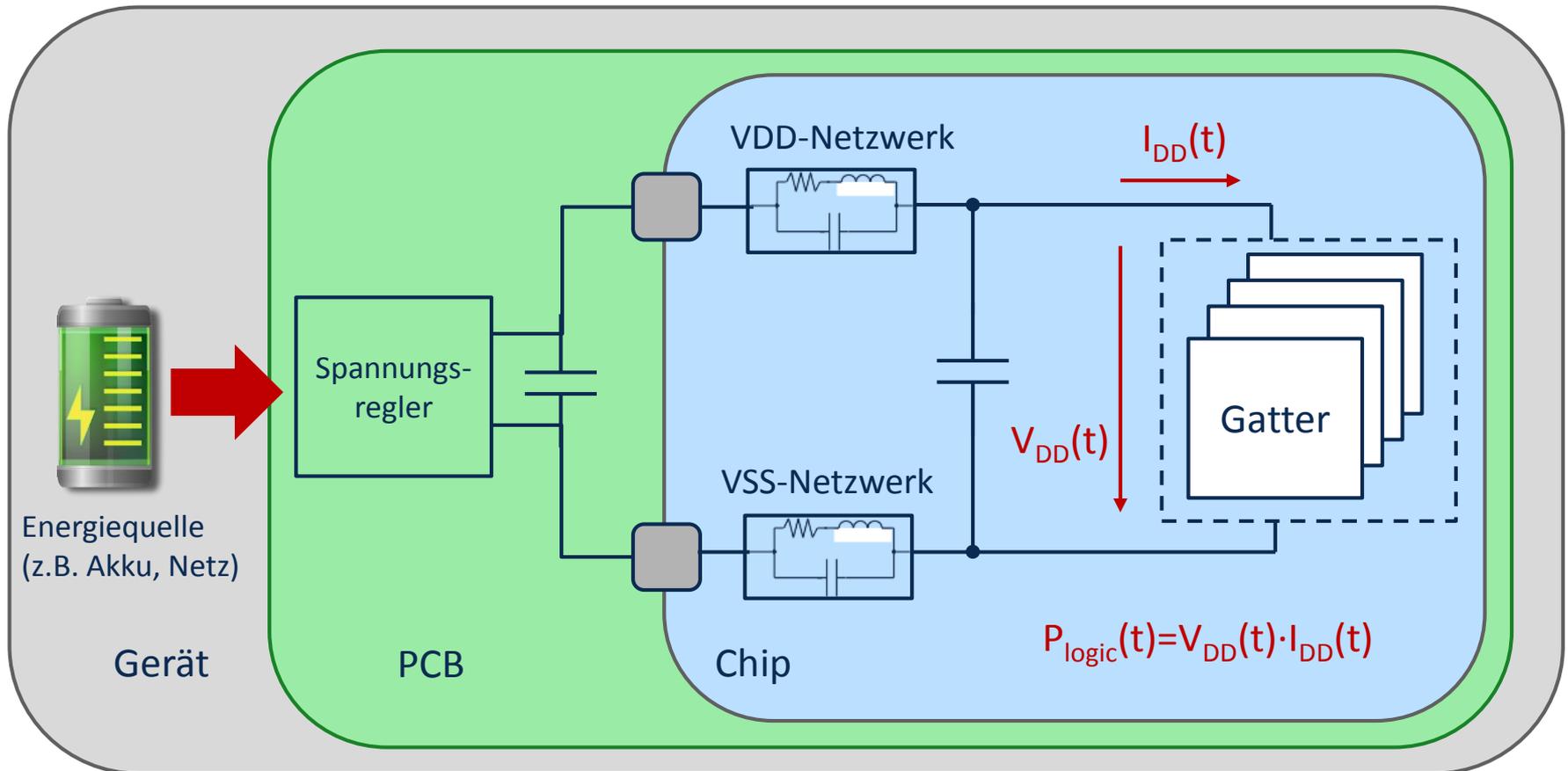
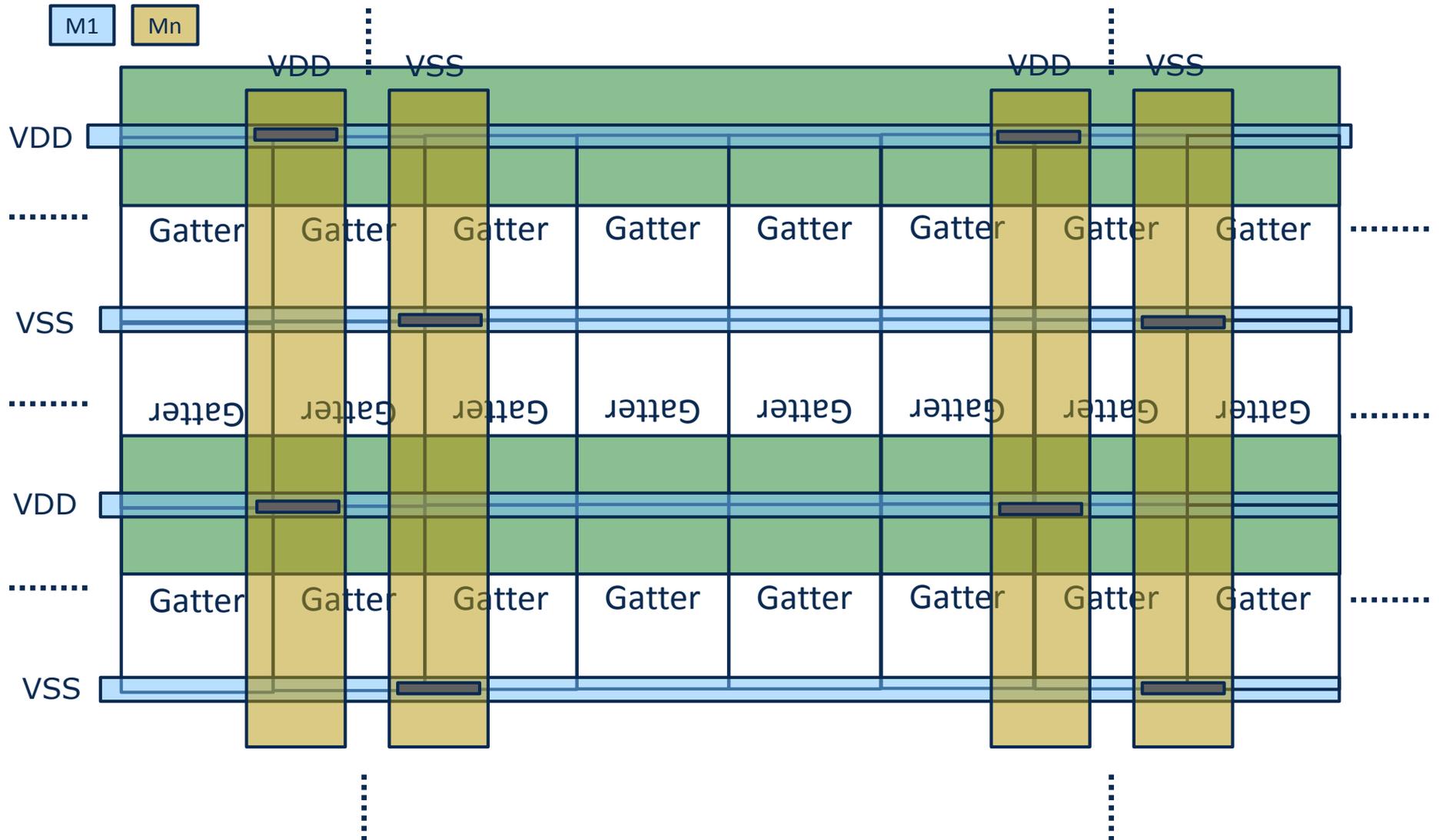
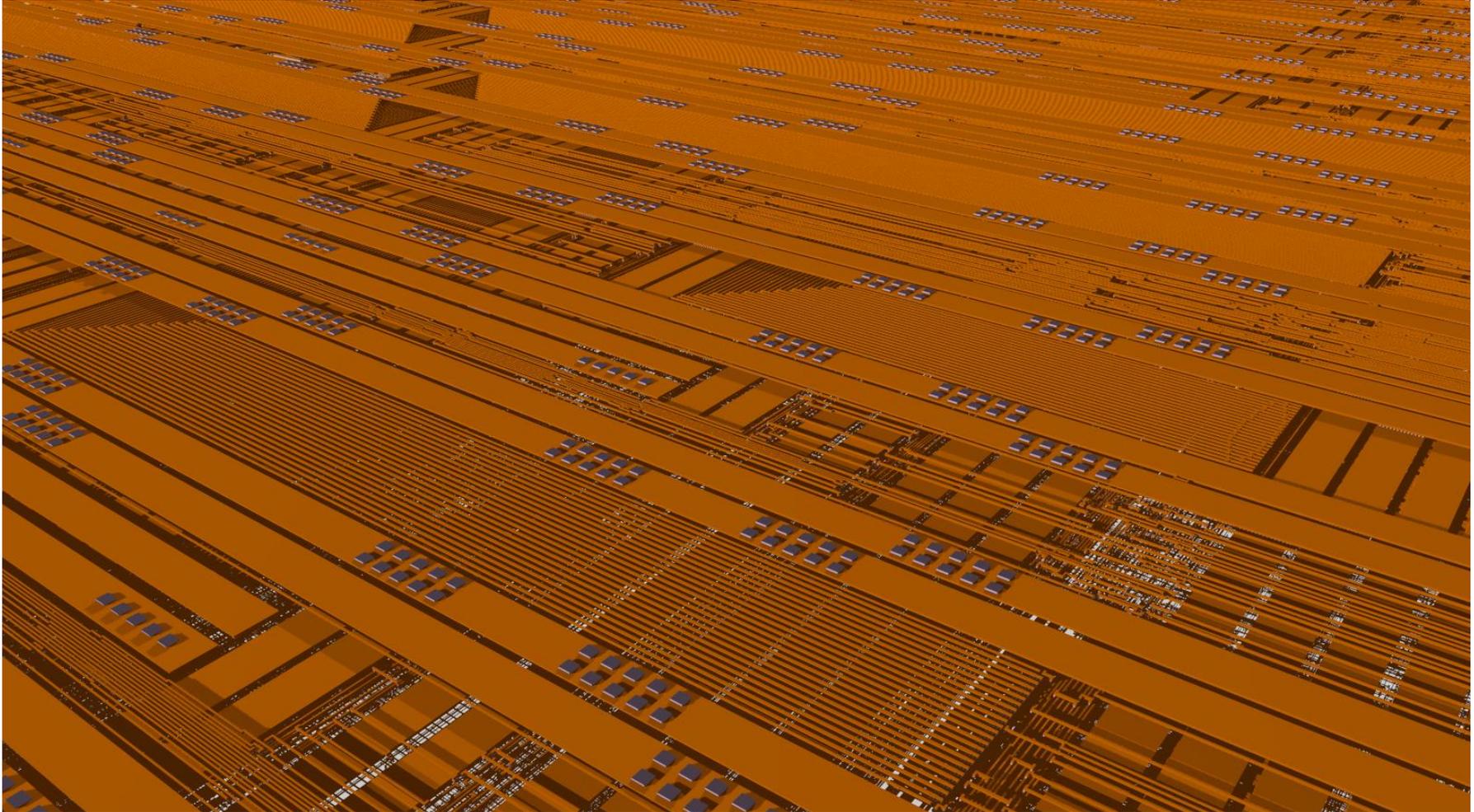
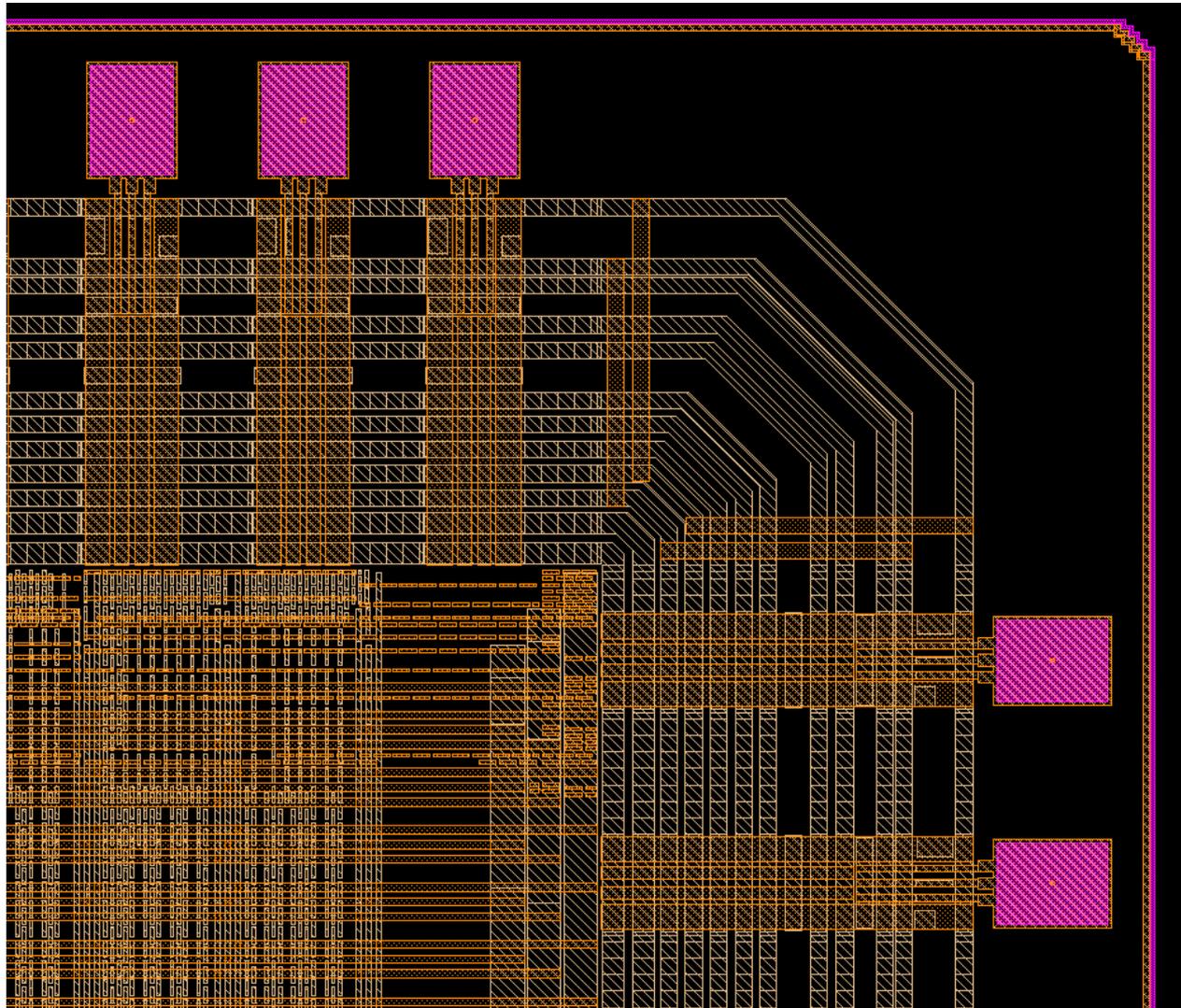


# Verlustleistung von digitalen CMOS Schaltungen

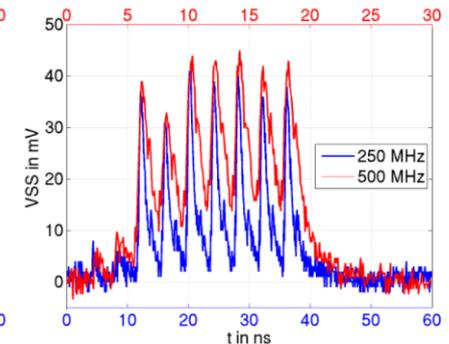
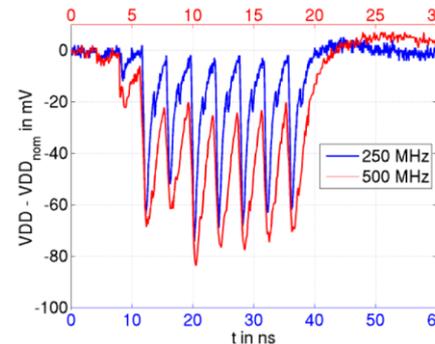
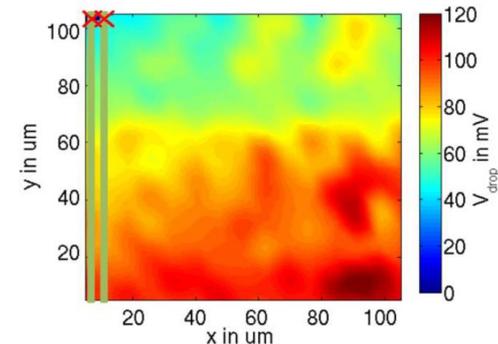
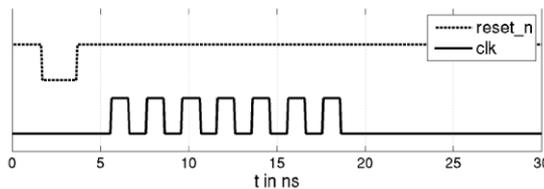








- IR-Drop beschreibt den Einbruch der Versorgungsspannung durch dynamische Stromaufnahme
  - Abhängig von Widerständen  $R$ , Induktivitäten  $L$  und Kapazitäten  $C$  im Netzwerk
  - IR-Drop ist ein **zeitliches** und **räumliches** Phänomen
- IR-Drop beeinflusst **Timing und Funktionalität** der CMOS Logik



[1] Dietel, S.; Hoppner, S.; Brauningner, T.; Fiedler, U.; Eisenreich, H.; Ellguth, G.; Hanzsche, S.; Henker, S.; Schuffny, R., "A compact on-chip IR-drop measurement system in 28 nm CMOS technology," in Circuits and Systems (ISCAS), 2014 IEEE International Symposium on , vol., no., pp.1219-1222, 1-5 June 2014

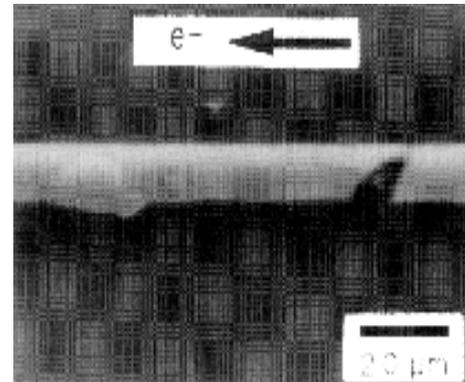
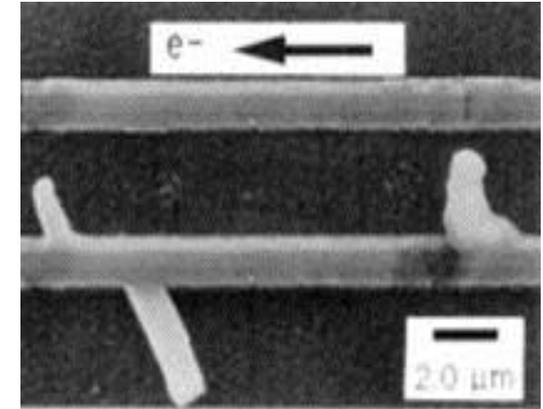
### • Berücksichtigung im Design Flow:

- Statische und dynamische IR-Drop Analyse basierend auf Power Analysen

• → Details siehe Vorlesung VLSI Prozessorentwurf

- Materialtransport durch Stromfluss in metallischen Leitern
- Kritisch bei:
  - Leiterstrukturen mit konstanter Stromrichtung
  - Kleinen Leitungsgeometrien
  - Geometrien
  - Hohen Stromdichten
  - Hohen Temperaturen
- Elektromigration beeinflusst die **Lebensdauer** eines Chips

“Metal hillock”  
→ Kurzschluss



“Metal void”  
→ Unterbrechung  
der Leitung

[1] – Synopsys University Courseware - 90-nm Physical Implementation Flow © 2015 Synopsys, Developed By: Vazgen Melikyan

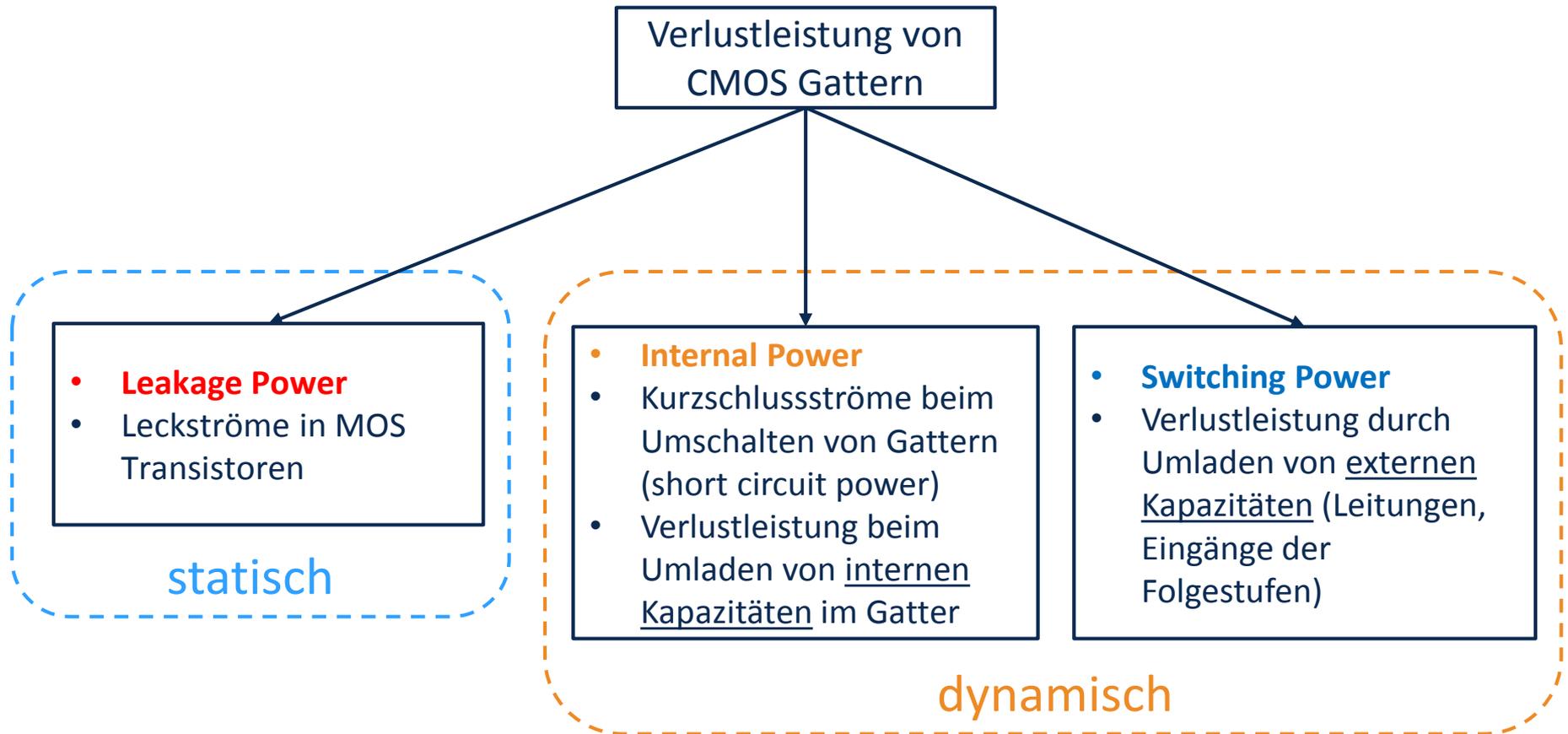
- **Berücksichtigung im Design Flow:**
  - Analyse der Elektromigration basierend auf Power Analysen

• → Details siehe Vorlesung VLSI Prozessorentwurf

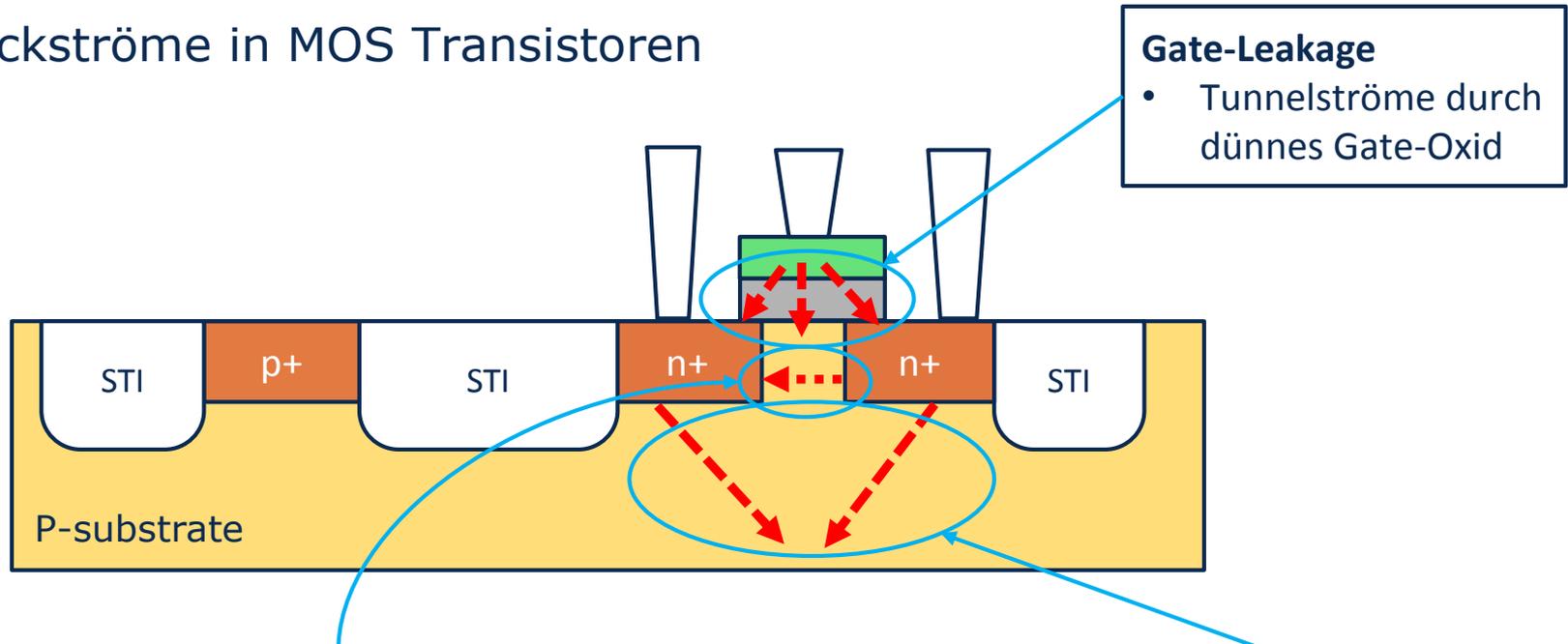
- Die Verlustleistungsaufnahme von CMOS Logik beeinflusst:
  - Das **Systemverhalten** (z.B. Akkulaufzeit)
  - Versorgungsspannung des Chips und damit **Funktionalität und Timing** (IR-Drop)
  - Die **Lebensdauer** des Systems (Elektromigration)
- Berücksichtigung der Verlustleistung bei der Dimensionierung des Versorgungsnetzwerkes
  - Aktive Elemente (z.B. Spannungsregler)
  - Passive Elemente (Layout, Leitungsbreiten, Kondensatoren)

**Eine akkurate Power Analyse ist wichtig!**

- Größen:
  - Stromstärke :  $I(t) = \frac{dQ(t)}{dt}$
  - Spannung:  $V(t)$
  - Verlustleistung:  $P(t)$
  - Energie:  $E(t) = \int_{t_0}^t P(t') dt'$



- Leckströme in MOS Transistoren



**Gate-Leakage**

- Tunnelströme durch dünnes Gate-Oxid

**Sub-Threshold Leakage**

- $V_{GS} < V_{th}$  (Kanal in weak Inversion)
- Diffusion von Ladungsträgern zwischen Drain und Source
- Exponentielle Abhängigkeit von  $V_{GS}$  und  $V_{DS}$  und  $T$

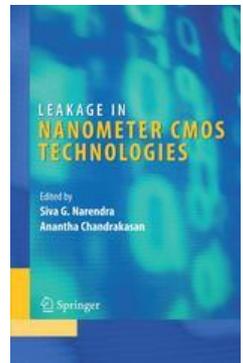
**Reverse Biased Junction Leakage**

- Leckströme durch gesperrte D-B Diode bzw. S-B Diode

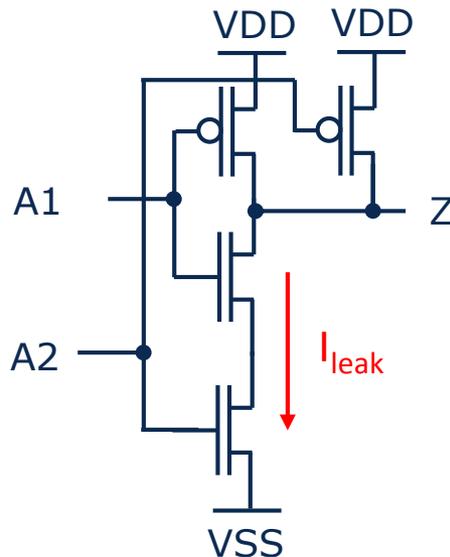
- Sub-Threshold Leckstrom:
  - $I_{leak,subth} \sim \phi_T^2 \cdot \frac{W}{L} \cdot e^{(V_{GS}-V_{th})/n\phi_T} \cdot (1 - e^{-V_{DS}/\phi_T})$  mit  $\phi_T = kT/q$
- **Exponentielle** Abhängigkeit von Spannung und Temperatur
- Leakage Power ist kritisch bei hohen  $V_{DD}$  und hohen Temperaturen
- Reihenschaltung von Transistoren zeigt überproportional geringere Leckströme

- **Berücksichtigung im Design Flow:**
  - Modellierung der Leakage Power in der jeweiligen PVT Corner

- Literatur:
  - Buchkapitel: Leakage in CMOS Circuits – An Introduction; D. Helms, E. Schmidt, and W. Nebel, SPRINGER
  - Buch: Leakage in Nanometer CMOS Technologies; Editors: Narendra, Siva G., Chandrakasan, Anantha P. (Eds.), SPRINGER



- Abhängigkeit des Leckstromes von
  - Transistortyp (NMOS, PMOS)
  - der **Verschaltung** gesperrter Transistoren
- → Leakage Power eines Gatters abhängig von der Eingangsbelegung

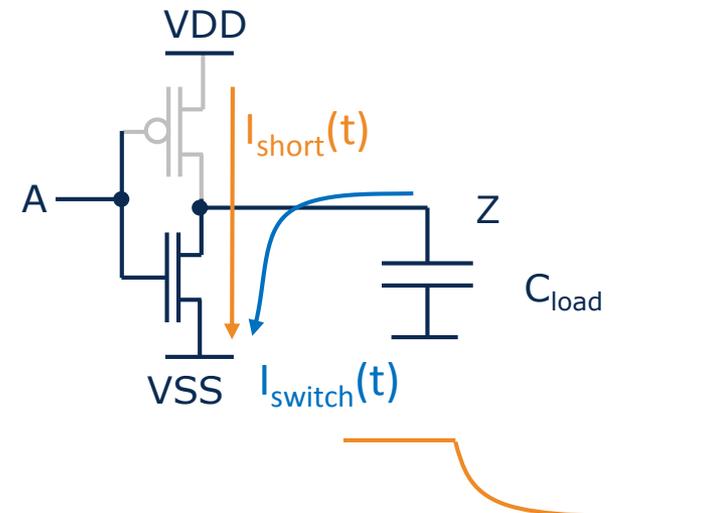
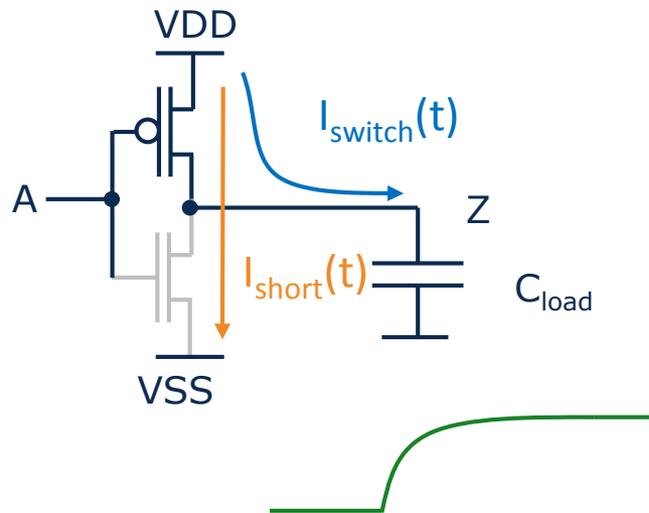


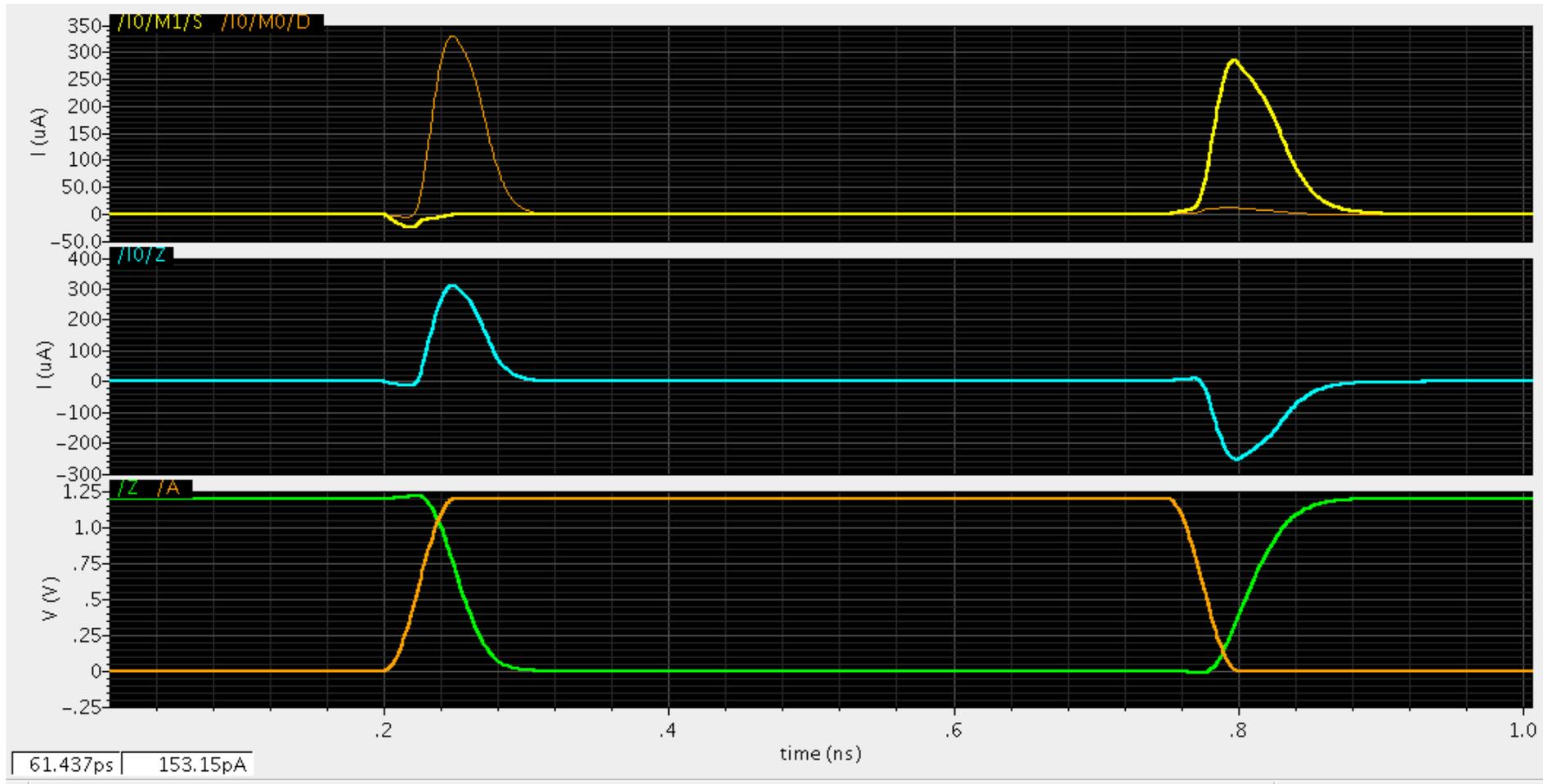
A1	A2	Z	$I_{leak}(\text{pA}) @25^\circ\text{C}$
0	0	1	9.2
0	1	1	49.6
1	0	1	61.6
1	1	0	124.1

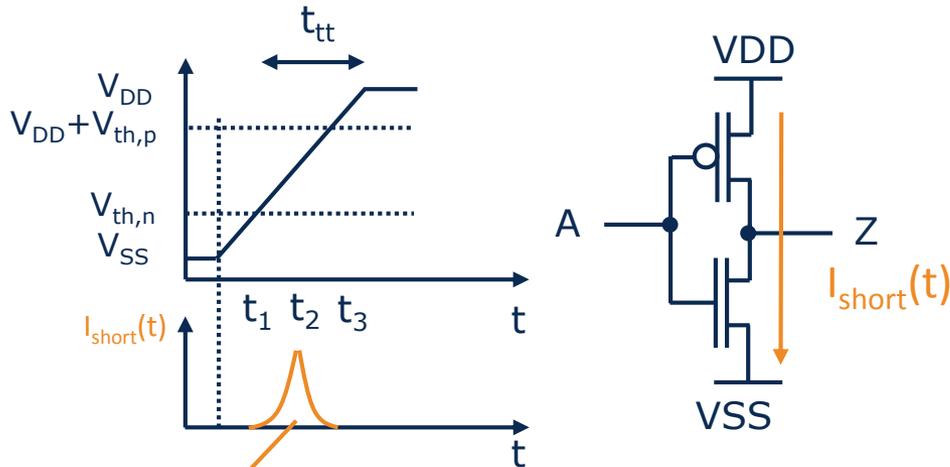
- **Berücksichtigung im Design Flow:**
  - Modellierung der Leakage Power in Abhängigkeit der Eingangssignale

- → Leakage Minimierung durch Wahl geeigneter Reset oder Low-Power Zustände in FlipFlops

- Dynamische Verlustleistung







$$Q_{short} = 2 \cdot \int_{t_1}^{t_2} \frac{\beta}{2} \cdot (V_{in}(t) - V_{th})^2 dt$$



$$Q_{short} = \frac{t_{tt}}{24} \cdot \frac{\beta}{V_{DD}} (V_{DD} - 2 \cdot V_{th})^3$$

• Annahmen:

- $\beta_n = \beta_p$
- $V_{th,n} = -V_{th,p}$

$$V_{in}(t) = \frac{V_{DD}}{t_{tt}} \cdot t$$

$$t_1 = \frac{V_{th}}{V_D} \cdot t_{tt}$$

$$t_2 = \frac{t_{tt}}{2}$$

• → Ladungsmenge  $Q_{short}$  pro Umschaltvorgang

[1] VEENDRICK, H.J.M. Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits. IEEE Journal of Solid State Circuits, New York, v.SC-19, n.4, p. 468-473, Aug. 1984.

- Energie pro Umschaltvorgang aus  $V_{DD}$ :  $E_{short} = V_{DD} \cdot Q_{short}$

- Verlustleistung:  $P_{short} = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}$

- Taktfrequenz  $f = 1/T_{clk}$
- Toggle-Rate  $\alpha$ 
  - $\alpha = 1 \rightarrow$  Daten Toggle in jedem Takt

$$P_{short} = \alpha \cdot f \cdot \frac{t_{tt}}{24} \cdot \beta \cdot (V_{DD} - 2 \cdot V_{th})^3$$

- **Berücksichtigung im Design Flow:**

- Modellierung der internen **Energie pro Schaltvorgang** eines Gatters für einen toggle am Input
- abhängig von der Signalflanke und der Signalbelegung anderer Inputs (siehe AOI12)
- PVT Corner

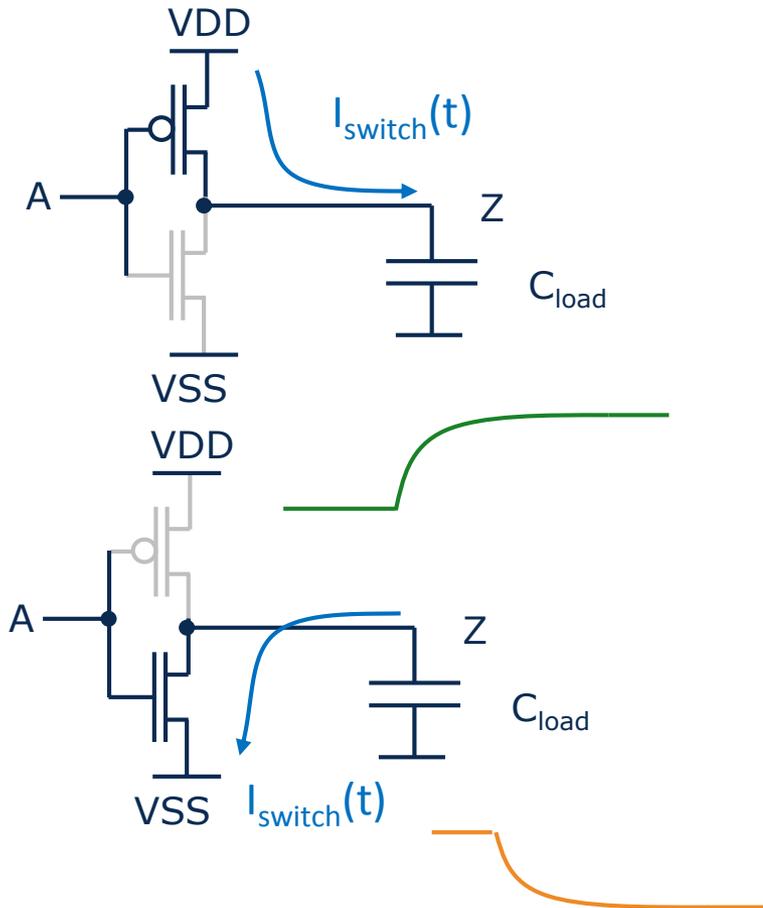
- Energie pro Umschaltvorgang aus  $V_{DD}$ :

$$E_{switch, rise} = V_{DD} \cdot Q_{load}$$

$$E_{switch, fall} = 0$$



$$P_{switch} = \alpha/2 \cdot f \cdot V_{DD}^2 \cdot C_{load}$$



## • Berücksichtigung im Design Flow:

- Berechnung der Energie pro Signalwechsel der externen Kapazitäten, Leitungen, Eingänge der Inputs
- abhängig von der Signalflanke
- PVT Corner

- Mittlere Verlustleistung einer CMOS Schaltung:

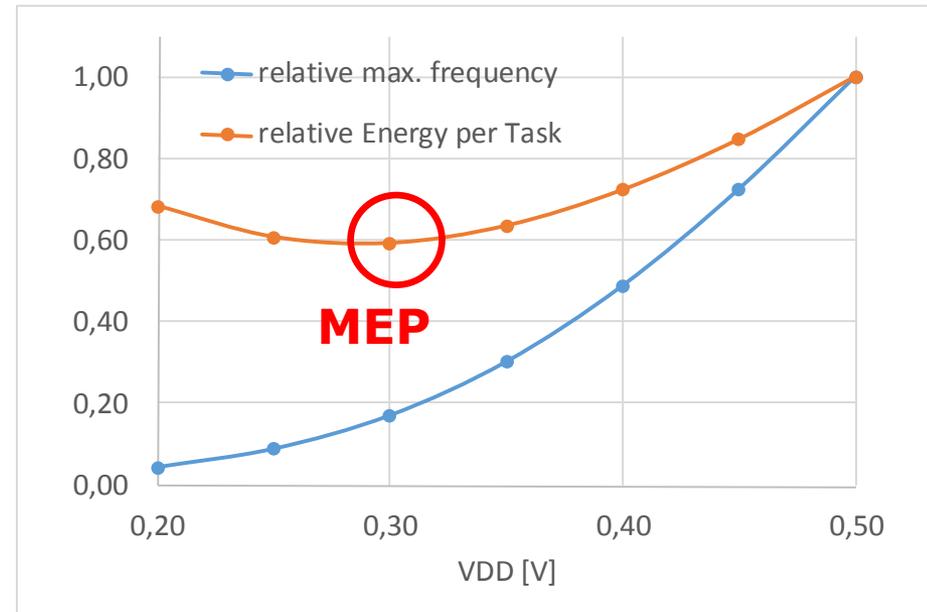
$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

- **Task:**
  - Abfolge von n Takten mit Frequenz  $f=1/T_{clk}$
  - mit  $n \cdot \alpha$  Toggles
- Energie pro Task:

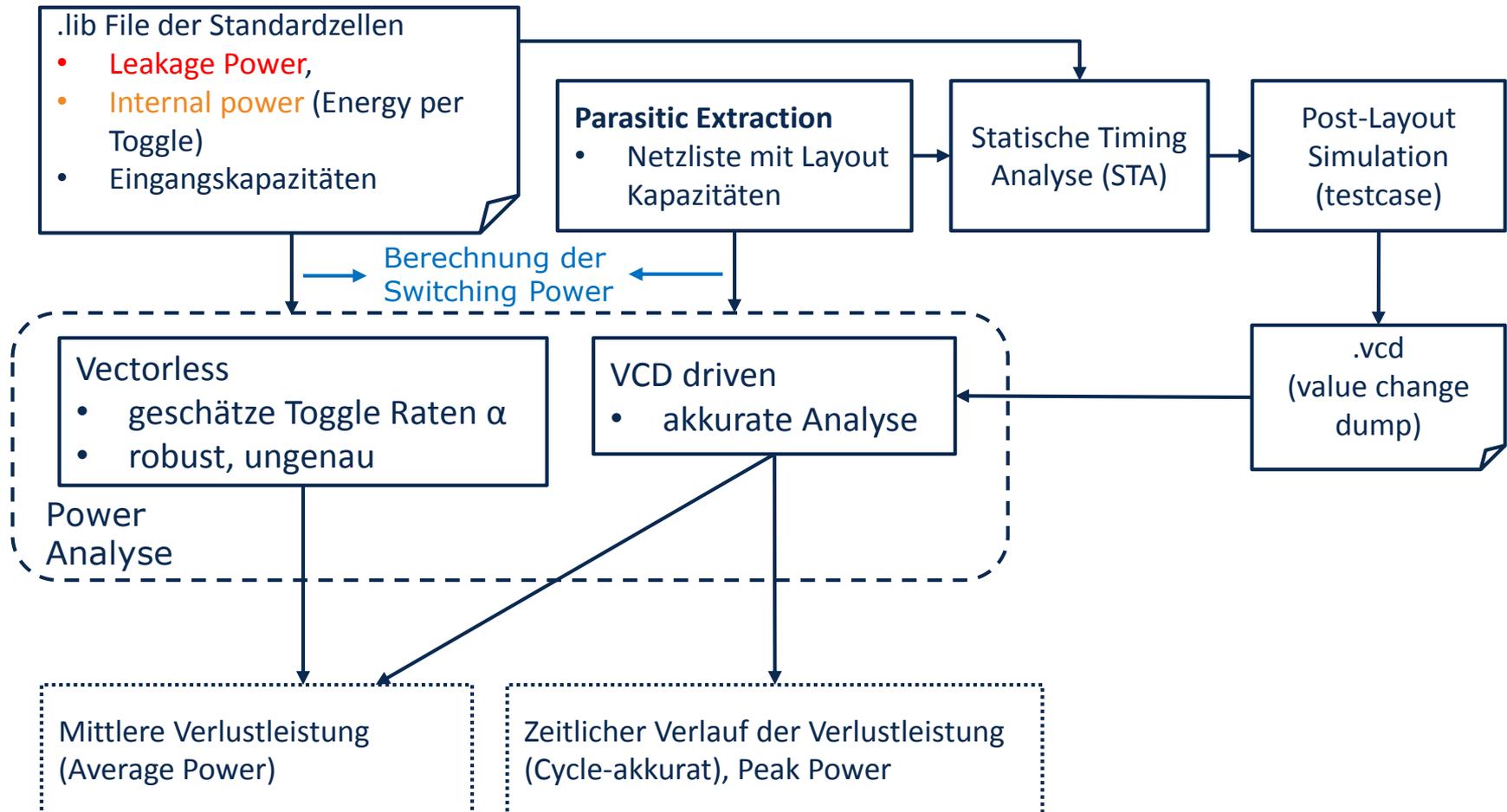
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

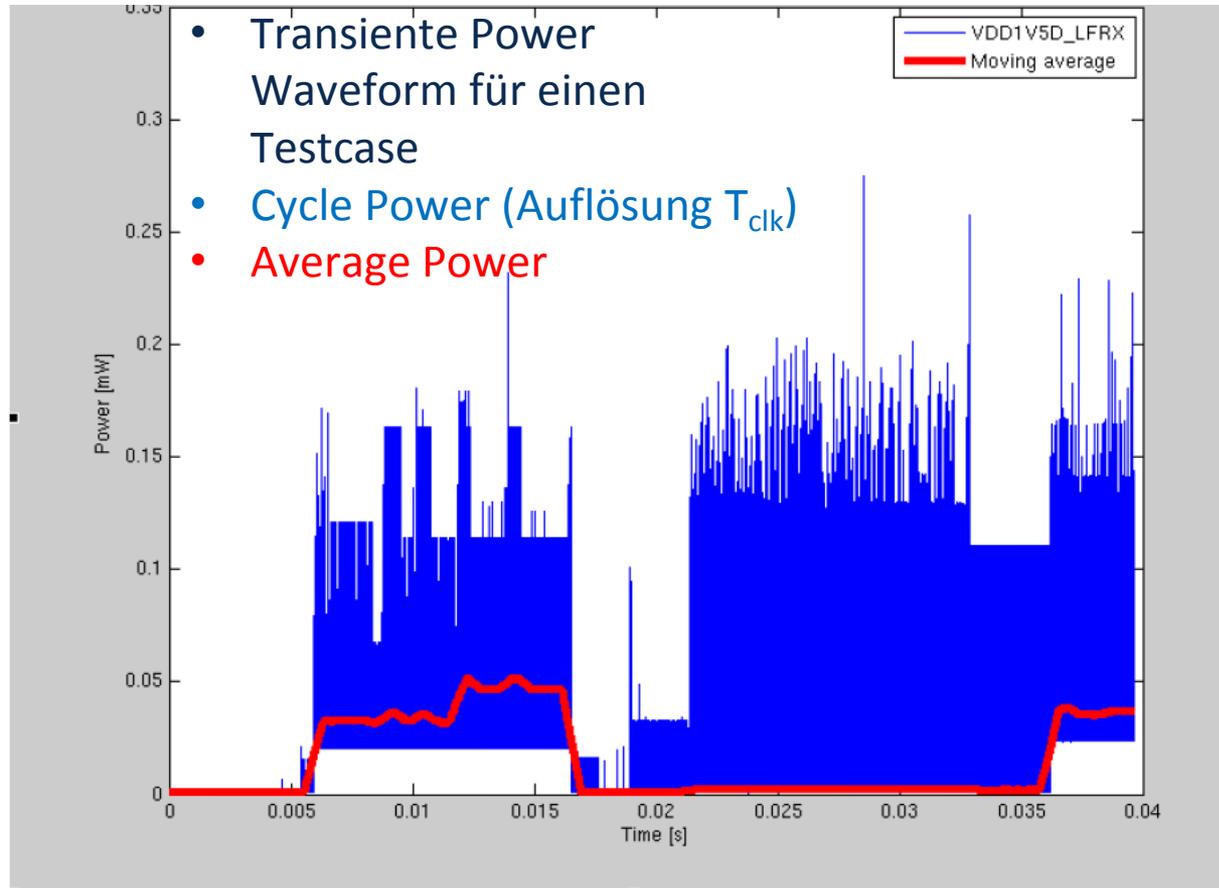
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot 1/f_{CLK}$$

- Beispiel CMOS Logik:
  - max. Taktfrequenz und Energie pro Task
- Hohe  $V_{DD}$ :
  - Hohe dynamische Energieaufnahme, schnelle Taktfrequenz
- Kleine  $V_{DD}$ :
  - Geringe dynamische Stromaufnahme
  - Langsame Taktfrequenz
  - → **Hohe Leakage Energie**

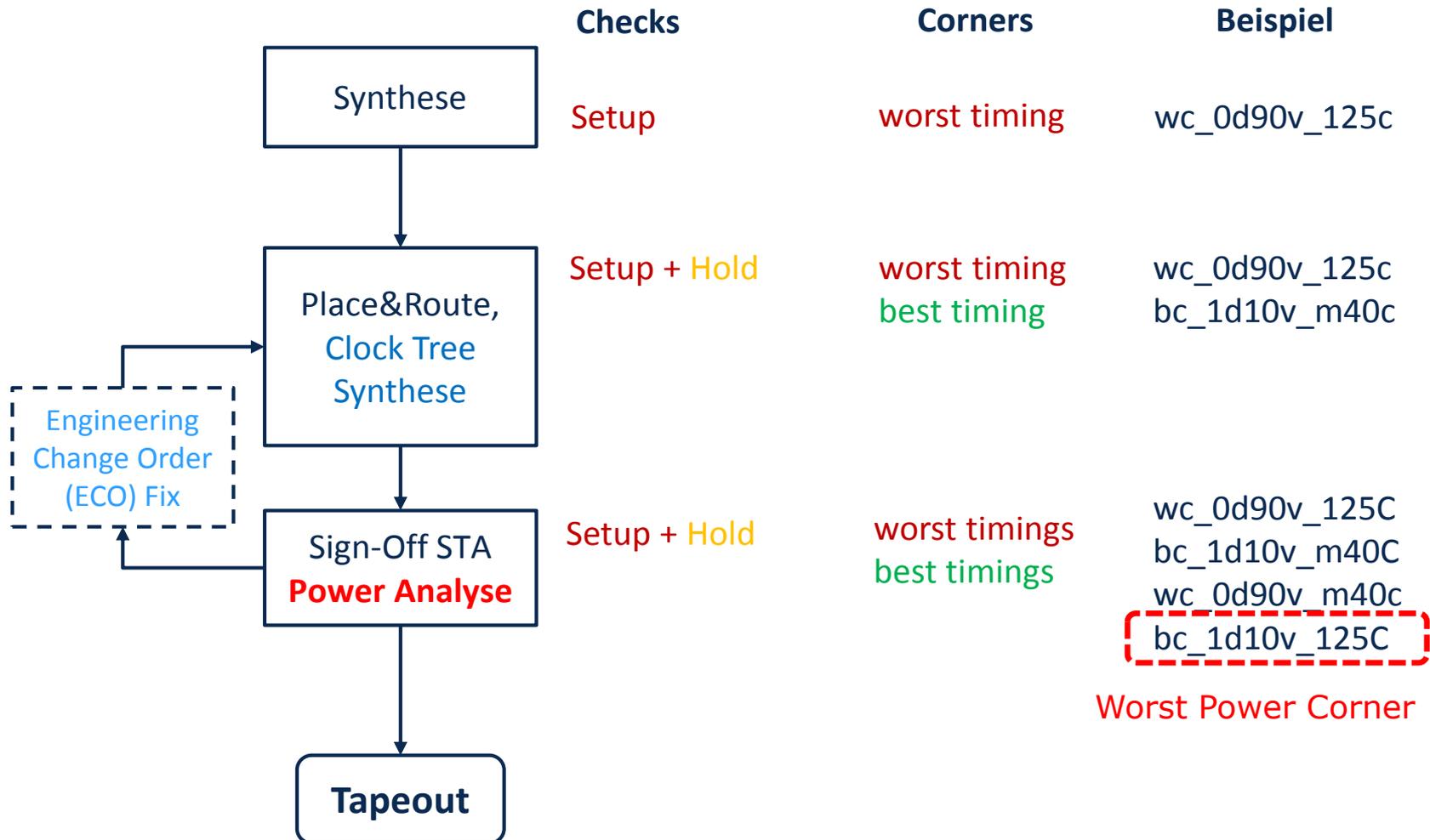


- Es existiert ein Arbeitspunkt mit minimaler Energie pro Task (Minimum Energy Point MEP)
- MEP meist im Sub-Threshold Bereich
- → Ultra-Low Power Schaltungsdesign im Sub-Threshold Betrieb



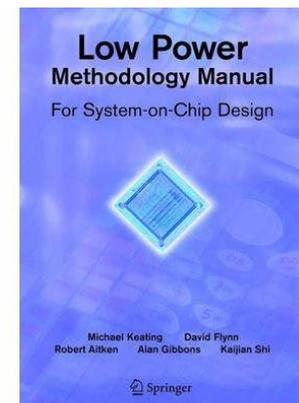


- → Details siehe Vorlesung VLSI Prozessorentwurf



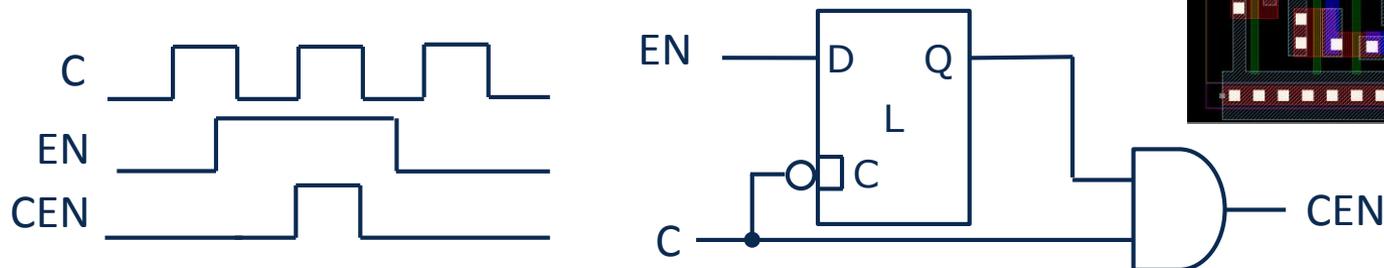
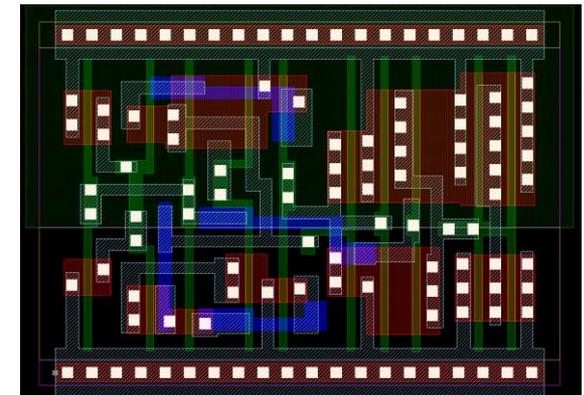
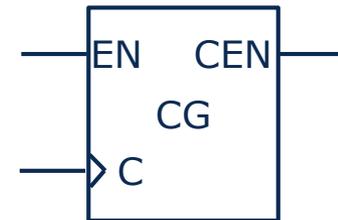
# Low Power Schaltungsentwurf

- Maßnahmen zur Reduktion der Verlustleistung und/oder zur Erhöhung der Energieeffizienz von CMOS Schaltungen
- Anwendung bei
  - Architekturentwurf → größte Effizienzgewinne möglich
  - Schaltungsimplementierung → transparente Implementierung
- Low Power Techniken:
  - Clock Gating
  - Multi-Vt Implementierung
  - Power-Shut-off
  - (Dynamic) Frequency Scaling (DFS)
  - (Dynamic) Voltage and Frequency Scaling (DVFS)
  - Adaptive Voltage and Frequency Scaling (AVFS)
- Literatur: Low Power Methodology Manual For System-on-Chip Design; Michael Keating, SPRINGER



$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

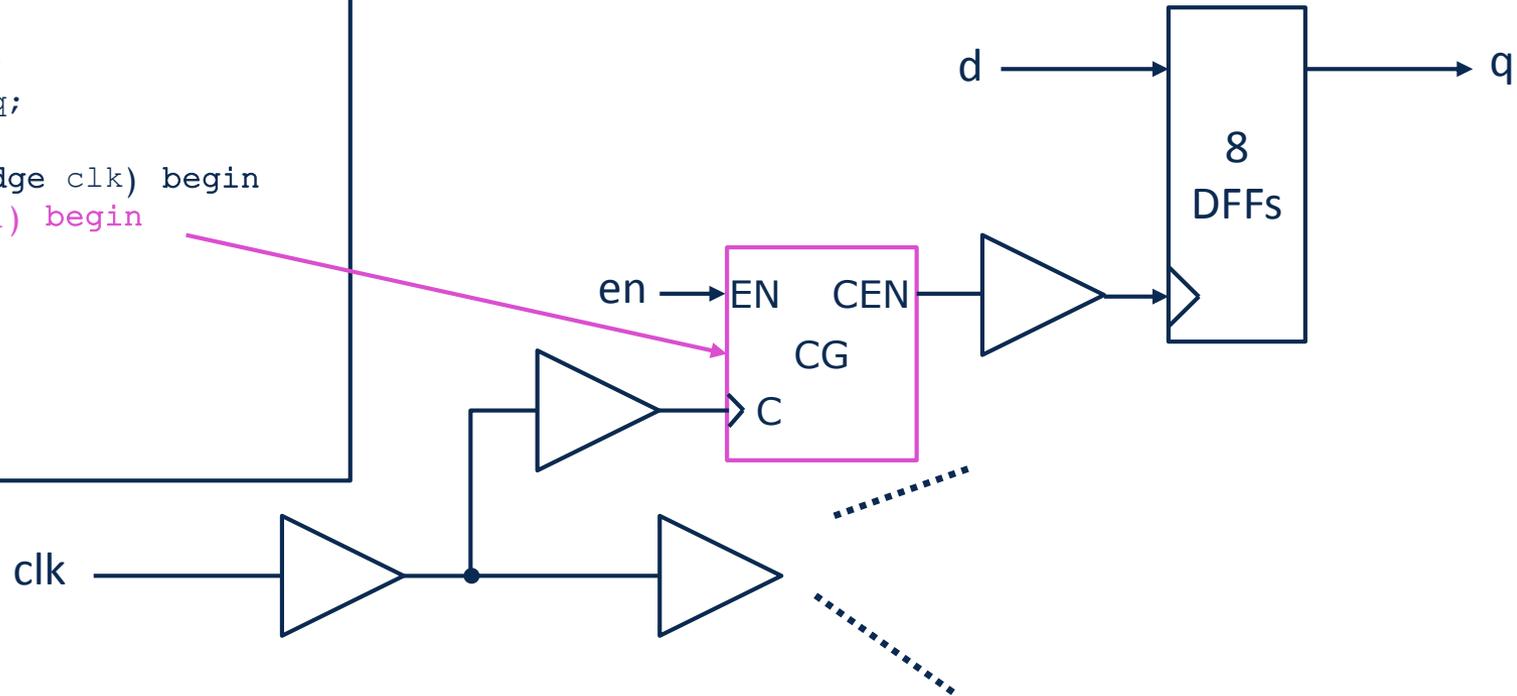
- Anhalten des Taktsignals wenn nicht benötigt
- → Reduktion der Toggle Rate von Taktnetzen.
- Einfügen von Clock Gate Zellen in das Taktnetzwerk
- Zyklen-akkurates Enable von Clock Gates
- Implementierung von Clock Gates
  - Manuelle Instanziierung in der RTL
  - Automatisiert durch die Synthese



```

module reg_gated (clk, en, d, q)
  input clk;
  input en;
  input [7:0] d;
  output [7:0] q;
  reg [7:0] q;
  always @(posedge clk) begin
    if (en==1'b1) begin
      q<=d;
    end
  end
  assign q=q;
endmodule

```



## • Vorteile:

- Einfache Implementierung
- Keine Funktionalen Änderungen im Design
- Sehr effizient

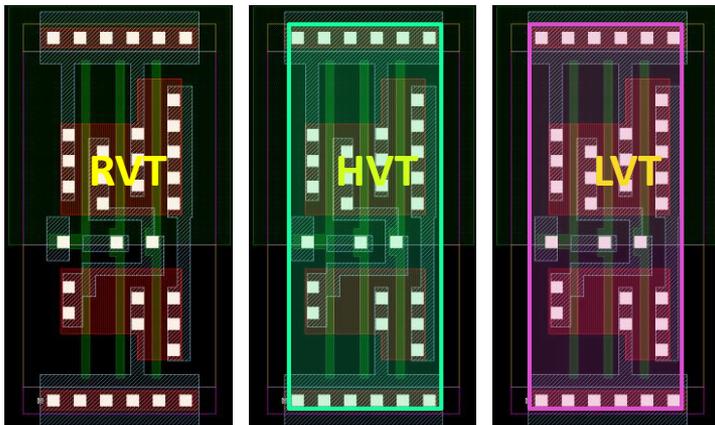
## • Nachteile:

- Layout Overhead
- Timing des Enable Signals kann kritisch sein
- Power Overhead bei hohen Toggle Raten

$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

- Kompromiss zwischen Leckströmen, Short Circuit Strömen und Delay der Standardzellen von Zellen
- Zusammenhang über Schwellspannung  $V_{th}$ 
  - High VT :  $V_{th} \uparrow \rightarrow td \uparrow, Q_{short} \downarrow, I_{leak} \downarrow$
  - Low VT :  $V_{th} \downarrow \rightarrow td \downarrow, Q_{short} \uparrow, I_{leak} \uparrow$
- Die höhere Treiberstärke von Low VT Gattern kann die notwendige Weite und damit  $C_{in}$  reduzieren.  $\rightarrow$  reduktion der Switching Power
- $\rightarrow$  Nutzung von Standardzellen mehrerer Schwellspannungen in einer Schaltung
- Nutzung von LVT Zellen nur in kritischen Timing Pfaden

- Multi-VT Bibliotheken: **Identisches Layout** der Zellen, Unterscheidung durch Marker Layer
- Tausch der Zellen durch das Synthese bzw. Place& Route Tool



	LVT	RVT	HVT
Rel. $V_{th}$	0.8	1.0	1.2
Area	1.0	1.0	1.0
Rel. Treiberstärke	1.2	1.0	0.8
Rel. Leckstrom	8.8	1.0	0.15

### • Vorteile:

- Einfache Implementierung
- Keine Funktionalen Änderungen im Design
- Sehr effizient

### • Nachteile:

- Höhere Maskenkosten (zusätzliche Masken für Schwellspannungsoption)
- Komplexeres Temperaturverhalten möglich (Temperaturinversion)

$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

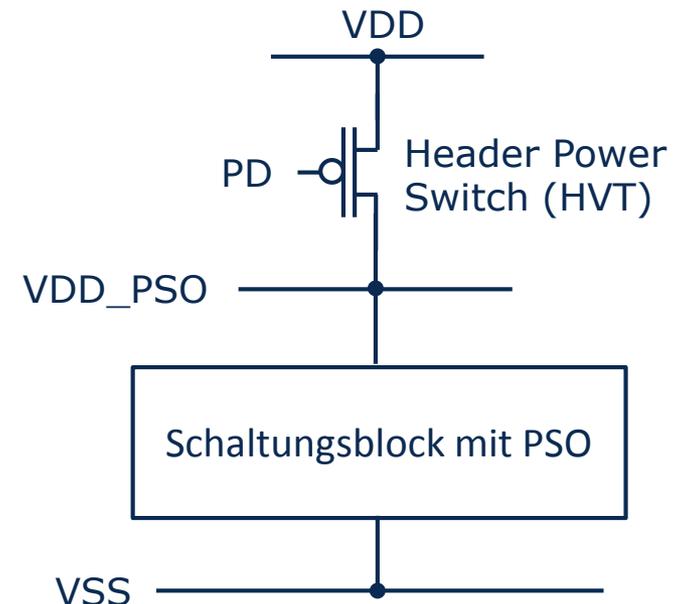
- Abschalten nicht aktiver Schaltungsteile durch Trennung von der Versorgungsspannung
- Nutzung von Power Switches (Header PMOS, oder Footer NMOS)
- Integration der Switches in das Power Mesh

### • Vorteile:

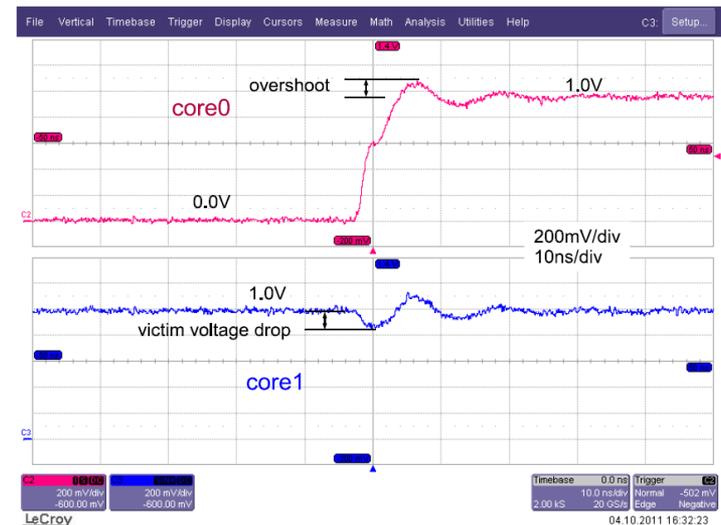
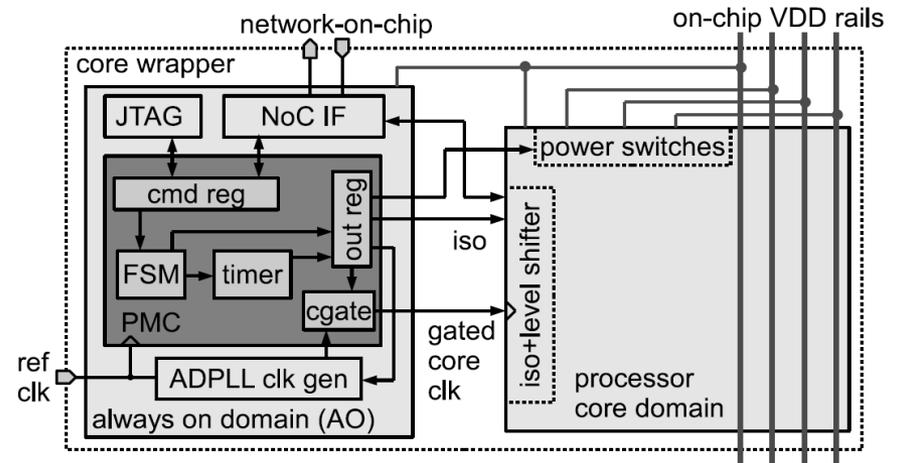
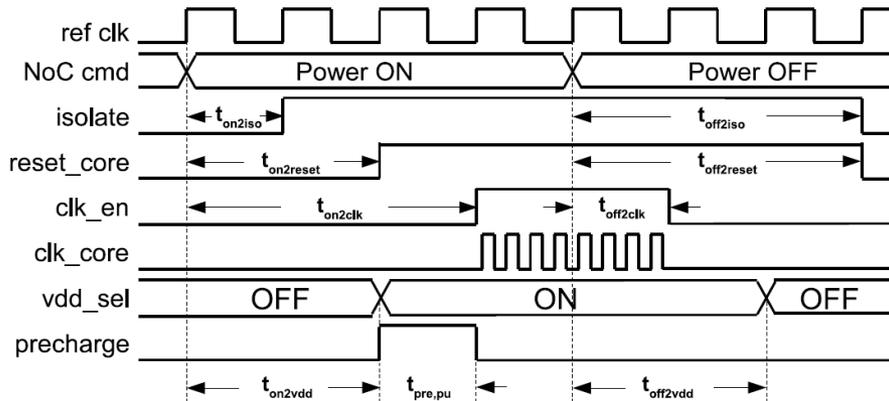
- Effiziente Reduktion des Leckstroms

### • Nachteile:

- Flächen-Overhead
- Komplexeres Power Mesh Design
- IR-Drop über dem Switch
- Architekturanpassung erforderlich (Power Management Controller, Berücksichtigung von PSO in der Ablaufsteuerung)



- Aktive Steuerung des PD Signals:
- zeitlicher Ablauf
- Isolation Logik (Verhindern von Treiben von X in aktive Logik)
- Ggf. Wiederherstellen von gespeicherten Daten nach dem Power-up



S. Höppner, C. Shao, H. Eisenreich, G. Ellguth, M. Ander, R. Schüffny, A Power Management Architecture for Fast Per-Core DVFS in Heterogeneous MPSoCs, IEEE International Symposium on Circuits and Systems, 2012, p. 261-264,

$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

- Dynamische Reduktion der Taktfrequenz von Schaltungsteilen bei geringem Datendurchsatz
- Erfolgt durch programmierbaren Taktgenerator (z.B. PLL)
- Reduktion der Peak Power
  - Entlastung des Power Mesh (IR-Drop, EM)
  - Entlastung von Spannungsversorgung und Kühlung des Systems
- Dauer eines Tasks verlängert sich
- → **DFS erhöht  $E_{task}$  durch erhöhte Leakage Energie**

## • Vorteile:

- Effektive Reduktion der Verlustleistung
- Kein Eingriff in die Spannungsversorgung

## • Nachteile:

- Benötigt programmierbaren Taktgenerator
- Architekturanpassung erforderlich (Ablaufsteuerung)
- Verschlechtert die Energieeffizienz

$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

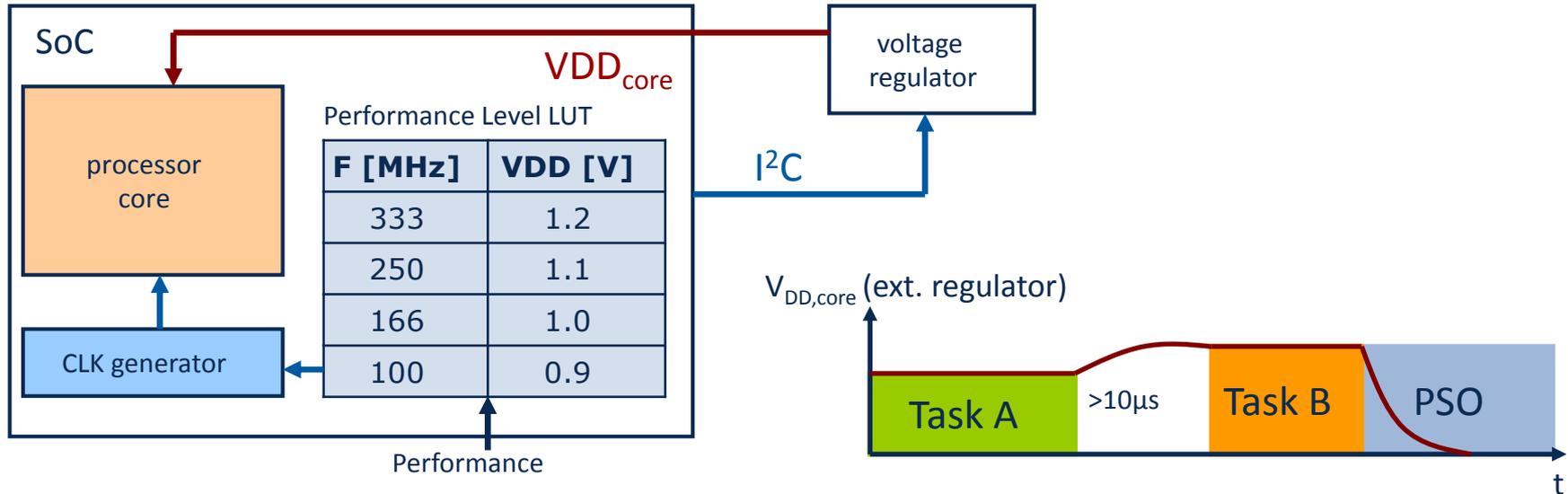
- Skalierung der Taktfrequenz bei gleichzeitiger Skalierung der Versorgungsspannung
- Anpassung der Performanz (Performance Level) des Systems an die aktuelle Anforderung
  - Hohe Performance :  $V_{DD} \uparrow, f \uparrow, E_{task} \uparrow$
  - Geringe Performance :  $V_{DD} \downarrow, f \downarrow, E_{task} \downarrow$
- Benötigt programmierbaren Taktgenerator und Spannungsversorgung

## • Vorteile:

- Effektive Reduktion der **Verlustleistung und Energie** pro Task

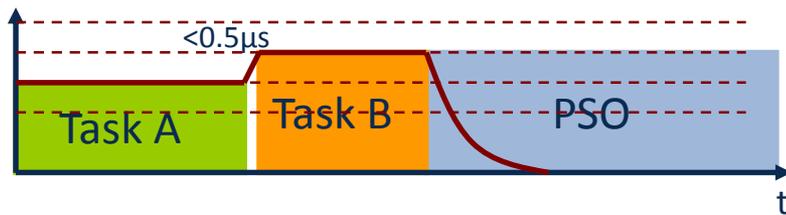
## • Nachteile:

- Benötigt programmierbaren Taktgenerator und Spannungsversorgung
- Komplizierte Sign-Off Analysen mit mehreren nominalen Spannungen
- Architekturanpassung erforderlich (Ablaufsteuerung)

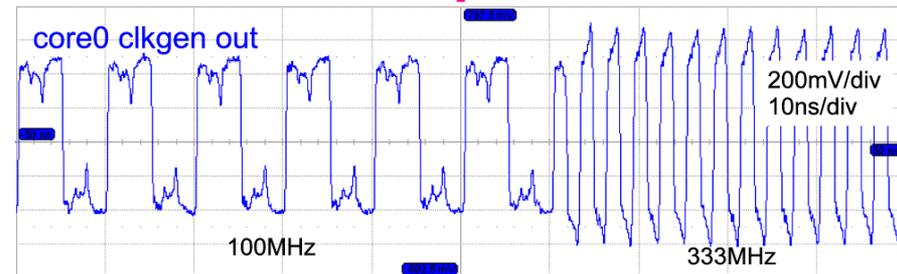
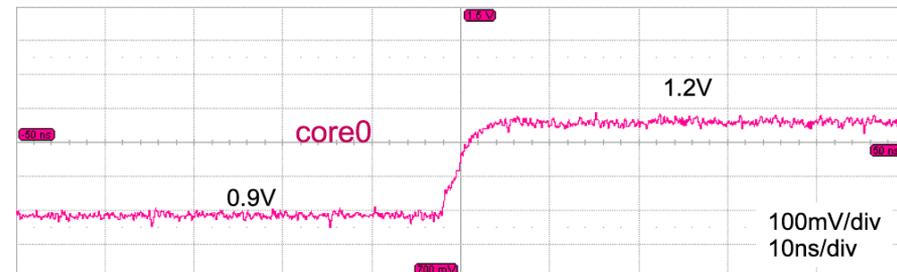
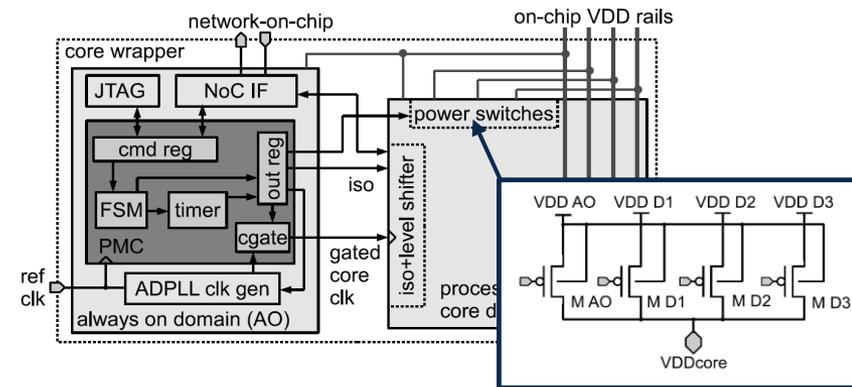


- Performance Level Lookup Tabelle
  - (V<sub>DD</sub>,f) Kombinationen
  - Festgelegt zur Entwurfszeit bzw. nach dem Test der Chips
  - Statisch im Betrieb
- Berücksichtigung des Spannungsbereichs beim Timing Sign-Off nötig (Hold Violations!)

core  $V_{DD}$  (discrete on-chip switching)



- Schnelles Umschalten zwischen verschiedenen Spannungen Spezielle Ablaufsteuerung nötig zur Reduktion des IR-Drops (pre-charge)
- In Kombination mit ADPLL Taktgenerator ist Wechsel des Performance Levels in  $<20ns$  möglich



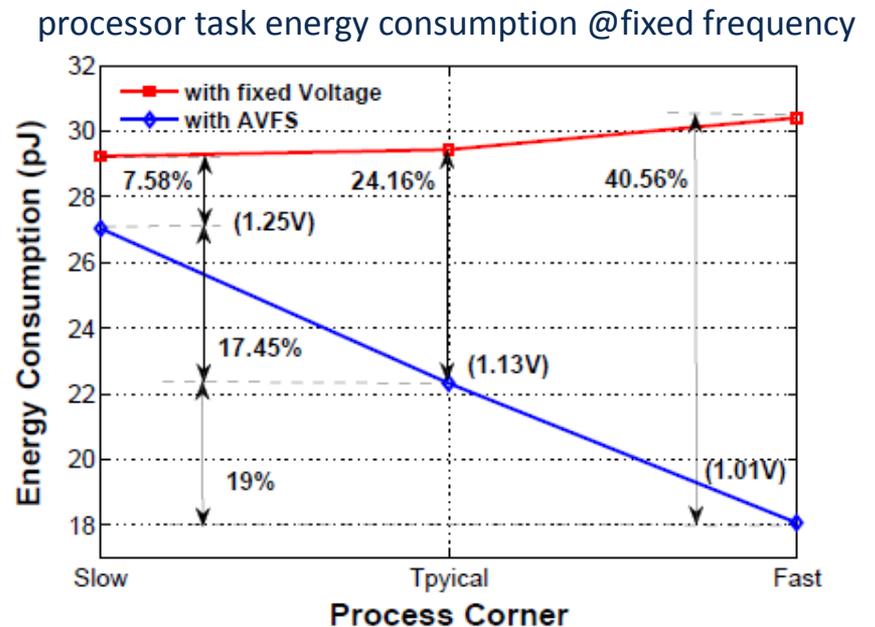
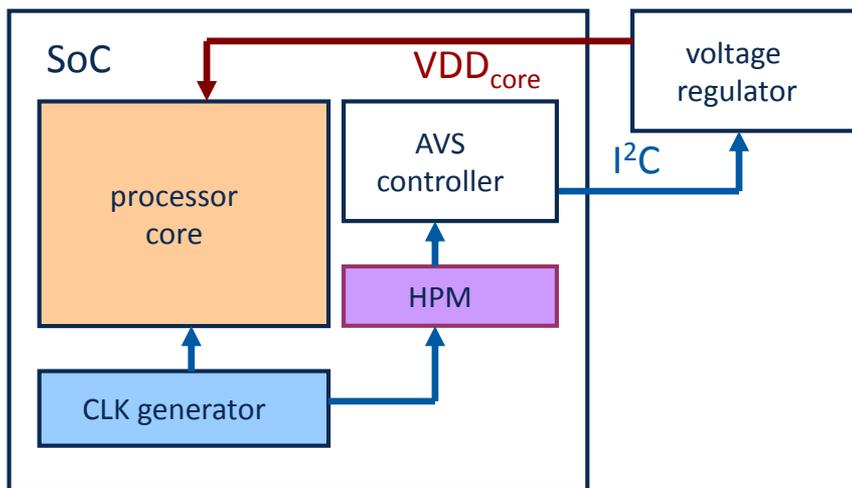
S. Höppner, C. Shao, H. Eisenreich, G. Ellguth, M. Ander, R. Schüffny, A Power Management Architecture for Fast Per-Core DVFS in Heterogeneous MPSoCs, IEEE International Symposium on Circuits and Systems, 2012, p. 261-264,

$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

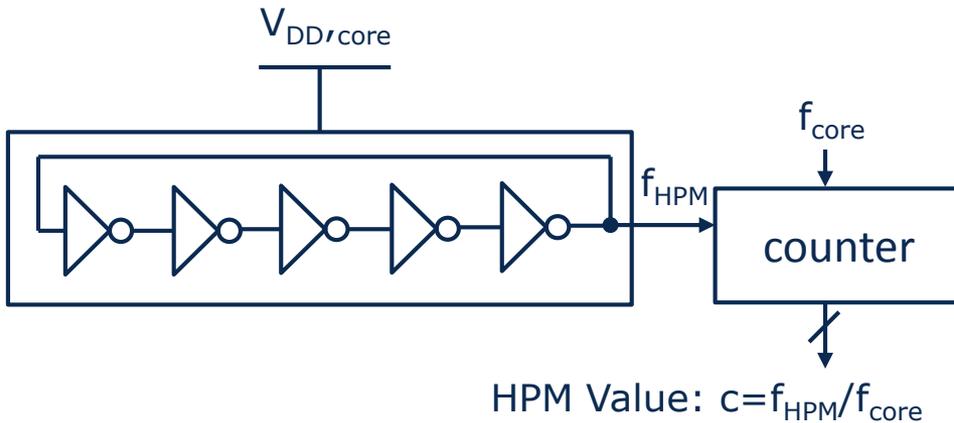
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

- Skalierung der Taktfrequenz gemäß der Performance Anforderung bei gleichzeitiger **autonomer, adaptiver Skalierung** der Spannung
- Anpassung der Performanz (Performance Level) des Systems
  - Hohe Performance :  $f \uparrow \rightarrow V_{DD} \uparrow, E_{task} \uparrow$
  - Geringe Performance :  $f \downarrow \rightarrow V_{DD} \downarrow, E_{task} \downarrow$
- Betrieb der Schaltung mit der **minimalen VDD** für die jeweilige Frequenz
- Benötigt programmierbaren Taktgenerator, Spannungsversorgung und Hardware Performance Monitor (HPM)

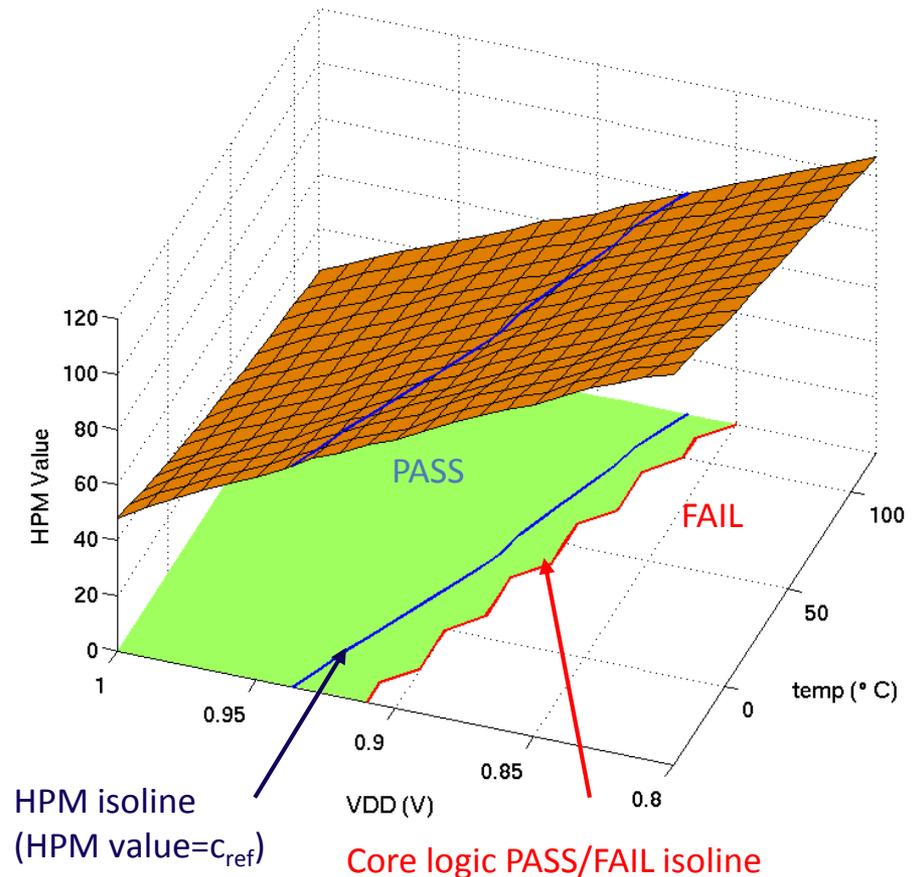
- Minimierung der Versorgungsspannung, bis das kritische Timing des Designs ausreichend ist für die aktuelle Taktfrequenz
- Kritisches Timing abhängig von:
  - Prozessrealisierung: individuell für jeden Chip
  - Temperatur : ändert sich im Betrieb
  - Versorgungsspannung: **Adaption durch AVFS**



- Ring-Oszillator (RO) HPM → Replica (Kopie) des kritischen Timings des Designs
- Vergleich des HPM Timings  $f_{\text{HPM}}$  mit fester Taktfrequenz  $f_{\text{core}}$  durch Zähler  $c$



- Kalibrierung:  $c = c_{\text{ref}}$
- Verhältnis aus kritischem Design Timing zu RO Periode
- AVFS Regelung von  $V_{\text{DD,core}}$ :
  - $c < c_{\text{ref}}$ :  $V_{\text{DD,core}} \uparrow$
  - $c > c_{\text{ref}}$ :  $V_{\text{DD,core}} \downarrow$

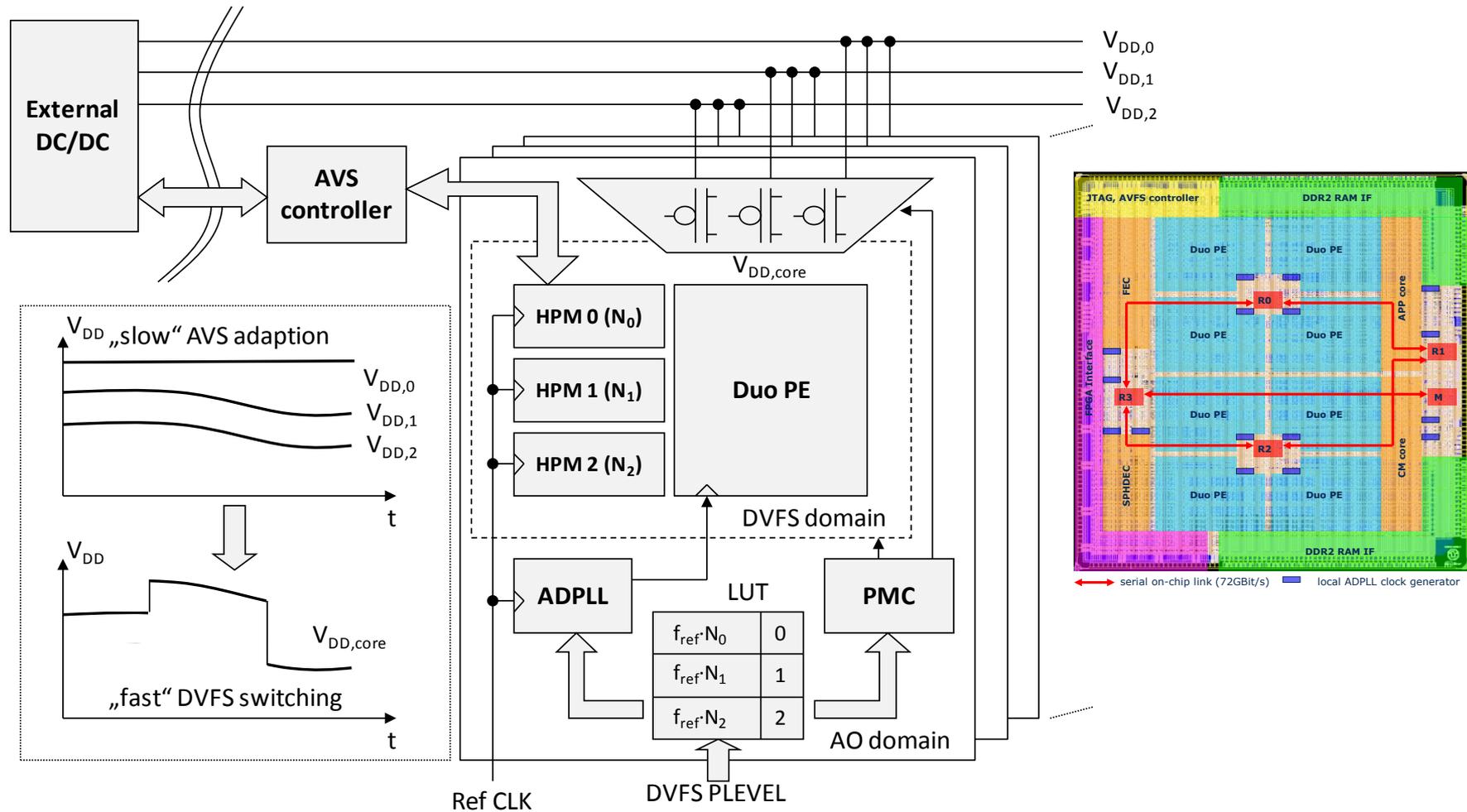


- **Vorteile:**

- Sehr effektive Reduktion der **Verlustleistung und Energie** pro Task individuell pro Chip
- Reduktion des Pessimismus durch PVT Corner durch Adaption
- Kein Einfluss auf Ablaufsteuerung wenn Frequenz nicht geändert wird (AVS)

- **Nachteile:**

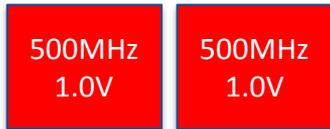
- Benötigt programmierbaren Taktgenerator, Spannungsversorgung und HPM
- HPM benötigt Kalibrierung (erhöht Kosten für Test)
- Komplizierte Sign-Off Analysen mit mehreren nominalen Spannungen
- Architekturanpassung erforderlich (Ablaufsteuerung) bei AVFS



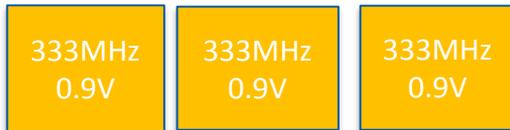
Noethen, B. and Arnold, O. and Perez Adeva, E. and Seifert, T. and Fischer, E. and Kunze, S. and Matus, E. and Fettweis, G. and Eisenreich, H. and Ellguth, G. and Hartmann, S. and Hoepfner, S. and Schiefer, S. and Schluessler, J.-U. and Scholze, S. and Walter, D. and Schueffny, R., 10.7 A 105GOPS 36mm<sup>2</sup> heterogeneous SDR MPSoC with energy-aware dynamic scheduling and iterative detection-decoding for 4G in 65nm CMOS, Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International (2014), 188-189,

- Berücksichtigung von Power Management Techniken auf Architekturebene.
- Planung der Maßnahmen notwendig
  - Steuerung des Power Managements (Hardware, Software)
  - Verifikation der Power Management Maßnahmen
  - Testbarkeit und Test
  - Off-Chip Power Supply
- Ansätze:
  - Dark Silicon:
    - Implementierung dedizierter Hardware Beschleuniger, Power-shut-off wenn nicht verwendet
  - Grey Silicon:
    - Versorgungsspannungsreduktion und Parallelisierung

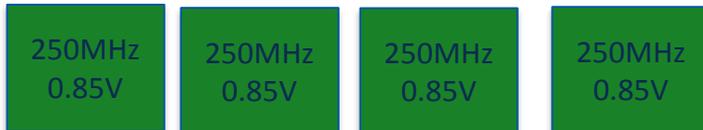
- Technologieskalierung Beispiel: 65nm → 28nm:
  - Fläche Faktor 1/4, Energie/Operation: Faktor 1/3
    - Erhöhung der Leistungsdichte
    - Nutzung von mehr Fläche (Parallelverarbeitung) zur Energiereduktion



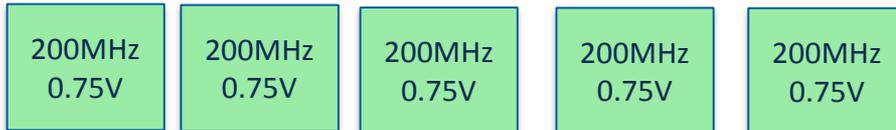
→ 1Gop/s, 2x112mW=**224mW**, **446pJ/op**



→ 1Gop/s, 3x61mW=**183mW**, **363pJ/op**



→ 1Gop/s, 4x40mW=**160mW**, **325pJ/op**



→ 1Gop/s, 5x25mW=**125mW**, **252pJ/op**

- Die Berücksichtigung der Verlustleistung von CMOS Schaltung ist notwendig für die Systemintegration
- Modellierung der Verlustleistung
  - Leakage Power
  - Internal Power
  - Switching Power
- Low Power Schaltungstechniken und Architekturen zur Reduktion der Verlustleistung und Erhöhung der Energieeffizienz