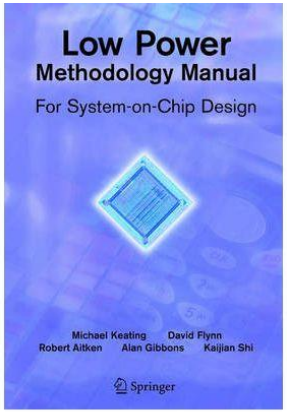
Low Power Schaltungsentwurf

- Maßnahmen zur Reduktion der Verlustleistung und/oder zur Erhöhung der Energieeffizienz von CMOS Schaltungen
- Anwendung bei
  - Architekturentwurf → größte Effizienzgewinne möglich
  - Schaltungsimplementierung → transparente Implementierung

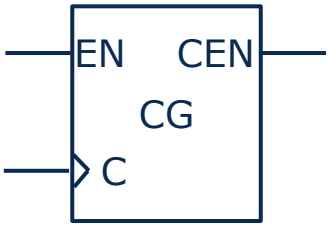
- Low Power Techniken:
  - Clock Gating
  - Multi-Vt Implementierung
  - Power-Shut-off
  - (Dynamic) Frequency Scaling (DFS)
  - (Dynamic) Voltage and Frequency Scaling (DVFS)
  - Adaptive Voltage and Frequency Scaling (AVFS)
  - Adaptive Body Biasing (ABB)

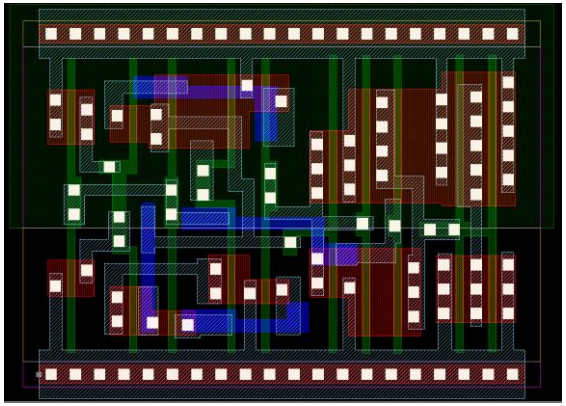
- Literatur: Low Power Methodology Manual For System-on-Chip Design; Michael Keating, SPRINGER

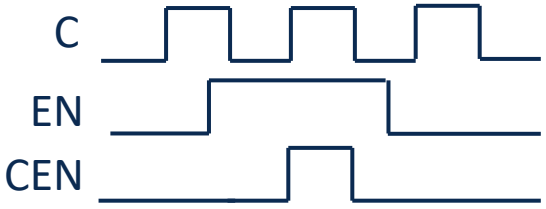


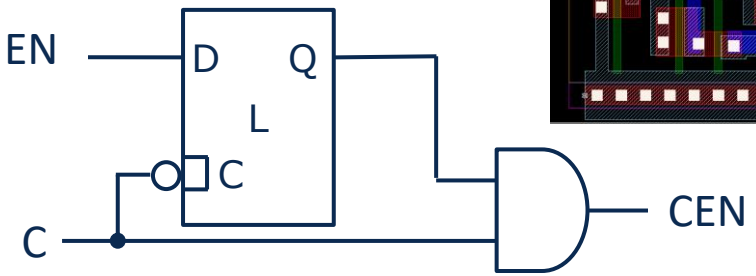
$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

- 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



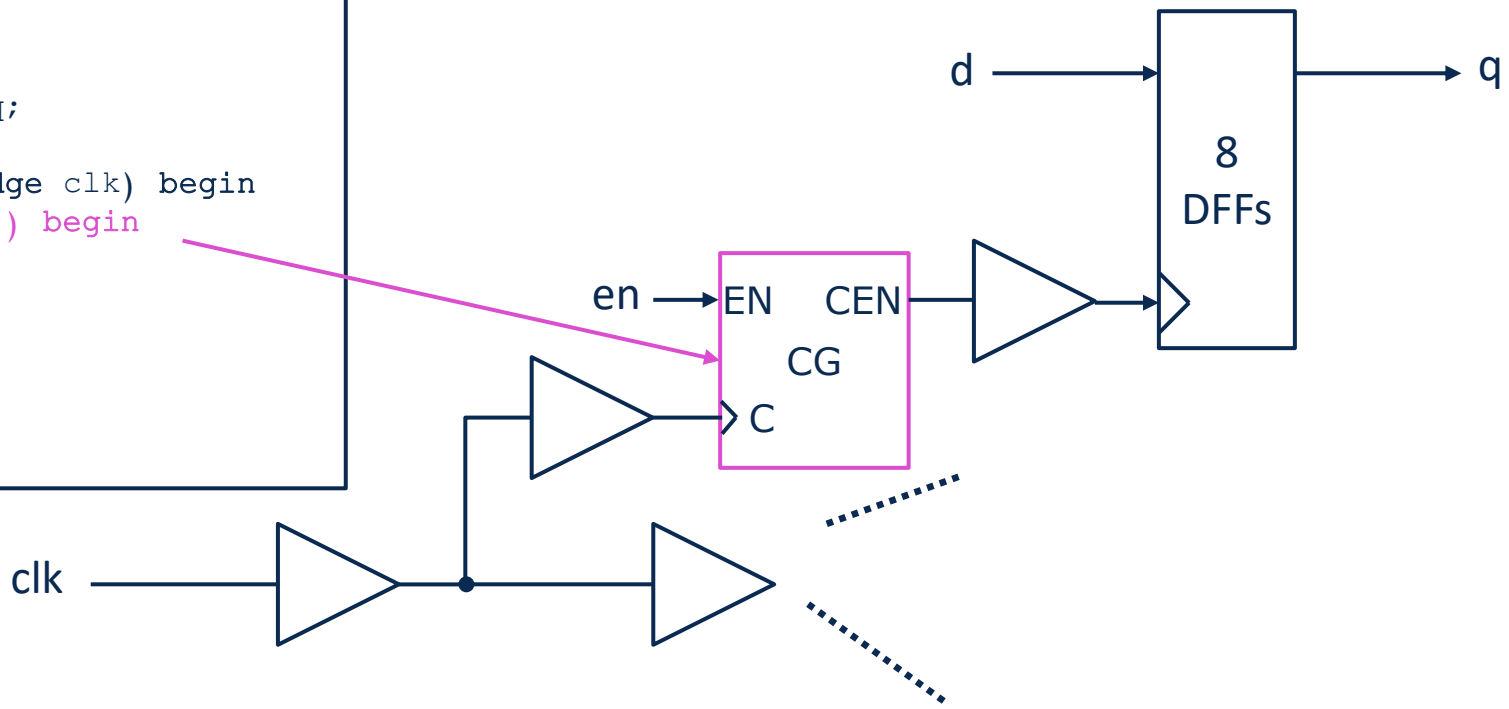






Beispiel: Clock Gating

```
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

Multi-Vt Implementierungen

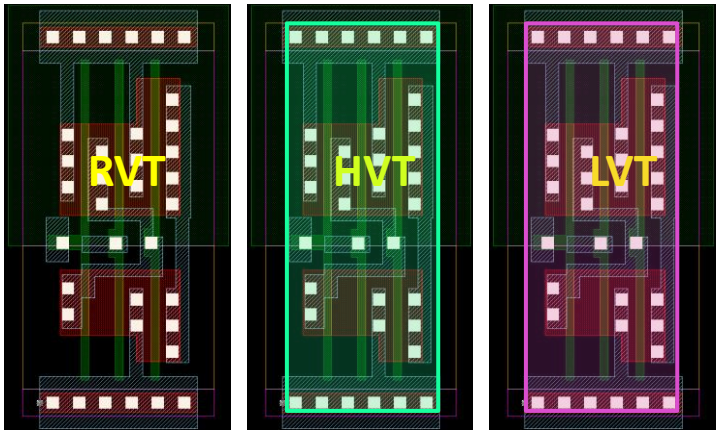
$$P(t) = \alpha \cdot f \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot f \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD})$$

- Kompromiss zwischen Leckströmen, Short Circuit Strömen und Delay der Standardzellen von Zellen
- Zusammenhang über Schwellspannung V_{th}
  - High VT : V_{th} ↑ → td ↑ , Q_{short} ↓, Ileak ↓
  - Low VT : V_{th} ↓ → td ↓ , Q_{short} ↑, Ileak ↑
- Die höhere Treiberstärke von Low VT Gattern kann die notwendige Weite und damit C_{in} reduzieren. → reduktion der Switching Power

- → Nutzung von Standardzellen mehrerer Schwellspannungen in einer Schaltung
- Nutzung von LVT Zellen nur in kritischen Timing Pfaden

Multi-Vt Implementierungen

- 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

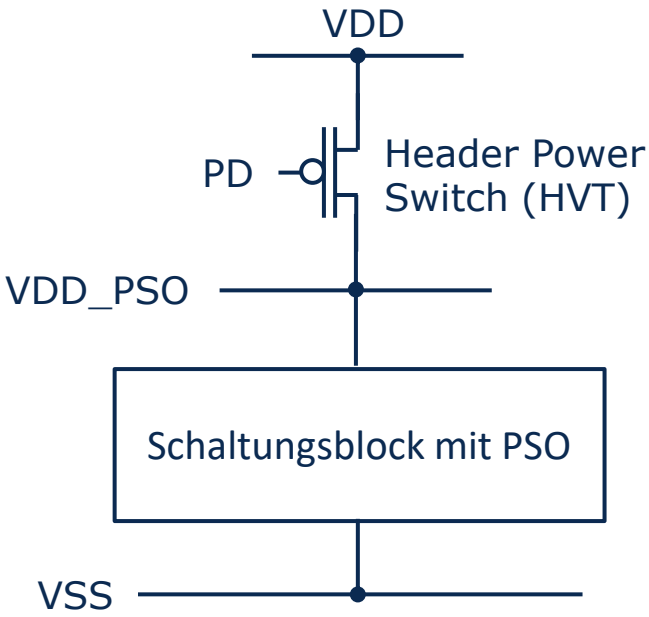
- **Vorteile:**
  - Einfache Implementierung
  - Keine Funktionalen Änderungen im Design
  - Sehr effizient

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

Power-Shut Off (PSO)

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

- 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

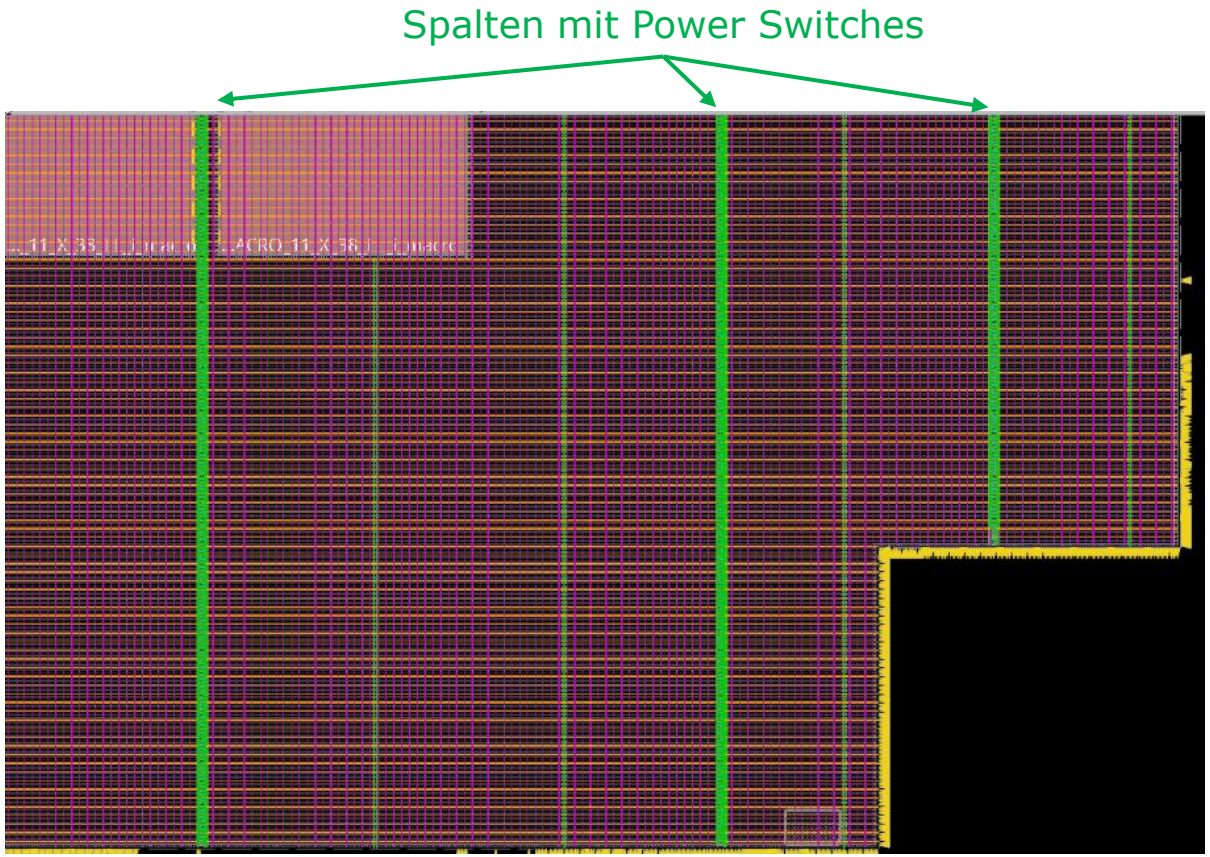


- **Vorteile:**
  - Effiziente Reduktion des Leckstroms

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

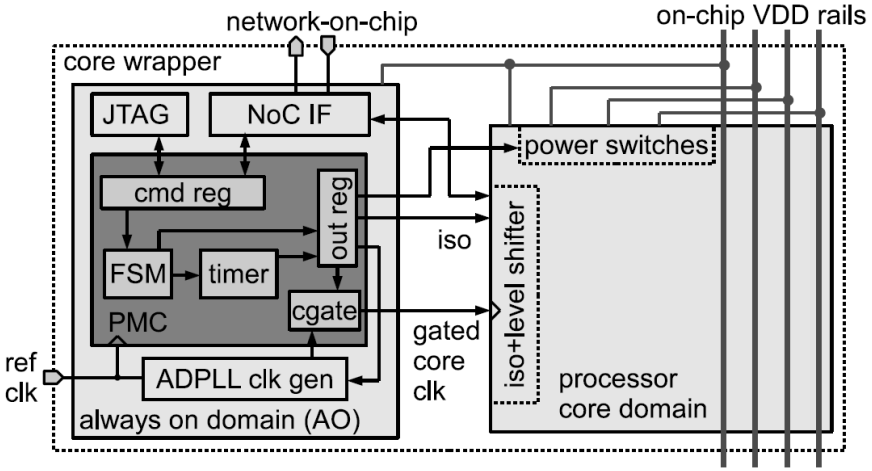
Power Switch Cell Beispiel

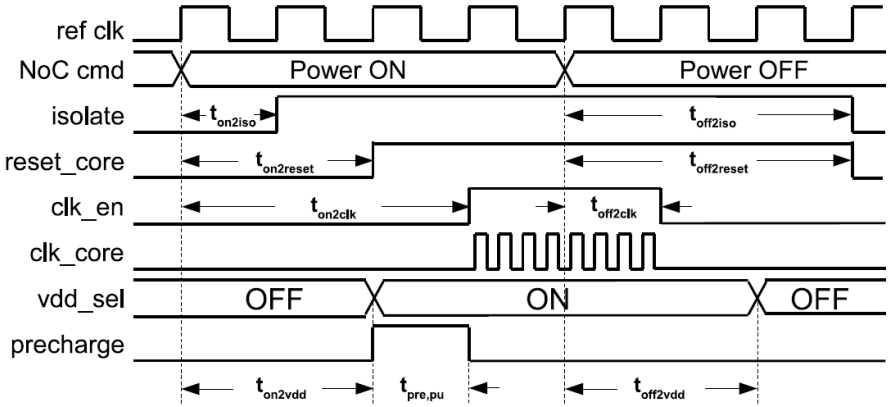
- Power-Switches werden räumlich verteilt in das Raster der Standardzellen eingefügt
- Typischerweise mehrere 100 bis 1000 Switch Zellen pro abschaltbarer Domäne
- Direkte Kontaktierung der Power-Rails der Standardzellen

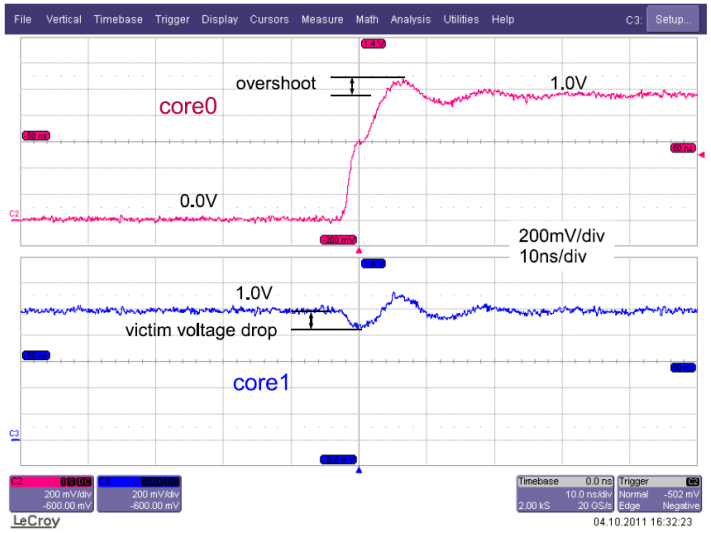


Beispiel: Power-Shut Off - Ablaufsteuerung

- 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,

Dynamic Frequency Scaling (DFS)

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

- 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

Dynamic Voltage and Frequency Scaling (DVFS)

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

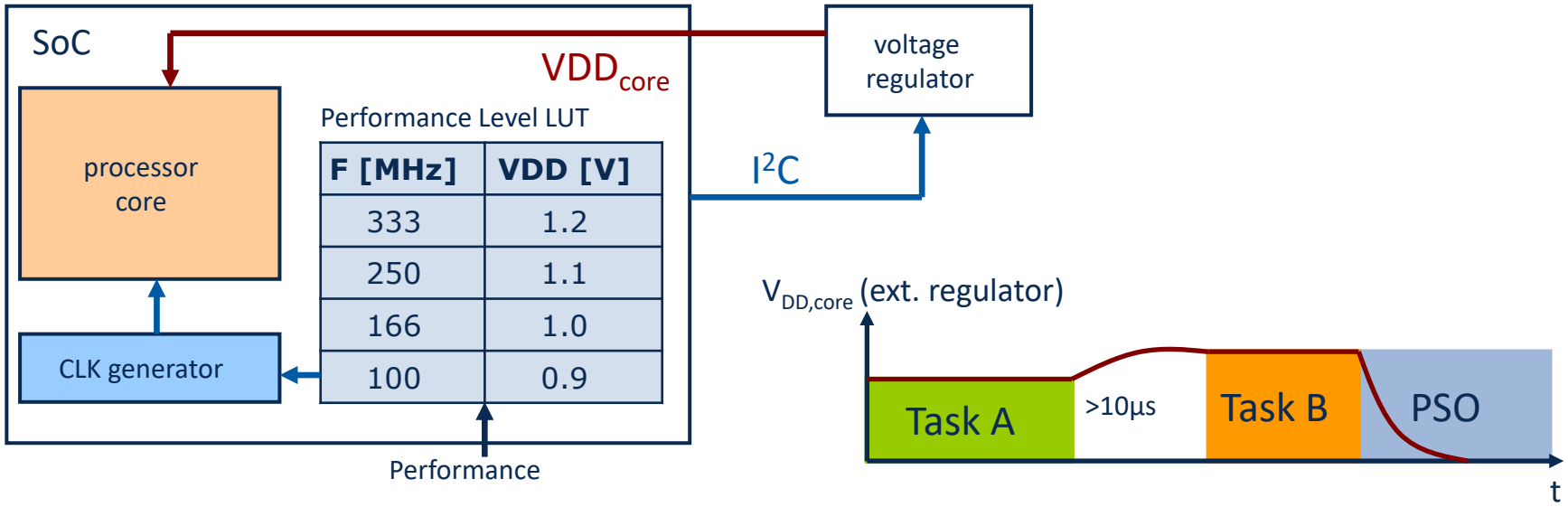
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

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

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

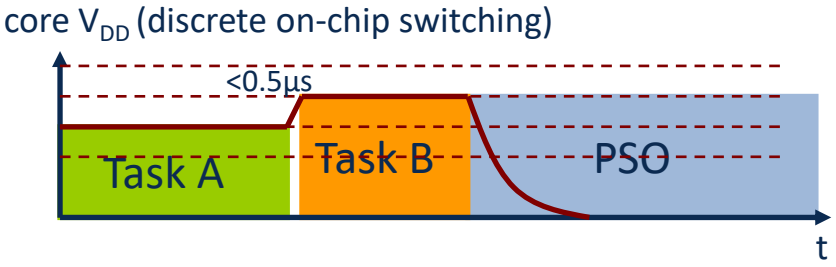
- **Nachteile:**
  - Benötigt programmierbaren Taktgenerator und Spannungsversorgung
  - Komplizierte Sign-Off Analysen mit mehreren nominalen Spannungen
  - Architekturanpassung erforderlich (Ablaufsteuerung)

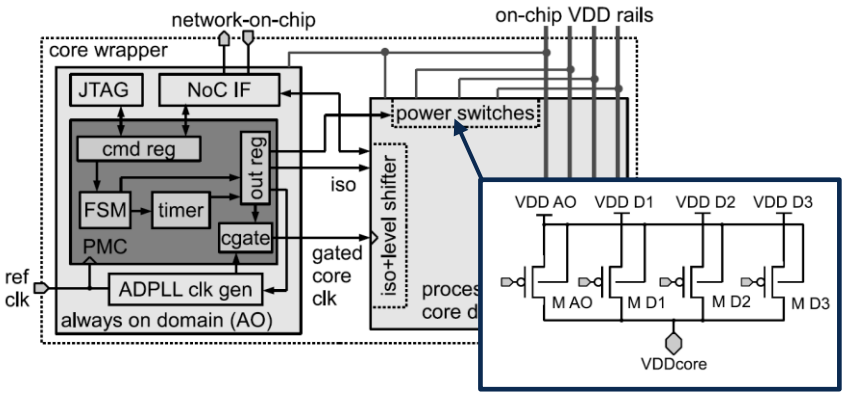
DVFS Architektur



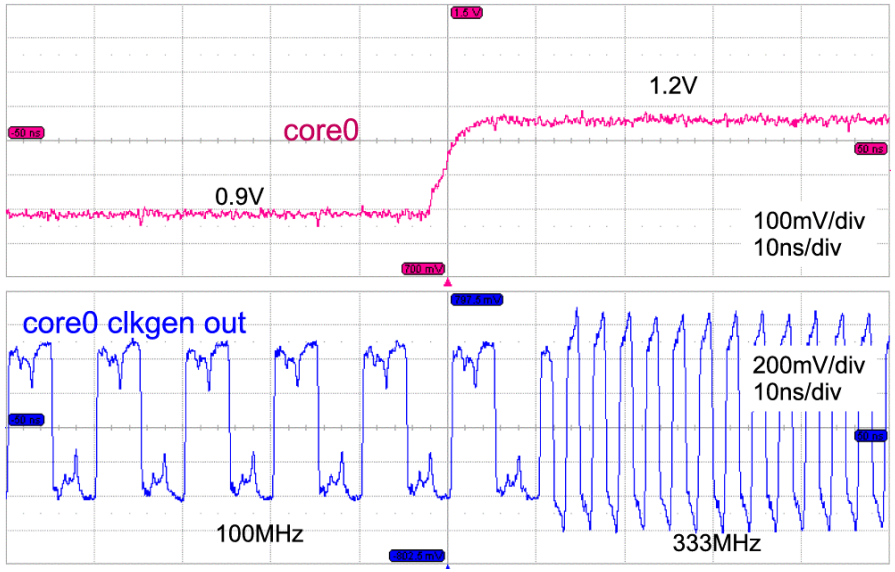
- Performance Level Lookup Tabelle
  - (V_{DD},f) Kombinationen
  - Festgelegt zur Entwurfszeit bzw. nach dem Test der Chips
  - Statisch im Betrieb
- Berücksichtigung des Spannungsbereichs bim Timing Sign-Off nötig (Hold Violations!)

Beispiel: Schnelles DVFS





- 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,

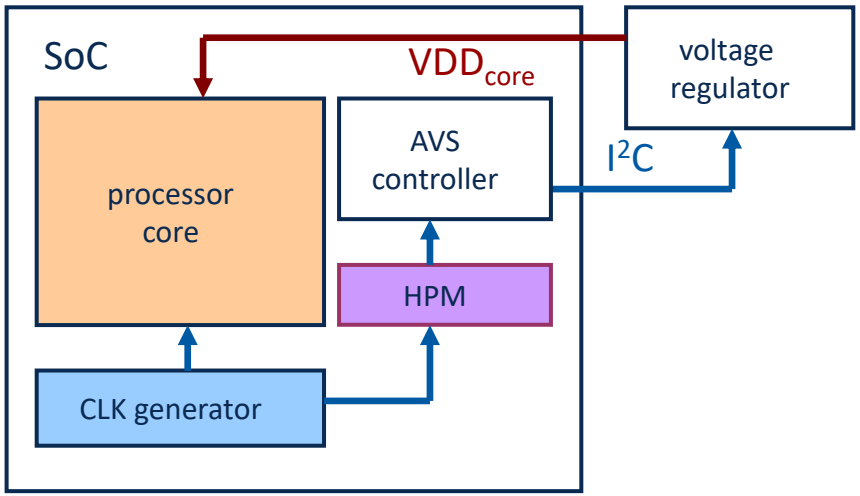
Adaptive Voltage and Frequency Scaling (AVFS)

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

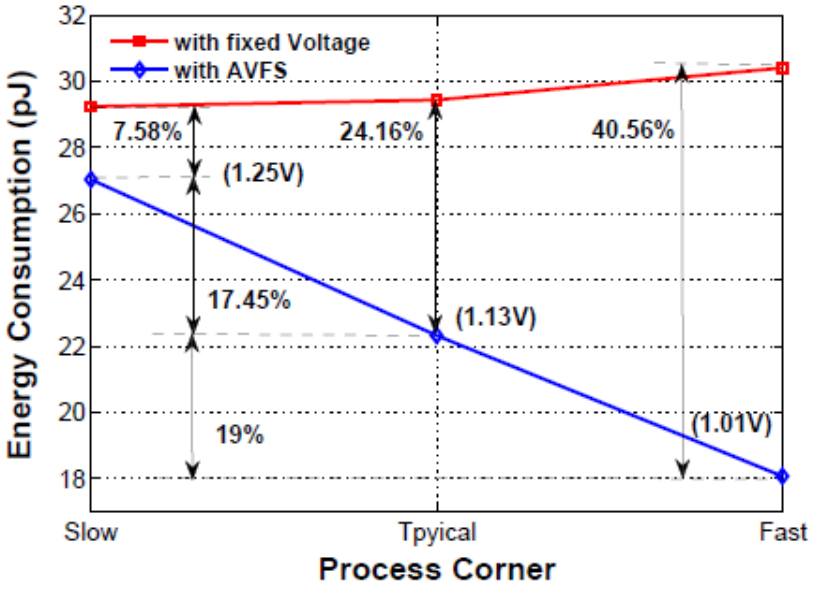
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

- Skalierung der Taktfrequenz gemäß der Performance Anforderung bei gleichzeitiger **autonomer, adaptiver Skalierung** der **Spannung VDD**
- Anpassung der Performanz (Performance Level) des Systems
  - Hohe Performance : f ↑→ V_{DD} ↑, E_{task} ↑
  - Geringe Performance : f ↓→ V_{DD} ↓, E_{task} ↓
- 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**



processor task energy consumption @fixed frequency



Kai Zheng, *Hardware Performance Monitor Design in Sub-100nm-Technologien*, Diplomarbeit, Oktober 2011, TU Dresden, Betreuer: H. Eisenreich, S. Höppner, und R. Schüffny

Vorteile und Nachteile von AVFS

- **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 Kalibierung (erhöht Kosten für Test)
  - Komplizierte Sign-Off Analysen mit mehreren nominalen Spannungen
  - Architekturanpassung erforderlich (Ablaufsteuerung) bei AVFS

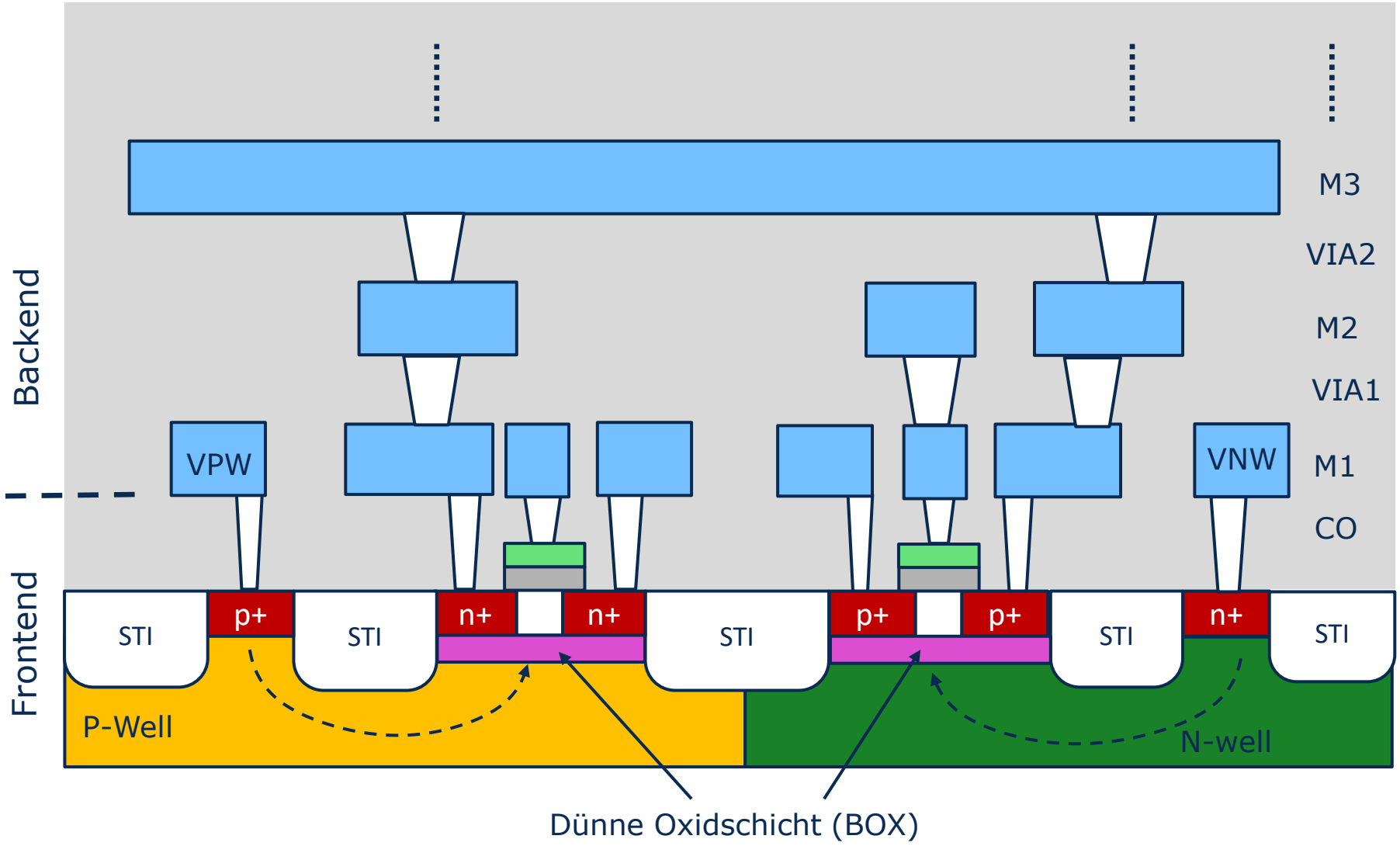
Adaptive Body Biasing (ABB)

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

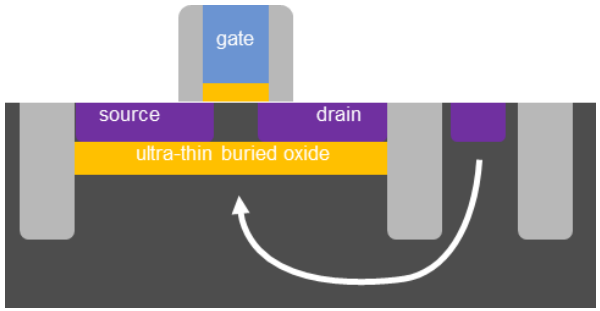
$$E_{task} = \alpha \cdot n \cdot V_{DD} \cdot Q_{short}(V_{DD}) + \alpha/2 \cdot n \cdot C_{eff} \cdot V_{DD}^2 + V_{DD} \cdot I_{leak}(V_{DD}) \cdot n \cdot T_{CLK}$$

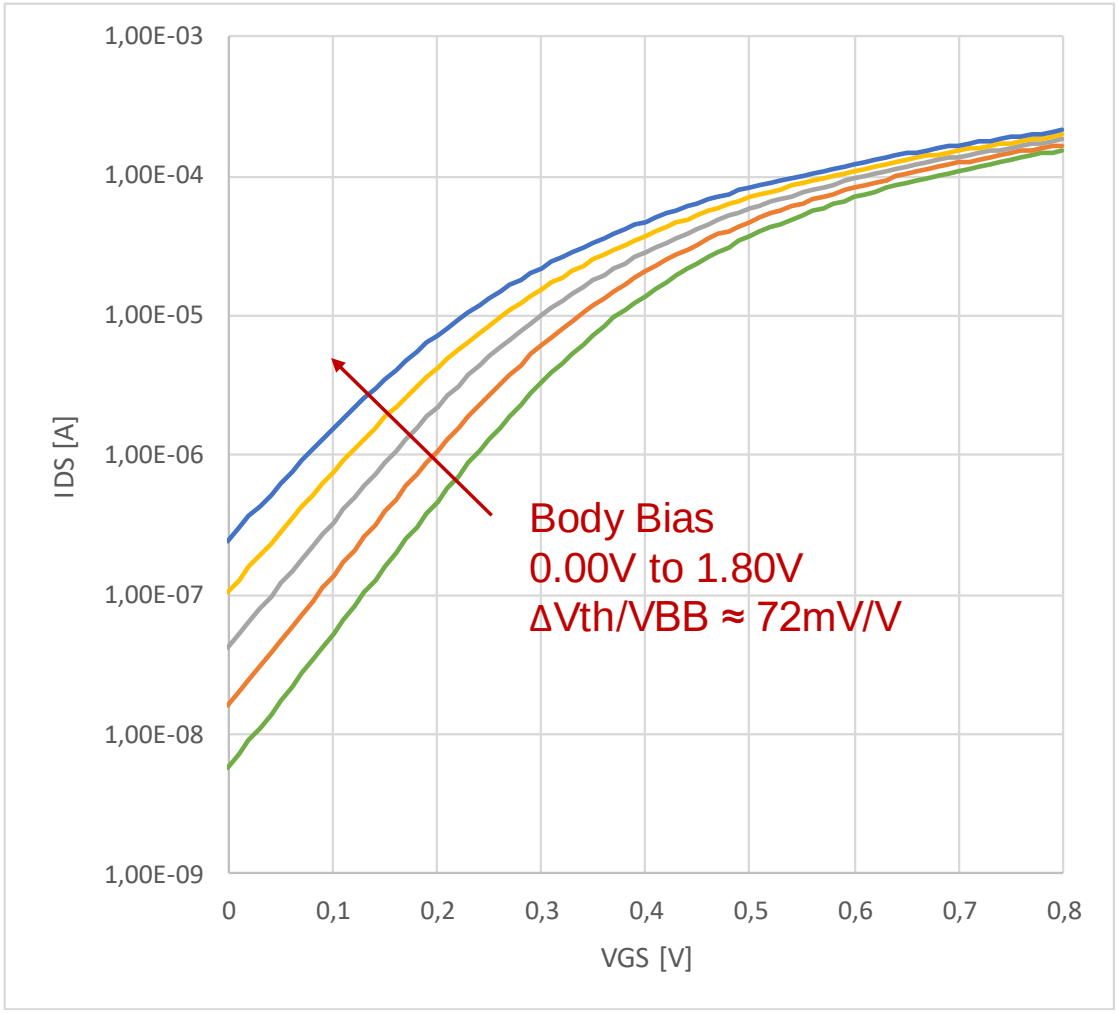
- Skalierung der Schwellspannung der Transistoren gemäß der Performance Anforderung durch **adaptive Skalierung** der Back-Gate **Spannungen (VNW, VPW) in FDOI Technologien**
- Adaptive Kompensation von
  - Prozessvariationen
  - Versorgungsspannungsvariationen (statisch, langsam)
  - Temperaturvariationen
- → Verbesserung der worst case performance und worst case Leckstrom
- → Erlaubt Betrieb bei kleineren Versorgungsspannungen
- → Reduziert die Nutzung von Standardzellen mit kleinerer Schwellspannung (Multi-Vt Implementierung)
- Benötigt Adaptive Body Bias Generator

CMOS Querschnitt (FDSOI Prozess)

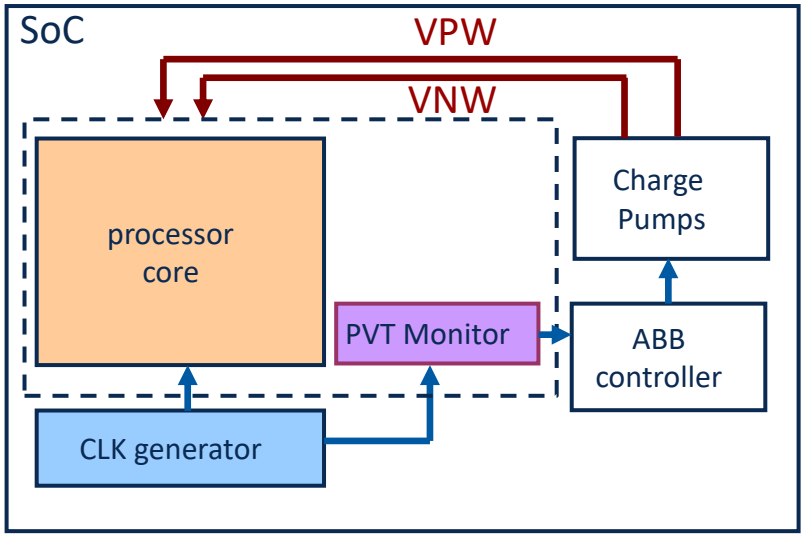


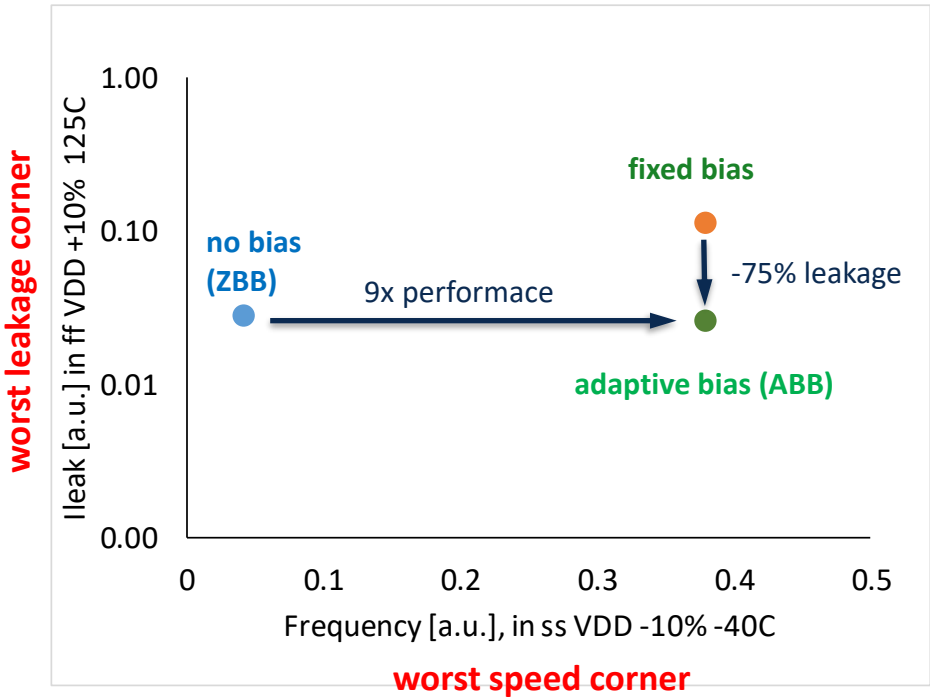
- Kontrolle der Transistorkennlinie über das Back-Gate (Body Bias)



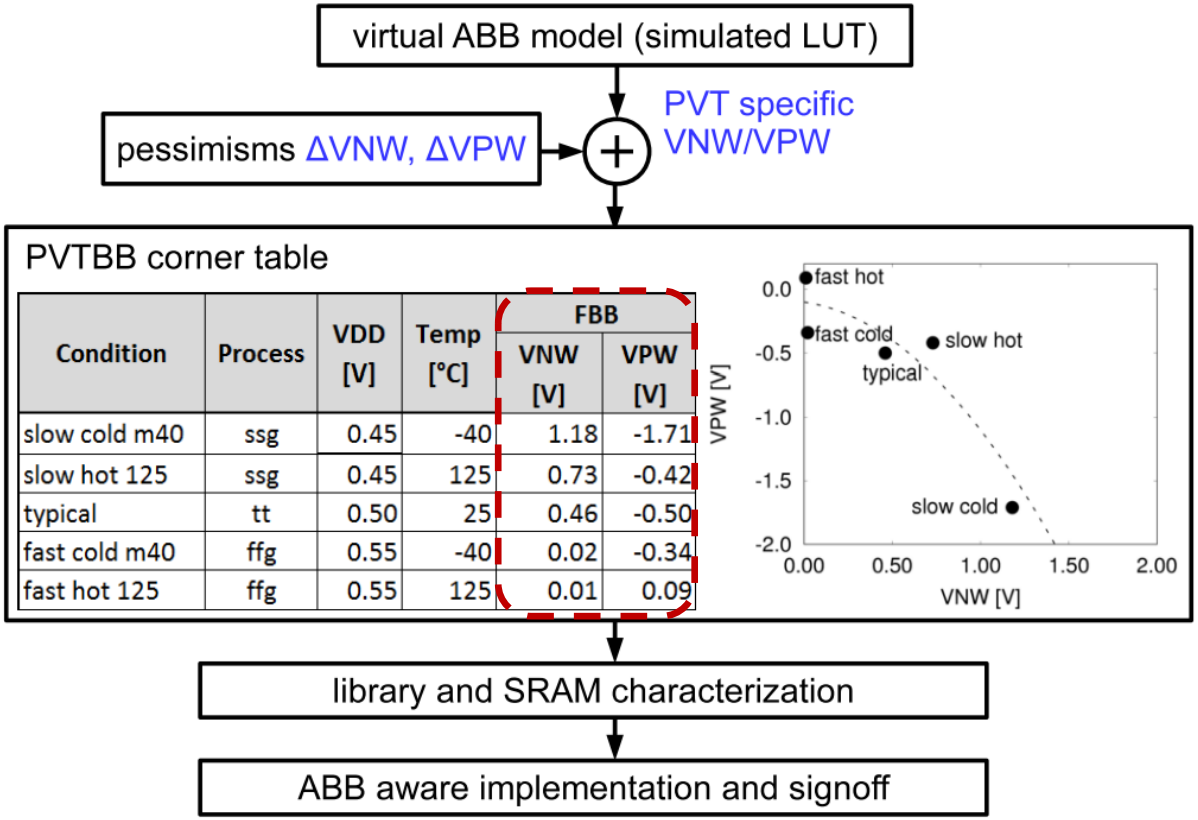


- Adaption der VNW und VPW Spannung basierend auf der Hardware-Performance
- Kritisches Timing abhängig von:
  - Prozessrealisierung: individuell für jeden Chip
  - Temperatur : ändert sich im Betrieb
  - Versorgungsspannung: ändert sich im Betrieb
  - Body-Bias Spannungen: **Adaption durch ABB**





- Berücksichtigung der Adaptiven VNW/VPW Spannungen im Design Flow
- PVTBB Corners für Synthese, P&R und Timing Analyse



Condition	Process	VDD [V]	Temp [°C]	FBB VNW [V]	FBB VPW [V]
slow cold m40	ssg	0.45	-40	1.18	-1.71
slow hot 125	ssg	0.45	125	0.73	-0.42
typical	tt	0.50	25	0.46	-0.50
fast cold m40	ffg	0.55	-40	0.02	-0.34
fast hot 125	ffg	0.55	125	0.01	0.09

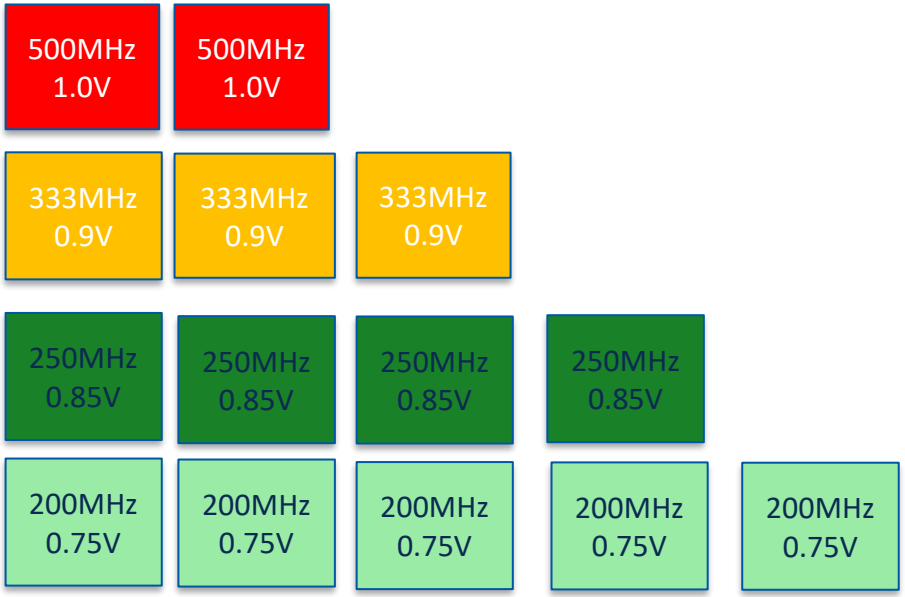
from: S. Höppner et al., "Adaptive Body Bias Aware Implementation for Ultra-Low-Voltage Designs in 22FDX Technology," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 10, pp. 2159-2163, Oct. 2020, doi: 10.1109/TCSII.2019.2959544.

Power Management auf Systemlevel

- 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**

Zusammenfassung

- 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