

Ontwerp van een RRAM geheugen voor ingebedde NV toepassingen

Wouter Diels
Alexander Standaert

Thesis voorge dragen tot het behalen
van de graad van Master of Science
in de ingenieurswetenschappen:
elektrotechniek, optie Elektronica en
geïntegreerde schakelingen

Promotor:
Prof. dr. ir. W. Dehaene

Assessoren:
Prof. dr. ir. R. Lauwereins
Prof. dr. ir. M. Verhelst

Begeleiders:
ir. B. Baran
dr. ir. S. Cosemans

© Copyright KU Leuven

Zonder voorafgaande schriftelijke toestemming van zowel de promotor als de auteurs is overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie i.v.m. het overnemen en/of gebruik en/of realisatie van gedeelten uit deze publicatie, wend u tot ESAT, Kasteelpark Arenberg 10 postbus 2440, B-3001 Heverlee, +32-16-321130 of via e-mail info@esat.kuleuven.be.

Voorafgaande schriftelijke toestemming van de promotor is eveneens vereist voor het aanwenden van de in deze masterproef beschreven (originele) methoden, producten, schakelingen en programma's voor industrieel of commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Voorwoord

Na bijna een jaar aan deze thesis gewerkt te hebben zouden we graag een aantal mensen in het bijzonder willen bedanken voor hun hulp en steun. Ten eerste prof. Dehaene voor het onderwerp aan te bieden en toe te wijzen aan ons. Ten tweede onze begeleiders Stefan en Burak.

Stefan, hartelijk dank voor de tijd die je wou vrij maken voor ons, het was heel aangenaam om een begeleider te hebben die zoveel kennis en ervaring heeft van het vakgebied.

Burak, thanks for always being available for us and for providing us the resources to get our thesis underway.

Verder zouden we graag Bert DeKnuydt willen bedanken voor zijn hulp met het Condor systeem. Tenslotte zouden we ook elkaar willen bedanken, we waren een goed complementair team.

*Wouter Diels
Alexander Standaert*

Inhoudsopgave

Voorwoord	i
Samenvatting	iv
Lijst van figuren en tabellen	v
Lijst van afkortingen en symbolen	viii
1 Inleiding	1
1.1 Doel en afbakening van dit werk	2
1.2 Structuur van de tekst	2
2 Geheugencel	5
2.1 Memristor	5
2.2 Memristorconfiguraties	7
2.3 Besluit	10
3 Geheugenarchitectuur	13
3.1 Cel	13
3.2 Branch	13
3.3 Local Block	14
3.4 Global Block	16
3.5 Besluit	17
4 Lastimpedantie-analyse	19
4.1 Algemene lasteigenschappen en -specificaties	19
4.2 Evalueren van de last	20
4.3 Vergelijking van verschillende types last	22
4.4 Besluit	28
5 Sense Amplifier analyse	31
5.1 Types SA	31
5.2 Offsetspanning	33
5.3 Sensitiviteitsanalyse	33
5.4 Paretosimulatie	40
5.5 Besluit	43
6 Omringende logica	45
6.1 Decoders	45
6.2 Buffers	49

6.3	BL en WL drivers	51
6.4	Passgates	51
6.5	Besluit	53
7	Timing en optimalisatie	55
7.1	Timing	55
7.2	Analyse verschillende geheugenconfiguraties	61
7.3	Besluit	63
8	Volledige ontwerp	67
8.1	Het finaal ontwerp	67
8.2	Vergelijking met de literatuur	70
8.3	Besluit	70
9	Besluit	71
A	Leon Chua's memristortheorie	75
B	Ladingsinjectie bij het gebruik van ideale SPICE bronnen.	77
C	IEEE Paper	79
	Bibliografie	85

Samenvatting

Nu het schalen van flash-geheugens op zijn limieten begint te stoten, is er nood aan een alternatief. Met eigenschappen zoals lage voedingsspanning, kleine geheugencel en hoge leessnelheid is RRAM één van de meest belovende kandidaten.

Deze thesis behandelt het ontwerp van het leescircuit van een 1Mbit RRAM geheugen. De grootste uitdaging in het ontwerp van dit resistief geheugen is variabiliteit. Deze variabiliteit is kritisch op twee plaatsen: ten eerste bij de resistieve deling tussen de geheugencel en een lastimpedantie. Ten tweede bij het latchen van de sense amplifier.

Dit werk introduceert drie innovaties. Een grondige analysemethode voor de optimale keuze van de lastimpedantie werd uitgewerkt en toegepast. Een referentieschema met parallelle geheugencellen resulteert in en betere referentiespanningsdistributie. Tenslotte wordt een analyse over de invloed van een overlappende werking van pass-gates en sense amplifier onder variabiliteit aangebracht. Verder wordt ook het ontwerp van alle andere belangrijke bouwblokken zoals decoders en buffers besproken.

Deze kennis wordt dan gebruikt in het ontwerp van een 1Mbit geheugen in 45nm PTM technologie. Het geheugen maakt gebruik van 512 sense amplifiers die telkens gekoppeld zijn aan 2 geheugenmatrices met 32WL en 32BL. Op een voedingsspanning van 1V heeft het geheugen een random-access-leessnelheid van 435MHz. Het energieverbruik per bitleesoperatie bedraagt bij deze voedingsspanning 0.51pJ. Binnen de afbakening van dit werk presteert de ontworpen schakeling beter dan flash-geheugens gevonden in de literatuur.

Lijst van figuren en tabellen

Lijst van figuren

2.1 Resistieve schakeling	6
2.2 MIM structuur	6
2.3 Model van het Pt-TiO ₂ -Pt staal	6
2.4 Forming,resetting en setting van een memristor	7
2.5 DC-analyse bipolaire/unipolaire memristor	8
2.6 Een 1T1R-configuratie	8
2.7 Read sneak leakage path	9
2.8 Half select problem	10
2.9 1T1R matrix	11
3.1 Een geheugencel en een branch	14
3.2 Een local block	15
3.3 Een local block	15
3.4 Datasignaal uitlezen	16
3.5 Referentiesignaal uitlezen	16
3.6 Een global block	17
4.1 Verschil in bitlinespanning in functie van lastweerstand	20
4.2 Testbench voor de lastimpedantie	21
4.3 Types lastimpedanties	22
4.4 Lineaire sweep van switchload	25
4.5 Lineaire sweep van biasload	25
4.6 Lineaire sweep van diodeload	26
4.7 Lineaire sweep van bulkload	26
4.8 BL-spanningsverdeling voor een biasload	27
4.9 Bitlinespanning referentiecellen	27
4.10 Sweep switchload over transistor lengte en breedte	28
4.11 BL-spanningsverdeling voor de finale lastimpedantie	29
5.1 Voltage mode sense amplifiers	32
5.2 Current mode sense amplifiers	32
5.3 De drain-input latch-type SA	33
5.4 SA offsetspanning	33

5.5	Sensitiviteitsresultaten: offsetspanning i.f.v. mismatchvariabelen	35
5.6	Foutief latchen door β -mismatch	36
5.7	Transiënte simulatie zonder en met overlap	37
5.8	Simulatieopstelling voor het RC-latch-effect	37
5.9	Simulatieresultaten voor het RC-latch-effect	38
5.10	Invloed capaciteit op RC-latch effect	38
5.11	Circuit voor analyse voorwaarden RC-latch-effect	39
5.12	De pareto-optimale sense amplifiers	42
6.1	Types decoders	45
6.2	Basis decoders	46
6.3	Vergelijking van decoder types	48
6.4	Glitch in NOR-gate	49
6.5	Gebufferde en ongebufferde signalen naar de referentie logica	50
6.6	BL- en WL-drivers	51
6.7	BL- en WL-drivers	52
6.8	nMOS passgate	52
6.9	pMOS passgate	53
6.10	Dode zones voor verschillende types passgates	54
7.1	Timingproblemen bij de bitline	56
7.2	Global block:logica	57
7.3	Global block:timing	57
7.4	Data-array:logica	58
7.5	Data-array:timing	59
7.6	Delay van WL-decoders en -buffers i.f.v. BL-decoders	59
7.7	Referentie-array:logica	60
7.8	Referentie-array:timing	60
7.9	SA:logica	61
7.10	SA:timing	61
7.11	Leescyclus	62
7.12	Delay, energieverbruik en oppervlakte van alle geheugenconfiguraties . .	63
7.13	Delay, energieverbruik en oppervlakte van alle geheugenconfiguraties . .	65
8.1	Resultaten speed-vdd test	68
8.2	Bl-spanningen i.f.v. Vdd	69
B.1	Ladingsinjectie: testcircuit	78
B.2	Ladingsinjectie: stroom	78

Lijst van tabellen

1.1	Technologieparameters	2
5.1	Sensitiviteitsanalyse van de minimale SA	36

LIJST VAN FIGUREN EN TABELLEN

5.2	Sensitiviteitsanalyse van de minimale SA met overlap tussen passenable en latchenable	40
5.3	Sensitiviteitsanalyse van de SA in het finale geheugen	41
6.1	Aantal gates in de grid decoder	47
6.2	Lasten aangedreven door de verschillende buffers	50
8.1	Transistor afmetingen eind ontwerp	68

Lijst van afkortingen en symbolen

Afkortingen

BL	Bit Line
CDF	Cumulative Distribution Function
GB	Global Block
HRS	High Resistive State
LB	Local Block
LRS	Low Resistive State
MTJ	Magnetic Tunnel Junction
NoBLpLB	Number of Bit Lines per Local Block
NoGB	Number of Global Blocks
NoWLpB	Number of Word Lines per Branch
PDF	Probability Density Function
PTM	Predictive Technology Model
RAM	Random Access Memory
RRAM	Resistive Random Access Memory
SA	Sense Amplifier
SL	Source Line
WL	Word Line

Hoofdstuk 1

Inleiding

Vandaag de dag is elektronica niet meer uit het leven weg te denken. Van de smartphone tot het digitaal horloge, van de bordcomputer in de moderne wagen tot de microprocessor in de vaatwasser, overal vind je wel elektronica terug. Sinds Gordon Moore ongeveer 50 jaar geleden de uitspraak deed dat het aantal transistoren op eenzelfde oppervlakte per twee jaar zou verdubbelen [18], is de industrie er over het algemeen goed in geslaagd dit te verwezenlijken. Dit leidde tot de snelle en uiterst complexe chips die we vandaag allemaal goedkoop aankopen.

Naarmate de processorkracht groter werd, steeg ook de vraag voor grotere en snellere geheugens om deze processorkracht ook effectief uit te buiten. Static Random Access Memory (SRAM) blijft een populaire keuze voor snelle ingebedde geheugens, maar heeft het nadeel vluchtig te zijn: eenmaal de voedingsspanning wordt afgeschakeld, verdwijnt de informatie. Flash-geheugens, door veel mensen gebruikt voor massa-opslag in USB-sticks of SSDs, hebben ook hun weg gevonden naar het ingebedde domein en behoren wel tot de klasse van niet-vluchige geheugens. Het blijkt echter bijzonder moeilijk om flash-geheugens verder te verkleinen [19].

Onderzoek naar nieuwe geheugens is dan ook onontbeerlijk. Zo zijn er al nieuwe kandidaten in opmars die hoopgevend zijn om te concurreren met (ingebedde) flash-geheugens. MRAMs (Magnetic RAMs) en in het bijzonder STT-RAM (Spin-Transfer Torque) zullen op termijn een belangrijke rol gaan spelen.

Een andere kandidaat is Resistive RAM (RRAM of ReRAM). Daar waar SRAM-en flash-cellen informatie bevatten via het al dan niet aanwezig zijn van lading, bevat een RRAM-cel informatie door een bepaalde elektrische weerstandswaarde aan te nemen. RRAM zou geen problemen hebben om nog even op de klassieke manier mee te schalen en is dus zonder meer een interessante piste om verder te onderzoeken. Bovendien zou RRAM gefabriceerd kunnen worden met goedkopere processen dan flash-geheugens: bij flash-geheugenfabricatie zijn vaak dure extra maskers vereist terwijl RRAM-fabricatie geïntegreerd kan worden met een standaard CMOS-productieproces.

1. INLEIDING

$A_{\beta n}$	2 % μm	β -Pelgrom constante voor nMOS-transistoren
$A_{\beta p}$	1,2 % μm	β -Pelgrom constante voor pMOS-transistoren
AV_{Tn}	2,82 mV μm	VT-Pelgrom constante voor nMOS-transistoren
AV_{Tp}	2,5 mV μm	VT-Pelgrom constante voor pMOS-transistoren
C_{WL}	0,18 fF/cel	WL-capaciteit, stijgt lineair met het aantal cellen eraan
C_{BL}	0,18 fF/cel	BL-capaciteit, stijgt lineair met het aantal cellen eraan
C_{inv}	0,35 fF	intrinsieke capaciteit van een CMOS inverter
μ_{HRS}	32500 Ω	verwachtingswaarde van een HRS geheugenelement
μ_{LRS}	7500 Ω	verwachtingswaarde van een LRS geheugenelement
σ_{HRS}	833 Ω	standaarddeviatie van een HRS geheugenelement
σ_{LRS}	833 Ω	standaarddeviatie van een LRS geheugenelement
V_{DD}	1 V	voedingsspanning
V_{SS}	0 V	grondspanning

Tabel 1.1: Numerieke technologieparameters waarvan gebruik gemaakt is in simulaties

1.1 Doel en afbakening van dit werk

Dit werk beschrijft het ontwerp van een 1Mbit RRAM-geheugen voor ingebetde toepassingen. De doelstelling is een pareto-optimaal circuit te ontwerpen. De pareto-doelstellingen zijn snelheid, dynamische energie en oppervlakte. Het ontwerp is ook gewapend tegen variabiliteit d.w.z. ongecorreleerde gedragsvariaties van componenten. De analyse focust op de leesbewerking, de schrijfbewerking valt buiten het bereik van dit werk. Er worden wel mogelijke oplossingen aangereikt, maar deze werden niet uitdrukkelijk onderzocht. Bij het ontwerp is ook aandacht besteed aan het vermijden van destructieve leescycli.

Voor de leesbewerking wordt het geheugenelement gemodelleerd als een weerstand waarvan de weerstandswaarde afhangt van de celtoestand. Om variabiliteit te onderzoeken, worden Monte Carlo simulaties uitgevoerd waarbij de weerstandswaarde een Gaussisch verdeelde variabele is.

Temperatuursvariaties werden niet systematisch onderzocht, maar aangezien temperatuur een globale variabele is en het systeem differentieel werkt, wordt niet verwacht dat de performantie aanzienlijk zal verminderen.

Alle analyses in dit werk zijn uitgevoerd met Spectre simulaties met 45nm PTM transistormodellen. In tabel 1.1 zijn technologieparameters te zien, waarvan meermaals gebruik wordt doorheen dit werk.

1.2 Structuur van de tekst

In hoofdstuk 2 wordt de technologie van een RRAM geheugen uiteengezet, alsook diens toepassingen. In hoofdstuk 3 wordt het geheugensysteem vanuit vogelperspectief besproken. Er wordt hier ook aangehaald wat de regelbare parameters zijn van de architectuur. Voor een robuuste, snelle en laag-energetische leesoperatie is het belangrijk het geheugenelement te combineren met een zorgvuldig gekozen impedantie,

dit wordt onderzocht in hoofdstuk 4. Uiteindelijk worden bits afgeleverd aan de uitgang van het systeem: de sense amplifier zorgt hiervoor en wordt besproken in hoofdstuk 5. In de geheugenstructuur worden ook bouwblokken zoals decoders, buffers en passgates gebruikt om op basis van het opgegeven adres de juiste cel aan te spreken; deze worden beschreven en geanalyseerd in hoofdstuk 6. In hoofdstuk 7 wordt de timing van controlesignalen onderzocht alsook de optimalisatie van het systeem door de architectuurparameters te tunen. Tenslotte wordt een overzicht gegeven van de resultaten van het volledige ontwerp in hoofdstuk 8.

Hoofdstuk 2

Geheugencel

De meeste geheugenschakelingen bestaan uit een verzameling individuele cellen die op een bepaalde manier informatie bevatten. In dit hoofdstuk wordt dieper ingegaan op de manier waarop een RRAM geheugencel informatie bevat.

2.1 Memristor

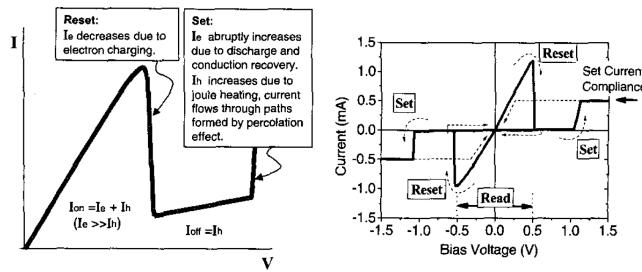
Het essentiële element van de RRAM geheugencel is de zogenaamde memristor. Volgens de originele memristortheorie (zie bijlage A) is dit de 4^e passieve component, naast de weerstand, spoel en condensator. De resistief schakelende elementen die in de praktijk gebruikt worden stroken echter niet met deze theorie, maar zullen in wat volgt toch memristors genoemd worden.

2.1.1 Fysische memristors

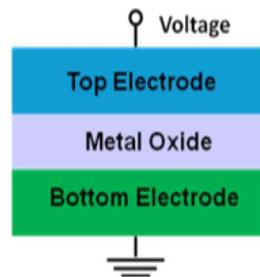
Er werd reeds langer (zelfs al sinds de jaren 60) opgemerkt dat sommige metaaloxides, die normaal gezien als elektrisch isolator functioneren, een plotselinge overgang kunnen vertonen naar een veel hogere staat van geleiding (zie figuur 2.1). Dit gebeurt veelal in een configuratie waarbij het oxide wordt geplaatst tussen 2 metalen (MIM configuratie)[29] (zie ook figuur 2.2). In die vroege jaren werd er reeds gesuggereerd dat deze structuren gebruikt konden worden voor geheugentoepassingen[24], maar de elementen bleken niet stabiel genoeg voor circuitimplementatie. Bovendien waren de siliciumgebaseerde geheugens in opmars, waardoor er geen nood was voor verder onderzoek.

In het begin van het nieuwe millennium wakkerde de interesse voor resistief schakelende elementen weer aan door enkele nieuwe onderzoekspublicaties zoals [2] en [32], waarin veel stabielere resistieve elementen werden gepresenteerd. Zhuang et al. brachten ook de term RRAM aan in hun artikel. In 2008 publiceerde een onderzoeksgroep van Hewlett-Packard een artikel waarin ze opmerkten dat het gedrag van hun Pt-TiO₂-Pt stalen een merkwaardige gelijkenis vertoonde met Chua's originele memristortheorie[26]. Uit de modellering van hun stalen argumenteerden ze dat dit een ideale memristor zou zijn en dat het effect meer uitgesproken is bij

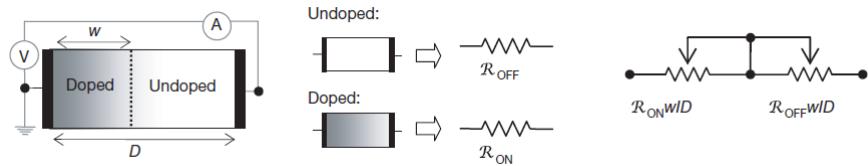
2. GEHEUGENCEL



Figuur 2.1: Overgang van hoge weerstand naar lage weerstand voor NiO bij DC analyse (unipolaire memristor), reproduced from[1]



Figuur 2.2: Metal-Insulator-Metal structuur, reproduced from[29]

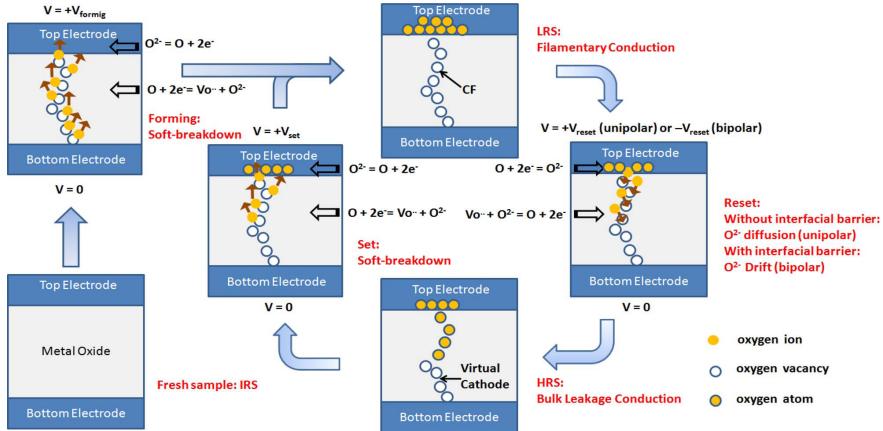


Figuur 2.3: Model van het Pt-TiO₂-Pt staal, reproduced from[26]

kleine afmetingen. Het titaniumoxide bestaat o.w.v. niet-idealiteiten uit 2 delen: zuiver TiO₂, een halfgeleider met hoge weerstand, en TiO_{2-x} met zuurstofafwezigheid (oxygen vacancies) met een veel lagere weerstand. Door een elektrisch veld aan te leggen worden zuurstofatomen weg of in het rooster getrokken en verandert de verhouding TiO₂ en TiO_{2-x} en dus ook de netto weerstand (zie figuur 2.3).

Voor deze opstelling geldt dan dat $M(q) = R_{off}(1 - \frac{\mu_v R_{on}}{D^2} q(t))$ met M de *memristance*, D de totale dikte van de titaniumoxidefilm, μ_v de mobiliteit van de zuurstofionen en $R_{on} \leq M \leq R_{off}$. Het dynamisch gedrag van de ogenblikkelijke weerstandswaarde is dus afhankelijk van het verloop van de stroom in de tijd en dit effect treedt des te meer op in het nanometerdomein.

Er zijn nog talloze andere materialen die schakelend weerstandsverdrag vertonen zoals nikkeloxide[1], hafniumoxide[5], aluminiumoxide[13],... Niet altijd kunnen de resultaten gemodelleerd worden volgens de originele memristortheorie, maar



Figuur 2.4: Illustratie van forming,resetting en setting, reproduced from[29]

desalnietemin zullen deze materiaalconfiguraties bruikbaar zijn in toepassingen en zullen ze in de rest van dit werk memristoren genoemd worden. Bij al deze MIM-configuraties blijft het mechanisme wel hetzelfde: na fabricatie is het oxidekristal intrinsiek zuiver, maar onder druk van een voldoende groot elektrisch veld zullen de zuurstofatomen losgerukt worden uit het rooster naar de anode. Het gebrek aan zuurstofatomen zorgt voor conductieve filamenten. Het element bereikt dan een laagresistieve staat (LRS). Deze zachte doorslag van het zuivere oxide wordt *forming* genoemd. Het proces is tot zekere hoogte omkeerbaar (*reset*), maar er zullen altijd meer defecten in het kristal zijn dan voor de forming. Dit betekent dus ook dat nadat de memristor één keer een forming- en resetproces is ondergaan en zich terug in een hogeresistieve staat (HRS) bevindt, er hierna een minder groot elektrisch veld nodig is om terug tot een LRS te komen. Dit proces heet *setting*. Deze drie processen zijn geïllustreerd op figuur 2.4.

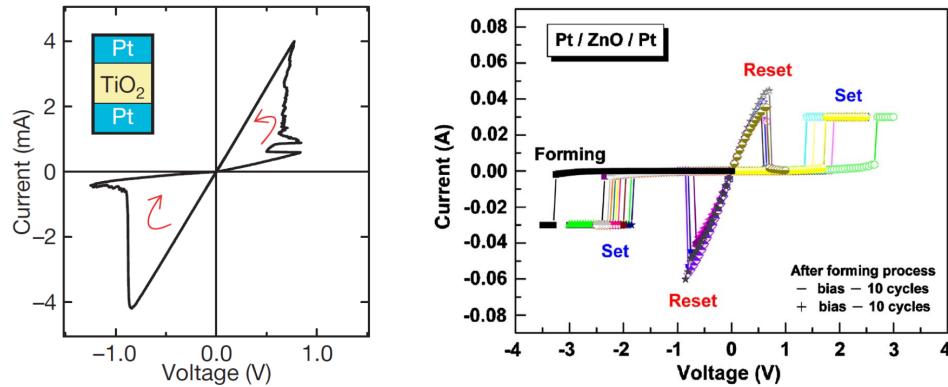
MIM-structuren gefabriceerd uit verschillende materialen hebben ook verschillende eigenschappen. Zo moet er onderscheid gemaakt worden tussen unipolaire schakelen en bipolaire. Bij bipolaire resistieve schakelen zal forming/setting optreden wanneer de aangelegde spanning een bepaalde polariteit heeft en resetting bij de omgekeerde polariteit. Bij unipolaire schakelen is de amplitude of de duur van de spanning doorslaggevend voor welke van de 3 processen zal optreden, niet de polariteit. DC-analyses voor bipolare en unipolaire memristors staan afgebeeld op figuur 2.5.

2.2 Memristorconfiguraties

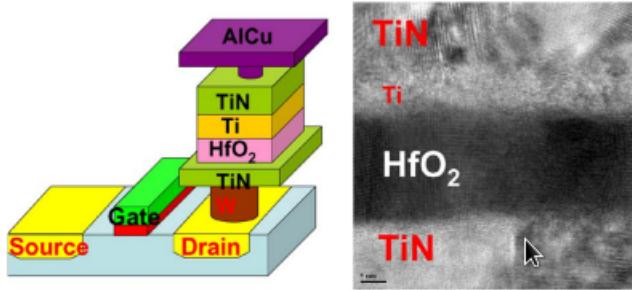
2.2.1 Algemene configuraties

Op basis van deze bevindingen kan de memristor gebruikt worden als geheugenelement: de MIM-configuraties hebben op z'n minst 2 resistieve toestanden, al zijn er artikels gepubliceerd waarbij ook tussenliggende toestanden gebruikt worden[17]. Als deze multiresistive states onder controle gehouden kunnen worden, kan een nog ho-

2. GEHEUGENCSEL



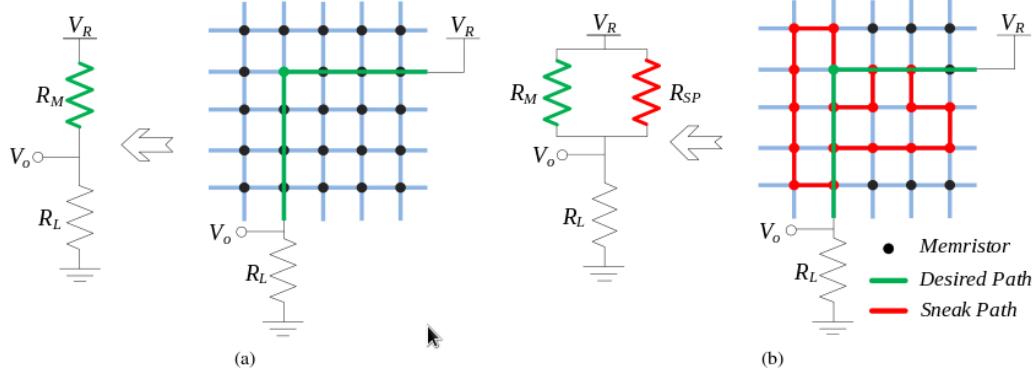
Figuur 2.5: DC-analyses bipolaire en unipolaire memrisors, reproduced from [26] & [3]



Figuur 2.6: Een 1T1R-configuratie, reproduced from[29]

gtere densiteit aan informatie gerealiseerd worden aangezien elke memristor meer dan 1 bit informatie zou bevatten (multi-level cells). Deze 2 toestanden kunnen gebruikt worden voor geheugen- en logicatoepassingen[23][21]. In geheugentoepassingen kan men onderscheid maken tussen 1T1R-, 1R- en 1D1R-configuraties[10]. Met een 1R- configuratie kan men de grootste densiteit van geheugen bereiken, alsook een betere schaling, maar deze configuratie heeft te kampen met cellen die half geselecteerd worden en lekstroom. Dit kan opgelost worden door een selectie-element aan de configuratie toe te voegen, zoals een diode of een transistor. De 1D1R configuratie kan echter enkel geïmplementeerd worden met unipolaire memristors[30]. Daarom is er in dit werk voor een 1T1R-configuratie geopteerd. De memristor waarop dit werk gebaseerd is, is immers een bipolaire hafniumoxide-memristor (zie figuur 2.6).

Naast geheugentoepassingen heeft de memristor ook potentieel in logicatoepassingen en er wordt zelfs gesproken over een mogelijke vervanger van de transistor[14].



Figuur 2.7: Read sneak leakage path, reproduced from[33]

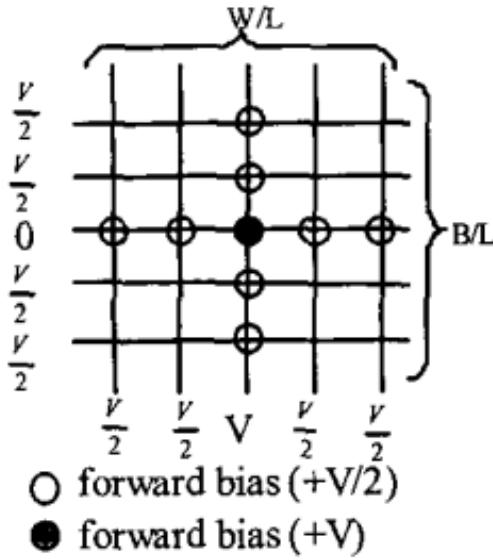
2.2.2 1R-cel

The 1R-configuratie bestaat uit 1 memristor. Deze cel kan dan verbonden worden aan een WL en BL. Deze structuur die beter bekend is als een *crossbar array* kan een geheugenmatrix vormen met een heel hoge densiteit. Een van de problemen waarmee deze configuratie te kampen heeft bij de leesoperatie is echter het *read sneak leakage path*[33](figuur 2.7). Bij de leesoperatie wordt er een spanningsdeling uitgevoerd tussen de memristor en een lastimpedantie. Bij een *crossbar array* gaan er echter lekstromen gevormd worden die door andere cellen vloeien. Dit zorgt voor een parallelle weerstand die heel afhankelijk is van de resistieve staat van de cellen in de geheugenmatrix. Volgens [31] kan dit probleem opgelost worden door de resterende WL te biasen of door de leescyclus in twee stappen uit te voeren. Bij deze laatste oplossing gaat men eerst de lekstroom van het circuit opmeten en deze vervolgens aftrekken van de totale leesstroom.

De schrijfoperatie heeft dan weer te kampen met een probleem dat het *write half-select problem* heet (figuur 2.8). Hierbij gaan cellen die op dezelfde BL of WL liggen als de geselecteerde cel, half geselecteerd zijn. De uitdaging hierbij is om te garanderen dat deze cellen niet van resistieve staat veranderen. Er bestaan verschillende biasing methodes [4] om dit probleem te overkomen. Bij het schrijven van meerderen cellen tegelijkertijd, kunnen methodes zoals *SET-before-RESET* of *ERASE-before-RESET*[31] gebruikt worden.

2.2.3 1T1R-cel

De 1T1R-configuratie bestaat uit 1 CMOS-transistor en 1 memristor in serie geschaald. De cel kan worden verbonden in een matrix zoals geïllustreerd in figuur 2.9. Een wordline wordt verbonden aan de gate van de transistor, de source van de transistor wordt verbonden aan een sourceline en tenslotte wordt een bitline gehangen aan de vrije terminal van de memristor. Ook lees- en schrijfoperaties worden geïllustreerd op figuur 2.9. De leesoperatie is in wezen een spanningsdeling tussen de impedantie van de cel en een andere lastimpedantie die de BL verbindt met de voedingsspanning.

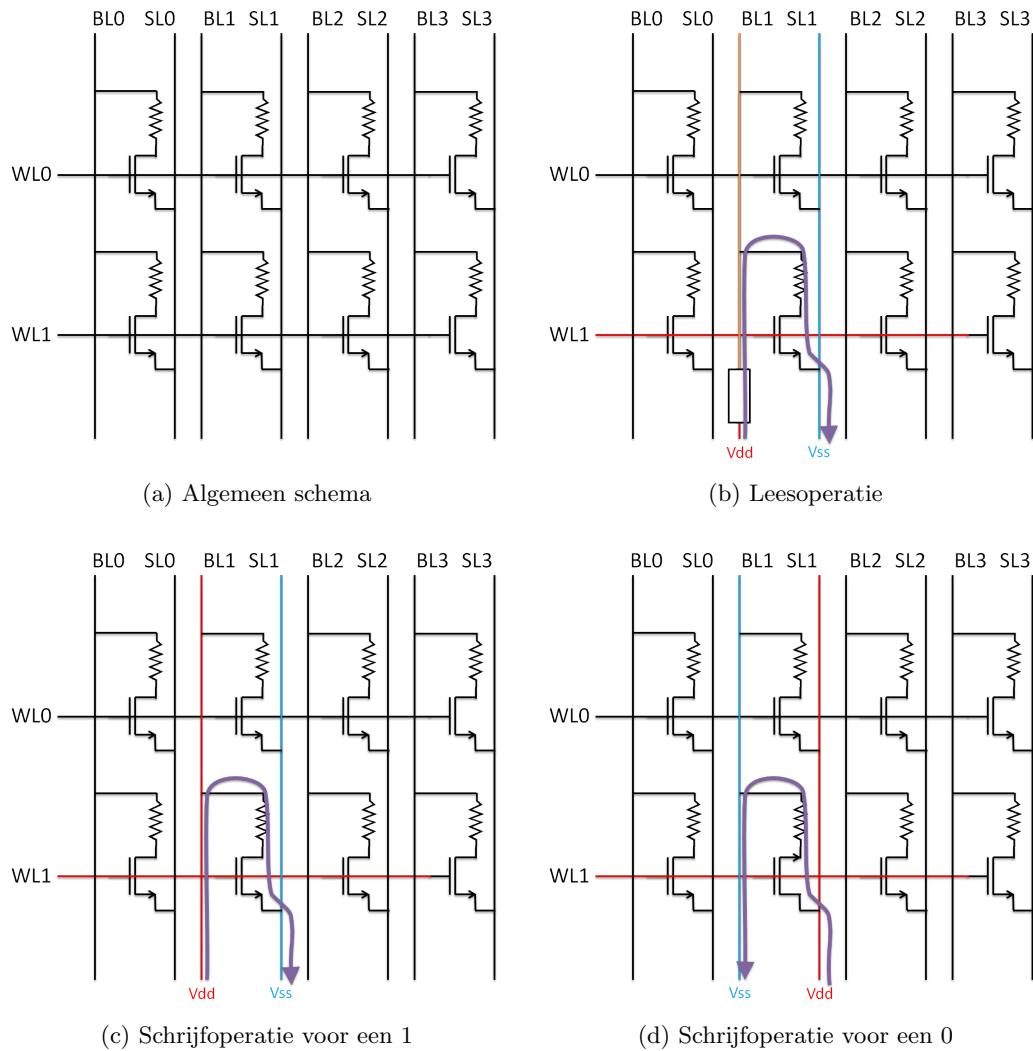


Figuur 2.8: Half select probleem, zwarte cellen zijn volledig geselecteerd, witte cellen zijn half geselecteerd, reproduced from[4]

Om een logische 1 te schrijven kan de BL rechtstreeks worden aangesloten aan de voedingsspanning. De stroom door de memristor is hierbij groter dan bij leesoperatie en de memristor (indien die zich nog niet in de gewenste resistieve staat bevond) gaat schakelen naar de gewenste resistieve staat. Om een logische 0 te schrijven staat er op de BL de grondspanning en op de SL de voedingsspanning. De stroom vloeit nu in de omgekeerde richting.

2.3 Besluit

De memristor is een theoretische passieve component die kan gemodelleerd worden via een verband tussen lading en elektrische flux. In de praktijk zijn er MIM-configuraties ontdekt die (gedeeltelijk) memristorkarakteristieken vertonen. Deze karakteristieken zijn bijzonder interessant voor geheugens: gecombineerd met een transistor vormt de memristor een 1T1R-geheugencel, die geïmplementeerd wordt in de volgende hoofdstukken.



Figuur 2.9: 1T1R matrix met verschillende operaties

Hoofdstuk 3

Geheugenarchitectuur

Om overzicht te houden op de geheugenarchitectuur, zijn bepaalde bouwblokken gedefinieerd. Het grootste bouwblok is de global block, dit bestaat uit twee local blocks en een sense amplifier. Local blocks zijn geheugencelmatrices met decoders en passgates er rond. Dit hoofdstuk bespreekt de algemene structuur alsook de vrijheidsgraden die in hoofdstuk 7 onderzocht worden om tot een optimaal werkend systeem te komen. Ten slotte zullen ook nog de bouwblokken vermeld worden die meer uitvoerig besproken worden in de volgende hoofdstukken.

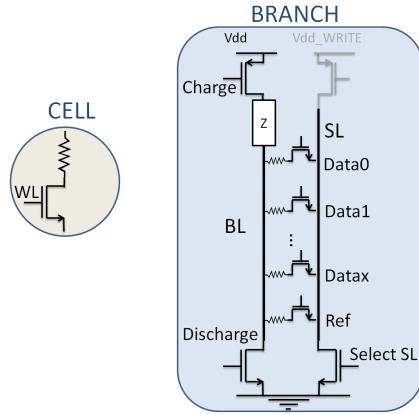
3.1 Cel

Dit elementaire bouwblok is besproken in sectie 2.2.3 en vormt de kern van het geheugensysteem. De cel bestaat uit een memristor en een transistor (1T1R-cel). De geheugencel heeft drie terminals: de gate van de transistor, die verbonden wordt met een wordline, de source van de transistor, die verbonden wordt met een sourceline en tenslotte de terminal van de memristor, die verbonden wordt met een bitline.

3.2 Branch

De branch en cel worden getoond in figuur 3.1. In een branch worden er een bepaald aantal datacellen verbonden aan één BL en één SL. Dit aantal wordt *Number of Word Lines per Branch* (NoWLpB) genoemd en is een van de vrijheidsgraden van deze geheugenarchitectuur. Naast alle datacellen is er ook nog één referentiecel - dit is een cel waarvan de resistieve staat voorgeschreven is - verbonden aan de BL en SL van de branch. Elke BL wordt via een pMOS-transistor (al dan niet met nog een impedantie tussenin) gekoppeld aan de voedingsspanning Vdd en via een nMOS-transistor aan de grondspanning Vss. In dit werk is er enkel een nMOS-transistor die de SL verbindt met Vss.¹ De nMOS-transistoren aan BL en SL fungeren als schakelaars, de pMOS-transistor wordt daarenboven ook gebruikt als impedantie voor een resistieve spanningsdeling (zie hoofdstuk 4).

¹In een volledig geheugensysteem zou de SL via een pMOS ook nog verbonden zijn met een niet onderzochte spanningsknoop Vdd_write. De pMOS zou dan worden aangezet voor schrijfwerking.



Figuur 3.1: Een geheugencel en een branch

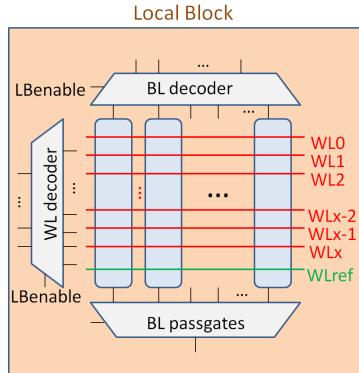
3.3 Local Block

Verschillende BLs en SLs worden samengebracht in een local block, waarvan de vrijheidsgraad *Number of BitLines per Local Block* (NoBLpLB) heet. In een LB bevinden er zich dus NoBLpLB x NoWLpB datacellen en NoBLpLB referentiecellen. Ook bevat een local block zowel BL- als WL-decoders. De afzonderlijke BLs worden via passgates verbonden tot een uitgangsknooppunt. De structuur van een local block wordt afgebeeld op figuur 3.2, een meer gedetailleerd beeld - zonder decoders - is getoond op figuur 3.3. De uitgangen van de WL-decoder sturen de data-WLs aan [eventueel met een buffer], de uitgangen van de BL-decoder activeren een spanningsdeling op de BLs.² De referentie-WL is via een extern signaal verbonden. Voor een gedetailleerdere beschrijving over hoe de decoderuitgangen gebruikt worden, zie sectie 7.1. Een LB heeft twee werkingsmodes: een mode waarbij er één datacel wordt aangesproken om een datasignaal aan de uitgang te verkrijgen en een mode waarbij er een bepaald aantal referentiecellen in parallel wordt aangesproken om een referentiesignaal aan de uitgang te verkrijgen.

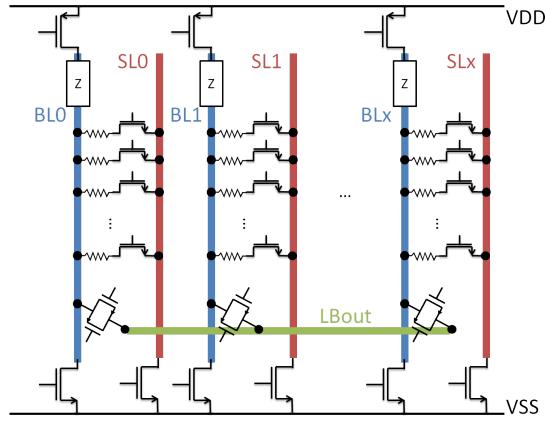
3.3.1 Dataspanning genereren uit datacel

Het datasignaal is de spanning op de BL wanneer er een resistieve deling aan de gang is, waarbij er stroom vloeit door één cel. De last die hangt aan de voedingsspanning, op figuur 3.4 voorgesteld als een pMOS-transistor en optionele extra impedantie, wordt aangeschakeld, alsook de nMOS-transistoren in de cel en aan de sourceline. Deze vormen met de weerstand van het geheugenelement zelf de equivalente totale weerstand R_{tot} . Er vloeit een stroom langs dit pad: $I = V_{dd}/R_{tot}$ en de spanning op de BL is $V = I \cdot R_{Req}$ met R_{Req} de equivalente weerstand van de SL-transistor- en

²Indien schrijfbewerking zou toegevoegd worden, zouden de uitgangen van de BL-decoder aan twee AND-poorten worden verbonden; bij leesoperatie brengt de uitgang van de ene AND-poort de resistieve deling op de BL teweeg, bij schrijfoperatie zet de uitgang van de andere AND-poort een pull-up-operatie van de BL naar V_{dd_write} op.



Figuur 3.2: Een Local Block



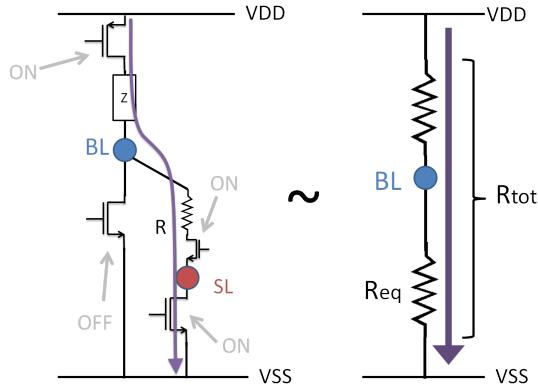
Figuur 3.3: Een meer gedetailleerde illustratie van een LB, decoders zijn weggelaten

celimpedantie. Om dit datasignaal op de uitgangsknoop van het LB te krijgen, wordt de bijhorende passgate van de BL in kwestie geactiveerd.

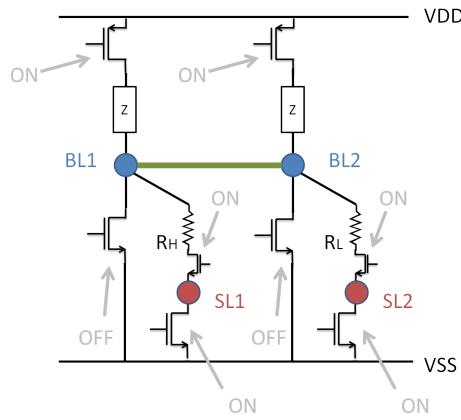
3.3.2 Referentiespanning genereren uit referentiecellen

Het referentiesignaal is een spanning die tussen de spanning van een LRS datasignaal en een HRS datasignaal moet liggen. Een dergelijk signaal kan verkregen worden door twee BLs kort te sluiten zoals op figuur 3.5. In dit ontwerp zal de kortsluiting gerealiseerd worden door de passgates van de BLs aan te zetten, zo komt het referentiesignaal bovendien ook op de uitgangsknoop te staan. In theorie is het voldoende om 2 BLs [de ene met een HRS cel en de andere met een LRS cel] kort te sluiten om het referentiesignaal te verkrijgen. Er zit echter op de resistieve geheugenelementen variabiliteit: er wordt aangenomen dat R_H normaal verdeeld is met $\mu = 32500\Omega$ en $\sigma = 833\Omega$. R_L is ook normaal verdeeld met $\mu = 7500\Omega$ en $\sigma = 833\Omega$. Dit betekent dat ook de data-signalen en referentie-signalen stochastische variabelen zijn. Door meerdere referentiebitlines kort te sluiten gaat de spreiding van

3. GEHEUGENARCHITECTUUR



Figuur 3.4: Spanningsdeling waarbij dataspanning over BL staat

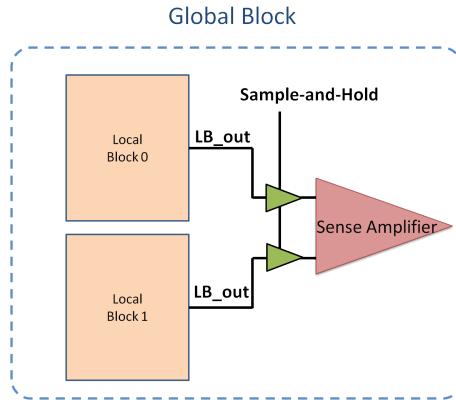


Figuur 3.5: Topologie om referentiesignaal te verkrijgen

het referentiesignaal dalen [maar het energieverbruik stijgen]. Bovendien kan men de verwachtingswaarde verschuiven door meer HRS (LRS) referentiegeheugenelementen te gebruiken dan LRS (HRS).

3.4 Global Block

Een global block bestaat uit twee LBs en een sense amplifier (SA) met bijhorende sample-and-hold-schakelaars (S&H). De S&H-schakelaars worden geïmplementeerd met passgates. In het ene LB gaat er een datasignaal geproduceerd worden, in het andere een referentiesignaal (zie figuur 3.6). Vervolgens gaat de SA dit kleine signaalverschil versterken tot een zuivere rail-to-rail output. Aan de uitgang van het GB verschijnen dan ook de opgevraagde bits. De laatste architectuurvrijheidsgraad is de *Number of Global Blocks* (NoGB), het totale geheugen bevat NoGB x 2 x NoBLpLB x NoWLpB datageheugencellen.



Figuur 3.6: Een global block

3.5 Besluit

De geheugenarchitectuur werd in vogelvlucht overlopen. Het kleinste bouwblok is de cel, deze wordt geplaatst in een branch, d.i. een combinatie van cellen aan een BL en SL, verbonden met schakelaars en impedanties aan VDD en VSS. Verschillende branches vormen samen met decoders en passgates een local block. Twee local blocks en een sense amplifier met bijhorende passgates worden gegroepeerd tot een global block. Het totale geheugen bestaat tenslotte uit een verzameling global blocks.

Hoofdstuk 4

Lastimpedantie-analyse

Om een cel uit te lezen wordt er een spanning gevormd op de bitline door middel van een spanningsdeling tussen twee impedanties, zoals besproken in sectie 3.3.1. De ene impedantie is de celimpedantie, hier valt niet veel aan te veranderen. De andere impedantie is de lastimpedantie, deze moet wel onderzocht worden met het oog op optimalisatie van snelheid, bitline spanningsverschil en spanningsval over de memristor. Ook belangrijk is dat de bereikte resultaten robuust zijn tegen variabiliteit.

4.1 Algemene lasteigenschappen en -specificaties

In deze eerste sectie bestuderen we de spanningsdeling tussen last- en celimpedantie als een simpel model: beiden worden gemodelleerd als een eenvoudige weerstand zoals op figuur 3.4 in sectie 3.3.1. Dit model zal inzicht geven over de invloed van de weerstandswaarden op de spanningsdeling voor geheugenspecificaties zoals verschil tussen opgewekte spanning bij HRS- en LRS-cel en settlingdelay. Een nominaal groot verschil in BL-spanning tussen een HRS- en LRS-cel komt overeen met een relatief groot verschil tussen data- en referentiesignaal. Deze stelling gaat echter enkel op wanneer de referentiespanning het gemiddelde is van HRS- en LRS-dataspanning alhoewel hier voor gezorgd kan worden, zoals aangehaald in sectie 3.3.2. Een groot verschil in data- en referentiespanning is robuuster tegen de offsetspanning van de sense amplifier wanneer variabiliteit in rekening wordt genomen. In het simpele model kan het verschil in bitlinespanning analytisch berekend worden:

$$\Delta V = \left(\frac{R_{HRS}}{R_{last} + R_{HRS}} - \frac{R_{LRS}}{R_{last} + R_{LRS}} \right) VDD \quad (4.1)$$

Voor constante waarden van R_{HRS} en R_{LRS} is er een maximum voor ΔV zoals duidelijk gezien kan worden op figuur 4.1. De sensitiviteit van de lastweerstand op het spanningsverschil moet men voorzichtig interpreteren. Op figuur 4.1 kan gezien worden dat de helling voor het maximum steiler is dan voorbij het maximum. Het is dus beter om een iets grotere lastweerstand te hebben dan een iets te kleine lastweerstand. Wanneer men deze weerstand naar transistorafmetingen vertaalt, kan

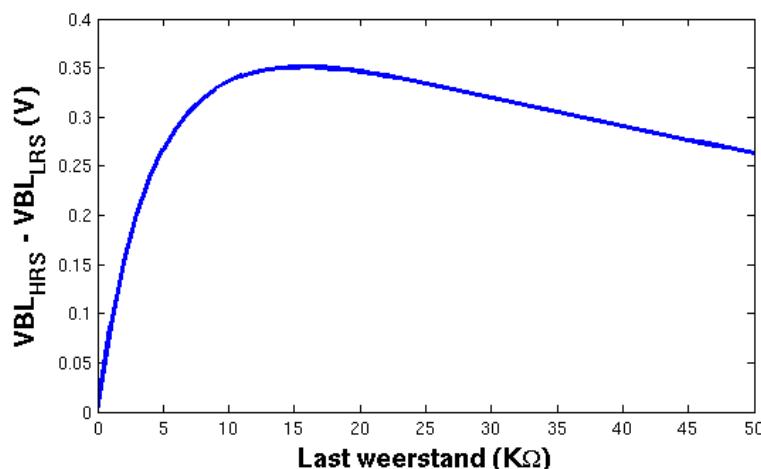
4. LASTIMPEDANTIE-ANALYSE

met dit op verschillende manieren realiseren. De aanweerstand van een transistor is omgekeerd evenredig met $\frac{W}{L}$. Een transistor met grote L heeft ook een grotere weerstand dan een transistor met kleine L, maar dezelfde $\frac{W}{L}$ -verhouding. Voor een transistor met minimale lengte kan de grootste aanweerstand dus enkel gerealiseerd worden voor minimale breedte. Zulk ontwerp is echter gevoeliger voor variaties dan transistoren met grotere breedtes.

De snelheid van het opladen van de bitline kan in het simpele model ook analytisch beschreven worden. De volgende vergelijking stelt het tijdstip voor na het aanschakelen van de voeding wanneer de bitline 99% is opgeladen.

$$t = -\ln(0.01) * RC \text{ met } R = \left(\frac{1}{R_{cel}} + \frac{1}{R_{last}}\right)^{-1} \quad (4.2)$$

Deze delay zal kleiner worden naarmate R kleiner wordt, resulterend in een kleinere lastimpedantie.



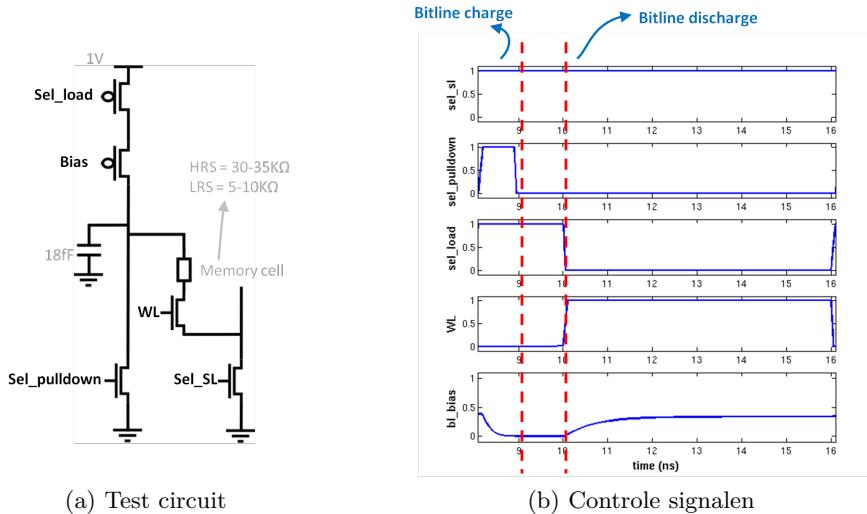
Figuur 4.1: Verschil in bitlinespanning in functie van lastweerstand, $R_{LRS} = 7,5k\Omega$ en $R_{HRS} = 32,5k\Omega$

4.2 Evaluieren van de last

Om verschillende lasten met elkaar te kunnen vergelijken, is het belangrijk om de resulterende grootheden in gelijkaardige simulatieomstandigheden te bekomen. Figuur 4.2 geeft de verschillende aspecten van de gebruikte simulatiesetup weer. Het testcircuit (figuur 4.2a) stelt een bitline voor met een capaciteit van 18fF, wat ruwweg overeenkomt met een bitline waaraan 100 cellen hangen. Aan deze bitline zijn naast de cel ook een last en een ontladintransistor aangesloten. De ontladingstransistor heeft minimale afmetingen. De nominale waarden voor LRS en HRS zijn $7.5k\Omega$ en $32.5k\Omega$. Tijdens Monte Carlo simulaties worden deze nominale waarden als verwachtingswaarde genomen van een Gaussische distributie met $\sigma = 0.833k\Omega$. Aan deze

memristorweerstand hangt een WL-transistor, die ook minimaal gehouden wordt. Samen vormen deze de geheugencel. De cel hangt aan de BL, WL en SL. Aan de SL is tenslotte nog een ontladingstransistor verbonden. Deze transistor wordt bewust groot gekozen zodat de equivalente weerstand van de onderste tak gedomineerd wordt door de weerstand van de geheugencel. De SL bevat in de simulatieopstelling ook een capaciteit van 18fF , al heeft dit geen significante invloed voor de leesbewerking; de ontladingstransistor aan de SL staat immers altijd aan. Tenslotte wordt de voedingsspanning altijd op 1V gehouden.

Figuur 4.2b stelt de sequentie voor van alle controlesignalen tijdens de simulatie. Eerst wordt de bitline volledig ontladen. Vervolgens is er een interval waarin enkel de SL-transistor aanstaat. Tenslotte worden last en cel aangeschakeld en wordt de bitline opgeladen. Op het einde van de simulatie kan men bij benadering stellen dat de BL volledig is opgeladen.



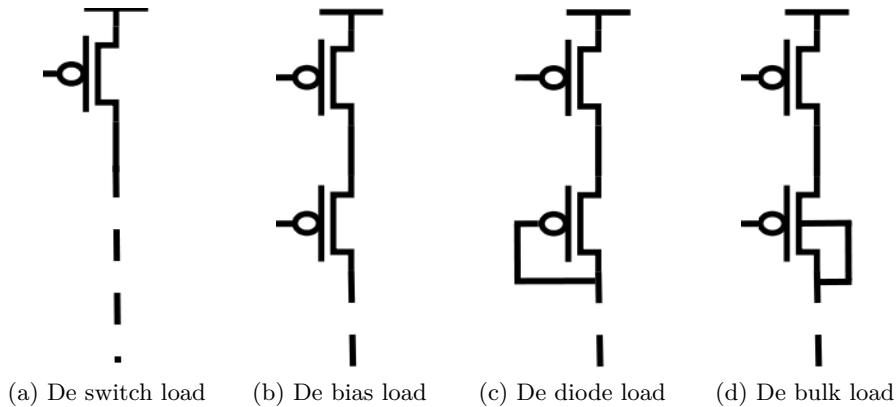
Figuur 4.2: Testbench voor de lastimpedantie

Eens een simulatie met een bepaalde impedantie uitgevoerd is, wordt deze last beoordeeld op basis van resulterende oppervlakte, BL-laadsnelheid, nominale BL-spanningsverschil en spanningsval over het geheugenelement. De oppervlakte wordt berekend op basis van de lengtes en breedtes van de lasttransistoren. De BL-laadsnelheid is de tijd die nodig is om de bitline 99% op te laden. Het nominale BL-spanningsverschil is het verschil van de spanning over 2 BLs - aan de ene hangt een cel in HRS, de andere een cel in LRS - wanneer de bitline 100% opgeladen is. De bitline wordt verondersteld 100% opgeladen te zijn op het einde van de simulatie, de simulatietyl wordt voldoende hoog gehouden om dit te garanderen. De spanningsval over het geheugenelement is belangrijk wat betreft destructieve leesoperaties. Een te grote spanning gedurende een te lange tijd kan een schakeling van resistieve toestand

veroorzaken. De numerieke waarde van de maximale spanningsval over de cel is heel erg afhankelijk van het type memristor. In dit onderzoek wordt uitgegaan van een maximum van 0,5V over de cel tijdens de leesbewerking[25].

4.3 Vergelijking van verschillende types last

Voor dit onderzoek worden vier mogelijke kandidaten van lastimpedanties vergeleken: de switchload (figuur 4.3a), de biasload (figuur 4.3b), de diodeload (figuur 4.3c) en de bulkload (figuur 4.3d)[22]. Eerst wordt er een lineare sweep gedaan van de verschillende lasten (sectie 4.3.1), waarbij enkel de breedtes en bias spanningen worden geswept. De lengtes van de transistoren worden minimaal gehouden om er voor te zorgen dat de lasttransistoren binnen de pitch van de bitline passen. Eens variabiliteit wordt toegevoegd aan de simulatie met Monte Carlo simulaties (sectie 4.3.2) zal echter blijken dat het verschil in BL-spanning te klein is en zal de lengte van de lasttransistoren ook moeten worden vergroot (sectie 4.3.3). Dit vermoeilijkt het layoutontwerp echter.



Figuur 4.3: De verschillende types lastimpedanties

4.3.1 Lineaire sweep van de lasten

De switchload bestaat uit één pMOS-transistor die volledig wordt aan- of afgeschakeld. Een lineaire sweep met een transistorbreedte tussen 100nm en 500nm werd uitgevoerd en is geïllustreerd in figuur 4.4. Bij het vergroten van de transistorbreedte zal de aanweerstand dalen en het verschil tussen de BL-spanningen ook. Als we deze last vergelijken met het simpele model uit sectie 4.1, zit de weerstandswaarde aan de linkerkant van de piek uit figuur 4.1. Bij het vergroten van de transistorbreedte zal de BL-spanning stijgen en de spanningsval over het geheugenelement dus ook. Verder volgt de settling-tijd ook het simpele model uit sectie 4.1, waarbij de settling-tijd daalt bij kleinere weerstandswaarden.

De biasload is een last met twee pMOS-transistoren in serie. De bovenste transistor wordt als een schakelaar gebruikt en dus volledig aan- of afgesloten. De gate van de onderste transistor wordt gebiased op een bepaalde spanning. Het voordeel van de biasload is dat men grotere weerstanden kan realiseren en dus de piek kan bereiken uit figuur 4.1. Dit kan men duidelijk zien op de xassen van figuur 4.5. Ook hier zijn de breedtes van de transistoren geswept tussen 100nm en 500nm. De biasspanning is tussen 0V en 0.4V geswept. Een hogere biasspanning heeft geen nuttige bijdrage. Omdat de kleinste weerstand die deze configuratie kan aannemen binnen deze sweeprange net iets groter is dan die van de switchload, is de biasload ook iets trager. De oplossingen waarbij dit het geval is, hebben echter een onbruikbaar verschil in BL-spanningen. De spanningsval over het geheugenelement is vergeleken met de switchload heel wat hoger maar voor de meeste oplossingen ligt ze nog steeds onder de limiet van 0.5V.

De diodeload bestaat ook uit twee transistoren waarbij de bovenste functioneert als schakelaar zoals bij de biasload. Bij de onderste transistor zijn drain en gate kortgesloten, dit noemt men een diode-geconnecteerde MOS-transistor. Uit de resultaten van de sweep (figuur 4.6) blijkt dat de settling met deze last heel snel is, maar het BL-spanningsverschil is te klein om bruikbaar te zijn.

De bulkload werd voorgesteld in de paper van Ren et al. [22] als een goede kandidaat omwille van zijn grote uitgangsimpedantie. Deze last bestaat uit een serieschakeling van een schakelaartransistor en een bulk-geconnecteerde transistor. Deze bulk-geconnecteerde transistor wordt op 0V gebiased aangezien dit optimale resultaten geeft. De breedtes van de transistoren werden gevareerd van 100nm tot 500nm. De resultaten van deze sweep zijn weergegeven in figuur 4.7. In de resultaten kan gezien worden dat deze last zich vergelijkbaar gedraagt als de biasload.

4.3.2 Het toevoegen van variabiliteit

Na een pareto-optimale selectie te hebben gemaakt van de oplossingen uit de vorige sectie, worden met deze oplossingen nieuwe simulaties gedaan waarbij ook variabiliteit in rekening gebracht wordt. De variabiliteit is toegevoegd op alle transistoren in het testcircuit en op de weerstandswaarde van de geheugenelementen. Voor de transistoren wordt er een Pelgrom constante voor V_t van $2.5mV\mu m$ gebruikt en voor β een van $1.2\%\mu m$ [15]. Voor de weerstandswaarde van de memristors wordt er een gaussische verdeling gebruikt met verwachtingswaardes $7.5k\Omega$ en $32.5k\Omega$ en met $\sigma = 0.833k\Omega$. Er worden telkens 500 Monte Carlo simulaties uitgevoerd per oplossing. Hierna worden de BL-spanningen van cellen met een HRS en LRS gefit op een Gaussische distributie. De oplossing met het grootste BL-spanningsverschil tussen de extrema van HRS en LRS is een biasload met een schakelaartransistorbreedte van 100nm, een biastransistorbreedte van 180nm en een biasspanning van 0V. De BL-spanning-distributies zijn geïllustreerd op figuur 4.8. Uit de CDF van deze verdelingen kan men besluiten dat het BL-spanningsverschil in 99.9²% van de

4. LASTIMPEDANTIE-ANALYSE

gevallen¹ groter zal zijn dan 65mV. Dit is niet bijzonder veel aangezien de distributie van de referentiespanning hier ook tussen moet passen en er daarna nog marge moet zijn voor de offsetspanning van de sense amplifier.

Figuur 4.9 stelt de distributie van de referentiespanning voor. De verschillende curves stellen referentiesignalen opgewekt met meerdere referentiecellen [bereik 2 tot 30, steeds evenveel LRS- als HRS-cellens] voor. Zoals gezien kan worden, heeft men een groot aantal cellen nodig om een distributie breedte (6σ) van 39mV te krijgen. Dit betekent dat de offsetspanning van de SA niet groter mag zijn dan 10mV, indien de vooraf vermelde biasload gebruikt wordt. Dit is een zeer strenge voorwaarde. Daarom wordt de constraint waarbij de transistorlengte minimaal gehouden wordt, opgeheven in de volgende sectie.²

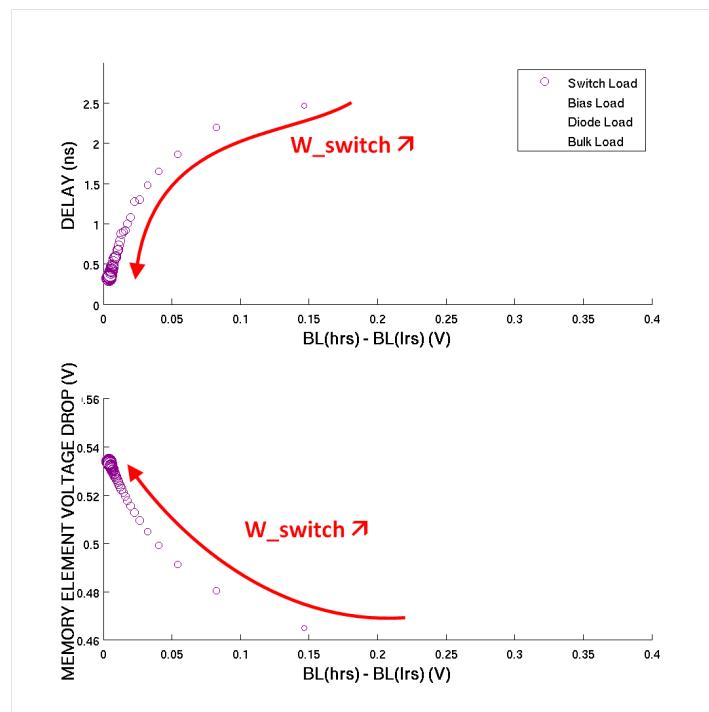
4.3.3 De transistor lengte vergroten

Om de nadelige effecten als gevolg van variabiliteit onder controle te houden moeten de transistoren vergroot worden. Twee opties worden hiervoor overwogen. De eerste is het toevoegen van een derde transistor in serie. Om dezelfde lastimpedantie te bekomen als voor 2 transistoren in serie, moeten de drie transistoren een grote breedte hebben. De $\frac{W}{L}$ -verhouding vergroten verlaagt de aanweerstand van de individuele transistoren, maar door ze in serie te schakelen blijft de equivalente weerstand voldoende groot. Omwille van de vergrote afmetingen zouden ze bovendien minder gevoelig zijn voor mismatch. Hierbij wordt wel verondersteld dat alle drie de transistoren zich in lineair gebied bevinden. Uit simulatieresultaten blijkt de onderste transistor zich in het near- tot subthresholdgebied te bevinden. De stroom in het subthreshold gebied varieert exponentieel met $V_{GS} - V_T$. De stroom en aanweerstand van de transistor zijn dus zeer gevoelig voor VT-variaties. Dit fenomeen ziet men niet bij 2 transistoren in serie, aangezien de transistoren zich hier in het lineare gebied situeren. Daarom wordt er om de mismatch onder controle te houden geopteerd voor een tweede optie, namelijk het vergroten van de transistorlengte. De lengte vergroten resulteert in een toename van de aanweerstand en vermindert variabiliteit. Omwille van de eenvoud, wordt er voor deze nieuwe ontwerpkeuze teruggegrepen naar de switchload.

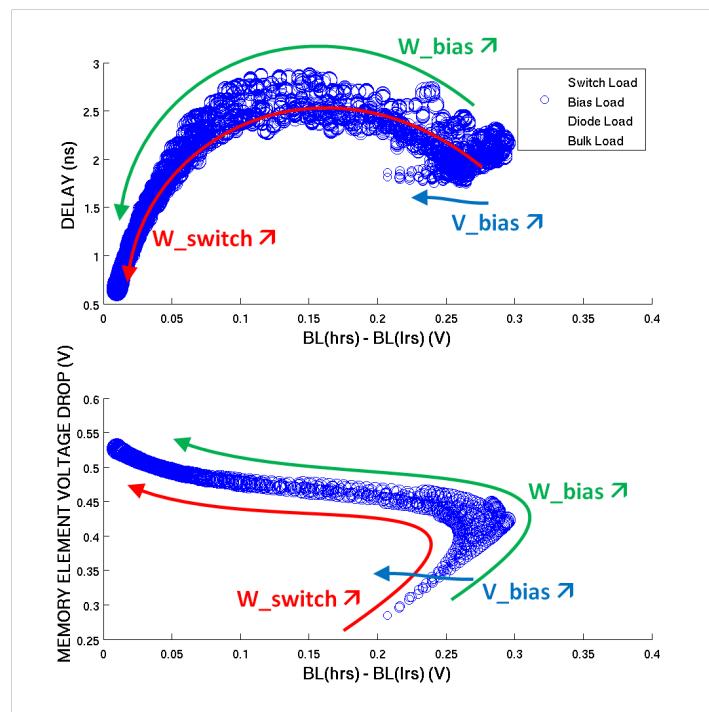
Figuur 4.10 geeft de resultaten weer van een sweep van verschillende lengtes en breedtes voor een switchload. De resultaten worden voorgesteld in functie van $\frac{W}{L}$ wat een indicatie is voor de weerstand van de transistor. In de bovenste figuur kan men duidelijk een maximum zien voor het verschil in BL-spanning zoals in sectie 4.1 werd voorspeld. Verder dient worden opgemerkt dat er een oplossing aan de linkerkant van het maximum gekozen moet worden aangezien de spanningsval over

¹ $CDF(V_{BL-RHS}) < 0,1\% - CDF(V_{BL-LHS}) > 99,9\%$

²Door met complementaire cellen te werken - bij elk datacel hoort een andere cel waarbij het geheugenelement zich in de andere resistieve toestand bevindt - kan het gebruik van referentiesignalen geëlimineerd worden: de SA vergelijkt in dit geval altijd een HRS-singaal met een LRS-singaal (het BL-spanningsverschil) en de offsetspanning is minder kritisch wanneer variabiliteit in rekening wordt genomen. Voor dit soort architectuur kan dus gerust een lastimpedantie met minimale lengte worden toegepast ten koste van meer oppervlakte.

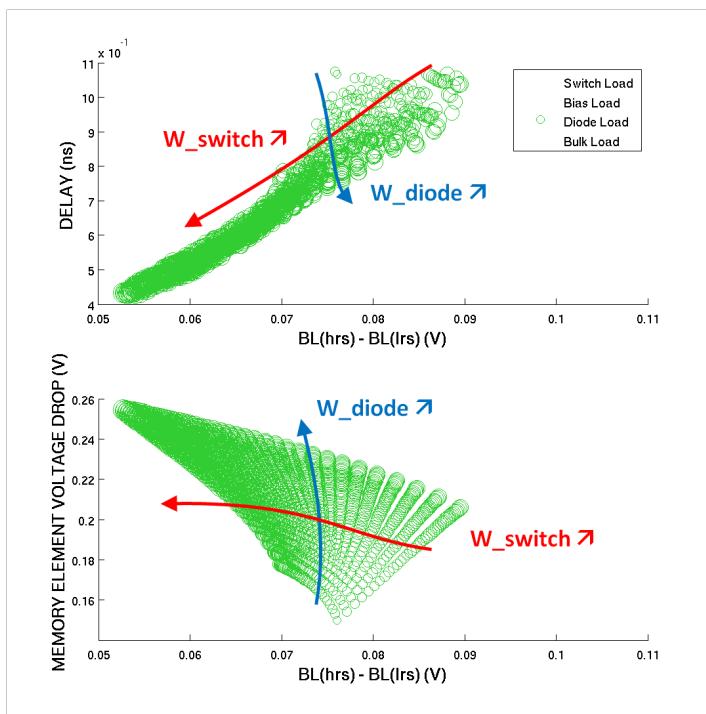


Figuur 4.4: Lineaire sweep van switchload

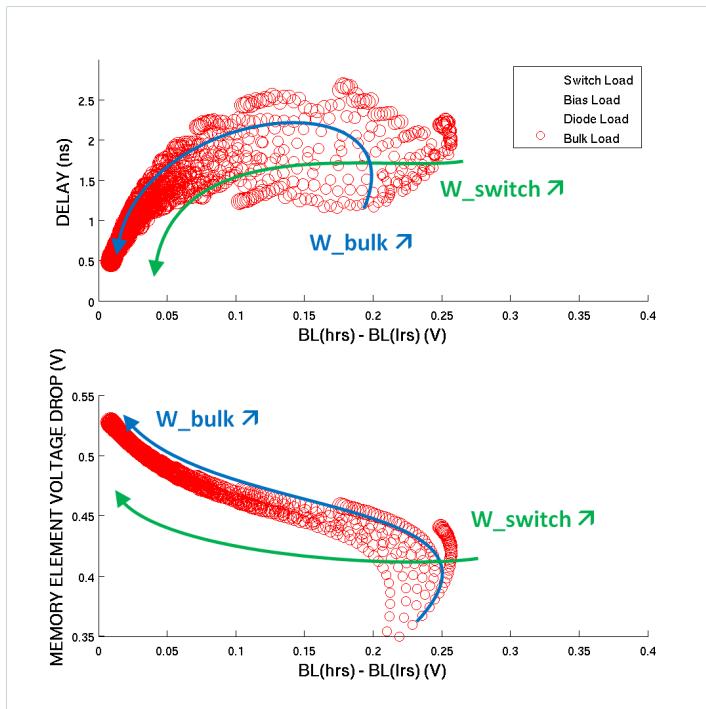


Figuur 4.5: Lineaire sweep van biasload

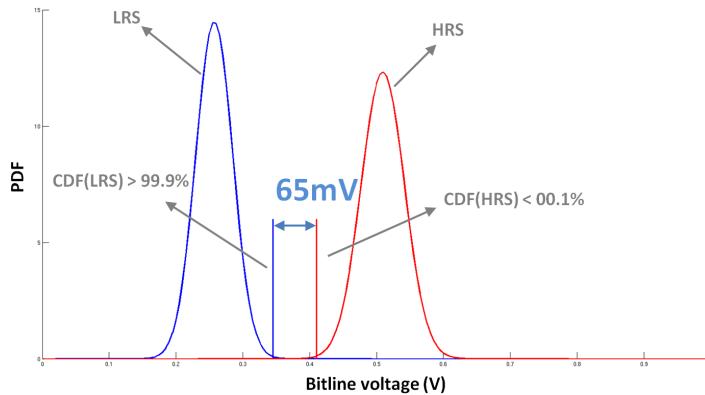
4. LASTIMPEDANTIE-ANALYSE



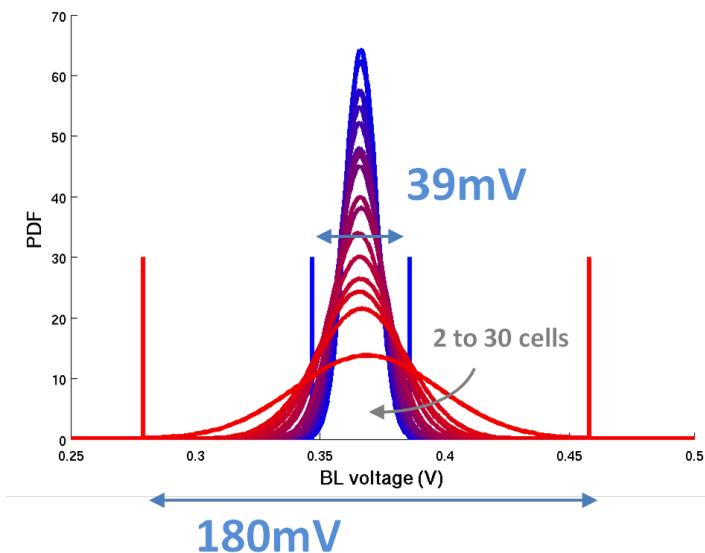
Figuur 4.6: Lineaire sweep van diodeload



Figuur 4.7: Lineaire sweep van bulkload



Figuur 4.8: BL-spanningsverdeling voor een biasload

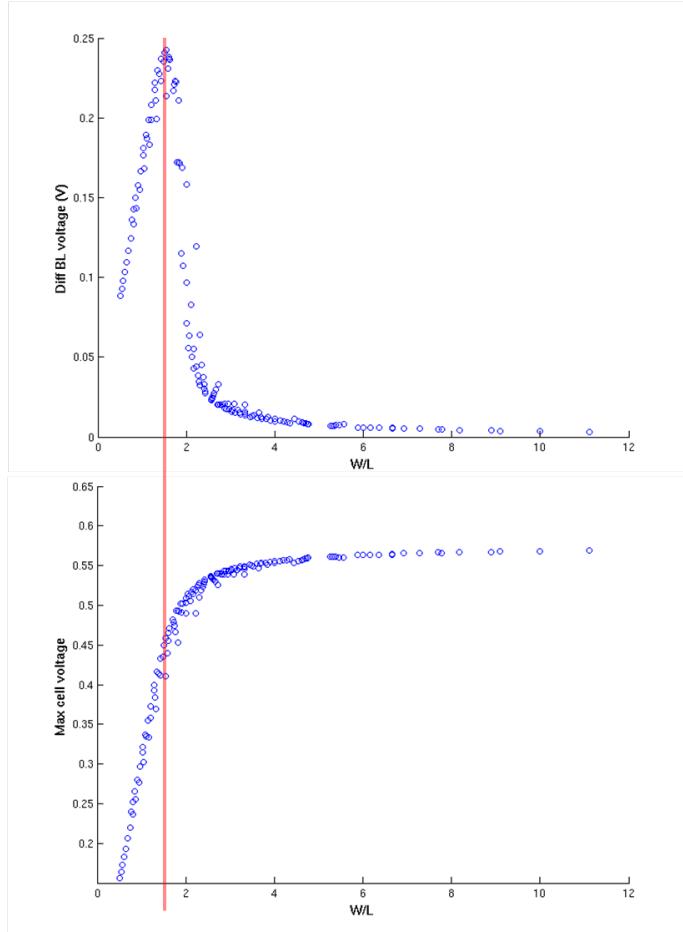


Figuur 4.9: Bitline spanningsdistributies voor een verschillend aantal referentiecellen

het geheugenelement van de oplossingen aan de rechterkant van het maximum te hoog zijn.

Voor de finale last wordt er gekozen voor een transistor met lengte 198nm en breedte 300nm. Op figuur 4.10 wordt deze aangeduid met de rode lijn. Op figuur 4.11 wordt de BL-spanningsdistributie van deze last getoond. Het minimale verschil in BL-spanning is bijna 200mV. De distributie van het referentiesignaal is ook aangegeven op deze figuur. De referentie bestaat uit 16 referentie cellen waarvan 6 in HRS en 10 in LRS. Het aantal cellen in RHS en LRS is zo gekozen dat de referentie distributie in het midden ligt tussen de distributies van de data signalen. Aangezien de standaarddeviatie op de BL-spanningen heel wat kleiner is, kan er gerust gekozen worden voor een last met een kleiner nominaal BL-spanningsverschil . Dit brengt 2

4. LASTIMPEDANTIE-ANALYSE

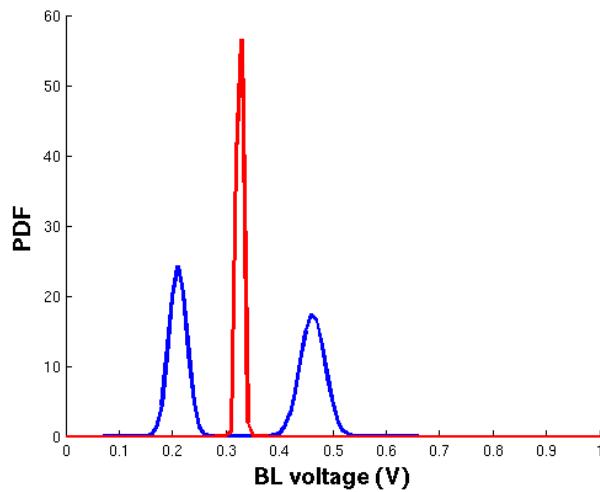


Figuur 4.10: Verschillende nominale oplossingen voor de switchload met variabele lengtes en breedtes

voordelen met zich mee. Zo is de spanningsval van de memristor lager wanneer men een last kiest links van het maximum in figuur 4.10. Voor dergelijke lastimpedanties moeten de BLs bovendien tot een lagere spanning opladen wat een energie winst oplevert. Ondanks deze voordelen werd er toch geopteerd voor de oplossing met het grootste BL-spanningsverschil.

4.4 Besluit

Verschillende kandidaten voor lastimpedanties werden overwogen. Aanvankelijk werd er getracht een last met minimale transistorlengtes te vinden, dit bleek echter niet haalbaar wanneer variabiliteit in rekening wordt genomen. Een enkele transistor met niet-minimale afmetingen bleek de beste resultaten te leveren wat betreft BL-spanningsverschil en spanningsval over geheugenelement. Deze voorwaarden stroken echter met het objectief om de settling tijd te minimaliseren. Het BL-spanningsverschil



Figuur 4.11: BL-spanningsverdeling voor de finale lastimpedantie

en de spanningsval van het geheugenelement zijn wel prioritair, dus werd er niet geoptimaliseerd voor delay.

Hoofdstuk 5

Sense Amplifier analyse

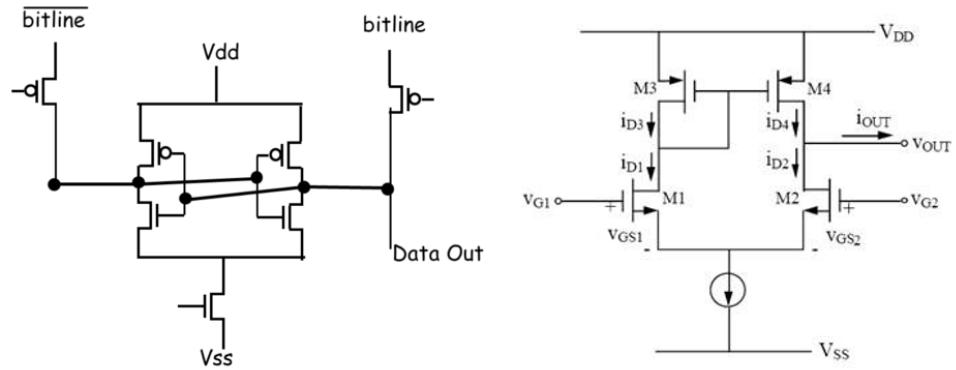
Een sense amplifier versterkt kleine signaalverschillen tot rail-tot-rail signalen. Aangezien de uitgangsignalen hiervan ook de uitgelezen bits zijn van het geheugen, is het uiterst belangrijk dat dit op een correcte manier gebeurt, ondanks alle mogelijke nadelige gevallen van variabiliteit. In dit hoofdstuk wordt de sense amplifier dus ook wat dieper onderzocht. Uiteindelijk wordt een SA ontwerpen die functioneert op een robuuste manier en die ook voldoende snel en laagenergetisch is.

5.1 Types SA

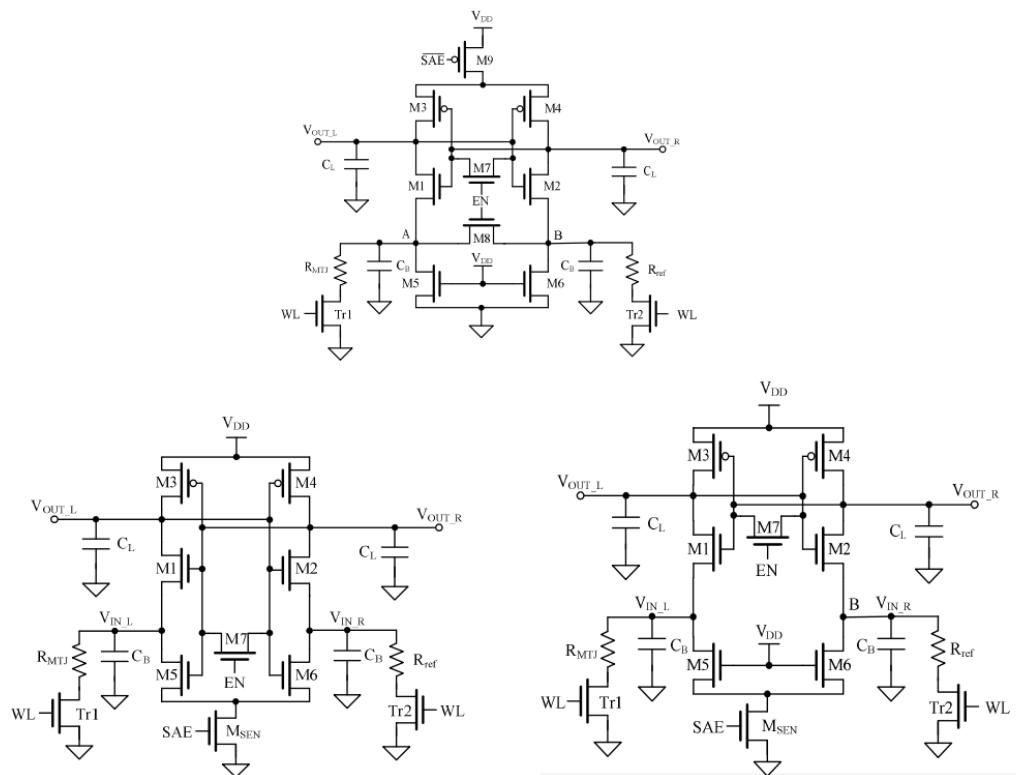
Er bestaan heel wat verschillende soorten sense amplifiers, traditioneel kunnen die op twee verschillende manieren geklassificeerd worden. Enerzijds kan men onderscheid maken tussen differentiële en niet-differentiële SA. Anderzijds kan men een opdeling maken in voltage en current SA. Een voorbeeld van een niet-differentiële SA is een inverter. De drempelspanning waarbij de inverter schakelt wordt bepaald door de groottes van de transistoren. Deze architectuur is echter bijzonder susceptibel aan globale variabelen zoals temperatuur en voedingsspanning. Een dergelijk probleem kan opgelost worden door een referentiesignaal te voorzien dat gelijke variaties ervaart als het datasignaal, als gevolg van deze globale schommelingen. Voor een dergelijke architectuur zijn uiteraard differentiële SAs nodig. Current en voltage SA verschillen in ingangsimpedantie: bij een hoge ingangsimpedantie spreekt men van voltage-sensing, bij een lage impedantie spreekt men van current-sensing. Voorbeelden van voltage- en current-mode sense amplifiers zijn geïllustreerd in figuren 5.1 en 5.2. Voor de architectuur van dit werk waarbij de data- en referentiesignalen opgewekt worden door een spanningsdeling met eigen gekozen lastimpedantie, is een SA nodig met een grote ingangsimpedantie.

Omwillen van eenvoud en goede performantie[8] wordt er in wat volgt voortgewerkt met de zogenoemde drain-input latch-type SA van figuur 5.3.

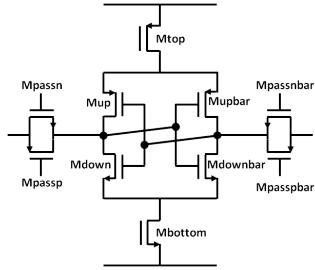
5. SENSE AMPLIFIER ANALYSE



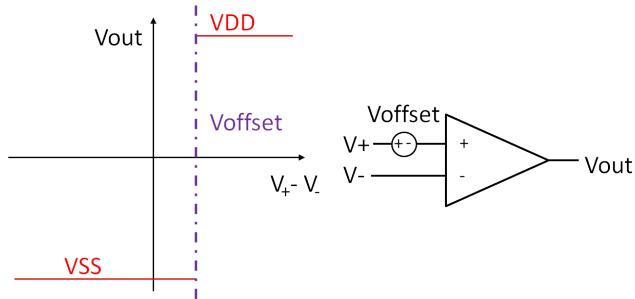
Figuur 5.1: Voltage mode sense amplifiers



Figuur 5.2: Current mode sense amplifiers, reproduced from[6]



Figuur 5.3: De drain-input latch-type SA



Figuur 5.4: Illustratie van offsetspanning

5.2 Offsetspanning

Een ideale sense amplifier zal voor elke twee ingangssignalen correct versterken, tenzij deze signalen dezelfde zijn. Dit is een conditie die de SA in een metastabiele toestand brengt. In de praktijk bestaat er echter wegens variabiliteit een onderlimiet voor het ingangsspanningsverschil waarbij er correct versterkt wordt. Deze limiet heet de offsetspanning en wordt geïllustreerd in figuur 5.4. De offsetspanning van een SA is in de ontwerpfase een stochastische variabele met gemiddelde 0V, pas nadat een chip gefabriceerd is ligt de offsetspanning definitief vast [al kan het zijn dat deze met de tijd nog verandert].

Er zijn 2 manieren waarop men de offsetspanning van een systeem kan aanpakken: ofwel ontwerpt men het systeem zodanig dat het verschil van de ingangssignalen van de SA groot genoeg is zodat het [in 99,x% van de gevallen] groter is dan de offsetspanning, ofwel bouwt men een mechanisme in waarbij de offsetspanning na fabricatie gemeten en vervolgens gecompenseerd wordt. In dit werk is gekozen voor de eerste oplossing. Hiervoor is het wel belangrijk te onderzoeken wat de verdeling is van de offsetspanning, dit wordt gedaan in de volgende sectie.

5.3 Sensitiviteitsanalyse

De SA wordt gerealiseerd als een circuit bestaande uit transistors. Elke transistor wordt in onze simulaties gekarakteriseerd door 2 stochastische parameters

5. SENSE AMPLIFIER ANALYSE

met een normale verdeling, nl. ΔV_t en $\Delta\beta$. De spreiding van deze verdelingen is gekend: $\sigma_{\Delta V_t} = \frac{A_{V_t}}{\sqrt{WL}}$ en $\sigma_{\Delta\beta} = \frac{A_\beta}{\sqrt{WL}}$. Met een sensitiveitsanalyse kan men uit deze standaardafwijkingen de standaardafwijking van de offsetspanning $\sigma_{V_{offset}}$ berekenen. Hierbij wordt verondersteld dat de stochastische variabele V_{offset} een lineaire combinatie is van de normaal verdeelde afwijkingen $(\Delta V_t)_i$ en $(\frac{\Delta\beta}{\beta})_i$:

$$V_{offset} = \sum_{i=1}^N a_i (\Delta V_t)_i + b_i (\frac{\Delta\beta}{\beta})_i.$$

a_i en b_i zijn de gevoeligheden van de offset naar de variatieparameters: $a_i = \frac{\partial V_{offset}}{\partial (\Delta V_t)_i}$ en $b_i = \frac{\partial V_{offset}}{\partial (\frac{\Delta\beta}{\beta})_i}$. Voor een dergelijke variabele geldt dan:

$$\sigma_{V_{offset}} = \sqrt{\sum_{i=1}^N a_i^2 (\sigma_{\Delta V_t})_i^2 + b_i^2 (\sigma_{\frac{\Delta\beta}{\beta}})_i^2}.$$

Er moet wel geverifieerd worden of de stelling dat er een lineaire afhankelijkheid is tussen V_{offset} en de variatieparameters gegroned is. Dit kan gedaan worden aan de hand van een analyse waarbij elke variatieparameter afzonderlijk geswept wordt, een noodzakelijke maar niet voloende vereiste.

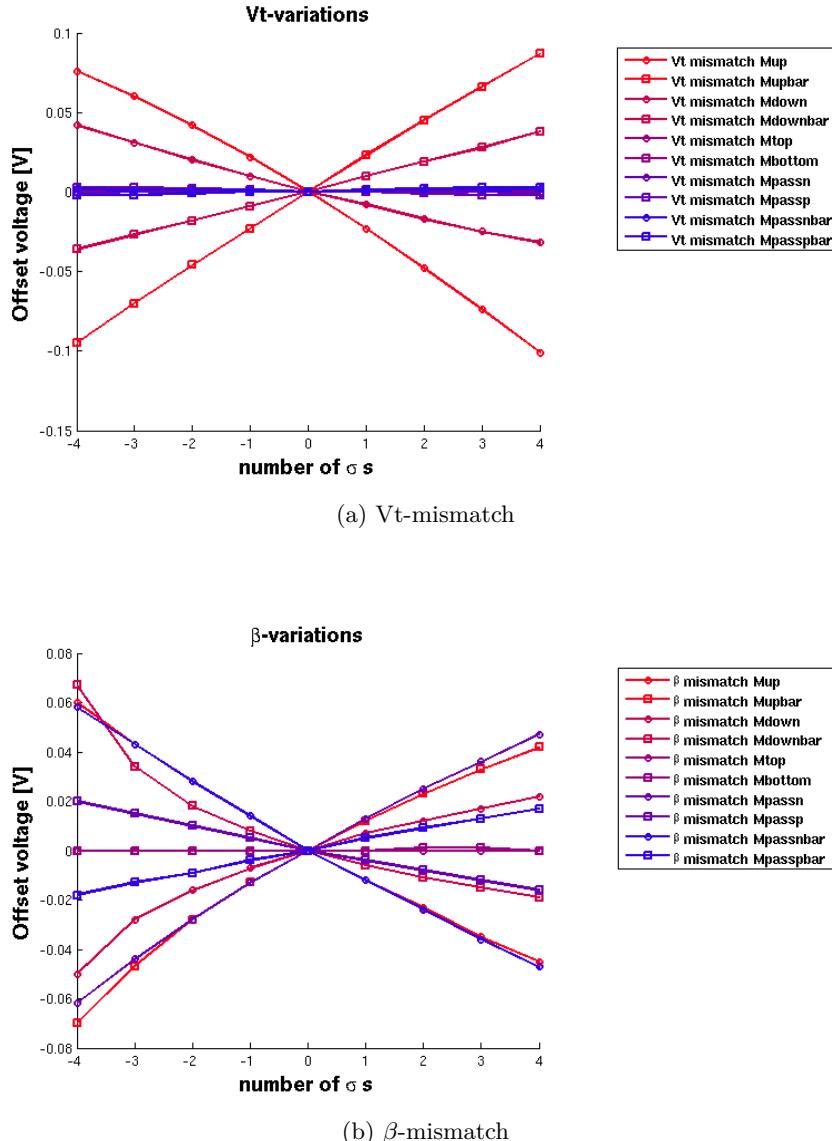
5.3.1 Sensitiviteitsanalyse op een minimale SA

In figuur 5.5 wordt het resultaat getoond voor een dergelijke analyse bij een SA met minimale afmetingen. Op deze grafiek kan gezien worden dat de offsetspanning lineair varieert met de variatieparameters en dat er dus voldaan is aan bovenstaande vereiste. Merk op dat de richtingscoëfficient van deze curves gelijk is aan $a_i(\Delta V_t)_i$ en $b_i(\frac{\Delta\beta}{\beta})_i$. In tabel 5.1 worden de resultaten en de resulterende standaardvariatie van de SA getoond. Er moet opgemerkt worden dat er bij deze simulatie slechts geswept werd voor de variatieparameters van -4σ tot 4σ . Dit is omwille van het feit dat voor de minimale transistoren de standaardvariatie het grootst is. In de Spectre-simulaties zouden transistoren voor te grote negatieve β -mismatch stroom leveren in de omgekeerde richting. Deze situatie zal fysisch nooit optreden.

Opmerkelijk bij deze analyse is dat er een significante bijdrage is van de passgates door β -mismatch. Een nadere observatie leert dat deze bijdrage optreedt door ladingsinjectie van de passgates die niet meer gematched is (zie figuur 5.6)¹. Hierbij moet wel worden opgemerkt dat er voor deze simulatie geen overlap is tussen het controlesignaal om de passgate aan te zetten en het signaal om de SA te activeren (zie figuur 5.7a). De reden hiervoor is dat als er overlap tussen deze signalen is, de SA ook de BL zou trachten op te laden. Hierbij zou men in eerste instantie verwachten dat er moet ingeboet worden aan snelheid en dat dit extra energie zou kosten.

Men kan argumenteren dat er een korte overlap zou kunnen toegelaten zijn, waarna er voldoende spanningsverschil tussen de 2 ingangs-uitgangsknopen zou opgebouwd

¹Voor de gebruikte transistormodellen treedt in simulaties deze ladingsinjectievariatie op door β -variaties. In werkelijkheid komt de exacte ladingsinjectieverdeling allicht niet overeen met de simulaties. Maar het is desalniettemin aannemelijk dat er ladingsinjectiemismatch kan optreden door afmetingvariaties van transistoren.



Figuur 5.5: Sensitiviteitsresultaten: offsetspanning i.f.v. mismatchvariabelen

zijn opdat de ladingsinjectie geen effect meer kan hebben op het eindresultaat (zie figuur 5.7b). Een tegenargument is dat de leescyclus door het opladen van de BL langer zou duren.

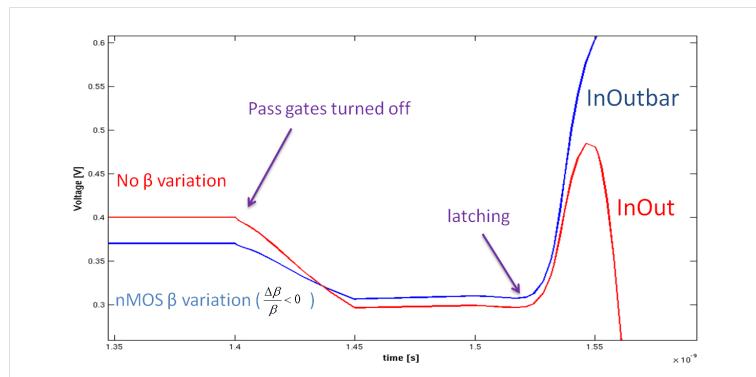
5.3.2 RC-latch-effect

De situatie waarbij er volledige overlap is tussen de controle signalen kan vereenvoudigd worden opgesteld met de situatie van figuur 5.8. De 2 passgates (die van

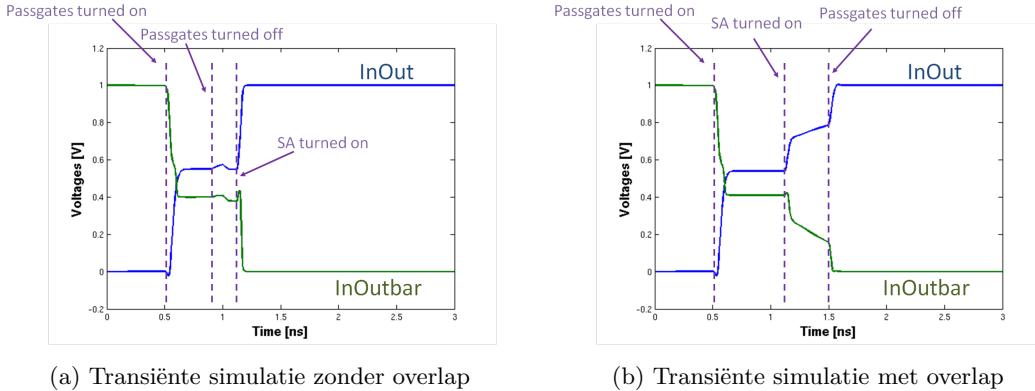
5. SENSE AMPLIFIER ANALYSE

Transistor	Parameter	Richtingscoëfficiënt [$\frac{mV}{\sigma}$]	W [nm]	L [nm]	σ
Mupbar	Vt	22.7	100	45	37.3mV
Mup	Vt	-22.3	100	45	37.3mV
Mupbar	β	13.6	100	45	17.9%
Mpassn	β	13.5	100	45	29.8%
Mpassbarn	β	-13.1	100	45	29.8%
Mup	β	-13.0	100	45	17.9%
Mdownbar	β	-9.4	100	45	29.8%
Mdown	Vt	-9.3	100	45	42.0mV
Mdownbar	Vt	9.2	100	45	42.0mV
Mdown	β	8.2	100	45	29.8%
Mpassp	β	-4.5	100	45	17.9%
Mpassbarp	β	4.4	100	45	17.9%
Mpassbarp	Vt	0.70	100	45	37.3mV
Mpassp	Vt	-0.70	100	45	37.3mV
Mbottom	β	0.083	100	45	29.8%
Mbottom	Vt	-0.033	100	45	42.0mV
Mpassbarn	Vt	0	100	45	42.0mV
Mpassn	Vt	0	100	45	42.0mV
Mtop	Vt	0	100	45	37.3mV
Mtop	β	0	100	45	17.9%
$\sigma_{V_{offset}}$:		45.7mV			

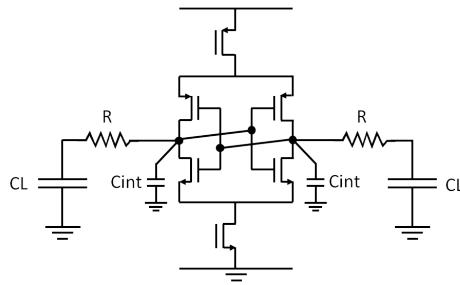
Tabel 5.1: Sensitiviteitsanalyse van de minimale SA



Figuur 5.6: Door β -mismatch is ladingsinjectie van de passgates niet meer gematched en gaat de SA foutief latchen



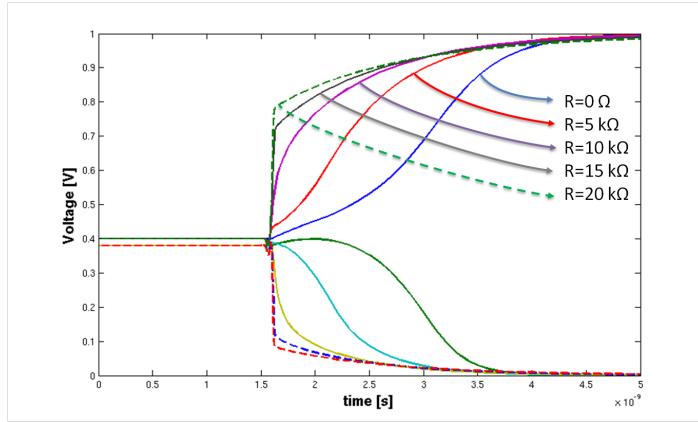
Figuur 5.7: Transiënte simulatie zonder en met overlap tussen passgate en SA operatie. Er zit geen variatie op de circuitelementen. De ingangs-uitgangsknopen van de SA zijn initieel opgeladen op 0V en 1V. De spanning op de LB-uitangsknopen bedraagt 0,4V en 0,55V.



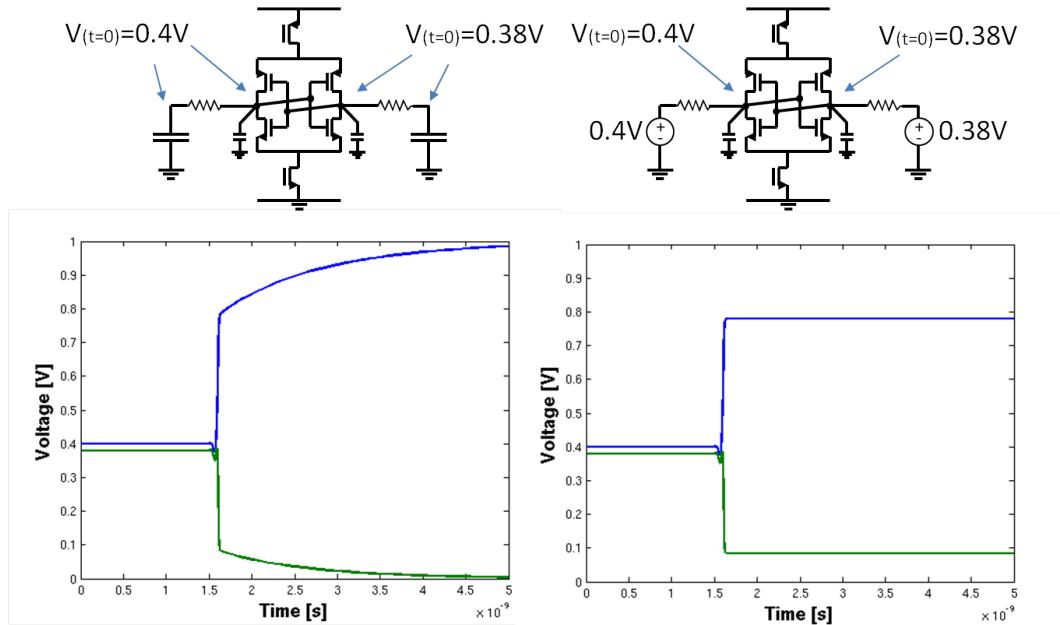
Figuur 5.8: Simulatieopstelling voor het RC-latch-effect

LB en SA) worden voorgesteld als een weerstand, de parasitaire capaciteit tussen de 2 passgates wordt verwaarloosd. CL bedraagt voor deze simulatie 46 fF, het equivalent voor een BL waaraan 256 cellen hangen. Cint bedraagt voor een SA met minimale transistorafmetingen 161 aF. Wanneer het dynamisch latch-gedrag bekijken wordt voor verschillende waarden van R, treedt er een merkwaardig effect op (zie figuur 5.9): voor voldoende grote waarden van R lijkt het alsof de grote capaciteit ontkoppeld is van de latch tot op een zeker tijdstip, waarna een veel tragere settling optreedt. De verklaring ligt in het feit dat CL zich voor hoge frequenties als een kortsluiting gedraagt (zie figuur 5.10), een plotse stroom vloeit door de weerstand waardoor er een spanningsval over de weerstand onstaat. Hierna bouwt er zich met een veel grotere tijdsconstante een spanning op over de capaciteit waardoor de ingangs-uitganksknopen volledig kunnen laden/ontladen tot VDD en VSS. Het gevolg van wanneer dit effect optreedt is dus dat het nuttige signaal zich snel - alsof er helemaal geen last aanhangt - en lineair opbouwt en dat er geen AC-signaal is over de BL-capaciteit. Een analyse van de respons van een RC-circuit op een

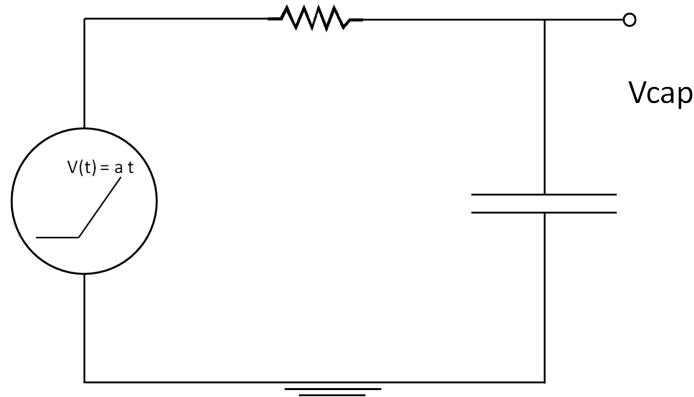
5. SENSE AMPLIFIER ANALYSE



Figuur 5.9: Simulatieresultaten voor het RC-latch-effect: de 2 ingangs-uitgangsknopen zijn voorgeladen op 400mV en 380mV. Na 1,6ns wordt de SA aangezet. De simulatie veronderstelt een SA waarbij er geen variaties zitten op de transistoren.



Figuur 5.10: Vergelijking situatie met voorgeladen (eindige) capaciteit en situatie met spanningsbron (oneindige capaciteit)



Figuur 5.11: Circuit voor analyse voorwaarden RC-latch-effect

lineair stijgende spanningsbron geeft meer duidelijkheid voor de voorwaarden waarop het RC-latch-effect optreedt (zie figuur 5.11). De respons van de spanning over de capaciteit is $V_{cap}(t) = at - aRC(1 - e^{-\frac{t}{CR}})$. Uit deze uitdrukking blijkt dat het RC-latch-effect optreedt wanneer de latch zonder last snel is ($a \ll 1$) en/of wanneer het RC-product hoog is ($RC \gg 1$). Wanneer het effect zich voordoet zijn latching en RC-respons onafhankelijke processen. Wanneer de voorwaarden niet meer zo uitgesproken zijn, gaan deze processen met elkaar interfereren en is het moeilijk dit gecombineerde proces wiskundig te beschrijven.

Conclusie van het RC-latch effect is dat de timing helemaal niet zo kritisch is: in theorie hoeft de overlap slechts even lang te duren als de delay van de SA wanneer er geen last op is aangesloten, maar het is niet erg als de overlap wat langer duurt. De passgates mogen ook minimaal zijn, om hun weerstand te vergroten zodat het effect kan optreden. In geval verder zou gewerkt worden met een SA zonder overlap met passgate-enable en SA-enable, zouden de passgates moeten geschaald worden om de mismatch te minimaliseren. Dit zou wel betekenen dat er per schakeling van de passgates een grotere hoeveelheid lading wordt geïnjecteerd.

5.3.3 Sensitiviteitsanalyse voor minimale SA - vervolg

In tabel 5.2 worden de resultaten van een nieuwe sensitiviteitsanalyse getoond voor een minimale SA, ditmaal waarbij er dus overlap is tussen passgate-enable en SA-enable. De spreiding van de offsetspanning is voor deze simulatieopstelling wel degelijk gedaald. Toch heeft de mismatch van de passgates nog steeds een significante bijdrage, dit kan verklaard worden a.d.h. van figuur 5.8: de 2 weerstanden zijn niet gematcht, deze mismatch treedt op door zowel β - als V_T -mismatch van de passgates. De bijdrage van de mismatch van de differentiële paren daalt dan echter weer door de interactie met de passgates.

5. SENSE AMPLIFIER ANALYSE

Transistor	Parameter	Richtingscoëfficiënt [$\frac{mV}{\sigma}$]	W [nm]	L [nm]	σ
Mupbar	Vt	15.3	100	45	37.3mV
Mup	Vt	-14.9	100	45	37.3mV
Mdownbar	Vt	10.3	100	45	42.0mV
Mdown	Vt	-9.9	100	45	42.0mV
Mpassn	β	8.4	100	45	29.8%
Mpassbarn	Vt	6.8	100	45	42.0mV
Mpassbarn	β	-6.8	100	45	29.8%
Mpassn	Vt	-6.7	100	45	42.0mV
Mupbar	β	6.4	100	45	17.9%
Mup	β	-6.1	100	45	17.9%
Mdownbar	β	-5.6	100	45	29.8%
Mdown	β	5.2	100	45	29.8%
Mpassp	Vt	0.23	100	45	37.3mV
Mpassbarp	Vt	-0.23	100	45	37.3mV
Mtop	Vt	0.17	100	45	37.3mV
Mpassp	β	-0.17	100	45	17.9%
Mpassbarp	β	0.17	100	45	17.9%
Mtop	β	0.12	100	45	17.9%
Mbottom	β	0.050	100	45	29.8%
Mbottom	Vt	0.033	100	45	42.0mV
$\sigma_{V_{offset}}$:		31.7mV			

Tabel 5.2: Sensitiviteitsanalyse van de minimale SA met overlap tussen passenable en latchenable

5.3.4 Sensitiviteitsanalyse voor gebruikte SA

In tabel 5.3 worden de resultaten van de sensitiviteitsanalyse getoond voor de SA die gebruikt wordt in het finale geheugenontwerp. Deze is gekozen aan de hand van de resultaten van de paretosimulatie in de volgende sectie. Er is voor deze SA geopteerd voor overlap tussen passgate- en SA-operatie. De passgates zijn opgeschaald om V_T - en β -mismatch in te perken. Ondanks deze verkleinde weerstand, treedt het RC-latch-effect nog steeds op voor deze SA. Er dient tenslotte opgemerkt te worden dat een voldoende grote SA (passgates niet inbegrepen) het RC-latch effect niet nodig heeft om snel te latchen: een dergelijke SA kan genoeg stroom leveren om de BL-capaciteit snel op te laden.

5.4 Paretosimulatie

In het beginstadium van een ontwerp is nog niet duidelijk wat de impedantie aan de BL wordt. Het is deze impedantie die bepaalt wat het spanningsverschil is tussen het datasignaal en het referentiesignaal aan de sense amplifier. Bovendien kan het zijn dat er midden in het ontwerp besloten wordt om een andere impedantie te kiezen

Transistor	Parameter	Richtingscoëfficiënt [$\frac{mV}{\sigma}$]	W [nm]	L [nm]	σ
Mup	Vt	-4.3	1700	45	9.0mV
Mupbar	Vt	4.3	1700	45	9.0mV
Mpassn	Vt	-3.8	500	45	18.8mV
Mpassbarn	Vt	3.7	500	45	18.8mV
Mpassbarn	β	-3.0	500	45	13.3%
Mpassn	β	3.0	500	45	13.3%
Mup	β	-1.8	1700	45	4.3%
Mupbar	β	1.8	1700	45	4.3%
Mdown	Vt	-1.1	1500	45	10.9mV
Mdownbar	Vt	1.1	1500	45	10.9mV
Mdown	β	0.83	1500	45	7.7%
Mdownbar	β	-0.83	1500	45	7.7%
Mpassp	Vt	0.17	500	45	16.7mV
Mpassp	β	0.17	500	45	8%
Mpassbarp	Vt	-0.17	500	45	16.7mV
Mpassbarp	β	-0.17	500	45	8%
Mtop	β	0.13	900	45	6.0%
Mtop	Vt	0.10	900	45	12.4mV
Mbottom	β	-0.067	500	45	13.3%
Mbottom	Vt	0.033	500	45	18.8mV
$\sigma_{V_{offset}}$:		9.6125mV			

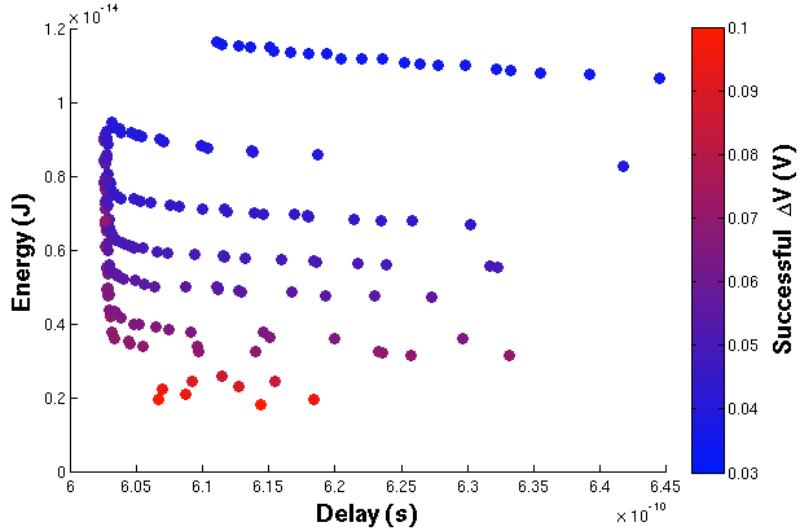
Tabel 5.3: Sensitiviteitsanalyse van de SA in het finale geheugen, er is overlap tussen de controlesignalen

om alsnog te optimaliseren naar een andere variabele. Natuurlijk is het mogelijk om één SA te gebruiken die voor elke impedantie een correcte en snelle werking zou garanderen. Dit zou echter een verspilling zijn van energie. In deze sectie wordt een pareto-oppervlak opgesteld waarbij er voor elk spanningsverschil de snelste en energieuwigste SA-ontwerpen worden gekozen.

5.4.1 Opstelling

Uit een verzameling van allerhande SA [dit zijn drain-input latch-type SA waarvan de transistoren verschillend geschaald zijn - differentiële paren hebben zelfde afmetingen] worden enkel de pareto-optimale SA uitgekozen. De pareto-criteria zijn ΔV , snelheid en dynamische energie.

Voor deze opstelling worden de passgates weggelaten van de SA [dit is geoorloofd zoals bleek uit de sensitiviteitsanalyse], de last aan de ingangs-uitgangsknopen is een simpele CMOS inverter. De knopen zijn voorgeladen op 2 spanningen: 0,4V en 0,4V - ΔV . Na 0,5ns wordt de SA aangezet en wordt de tijd gemeten tot wanneer de ingangs-uitgangsknopen geladen of ontladen zijn tot 99,9% van hun finale waarde (VDD of VSS). Dit is wellicht een te strenge methode om de snelheid van de SA



Figuur 5.12: De pareto-optimale sense amplifiers

te bepalen aangezien de inverters al eerder zullen schakelen. Indien de snelheid van de 2 knopen verschilt, zal de traagste tijd genomen worden. De dynamische energie wordt opgemeten van het moment dat de SA wordt aangeschakeld tot dit tijdstip. Ook het statisch vermogen van de SA wordt opgemeten vanaf wanneer de ingangs-uitgangsknopen VDD en VSS bereiken. Uiteraard wordt ook geverifieerd of de SA wel correct heeft gelatcht.

Per sense amplifier worden er 250 Monte Carlo simulaties uitgevoerd met deze opstelling. Indien de SA niet elke keer correct functioneerde, wordt de SA verworpen. Latchte de SA wel elke keer correct, wordt het gemiddelde van de delay, dynamische energie en statisch vermogen opgeslagen. Merk op dat de absolute waarden die opgeslagen worden niet zo nuttig zijn. Voor een praktisch ontwerp is de worst-case delay van de SA belangrijk, niet zozeer de gemiddelde. Deze gemiddelde waarden geven wel een beeld van hoe de SAs zich onderling meten.

5.4.2 Resultaten

Op figuur 5.12 zijn de pareto-optimale resultaten getoond van de groep sense amplifiers. Het doel van deze simulatie is veeleer om de transistorafmetingen te situeren in functie van deze optimalisatievariabelen. Voor deze simulatieopstelling kan men enkel zeggen dat de kans dat de offsetspanning lager is dan ΔV minstens $1 - \frac{1}{250}$ is. Dit is een veel te kleine garantie voor een sense amplifier die misschien wel miljoenen keren zal gefabriceerd worden. Voor meer informatie over de verdeling van de offsetspanning te krijgen moet de standaardafwijking berekend worden met de sensitiviteitsanalyse.

5.5 Besluit

In dit hoofdstuk werd dieper ingegaan op de sense amplifier, die het kleine spanningsverschil tussen datasignaal en referentiesignaal correct moet versterken tot VDD en VSS. De belangrijkste eigenschap van de SA is de offsetspanning welke het gevolg is van variaties op transistorafmetingen en -karakteristieken. Deze kan voldoende klein gemaakt worden door de transistoren voldoende op te schalen. De offsetspanning kan statistisch beschreven worden met behulp van een sensitiviteitsanalyse. Tenslotte worden er ook uit een grote groep SA de pareto-optimale gekozen. De resultaten geven een idee van de grootteordes van transistorafmetingen voor een bepaalde offsetspanning, snelheid en dynamische energie.

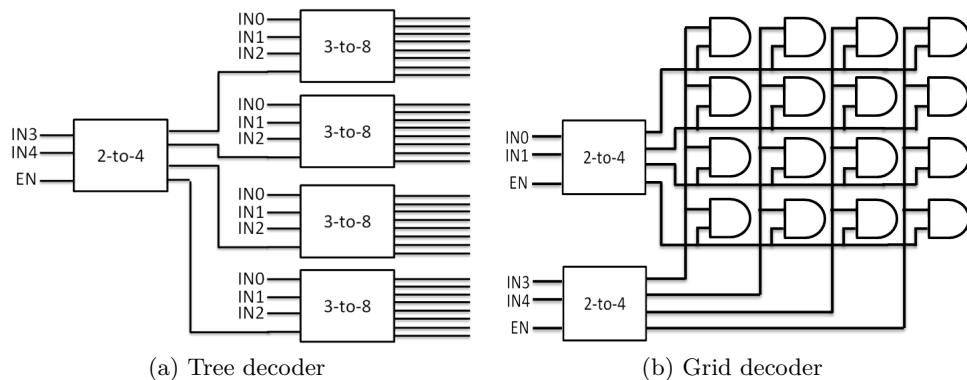
Hoofdstuk 6

Omringende logica

Het geheugensysteem maakt gebruik van bouwblokken zoals decoders, buffers en passgates. In dit hoofdstuk worden deze componenten van naderbij onderzocht.

6.1 Decoders

Een decoder is een logische schakeling: op basis van een geëncodeerde bus van ingangen brengt de decoder één uitgang actief hoog; uit een combinatie van N ingangen, gaat er steeds één van 2^N uitgangen actief hoog worden¹. Om alle mogelijke geheugenconfiguraties (aantal WLs en BLs) te kunnen exploreren in sectie 7.1.1, werd er een gamma aan decoders ontworpen gaande van een 2-naar-4 decoder tot en met een 9-naar-512 decoder. Grottere decoders worden opgebouwd uit kleinere decoders. Dit kan gedaan worden op 2 manieren; volgens een boompatroon (6.1a) of volgens een gridpatroon (6.1b). In de volgende secties wordt het ontwerp van beide manieren toegelicht en vergeleken.

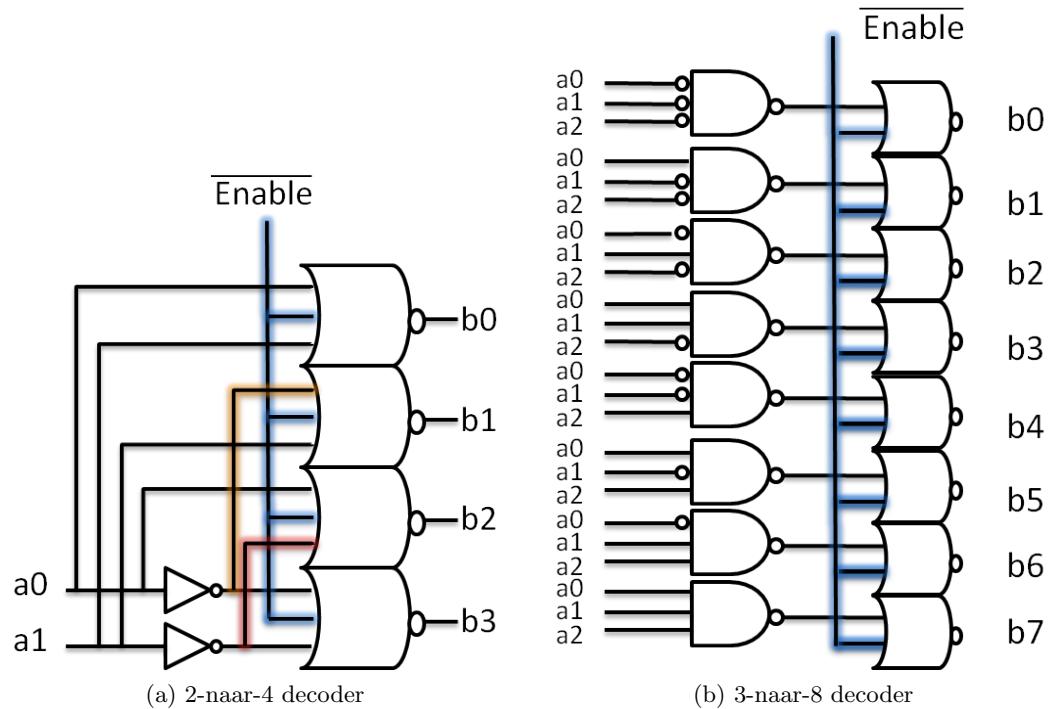


Figuur 6.1: Opbouw voor grotere decoders

¹Er is nog een enable-controlesignaal, wanneer dit actief laag is, worden alle uitgangen laag.

6.1.1 De tree decoder

De tree decoder is een decoder met een meerlaagse structuur die zich uitwaaierd naar de uitgangen. De basisblokken van deze decoder zijn een 2-naar-4 decoder (figuur 6.2a) en een 3-naar-8 decoder (figuur 6.2b). Het principe van de tree decoder is als volgt: met x -naar- 2^x en y -naar- 2^y decoders kan men een $x + y$ -naar- 2^{x+y} realiseren door ze te cascaderen. De MSBs worden aangesloten aan de eerste laag. Dit wordt geïllustreerd in de vorm van een 5-naar-32 decoder in figuur 6.1a.



Figuur 6.2: Basis decoders

6.1.2 De grid decoder

De grid decoder heeft een tweelaagse structuur. De eerste laag bestaat uit een aantal 2-naar-4 en/of 3-naar-8 decoders die in parallel staan. De verschillende uitgangen van deze eerste laag worden dan met AND-gates samen gevoegd in een tweede laag. Om glitches te voorkomen is het belangrijk dat al de signalen gelijktijdig binnen komen in de AND-gates, daarom werd de topologie van de 2-naar-4 decoder van figuur 6.2a veranderd tot een NAND-NOR topologie zoals die van de 3-naar-8 decoder in figuur 6.2b. Bij grotere decoders worden de uitgangen van de eerste laag aangesloten aan een groot aantal AND-gates. Omwille van deze grote fan-out moeten de uitgangen van de eerste laag gebufferd worden met overeenkomstig geschaalde inverters. Omdat er dus toch al minstens 2 inverters voor de AND-gates worden geplaatst, worden deze

# inputs decoder	# 2-naar-4 decoders	# 3-naar-8 decoders	# AND-gates
4	2	0	16
5	1	1	32
6	0	2	64
7	2	1	128
8	1	2	256
9	0	3	512

Tabel 6.1: Aantal gates in de grid decoder

geïmplementeerd als NOT-gate + NOR-poort i.p.v. NAND-poort + NOT-poort. Op deze manier wordt er één inverter uitgespaard. Tabel 6.1 geeft de hoeveelheid basisdecoders weer in de eerste laag van de grid decoder en het aantal AND-gates de tweede laag, in functie van het aantal inputs.

6.1.3 Vergelijkende studie

Eens ontworpen, kunnen de tree en grid decoders met elkaar vergeleken worden. Naast oppervlakte, energie en delay worden ook glitches, mismatch en delay tussen verschillende adressen onderzocht.

Zoals in figuur 6.3a gezien kan worden, schaalt de oppervlakte van de grid decoder veel minder dan die van de tree decoder bij een groot aantal ingangen. De plotse stijging in de oppervlakte van de tree decoder met 8 ingangen kan verklaard worden door het gebruik van een extra laag in de boomstructuur.

Het energieverbruik wordt vergeleken in figuur 6.3b. De grid decoder verbruikt minder energie in functie van het aantal ingangen. Sommige signalen in de tree decoder zullen diep doorrimpen, waardoor er meer gates zullen schakelen dan in de grid decoder. Het overbodig schakelgedrag van de tree decoder zou tot op zekere hoogte ingeperkt kunnen worden door de topologie van de basisdecoders (figuur 6.1) te wijzigen zodat de enable vooraan komt te staan.

De delay van de decoders kan afgelezen worden in figuur 6.3c. Beide types decoders hebben ongeveer dezelfde delay. Bij grotere grid decoders kan de extra delay verklaard worden door de extra latency van de buffers.

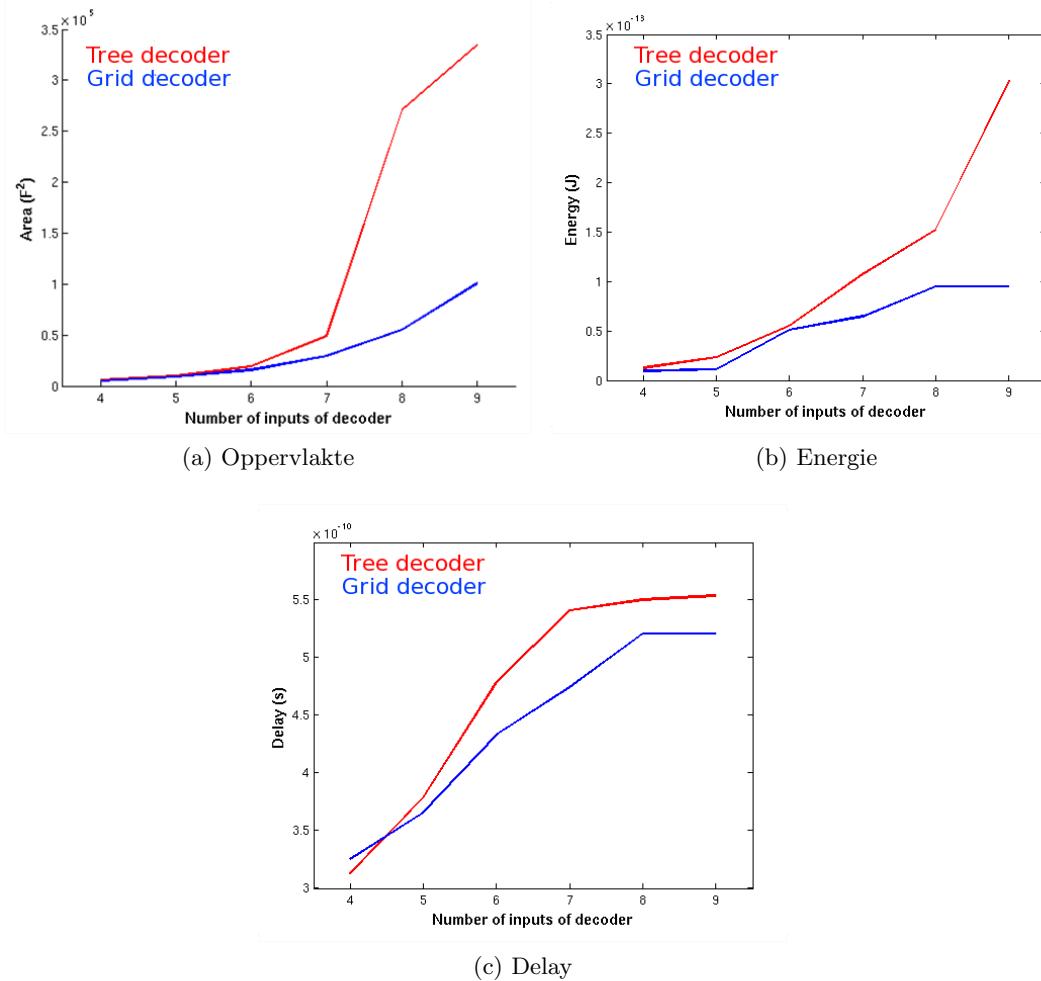
Verder werden glitches onderzocht. In beide types decoders kunnen er glitches opduiken. De oorsprong van dit probleem ligt bij het gebruik van de NOR-gates. Wanneer de twee ingangs-signalen van de NOR-gate niet gelijktijdig toekomen (zie figuur 6.4), kan de uitgang van de gate tijdelijk actief hoog getrokken worden, om vervolgens weer laag getrokken te worden. Bij de tree decoder is het nagenoeg onmogelijk om deze ongelijktijdigheid te voorkomen: sommige signalen moeten immers meer lagen doorlopen dan andere vooraleer de NOR-ingang bereikt wordt. Bij de grid decoder kan een glitch opduiken als de buffers die de tweede laag aansturen een asymmetrische delay hebben. Dit kan bijvoorbeeld voorkomen bij een 5-naar-32 decoder. De uitgangen van de 2-naar-4 decoder en de 3-naar-8-decoder in deze decoder

6. OMRINGENDE LOGICA

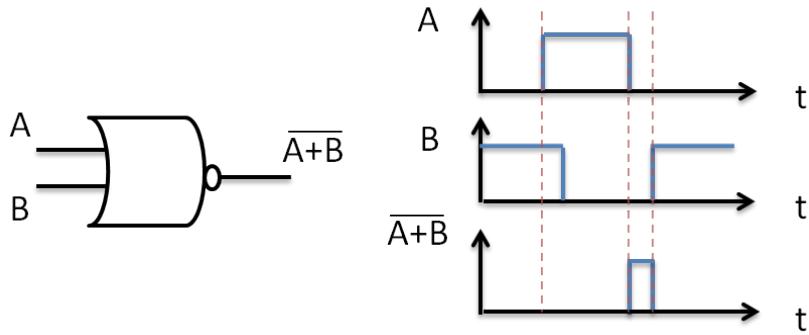
hebben een andere last. Sommige buffers kunnen echter suboptimaal ontworpen worden zodat de NOR-ingangssignalen alsnog ongeveer gelijktijdig toekomen.

Na een snelle mismatch-analyse blijkt dat de grid decoder minder variatie toont in dynamisch energieverbruik en delay dan de tree decoder. Tenslotte heeft de tree decoder een grotere spreiding wat delay betreft, afhankelijk van het vorige en huidige adres: wanneer er slechts één adresbit schakelt, kan het voorkomen dat dit signaal slechts kort moet doorimpelen tot de uitgang. Dit ziet men minder in de grid decoder.

Na het vergelijken van beide decoders wat oppervlakte, energie, delay, glitches, mismatch en delay betreft, komt de grid decoder er als beste uit en zal deze dan ook gebruikt worden in het finale ontwerp.



Figuur 6.3: Vergelijking van decoder types



Figuur 6.4: Glitch in NOR-gate

6.2 Buffers

Kleine CMOS gates kunnen slechts een beperkte hoeveelheid stroom leveren. Wanneer de uitgang van deze gates aangesloten wordt op een grote capacitieve last, duurt het lang voordat deze last op- of ontladen is. Soms is het niet aangewezen om deze gate zelf te vergroten, hierdoor neemt de intrinsieke last immers toe en gaan de gates van de vorige trap mogelijk niet meer genoeg stroom kunnen leveren om de ingang van de gate snel te sturen. In dit geval is het aangewezen om de uitgang van de gate te bufferen. Een ideale buffer heeft geen ingangscapaciteit en kan oneindig veel stroom leveren om eender welke last onmiddellijk te sturen. In de praktijk worden buffers geïmplementeerd door een even aantal invertoren te cascaderen. De eerste inverter in de ketting is klein genoeg zodat de gebufferde gate hier geen last van ondervindt, de volgende inverters in de ketting worden systematisch opgeschaald zodat de laatste inverter in de ketting voldoende stroom kan leveren om de last te sturen. Buffers worden op drie plaatsen in de geheugen architectuur gebruikt. Ten eerste om de wordlines aan te sturen. Ten tweede om de referentielogica aan te sturen en tenslotte tussen de eerste en tweede laag in de grid decoders. Tabel 6.2 geeft een overzicht van de oorsprong van de last en de waarde van de last die de verschillende buffers moeten aansturen.

De buffers werden ontworpen met de methode van logical effort[27] waarbij het aantal stages en de sizing van elke stage werd bepaald volgens het volgende stappenplan:

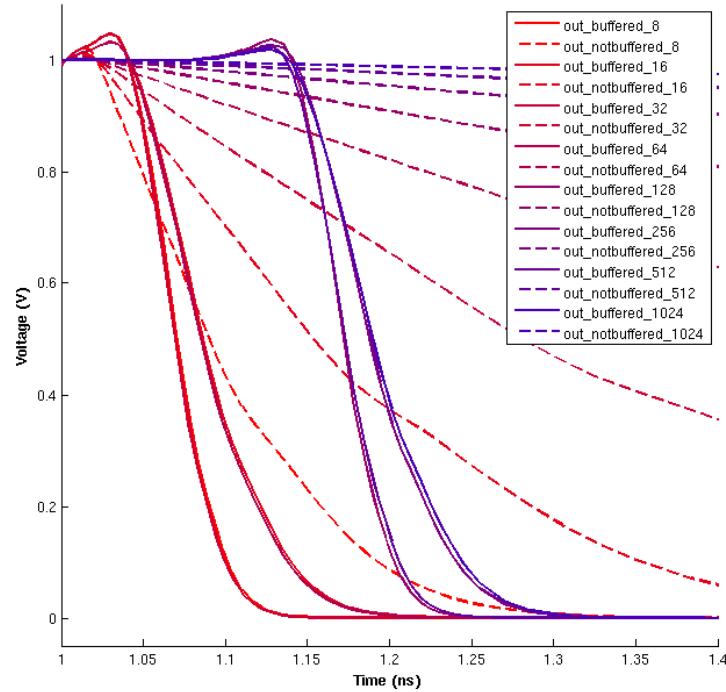
1. Bepaal de path effort $F = GH$ waarbij $G = 1$ aangezien we enkel met inverters werken en $H = \frac{C_{out}}{C_{in}}$
2. Het aantal stages wordt bepaald door $\hat{N} = \log_4 F$. Hierbij werd er voor een stage effort van 4 gekozen voor een optimale delay. \hat{N} wordt dan afgerond tot een even getal N voor de wordline- en referentiebuffers en tot een oneven getal voor de decoderbuffers.
3. N wordt dan gebruikt om een nieuwe stage effort \hat{f} te berekenen met de formule $\hat{f} = F^{1/N}$.

	Last	Capaciteit
Wordlinebuffer	1 Transistorgate	$\#BL * 0,18fF$
Referentiebuffer	1 Nor + 1 Inv	$\#BL * 0,93fF$
Decoderbuffer	1 Nor	$(4-64)^2 * 0,58fF$

Tabel 6.2: Lasten aangedreven door de verschillende buffers

4. Tenslotte kunnen de groottes van de verschillende invertoren in de chain berekend worden met de nieuwe stage effort $\hat{f} = gh$.

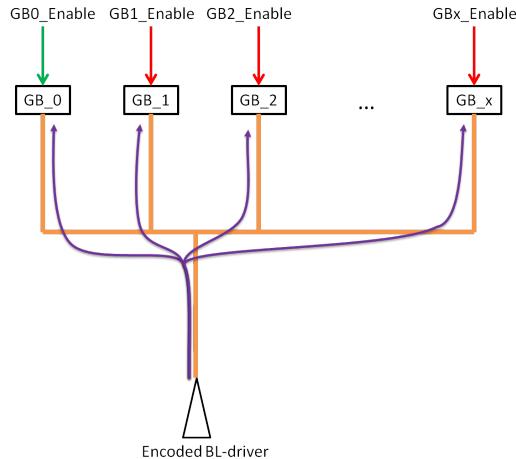
Figuur 6.5 illustert de ongebufferde en gebufferde signalen die naar de referentielogica gaan, voor verschillende veelvouden van elementaire last (een NOR-gate en een NOT-gate).



Figuur 6.5: Gebufferde en ongebufferde signalen naar de referentie logica - ongebufferde signalen ontladen traag door de zware last, buffers kunnen snel deze last ontladen/opladen. Er zit vertaging op het signaal op heel zware lasten omdat de buffer bestaat uit een groot aantal inverters

6.3 BL en WL drivers

Elk global block (GB) heeft ingangslijnen voor geëncodeerde BL- en WL-signalen. Het zou energieverlaging zijn om deze signalen naar alle GB te laten propageren, terwijl er slechts 1 GB in werking zal treden omdat het bijhorende *GB_Enable*-controlesignaal actief hoog is. Deze vertakte lijnen, die over heel de chip lopen zouden immers onnodig een aanzienlijke capaciteit opladen (zie figuur 6.6). Het is beter om deze BL- en WL-signalen tussenin nog te laten bufferen zoals op figuur 6.7. Deze drivers kunnen eenvoudig gerealiseerd worden met AND-gates in een NAND-gate + NOT-gate implementatie. De energiewinst hangt natuurlijk samen met de onwerpparameters NoGB, NoBLpLB en NoWLpB: er zijn $NoGB \cdot GB_Enable$ signalen, die elk $\log_2 NoBLpLB + \log_2 NoWLpB$ drivers moeten aansturen. Deze extra benodigde energie moet kleiner zijn dan de energie die wordt gewonnen door een groot deel capaciteit van de lijnen te vermijden. Exacte waarden voor deze capaciteiten kunnen maar worden bepaald door parasitaire layout extraction, wat in dit werk niet is uitgevoerd.

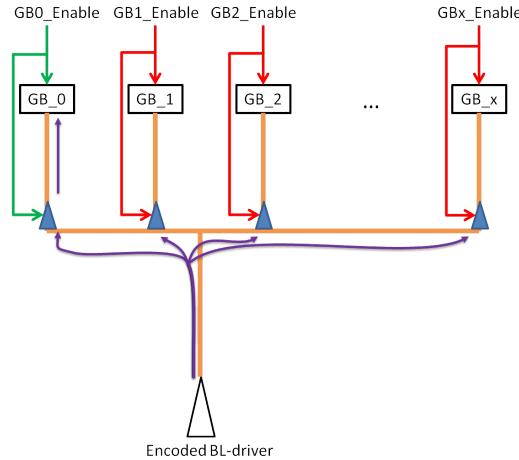


Figuur 6.6: Off-chip driver moet onnodig veel last opladen

6.4 Passgates

Passgates zijn schakelaars die de spanning van een laagimpedant knooppunt doorlaten naar een hoogimpedant knooppunt. Idealiter heeft de passgate geen weerstand wanneer hij aanstaat. In de praktijk zal er altijd een beetje weerstand zijn, dit heeft als gevolg dat er een kleine RC-delay vooraleer het hoogimpedante punt is geladen/ontladen tot de waarde van het laagimpedante punt. Een passgate kan gerealiseerd worden met een nMOS transistor, een pMOS of een combinatie van beiden. In wat volgt worden de verschillende scenario's besproken die kunnen optreden wanneer de schakelaar aangezet wordt, de passgate zal immers niet altijd stroom kunnen leveren om het hoogimpedante knooppunt te (ont)laden.

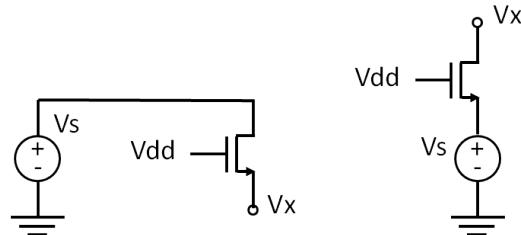
6. OMRINGENDE LOGICA



Figuur 6.7: Tussenin wordt gebufferd, zodat niet hele lijnen moeten worden opgeladen

6.4.1 nMOS passgate

De opstelling is het circuit van figuur 6.8. Afhankelijk van de (initiële) waardes van V_x en V_s , kunnen volgende situaties optreden.

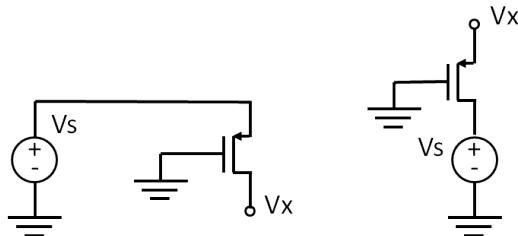


Figuur 6.8: nMOS passgate opstelling. a: $V_s > V_x$, b: $V_s < V_x$

- $V_x > V_s$ en $V_{dd} - V_s > V_{tn}$: de transistor ontladt V_x tot V_s .
- $V_x > V_s$ en $V_{dd} - V_s < V_{tn}$: de transistor staat af, er gebeurt niets.
- $V_x < V_s$ en $V_s < V_{dd} - V_{tn}$: de transistor levert stroom tot V_x is opgeladen tot V_s .
- $V_x < V_s$, $V_s > V_{dd} - V_{tn}$ en $V_x < V_{dd} - V_{tn}$: de transistor gaat V_x opladen tot $V_{dd} - V_{tn}$.
- $V_x < V_s$, $V_s > V_{dd} - V_{tn}$ en $V_x > V_{dd} - V_{tn}$: de transistor staat af, er gebeurt niets.

6.4.2 pMOS passgate

De opstelling is het circuit van figuur 6.9. Er kunnen wederom verschillende situaties optreden.



Figuur 6.9: pMOS passgate opstelling. a: $V_s > V_x$, b: $V_s < V_x$

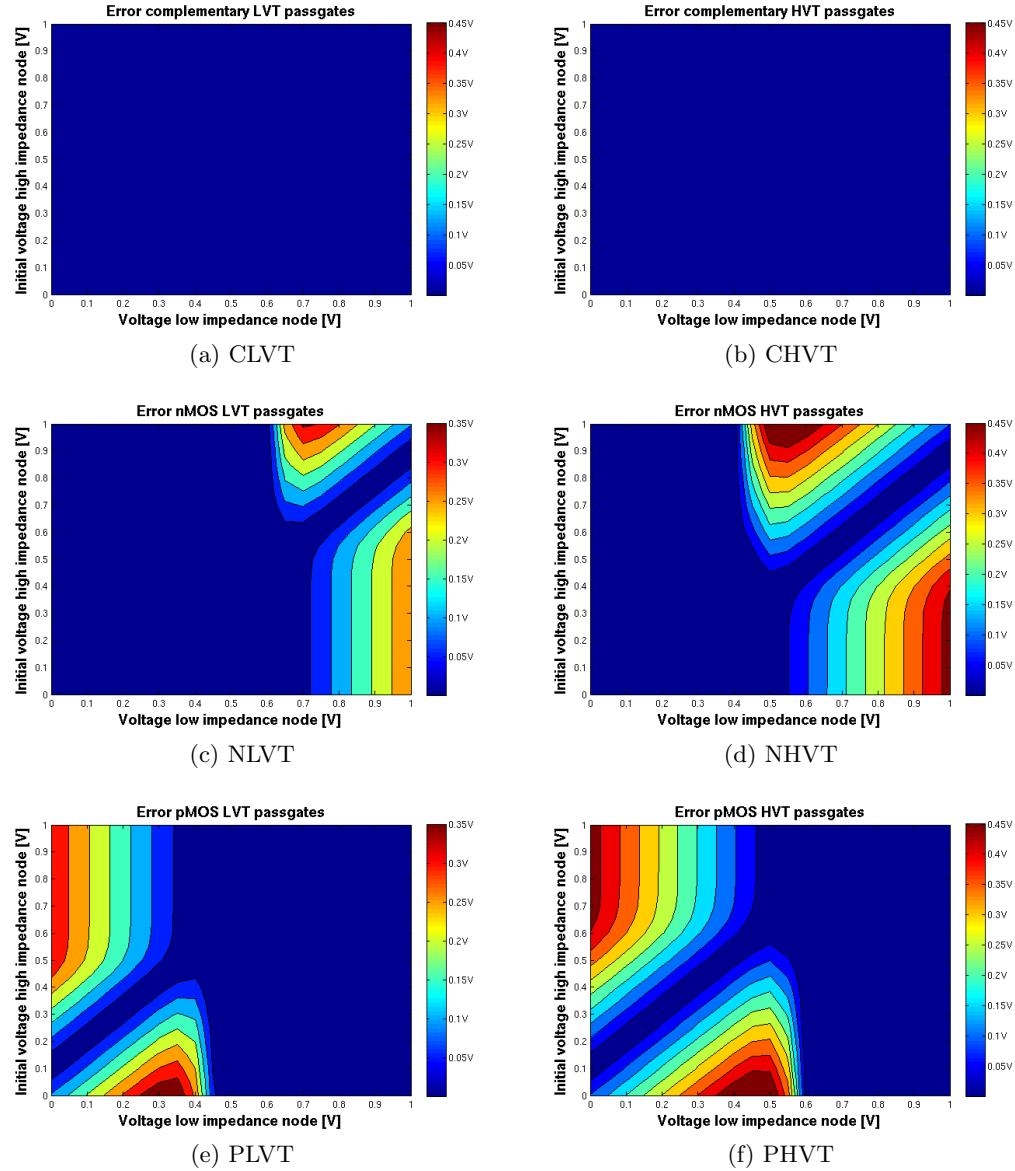
- $V_s > V_x$ en $V_s > |V_{tp}|$: de transistor laadt V_x op tot V_s .
- $V_s > V_x$ en $V_s < |V_{tp}|$: de transistor staat af, er gebeurt niets.
- $V_s < V_x$ en $V_s > |V_{tn}|$: de transistor levert stroom tot V_x is ontladen tot V_s .
- $V_s < V_x$, $V_s < |V_{tn}|$ en $V_x > |V_{tp}|$: de transistor gaat V_x ontladen tot $|V_{tp}|$.
- $V_s < V_x$, $V_s < |V_{tn}|$ en $V_x < |V_{tn}|$: de transistor staat af, er gebeurt niets.

Er zijn dus spanningszones waarvoor de passgate niet functioneert. Het is belangrijk te weten wat deze zones zijn, zodat het circuit ontworpen wordt om deze zones te vermijden. Op figuur 6.10 worden deze zones in kaart gebracht voor een nMOS, een pMOS en een complementaire passgate, zowel voor LVT als HVT transistoren. De resultaten komen voort uit een transiënte simulatie van 1 ns, de bijdrage van lekstromen is op deze korte tijd verwaarloosbaar. Er dient opgemerkt te worden dat de passgates niet (of slechts even) aanstaan voor sommige zones, maar dat de uiteindelijke fout nog meevalt. Als V_s bijvoorbeeld 0,9V bedraagt voor een nMOS passgate en V_x initieel 0,85V (met V_{dd} 1V), levert de transistor geen stroom, maar bedraagt de uiteindelijke fout 'slechts' 50mV. In deze fout zit nog geen ladingsinjectie inbegrepen, eenmaal de passgate afschakelt is de injectie onvermijdelijk. Zowel de LVT als de HVT complementaire passgates vertonen geen dode zones. Voor performantieredenen is geopteerd voor de LVT transistoren in het geheugenontwerp.

6.5 Besluit

Bij het ontwerpen van een geheugen is er nood aan een aantal logische bouwblokken: decoder, buffers, passgates. Ontwerpproblemen en -keuzes werden geanalyseerd en toegelicht. Bij de decoders werd er gekozen voor een grid architectuur om glitches te vermijden. Buffers werden ontworpen op basis van logical effort, waarbij het aantal

6. OMRINGENDE LOGICA



Figuur 6.10: Dode zones voor verschillende types passgates - de contouren stellen het verschil voor tussen V_s en V_x na 1 ns

trappen werd bepaald in functie van de last. En tenslotte bleek uit de analyse van de passgates dat er nood was aan complementaire pass-gates met lage V_t.

Hoofdstuk 7

Timing en optimalisatie

Voor een correcte werking van het geheugen, is het van belang dat de verschillende controlesignalen in een welbepaalde volgorde verwerkt en doorgegeven worden. Door al de signalen even snel te maken als het kritisch pad is er ruimte voor optimalisatie. In het eerste deel van dit hoofdstuk zal de invloed van architectuur en sizing onderzocht worden op de timing van de signalen. De constraints en vrijheidsgraden die hieruit volgen zullen dan gebruikt worden in het tweede deel van dit hoofdstuk om een optimale architectuur te bepalen.

7.1 Timing

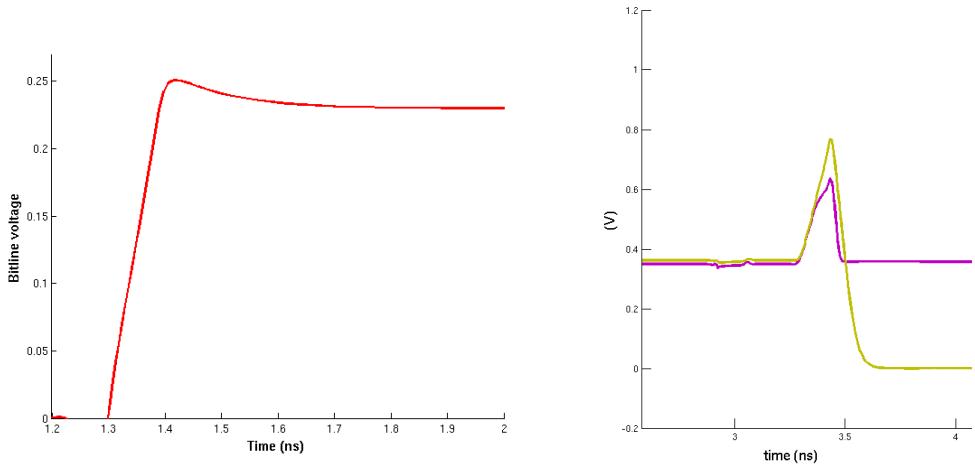
Het ontwerp van het geheugen in dit werk gaat tot het niveau van het global block (GB) 3.4, hierbij wordt de veronderstelling gemaakt dat alle signalen tegelijkertijd binnen komen in het GB. Hierna propageren de signalen door logische poorten tot ze verschillende transistoren rond de BL aansturen. De aansturing van deze transistoren leidt tot de eerste kritische timing constraints. Vervolgens worden de passgates en SA aangesloten, deze zullen de tweede timing constraints bevatten.

7.1.1 Kritische timing voor het (de)selecteren cell

Timingproblemen die het gevolg zijn van het (de)selecteren van de cel ontstaan door een verschil in timing voor (de)selecteren van de lastimpedantie en cel. Indien de load geactiveerd wordt voor de cel geactiveerd is, zal de bitline vroegtijdig beginnen opladen naar de voedingsspanning. Wanneer de cel dan geselecteerd is zal de bitline naar een spanning getrokken worden die overeen komt met de referentiespanning of een spanning van een cel in RHS of LRS. Afhankelijk van het tijdsverschill tussen deze twee gebeurtenissen, zal de bitline terug omlaag getrokken worden, wat resulteert in een energieverspilling. Dit wordt geïllustreerd in figuur 7.1a. Indien de cel gedeselecteerd wordt voordat de load gedeselecteerd is, zal de bitline ook opladen naar de voedingsspanning. Dit heeft als gevolg dat het ontladen van de bitline langer zal duren of het risico op een leesverstoring vergroot zal worden. Het overbodig opladen resulteert eveneens in een energieverspilling. De passgate die aan de BL

7. TIMING EN OPTIMALIZATIE

hangt, wordt aangestuurd door de controlelogica. Door de opbouw van deze logica, heeft de uitgangsknoop achter de passgate één inverterdelay tijd om te ontladen. Vaak is dit niet voldoende, en zal de uitgangsknoop van het LB niet volledig ontladen zijn. Dit wordt geïllustreerd in figuur 7.1b. Dit heeft geen nadelige gevolgen omdat de capaciteit op dit knooppunt heel klein is en er bijgevolg een verwaarloosbare ladingsherververdeling is in de volgende leescyclus.



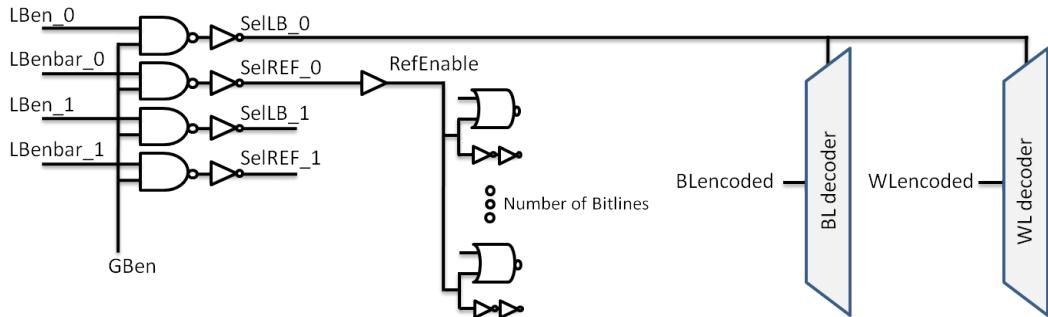
(a) Overshoot van BL omdat load geactiveerd wordt voor WL geactiveerd is

(b) Overshoot van BL omdat load degeactiveerd wordt nadat WL degeactiveerd is

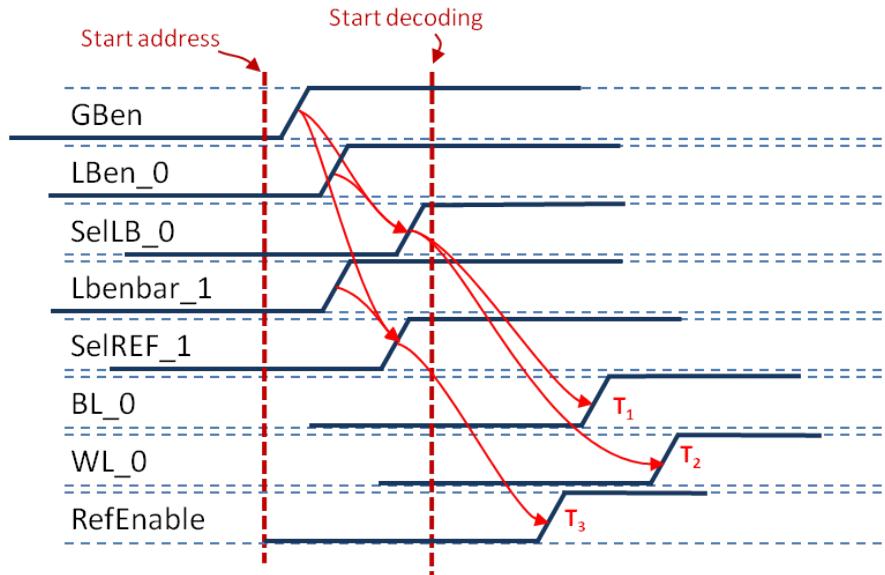
Figuur 7.1: Timingproblemen bij de bitline

Er wordt verondersteld dat alle adressignalen tegelijkertijd in het GB aankomen. De timinganalyse start dan ook vanaf dit tijdstip. Het circuit- en timingsdiagram van de logica in het GB worden geïllustreerd in figuren 7.2 en 7.3. T1 en T2 stellen de momenten voor dat de signalen uit de BL- en WL-decoder komen. T3 stelt het moment voor dat het signaal uit de referentiebuffer komt. T1 en T3 zouden op hetzelfde moment moeten aankomen om een optimale timing te hebben. Indien dit niet het geval is, zullen de referentie-BL al geactiveerd zijn voordat de data-BL opgeladen wordt. Indien er een groot aantal referentiebitlines zijn, zal dit resulteren in een grote energieverspilling. Er zijn twee opties om de correcte volgorde tussen T1, T2 en T3 te garanderen. De eerste is het kiezen van een kleine BL-decoder en een grote WL-decoder. Dit zal voor een kleine T1 zorgen door een kleine delay in de bitlinedecoder. Dit zal tevens een grotere T3 geven omdat de referentiebuffer een grotere last heeft om op te laden. Een evenwicht kan zo gevonden worden om T1 en T3 op hetzelfde moment te doen verschijnen. Deze eerste optie beperkt de mogelijke architectuur aanzienlijk en zal het verwezenlijken van andere timingconstraints voor T2 onmogelijk maken, zoals later zal blijken.

De tweede optie voor het matchen van T1 en T3 is het uitstellen van T3. Dit uitstel kan gerealiseerd worden door het invoeren van delayelementen of door de buffer suboptimaal te ontwerpen. Het invoeren van delay(vertragings)elementen zorgt in de praktijk echter meestal voor een te grote bijkomende delay. Daarom werd er in het finale ontwerp een buffer ontworpen die niet optimaal is wat snelheid betreft. Om het energieverbruik van de referentiebitlines verder in te perken werden niet al de bitlines in de array gebruikt voor het genereren van het referentiesignaal.



Figuur 7.2: Global block logica



Figuur 7.3: Timing global block

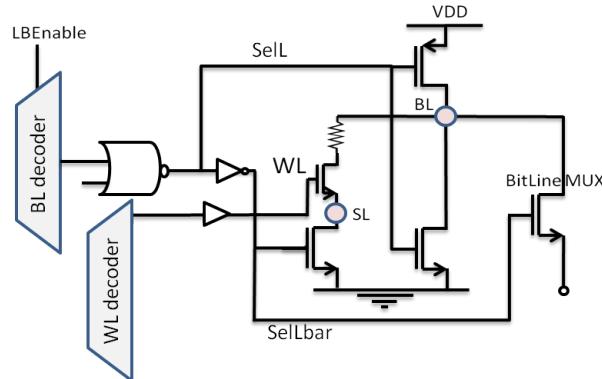
Eens de signalen uit de decoders komen worden deze gestuurd naar de controlelogica voor de memory array. Het circuit- en timingdiagram hiervan staan geïllustreerd in figuren 7.4 en 7.5. De cel zou vroeger dan of gelijk met de last moeten aangeschakeld

7. TIMING EN OPTIMALISATIE

worden. Op het timingdiagram wordt dit geïllustreerd als $T_4 = T_5 = T_6$. Door de implementatie van de logica is dit niet mogelijk aangezien er altijd een inverterdelay verschil is tussen T_4 en T_5 . Deze vertraging is minimaal en kan getolereerd worden omdat de bitline in elk geval moet opgeladen worden tot minimum V_{LRS} . Bij lage voedingsspanningen wordt dit probleem echter meer uitgesproken. T_6 wordt bepaald door WL-decoder en -buffers. Deze vertraging zou zodanig ontworpen moeten worden dat deze vroeger dan of gelijk met T_5 valt. Bij het afschakelen van de cel zijn de omgekeerde voorwaarden nodig. De last zou namelijk vroeger dan of gelijk met de cel moeten afgeschakeld worden. Deze voorwaarde is voldaan als T_7 voor T_8 en T_9 komt. Door de inverter is T_7 altijd voor T_8 . T_9 daarentegen wordt bepaald door de WL-decoder en -buffer en zou voor T_7 moeten komen.

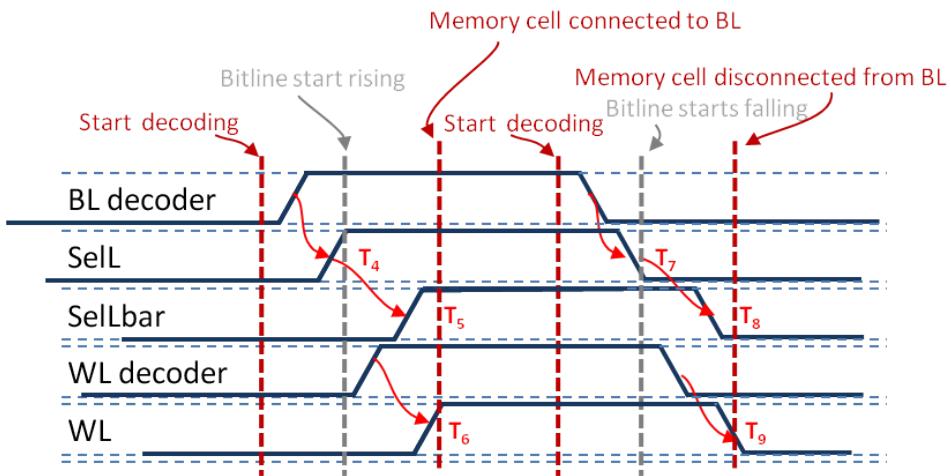
Zoals uit figuur 7.4 blijkt, zijn de meeste controlesignalen om een datacel uit te lezen niet onafhankelijk, enkel de timing van het WL-signaal kan vrij aangepast worden. Deze moet geselecteerd worden vooraleer de sourceline geselecteerd is en mag pas gedeselecteerd worden nadat de last gedeselecteerd is.

De timing van de wordline wordt expliciet bepaald door de grootte van de WL-decoder en impliciet door de grootte van de BL-decoder. De BL-decoder bepaalt namelijk de delay van de WL-buffer. Figure 7.6 geeft de delay van verschillende groottes van WL-decoders en -buffers i.f.v. verschillende groottes van BL-decoders weer.

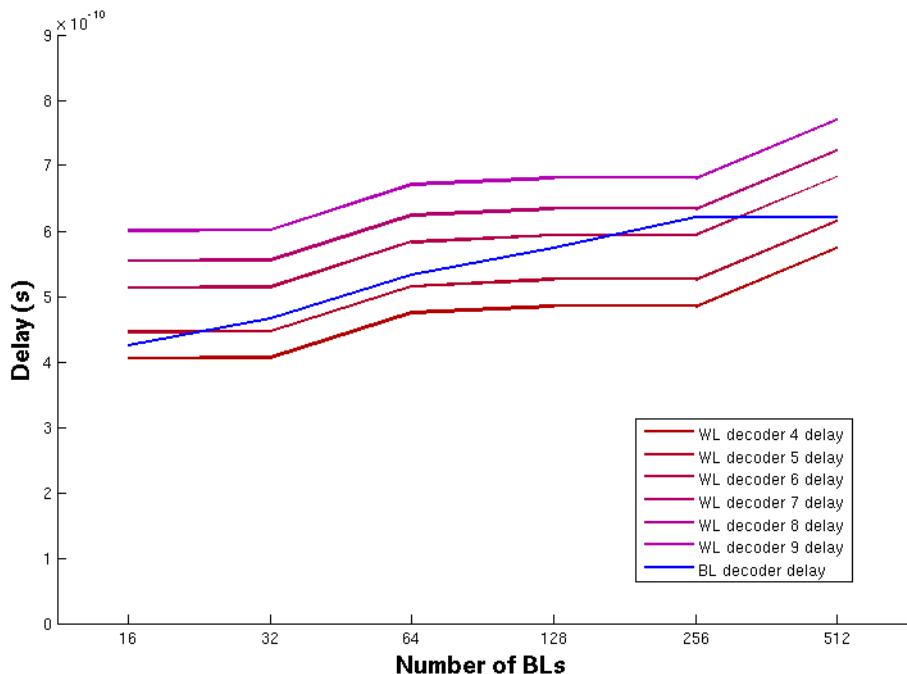


Figuur 7.4: Controlelogica data-array

De timingvoorwaarden voor het selecteren en deselecteren van de referentiecellen zijn dezelfde als die van de datacellen. Het circuit- en timingdiagramma hiervan zijn geïllustreerd in figuren 7.7 en 7.8. Anders als bij de datacellen is er aan de timingvoorwaarden al automatisch voldaan met deze logica. Dit omdat de WLs worden aangestuurd door een signaal dat rechtstreeks van de BL-decoder komt. Dit signaal wordt dan vertraagd met twee invertoren om de juiste timing te verwezenlijken.



Figuur 7.5: Timing controlelogica data-array

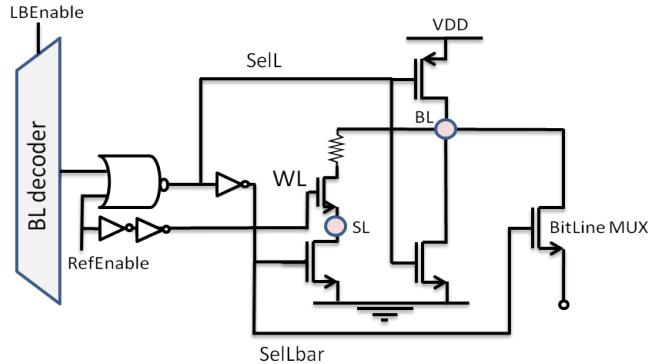


Figuur 7.6: Delay van WL-decoders en -buffers i.f.v. aantal bitlines

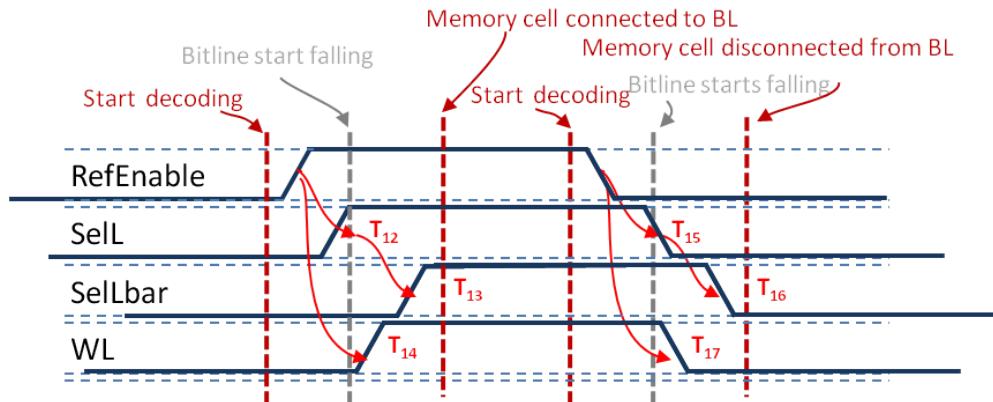
7.1.2 Kritische timing voor het uitlezen van de cel

Eens de cel geselecteerd is wordt de bitline opgeladen. De volgende stap is dit signaal te voeden aan de sense amplifier. Het signaal wordt eerst door een eerste passgate geleid om uit het local block te geraken. Vervolgens wordt het signaal door

7. TIMING EN OPTIMALISATIE



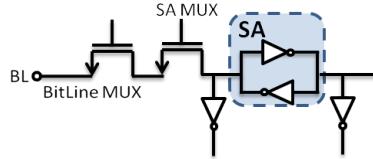
Figuur 7.7: Controlelogica referentie-array



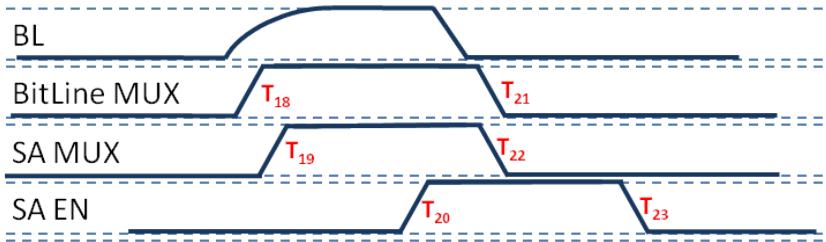
Figuur 7.8: Timing controlelogica referentie-array

een tweede passgate geleid die als sample-and-hold dient voor de sense amplifier. Figuren 7.9 en 7.10 illustreren het circuit en timing rond de sense amplifier. Eens de BL wordt aangesproken wordt de eerste passgate automatisch geactiveerd zoals uitgelegd in de vorige paragrafen. T19 stelt het tijdstip voor wanneer de tweede passgate aangezet moet worden. Deze timing is niet cruciaal: de tweede passgate mag zowel voor als na de eerste passgate geactiveerd worden. Het tijdstip waarop deze passgate wordt afgeschakeld (T22) is daarentegen wel belangrijk. Dit moet namelijk gebeuren voordat de eerste passgate afgesloten is (T21), indien niet zullen er 2 ladingsinjectie optreden i.p.v. één. Om een zo snel mogelijke latching van de sense amplifier te verkrijgen, is het tijdstip waarop de sense amplifier (T20) geactiveerd wordt belangrijk. Wanneer de sense amplifier juist wordt aangesloten treedt er het RC-latch-effect op waarbij de SA zich gedraagt alsof er geen last aan hangt. Dit effect werd beschreven in sectie 5.3.2. Na deze snelle fase, gaan de ingangs-uitgangsknopen van de SA veel trager opladen en gaat de SA de BL ook op- of ontladen. Om een snelle latching te verkrijgen moet de tweede passgate dus zo snel mogelijk na de snelle fase afgeschakeld worden. Eens de ingangs-uitgangsknopen van de SA gelatcht

zijn, mag de SA gedesactiveerd worden.



Figuur 7.9: Logica rond SA



Figuur 7.10: Timing logica rond SA

7.2 Analyse verschillende geheugencoreguraties

Het finale geheugen is 1Mbit groot. Heel wat configuraties zijn mogelijk om dit te verwezenlijken. Om deze mogelijkheden wat in te perken wordt volgende beperking opgelegd: het aantal WLs moet groter dan of gelijk zijn aan het aantal BLs. Bij deze configuraties zal het ontladen van de bitline sneller verlopen dan bij configuraties met meer BLs dan WLs. Dit levert 20 mogelijke configuraties voor NoWLpB, NoBLpLB en NoGB. Deze configuraties worden vergeleken op basis van oppervlakte, energieverbruik en leessnelheid.

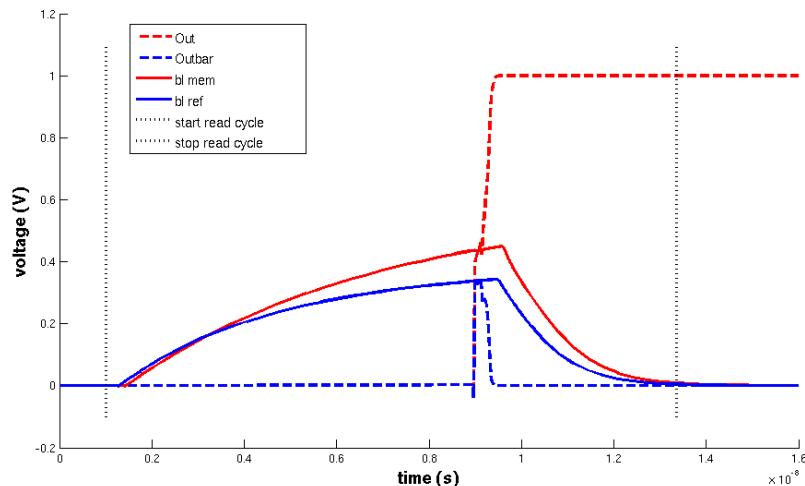
7.2.1 Evaluatie criteria voor de geheugencoreguraties

De oppervlakte wordt berekend op basis van de lengtes en breedtes van het totaal aantal transistoren (behalve de celtransistoren). Verbindingslijnen worden niet meegerekend in de berekeningen van de oppervlakte van de logica. Voor de lengte van de geheugencellen wordt $1.5*6F$ genomen en voor de breedte wordt $2*6F$ genomen [9]. Hoewel deze afmetingen voor een MTJ geheugen cell zijn, geven ze een goede schatting van de oppervlakte van een 1T1R-cel. Deze oppervlakte omvat ook de oppervlakte van bitline, wordline en sourceline meegerekend.

Het energie verbruik wordt berekend door de stroom van de voedingsspanning te integreren over de tijd en te vermenigvuldigen met de voedingsspanning. De signalen die binnen komen in een global block zijn in SPICE simulaties afkomstig van ideale spanningsbronnen. Dit heeft als gevolg dat er een ladingsinjectie optreedt naar

7. TIMING EN OPTIMALIZATIE

de voedingsspanningsbron (zie bijlage B). Dit heeft een beperkte invloed op de energieberekeningen. Deze invloed wordt evenwel niet meegenomen in de analyse. De leessnelheid is afhankelijk van de verschillende controlesignalen in de leescyclus. De simulatie-opstelling is als volgt: de leescyclus begint wanneer de signalen binnen komen in het global block. De SA wordt aangezet wanneer het verschil tussen de data- en referentiesignaal 100mV bedraagt. Omdat er niet gewacht wordt op het tijdstip wanneer de BL volledig opgeladen is, wordt het verschil in leessnelheid tussen geheugens met een klein aantal wordlines en geheugens met een groot aantal wordlines vergroot. Geheugens met iets meer wordlines worden zo in de race gehouden. In het finale geheugenontwerp zal de leessnelheid verder opgedreven worden door dit 100mV spanningsverschil te verkleinen. Verder wordt er ook altijd een HRS-cel uitgelezen aangezien deze de bitline langer moet opladen om tot aan de 100mV verschil drempel te komen wat een realistischere leessnelheid geeft. De leescyclus eindigt wanneer de BL-spanning terug naar de grond is getrokken. Figure 7.11 illustreert de hele leescyclus.



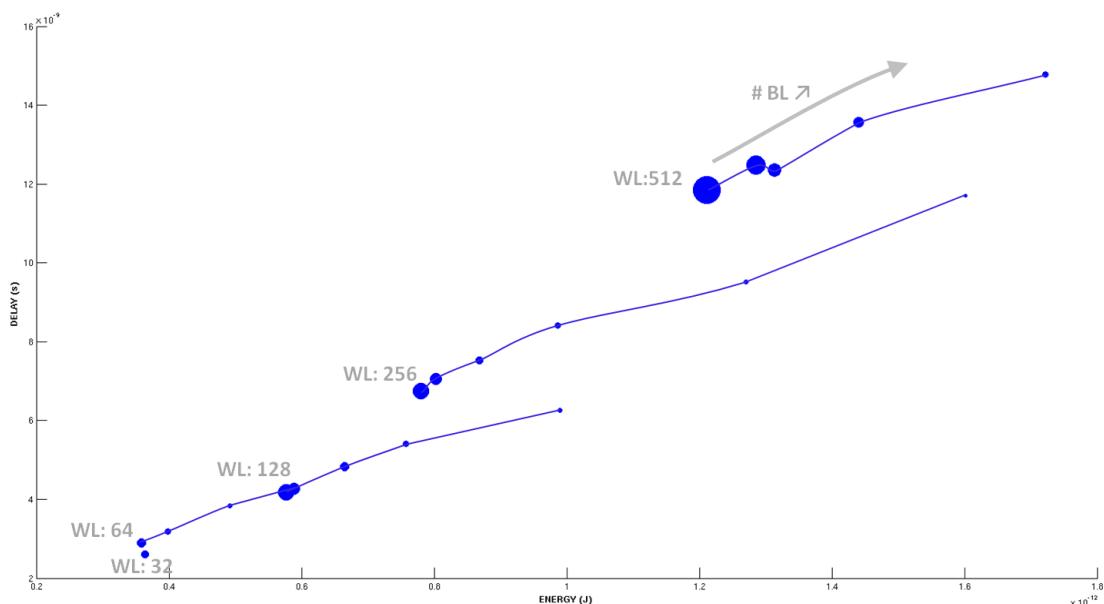
Figuur 7.11: Leescyclus

7.2.2 Vergelijking van de geheugenconfiguraties

Er werden 20 mogelijke geheugenconfiguraties geselecteerd als kandidaat voor het finale ontwerp, hun positie in de evaluatieruimte wordt getoond in figuur 7.13. Hierop staat het energieverbruik op de x-as, delay op de y-as en de oppervlakte wordt weergegeven door de grootte van de bolletjes. Het aantal effectief gebruikte referentiecellen wordt constant gehouden voor de verschillende configuraties. Dit wordt gedaan om het energieverbruik te verkleinen en omdat men maar een beperkt aantal cellen nodig heeft om een goede referentiedistributie te verkrijgen. De delay wordt voornamelijk bepaald door het opladen van de bitlines, die ook de grootste

energieverbruikers zijn. De snelheid van de bitlines wordt dan weer bepaald door het aantal wordlines, dit kan gezien worden in figuur 7.12. Het aantal bitlines beïnvloedt dan weer meer het energieverbruik. Dit extra energieverbruik gaat in de eerste plaats naar de WL-buffers, in mindere mate naar de bitline decoders en nog minder naar de bitline zelf. Bij alle geheugenconfiguraties komt het vermogenverbruik voornamelijk van de geheugecel, vervolgens in dalende lijn van de logica, de buffers en de sense amplifiers. De oppervlakte wordt bepaald door het aantal global blocks en de grootte van de decoders. Een groot aantal wordlines in combinatie met een klein aantal bitlines geeft de noodzaak aan een groot aantal global blocks, dit heeft een groot oppervlak als gevolg.

Er kan dus besloten worden dat een geheugen configuratie bestaande uit een klein aantal wordlines en evenveel bitlines een optimum zal geven voor energieverbruik en delay, maar een suboptimum voor oppervlakte.



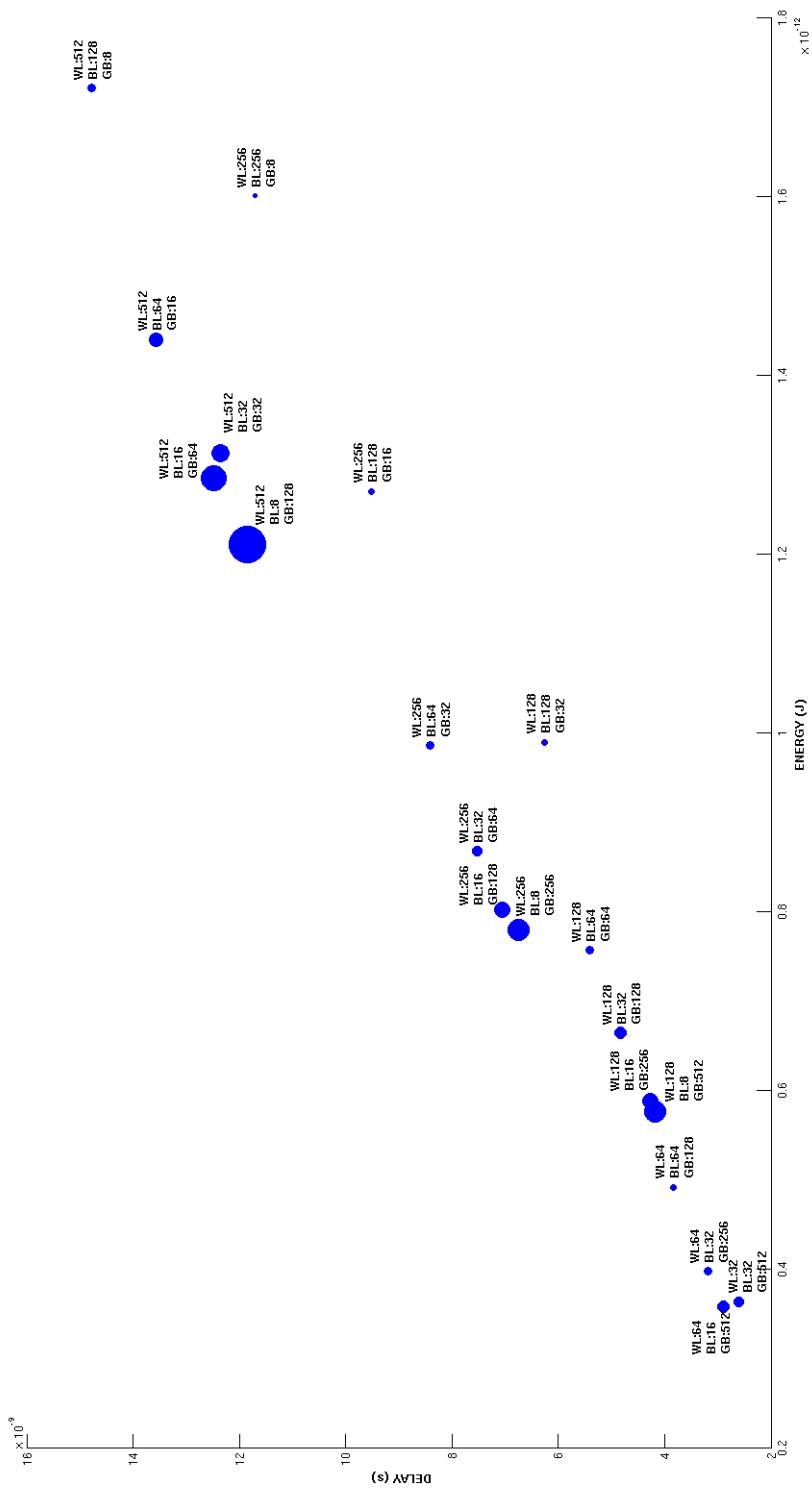
Figuur 7.12: Invloed #BL en #WL op delay, energieverbruik en oppervlakte voor verschillende geheugenconfiguraties van 1Mbit. Kleinste transistor-gate-oppervlakte is $0.009mm^2$, grootste transistor-gate-oppervlakte is $0.68mm^2$

7.3 Besluit

In dit hoofdstuk werd de timing van alle logica in de geheugenarchitectuur in kaart gebracht. Hierbij werd er gekeken naar wat de gewenste opeenvolging van signalen is en hoe dit problemen of beperkingen in de architectuur kan teweegbrengen. Vervolgens werd met deze kennis een aantal geheugenconfiguraties ontworpen en vergeleken. De conclusie is dat een kleiner aantal woordlines en bitlines een optimale

7. TIMING EN OPTIMALIZATIE

snelheid en energieverbruik voor een GB geven. Op het vlak van oppervlakte leidt dit tot een suboptimum.



Figuur 7.13: Delay, energieverbruik en oppervlakte van verschillende geheugenconfiguraties van 1Mbit. Kleinste transistor-gate-oppervlakte is 0.009mm^2 , grootste transistor-gate-oppervlakte is 0.68mm^2

Hoofdstuk 8

Volledige ontwerp

8.1 Het finaal ontwerp

Het finale ontwerp is een 1Mbit geheugen. Het bestaat uit 512 global blocks (GB) en local blocks zijn opgebouwd uit 32 BLs en 32 WLs. Op elke BL is er plaats voorzien voor één referentie cel. Van de 32 BL worden er 16 gebruikt voor het genereren van het referentiespanning. En van de 16 gebruikte cellen zijn er 6 in HRS en 10 en LRS. Dit om de referentiespanningsverdeling te centreren tussen de BL-spanningen voor cellen in HRS en LRS. De afmetingen van alle transistoren staan in tabel 8.1.

Op dit geheugen wordt een speed-vdd-test uitgevoerd. Dit is een test waarbij de voedingsspanning wordt verlaagd en er vervolgens geverifieerd wordt aan welke snelheid de leescyclus nog kan uitgevoerd worden. Hierbij wordt de dutycycle¹ spanningsdeling-latching manueel gekozen op basis van het circuitgedrag op de voedingsspanning. Dit geeft natuurlijk een meer optimistisch beeld dan dat er een kloksignaal met een bepaalde frequentie zou verwerkt worden door digitale logica om deze dutycycle te bekomen. De delay van de logica zou overigens niet lineair schalen met de voedingsspanning, waardoor de dutycycle sowieso niet constant blijft. Figuur 8.1 toont de resultaten van deze test. Voor elk vakje in de shmoo plot werden 100 Monte Carlo simulaties uitgevoerd, een groen vakje stelt 100 geslaagde leesoperaties voor, bij een rood vakje is er minstens 1 leescyclus foutief verlopen. Zoals men duidelijk kan zien daalt de leessnelheid bij het verlagen van de voedingsspanning. Dit komt door een combinatie van 2 factoren. Ten eerste wordt de logica trager, dit heeft als gevolg dat de spanningsdeling later wordt uitgevoerd na het schakelen van de controlesignalen. Bovendien komt het signaal aan de gate van de ChargeBL- en DischargeSL-transistor eerder aan dan het WL-signaal. Hierdoor verschijnt er een overshoot bij het laden van de bitline zoals in het vorig hoofdstuk werd geïllustreerd in figuur 7.1a. De overshoot treedt eerder op voor LRS-cellens omdat de BL hierbij minder lang moet opladen. Door de overshoot van de BL-spanning moet men langer wachten vooraleer de sense amplifier mag aangezet worden. De tweede factor die de leessnelheid doet vertragen is de SA zelf. Bij een voedingsspanning van 1V kan deze

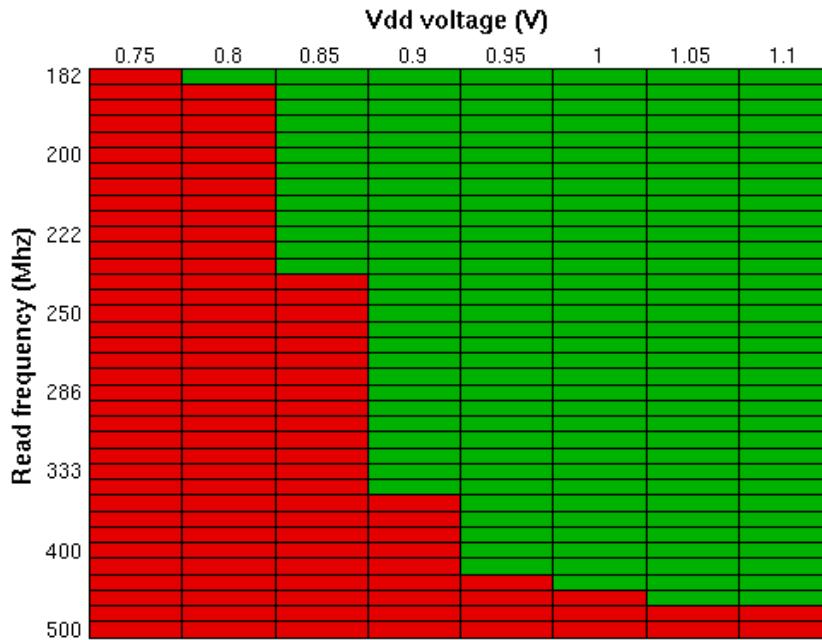
¹De verhouding van de tijd waarbij de spanningsdeling gebeurd op de totale leestijd. De totale leestijd is de som van de spanningsdelingstijd en de latchingstijd.

8. VOLLEDIGE ONTWERP

Transistor	L (nm)	W (nm)
ChargeBL	195	300
DischargeBL	45	100
DischargeSL	45	500
Sa enableP	45	900
Sa enableN	45	500
Sa P	45	1700
Sa N	45	1500
Mux LB	45	200
Mux GB	45	100

Tabel 8.1: Afbeeldingen van de transistoren in het eind ontwerp

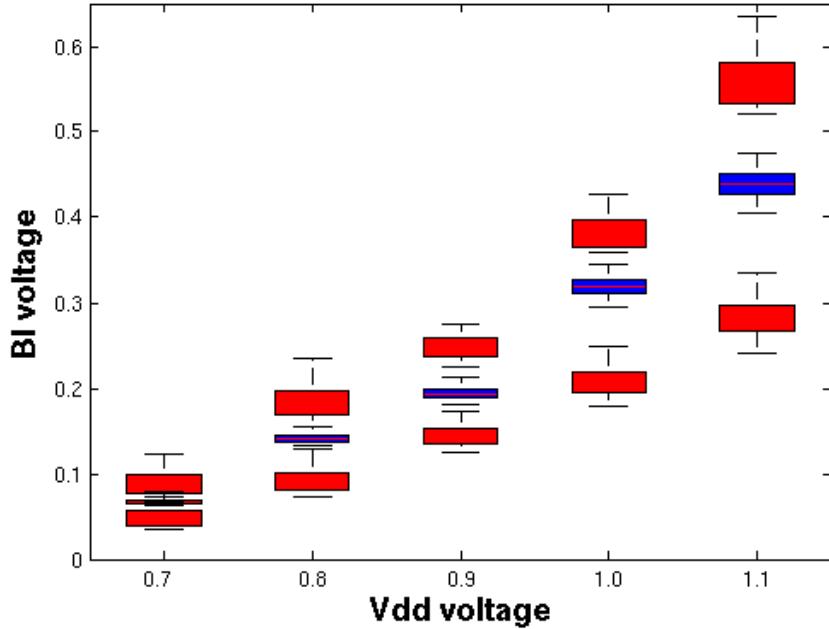
binnen de 0.25ns schakelen. Bij lagere voedingsspanningen kan de SA minder stroom trekken, dit kan in combinatie met mismatch resulteren in een latching van 2ns.



Figuur 8.1: Resultaten speed-vdd test

Verder kan men ook zien dat de schakeling een voedingsspanning hoger dan of gelijk aan 0.8V nodig heeft om correct te kunnen werken. De verklaring hiervoor kan gezien worden in figuur 8.2. Deze figuur stelt de distributie voor van de BL-spanningen van een cel in RHS, een cel in LRS en de referentiespanning in functie

van verschillende voedingsspanningen. Een duidelijke trend bij het verlagen van de voedingsspanningen is dat deze distributies dichter bij elkaar komen te liggen en dat een voedingsspanning van 0.8V wel degelijk een limiet is. Aangezien de extrema's van de distributies bij een voedingsspanning van 0.8V zo dicht bij elkaar zitten, wordt er verwacht dat de schakeling occasioneel zal falen omdat de SA ontworpen is voor een ΔV van 35mV. Het incorrect latchen is echter niet geobserveerd bij de 100 Monte Carlo simulaties. Als men naar de distributies kijkt voor een voedingsspanning van 1V zal men opmerken dat deze niet dezelfde zijn als de distributies getoond in hoofdstuk 4 (figuur 4.11). De reden hiervoor is dat men in de speed-vdd-test niet wacht tot de bitline volledig is opladen, wat een tijdwinst oplevert. Ook werd er in hoofdstuk 4 gesuggereerd dat een energiewinst zou bereikt kunnen worden door een andere last te kiezen (sectie 4.3.3). Hoewel dit mogelijk is, heeft dit wel als nadeel dat de schakeling minder tolerant is voor voedingsspanning variaties. Tenslotte moet ook vermeld worden dat de speed-vdd-test uitgevoerd werd op een (SPICE) temperatuur van 30°C. Moest deze schakeling worden geïmplementeerd in een processor is de kans groot dat dit onderhevig zal zijn aan temperaturen tussen de -40°C en 85°C wat ook tragere leessnelheden zal opleveren. Hoe traag werd echter niet onderzocht.



Figuur 8.2: Bl-spanningen voor Referentie, HRS en LRS i.f.v. Vdd

Het totale energieverbruik van een leescyclus bij een voedingsspanning van 1V is gemiddeld 0.51pJ. Hierbij gaat 25% van de energie naar de logica, 2% naar de sense amplifier, 65% naar de stroomdeling en 8% naar de buffers. Hierbij werden de decoderbuffers bij logica gerekend.

8.2 Vergelijking met de literatuur

Het vergelijken van 2 geheugens is geen evidentie. Ten eerste verschillen vaak de vooropgestelde chipspecificaties, ten tweede verschillen vaak de technologieën en tenslotte geven veel papers niet alle resultaten weer om goed te kunnen vergelijken. Chipspecificaties hangen af van de noden van de applicatie van de chip. Voor automotive applicaties bijvoorbeeld is er vooral nood aan geheugens die bij hoge temperaturen een hoge betrouwbaarheid hebben. Medische toepassingen daarentegen hebben nood aan low-power chips. Onder verschillende technologieën kan er een onderscheid gemaakt worden tussen de technologie van de logica en van het geheugen. Zo wordt er vaak voor verschillende toepassingen een andere soort NOR-flash geheugencel gebruikt: charge-trapping-cellen worden gebruikt voor betere betrouwbaarheid, split-gates-cellen voor hoge performantie [11]. De schakeling ontworpen in dit werk kan bij de snellere geheugens worden gecategoriseerd wanneer men vergelijkt met NOR-geheugens in de industrie [12][16][20]. Hierbij werd er gekeken naar de random-access-leessnelheid. Vaak kan een groot verschil in leessnelheid (gedeeltelijk) verklaard worden door de bitlinecapaciteit.

Energieverbruik kan berekend worden d.m.v CV_{vdd}^2 . Meestal wordt energieverbruik niet vermeld in papers maar gezien de lage voedingsspanning bij 45nm technologie, kan men vermoeden dat de schakeling in dit werk ook bij de meer energieuinige schakelingen hoort. De meeste schakelingen gevonden in de literatuur hebben namelijk een hogere voedingsspanning. De werking van het ontworpen RRAM geheugen op verschillende temperaturen werd in dit werk niet onderzocht. Naast verschillen in de werking van de logica zal ook de memristor onderhevig zijn aan temperatuursveranderingen. Volgens [28] zal bij een HfO_2 geheugencel de R_{OFF}/R_{ON} -verhouding dalen bij stijgende temperatuur. Ondanks deze daling in prestatie ziet men in de industrie toch RRAM-chips opduiken die functioneren bij hoge temperaturen². Men kan besluiten dat met de afbakening in het achterhoofd de schakeling een goede prestatie levert t.o.v. schakelingen in de literatuur.

8.3 Besluit

Een finale schakeling werd ontworpen en geëvalueerd. Hierbij werd er voornamelijk gekeken naar de prestatie onder verschillende voedingsspanningen. Er werd een absolute limiet van minimum 0.8V gevonden voor de voedingsspanning en verklaard. Verder werd een vergelijkende studie uitgevoerd met NOR-flash schakelingen in de literatuur en op basis hiervan werd er besloten dat de schakeling in dit werk een goed alternatief is op het vlak van snelheid en energieverbruik. Andere aspecten kunnen niet vergeleken worden aangezien deze niet onderzocht werden in dit werk.

²<http://www.crossbar-inc.com/markets/automotive.html>

Hoofdstuk 9

Besluit

In dit werk werd een RRAM leescircuit ontworpen. De 1T1R-cel die de informatie bevat bestaat uit een minimale transistor en een resistief geheugenelement, de memristor (**hoofdstuk 2**). Voor leessimulaties werden de geheugenelementen gemodelleerd als weerstanden waarvan de weerstandswaarden gebaseerd zijn op hafniumoxidememristoren.

Cellen worden in geheugenmatrices gegroepeerd (**hoofdstuk 3**), aan deze matrizes worden decoders en passgates toegevoegd die samen een local block (LB) vormen. Een local block kan aan diens uitgang zowel een datasignaal leveren als een referentiesignaal. De uitgangen van 2 LBs vormen de ingangs- en referentiespanning van een sense amplifier; de combinatie van twee LBs en een SA heet global block. Data- en referentiesignalen worden verkregen via een spanningsdeling met een bepaalde lastimpedantie en celimpedantie. Voor het datasignaal vindt er een spanningsdeling plaats op één BL, deze BL-spanning wordt naar de uitgang van het LB overgebracht door de bijhorende passgate te activeren. Voor het referentiesignaal worden er op meerdere BLs spanningsdelingen uitgevoerd, de referentiespanning wordt gevormd door de BLs kort te sluiten aan de uitgang met de passgates. Door meerdere referentiecellen te gebruiken kan de distributie van het referentiesignaal gemanipuleerd worden.

Om een zo groot mogelijk verschil tussen data- en referentiespanning te bekomen blijkt er een optimale lastimpedantie (**hoofdstuk 4**). Verschillende topologieën van impedanties zijn onderzocht naar BL-spanningsverschil, snelheid en spanningsval over geheugenelement, alsook de invloed van variabiliteit hierop. Omwille van dit laatste is het niet mogelijk om met transistoren met minimale lengtes te werken. Een enkele transistor met niet-minimale afmetingen die gebruikt wordt als schakelaar blijkt er als beste uit te komen.

De sense amplifier moet het kleine spanningsverschil tussen data- en referentiesignaal correct versterken tot de voedingsspanning (**hoofdstuk 5**). De gebruikte topologie voor de SA in dit werk is de drain-input latch-type SA. De belangrijkste eigenschap van een SA wat correcte werking betreft is de offsetspanning. De distri-

9. BESLUIT

butie hiervan wordt in kaart gebracht met sensitiviteitsanalyses. Door korte overlap tussen passgate- en SA-enable-singaal kan de spreiding van de offsetspanning kleiner gemaakt worden. Het RC-latch-effect zorgt ervoor dat de SA gedurende de snelle fase van de versterking geen invloed merkt van de grote BL-capaciteit, waaraan die blootgesteld wordt tijdens de overlap. Op basis van een lineaire sweep van transistorafmetingen werden pareto-optimale SAs bepaald wat snelheid, dynamische energie en offsetspanning betreft.

Naast lastimpedantie en SAs is er in het geheugen ook nood aan omringende logica zoals buffers, passgates, decoders,... Deze werden onderzocht in **hoofdstuk 6**.

Hoofdstuk 7 brengt de timing van alle signalen in het geheugen in kaart. Hierbij werd er gekeken welke beperkingen er opgelegd moeten worden aan de architectuur om een juiste timing te hebben. Met deze kennis worden een aantal geheugens ontworpen van 1Mbit en met elkaar vergeleken op vlak van snelheid, energieverbruik en oppervlaktegebruik.

Uiteindelijk wordt een geheugenarchitectuur met 32 woordlijnen, 32 bitlijnen en 512 global blocks gekozen als eindontwerp. Hierop wordt er een speed-vdd-test (**hoofdstuk 8**) uitgevoerd en wordt de prestatie van het circuit vergeleken met de literatuur. Men kan besluiten dat met de afbakeningen en beperkingen in het achterhoofd de schakeling een goede prestatie levert t.o.v. schakelingen in de literatuur.

Bijlagen

Bijlage A

Leon Chua's memristortheorie

In 1971 publiceerde Leon Chua, een Amerikaans onderzoeker in o.a. niet-lineaire circuittheorie, een artikel waarin hij opmerkte dat er voor de 4 fundamentele circuitvariabelen (de spanning v , stroom i , lading q en flux ϕ^1) van 6 mogelijke onderlinge relaties er slechts 5 gekend waren:

$$\begin{aligned} q(t) &= \int_{-\infty}^t i(\tau) d\tau \\ \phi(t) &= \int_{-\infty}^t v(\tau) d\tau \\ v(t) &= R * i(t) \\ q(t) &= C * v(t) \\ \phi(t) &= L * i(t) \end{aligned}$$

volgen uit de wetten van Maxwell en uit de definities van de weerstand, spoel en condensator, maar er ontbrak dus nog een relatie tussen ϕ en q [7]. Hij suggereerde dat er een 4e nog niet ontdekte passieve 2-pool moest bestaan die dit verband herbergde. Hij stelde dat $M(q) = \frac{d\phi(q)}{dq}$ met M de *memristance*. Hieruit volgt dat voor dit element $v(t) = M(q(t))i(t)$.

Indien er een lineair verband bestaat tussen ϕ en q , gedraagt dit element zich als een gewone weerstand. Enkel wanneer er een niet-lineair verband bestaat, beginnen er zich interessante fenomenen voor te doen. Zo gedraagt het element zich ogenblikkelijk als een weerstand, maar gaan deze weerstandswaarde variëren in de tijd aan de hand van de stroom die er doorgelopen heeft.

Gebaseerd op deze conclusie doopte hij deze component de memristor, een contractie van memory en resistor. Chua beëindigde zijn artikel met te erkennen dat er op dat moment nog geen fysische memristor was ontdekt, maar dat dit in de toekomst wel kon gebeuren, al dan niet zelfs per ongeluk. Hij gaf aan dat er misschien al in die tijd materialen met memristorkarakteristieken gebruikt werden, maar dat men hier over keek. Hij zou gelijk krijgen (min of meer).

¹ $\phi(t) = \int_{-\infty}^t v(\tau) d\tau$, voor een ideale inductantie is dit hetzelfde als magnetische flux

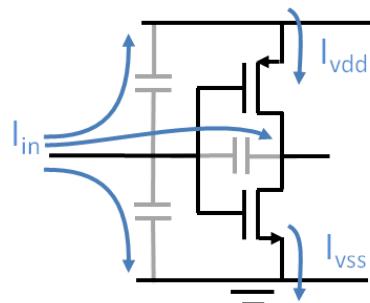
Bijlage B

Ladingsinjectie bij het gebruik van ideale SPICE bronnen.

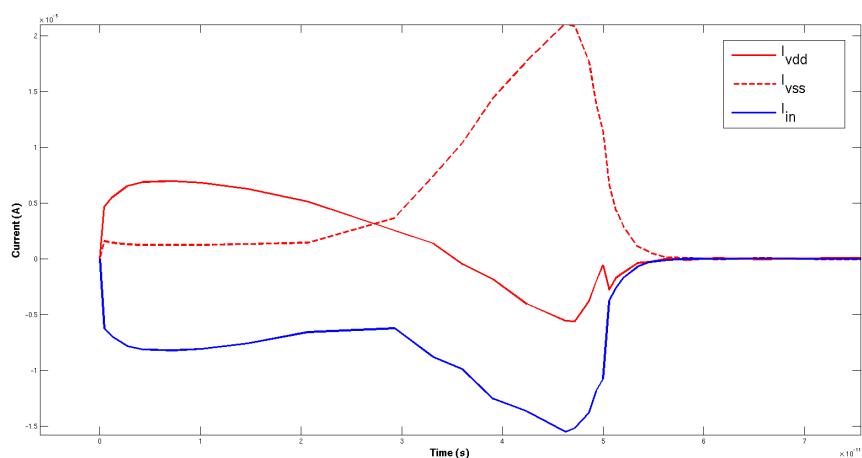
Bij het berekenen van de energie consumptie van verschillende bouwblokken, valt het op dat de voeding stroom opneemt i.p.v. levert bij het schakelen van de ideale spice bronnen. Dit komt door een ladingsinjectie van de ideale spice bronnen. Om dit te verificeren, worden de stromen van een simpele inverterschakeling bestudeerd. Figuur B.1 stelt een simple inverterschakeling voor met relevante parasitaire capaciteiten. Bij het aanleggen van een stapfunctie aan de inverter vloeit er een stroom door de capaciteiten. Omdat de positieve stroom die geobserveerd wordt in de voeding afkomstig is van de ingang, moet de som van de stromen door de ingang, voeding en grond nul zijn. De stroom door de ingang kan in SPICE opgemeten worden door een weerstand met resistieve waarde gelijk aan nul, in serie met de ingang te zetten.

Figure B.2 toont deze drie stromen. In het eerste deel van de figuur (tot tijdstip 30ps) is de spanning aan de input al aan het stijgen maar de invertor is nog niet aangeschakeld. De voedings- en grondstromen komen dan puur van de ingang. Vanaf tijdstip 30ps, is de invertor aan het schakelen en is er een aandeel van de stroom in de grond die afkomstig van de voeding is. De som van de drie stromen is ten alle tijden gelijk aan nul. We kunnen dus besluiten dat er een ladingsinjectie is van ideale spice bronnen in het circuit. Dit heeft een invloed op de stromen en daardoor ook op de energie berekeningen, maar dit verwaarlozen we bij onze berekeningen.

B. LADINGSINJECTIE BIJ HET GEBRUIK VAN IDEALE SPICE BRONNEN.



Figuur B.1: Testcircuit ladingsinjectie



Figuur B.2: Stromen in circuit

Bijlage C

IEEE Paper

Design of 1Mbit RRAM memory to replace eFlash

Diels Wouter, Standaert Alexander

Abstract—A 1Mbit RRAM memory in 45nm technology is presented. The focus lies on read reliability. To overcome variability a tuned reference signal is generated by connecting multiple reference cells in parallel. A bitline load has been designed to obtain maximum bitline voltage difference. Sense amplifier performance has been improved by allowing overlap between passgate-enable and latch-enable signals, this overlap gives rise to a nonlinear phenomenon, the RC-latch-effect. Write operation has not been included in the design and the results are based on circuit simulations.

May 20, 2014

I. INTRODUCTION

NON-volatile memories such as flash are widely used for mass storage devices, but are also steadily finding their way into the embedded domain. However, as discussed in [1], it is getting difficult to fabricate reliable flash memories in deep-submicron. It is argued that the scaling of flash-memories will not last for more than a few technology nodes. RRAM memories, in which information is stored in the resistive state of a memristor, would be able to scale further, due to the fact that it is a nanopartical based memory[2]. Furthermore, the memristor fabrication can easily be integrated in a standard CMOS fabrication process. In this work, a RRAM memory has been designed, armed against intra-die variations. A read access time of 2.3ns and an energy consumption of 0.51pJ per bit access makes it faster and less energy greedy than conventional flash memories. This paper is structured as follows: Section II discusses the general memory architecture. In section III, the tuning of the reference voltage is explained. Section IV presents the results of the load analysis, in which an optimal load impedance is chosen for sufficiently large voltage differences for the sense amplifier and sufficiently low voltage drops over the memristor. In section V, some techniques for decreasing the offset voltage of the sense amplifier will be explained. Finally, section VI summarizes the results of the complete memory.

II. GENERAL ARCHITECTURE

The general architecture can be seen in figure 1. The memory consists of 512 global blocks (GB). Each GB consists of two local blocks (LB), which in turn consists of 32 bitlines (BL) and sourcelines (SL) and 32 wordlines (WL). A branch is defined as a collection of memory cells connected to one BL and one SL.

A. Branch

In a branch 32 1T1R cells are connected to a BL and a SL. The transistor terminal of the cell is connected to the SL and the memristor terminal of the cell is connected to the BL. The memristors can be either in high or low resistive state (HRS or

LRS). Besides these 32 data cells, there is also one reference cell in the branch, its gate is connected to the reference WL. Connected to the BL is a load. The load consists out of a single pMOS transistor that is connected to the supply voltage, which is switched on and off by means of digital logic. Also connected to the BL is a nMOS transistor that serves as a switch to the ground voltage. An nMOS switch is also placed between the SL and the ground voltage. This switch is needed if a write circuit is implemented so that the current can also flow in reverse direction though the memristor.

B. Local block

A local block (LB) consists of 32 branches, BL & WL decoder and passgates on the BL of each branch (BL mux). The passgates are connected to the output node LBout. To read out a data cell, the appropriate WL is brought to the supply voltage. The BL-load and SL-transistor of the appropriate BL and SL are switched on. A current will flow from the supply voltage through the load, the cell and SL-switch to the ground voltage and a voltage will appear on the BL node. The passgate of the BL is then turned on and the BL voltage is passed to the output of the LB. Reference signals are generated by activating the reference WL and turning the BL loads and SL switches on. The BLs are then shorted by turning on all the appropriate passgates. This averages the BL voltages to a signal which has a value between the BL voltage of a cell in

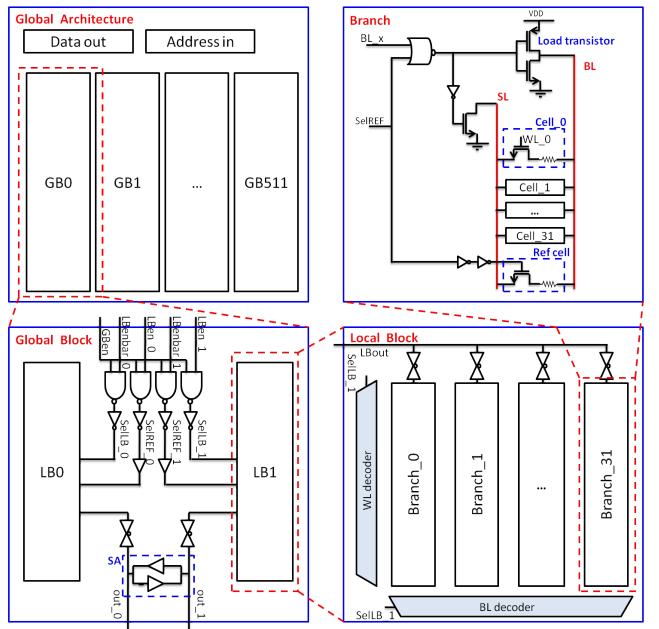


Fig. 1: Overview architecture

HRS and LRS. To save energy, not all 32 reference cells in a LB are used to generate the reference signal.

C. Global block

A global block (GB) consists out of two LBs and a sense amplifier. If one LB produces a data signal at its output, the other will produce a reference signal and vice versa. While reading a cell in the memory circuit, only one GB is active. This is guaranteed by the GBen signal.

III. TUNING THE REFERENCE SIGNAL DISTRIBUTION

Due to intra-die variations, the properties of different components in the circuit vary. In this work, variations of the transistor parameters V_T and β are considered and modeled with normal distributions. The variations of the resistive value of the memristor are also considered, normal distributions are used to model both HRS and LRS. Due to these component variations, signals such as the data and reference signal also have a distribution. The distribution of the reference signal however, can be tuned unlike the distributions of the data signal. Recall that the reference signal is generated by shorting active BLs using passgates. Shorting a BL with an addressed HRS reference cell with a BL with a LRS reference cell would suffice for producing a voltage lying between a HRS data voltage and a LS data voltage. By using this shorting technique however the mean of the reference signal PDF would not lie exactly between the means of the HRS data PDF and LRS data PDF. By implementing more than 2 reference cells for the reference signal, and having more HRS (LRS) cells than LRS (HRS) cells, the mean of the reference signal PDF can be shifted. Furthermore, the distribution will have a smaller spread by implementing a bigger amount of reference cells. One should not implement too many reference cells however, since energy consumption (for each active reference cell, current flows through its corresponding bitline) increases drastically. In this design 16 reference cells in a LB are addressed for generating the reference signal, the remaining 16 serve as dummies. Of the 16 active reference cells, 6 are HRS and 10 are LRS.

IV. LOAD ANALYSIS

Due to the aforementioned variations in the circuit, the data and reference signals should be designed in such a way that they are sufficiently far apart. The minimal distance of these two signals directly determines the maximum offset voltage the SA can have. Moreover, the distributions of these signals cannot overlap or a correct reading of the cell is impossible. The load impedance influences the value of the data and reference voltages, it also determines the settling time of the charging of the BL and the voltage drop over the memristor. Destructive reads might occur because of a too high voltage drop over the memristor. As it turns out, fast settling is not compatible with low memristor voltage drop and large voltage difference. Because the latter are imperative for a reliable memory and the former is not, settling time had no bearing on the final choice of load impedance. Several different types of load were considered [3] but the best performance with

variability was achieved with a single transistor load. Figure 2 shows nominal HRS/LRS voltage differences and memristor voltage drop for several pMOS transistor loads with different sizes.

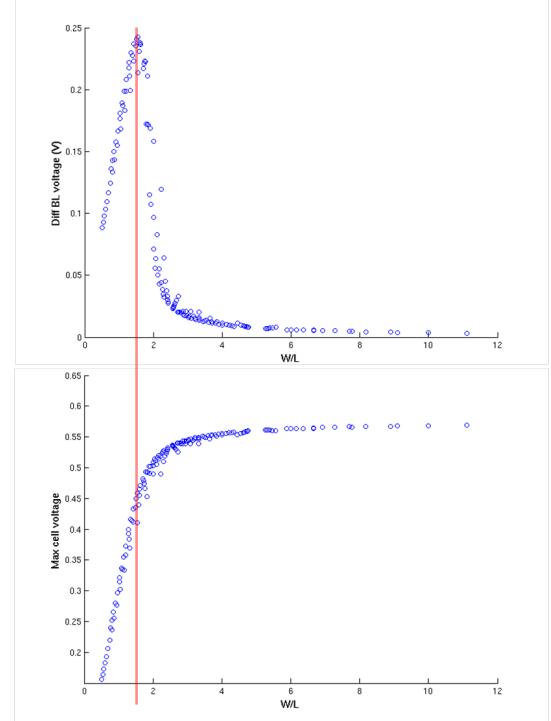


Fig. 2: Nominal HRS/LRS BL voltage difference and voltage drop over memristor for pMOS transistor loads

In the end a transistor with a width of 300nm and a length of 198nm was chosen. In figure 3 the distributions of the data and reference voltages are displayed for this load.

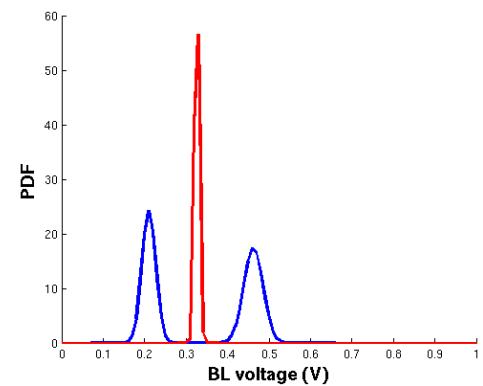


Fig. 3: Distributions of the reference signal and HRS & LRS data signals

V. SENSE AMPLIFIER OVERLAP TECHNIQUES

This design uses the drain-input latch-type SA (see figure 4). Its input/output nodes are connected to the output nodes of the local block through complementary passgates. There

are two ways to implement the latch timing cycle: one could separate the pass operation from the latch operation or could allow overlap between these two operations.

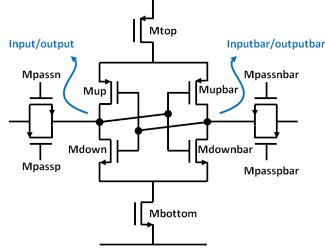


Fig. 4: Drain-input latch-type sense amplifier[4] used in this design

A. No overlap: disconnect inputs before activating SA current source

The SA can be activated after the pass-gates are disconnected. Using this control scheme, the SA would be separated from the local block when enabled. Offset voltage spread is mainly determined by the Δ_{V_T} and Δ_β variations of the differential pairs of the SA. These contributions can be decreased by increasing the sizes of the differential pair transistors. Sizing up the top and bottom transistor has more influence on the latching speed rather than the offset voltage. There is also a slightly surprising contribution on the offset voltage by the passgates. This can be explained by the charge injection of the passgates: when the passgates are turned on, the output voltage of the local block is passed on to the input/output node of the SA almost perfectly - whether there are variations on the transistors or not. When the passgates are turned off, a charge injection occurs on the SA input/output node - distorting the original voltage. The SA operates differentially though, so as long as this charge injection is matched at the two input/output nodes there would be no problem. β mismatch of the passgate transistors however results in charge injection mismatch (see figure 5). Hence the contribution of the passgates to the offset voltage spread. This mismatch can be reduced by sizing up the passgates transistors. More charge is injected when the passgates are turned off, but the distortion difference at the two sides is reduced. This reduces the offset contribution of the passtransistors.

B. Overlap: inputs are connected when SA current source is already activated

If the passgates are still turned on when the SA starts latching, the voltage difference of the input/output nodes has not experienced charge injection mismatch. When the voltage difference on the SA internal nodes has been sufficiently amplified, the passgates are turned off. Charge injection mismatch will occur, but it will not change the outcome of the latching anymore. During the overlap of the pass and latching operation, the passgates can be modeled by resistors. Neglecting the capacitance of the output node (LBout) of the

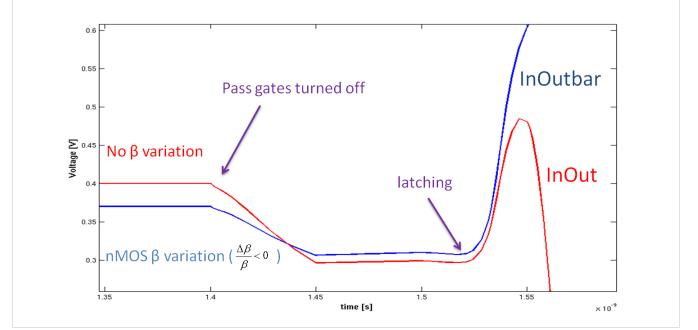


Fig. 5: Charge injection mismatch due to β variations of passgate transistors causes incorrect latching

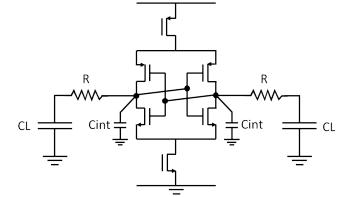


Fig. 6: Simplified circuit when overlap between pass enable and latch enable is applied

local block, the situation can be depicted as in figure 6. CL is the bitline capacitance.

One could suspect that a large BL capacitance would significantly increase the latching time. While this is true for small values of R, as can be seen in figure 7, for greater values the latching goes through two phases: during the first phase, it appears as if the big capacitance is decoupled from the SA. After this fast phase, settling is much slower.

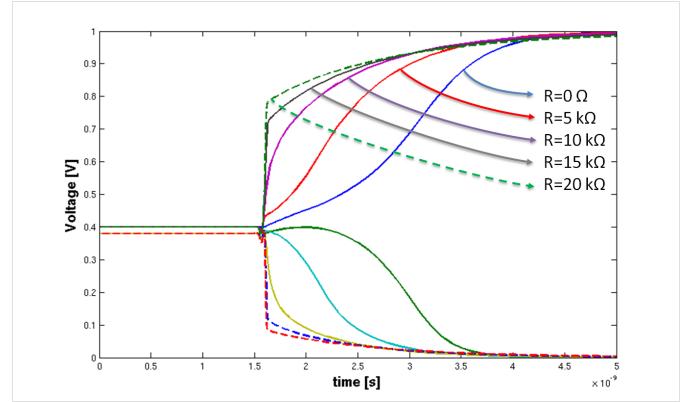


Fig. 7: Transient simulation for the schematic of figure 6 for different values of R, $CL = 46fF$, SA is minimal, no variations are included in simulation

This RC-latch effect arises when the RC-product of the pass-gate resistance and the BL capacitance is large. When it occurs, the capacitance behaves as a short-circuit and current flows through the resistor and a corresponding voltage drop over the resistor builds up. Afterwards the load capacitance charges itself at a much lower time constant.

The overlap between the enabling of the passgates and the SA thus needn't be this great: a load-less SA latching delay suffices. For this situation V_T and β mismatch of the passgate transistors results in mismatch of the resistors. The interaction between these resistors and the SA is strongly non-linear. Depending on the precharged values of the load capacitances, this R mismatch can result in incorrect latching. There is thus still a contribution of the passgates to the offset voltage spread. The interaction between R and the SA reduces the contributions of the differential pair mismatch however. For a minimal SA, the offset voltage spread is smaller with overlap than without. By sizing all the transistors, this spread can be reduced to the desired level. Speed can be improved by using the RC-latch effect which implies using small pass-gate transistors, or by sizing the rest of the SA. This allows the SA to deliver a large current to charge the BL capacitance.

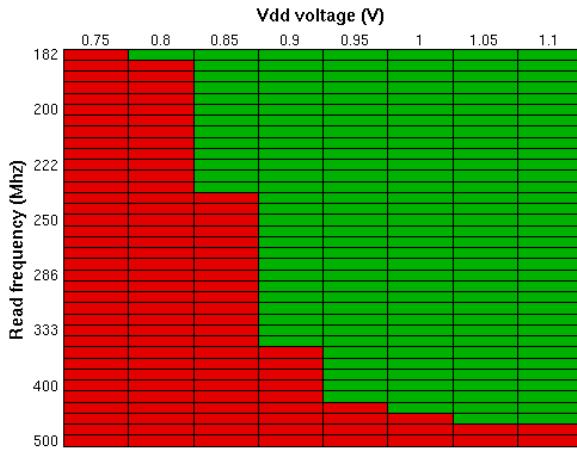


Fig. 8: Results read access time test. Green successful simulation. Red unsuccessful simulation

On figure 8 it can clearly be seen that the circuit can only operate down to a supply voltage of $0.8V$. An explanation for this is given in figure 9. As vdd is decreased, the bitline voltage distributions of the reference signal and HRS & LRS data signals also decrease. Eventually the bitline voltage distributions overlap and there is a high probability of circuit failure.

VI. READ THROUGHPUT RESULTS

A read access time simulation test was performed on the complete circuit (Figure 8). In this test the supply voltage was decreased and the maximum read throughput was determined. The test was performed at a (SPICE environment) temperature of 30°C . At each point in the shmoo plot, 100 Monte Carlo simulations were performed. At 1V Vdd, the circuit achieves a read access time of 2.3ns . during the read cycle, its energy consumption is 0.51pJ . Most of the energy (65%) is consumed by the bitlines. 25% of the energy consumption is due to the logic, while the buffers and SA together take the remaining 10%. As the supply voltage goes down, the read throughput also goes

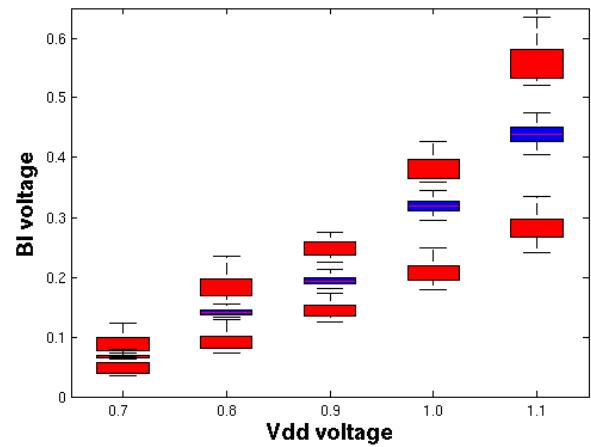


Fig. 9: BL voltage distribution for reference signal and HRS & LRS signals in function of vdd

down. This is caused by a combination of phenomena. First of all the logic operates slower. This causes a timing issue in which the bitline load transistor is activated before the cell. This timing issue results in a rapid increase in bitline voltage which causes a longer settling time of the bitline when the cell is eventually selected. A second reason for the decrease in read throughput at a lower supply voltage, is the slower latching of the SA.

VII. CONCLUSION

A 1Mb RRAM memory has been designed and presented. This memory design employs several techniques to improve performance and reliability. A good reference voltage distribution was achieved by averaging the signal from 10LRS and 6HRS cells. A load impedance was carefully chosen to maximize HRS & LRS voltage differences while keeping the memristor voltage drop sufficiently low to avoid resistive switching and thus destructive reading. Overlap between passgates and sense amplifier operation reduces offset voltage spread as it avoids differential charge injection. Although keeping the passgates enabled while activating the SA does mean current flows from SA to BL, the impact of this on SA speed and energy is small, as $R_{passgate} \cdot C_{BL}$ is much larger than the overlap time.

REFERENCES

- [1] K. Prall and K. Parat, "25nm 64gb mlc nand technology and scaling challenges invited paper," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, Dec 2010, pp. 5.2.1–5.2.4.
- [2] H. Nazarian, "Crossbar resistive memory: The future technology for nand flash," *white paper from crossbar*, 2013.
- [3] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. Yang, and D. Markovic, "A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer rams (stt-rams)," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, March 2012, pp. 275–282.
- [4] S. Cosemans, "Variability-aware design of low power sram memories," Ph.D. dissertation, KULeuven, 2010.

Bibliografie

- [1] I. Baek, M. Lee, S. Seo, M.-J. Lee, D. Seo, D. S. Suh, J. Park, S. Park, T. Kim, I. Yoo, U.-i. Chung, and J. Moon. Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses. In *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, pages 587–590, Dec 2004.
- [2] A. Beck, J. G. Bednorz, C. Gerber, C. Rossel, and D. Widmer. Reproducible switching effect in thin oxide films for memory applications. *Applied Physics Letters*, 77(1):139–141, 2000.
- [3] W.-Y. Chang, Y.-C. Lai, T.-B. Wu, S.-F. Wang, F. Chen, and M.-J. Tsai. Unipolar resistive switching characteristics of zno thin films for nonvolatile memory applications. *Applied Physics Letters*, 92(2), 2008.
- [4] Y.-C. Chen, C. Chen, C. T. Chen, J. Y. Yu, S. Wu, S. L. Lung, R. Liu, and C.-Y. Lu. An access-transistor-free (0t/1r) non-volatile resistance random access memory (rram) using a novel threshold switching, self-rectifying chalcogenide device. In *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International*, pages 37.4.1–37.4.4, Dec 2003.
- [5] Y.-Y. Chen, L. Goux, L. Pantisano, and J. Swerts. Fully cmos beol compatible hfo₂ rram cell, with low program current, strong retention and high scalability, using an optimized plasma enhanced atomic layer deposition (peald) process for tin electrode. In *Interconnect Technology Conference and 2011 Materials for Advanced Metallization (IITC/MAM), 2011 IEEE International*, pages 1–3, May 2011.
- [6] C.-T. Cheng, Y.-C. Tsai, and K.-H. Cheng. A high-speed current mode sense amplifier for spin-torque transfer magnetic random access memory. In *Circuits and Systems (MWSCAS), 2010 53rd IEEE International Midwest Symposium on*, pages 181–184, Aug 2010.
- [7] L. . CHUA. Memristor-the missing circuit element. *IEEE Transactions of circuit theory*, September 1971.
- [8] S. Cosemans. *Variability-aware design of low power SRAM memories*. PhD thesis, KULeuven, 2010.

BIBLIOGRAFIE

- [9] S. Cosemans. Intro regarding read sensing schemes. powerpoint presentation, 2013.
- [10] Y. Deng, P. Huang, B. Chen, X. Yang, B. Gao, J. Wang, L. Zeng, G. Du, J. Kang, and X. Liu. Rram crossbar array with cell selection device: A device and circuit interaction study. *Electron Devices, IEEE Transactions on*, 60(2):719–726, Feb 2013.
- [11] H. Hidaka. Evolution of embedded flash memory technology for mcu. In *IC Design Technology (ICICDT), 2011 IEEE International Conference on*, pages 1–4, May 2011.
- [12] M. Jefremow, T. Kern, U. Backhausen, J. Elbs, B. Rousseau, C. Roll, L. Castro, T. Roehr, E. Paparisto, K. Herfurth, R. Bartenschlager, S. Thierold, R. Renardy, S. Kassenetter, N. Lawal, M. Strasser, W. Trottmann, and D. Schmitt-Landsiedel. A 65nm 4mb embedded flash macro for automotive achieving a read throughput of 5.7gb/s and a write throughput of 1.4mb/s. In *ESSCIRC (ESSCIRC), 2013 Proceedings of the*, pages 193–196, Sept 2013.
- [13] K. M. Kim, B. J. Choi, B. W. Koo, S. Choi, D. S. Jeong, and C. S. Hwang. Resistive switching in pt/al₂o₃/tio₂/ru stacked structures. *Electrochem. Solid State Lett.*, 9G343–G346, 2006.
- [14] P. J. Kuekes, D. R. Stewart, and R. S. Williams. The crossbar latch: Logic value storage, restoration, and inversion in crossbar circuits. *Journal of Applied Physics*, 97(3):–, 2005.
- [15] K. J. Kuhn. Variation in 45nm and implications for 32nm and beyond. powerpoint presentation, 2009.
- [16] C. Lee, S. H. Baek, and K.-H. Park. A hybrid flash file system based on nor and nand flash memories for embedded devices. *Computers, IEEE Transactions on*, 57(7):1002–1008, July 2008.
- [17] L. Liu, Y. Hou, D. Yu, B. Chen, B. Gao, Y. Tian, D. Han, Y. Wang, J. Kang, and X. Zhang. Multilevel set/reset switching characteristics in al/ceox/pt rram devices. In *Electron Devices and Solid State Circuit (EDSSC), 2012 IEEE International Conference on*, pages 1–3, Dec 2012.
- [18] G. E. MOORE. Cramming more components onto integrated circuits. *Electronics*, April 1965.
- [19] K. Prall and K. Parat. 25nm 64gb mlc nand technology and scaling challenges invited paper. In *Electron Devices Meeting (IEDM), 2010 IEEE International*, pages 5.2.1–5.2.4, Dec 2010.
- [20] P. Pulici, T. Lessio, A. Vigilante, G. Vanalli, P. Stoppino, G. Ripamonti, A. Losavio, and G. Campardo. 1.2v nor flash memory in system-in-package. *Electronics Letters*, 42(23):1334–1335, November 2006.

- [21] T. Raja and S. Mourad. Digital logic implementation in memristor-based crossbars. In *Communications, Circuits and Systems, 2009. ICCCAS 2009. International Conference on*, pages 939–943, July 2009.
- [22] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. Yang, and D. Markovic. A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer rams (stt-rams). In *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, pages 275–282, March 2012.
- [23] G. Rose, J. Rajendran, H. Manem, R. Karri, and R. Pino. Leveraging memristive systems in the construction of digital logic circuits. *Proceedings of the IEEE*, 100(6):2033–2049, June 2012.
- [24] J. Simmons and R. R. Verderber. New thin-film resistive memory. *Radio and Electronic Engineer*, 34(2):81–89, August 1967.
- [25] R. D. Stefan Cosemans. Verilog-a implementation of hourglass model. powerpoint presentation, 2012.
- [26] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. The missing memristor found. *Nature*, 453(7191):80–83, 2008.
- [27] I. Sutherland, B. Sproull, and D. Harris. *Logical Effort: Designing Fast CMOS Circuits*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.
- [28] C. Walczyk, D. Walczyk, T. Schroeder, T. Bertaude, M. Sowinska, M. Lukosius, M. Fraschke, D. Wolansky, B. Tillack, E. Miranda, and C. Wenger. Impact of temperature on the resistive switching behavior of embedded hfo₂ -based rram devices. *Electron Devices, IEEE Transactions on*, 58(9):3124–3131, Sept 2011.
- [29] H. S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. Chen, and M.-J. Tsai. Metal oxide rram. *Proceedings of the IEEE*, 100(6):1951–1970, June 2012.
- [30] D. Wouters. Oxide resistive ram (oxrram) for scaled nvm application. powerpoint presentation, 2009.
- [31] C. Xu, X. Dong, N. Jouppi, and Y. Xie. Design implications of memristor-based rram cross-point structures. In *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*, pages 1–6, March 2011.
- [32] W. Zhuang, W. Pan, B. D. Ulrich, J. Lee, L. Stecker, A. Burmaster, D. Evans, S. Hsu, M. Tajiri, A. Shimaoka, K. Inoue, T. Naka, N. Awaya, A. Sakiyama, Y. Wang, S. Liu, N. Wu, and A. Ignatiev. Novel colossal magnetoresistive thin film nonvolatile resistance random access memory (rram). In *Electron Devices Meeting, 2002. IEDM '02. International*, pages 193–196, Dec 2002.
- [33] M. A. Zidan, H. A. H. Fahmy, M. M. Hussain, and K. N. Salama. Memristor-based memory: The sneak paths problem and solutions. *Microelectronics Journal*, 44(2):176 – 183, 2013.

Fiche masterproef

Studenten: Wouter Diels
Alexander Standaert

Titel: Ontwerp van een RRAM geheugen voor ingebetde NV toepassingen

Engelse titel: Design of a RRAM memory for embedded NV applications

UDC: 621.3

Korte inhoud:

RRAM is een veelbelovende technologie voor het maken van embedded NV geheugens. Deze thesis behandelt het ontwerp van het leescircuit van een 1Mbit RRAM geheugen. Er wordt een geheugenarchitectuur ontworpen dat opgebouwt is uit cellen, branches, local blocks en global blocks. Hierin wordt er een uitgebreide analyse gedaan op het ontwerp en keuze van componenten zoals lastimpedantie en sense amplifier. Nadat omringende logica zoals buffers en decoders ontworpen zijn, wordt er onderzocht wat de optimale geheugenconfiguratie is. Deze optimale configuratie wordt dan onderworpen aan een speed-vdd-test en vergeleken met schakelingen in de literatuur.

Thesis voorgedragen tot het behalen van de graad van Master of Science in de ingenieurswetenschappen: elektrotechniek, optie Elektronica en geïntegreerde schakelingen

Promotor: Prof. dr. ir. W. Dehaene

Assessoren: Prof. dr. ir. R. Lauwereins
Prof. dr. ir. M. Verhelst

Begeleiders: ir. B. Baran
dr. ir. S. Cosemans