Leckstromreduzierung in Nanometer-Technologien ohne Performanceverluste

Frank Sill[*], Claas Cornelius[+], Dirk Timmermann[+]

[*] Department of Electrical Engineering, Universidade Federal de Minas Gerais (UFMG), Av. Antônio Carlos 6627, 31270-010 Belo Horizonte, Brasilien

[+] Institut für Angewandte Mikroelektronik und Datentechnik, Fakultät für Informatik und Elektrotechnik, Universität Rostock, Richard-Wagner-Str. 31, 18119 Rostock-Warnemünde, Deutschland

1. Einleitung

Die fortschreitende Verkleinerung der Strukturgrößen im Chipdesign ermöglicht immer größere Integrationsdichten, wodurch die kontinuierlich steigenden Performanceanforderungen erfüllt werden können. Gleichzeitig treten jedoch in den aktuellen Nanometer-Technologien Effekte in den Vordergrund, welche bisher vernachlässigt werden konnten. Dies zeigt sich insbesondere beim exponentiellen Anstieg der Leckströme innerhalb der vergangenen Technologie-Generationen. So verursachen diese Ströme in aktuellen Designs bis zu 50 % des gesamten Leistungsverbrauchs, wobei die Tendenz auch weiterhin steigend ist (Borkar 2005).

Ein verbreiteter Ansatz zur Leckstromreduzierung ist die Verwendung von Logikgattern, die sich in Berechnungszeit und Leckstrom unterscheiden (Wang et al. 1998, Sultania et al. 2004). Dabei werden die Logikgatter, welche eine geringe Verzögerungszeit aber einen hohen Leckstrom haben, in den zeitkritischen Abschnitten der Schaltung verwendet. Im Gegensatz dazu kommen in den unkritischen Abschnitten der Schaltung die Logikgatter mit geringem Leckstrom und hoher Verzögerungszeit zum Einsatz. Somit kann der Leckstrom bei konstanter Schaltungsperformance reduziert werden. Die unterschiedlichen Gattertypen werden mit Transistoren realisiert, die sich in der Schwellspannung („*dual threshold*“-CMOS-Ansatz) oder in der Gate-Oxiddicke T_{ox} („*dual* T_{ox}“-CMOS-Ansatz) unterscheiden. Zur Vereinfachung werden diese Techniken im Folgenden als DxCMOS-Ansätze zusammengefasst.

In der vorliegenden Arbeit wird eine Erweiterung dieser Ansätze vorgestellt, welche auf einer Umstrukturierung der Logikgatter beruht. Ferner wird ein Algorithmus zur Zuweisung der Gattertypen innerhalb der Schaltungen eingeführt. Mit Hilfe dieses neuen Ansatzes kann der Leckstrom der Schaltungen im Durchschnitt um 60 % verringert werden, wobei die

Performance konstant bleibt. Darüber hinaus wird im Vergleich zu den DxCMOS-Techniken eine zusätzliche Leckstromreduzierung von durchschnittlich 24 % erreicht.

2. Der „*Mixed Gates*“-Ansatz

Ein Hauptproblem der DxCMOS Ansätze besteht darin, dass sie entweder nur auf Gatter- oder nur auf Transistorebene arbeiten. So werden bei Ansätzen auf der Gatterebene nur Gatter verwendet, die jeweils aus einem einzigen Transistortyp bestehen. Dies schränkt die Freiheitsgrade bei der Optimierung erheblich ein. Wird hingegen ein Ansatz auf Transistorebene verwendet, ist ein immenser Anstieg des Rechenaufwands zu verzeichnen.

Die Idee des „*Mixed Gates*“-Ansatzes besteht darin, die Vorteile der Ansätze auf Gatter- und Transistorebene zu vereinen (Sill et al. 2007). Dabei werden beim „*Mixed Gates*“-Ansatz die Gatter aus zwei verschiedenen Transistortypen generiert. Im Vergleich zu den DxCMOS-Ansätzen auf ermöglicht dies eine größere Anzahl von unterschiedlichen Gattertypen. Gleichzeitig bleibt der Aufwand zur Reduzierung des Leckstroms geringer als bei einem Ansatz auf Transistorebene. Die verwendeten Transistoren unterscheiden sich sowohl in ihrer Schwellspannung V_{th} als auch in der Dicke der Oxidschicht T_{ox}. Daher können gleichzeitig die beiden Leckstromkomponenten mit dem größten Einfluss, der *subthreshold leakage current* und der *gate oxide leakage current*, reduziert werden.
Der „*Mixed Gates*“-Ansatz ermöglicht für jedes Gatter die Realisierung mit dem HVTO-, dem MG- und dem F-MG-Gattertyp. Dabei hat der HVTO-Gattertyp die längste Berechnungszeit t_d und den geringsten Leckstromverbrauch I_{leak}. Er besteht vollständig aus „*high-*V_{th}*/*T_{ox}“-Transistoren mit hohem V_{th} und dickem T_{ox} (siehe Abb. 1a). Der MG-Gattertyp verwendet vereinzelt „*low-*V_{th}*/*T_{ox}“-

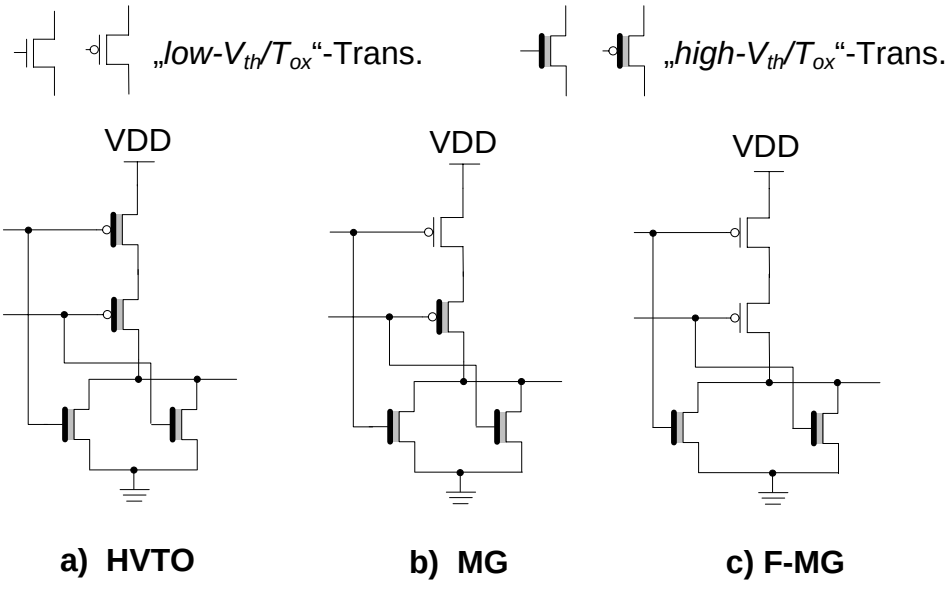

Abb. 1: NOR2 realisiert in der „*Mixed Gates*“-Technik

Transistoren mit geringem V_{th} und dünnem T_{ox} (siehe Abb. 1b). Dadurch kann die Verzögerungszeit für die ungünstigste Kombination der Eingangssignale, dem so genannten *worst case*, verringert werden. Hierbei ist anzumerken, dass bei der späteren Optimierung üblicherweise nur der *worst case* betrachtet wird. Der F-MG-Gattertyp hat schließlich die gleiche Berechnungszeit, wie ein Gattertyp, der nur aus „*low-V_{th}/T_{ox}*“-Transistoren besteht. Gleichzeitig ist jedoch der Leckstromverbrauch des F-MG-Typs niedriger. Im F-MG-Gattertyp besteht der Transistorpfad, der im *worst case* die Berechnungszeit des Gatters bestimmt, komplett aus „*low-V_{th}/T_{ox}*“-Transistoren. Die restlichen Pfade bestehen aus „*high-V_{th}/T_{ox}*“-Transistoren (siehe Abb. 1c).
Um den Leckstrom einer Schaltung zu reduzieren, werden die kritischen Pfade ermittelt. Hierbei handelt es sich um die Pfade von den Eingängen zu den Ausgängen der Schaltung, welche die längste Berechnungszeit t_{pfad_krit} benötigen. Da diese die Gesamtberechnungszeit der Schaltung bestimmen, werden alle Gatter der kritischen Pfade mit dem F-MG-Gattertyp, welcher die geringste Verzögerungszeit hat, realisiert. In den restlichen Pfaden werden die HVTO- bzw. MG-Gatter implementiert. Dies geschieht solange, bis die langsamen Pfade die gleiche Verzögerungszeit wie die kritischen Pfade haben, bzw. alle Gatter modifiziert wurden. Somit hat die resultierende Schaltung die gleiche maximale Verzögerungszeit und damit die gleiche Performance wie eine nichtoptimierte Schaltung, wobei jedoch der Leckstrom reduziert wird. Ein Beispiel für eine „*Mixed Gates*“-Schaltung ist in Abb. 2 dargestellt.

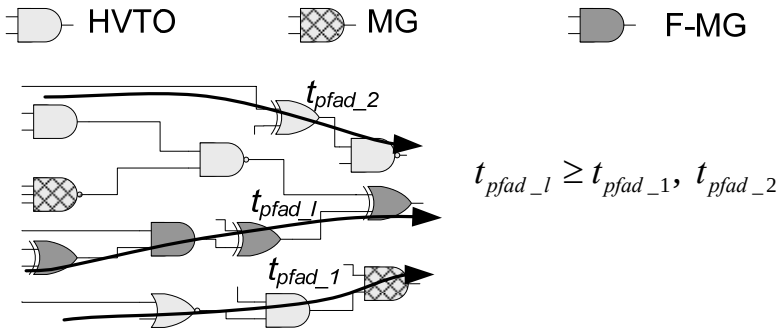

Abb. 2: Beispielrealisierung einer „Mixed Gates“-Schaltung. In den unkritischen Pfade sind HVTO- bzw. MG-Gatter implementiert, wobei die Verzögerungszeiten t_{pfad_n} dieser Pfade geringer sind als die größte Pfadverzögerungszeit t_{pfad_krit}.

3. Der Zuweisungsalgorithmus

Der Zuweisungsalgorithmus bestimmt, mit welchem Typ jedes Gatter innerhalb der Schaltung realisiert wird. Dabei ergibt sich die Effektivität des Algorithmus durch die Anzahl der Gatter, denen der HVTO bzw. MG-Gattertyp zugewiesen wird. Je höher diese Anzahl ist, umso größer ist die Reduzierung des Leckstroms gegenüber einer nichtoptimierten Schaltung. Um eine hohe Effektivität zu erreichen, basiert der Algorithmus auf einem prioritätenbasierten Ansatz. Dabei wird

jedem Gattertyp eines Gatters ein Bewertungsfaktor Ψ zugewiesen, der sich aus den Gattereigenschaften und der Position des Gatters innerhalb der Schaltung ergibt. Anschließend werden in Abhängigkeit dieses Bewertungsfaktors die Gattertypen bestimmt.
Der Bewertungsfaktor des verwendeten Zuweisungsalgorithmus setzt sich aus folgenden Parametern zusammen:

- t_{slack} – Zeit, um welche die Verzögerungszeit des Gatters erhöht werden kann, ohne dass die Schaltungsperformance verringert wird
- t_{d_diff} – Änderung der Gatterverzögerungszeit durch die Zuweisung des neuen Gattertyps
- I_{leak_diff} – Änderung des Gatterleckstroms durch die Zuweisung des neuen Gattertyps
- n_p – Anzahl der Pfade, in denen sich das Gatter befindet

Um eine Wichtung der Parameter untereinander zu vereinfachen, werden alle Werte relativ zu den maximal und minimal möglichen Werten innerhalb der Schaltung gesetzt. Dies führt zu den relativen Parametern t_{slackr}, $t_{d\_diff\_r}$, $I_{leak\_diff\_r}$ und n_{P_r}. Da alle Gatter der Schaltungen mit F-MG-Gattertyp initialisiert werden, bekommen diejenigen Gattertypen einen hohen Bewertungsfaktor, welche:

- sich in wenig Pfaden befinden (kleines n_P),
- eine hohe Leckstromdifferenz I_{leak_diff} vorweisen,
- eine geringe Verzögerungszeitdifferenz t_{d_diff} haben und
- einen großen t_{slack}-Wert haben.

Das Problem der relativen Parameter sind jedoch Extremwerte, denn verfügt die Schaltung über einzelne Gattertypen mit sehr großen bzw. sehr kleinen Parameterwerten, so unterscheiden sich die relativen Parameterwerte der restlichen Gattertypen kaum. Diese ungleichmäßige Verteilung der Werte ist in Abb. 3 exemplarisch für die beiden ISCAS-Benchmarkschaltungen c432 und c880 dargestellt (Hansen et al. 1999). Es werden die Differenzwerte I_{leak_diff} der Typen aller Gatter bei der Initialisierung der Schaltungen verglichen. Zur Vereinfachung wurden die Werte in Cluster unterteilt. Es ist ersichtlich, dass die relativen Parameterwerte der meisten Gattertypen sehr dicht beieinander liegen, während einzelne Parameterwerte sehr groß bzw. sehr klein sind. Somit unterscheiden sich auch die meisten Bewertungsfaktoren der einzelnen Gattertypen nur sehr wenig. Anderseits haben einige Gattertypen sehr große Bewertungsfaktoren. Dies resultiert in einer Reduzierung der Qualität der Ergebnisse des Algorithmus. Daher empfiehlt es sich, die einzelnen Parameter anzupassen, wobei jeder Parameter entsprechend einer Verteilungsfunktion erhöht wird. Die Verteilungsfunktion orientiert sich an der Häufigkeit, mit der innerhalb der Schaltung Gattertypen einen Parameter innerhalb eines Wertebereiches haben. Hierbei gilt, je größer die Anzahl der Gattertypen mit Parametern in einem bestimmten Wertebereich ist, desto stärker werden diese Parameter erhöht. Dadurch werden

die relativen Unterschiede der Parameter größer, während die Unterschiede zu den maximalen und minimalen Parametern geringer werden. Somit folgt für den Bewertungsfaktor Ψ:

$$\Psi = \kappa_{np}\left(1-\left[n_{P\_r}+n_{P\_add}\right]\right)+\kappa_{Ileak}\left[I_{leak\_diff\_r}+I_{leak\_diff\_add}\right]...$$
$$...+\kappa_{td}\left(1-\left[t_{d\_diff\_r}+t_{d\_diff\_add}\right]\right)+\kappa_{slack}\left(t_{slack\_r}+t_{slack\_add}\right)$$

$$\text{mit: } X_{\_r}=\frac{X_{P\_max}-X_P}{X_{P\_max}-X_{P\_min}}, X\in n_P, I_{leak\_diff}, t_{d\_diff}, t_{slack}$$

Hierbei sind n_{P_max} die maximale und n_{P_min} die minimale Anzahl von Pfaden, in denen sich ein Gatter der Schaltung befindet. $I_{leak\_diff\_max}$ und $I_{leak\_diff\_min}$ sind die maximal und minimal möglichen Differenzen des Leckstroms zweier Typen eines Gatters innerhalb der Schaltung. Weiterhin sind $t_{d\_diff\_max}$ und $t_{d\_diff\_min}$ die maximal und minimal möglichen Differenzen der Verzögerungszeit zweier Typen eines Gatters der Schaltung. Die Variablen κ_{np}, κ_{Ileak}, κ_{td} und κ_{slack} dienen zur Wichtung der einzelnen Parameter und müssen empirisch bestimmt werden. Ferner sind n_{P_add}, $I_{leak\_diff\_add}$, $t_{d\_diff\_add}$ und t_{slack_add} die jeweils zu addierenden Werte, welche sich aus Verteilungsfunktion der Parameterwerte ergeben.

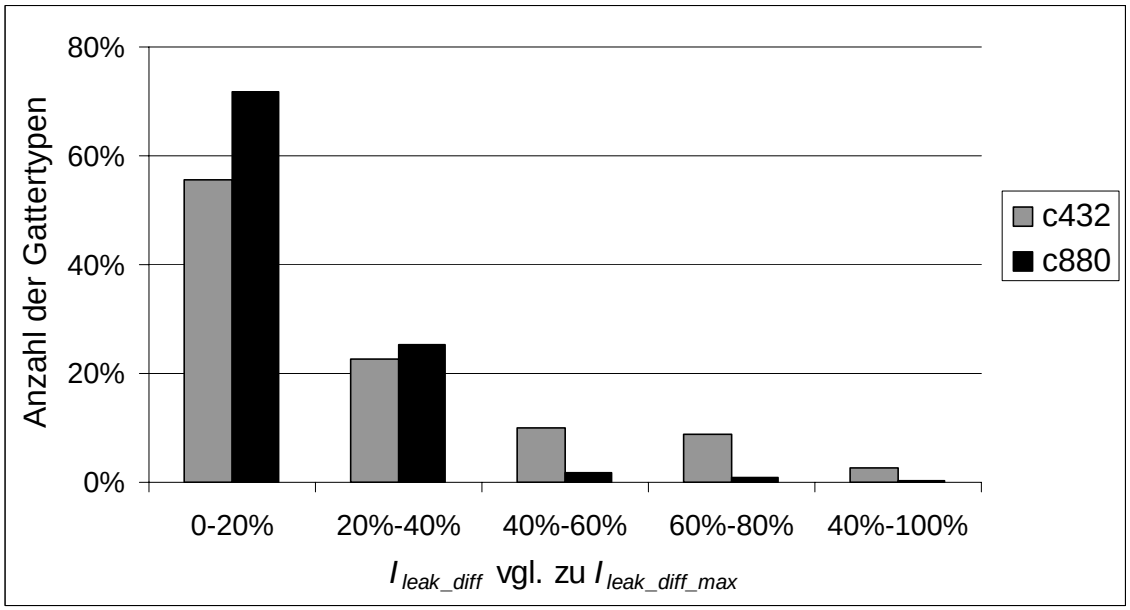

Abb. 3: Verteilung des Wertes I_{leak_diff} der Gattertypen aller modifizierbaren Gatter bei der Initialisierung der ISCAS-Schaltungen c432 und c880 mit LVTO-Gattertypen

4. Ergebnisse und Vergleich

Um den „*Mixed Gates*“-Ansatz zu verifizieren, wurde der vorgestellte Algorithmus implementiert und auf Basis einer modifizierten Technologie (Sill

et al. 2007) eine „*Mixed Gates*“-Gatterbibliothek generiert. Die Verifizierung erfolgte anhand der bekannten ISCAS-Benchmarkschaltungen (Hansen et al. 1999). Dabei wurde für jede Schaltung neben einer nichtoptimierten und der „*Mixed Gates*“-Version auch eine DxCMOS-Variante erstellt. Ferner galt als Bedingung, dass alle Implementierungen einer Schaltung die gleiche maximale Verzögerungszeit haben.
Aus den Simulationsergebnissen folgt, dass der Leckstrom der Schaltungen mit dem DxCMOS-Ansatz um durchschnittlich 46 % und mit dem „*Mixed Gates*“-Ansatz um durchschnittlich 60 % reduziert wird. Darüber hinaus zeigt der Vergleich des „*Mixed Gates*“-Ansatzes mit dem DxCMOS-Ansatz, dass der Leckstrom um bis zu 29 % zusätzlich reduziert werden kann (c1355). Im Durchschnitt ist der Leckstrom in den DxCMOS-Schaltungen um etwa 24 % größer, wobei der Unterschied mindestens 17 % beträgt (c2670). Somit wurde nachgewiesen, dass durch den „*Mixed Gates*“-Ansatz eine deutliche Reduzierung des Leckstroms gegenüber bekannten DxCMOS-Ansätzen erreicht werden kann.

5. Zusammenfassung

In der vorliegenden Arbeit wurde ein Ansatz zur Leckstromreduzierung vorgestellt. Dieser „*Mixed Gates*“-Ansatz basiert auf der Realisierung von Gattern mit unterschiedlichen Transistortypen. Ferner wurde ein prioritätenbasierter Zuweisungsalgorithmus präsentiert. Durch Simulationen konnte gezeigt werden, dass der „*Mixed Gates*“-Ansatz gegenüber nichtoptimierten Schaltungen eine durchschnittliche Leckstromreduzierung von 60 % ermöglicht. Zusätzlich wird im Vergleich zu bekannten Leckstromreduzierungstechniken eine Verbesserung von durchschnittlich 24 % erreicht.

6. Literatur

Borkar Y. S. (2005), VLSI Design Challenges for Gigascale Integration, keynote address at 18th Conference on VLSI Design, Kolkata, India.

Hansen M., Yalcin H., Hayes J. P. (1999), Unveiling the ISCAS-85 Benchmarks: A Case Study in Reverse Engineering, *IEEE Design and Test*, vol. 16, no. 3, S. 72-80, Juli-September,.

Sill F., Jiaxi Y., Timmermann D. (2007), Design of Mixed Gates for Leakage Reduction, Proc. 2007 Great Lakes Symposium on VLSI (GLSVLSI), S.263-268, Stresa-Lago Maggiore, Italien.

Sultania A. K., Sylvester D., Sapatnekar S. S. (2004), Tradeoffs between date oxide leakage and delay for dual Tox circuits, Proc. 41st Conference on Design Automation (DAC), San Diego, USA.

Wang Q., Vrudhula S. B. K. (1998), Static power optimization of deep submicron CMOS circuits for dual VT technology, Proc. Internat. Conference on Computer-Aided Design (ICCAD), S. 490-496, San Jose, USA.