KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Reliability characterization of STT-MRAM magnetic memory

The impact of self-heating

Simon Van Beek

Supervisor:
Prof. dr. ir. Guido Groeseneken

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

August 2018

Reliability characterization of STT-MRAM magnetic memory

The impact of self-heating

Simon VAN BEEK

Examination committee:
Prof. dr. ir. Carlo Vandecasteele, chair
Prof. dr. ir. Guido Groeseneken, supervisor
Prof. dr. ir. Jo De Boeck
Prof. dr. ir. Kristin De Meyer
Prof. dr. ir. Sofie Pollin
Dr. ir. Koen Martens

Prof. dr. ir. Dirk Wouters
(RWTH Aachen)
Prof. dr. Thibaut Devolder
(Université Paris-Sud)
Dr. ir. Robin Degraeve
(imec)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

August 2018

© 2018 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Simon Van Beek, Kasteelpark arenberg 10, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Acknowledgements

First and foremost, I am grateful to Prof. Guido Groeseneken and Prof. Dirk Wouters for giving me the opportunity to pursue this PhD at imec. I appreciate the guidance and support throughout this thesis. I would also like to acknowledge all the members of the jury. Thank you for your interest in this thesis, your time to carefully examine the manuscript and your constructive comments and suggestions.

Exceptional thanks are given to Koen Martens, my daily supervisor during the first years of my PhD. His down-to-earth criticism and exceptional knowledge, raised the bar time and time again. Many thanks for reading and improving my abstracts and publications. A particular thanks goes to Koen Martens and Sven Cornelissen, who have helped me extensively writing my IWT funding proposal.

I am extremely grateful to Robin Degraeve, who won the lottery and was appointed to be my new daily supervisor. Next to all the other PhD students you are supervising, you managed to free up valuable time to read and improve this manuscript. Thank you for many fruitful discussions about the breakdown model.

I would also like to thank Philippe Roussel. Without a doubt you are imec's biggest mathematical treasure, although I can only say this with 95 % confidence. Thanks for your endless help with all the statistics, which was of immeasurable importance to this work.

My presentations would not have been the same without Erik Bury. Thanks for sitting next to me and for the countless last minute adjustments to my conference, PTW and even my PhD defence presentation. "Ik ben blij dat

gij in mijn team zit". Which brings me to the DRE-team, such a diverse and complementary team, which I am very grateful to be part of. Thank you all for the help, discussions, team events, lunches ...

And then there are PhD-students, too many to name. I am lucky imec has such a fantastic PhD community. Thanks for sharing struggles and frustrations, celebrating achievements or drinking beers at conferences or after work receptions. Special thanks for the post-conference trips to Marko, Alexandre, Adrian, Gerhard, Vamsi, Kent, Roman, Sofie, Bart, Eline, Olivier.

Karen Levrie, thanks for booking my agenda on Tuesday and Thursday for a noon run, with a complementary post-run lunch. You were always one step ahead and the ideal person I could shoot all my administrative questions at.

Frank Gijbels, thank you for your support in the electronic workplace. You always seemed to be able to squeeze in your agenda some urgent request of mine. I am also grateful to the amsimec team, allowing me to tune and tailor the experimental setups.

Ook zou ik graag mijn familie en vrienden willen bedanken. Hun onrechtstreekse bijdrage aan dit werk is onbeschrijfbaar. Hoe gemakkelijk is het om de dagelijkse worstelingen en up's and down's te vergeten tijdens een paasweekend, citytrip, roadtrip, zwemmeke, gastronomisch diner... Graag wil ik ook mijn mama bedanken voor al haar tijd en geduld met mij. Mijn gedachten zijn ook bij mijn papa, ik weet hoe trots je zou zijn dat ik in je voetsporen treed.

En last but not least wil ik mijn principessa Lauren bedanken. Bedankt voor je oprechte interesse en ondersteuning, onze mooie momenten en reizen, en ook om wakker en geinteresseerd te blijven tijdens de rit naar de kust, toen ik je mijn doctoraat probeerde uit te leggen. Je bent mijn grootste fan, ik zie je graag.

Simon
August 2018

Abstract

In this thesis, the impact of self-heating on the reliability of the emerging STT-MRAM memory technology is studied. Memory devices are necessary to ensure the performance of computational electronic devices. In conventional memory devices, the performance is improved through scaling of the device dimensions. Increasingly multi-functional portable electronic devices demand a continued performance improvement at decreasing power densities. However, satisfying this demand by continued scaling of the conventional memory hierarchy becomes difficult. Among the emerging memory technologies being investigated for future generation applications is spin-transfer torque magnetic RAM (STT-MRAM). With properties including high endurance, high speed, non-volatility, CMOS compatibility at low power consumption, MRAM offers great opportunity. In addition, this magnetism-based memory, as opposed to conventional charge-based, has great scaling potential.

To enable this technology, reaching the reliability specifications is crucial. The tunnel barrier in STT-MRAM is 1 nm thick, through which high operating current densities of 1 MA/cm^2 flow, potentially causing breakdown and limiting the lifetime. The lifetime of the devices is estimated relying on models built for CMOS technologies, that do not take into account the contribution from self-heating. Self-heating is the phenomenon whereby the on-state device temperature increases as a result of the large currents flowing. In STT-MRAM, the impact of how the generated self-heating affects breakdown and the corresponding lifetime extrapolations has thus far not been studied in detail.

We have extended the breakdown model by incorporating the effect of self-heating, which is shown to increase device temperature by as much as 300 °C during measurement. We establish that the tunnel barrier thickness and device dimensions play dominant roles in the self-heating mechanism. An in-depth analysis corroborated by large measured breakdown statistics using (1) a breakdown time range of more than 11 orders of magnitude and (2) different tunnel barrier thicknesses is performed. In addition, we explain how oxygen

diffusion reduces the resilience to breakdown. Furthermore, we find that scaling down the device dimensions and tunnel barrier thickness will further increase the reliability margin between breakdown and writing current.

In addition, STT-MRAM is a non-volatile memory and therefore retains its memory state when the power is turned off, i.e. off-stress. Insufficient data retention can result in loss of this non-volatility. Data retention is defined by the energy barrier at off-stress conditions. Extracting this energy barrier requires switching the device at accelerated conditions and extrapolating to off-stress conditions, relying on simplified switching models. We perform a thorough evaluation of this extrapolation, and provide a baseline for validating which model most accurately describes the extrapolation to off-stress conditions. To achieve this baseline, a combination of temperature, magnetic field and current acceleration, including self-heating, are implemented.

Beknopte samenvatting

In dit proefschrift wordt de impact van zelf-geïnduceerde opwarming op de betrouwbaarheid van de nieuwe STT-MRAM geheugen technologie bestudeerd. Geheugens zijn noodzakelijk om de performantie in rekenkrachtige elektronische apparaten (Engels: device) te verzekeren. In conventionele geheugens is de performantie verbeterd door het verkleinen van de apparaten. De vraag naar betere performantie, maar ook lage vermogens, in een wereld van meer en meer multifunctionaliteit in draagbare elektrische apparaten, maakt het moeilijk voor de conventionele geheugen hiërarchie om deze vraag te verwezenlijken. Het continue verkleinen van deze apparaten wordt moeilijk. Dit opent deuren voor nieuwe technologieën zoals de spin-overbrengende koppel magnetische RAM (Engels: spin-transfer torque magnetic RAM). Met eigenschappen als hoog uithoudingsvermogen (Engels: endurance), hoge snelheid, niet-vluchtigheid en CMOS compatibiliteit heeft het een groot potentieel in verschillende toepassingsgebieden. Bovendien, is deze technologie gebaseerd op magnetisme in plaats van op elektrische ladingen, wat meer toekomst biedt voor het verder verkleinen van de apparaten om zo alsnog de performantie te verbeteren.

Om deze technologie waar te maken, is het halen van de betrouwbaarheidscriteria cruciaal. De tunnel barrière in STT-MRAM is slechts 1 nm dik, door dewelke hoge stroomdichtheden vloeien in de orde van 1 MA/cm² en doorslag (Engels: breakdown) van de tunnel barrière kunnen veroorzaken en zo de levensduur limiteren. De levensduur van de geheugens is geschat door gebruik te maken van modellen gemaakt voor CMOS technologieën, dewelke geen bijdrage van de zelf-geïnduceerde opwarming (Engels: self-heating) in rekening brengen. Self-heating is het fenomeen waarbij de temperatuur van het geheugen toeneemt door de hoge elektrische stromen. Er zijn momenteel geen studies over hoe deze self-heating de breakdown en de gerelateerde levensverwachting zal beïnvloeden.

We hebben het effect van self-heating toegevoegd aan het breakdown model. Tijdens een breakdown meting kan de temperatuur van het geheugen stijgen

tot wel 300°C. Overigens spelen de dikte van de tunnel barrière en de grootte van de apparaten een belangrijke rol in het self-heating mechanisme. We doen een uitgebreide analyse, bekrachtigd met grote gemeten breakdownstatistieken, die gebruik maken van (1) een grote tijdsschaal van breakdowntijden van meer dan 11 orde groottes en (2) verschillende diktes van de tunnel barrière. Verder verklaren we hoe zuurstofdiffusie zorgt voor een gereduceerde weerstand tegen breakdown. Bovendien, vinden we dat het verder verkleinen van de oppervlakte en de dikte van de tunnel barrière resulteert in een verbetering van de betrouwbaarheidsmarge tussen breakdown en schrijven van het geheugen.

Daarnaast is STT-MRAM een niet-vluchtig geheugen en kan daarom zijn geheugentoestand onthouden, zelfs wanneer het apparaat afstaat. Onvoldoende data retentie zorgt voor het verlies van de niet-vluchtigheidseigenschap. Data retentie is bepaald door de energie barrière bij af-toestandscondities. Extraheren van deze energie barrière vereist acceleratiemetingen, waarna een extrapolatie met behulp van modellen naar af-toestandscondities noodzakelijk is. Wij evalueren deze extrapolatie grondig en voorzien een basis voor het valideren welk model het beste de extrapolatie naar af-toestandscondities beschrijft. Hiervoor maken we gebruik van een combinatie van temperatuur, magnetisch veld en stroom acceleratie, samen met de zelf-geïnduceerde opwarming.

List of Abbreviations

0T1MTJ	zero transistor one-magnetic tunnel junction.
1T	One-transistor.
1T1C	One-transistor one-capacitor.
1T1MTJ	One-transistor one-magnetic tunnel junction.
6T	Six-transistors.
AHI	Anode hole injection.
AHR	Anode hydrogen release.
AP	Anti-parallel.
BD	Breakdown.
BE	Bottom electrode.
BEOL	Back-end-of-line.
BL	bitline.
CB	Confidence bound.
CD	Critical diameter.
CDF	Cumulative distribution function.
CIPT	Current in-plane tunneling.
CMOS	Complementary Metal-Oxide-Semiconductor.
CPU	Central processing unit.
CVS	Constant voltage stress.
DC	Direct current.
DMM	Digital multimeter.
DOS	Density of states.
DRAM	Dynamic random access memory.
DT	Direct tunneling.
DUT	Device-under-test.
DW	Domain wall.

eCD	Electronic critical dimension.
FEOL	Front-end-of-line.
FL	Free layer.
FN	Fowler-Nordheim.
HDD	Hard disk drive.
HL	Hard layer.
HM	Hard mask.
IBE	Ion Beam Etch.
IoT	Internet-of-things.
LLG	Landau-Lifshitz-Gilbert.
MC	Monte-Carlo.
ML	Maximum likelihood.
MRAM	Magnetic random access memory.
MS	Macrospin.
MTJ	Magnetic Tunnel Junction.
NVM	Non-volatile memory.
P	Parallel.
pdf	Probability density function.
PECVD	Plasma-enhanced chemical vapor deposition.
PMA	Perpendicular Magnetic Anisotropy.
PVD	Physical vapor deposition.
RA	Resistance area.
RF	Radio frequency.
RIE	Reactive Ion Etch.
RKKY	Ruderman, Kittel, Kasuya and Yosida.
RL	Reference layer.
RR	Ramp-rate.
RVS	Ramped voltage stress.
SAF	Synthetic antiferromagnet.
SL	source line.
SOT	Spin-orbit torque.
SRAM	Static random access memory.

SSD	Solid-State Drives.
STT	Spin-transfer torque.
STT-MRAM	Spin-transfer torque magnetic random access memory.
TCR	temperature coefficient of resistance.
TE	Top electrode.
TEM	Transmission electron microscopy.
TMR	Tunneling magnetoresistance.

List of Symbols

A	Area	m^2
A	Material and stress-dependent prefactor	$-$
A_{BDspot}	Area of breakdown spot	m^2
A_{ex}	Exchange stiffness constant	J/m
A	Inelastic tunneling fraction prefactor	$-$
α	Temperature coefficient	$1/K$
α	ratio between Q_p and Q_{BD}	$-$
α	Damping coefficient	$-$
α_T	Temperature acceleration coefficient	eV
β	Weibull slope	$-$
CB	Confidence bound	$-$
χ^2	Chi-square distribution	$-$
C_s	Heat capacity	J/Km^3
Δ	Thermal stability	kT
$\varnothing$	Diameter	m^2
$D_{ot,tot,crit}$	Critical oxide trap density	C/m^2
ΔT_{SH}	Self-heating Temperature	K
δ_w	Domain wall width	m
E_b	Energy barrier	J/m^3

E_{ox}	Electric oxide field	V/m
ϵ	Energy contribution	J/m^3
η	Scale factor or 63 %-value	s
$F(t)$	Cumlative distribution function (cdf)	$-$
$f(t)$	Probability density function (pdf)	$-$
f_0	Attempt frequency	$1/s$
G	Total free energy	J
γ	Field acceleration parameter	$-$
γ_0	Product of gyromagnetic ratio and permeability of vacuum	$rad\,m\,s^{-1}A^{-1}$
$H(\theta)$	Hessian matrix	$-$
ΔH_0	Zero-field activation energy	eV
$H_{AP-to-P}$	Switching field from AP to P	A/m
H_{BE}	BE height	m
$H_{BE,ext}$	BE extension height	m
H_c	Coercivity field	A/m
H_d	Demagnetizing field	A/m
H_{eff}	Effective magnetic field	A/m
H_{off}	Offset field	A/m
$H_{P-to-AP}$	Switching field from P to AP	A/m
H_{TE}	TE height	m
$H_{TE,ext}$	TE extension height	m
$I(\theta)$	Fisher information matrix	$-$
I_c	Critical switching current	A
J_{FN}	Fowler-Nordheim current density	A/m^2
k	Defect reaction probability	$-$

k_b	Boltzmann Constant	eV/K
K_{eff}	Effective anisotropy energy	J/m^3
κ_{th}	Thermal conductivty	W/Km^2
$\lambda(t)$	Switching rate	$1/s$
Λ	Maximum likelihood function	$-$
l_w	Domain wall length	m
M	Magnetization	A/m
m	Trap generation rate	$-$
M_s	Magnetization saturation	A/m
$\boldsymbol{\mu}$	Magnetic moment	Am^2
n	Power law exponent	$-$
N_{ij}	Demagnetizing tensor	$-$
P	Power	W
$P(t)$	Probability particle did not switch	$-$
ϕ_T	Arrhenius activation constant of power-law	eV
φ_T	Linear slope of Δ afo temperature	kT/K
$\overrightarrow{q}$	Heat flux density	W/m^2
Q	Total heat flux	W
Q_{BD}	Charge to breakdown	C/m^2
$Q_{p,crit}$	Critical hole flunece	C/m^2
R	Resistance	Ω
ρ_0	Resistivity	Ωm
R_{2Dax}	2D axis-symmetric radius	m
RA	Resistance area product	$\Omega\mu m^2$
R_{AP}	Anti-parallel resistance	Ω
R_{BD}	Post-breakdown resistance	Ω

R_{BDpath}	Post-breakdown resistance	Ω
R_{BDspot}	Post-breakdown resistance	Ω
R	Magnetic field ramp-rate	T/s
R_{MgO}	Resistance of pristine MTJ	Ω
R_{MTJ}	MTJ resistance	Ω
R_P	Parallel resistance	Ω
RR	Ramp-rate	V/s
R_{th}	Thermal resistance	K/W
σ_w	Domain wall energy	J/m^3
T	Temperature	K
t	Time	s
$T_{ambient}$	Ambient Temperature	K
τ_{pw}	Pulse width	s
t_{BE}	BE thickness	m
$t_{BE,ext}$	BE extension thickness	m
TCR	Temperature coefficient of resistance	Ω/K
∇T	Local temperature gradient	K
$\hat{\theta}_{ML}$	Maximum likelihood estimator	$-$
TMR	Tunneling magnetoresistance	-
T_{MTJ}	MTJ Temperature	K
t_{ox}	Oxide thickness	m
T_{SH}	Self-heating temperature	K
t_{TE}	TE thickness	m
$t_{TE,ext}$	TE extension thickness	m
$var(\theta)$	Variance matrix	$-$
V_{div}	Voltage division ratio	$-$

V_{max}	Maximum tolarable voltage	V
V_{MTJ}	MTJ Voltage	V
V_{ref}	Reference voltage	V
$W(t)$	Cumulative distribution function rescaled to Weibull scale	$-$
W_{BE}	BE width	m
$W_{BE,ext}$	BE extension width	m
W_{TE}	TE width	m
$W_{TE,ext}$	TE extension width	m
ζ	Defect generation efficiency	$-$

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In today's world, the number of electronic devices steadily increases. From the smartphone in your pocket to an autonomous driving car. In addition, the performance and functionality has grown tremendously. From a Nokia 3310 cellphone in 2000, to an iPhone X smartphone with fingerprint scanner and capability of capturing 4K-video. The semiconductor industry made these evolutions possible by designing an ingenious architecture built on a central processing unit (CPU), that executes computations at high speed, together with different types of memory devices serving as working memory or data storage, where each type of memory device has its own strengths and weaknesses.

In the future world, the demand of increasing performance at low power, new applications and connecting the world, requires maximum exploitation of the conventional computation systems, but also offers great opportunities for new emerging memories to shine where other memories fade.

1.1 Evolution of electronic memory

In the conventional memory architecture, the different types of memory devices have their strengths and weaknesses. They can be categorized in volatile and non-volatile (Fig. 1.1), where volatile memories cannot retain their memory state when the power is turned off. The most prominent volatile memories are static random access memory (SRAM) and dynamic RAM (DRAM). Non-volatile memories are generally used for data storage, we consider Flash and

hard disk drives (HDD). After briefly reviewing the existing memory devices, the limitation of the conventional memory hierarchy is discussed.

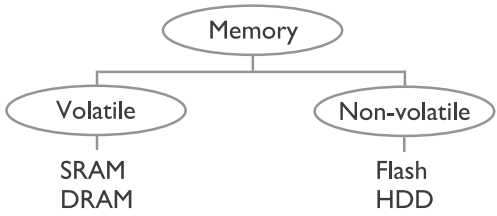

Figure 1.1: *General overview of the most prominent semiconductor memories.*

1.1.1 Existing memory devices

Static RAM (SRAM)

A conventional SRAM cell is composed of six transistors (6T). Four transistors (T1-T4) serve as a storage cell, where a bit (0 or 1) is stored. These transistors form two cross-couple inverters (Fig. 1.2). The two additional transistors (T5, T6) control the access to a storage cell during read and write operations. It is the fastest available type of memory and performs write and read operations in the order of a few nanoseconds. Therefore it finds its application as a working cache memory close to the CPU. A major drawback is the required chip area, due to the large number of transistors, a single SRAM cell can take up to 150 F^2, where F is the minimum lithographic feature size dictated by the technology node. The term "static" refers to the absence of a required periodical refresh, which is necessary in dynamic RAM.

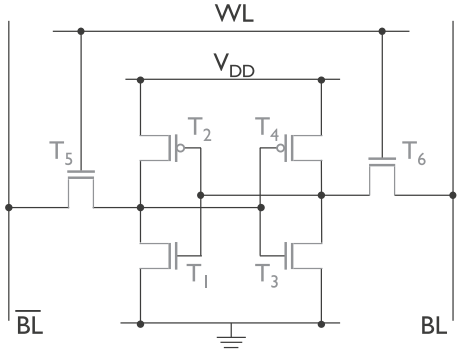

Figure 1.2: *Circuit diagram of a 6T SRAM cell.*

Dynamic RAM (DRAM)

DRAM only consists of one transistor and one capacitor (1T1C), see Fig. 1.3. The bit is stored on the capacitor, which can be either charged or discharged via the access transistor. Because the electric charge on the capacitor leaks off, it is required to refresh, i.e. recharge, the capacitor. In addition, the charging process is slower than SRAM, typically in the range of tens to hundreds of nanoseconds. The DRAM structure is very simple and allows to reach very high densities at low cost per bit.

Both SRAM and DRAM have fundamentally "unlimited" cycling endurance ($> 10^{15}$). Reading and writing to these memories constantly, would still result in a product lifetime of 10 years. Furthermore, these memories are volatile. Turning off the power results in a loss of the stored data.

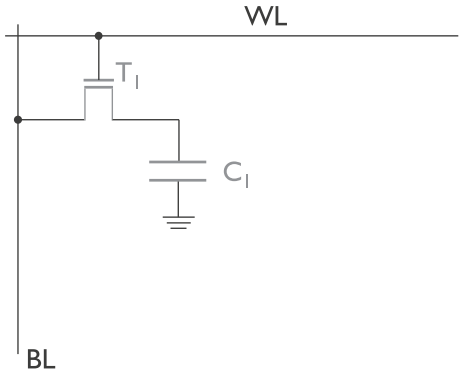

Figure 1.3: *Circuit diagram of a 1T1C DRAM cell.*

Flash

Flash is a non-volatile, charge-based memory. A typical flash memory cell is made of one transistor (1T) with two gates, a control gate and an extra floating gate (Fig. 1.4(a)). The floating gate is located between the control gate and the transistor channel. Note that the floating gate is fully surrounded by oxide, and therefore charges can be trapped within the floating gate. When applying a large, positive voltage to the control gate, electrons can tunnel through the oxide and get trapped in the floating gate. These trapped electrons screen, i.e. partially cancel, the electric field from the control gate, thus increasing the threshold voltage V_T of the cell. To erase the trapped electrons in the floating

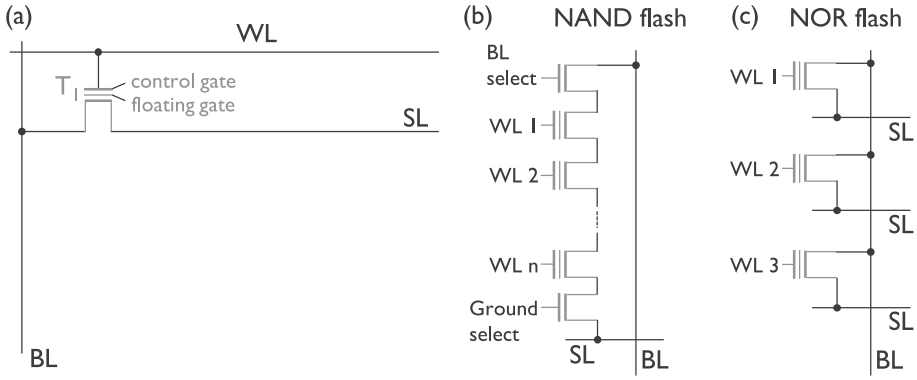

Figure 1.4: *(a) Circuit diagram of a 1T Flash cell with a control gate and a floating gate on which charges are stored. (b) NAND and (c) NOR Flash memory array (Replotted from [84]).*

gate, a large, negative voltage is applied to the control gate, which repels the trapped electrons from the floating gate. Depending on the amount of charges stored in the floating gate, and the related change in threshold voltage, it will result in a conductive transistor channel (bit '1') or a non-conductive channel (bit '0').

In general, Flash memories are divided into two main categories depending on their configuration, i.e. NAND and NOR (Fig. 1.4(b,c))). In NOR architectures, the transistor cells are connected in parallel to the bit lines, allowing each cell to be read and programmed independently. This configuration resembles a NOR-gate. In NAND architectures, the transistors cells are connected in series, which resembles the working of a NAND-gate. This configuration takes up less space than NOR. The data access, however, can only be serially, which increases the read latency. Considering the area advantage and slower read, NAND Flash is more suitable for large data storage as in Solid-State Drives (SSD). NOR Flash on the other hand, is used for code storage, where frequent and fast data readout is required.

Hard Disk Drive (HDD)

A HDD is a magnetic-based storage device. The information is stored by the direction of the magnetization in a ferromagnetic thin film. This is the cheapest type of memory and is used for very large storages ($> 1\,\mathrm{TB}$). To read and write the bit a mechanical arm with an electromagnetic head is used. Due to the

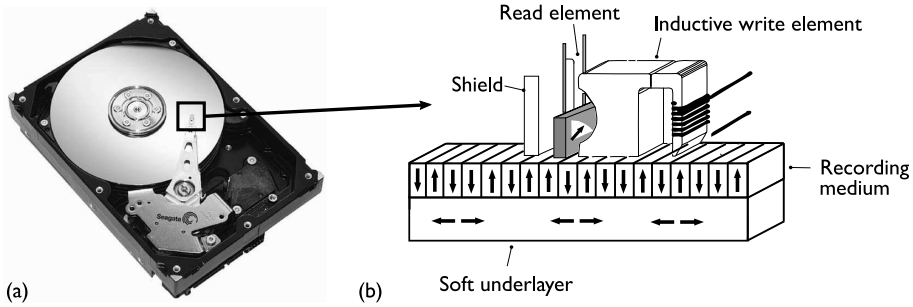

Figure 1.5: *(a) A $3\frac{1}{2}$-inch HDD (courtesy of Seagate Technology). (b) Schematic of perpendicular magnetic recording. Replotted from [23].*

mechanical interactions, the HDD device has slow writing/reading speeds in the order of milliseconds.

1.1.2 Memory hierarchy - filling the memory gap

In a typical Von Neumann architecture, an instruction fetch and a memory write/read operation cannot occur at the same time, because they share a common bus. To mitigate the bottleneck of a fast working CPU and a slow data transfer to the main memory, different cache levels are implemented between the CPU and the main memory. The closer the cache is to the CPU, the smaller and faster the memory is, resulting in a memory hierarchy (Fig. 1.6). The hierarchy separates the memories based on their capacity and speed. At the top we find the CPU, with below typically 3 levels of very fast SRAM-cache, each level with larger capacity and optimized differently. One step below is the DRAM, which is slower, but has high density and lower cost. The next two levels are NAND flash and HDD, large, cheap and slow memories for data storage. In between DRAM and NAND flash there is a gap, causing an overall performance limitation. Namely, when data is transfered from DRAM to NAND, the CPU has to wait.

It is in the lower cache level of SRAM (L2-L3), DRAM-level or part of the gap that STT-MRAM has good potential to be used.

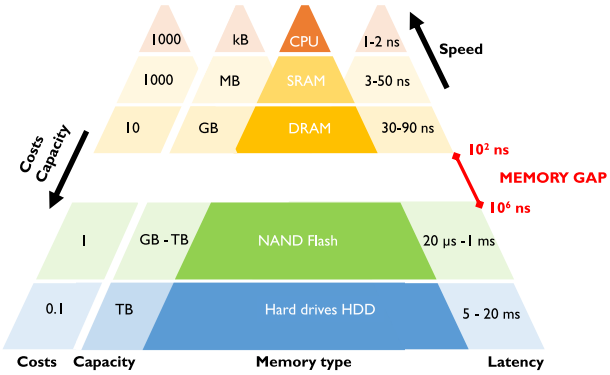

Figure 1.6: *Memory hierarchy in the existing computational system where a latency gap is observed between the DRAM and NAND Flash. Replotted from [18].*

1.2 Spin-Transfer Torque Magnetic RAM (STT-MRAM) as a universal memory

STT-MRAM is a promising memory technology, considered to be used as a universal memory, because of its high endurance [58], high speed [53], low voltage operation [53, 54] and CMOS compatibility. In addition, STT-MRAM is non-volatile, which results in an even wider range of applicability. From replacing embedded Flash to replacing lower SRAM cache (L2-L3). Furthermore, STT-MRAM could be used in new emerging markets like the internet-of-things (IoT), which is a collection of smart devices connected via the internet. For these applications ultra-low power is imperative.

The principal element in the MRAM technology is the magnetic tunnel junction (MTJ). The MTJ consists of 2 ferromagnetic layers with in between a tunnel barrier (Fig. 1.7(a)). One ferromagnetic layer is called the reference layer (RL), the other the free layer (FL). Information is stored in the orientation of the magnetization of these ferromagnetic layers. The RL is the most stable and its magnetic orientation remains pinned, whereas the magnetic orientation of the FL can switch. Depending on the configuration of the orientations between FL and RL, the resistance of the MTJ is high or low for an anti-parallel (AP) and parallel (P) orientation, respectively. Switching from state to state is possible via magnetic field (Fig. 1.7(b)) or electric voltage/current (Fig. 1.7(c)). Furthermore, MRAM is non-volatile, when it is not stressed it can retain its state. As such, MRAM is a simple 2-terminal resistor and is used together with a transistor (1T1MTJ).

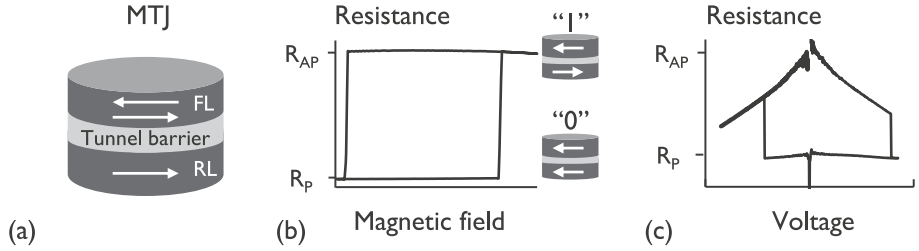

Figure 1.7: *(a) The MTJ is composed of 2 ferromagnetic layers and in between a tunnel barrier. One bi-stable ferromagnetic layer is the free layer (FL), the other pinned layer is the reference layer (RL). The magnetization of the FL can be switched by a magnetic field (b) or voltage/current (c). The magnetic orientation between RL and FL can be either parallel (P) or anti-parallel (AP). We distinguish a high resistive state "1", i.e. R_{AP}, and a low resistive state "0", i.e. R_P.*

In section 1.2.1, we briefly discuss some important steps in the evolution of MRAM from a field-driven to a current-driven memory. Next, in section 1.2.2, we elaborate on the trade-offs, and in section 1.2.3, important reliability aspects are considered.

1.2.1 Evolution of MRAM

MRAM has already seen a remarkable evolution since the product development by Everspin around 2004 [80]. At the base of the technology is the magnetic tunnel junction (MTJ), which in its most basic form consists of 2 ferromagnetic layers with in between a dielectric tunnel barrier. Information is stored in the magnetization of the ferromagnetic layers, resulting in a high and low resistive state.

The most important members of the MRAM family are illustrated in Fig. 1.8. The magnetization orientation of the MTJ is represented by the white arrow. We consider field-driven and current-driven MRAM, Fig. 1.8(a) and Fig. 1.8(b,c), respectively. In field-driven MRAM, the conventional MRAM, the magnetic field to switch the magnetic orientation of the MTJ is generated by sending current through the bit lines, i.e. the so-called Oersted field. For current-driven switching we distinguish the spin-transfer torque (STT) and the spin-orbit torque (SOT). We briefly explain these members of the MRAM family.

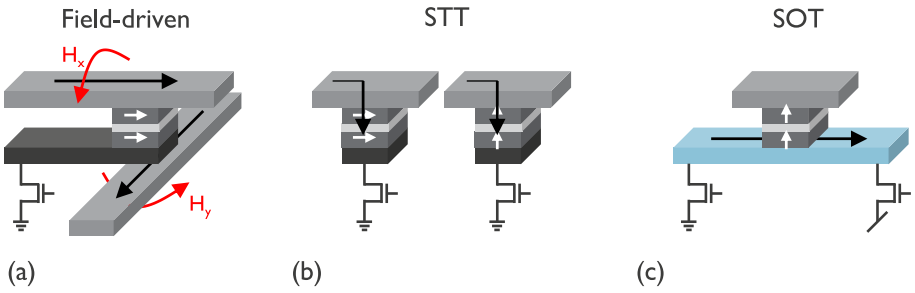

Figure 1.8: *Members of the MRAM family. (a) Field-driven MRAM makes use of Oersted fields (red arrows) generated by sending current through the bitlines. (b) Current-driven MRAM makes use of torques (STT) exerted by the current, which flows through the device. (c) current-driven MRAM makes use of torques (SOT) exerted by the current, which flows underneath the device. The white arrows represent the magnetization of the RL and FL.*

Conventional magnetic field-driven MRAM

In the first generation of MRAM, Oersted fields, generated by sending a large current through the bit lines, induce switching of the in-plane MTJ-state (Fig. 1.8(a)). The major limitation of this technology is the limited scalability to a device area of $60 \times 150\,\mathrm{nm}^2$, because at these dimensions the current density ($> 10^7\ A/cm^2$) necessary to generate strong enough Oersted fields, causes failure due to electromigration. This technology is still being produced by Everspin for niche applications in e.g. aerospace [39] and automotive [40], but there is no further development for larger markets.

Current-driven STT-MRAM

Also current can be used directly to switch the magnetization of the FL. Slonczewski at al. have found that the spin of the electron can exert a torque on the FL magnetization, i.e. the spin-transfer torque (STT) effect [99]. On nanometer scale, it is more effective to exert torque by spin-transfer than by the generated Oersted fields in the nearby bitlines in field-driven MRAM, resulting in better scalability. Although, high current densities ($>1\,\mathrm{MA/cm}^2$), flowing now through the MTJ, are required to switch the magnetization state of the FL and will heat up the device and can cause breakdown of the tunnel barrier.

First developments have been done on the same in-plane MTJs used in the field-driven MRAM, but this technology did not prove to be scalable below

$30 \times 30\,\mathrm{nm}^2$ [34], because the data retention below these dimensions could not be guaranteed. Recent discoveries, however, offer a solution using a perpendicular magnetic orientation [52].

Based on these perpendicular MTJs, the development of the STT-MRAM technology gained around-the-world interest by research centers and universities [136, 52, 85, 53, 118], but also prominent companies have their own MRAM program, like Qualcomm [59], Samsung [100], Toshiba [22], GlobalFoundries [45] and Taiwan semiconductor manufacturing company (TSMC) [112].

3-terminal SOT-MRAM

An interesting new type of device has entered the MRAM family recently. It is a three-terminal device, where the switching is induced by means of spin-orbit torque (SOT). In addition, the read and write path are decoupled. The high current densities required to switch the free layer of the MTJ do not need to pass through the thin tunnel barrier, because current flows in the metal path under the MTJ, see blue metal line in Fig. 1.8(c)). This way, the risk of breakdown of the tunnel barrier is significantly reduced. Moreover, the switching mechanism is fast and sub-nanosecond. However, this type of device is still in the research phase and we will not further discuss or analyze the SOT-MRAM.

1.2.2 Trade-offs

STT-MRAM is sometimes considered as the universal memory, because MRAM has high speed, high endurance, low power and non-volatility. However, there exist many trade-offs between these properties, where the STT-MRAM will have to compromise. For example, for good data retention, a very stable state with a high energy barrier is necessary, which is hard to switch, and thus will require high current or slower switching speeds. A higher current in its turn will cause substantial heating (self-heating) of the MTJ and accelerate breakdown. Several of these trade-offs can be illustrated with a "trilemma triangle" (Fig. 1.9). On the corners are three properties of interest, retention, density and write speed. On the lines are the parameters that have an opposite effect on the property the line connects (MTJ critical diameter (CD), energy barrier and write current). For high density you require small MTJ CD, where CD is the critical dimension, but larger MTJs will have better data retention. One should not fixate on these dependences in the trilemma triangle, it illustrates that trade-offs exist.

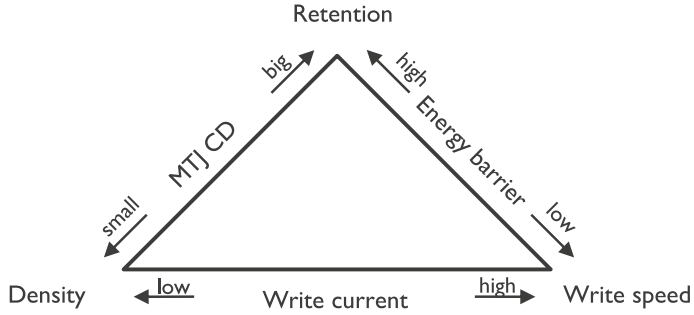

Figure 1.9: *Illustration of trade-offs seen in STT-MRAM. Three properties of interest: data retention, memory density and write speed are bound to compromise, optimizing one can result in weakening the other. In between the properties are shared parameters like the energy barrier, the MTJ critical dimension (CD) and the write current. Replotted from [15].*

Despite the trade-offs, being tunable and having an overall good performance make MRAM suitable for a large number of applications.

1.2.3 Reliability concerns for STT-MRAM

A reliable STT-MRAM has to fulfill the required operation for 10 years. To test whether STT-MRAM is reliable, two steps are required. (1) Accelerate the failure mechanisms to enable failure observation and failure analysis. (2) Extrapolate these failure mechanisms back to operation conditions, relying on accurate models. Among the reliability concerns we distinguish these important three:

1. Breakdown of the tunnel barrier causes limited endurance specifications. In contrast to the negligible current that flows in CMOS dielectrics, STT-MRAM operates at high current densities, causing breakdown of the tunnel barrier. In STT-MRAM, there is insufficient in-depth analysis of the breakdown dependencies and modeling, and more importantly how self-heating will affect the breakdown model.

2. Insufficient data retention results in loss of the non-volatility property. Data retention is determined by the energy barrier needed to overcome at off-stress operation conditions. There are many publications on switching

mechanisms, switching models and acceleration techniques. There is, however, no consensus on which model is correct and more importantly none of the existing models take into account the effect self-heating will have on the extraction of the energy barrier.

3. Variability causes each device to operate differently. The uniformity of the STT-MRAM process is crucial, and for this a very stable process and high density test vehicles are required. How self-heating in these dense arrays can impact the lifetime and operation is not known. In addition, taking into account the variability by making use of safety margins for large memory sizes can affect reliable read or write operations, causing unwanted errors and failures.

1.3 Objective and outline

In this thesis, we make use of our in-house STT-MRAM technology, studying isolated devices and 1024 x 1024 arrays. Although the STT-MRAM process cannot compete with the fabrication process of first-class companies, significant learning is achieved on the first two reliability concerns, breakdown and data retention. Operating at high current densities causes significant self-heating. The main goal is to characterize and model the effect of self-heating, in order to make accurate lifetime predictions for (1) breakdown and (2) data retention.

Chapter 2 introduces the basic properties of the MTJ. The full stack configuration is discussed in detail and it is explained why for STT-MRAM is opted for these specific materials and functional layers. In addition, the complex process is elaborated, which requires a 400° C thermal budget to ensure back-end-of-line compatibility.

Self-heating simulations show that STT-MRAM substantially heats in the matter of nanoseconds (**chapter 3**). The bad thermal conducting layers near the tunnel barrier, where the heat is generated, are at the origin of this fast heating and high thermal resistance. Furthermore, the impact of the surroundings is investigated, which results in an additional slower heating process. These simulations, however, rely on poorly known thermal conductivities, considering the many nanometer thick layers in the stack. Therefore, two indirect measurements to characterize the self-heating are proposed, (i) using breakdown and (ii) retention statistics. Self-heating is estimated to have significant impact on both operation as accelerated conditions.

The first important reliability concern is breakdown of the nanometer thick MgO tunnel barrier. In **chapter 4**, we introduce the concepts of breakdown distribution and the lifetime extrapolation approach. We have developed a new maximum likelihood fitting method, which simultaneously fits the breakdown distribution as well as the voltage acceleration. In addition, making use of a developed and in-depth analyzed pulsed breakdown-based measurement method, we demonstrate that the MgO degradation is cumulative in nature, and that a power-law voltage acceleration model best describes the lifetime data.

The measurement techniques elaborated in chapter 4 are used to study the influence of different processing steps and stack configurations to find that the functional layers near the MgO tunnel barrier significantly impact the breakdown characteristics (**chapter 5**). We propose an oxygen scavenging model to explain the increased susceptibility to breakdown observed when using the standard Ta-based functional layers. Furthermore, we demonstrate that taking into account self-heating is imperative to make accurate lifetime predictions. In addition, we elaborate on the in-depth temperature analysis on devices with different MgO thickness, which is required to isolate and characterize the self-heating in STT-MRAM in an indirect manner.

The second reliability concern is data retention. Accurately extracting the thermal stability, and thus characterizing the data retention, requires large switching statistics, obtained by accelerating switching with magnetic field, current or temperature (**chapter 6**). Based on two switching mechanisms, i.e. uniform switching and domain wall mitigated switching, two frequently used models are studied and their impact on the thermal stability extraction are discussed. Furthermore, using a statistical method and including the impact of self-heating at accelerated conditions, allows to discriminate between the switching models and results in an additional indirect self-heating characterization method.

Finally, **chapter 7** provides a concluding overview of the main results and highlights the prospects for future work.

Chapter 2

Theory on MTJ

A reliable STT-MRAM technology is composed of many nanometer thick layers. All layers contribute to a 400 °C compatible process and contribute to achieve good performance for all basic properties.

2.1 Introduction

A lot of engineering efforts have been made to ensure the STT-MRAM technology fulfills all requirements of non-volatility, high speed, high endurance, low power and CMOS technology compatibility. These requirements translate into the basic operating properties of the magnetic tunnel junction (MTJ), the key element in STT-MRAM technology. Furthermore, these operating properties cannot degrade during the BEOL 400 °C processing thermal budget. In addition, the basic operating properties result in engineering trade-offs, meaning faster MTJs can lead to a decrease in breakdown reliability, or a very stable non-volatile MTJ requires higher power to force switching. Engineering the stack to obtain the best of all worlds is not the main topic of this thesis. In this chapter, the focus is on establishing a better understanding of the mechanisms that influence the MTJ properties.

We first discuss the basic MTJ properties that allow reading, writing, retaining the MTJ-state (section 2.2). Next, advanced properties and important adjustments to the simple MTJ-stack are elaborated in section 2.3. Furthermore, the most important processing steps are considered in section 2.4.

2.2 Basic MTJ properties

In the heart of the STT-MRAM technology one finds the magnetic tunnel junction, which is composed of three main layers, as illustrated in Fig. 2.1(a): a magnetic free layer (FL), a tunnel barrier and a magnetic pinned reference layer (RL). These layers are only 1-2 nm thick and patterned into circular devices with diameters between 45-500 nm.
The FL has a bi-stable magnetic state and can be switched by a magnetic field or current, whereas the RL remains fixed. Changes in the relative orientation of the magnetization of the FL and the RL, result in a large magnetoresistance (MR) effect [136]. A parallel (P) and anti-parallel (AP) configuration corresponds with a low and high resistive state, respectively. The MR effect makes reading "0" and "1" possible, and in the case of an MTJ, which has a tunnel barrier, becomes tunneling magnetoresistance (TMR).

In this section, we will discuss some of the basic properties and parameters of the MTJ. How to read by TMR (Sec. 2.2.1), how writing an MTJ can be induced by STT (Sec. 2.2.2), how the MTJ can retain its state due to thermal stability (Sec. 2.2.3) and how the resistance area (RA) product, together with the MTJ CD are important to have an optimal synergy with the CMOS technology in section 2.2.4.

2.2.1 Principles of read by TMR

In a simplified schematic, depicted in Fig. 2.1(b), this tunneling magnetoresistance (TMR) effect is explained. As opposed to a non-magnetic metal, in a spin-polarized ferromagnetic metal, there are more states of one spin direction than the other at the Fermi level (Fig. 2.1(b)). When there are more states of spin-up, the magnetization of the RL or FL, will be up. The electrons "travel" from the FL to the RL by tunneling through the barrier, while conserving their spin state. When the FL and RL are in parallel alignment the spin-up electrons tunnel from a low density of states (DOS) to a low DOS, whereas the spin-down tunnel from a high DOS to a high DOS, resulting in a low resistance (R_P), bit-state "0". For an anti-parallel alignment, the electrons either leave from or tunnel to a low DOS, resulting in a high resistance (R_{AP}), bit-state "1". The

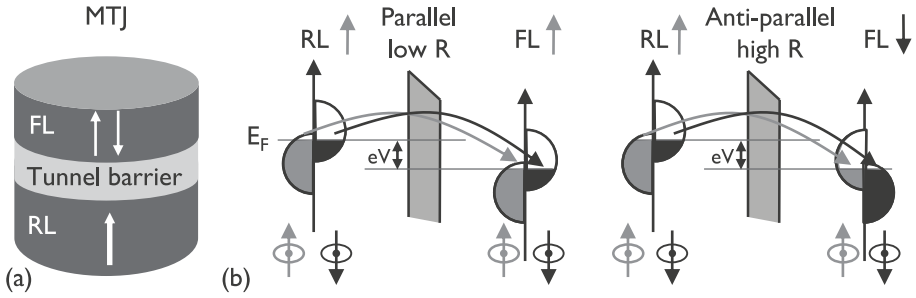

Figure 2.1: *(a) Functional layers in an MTJ, the ferromagnetic free- (FL) and reference layer (RL), with in between a tunnel barrier. The magnetic orientations are depicted as arrows, where the FL has a bi-stable magnetic state. (b) Schematic representation of the tunnel magnetoresistance in the case of two identical ferromagnetic layers. The tunneling process conserves the spin. When the electron states on each side of the barrier are spin-polarized, then electrons will find more easily free states to tunnel to when the magnetizations are parallel (left), than when they are anti-parallel (right). Replotted from [16].*

TMR, a typical figure of merit, is then given by:

$$TMR = \frac{R_{AP} - R_P}{R_P} \tag{2.1}$$

The TMR effect was first demonstrated by Juliere in 1975 [56]. Values around 14 % were obtained. Breakthrough came in 2004, using crystalline MgO barriers, where MgO acts as a near-perfect spin-filter, and huge TMR values in excess of 200 % are achieved at room temperature [136]. A thin MgO barrier is currently used in magnetic read heads for hard disk drives and in the STT-MRAM.

In the case of a perpendicular MTJ, as was depicted in Fig. 2.1(a), the magnetization preferentially likes to be up or down. This preferential direction is called the *easy-axis*. For these MTJs, it is more difficult to force the magnetization to be in the plane, i.e. the hard axis.

An MTJ can be switched by magnetic field and current. Magnetic field switching can be studied in typical TMR-loops [see Fig. 2.2(a)]. To obtain this typical hysteresis curve for ferromagnetic elements, the magnetic field is swept from high, negative to high, positive fields and back, keeping the magnetic field oriented parallel with the easy-axis, in this case this is perpendicular to the plane. In a TMR-loop we can distinguish the following important parameters: R_P, R_{AP}, TMR, coercive field (H_c) and offset field (H_{off}). Where H_c and

H_{off} are defined as:

$$H_c = \frac{H_{P-to-AP} - H_{AP-to-P}}{2} \tag{2.2}$$

$$H_{off} = H_{P-to-AP} - H_c \tag{2.3}$$

with $H_{P-to-AP}$ and $H_{AP-to-P}$ the magnetic field at which switching from P-to-AP and AP-to-P occurs, respectively. Ideally the offset field is kept zero at operating conditions. H_{off} originates from a stray field created by the pinned magnetic layer, which acts upon the free layer. The stack needs to be engineered to keep it close to zero.

Switching by means of a magnetic field was integrated in the first MRAM technology. The magnetic field in this conventional MRAM technology, is generated by sending a large current through a nearby wire. This so-called Oersted field, named after its discoverer in 1820, depends on the total current flowing through the wire and the radial distance at which the field is measured. The conventional MRAM does not scale to lower dimensions, because the critical current density required for switching becomes too high. The conventional MRAM is therefore only used in niche products with low memory capacity and density.

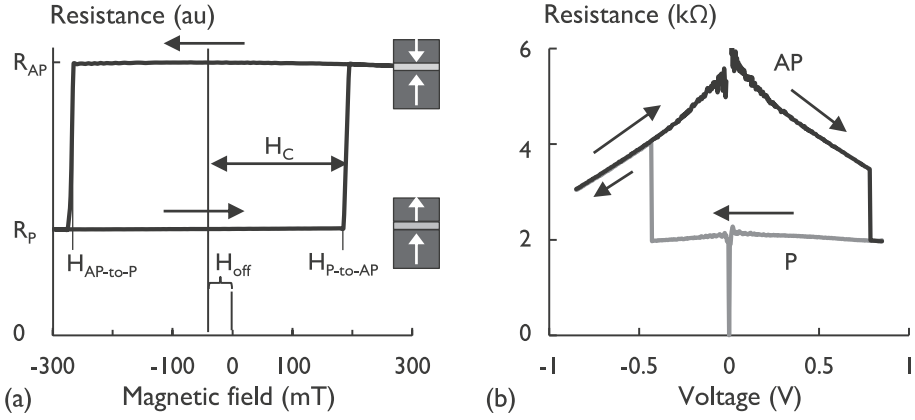

Figure 2.2: *(a) Example of a typical TMR loop. All important parameters are depicted on the figure: parallel and anti-parallel resistance (R_P and R_{AP}), the coercive- and offset field (H_c and H_{off}). (b) Example of a R-V trace. The resistance depends on the stress voltage. This bias dependence is more severe in the AP-state.*

In principle, the higher the TMR, the better. From application perspective, a practical TMR range is between 100 and 200 %.

2.2.2 Principles of write by STT

Since Slonczewski's discovery of the spin-transfer torque [99, 7], a two-terminal STT-MRAM can be switched solely by current pulses. Electrons exert a torque on the magnetization of the FL. By sending sufficient current through the MTJ the state of the FL can be switched. In order to switch, the STT has to overcome a damping term. This damping term tends to relax the magnetization back to the easy-axis. In case of high damping, more current is required in order to switch the MTJ-state. The magnetization dynamics and the effect of damping on switching is elaborated in Sec. 6.2.

In Fig. 2.2(b), an example of the resistance as a function of DC voltage is plotted. This R-V-trace shows the bias dependence of especially the anti-parallel resistance (R_{AP}). At higher bias the decrease in R_{AP} will result in a decrease in TMR.

High current densities are required to switch the MTJ, a target number to reach is $1\,\mathrm{MA/cm^2}$.

2.2.3 Principles of storage by thermal stability

A major advantage of the STT-MRAM technology is its non-volatility. The FL is bi-stable and retains its state when the power is switched off. In addition to the advantage of not losing data during a power outage, this also enables low power consumption, which is very important for portable electronics.

Data retention for STT-MRAM is characterized by the thermal stability factor, which is the ratio between the activation energy needed to switch states and the available thermal energy:

$$\Delta = \frac{E_b}{k_b T}\,, \tag{2.4}$$

here E_b is the energy barrier necessary to switch from AP-to-P or P-to-AP state and k_b the Boltzmann constant. For the required non-volatile operation, Δ values in the range of $\Delta = 50$ are required. Δ is defined at operating temperatures. Switching and thermal stability will be investigated in more detail in chapter 6.

2.2.4 Optimal RA product and MTJ CD for CMOS compatibility

Two other important parameters of the MTJ are the resistance-area product or RA and the critical diameter or CD. The CD is the diameter of the MTJ pillar at the MgO level. The RA parameter depends exponentially on the oxide thickness [136] and is used to characterize the absolute resistance of the stack, independent of the device area:

$$R_p = \frac{RA}{A} \sim \frac{e^{C_1 \cdot t_{MgO}}}{A}\,, \tag{2.5}$$

with C_1 a material constant and t_{MgO} the MgO thickness. The RA and the MTJ CD determine the resulting R_p of the device. An optimal resistance to cooperate with CMOS should be larger than 1 kΩ to be easily discriminated from the periphery resistances from the transistors. In addition, the resistance should be as low as possible to enable a very fast read out. For an MTJ with diameter (⌀) 60 nm, an ideal RA would be around $10\,\Omega\mu m^2$, which corresponds with an MgO thickness of 1 nm. However, scaling down the MTJ diameter (⌀< 20 nm), would result in too high resistance ($>\ 10\,k\Omega$), if reading speeds in nanosecond range are required. A lower RA of $4\,\Omega\mu m^2$ is better suited for these cells.

For patterned devices, i.e. thin films deposited over 300 mm wafers, the RA can be measured with a current in-plane tunneling (CIPT) technique [126]. This non-destructive technique relies on different probe spacings, in the range of 1 μm to 10 μm, between 12 in-line probe needles. CIPT calculates, using a four-point-probe method, the sheet resistance of the conducting, ferromagnetic layer above MgO and below MgO, and also the RA.

In this thesis, we study patterned devices. In this case, the RA is estimated via the resistance and area of a ⌀ 500 nm device. We have assessed the area of this device via a TEM micrograph. Deviation from the measured area with several nanometers will not significantly affect the resistance in a ⌀ 500 nm device, which would not be the case for a ⌀ 60 nm device. The RA is dominated by the thickness of the MgO and for 1 nm MgO the relevant RA is in the order of $10\,\Omega\mu m^2$ or smaller.

The electronic CD (eCD) is an estimation of the CD, based on the RA-value of patterned devices. We derive the eCD by calculating the area via Eq. 2.5, given the measured R_p and RA derived from the ⌀ 500 nm device. In the performed

measurements of this thesis, we report the nominal value of the CD printed in the layout and not the eCD, unless noted.

2.3 Advanced MTJ properties

The basic MTJ, discussed in Section 2.2, consists of two CoFeB layers and a MgO tunnel barrier. A perfect MgO-CoFeB interface is necessary to obtain high TMR. However, the final stack is much more complex than these three layers. Several changes have been made to the initial concept to make an STT-MRAM cell functional, meaning it can be fully integrated in the CMOS process, operate at low power, provide sufficient data retention and reliability. We will briefly discuss the introduction of perpendicular magnetization, synthetic antiferromagnet (SAF) and double MgO.

2.3.1 Perpendicular magnetic anisotropy (PMA)

Magnetic anisotropy means that the magnetic orientation of a ferromagnetic or antiferromagnetic layer lies along a preferential direction, i.e. easy-axis. Among different sources of anisotropy, magnetocrystalline anisotropy, shape anisotropy and interface anisotropy are the most important.

In magnetocrystalline anisotropy the magnetization aligns itself along a preferred crystallographic direction. Most magnetic materials show some magnetocrystalline anisotropy, however, a polycrystalline sample with no preferred orientation of its grains will have no overall crystalline anisotropy.

Nevertheless, the magnetization will depend on the shape of the sample. If the sample is not spherical, then it will be easier to magnetize it along the longest axis. This is called shape anisotropy and in thin film ferromagnets it forces the magnetization to be in the plane, since the perpendicular axis, i.e. along the thickness, is the shortest axis. The shape anisotropy is a result of the demagnetizing field, which is in opposite direction and proportional to the magnetization and a demagnetizing factor:

$$H_{di} = -N_{ij} M_j \;\; i,\, j = x, y, z\,, \tag{2.6}$$

where N_{ij} is the demagnetizing tensor, which is generally represented by a symmetric 3×3 matrix, determined by the shape of the sample, and is smallest

in the direction of the long axis of a thin film. For a spherical shape it is 1/3 in all directions, but for a perfect thin film it is 0 in the plane and 1 perpendicular to the plane. The first implementations of MRAM made use of in-plane magnetized CoFeB on elliptical shaped thin films. Introducing a large aspect ratio in the plane forces the magnetization in the direction of the long axis of the ellipse.

In the case of CoFeB it is found that perpendicular magnetic anisotropy (PMA) emerges if the layer is thinner than 1.6 nm [133, 52]. The PMA originates from the interface anisotropy, which takes over from the shape anisotropy. For very thin films the surface is more important than the bulk, and the surface atoms are in highly anisotropic surroundings, caused by orbital hybridization at the MgO/CoFeB interface [52], resulting in PMA. In addition, perpendicular MTJs offer better scaling potential for the following reasons: (1) Reducing MTJ CD, reduces the anisotropy energy and limits the downsize scalability to 800 nm^2 [34]. (2) The PMA does not suffer from the demagnetization term for the switching current, resulting in lower switching current for the same thermal stability. (3) Even though the PMA reduces when reducing the MTJ CD, the PMA is expected to be sufficient for downsizing the MTJ CD below 40 nm, certainly after the introduction of the double MgO interface (Sec. 2.3.3).

2.3.2 Synthetic antiferromagnet (SAF)

The synthetic antiferromagnet or SAF is introduced to stabilize the reference layer. To obtain PMA, the CoFeB reference layer has to remain thinner than 1.6 nm. The PMA is a result from interface anisotropy of the CoFeB and MgO interface. However, this PMA is not stable enough to serve as a pinned layer, e.g. the CoFeB from the FL has approximately the same thickness and interface. To improve the stability, without introducing a large stray field H_{off}, the RL is therefore coupled to a SAF structure.

The SAF structure consists of Co/Pt multilayers and a Ru coupling layer. Depending on the thickness of the Ru, the two Co/Pt layers are coupled ferromagnetically or anti-ferromagnetically, see Fig. 2.3(a). This is the so called RKKY exchange interaction, named after its discoverers Ruderman, Kittel, Kasuya and Yosida [92, 62, 135]. We will also refer to this SAF structure as the hard layer (HL). It is the Co in the Co/Pt multilayers that results in this high antiferromagnetic exchange coupling [9].

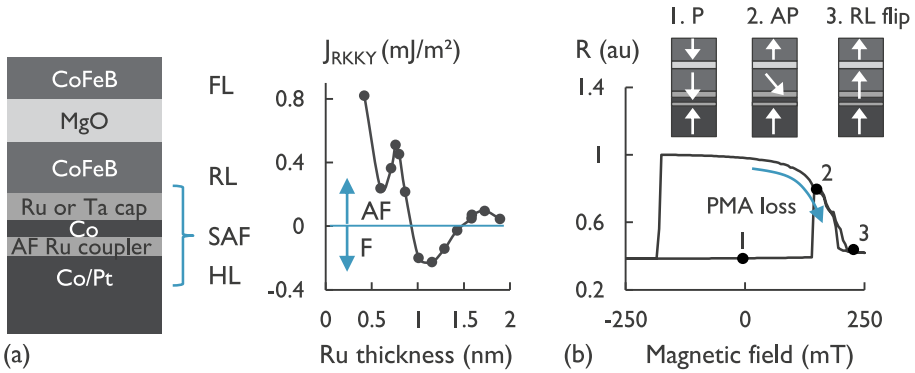

Figure 2.3: *(a) Introduction of synthetic anti-ferromagnet (SAF) by coupling a Co/Pt multilayer with a thin Ru layer. This coupler layer results in a RKKY exchange interaction (J_{RKKY}), which depends on the thickness of the Ru [9]. The magnetic coupling oscillates between anti-ferromagnetic (AF) and ferromagnetic (F). (b) Example of a badly-designed TMR loop. The PMA in the RL is unstable, resulting in a decrease in R_{AP} (2) and eventually a flip of the RL magnetization (3) at even higher magnetic fields.*

Furthermore, designing a thin SAF is necessary to achieve high TMR. In bottom-pinned MTJs, the CoFeB-MgO interface is only deposited after the SAF layer. Roughness at MgO level, deteriorates the TMR and depositing a thick SAF will increase the roughness [61]. In addition, designing a thin MTJ stack, will also be beneficial for the MTJ patterning, since less material has to be removed.

Engineering the thinnest stack as possible [61] and processing at high annealing temperature >375 °C, destabilizes the reference and hard layer. This is illustrated in Fig. 2.3(b), where we see an example of an unacceptable TMR loop, due to RL instability. For high applied magnetic field the anti-parallel coupling is not strong enough and the RL starts to align with the external magnetic field. This RL flip leads to a parallel alignment of the RL and FL, i.e. the parallel resistance.

2.3.3 Double MgO

When scaling down the MTJ size below 40 nm, the FL becomes unstable and cannot assure the required data retention. Therefore, a second CoFeB-MgO interface is introduced, resulting in the double MgO [95], see Fig. 2.4(a). Since

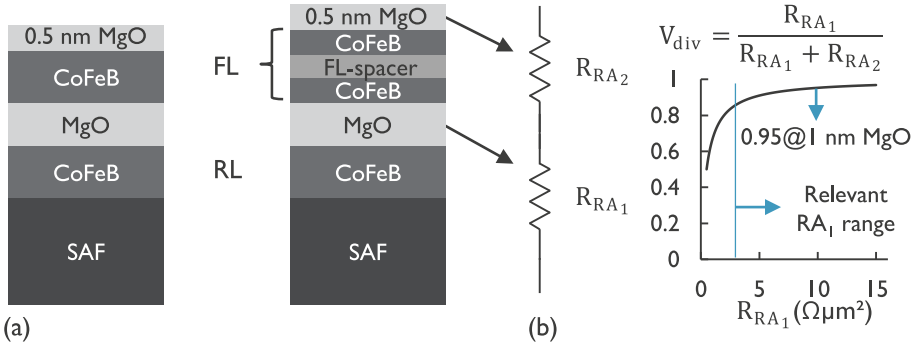

Figure 2.4: *(a) Schematic of the double MgO, with a 0.5 nm thin MgO-capping layer. (left) single CoFeB free layer and (right) composite free layer with a non-magnetic spacer layer in between 2 CoFeB layers. (b) The second MgO layer acts as a parasitic series resistance (R_{RA_2}), however in the relevant RA range of the first MgO, more than 80 % voltage drops over the first MgO (R_{RA_1}).*

the PMA comes from the interface between MgO and CoFeB, this is enforced by introducing two interfaces. The two CoFeB-MgO interfaces provide enough thermal stability even below a CD of 40 nm.

To improve the thermal robustness of the free layer and make it 400 °C compatible, a composite free layer is required. The composite free layer is composed of thin CoFeB layers with in between a non-magnetic spacer layer, conventionally Ta-based. The Ta spacer is hypothesized to act as a boron, as well as an oxygen getter during the high temperature anneal [25]. The necessity and impact of the high temperature anneal are further discussed in section 2.4.1.

The second MgO layer is kept very thin (0.5 nm) to make sure there is no TMR penalty, due to a series resistance effect. At these thicknesses the RA of the MgO layer would be below $0.5\,\Omega\mu m^2$. It is then estimated that 95 % of the applied voltage will drop over the first MgO layer for a 1 nm MgO. For thinner MgO, still in the relevant RA-range, this ratio drops to 85 % for an RA of $3\,\Omega\mu m^2$.

By introducing PMA, a SAF-structure and a double MgO configuration, the operation and stability requirements are met: high TMR due to a good interface between crystalline MgO and CoFeB, high thermal stability by providing two MgO-CoFeB interfaces to the FL and a stable pinned layer by coupling of the RL to a SAF HL.

2.4 MTJ process

As discussed previously, the STT-MRAM stack consists of a lot of nanometer thick layers stacked on top of each other. Processing this stack is composed of following important steps: stack deposition, annealing, hard mask (HM) opening, MTJ patterning and encapsulation. A very important strive is to demonstrate a thermal budget of 400°C, such that the technology is fully BEOL compatible. For this a lot of engineering efforts have to be made to the stack. The important processing steps are illustrated in Fig. 2.5(a) and will be briefly discussed in the following sections. A TEM micrograph of a processed ⌀ 60 nm device is shown in Fig. 2.5(b). As opposed to the simplified schematic, the patterning leaves behind a tapered sidewall with a tapering angle around 10-20°.

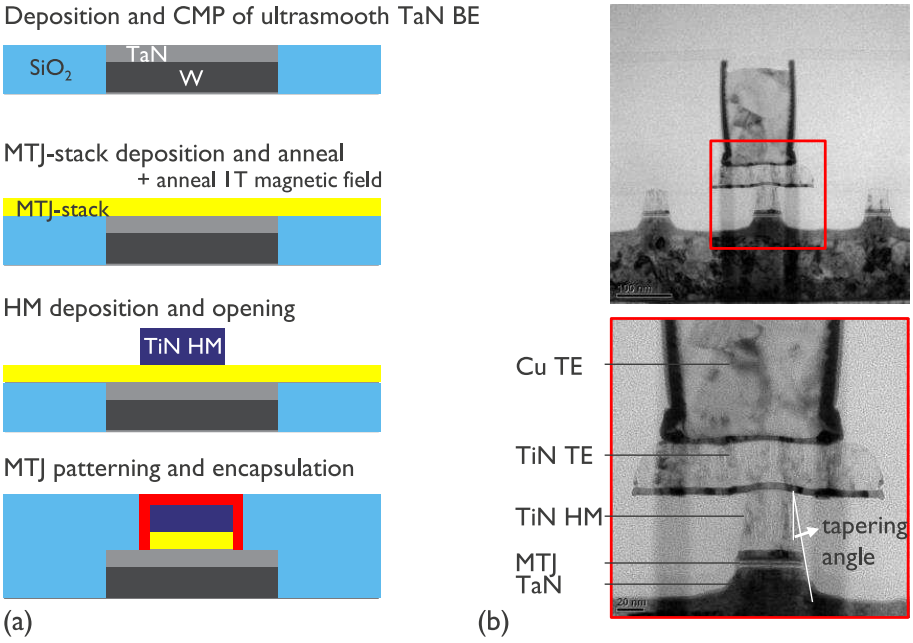

Figure 2.5: *(a) Simplified schematic of the MTJ process. (b) TEM micrographs of a ⌀ 60 nm nominal device shows the tapered side wall of the MTJ pillar.*

2.4.1 MTJ deposition and annealing

Full STT-MRAM stacks are deposited by magnetron sputtering in a Canon-Anelva EC7800 cluster tool on 300 mm Si wafers. The stacks are post-growth annealed in a 1 T perpendicular magnetic field in vacuum at temperatures ranging from 300°C to 400 °C. The MTJ deposition begins on top of the BE. To ensure reduced roughness at the CoFeB/MgO interface we start from an ultrasmooth BE, by depositing TaN on top of the W BE, it can be polished to a roughness of only 0.05 nm [61]. In the case of bottom pinned configuration, the HL and the RL need to be deposited before the MgO/CoFeB interface. This increases the roughness considerably, resulting in a reduction of the TMR. Another way to reduce the roughness at the MgO/CoFeB interface is by using a top pinned stack, where the RL and HL are on top of the MgO. However, the HL is not stable in this case, since the CoFeB RL layer does not provide a stable enough template for the HL. When annealed at 400°C, the HL fails. For a 400 °C compatible top-pinned stack, different solutions are necessary, as presented in [111]. We will mainly focus on the bottom-pinned alternative.

Within the stack there are quite a few ultrathin spacers and seed layers. They have various roles. Seed layers enable the growth of the desired crystal orientation. Spacer layers serve as coupling layers as in the case of the Ru in the SAF or Ta between the Co/Pt SAF and CoFeB RL. The Ta next to the CoFeB also serves as a boron getter. CoFe cannot be deposited amorphously, therefore boron is added. However, the boron does not contribute to the magnetization. During the anneal it is hypothesized that boron diffuses out towards the Ta spacer, at the same time, the MgO crystal orientation acts as a template for the CoFeB. This way a high quality bcc (001) crystal texture is obtained for both MgO and CoFeB, with a good lattice match. Normally CoFe would crystallize into a fcc (111) texture, which results in low TMR.

2.4.2 Hard mask opening

A good hard mask has some important specifications. It needs to be hard material, being able to withstand the MTJ etch. It needs to be conductive, making sure no extra series resistance is added to the MTJ. Furthermore, the hard mask needs to be resistant against oxidation. For all these reasons we make use of TiN.

100 nm TiN is deposited on the MTJ stack and afterwards opened with a 193i lithography step. Now the MTJ stack can be etched to small pillars.

2.4.3 MTJ patterning and encapsulation

One of the most important process steps is the MTJ patterning. We consider two common etch techniques, reactive ion etch (RIE) and ion beam etch (IBE). The advantage of RIE is that it makes use of a high density plasma, which allows processing very small pitches. However, controlling the redeposition of etched material on the side walls is challenging, because the high density plasma provides stimulated physical and/or chemical interactions across the whole surface. In contrast, IBE is a directional etch, that physically sputters the MTJ. The directionality leads to very controlled etching of the pillar and allows clearing of the sidewalls, which prevents shorts and also removes induced damage. However, due to the shadowing effect, i.e. when the MTJ pillars are positioned at narrow pitch, the beam may sputter neighboring pillars.

Furthermore, after etch an additional post-etch treatment removes the damage inflicted on the sidewall and prevents the creation of a short across the MgO barrier.

Finally, the MTJ pillar is encapsulated with Si_3N_4 by plasma-enhanced chemical vapor deposition (PECVD), where Si is sputtered in the presence of a nitrogen gas flow. Currently, this process is changed for the IBE to a physical vapor deposition (PVD), where Si combines with N and gets deposited along the pillar surface (here a pure N_2-flow is used). The encapsulation protects the MTJ against moisture.

The patterning processes results in tapered sidewalls of the pillar with a tapering angle ranging from 10-20°.

2.4.4 Conclusion

The knowledge of the full configuration of the stack is important. We will see in Sec. 3.2.3 that the TaN ultrasmooth BE and the TiN HM are bad thermal conductors and result in a fast build up of temperature at the MgO level. In addition, the processing and changes to the stack have a significant influence on the reliability, as will be discussed in chapter 5.

The STT-MRAM stack has seen considerable process and stack adaptations, a lot of engineering is required to optimize all important parameters like PMA,

reference layer stability, high TMR, low switching power, reliability. In this thesis, we do not report on all these iterations that are necessary to engineer the STT-MRAM stack. However, we unravel fundamental mechanisms necessary to explain the reliability of the STT-MRAM technology.

2.5 Main studied STT-MRAM stacks in this thesis

Throughout this thesis multiple STT-MRAM stacks have been studied. The major changes are arranged per chapter in Fig. 2.6. Note that processing conditions have continuously been optimized.

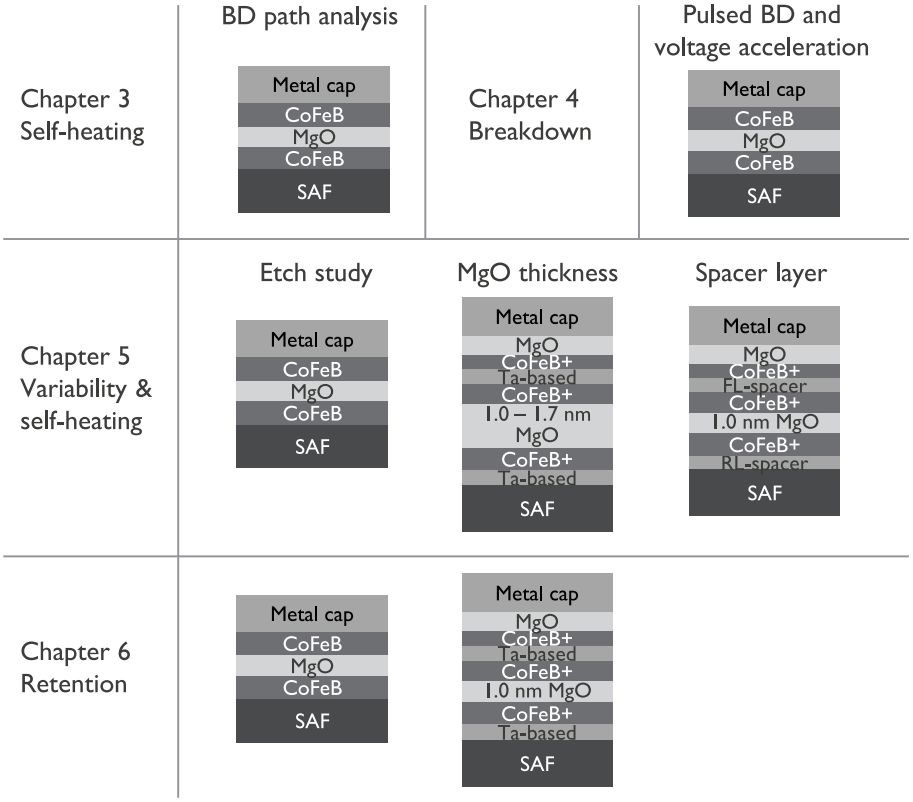

Figure 2.6: *Summary of the major STT-MRAM stacks that have been studied in this thesis.*

2.6 Summary

In this chapter, we have discussed the basic properties in MRAM that enable reading, writing and retaining the bit-state by TMR, STT and thermal stability, respectively. In addition, we consider the resistance area (RA) product, together with the MTJ CD, and explain how these two are important to have an optimal synergy with the CMOS technology. Next, important contributions necessary to achieve a functional STT-MRAM-stack are elaborated, like the introduction of SAF to stabilize the pinned layer and a double MgO to improve the thermal stability of the free layer.

The most important processing steps have been discussed. The MTJ has to withstand a 400 °C annealing temperature, such that STT-MRAM becomes BEOL compatible.

Chapter 3

Modeling and characterization of self-heating

The many nanometer thick layers in STT-MRAM affect not only the electric and magnetic, but also the thermal properties. Since excessive temperature deteriorates all the important MTJ properties, a theoretical model on MTJ cannot be complete without accurate self-heating characterization.

3.1 Introduction

The high current densities required in STT-MRAM can cause significant self-heating. The stack consists of several materials close to the MgO tunnel barrier, where the heat is generated, that poorly conduct heat. Only further away from the MgO are the bit and source lines with good thermal properties. Simulations in literature already indicate very fast and substantial heating, but do not investigate as to why and how the MTJ surroundings will impact the self-heating. In this chapter we analyze the role of self-heating and develop methods to characterize self-heating.

In section 3.2, we introduce the self-heating model and simulate the thermal resistance of the MTJ, using a 2D axis-symmetric model. Due to the high current densities required to switch the MTJ, self-heating impacts the MTJ

properties. We find that temperatures during a breakdown stress can reach up to 500 K and find that the thermal resistance is affected by the thermal boundaries of the simulations. However, these simulations are based on poorly known thermal conductivities of the thin layers in the MTJ-stack. Therefore, in section 3.3, we develop methods to assess the thermal resistance experimentally.

3.2 How self-heating impacts MTJ operation

High voltage pulses are applied to enable switching. As a result, current densities larger than 1 MA/cm^2 are required. The resulting non-negligible power will heat up the MTJ. In Sec. 3.2.4, we will show that self-heating severely impacts the reliability properties and estimating the contribution of self-heating is therefore important.

First, we introduce the self-heating model (Sec. 3.2.1). Then, in Sec. 3.2.2, we explain how to transform the 3D layout into a 2D axis-symmetric geometry. This geometry simplifies and speeds up the simulations. Furthermore, the impact of the thermal boundaries is discussed in Sec. 3.2.3.

3.2.1 Introduction of the self-heating model

As the MTJ is powered up, Joule-heating will occur due to resistive losses. This power dissipation will subsequently lead to a (local) temperature increase, depending on the heat conductivity of the system. The MTJ temperature can be written as:

$$T_{MTJ} = T_{ambient} + \Delta T_{SH}\,, \tag{3.1}$$

with $T_{ambient}$ the ambient temperature and ΔT_{SH} the temperature increase of the MTJ due to self-heating.
The steady-state ΔT_{SH} is assumed to be proportional to the dissipated power ΔP and the thermal resistance R_{th} between the MTJ and the external boundaries of the system:

$$\begin{aligned} \Delta T_{SH} &= \Delta P \cdot R_{th} \\ &= \frac{V_{MTJ}^2}{R_{MTJ}} R_{th}\,. \end{aligned} \tag{3.2}$$

The major challenge, associated with the self-heating assessment, is the evaluation of this R_{th}. At nanometer scale the material properties are changing

compared to macroscopic scale. The materials will not behave as bulk materials, but additional phonon scattering effects will come into play. Therefore, the exact thermal conductance of the complex MTJ stack is not known. The thermal conductivity values used in our simulations are listed in Table 3.1 [11, 64, 72, 81].

Material	Occurrence	κ ($WK^{-1}m^{-1}$)	Reference
SiO_2	Isolation	1.4	[COMSOL]
MgO	Tunnel barrier	48	[NamKoong2009]
W	BE	40	[Lu09]
Cu	TE	250	[Lu09]
TiN	HM, TE	11	[Kim12]
TaN	Ultrasmooth BE	3	[Grayeli11]
MTJ metals	Next to MgO	10	chosen

Table 3.1: *Thin-film thermal conductivity values for the materials occurring in the STT-MRAM-stack*

On large-scale devices, the steady state heat flow can be treated by the continuum model, i.e. Fourier's law, which is expressed in differential form as [43]:

$$\overrightarrow{q} = -\kappa_{th} \nabla T \,, \tag{3.3}$$

with $\overrightarrow{q}$ being the local heat flux density, κ_{th} the material's thermal conductivity and $-\nabla T$ the negative local temperature gradient. The previous conductance equation, written in terms of extensive properties, can be reformulated in terms of intensive properties, such that the equation expresses a quantity independent of distance. This is a similar approach as for Ohm's law for electrical resistance. For the heat equation, this results in:

$$\Delta T = R_{th} \cdot Q \,, \tag{3.4}$$

with Q the total heat flux and R_{th} the thermal resistance. A large thermal resistance can result in a large increase in temperature. The time-dependent heat equation is described as:

$$C_s \frac{\partial T}{\partial t} = \nabla \cdot (\kappa_{th} \nabla T) + \overrightarrow{q} \,, \tag{3.5}$$

where C_s is the heat capacity per unit volume. This equation describes how the heat will be distributed locally around the heat source $\overrightarrow{q}$. It is this equation that is solved, using the finite element method.

3.2.2 From a 3D model to a 2D axis-symmetric model

To simplify the model, we make use of a 2D axis-symmetric model. Such a model is easier to parametrize and the required numerical effort is significantly reduced. The results can then be rotated over 360°. This approach is straightforward for the small circular pillar MTJ. However, incorporating the top electrode (TE) and bottom electrode (BE) into a 2D axis-symmetric model, requires more effort, because they form a cross structure, which in case rotated over 360°results in an incorrect representation. As such, approximations have to be made.

Fig. 3.1(a) illustrates the full 0T1MTJ cell with the cross structure of the BE and TE. The arrows illustrate the dimensions of the metal lines. At the MTJ, i.e. in the middle of the cross, the width of the BE and TE is given by W_{BE} and W_{TE}, respectively. The BE and TE broaden to a width $W_{BE,ext}$ and $W_{TE,ext}$, respectively. The length of the BE and TE is given by H_{BE}, H_{TE}, and $H_{BE,ext}$, $H_{TE,ext}$ for the broadened region. In Fig. 3.1(b), the cross section through the MTJ is shown. The MTJ pillar is tapered with a tapering angle of 15°. The white arrows represent the heat flow, which will be dominated by the good thermal conducting materials, i.e. the metals. For a schematic representation of the equivalent thermal circuit we distinguish the thermal resistance of the W BE, the ultrasmooth TaN BE, the MTJ, the TiN HM, the TiN TE, the Cu TE and the SiO_2 isolation. The heat is generated mostly around the MgO barrier, where most of the voltage drops.

The generated heat at the MgO layer preferentially flows in materials with high thermal conductivity. The SiO_2 surroundings have poor thermal conductivity, therefore the heat will mainly stay in the metal layers. In a 2D axis-symmetric model, the heat flow through the metal layers needs to be maintained. To achieve this, the real structure is transformed to an equivalent axis-symmetric configuration. This is done as follows: the 4 contact lines of the cross structure are folded up and down to align them with the MTJ pillar (Fig. 3.2, Fig. 3.3(a)). The 2 TE-lines fold upwards and the 2 BE-lines fold downwards, the length of the TE and BE, becomes the height H_{TE}, $H_{TE,ext}$, H_{BE} and $H_{BE,ext}$. The cross section of the resulting TE and BE line is now twice the one of 1 TE-line and twice the one of 1 BE-line, respectively. Finally, the area is maintained when transforming the rectangular cross section into a circular cross section with radius R_{2Dax} (Fig. 3.3(b)). From top to bottom we find: Cu TE extension, Cu TE, TiN TE, TiN HM, MTJ stack, TaN BE, W BE and W BE extension, where the height of the layers is equal to the length of the metal path.

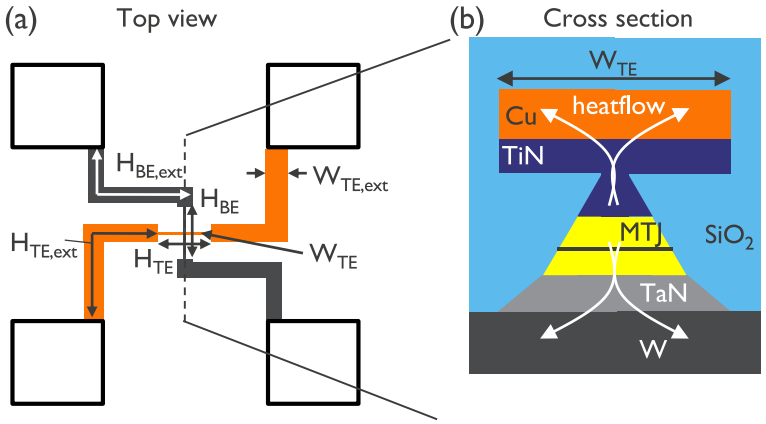

Figure 3.1: *Schematic of the 0T1MTJ layout. (a) Top view of the 4-point cross structure. (b) Cross section view. The white arrows represent the heat flow.*

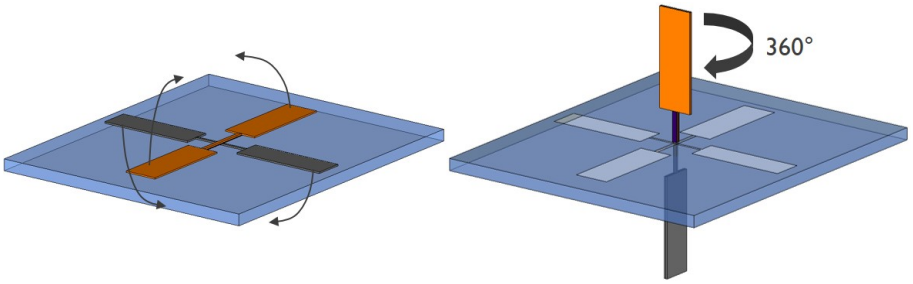

Figure 3.2: *Schematic of the transformation to a 2D axis-symmetric model. The 4 contact lines of the cross structure are folded up and down to align them with the MTJ pillar. In this configuration, 360°-rotation is possible. Note that the cross section area becomes circular, instead of rectangular. The area is maintained however, by adjusting the radius of the rotated cylinder (see Fig. 3.3(b)).*

In this 2D axis-symmetric model, the heat flow through the metal layers is maintained. Only the heat flow through the poor thermal conducting layers is slightly different. This approach, results in the equivalent 2D axis-symmetric model shown in most right schematic in Fig. 3.3(a), which after rotation over 360° around the center dashed-dotted line results in an equivalent 3D model.

To solve Eq. 3.5, we make use of finite element simulations, provided in the COMSOL software environment [78]. Our geometry parameters are based on layout and TEM information. The nominal values are summarized in Table 3.2.

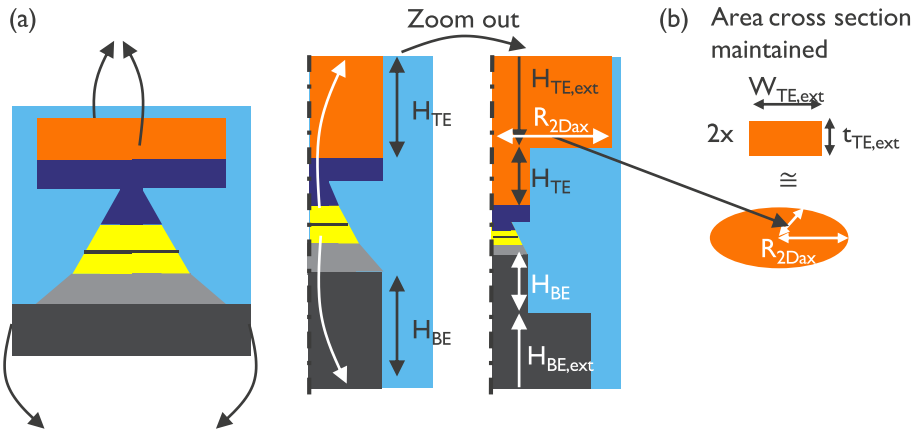

Figure 3.3: *Schematic of the 2D axis-symmetric model. Due to the cross structure (Fig. 3.1(a)), the schematic is not 2D axis-symmetric. However, since the the heat mainly flows through the metal layers, we can transform the cross structure to a 2D axis-symmetric scheme, by folding the TE and BE upwards and downwards, respectively. A zoom out reveals the broadening of TE and BE to TE_{ext} and BE_{ext}. (b) We maintain the cross section area in which the heat flows through TE and BE (2 times, because of 4 point structure), resulting in the 2D axis-symmetric radius R_{2Dax}.*

Material	Occurrence	Height	Cross section
W	BE	3 µm	250 x 160 nm²
W	BE_ext	30 µm	3000 x 160 nm²
TaN	Ultrasmooth BE	17-20 nm	Determined by MTJ etch and tapering angle (15°)
Stack below MgO		12 nm	
MgO	Tunnel barrier	1 nm	
Stack above MgO		11 nm	
2nd MgO	Double MgO	0.5 nm	
TiN	Hard mask	30-70 nm	
TiN	TiN TE	50 nm	500 x 50 nm²
Cu	TE	3 µm	400 x 157 nm²
Cu	TE_ext	30 µm	3500 x 157 nm²

Table 3.2: *Most important nominal dimensions used for simulations of the STT-MRAM-stack*

In each simulation we apply a square voltage pulse and simulate for a stress time of 1 ms. In Fig. 3.4(a), the temperature and voltage at MgO level are plotted as a function of time for a device with a diameter of 150 nm with all the nominal dimensions from Table 3.2. A fast (<1 ns) and a slow heating component can be observed, similar as in [101]. The fast heating component constitutes 70 %-95% of the total heating, depending on the diameter of the MTJ pillar. For our simulated stack, 70% for a 150 nm device, whereas already 95% for a 45 nm device (not shown). The origin of the fast and slow heating component is explained below.

The MTJ stack is surrounded by layers with poor thermal conductivity like the TiN HM, TaN ultrasmooth BE and the SiO_2 isolation. These layers prevent the generated heat at the MgO level from escaping efficiently. After the start of the voltage pulse, the heat flow away from the MgO is limited, resulting in a fast increase of the surrounding temperature. Once the heat reaches the good thermal conducting Cu and W layers, the heat can escape more easily, resulting in the slower heating component.

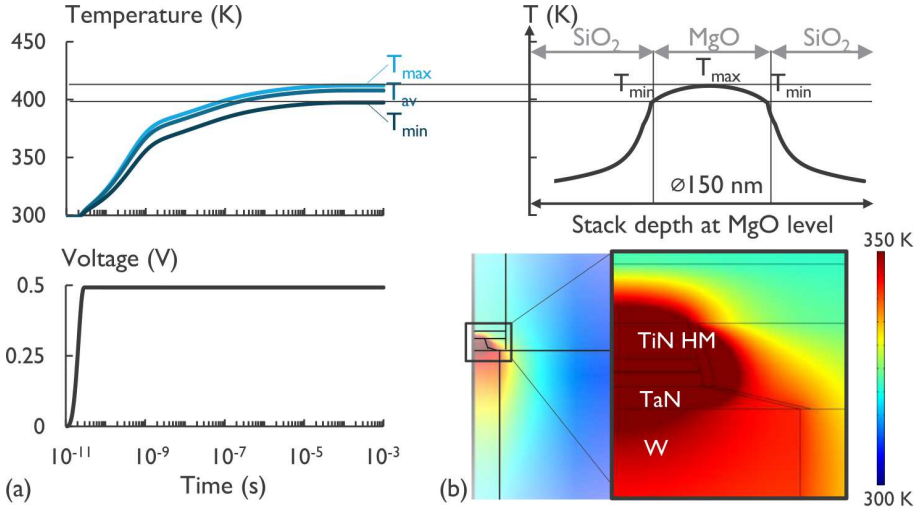

Figure 3.4: *Example of thermal simulation of full 2D axis-symmetric model for a ⌀150 nm device. (a) Temperature and voltage at the MgO level. (b) Bottom: temperature color map at 1 ms. Top figure shows the temperature cross section at MgO level at 1 ms.*

Steady-state is reached after 100 μs (Fig. 3.4(a)). We will see in next section that both the time to reach steady-state and the temperature level at steady-state depend on the geometry of the SiO_2 isolation and TE and BE height, i.e. the total volume of the cell and the location of the thermal boundaries.

In Fig. 3.4(b), the corresponding temperature profile at 1 ms is shown for the full stack (bottom) and at MgO level (top). In the MgO, the minimum temperature T_{min} is 4 % lower than the maximum temperature T_{max}. Thus the temperature profile in the MgO is quite uniform, with a maximum in the middle of the pillar. Of course this simulation assumes that the MgO thickness is uniform over the device area, resulting in a homogeneous distribution of the current. Outside the MgO, i.e. in the SiO_2 isolation, the temperature quickly drops. The impact of the thermal boundaries will be discussed in more detail in the next section.

3.2.3 The importance of the thermal boundary distance

To illustrate the importance of the distance between device and the thermal boundaries, we start from a cell with boundaries close to the MTJ. In this case the ambient temperature is forced very close to the device, resulting in the lower trace in Fig. 3.5(a).

The upper trace takes into account the total TE and BE dimension, up to the contact pads, namely H_{TE}, $H_{TE,ext}$, H_{BE} and $H_{BE,ext}$, and takes into account an SiO_2 isolation width of 10 μm ($W_{isolation}$). The total TE and BE height is $H_{BE,tot} = H_{TE,tot} = 33\,\mu$m, with a broadening of the cross section after $H_{BE} = H_{TE} = 3\,\mu$m, see Fig. 3.3(a) ($H_{BE,ext} = H_{TE,ext} = 30\,\mu$m).

Putting the thermal boundaries further from the device increases the time to reach steady-state. Reaching steady-state depends thus on both the SiO_2 isolation width and the height of the TE and BE. In Fig. 3.5(b) the impact of the height of TE and BE, as well as the total width of SiO_2 isolation on the simulated thermal resistance is plotted. The thermal resistance is extracted at 1 ms as the ratio of the temperature increase and the power:

$$R_{th} = \frac{\Delta T}{P}\,. \tag{3.6}$$

The R_{th} is normalized by the value for 10 nm TE and BE height and an isolation width of 1 nm from the TE and BE edge, i.e. thermal boundaries close to the MTJ. We have simulated 2 areas and find that the impact of the distance to

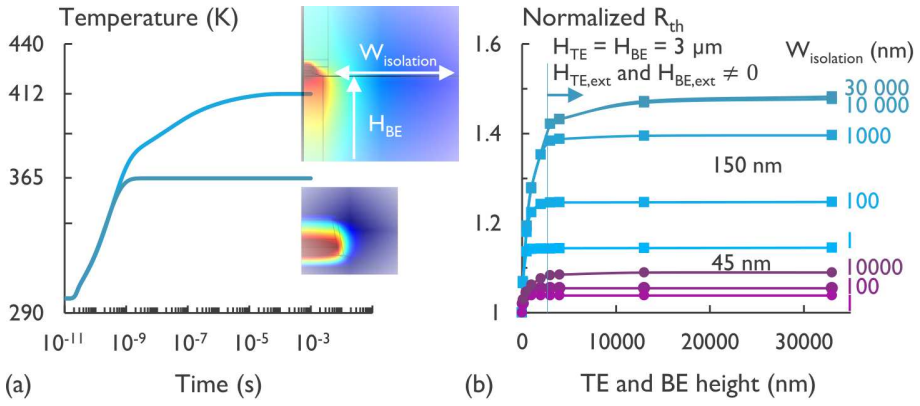

Figure 3.5: *Impact of the distance to the thermal boundary. (a) Temperature as a function of time for thermal boundaries at 1 nm from the TE and BE edge (lower trace) or for the full 0T1MTJ structure (upper trace). (b) Normalized R_{th} for 2 sizes (⌀ 150 nm (blue) and ⌀ 45 nm (purple)). R_{th} is normalized by the condition: $H_{TE} = H_{BE} = 10\,nm$, $H_{TE,ext}, H_{BE,ext} = 0\,nm$ and SiO_2 isolation width $W_{isolation} = 1\,nm$ from the TE and BE edge. When the thermal boundaries are further away from the heat source, R_{th} increases, especially for the large devices.*

the thermal boundary is more severe for a larger MTJ (from Fig. 3.5(b)).

The thermal resistance stabilizes around 10 μm of SiO_2 isolation width and also around 2.5 μm of TE and BE height. To conclude, it is important to simulate with the correct thermal boundaries, certainly for the larger devices.

3.2.4 Temperature effects on device performance

Temperature will impact the device performance and reliability. The switching voltage and breakdown voltage are accelerated by temperature. Self-heating is not negligible and has to be taken into account. In Fig. 3.6(a), the maximum temperature in the structure (at MgO level) in steady-state, is simulated as a function of voltage for different areas (using the full layout dimensions, see Table 3.2). A maximum temperature of more than 500 K at 1 V is predicted for all simulated areas. In this stress voltage range, the device also breaks down, see Fig. 3.6(b).

In Fig. 3.6(b), the 63 % breakdown voltages are shown as a function of temperature for devices with the same electrical CD as simulated in Fig. 3.6(a). Temperature accelerates breakdown, however due to the significant self-heating, the impact of temperature on breakdown is underestimated when only considering the external temperature. The effect of the self-heating on the breakdown modeling and lifetime extrapolation is elaborated in section 5.5.4.

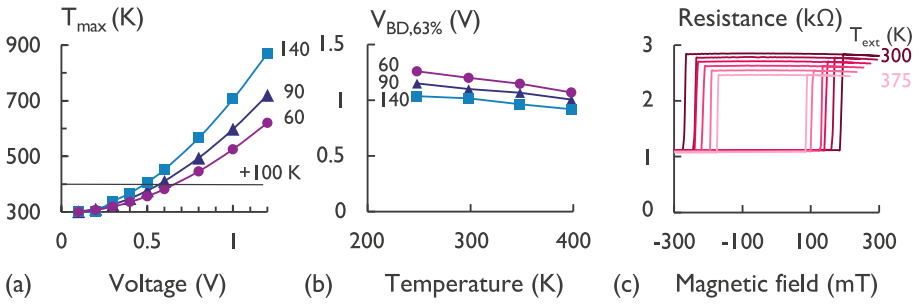

Figure 3.6: *(a) Maximum temperature shows a quadratic dependence on voltage for 3 sizes. (b) Temperature dependence of the 63 % breakdown voltage for 3 sizes (eCD = 60, 90 and 140 nm). (c) Temperature reduces mostly R_{AP}. As a result, the MR degrades with temperature, as well as the coercive field (⌀ 75 nm, 300 to 375 K).*

Finally, in Fig. 3.6(c), TMR loops for temperature in range 300 K to 373 K are plotted. Temperature decreases the R_{AP} and coercivity of the device.

To summarize, self-heating has a significant impact on the MTJ performance. Simulations indicate that a ⌀ 140 nm MTJ already heats up by 100 K at 0.5 V. At the same stress voltage, larger areas heat up significantly more. However, these simulations rely on poorly known thermal conductivities for the nanometer thick MTJ structure. Therefore, precise determination of the actual temperature based on experiments of the actual structure is mandatory to correctly include the self-heating effect in lifetime measurements. For this, in the next section we suggest a number of conceptual methods to experimentally determine the internal temperature at operation voltage.

3.3 Experimental issues in determining MTJ temperature

Self-heating will play an important role in the device performance. Measuring the temperature of an MTJ during operation is thus of fundamental importance.

Several methods have been proposed to measure the MTJ temperature. These methods can be divided in direct and indirect methods.

In direct measurements, the thermal resistance of one device is directly measured. Indirect measurements are based on matching statistical switching measurements to a statistical switching model [120]. Another indirect method is the self-heating simulation based on numerically solving the heat diffusion equation [51, 58]. As discussed in Sec. 3.2 however, the material conductivities for all the nanometer thick layers in the STT-MRAM device are poorly known.

The main problem for a direct measurement is the voltage and temperature dependence of both the parallel and anti-parallel resistance. A lot of models try to explain the voltage and temperature effect simultaneously [105, 12, 70, 87]. However, considering the full STT-MRAM stacks and the unknown effects of patterning, relying on these models can be precarious. In addition, these models do not take into account the temperature increase due to an applied voltage bias.

A proposed direct method makes use of spin wave noise spectroscopy [68]. The authors decouple the effects of all the various magnetic layers. The authors determine then, based on the thermal signature of the magnetization, the MTJ temperature. Besides that the theory is quite complex, the confidence interval is more than 300 K, making this technique of no practical use.

In this thesis, we attempted to develop 2 direct measurement methods, which unfortunately did not prove to be successful. An intuitive measurement would consist of determining the MTJ resistance as function of time while the temperature ramps-up due to self-heating. Since this ramp-up is expected to happen in sub-ns range (Fig. 3.4), this measurement is impractical to execute. A second method relies on the use of a breakdown path inside the MTJ as a thermometer, independent of the magnetization (Sec. 3.3.1). The breakdown path behaves metallic-like and generates heat like in a pristine MTJ.

Furthermore, we have developed 2 indirect methods, that offer more potential. The first one makes use of the breakdown statistics, in contrast to relying on switching models (Sec. 3.3.2). This method requires prior knowledge of the breakdown mechanism and statistics, which will be elaborated in chapter 4 and 5, therefore the method will only be practically demonstrated in Sec. 5.5. The second indirect method compares the thermal stability derived from current switching and magnetic field switching (Sec. 3.3.3). The difference in the thermal

stability can be explained by self-heating. Prior knowledge of thermal stability of chapter 6 is necessary, therefore this method is executed in Sec. 6.8.

3.3.1 Direct measurement using a breakdown path

This method aims at extracting the thermal resistance using a breakdown path (BD path) as follows:
(i) We generate deliberately a BD path in a device.
(ii) We measure the temperature coefficient of the post-BD resistance.
(iii) We derive the thermal resistance of the BD path.

To apply this method, we make use of cells with a four point measurement structure to accurately determine the resistance of the MTJ. The BD path will act as an internal, metal-like heat source. Note that this internal heat source in the broken MgO is at the same place as in a pristine MTJ. In addition, due to breaking down the MgO layer, the device is **magnetically dead** and the MTJ behaves as a metal connection between TE and BE.

(i) Generating the BD path

After breakdown, the device has a small resistance around 40 Ω, resulting from the BD path and the metal pillar.

(ii) Measuring the temperature coefficient of the resistance

Performing R-V measurements after breakdown at different temperatures, results in a linear increase of the post-BD resistance with temperature (see inset Fig. 3.7(a)), the slope is called the temperature coefficient of resistance (TCR).

During an R-V measurement, the applied power will result in an increased temperature, via self-heating. As can be seen from Fig. 3.7(a), the resistance depends linearly on power. Extrapolated to zero power, R_0 is obtained, R_0 is then used for the calculation of the TCR. In fig. 3.7(b), the TCR for $\varnothing$ 60 nm devices are shown as a function of breakdown voltage, no clear trend is observed.

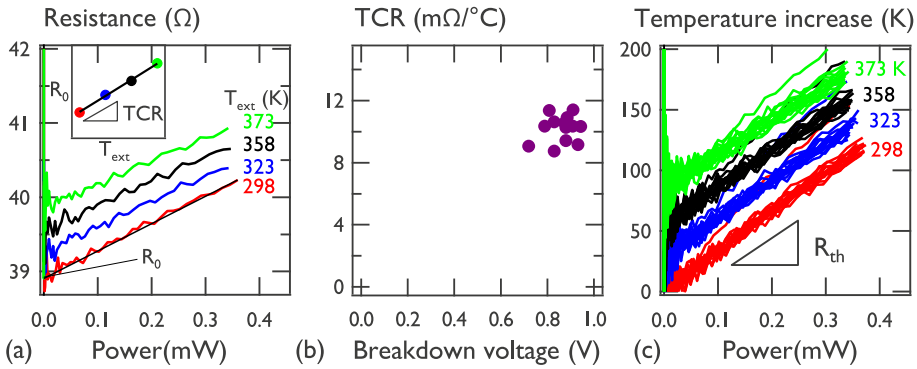

Figure 3.7: *Experimental determination of the thermal resistance of the BD path in ⌀ 60 nm devices. (a) Post-BD resistance as a function of power for 4 temperatures (298, 323, 358, 373 K). R_0 is determined as the intercept of a linear fit between resistance and power. (inset) Principle of extraction of the TCR. (b) TCR of ⌀ 60 nm broken devices as a function of breakdown voltage (current compliance during breakdown was 10 mA). (c) Calculation of the temperature increase ΔT via Eq. 3.7, R_0 at 298 K is used for all temperatures. The slope corresponds to the thermal resistance.*

(iii) Calculating the thermal resistance of the BD path

Finally, the temperature during the R-V can be calibrated as:

$$\Delta T = \frac{R - R_0}{TCR} \,. \tag{3.7}$$

From Eq. 3.6 we know that the ratio of ΔT and power is equal to the thermal resistance, which is the slope of the curves in Fig. 3.7(c).

Unfortunately, due to the low resistance of the BD path versus the resistance of the metal pillar, the assumption that the heat is mostly generated in the broken MgO, is not valid. There will be a more uniform heating throughout the device. Therefore, we cannot rely on the extracted thermal resistance.

We attempted to overcome this issue by changing the BD path resistance. By adjusting the current compliance, we hypothesize that the size of the BD spot can be controlled. Small BD spots have a higher resistance and therefore most of the voltage drops over the BD path. Small BD spots will thus generate more heat around the BD path.

Adjusting BD spot size by limiting the current compliance

By reducing the current compliance to values close to the current just before breakdown, one can limit the size of the BD spot. In Fig. 3.8, breakdown traces are shown as function of voltage. The post-BD resistance depends on the current compliance.

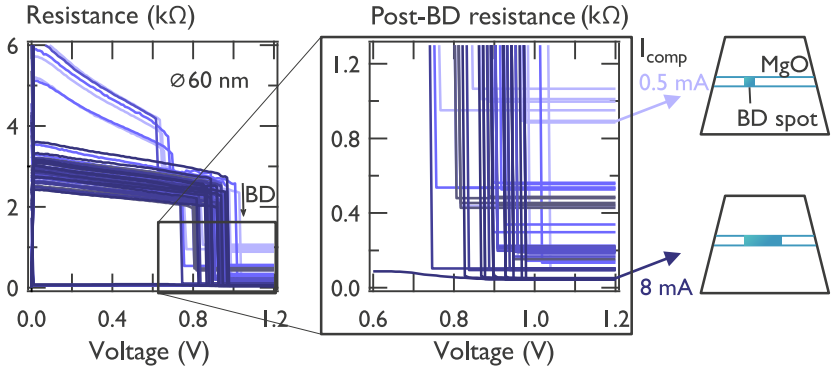

Figure 3.8: *R-V breakdown traces for ⌀ 60 nm devices with different current compliances (I_{comp} = 0.5 (bright), 0.8, 1, 2, 4 and 8 mA (dark)). The size of the BD spot can be controlled by the current compliance. Lower I_{comp} results in larger resistance after BD and thus smaller BD spots (illustrated by the schematic on the right).*

The total resistance can be derived from the BD path in parallel with the remaining MgO. In case the area of the BD spot becomes negligibly small, the total resistance increases to the pristine MTJ with the resistance determined by the entire MgO layer.

Two issues occur due to these small BD spots. (1) High power causes further degradation of the BD path and extension of the BD spot (see traces with I_{comp} =2 mA in Fig. 3.9(a)). For the smallest BD spot (2) The remaining MgO is still magnetically functioning, such that switching events can happen (left graph in Fig. 3.9(b)). We discuss these issues in more detail below.

To solve the first measurement issue, we limit the applied power and thus reduce the voltage range in the R-V sweeps after breakdown. The second issue can be solved by applying an external magnetic field. We verify whether the jumps in Fig. 3.9(b) are caused by the switching from AP to P state as follows: the

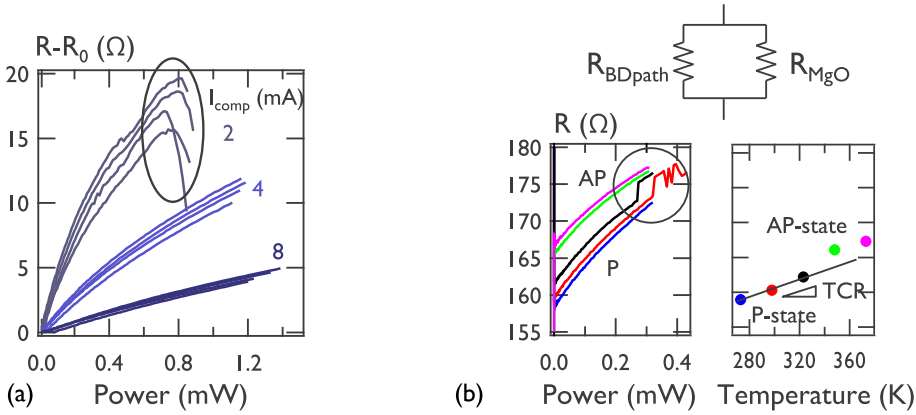

Figure 3.9: *(a) ΔR-P traces for TCR extraction can induce further expansion of the BD spot at high power (⌀ 60 nm at 3 current compliances). (b) The TCR extraction can be influenced by switching from P-to-AP at higher temperatures and higher power (⌀ 60 nm device with $I_{comp} = 2\,mA$).*

total post-BD resistance is

$$R_t = \frac{1}{\frac{1}{R_{BD}} + \frac{1}{R_{MgO}}}\,. \tag{3.8}$$

R_{MgO} is derived from a working ⌀ 60 nm device, the R_{AP} and R_P are around 5 kΩ and 3 kΩ, respectively. Considering that the resistance of the BD path for I_{comp} =2 mA is around 170 Ω (see Fig. 3.9(b)), then substituting R_{AP} and R_P as R_{MgO}, the total resistance changes with 5 Ω, which matches the observed resistance jumps.

The fact that switching events can occur in a broken device, is an indication the device is not magnetically dead. The closer the breakdown compliance approaches the current just before breakdown, or in other words the smaller the BD spot, the more the MgO layer contributes to the total resistance after BD.

Following the extraction method from Fig. 3.7, we now observe that the TCR depends on the BD spot size (Fig. 3.10(a)). To explain this dependence, we simulate the BD path as follows: (1) The BD path has the same electrical and thermal conductivity as the surrounding metal. (2) The MgO has for the simulation zero temperature dependent electrical conductivity. The results are shown in Fig. 3.10.

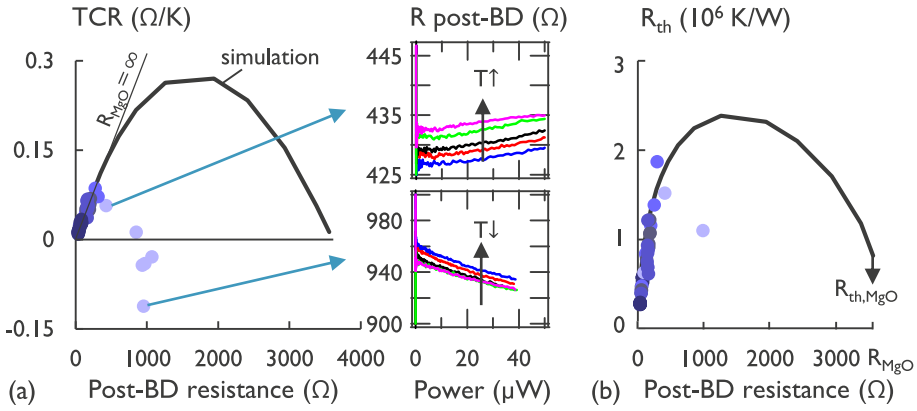

Figure 3.10: *Extracted TCR (a) and thermal resistance R_{th} (b) for BD paths in the ⌀ 60 nm devices from Fig. 3.8. Solid line is the simulation for increasing BD spot size, where the MgO has no temperature dependence. At higher post-BD resistance, i.e. smaller BD spot size, the TCR is dominated by the MgO. Inset: sample of TCR extraction for positive (upper) and negative (lower) TCR.*

The TCR first increases with the post-BD resistance (Fig. 3.10(a)), which is related to the increasing resistance of the breakdown path when the BD spot size decreases. The dependence is explained as follows:

$$\begin{aligned} R_{BDspot} &\approx \rho_0 \frac{l}{A_{BDspot}} \cdot \left(1 + \alpha \Delta T\right) \\ TCR = \frac{R_{BDspot} - R_0}{\Delta T} &\approx \frac{\rho_0 l}{A_{BDspot}} \cdot \alpha \,, \end{aligned} \tag{3.9}$$

where ρ_0 is the resistivity of the BD path at reference temperature, l is the length of the BD path, i.e. MgO thickness, A_{BDspot} is the area of the BD spot and α is the temperature coefficient. If the BD spot decreases, the TCR will increase, and since the post-BD resistance has a $1/A_{BDspot}$ dependence, the TCR will increase linearly with post-BD resistance. However, there is a tipping point, at this point the resistance of the BD path becomes comparable to the resistance of the surrounding MgO layer. Therefore, an increasing amount of current will pass through the MgO, which has no temperature dependence in the simulation and in reality a negative TCR, see the negative TCR data points for high post-BD resistances in Fig. 3.10(a). As a result, at higher resistance after BD, the TCR decreases.

Furthermore, the extracted thermal resistance has a similar dependence and

depending on the post-BD resistance, i.e. the BD spot size, the thermal resistance can be higher or lower than a pristine MTJ with thermal resistance $R_{th,MgO}$ and resistance R_{MgO}. Knowing the actual BD spot size can allow calibration of the simulations with the measured data, however from TEM analysis the size of the BD cannot be extracted. Therefore, we are left with two unknown parameters: (1) the area of the BD spot and (2) the RA of the post-BD MTJ pillar, i.e. the layers without MgO. To uniquely fit the data, it is required to determine at least one.

In its present form, this method cannot be used to extract the thermal resistance. For a large BD spot, the assumption that the heat is mostly generated around the BD path is violated, resulting in an underestimation of the thermal resistance. In case the BD spot is too small, the assumption that the device is magnetically dead and as such, on a thermal level the MTJ behaves now as a metal, is violated. Therefore, we decided not to further develop this method to extract the thermal resistance in a direct way.

3.3.2 Indirect measurement using different oxide thicknesses

In this indirect method, we will make use of the breakdown statistics and breakdown modeling of the MTJ to determine the internal temperature and thermal properties. Due to self-heating, the temperature just before breakdown is different from the external temperature, therefore the temperature acceleration of the time-to-breakdown needs to be corrected for this change in temperature. In thicker oxides, it can be justified to assume that the MTJs heat up only slightly, because the current is low (Fig. 3.11(a)).

The thermal resistance is derived as follows: (1) determination of the temperature acceleration of breakdown in the thick MgO. (2) Voltage extrapolation of the temperature acceleration α_T to a voltage stress range where the thin MgO breaks down. (3) Correct the observed temperature acceleration obtained from only the external temperature in time-to-breakdown measurements on thin MgO, such that it matches with the extrapolated α_T, using the power right before breakdown as input parameter and R_{th} as fitting parameter (see Fig. 3.11(b)). As a result the actual device temperature right before breakdown is determined.

Because knowledge acquired in chapter 4 and 5 is necessary to fully understand this method, it will be elaborated in section 5.5.

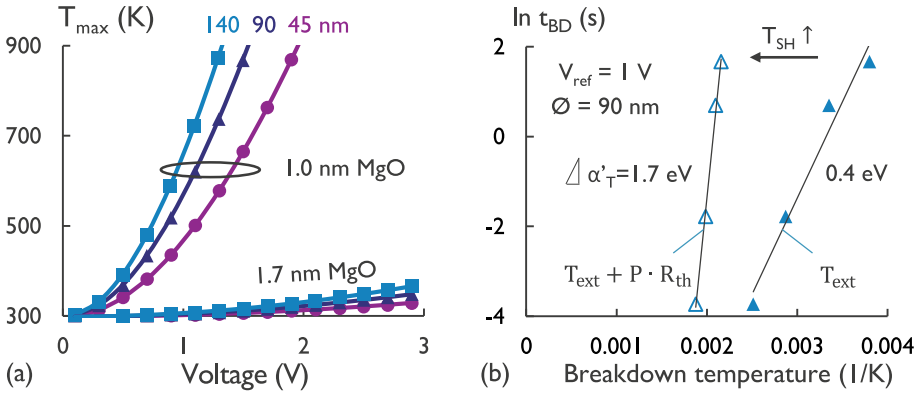

Figure 3.11: *(a) Simulated maximum temperature in 1.0 and 1.7 nm MgO for 3 sizes as a function of voltage. (b) Methodology where R_{th} is fitted for a specific dataset to result in a temperature slope $\alpha_T \approx 1.7\,eV$, which was derived from the thick MgO.*

3.3.3 Indirect measurement via thermal stability

In the final conceptual method, the thermal stability Δ of the AP- or P-state is used as temperature sensor. The thermal stability depends on the energy barrier between the AP- and P-state. To characterize the thermal stability, (i) accelerated stress conditions are necessary and (ii) extrapolation to off-stress is required. At accelerated stress conditions the switching is fitted by a switching model. There are two common models available, a macrospin model, where the magnetization behaves as a single spin, and a domain wall mitigated model, where a domain wall is nucleated and propagates through the device. There is, however, no consensus on which model is applicable for our MTJ devices.

Switching between these states can be induced by a magnetic field or current. In the case of current-induced switching, the MTJ is expected to experience self-heating before switching. By comparing the extracted Δ from a magnetic field-induced switching with current-induced switching, the self-heating component can be extracted and the internal temperature can be determined.

More information regarding magnetization dynamics and the switching models is required to fully understand all the switching statistics and fitting procedures used in this method. Therefore, this method will be elaborated in section 6.8, after a thorough discussion in chapter 6.

3.4 Conclusions

Thermal simulations indicate that self-heating can result in temperatures of 500 K at breakdown stress conditions. In these simulations the thermal boundaries affect mostly the time constants needed to reach a steady-state temperature. The thermal resistance derived from these simulations depends, however, on poorly known thermal conductivity values for the multiple thin film layers of which the STT-MRAM is composed.

In order to assess the thermal resistance, 2 indirect conceptual methods have been developed. These methods rely on the breakdown statistic and the thermal stability extraction method, respectively. We will demonstrate these methods in Sec. 5.5 and 6.8. An accurate estimation of the thermal resistance and self-heating, is necessary to correctly assess the impact on the MTJ properties.

Chapter 4

Breakdown analysis of MgO

A new pulsed-based breakdown measurement method demonstrates that MgO degradation is cumulative in nature. The voltage power-law describes best the lifetime data.

4.1 Introduction

Dielectric breakdown has been intensively investigated for gate oxides like SiO_2 and high-k dielectrics [129, 57, 75, 31]. Also low-k dielectrics between interconnects in the BEOL have been intensively investigated [26]. Prior obtained knowledge can be used to understand breakdown of the MgO in STT-MRAM.

For transistors, the FEOL SiO_2 oxide thickness has been scaled drastically until 2005. In 2005 SiO_2 was replaced by a SiO_2/high-k dielectric stack with equivalent oxide thickness $< 1\,\text{nm}$, but the physical thickness increased to $> 2\,\text{nm}$. In Fig. 4.1, the evolution of the gate oxide thickness is shown. The implementation of high-k dielectrics in the gate oxide increased the physical oxide thickness without reducing the stack capacitance. This thickness increase resulted in better dielectric breakdown reliability. Gate oxide breakdown became a less critical problem.

In an MRAM stack, the physical thickness of the MgO tunnel barrier is only 1 nm. This thickness is chosen to get an optimal resistance for integration

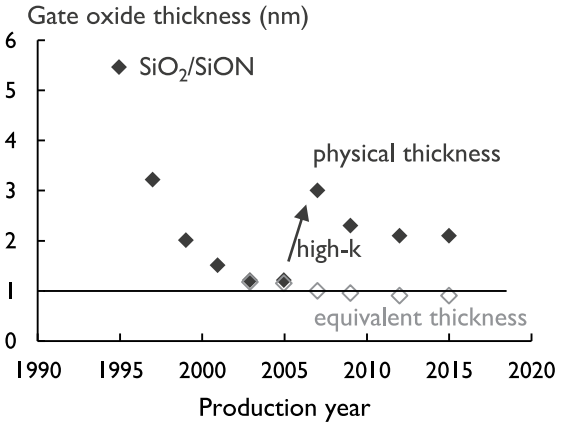

Figure 4.1: *Evolution of the physical and equivalent oxide thickness over the last two decades of MOSFET scaling. The scaling trend shows a clear stagnation after the 90 nm node in 2003. The introduction of high-k in the gate oxide stack results in low equivalent thickness, despite thicker physical dielectric thickness. The MgO used in STT-MRAM is only 1 nm thin. Replotted from [14].*

with CMOS. To switch an MTJ, high current pulses of several MA/cm^2 are necessary. Therefore, breakdown is a very important reliability concern for STT-MRAM. The stress can cause breakdown of the ultra-thin MgO. Another major concern is the strong variability in breakdown that comes with such thin dielectrics.

In this chapter, we review the Weibull distribution and lifetime requirements in Sec. 4.2. Furthermore, we develop a new maximum likelihood fitting method, which simultaneously fits the full dataset for the Weibull slope β, the reference scale parameter η_r and the voltage acceleration exponent n (Sec. 4.3). This simultaneous fit is more robust and takes into account the anomalies present due to variations in the device average stress voltage and the device series resistance.
Furthermore, we discuss the breakdown mechanism in thin dielectrics, more specifically the formation of a percolation path (Sec. 4.4).
We do not simply report the breakdown analysis of the MgO layer for different wafers, but perform an in-depth analysis of both the breakdown measurement techniques as well as the acceleration models. This results in two main observations:
(1) constant voltage stress, ramped voltage stress and pulsed voltage stress, result in the same extracted breakdown parameters, within the error bar of the

measurements (Sec. 4.5).
(2) The breakdown time can be expressed as the cumulative pulse time. This demonstrates the cumulative nature of the oxide degradation process, and excludes the effect of charging and relaxation effects (Sec. 4.5.3).
Finally, in Sec. 4.6, different acceleration models are discussed. In addition, we show that the power-law best describes the voltage acceleration, using a developed pulsed breakdown technique, capable of measuring, within a reasonable time, breakdown times which span eleven decades in time.

4.2 Weibull distribution and lifetime requirements

If we could stress the exact same device multiple times with the same stress conditions, the time to breakdown would be statistically distributed. The probability density function (pdf) $f(t)$ is used to describe the probability $f(t)\,dt$ that a device will fail between time t and $t + dt$. The total probability or the area under this curve is equal to 1 for infinite time. The progress of failure is established by the cumulative distribution function (cdf), $F(t)$, which is simply the integral of the pdf. $F(t)$ gives the probability that a device will fail before a certain time t:

$$F(t) = \int_0^t f(\tau) d\tau \,. \tag{4.1}$$

In case we investigate the failure of a system, the system can be a parallel system or a series system. In a parallel system the system can still function even if a specific unit from this system has failed. An example of such a system is Christmas lights that are connected in parallel. Failure of such a parallel system can be expressed with the product of all the individual cumulative distribution functions $F_i(t)$:

$$F_S(t) = \prod_{i=1}^{N} F_i(t) \,. \tag{4.2}$$

If we were to connect our Christmas lights in series, the whole chain would already fail if a single light fails. Now the failure of the series system is given by:

$$F_S(t) = 1 - \prod_{i=1}^{N} \big(1 - F_i(t)\big) \,. \tag{4.3}$$

The series system will be used later on to prove that area scaling results in a vertical shift of the Weibull distribution. First, we will discuss in more detail the Weibull distribution. It is known from literature that intrinsic dielectric breakdown follows a Weibull distribution, which is typical for weakest link

processes [124]. The pdf and cdf are given as:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \tag{4.4}$$

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \tag{4.5}$$

with β the shape factor or Weibull slope and η the scale factor or the 63%-value. Equation 4.5 can be rewritten, introducing the Weibit W(t):

$$W(t) = \ln\left[-\ln\left(1 - F(t)\right)\right] = \beta \cdot \ln t - \beta \cdot \ln \eta_k\,, \tag{4.6}$$

where W(t) is linear as a function of $\ln t$, with a slope β. This is the preferred method to plot breakdown distributions and it will be used in this thesis (Fig. 4.2(a)). η_k is then the 63%-value, i.e. the time at which 63 % of devices have failed while being stressed by a constant stress condition k.

These derivations were performed in the case we apply the same stress conditions. However, a similar Weibull distribution is found when ramping the stress voltage until failure. Measuring this way, results in a statistical distribution of the

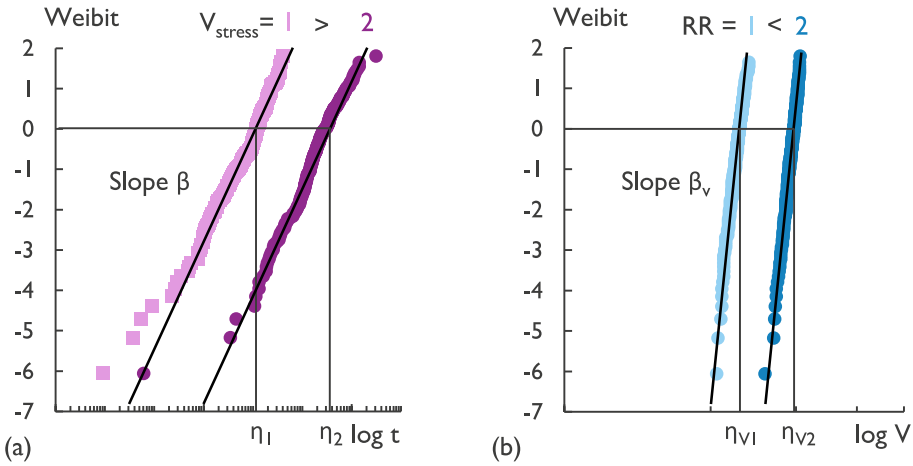

Figure 4.2: *Example breakdown distributions for 1.7nm thick MgO, which are well described by a Weibull distribution. (a) Breakdown time distribution for MTJs measured during 2 constant stress conditions V_{stress} 1 and 2 (V_{stress} 1 > V_{stress} 2). (b) Breakdown voltage distribution for MTJs measured during a ramped voltage stress for 2 ramping speed conditions (ramp-rate RR 1 and 2, RR 1 < RR 2).*

breakdown voltage (Fig. 4.2(b)):

$$W(V) = \ln\left[-\ln\left(1 - F(V)\right)\right] = \beta_v \cdot \ln V - \beta_v \cdot \ln \eta_v \,, \tag{4.7}$$

where W(V) is linear as a function of $\ln V$, with a slope β_v and 63 %-breakdown voltage η_v, considering a ramped voltage stress with a ramp-rate v.

These breakdown distributions follow the Poisson area scaling rule. This means that the breakdown distributions, F_{A_1} and F_{A_2}, of two devices with a different area, A_1 and A_2, respectively, are shifted on a Weibull plot over a vertical interval equal to the logarithm of the ratio of the areas. This can be proven by considering the reliability of a larger area A_1 as the reliability of a series of n devices with Area $A_2 = A_1/\mathrm{n}$, which all have the same Weibull statistics.

$$\begin{aligned} \ln[1 - F_{A_1}(t)] &= \sum_n \ln[1 - F_{A_2}(t)] \\ \ln[1 - F_{A_1}(t)] &= n \, \ln[1 - F_{A_2}(t)] \end{aligned} \tag{4.8}$$

$$\begin{aligned} \ln\big(-\ln[1 - F_{A_1}(t)]\big) &= \ln\left(\frac{A_1}{A_2}\right) + \ln\big(-\ln[1 - F_{A_2}(t)]\big) \\ W_1(t) &= \ln\left(\frac{A_1}{A_2}\right) + W_2(t) \,. \end{aligned} \tag{4.9}$$

Typically, we do not have measurements of the actual area of the devices, and since the etch process has a strong impact on the resulting MTJ pillar diameter, relying on the nominal area would result in an error. We therefore make use of the measured parallel resistance and the RA. The RA is assumed to be constant for the different areas. Using Eq. 2.5, the area scaling term can be written as a fraction of resistances instead of areas:

$$\begin{aligned} \ln\left(\frac{A_1}{A_2}\right) &= \ln\left(\frac{\frac{RA}{R_{p1}}}{\frac{RA}{R_{p2}}}\right) \\ &= \ln\left(\frac{R_{p2}}{R_{p1}}\right) . \end{aligned} \tag{4.10}$$

In Sec. 5.5, we will explain how self-heating of the MTJs causes an apparent deviation from this Poisson area scaling rule.

To estimate the cumulative distribution F(t), it is necessary to rank the data from first failed to last failed. In case our breakdown distribution contains N

devices, an easy way to estimate the cdf of the ith value, would be by converting to cumulative percentages (i/N). It has been shown, however, that by using the cumulative percentages, the low and high ordered values show an unacceptable bias. The exact ranking distribution is described by a beta-distribution [60]. We will use the median value of this distribution. More specific, we will make use of an approximation of the median value, introduced by Benard [6]:

$$F(t_i) = \frac{i - 0.3}{N + 0.4}\,, \tag{4.11}$$

This ranking algorithm is used to calculate the Weibit values of the breakdown distributions in Fig. 4.2. Since time-to-breakdown is a widely spread statistical parameter, sufficient data is necessary to predict whether a device will operate reliable.

In general, reliability is defined as "the probability that an item will perform a required function under stated conditions for a stated period of time" [30]. For STT-MRAM the required function is to perform as a working memory cell that can cycle and has an acceptable resistance and TMR. In STT-MRAM, the MTJ works under pulsed voltages for both read and write. The stated period of time will be the lifetime of the product and is generally taken to be 10 years. The lifetime is considered as an endurance test, since STT-MRAM is a cycling memory technology that could replace DRAM or SRAM. For these memories unlimited endurance is required, often stated as 10^{15} cycles. Because 10^{15}, 100 ns write pulses, is equivalent to practically 10 years of constant stress lifetime, we will keep reporting a lifetime of 10 years.

Finally, the requirements depend on which type of memory technology STT-MRAM will replace. Actual endurance requirement benchmarked on modern processors, are summarized in Table 4.1. For example the conditions for replacing an SRAM L2-cache would result in an actual endurance requirement of 10^{12} cycles for each MTJ cell in a 1 Mbyte memory [59] (see Table 4.1). Lower level caches are smaller in size, but need to operate faster. An L3-cache implementation of STT-MRAM would have a memory size of 6 MB and an actual endurance requirement of 10^{10} cycles. In other cases, like for embedded NVM implementations of STT-MRAM, such as in wearables and Internet-of-things [69, 73], the actual required endurance is also below 10^{15} cycles [59].

Actual breakdown measurements occur at accelerated stress conditions. Otherwise the stress until breakdown would require 10 years. In addition, the number of cells is limited compared to the given memory size. Making a correct lifetime prediction given the required specifications is approached as

Example use case	Memory size (MB)	Assumptions	10 year endurance requirement
L2 cache	1	10^8 access/sec, 40% write traffic	$7.7 \cdot 10^{11}$
L3 cache	6	10^7 access/sec, 40% write traffic	$1.3 \cdot 10^{10}$
Unified eNVM	32	1.6 GBps, 64-bit I/O, 100% write traffic	$1.6 \cdot 10^{10}$
IOT unified	1	400 MBps, 64-bit I/O, 1% duty cycle	$1.3 \cdot 10^{9}$
R.A.A.	n/a	50 ns attack period, 100% duty cycle	$6.3 \cdot 10^{15}$

Table 4.1: *Example cases for STT-MRAM usage and the specific actual endurance requirements. All cases assume 100 % uninterrupted operation/utilization. replotted from [59]*

follows: (1) the breakdown times are extrapolated from accelerated conditions to operation conditions, (2) the breakdown times undergo percentile scaling from 63 % failures to e.g. 1 ppm (Fig. 4.3).

For an accurate lifetime extrapolation we need to measure the breakdown time at different stress conditions. This voltage acceleration is described by a model, and in literature there are multiple models that describe the accelerated data well, but give significantly different results at extrapolated operation conditions. For Fig. 4.3, a power-law is used. Later, in section 4.6, we will discuss and analyze the most common models. First, in the following section, we will elaborate on an all-in-one fitting procedure to fit the resulting breakdown distributions.

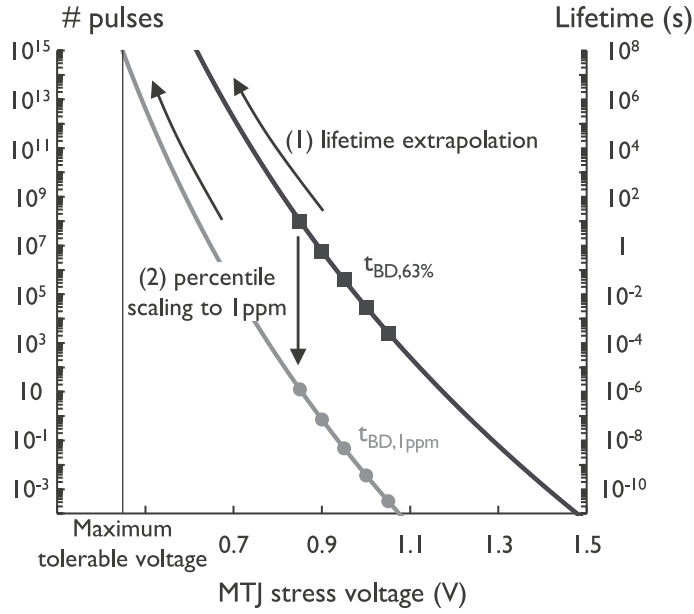

Figure 4.3: *The lifetime extrapolation consists of (1) extrapolating to the 10 year lifetime criterion using a voltage acceleration model, e.g. a power-law, (2) scaling from the measured time-to-63 %-breakdown, to the required percentile, e.g. 1 ppm. As a result we find the maximum tolerable operating voltage.*

4.3 New all-in-one maximum likelihood fitting of the breakdown distribution

Breakdown is a statistical phenomenon. It can be well described by a Weibull distribution and to determine the Weibull parameters for breakdown datasets obtained under various experimental conditions, we make use of a maximum likelihood (ML) fitting method. In general, each Weibull distribution is fitted, per experimental condition k, with a common Weibull slope β and stress voltage dependent scale factor η_k. Only afterwards, based on the fitted scale factors or the 63 %-values for the different stress conditions, the voltage acceleration is fitted. Using the voltage acceleration, a lifetime can be predicted at operating stress conditions.

We have developed a more robust fitting procedure by including the voltage acceleration in the principal maximum likelihood fit. As a result, the

Weibull parameters, as well as the voltage acceleration parameter are fitted simultaneously for the full dataset.

The derivation of the basic equations is discussed in section 4.3.1. In addition, we include a voltage acceleration model to the maximum likelihood fit. Furthermore, we will discuss the estimation of the error on the fitted parameters in Sec. 4.3.2, and in Sec. 4.3.3 we show how to implement the ML for a ramped voltage dataset.

4.3.1 Maximum likelihood estimation for a constant voltage stress

In case the breakdown times are acquired with sufficient time resolution, the maximum likelihood method is able to fit both censored data and accelerated lifetime data. In an ML fit the breakdown distribution parameters are estimated that best describe the dataset. This is done by maximizing the likelihood function:

$$\Lambda = \prod_{i}^{N} f(t_i)\,, \tag{4.12}$$

with $f(t)$ the probability density function (pdf), in this case a Weibull pdf. The likelihood function is the probability that the postulated distribution would give the observed failure times t_1, t_2, ..., t_N, and is a function of the breakdown distribution parameters. In the case of censored data this becomes [91]:

$$\Lambda = \prod_{i}^{N} \Big(\delta_i \cdot f(t_i) + (1-\delta_i)\cdot(1-F(t_i))\Big)\,, \tag{4.13}$$

with $F(t)$ the cumulative failure function and δ_i denotes whether the device failed at time t_i ($\delta_i = 1$), or had to be censored at time t_i ($\delta_i = 0$). Censoring occurs typically when the accelerated test is stopped before all devices have failed, or when a device-under-test (DUT) fails due to a different failure mechanism. Next, we substitute the f and F with a Weibull distribution, Eq. 4.4 and 4.5 respectively. Finally, we take the logarithm of the likelihood function:

$$\ln(\Lambda) = \ln(\beta)\cdot\sum_{i}^{N}\delta_i - \beta\cdot\sum_{i}^{N}\ln(\eta_i)\cdot\delta_i + (\beta-1)\cdot\sum_{i}^{N}\ln(t_i)\cdot\delta_i - \sum_{i}^{N}\frac{t_i^{\beta}}{\eta_i^{\beta}}\,. \tag{4.14}$$

η_i is the scale factor for stress condition i. When the experimental data points t_i have been acquired on different areas or different acceleration conditions, the

distribution function has to be transformed with the appropriate scaling laws. Under the assumption of a power-law voltage acceleration, η_i becomes:

$$\eta_i = \eta_r \left(\frac{V_i}{V_r} \right)^n , \tag{4.15}$$

where V_r is the reference stress voltage, η_r the reference scale parameter at a stress voltage V_r, and n is the power-law voltage acceleration exponent, with $n < 0$. Including the Poisson area scaling rule (Eq. 4.9), η_i becomes:

$$\ln(\eta_i) = \ln(\eta_r) - \frac{1}{\beta} \ln \left(\frac{A_i}{A_r} \right) + n \cdot \left(\frac{V_i}{V_r} \right), \tag{4.16}$$

where A_r is the reference area and the corresponding Weibull distribution:

$$W(t_i) = \ln(-\ln(1 - F_{CVS})) = \beta \cdot \left[\ln(t_i) - \ln(\eta_r) - n \ln \left(\frac{V_i}{V_r} \right) \right] + \ln \left(\frac{A_i}{A_r} \right), \tag{4.17}$$

By inserting Eq. 4.16 into Eq. 4.14, a full dataset can be fitted simultaneously for β, η_r and the power-law exponent n for a reference stress voltage V_r. The simultaneous fit can no longer be solved using a one-dimensional numerical root solver, as in [91], but instead using a multi-dimensional maximizer, finding the fitting parameters β, η_r and n that maximize the likelihood function Eq. 4.14. The simultaneous fit allows for a more robust fit, certainly if the dataset contains a lot of anomalies, as will be discussed in section 4.6.4.

The above derivation assumes sufficient time resolution of all the breakdown times, i.e. "continuous" monitoring. In the case of measurements at fixed inspection times, more than 1 failure can be found at the same time. The maximum likelihood function in Eq. 4.14 cannot be used. Instead, the maximum likelihood function becomes:

$$\Lambda = \prod_{i=1}^{K} \left(F(t_i) - F(t_{i-1}) \right)^{k_i} \cdot \left(1 - F(t_{i-1}) \right)^{l_i} , \tag{4.18}$$

where the difference in the cumulative distribution function between two inspection times is powered to the number of failures detected in that period k_i and K is the number of inspection times. l_i is the number of censored devices at inspection time t_i.

4.3.2 Calculating the confidence bounds on the estimated breakdown parameters

In addition, the ML method allows to calculate the confidence limits on the estimated parameters. To estimate the error of the fit, we use the variance of an ML estimator ($\hat{\theta}_{ML}$). For the breakdown distributions, $\hat{\theta}_{ML}$ can be $\hat{\beta}$, $\hat{\eta_r}$ or $\hat{n}$. The variance is calculated by the inverse of the Fisher information matrix $I(\theta)$, where this matrix is the negative of the expected value of the Hessian matrix $H(\theta)$ of the maximum likelihood function around the maximum (in $\hat{\beta}$, $\hat{\eta_r}$, $\hat{n}$)[44]:

$$\begin{aligned} var(\theta) &= \big[I(\theta)\big]^{-1} \\ &= \big(-E\big[H(\theta)\big]\big)^{-1} \\ &= \left(-E\left[\frac{\partial^2 \ln \Lambda(\theta)}{\partial\theta\partial\theta'}\right]\right)^{-1}. \end{aligned} \tag{4.19}$$

In short, the likelihood function around the maximum is approximated by a multivariate normal distribution. The steeper the likelihood function around the maximum, the higher the accuracy of the fit. This way a 95 % elliptical confidence contour can be determined. The 95 % confidence bound ($CB_{95\%}$) on the model parameters is then determined by the square root of the diagonal elements of the variance and a factor taking into account the 95 % confidence and the number of fitted parameters:

$$CB_{95\%} = \hat{\theta} \pm \sqrt{\chi^2(95\,\%, 2) \times var(\hat{\theta})}\,, \tag{4.20}$$

with χ^2 the chi-square distribution.

4.3.3 The maximum likelihood estimation for a ramped voltage stress

From a ramped voltage stress we extract a breakdown voltage V_i for a given applied ramp-rate RR_i. The breakdown voltages are again Weibull distributed (Eq. 4.7). Furthermore, given the ramp-rate RR_i, an equivalent time to breakdown at stress V_i can be derived [63]:

$$t_i = \frac{V_i}{RR_i \cdot (1-n)}\,, \tag{4.21}$$

where n is the power-law exponent. The equivalent Weibull distributions can be compared by substituting this equivalent time (Eq. 4.21) at breakdown voltage

V_i, in the breakdown time distribution (Eq. 4.17) without area scaling:

$$\begin{aligned} W(V_k) = \ln(-\ln(1-F_{RVS})) &= \beta \cdot \left[\ln\left(\frac{V_i}{RR_i \cdot (1-n)}\right) - \ln(\eta_r) - n \ln\left(\frac{V_i}{V_r}\right)\right] \\ &= \beta \cdot \ln\left(\frac{V_r^n}{RR_i \cdot (1-n) \cdot \eta_r \cdot V_i^{n-1}}\right) \\ &= (1-n) \cdot \beta \cdot \left[\ln V_i - \frac{1}{1-n} \cdot \ln\left(V_r^{1-n} \frac{RR_i \cdot (1-n)\eta_r}{V_r}\right)\right] \\ &= (1-n) \cdot \beta \left[\ln(V_i) - \ln V_r - \frac{1}{1-n} \cdot \ln\left(\frac{RR_i}{RR_r}\right)\right] \\ &= \beta_v \cdot \left[\ln(V_i) - \ln \eta_{v,ref} - n_v \cdot \ln\left(\frac{RR_i}{RR_r}\right)\right]. \end{aligned} \tag{4.22}$$

The original Weibull parameters β, η, n are converted to β_v, $\eta_{v,ref}$ and n_v. β_v is the Weibull slope of the V_{BD}-distribution, $\eta_{v,ref}$ is the reference breakdown voltage, n_v is the ramp-rate power-law exponent and RR_r is a chosen reference ramp-rate. These parameters are correlated with the constant voltage stress parameters:

$$\begin{aligned} \beta_v &= (1-n) \cdot \beta \\ \eta_{v,ref} &= V_r = \eta_r \cdot RR_r \cdot (1-n) \\ n_v &= \frac{1}{1-n} \end{aligned} \tag{4.23}$$

4.4 Breakdown mechanism

It is widely accepted that breakdown happens by the formation of a percolation path [29]. The percolation model assumes traps being generated at random positions inside the oxide. If the traps are close together, i.e. they "overlap", conduction between these traps increases exponentially, see Fig. 4.4(a). Breakdown is then defined when a conduction path bridges over the oxide. It is found that the total number of generated traps at breakdown is Weibull distributed. The relationship between the critical oxide trap density $D_{ot,tot,crit}$ and the charge to breakdown Q_{BD} is given by [29]:

$$D_{ot,tot,crit} = A \cdot (Q_{p,crit})^m = A \cdot (Q_{BD} \cdot \alpha)^m . \tag{4.24}$$

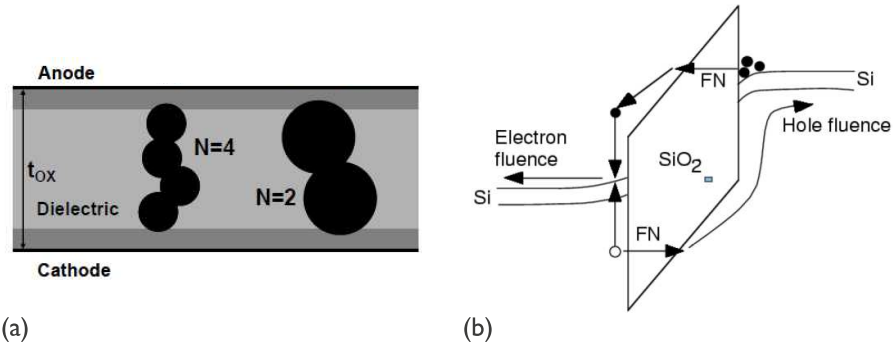

Figure 4.4: *(a) A percolation path is formed when traps form a conduction path from the anode to the cathode. Two trap sizes are depicted. (b) Schematic illustration of the anode hole injection model. Injected electrons reach the anode with high energy and can generate hot holes that can tunnel back to the cathode [28].*

In this empirical power-law, α is the ratio between the hole (Q_p) and the electron (Q_{BD}) fluence, A is a material and stress-dependent prefactor and m is the "trap generation rate". α is the probability for an injected electron to create a hole at the anode that can tunnel back to the cathode. This mechanism is described in Fig. 4.4(b) and well studied for thicker oxides, where the hole current could be experimentally determined [30]. The hole fluence results in the defect generation following a power-law trend with time. The empirical power-law is best suited for ultra thin gate oxides [29, 31, 106]. However, there are also reports that suggest the underlying physical defect generation mechanism is linear with respect to fluence, i.e. $m = 1$ [103, 130]. If $D_{ot,tot,crit}$ is Weibull distributed, then following 4.24 the Q_{BD}-distribution is also Weibull distributed with a Weibull slope of:

$$\beta_Q = \beta_{ot} \cdot m \tag{4.25}$$

In the analytical model elaborated in [32], generated traps have a probability to form a conductive path of N defects. Each conductive path of N defects has a probability to cover a distance larger than the oxide thickness and cause breakdown. These two probabilities result in the Weibull distribution for critical trap density, where β_{ot} is equal to the minimum number of traps N_{min} needed to form a BD path. $N_{min}(t_{ox})$ is proportional to the oxide thickness. The Weibull slope we can observe in measurements, following Eq. 4.25, depends thus on m and $N_{min}(t_{ox})$.

For ultra-thin oxides, like the studied 1 nm MgO, 1 or 2 defects could be enough to form a breakdown path. For these thin oxides the statistical spread of

the critical electron trap density is large even without any extrinsic defects. This inherent statistical property of the degradation could be mistakenly interpreted as extrinsic breakdown.

In Fig. 4.5, typical normalized breakdown time distributions are shown for SiO_2 gate oxide thickness from 1.7 nm to 7.8 nm [129]. As expected from the percolation theory, the Weibull slope decreases with oxide thickness. At 1 nm SiO_2, intrinsic Weibull slopes < 1 are expected and found [31], which is in direct contradiction to the claim of a linear trap generation in Eq. 4.24. In this case, if $m = 1$, β_Q has to be ≥ 1, since at least 1 defect is necessary to form the percolation path ($\beta_{ot} \geq 1$, Eq. 4.25).

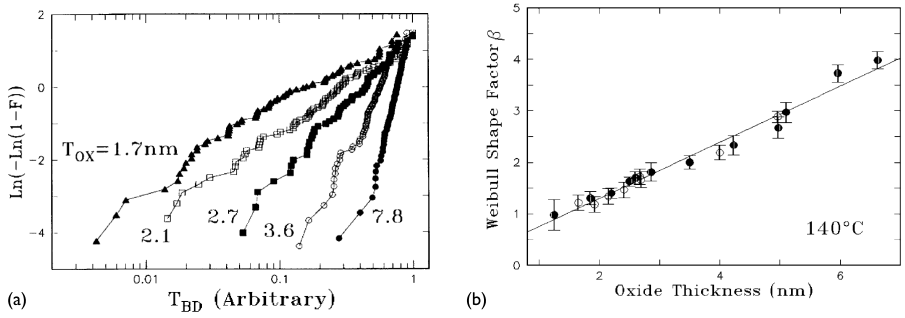

Figure 4.5: *(a) typical normalized breakdown time distributions in SiO_2 are shown for gate oxide thickness from 1.7 nm to 7.8 nm. The percolation model explains the decreasing Weibull slope with decreasing oxide thickness. (b) Weibull slopes as a function of oxide thickness. At 1 nm physical thickness Weibull slopes are < 1. Replotted from [129].*

Although there is a consensus on the breakdown mechanism, there is still a debate on how these traps are generated. It is very difficult to prove experimentally what traps or defects are generated during the stress and in a practical time range the models are hard to statistically differentiate. This will be further elaborated in section 4.6. In the next section, we will first focus on ways to measure experimentally breakdown in MgO (Sec. 4.5).

4.5 Breakdown measurements

Breakdown measurements are a required step in order to estimate the device lifetime. These measurements occur at accelerated conditions, i.e. higher voltage

and/or temperature, since at operating conditions the device will only fail after years. In order to make an accurate lifetime prediction at operating conditions, sufficient breakdown data has to be collected at different accelerated conditions. In this section, we will discuss and compare the most common measurements to analyze dielectric breakdown at accelerated conditions (see Fig. 4.6).

In section 4.5.1, we demonstrate why the constant voltage stress (CVS) is not suited to study the large variability in breakdown of the MTJs. A ramped voltage stress (RVS), see Fig. 4.6(b), is easy to automate and capable of handling the large variability in breakdown (Sec. 4.5.2). Furthermore in Sec. 4.5.3, we develop a dedicated experimental setup to apply voltage pulses for a pulsed breakdown (pulsed BD) measurement (see Fig. 4.6(c)). We find no influence of pulse width and duty cycle on the cumulative breakdown time and thus conclude that the degradation process is cumulative in nature and has no measurable relaxation mechanisms. In section 4.5.4, we compare the different techniques and summarize their advantages and disadvantages. Finally, we discuss how we perform breakdown measurements within the Mbit array.

4.5.1 Constant voltage stress

In a constant voltage stress (CVS), a constant voltage is applied over the MTJ and the time until breakdown is monitored (Fig. 4.6(a)). CVS is appropriate when the degradation is determined by voltage, in contrast to current. Since the MgO barrier is very thin, electrons are expected to directly tunnel through the

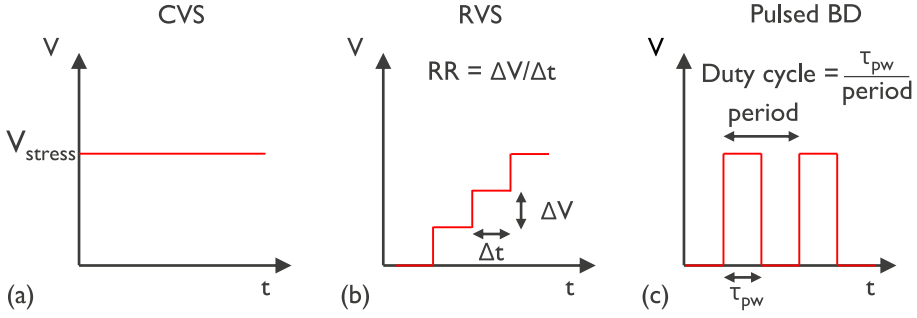

Figure 4.6: *(a) During CVS, a constant voltage is applied to the MTJ until breakdown. (b) For RVS, the voltage is ramped up with a constant voltage step ΔV with time steps Δt. The ramp-rate (RR) is $\Delta V/\Delta t$. (c) In pulsed breakdown stress, pulses are applied with 100 ns pulse width (τ_{pw}) and 300 ns period.*

oxide barrier. The electron energy at the cathode depends directly on the applied voltage, consequently the assumption that voltage determines the degradation is reasonable. In thicker oxides ($t_{ox} \geq 3\,\text{nm}$) the degradation depends more on the electric field than the voltage, because for these cases the electron current is dominated by field driven Fowler-Nordheim (FN) tunneling and the energy of the electrons is determined by the gain and relaxation happening after the electron enters the conduction band of the oxide [97].

We already have introduced that the Weibull slope depends on the oxide thickness in Sec. 4.4. Therefore, less defects are necessary to form a percolation path in thin oxides. This leads to single and double defect breakdown paths. Thin oxides are thus intrinsically correlated with a low Weibull slope and a very high statistical variability in breakdown times. In Fig. 4.7(a), Weibull plots for different CVS on a MgO tunnel barrier of 1 nm are depicted. The Weibull slope in this case is only 0.35 and therefore, for one stress voltage, the measured breakdown time of a small number of devices, can already vary over 5 orders of magnitude. The CVS is therefore limited by timeouts and early fails. Reasonable time range for accessible direct current (DC) parameter analyzers are shown in Fig. 4.7(b). The measurable time range is limited to 5-6 orders of magnitude. To make an accurate acceleration estimation, this is certainly a limiting factor.

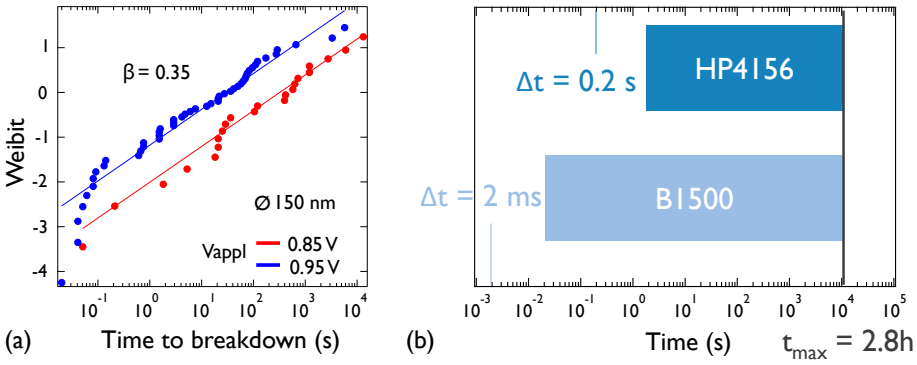

Figure 4.7: *(a) Example of breakdown time distributions for a CVS with 2 stress conditions ($V_{applied} = 0.85, 0.95V$). The 1 nm MgO has a low Weibull slope, resulting in 5 orders of magnitude in breakdown time for one stress condition. (b) Measurable time range (filled area) for a constant voltage measurement with the HP4156 and the B1500. Timeouts are defined from $t_{max} = 2.8$ hours.*

4.5.2 Ramped voltage stress

In a ramped voltage stress (RVS), the voltage is ramped with a constant voltage step ΔV per time interval Δt until the device breaks (Fig. 4.6(b)). The ramp-rate (RR) is $\Delta V/\Delta t$. RVS guarantees breakdown within a pre-defined time and without pre-defined stress conditions, which allows measuring large sample sizes with strong variation, i.e. low β.

Measuring at different ramp-rates provides additional information. In Kerber et al., it has been derived that the resulting breakdown voltage in an RVS stress can be converted into a CVS breakdown time [63]:

$$t_{CVS} = \frac{V_{CVS}}{RR \cdot (1-n)} \left(\frac{V_{BD}}{V_{CVS}} \right)^{1-n} \tag{4.26}$$

Here t_{CVS} is the equivalent breakdown time as if it would be measured by a CVS with a stress voltage V_{CVS}. The most important assumption here is that we make use of a power-law voltage acceleration with exponent n, with $n < 0$ and for room temperature in the range -60 to -40:

$$\eta_{CVS} = \eta_{RVS} \left(\frac{V_{RVS}}{V_{CVS}} \right)^{-n} \tag{4.27}$$

Eq. 4.26 is derived by considering a stepped RVS as a series of CVS at different V_{stress}-levels. Each CVS-level takes a time Δt and converting this stress time for each step to an equivalent stress time at a stress voltage V_{CVS}, using the power-law Eq. 4.27, results in:

$$\begin{aligned} t_{CVS} &= \sum_{i=1}^{N} \Delta t \left(\frac{V_i}{V_{CVS}} \right)^{-n} \\ &= \frac{V_{CVS}}{RR} \left(\frac{\Delta V}{V_{CVS}} \right)^{1-n} \sum_{i=1}^{N} i^{-n} \,. \end{aligned} \tag{4.28}$$

Equation 4.28 is only equal to Eq. 4.26 for large N and thus small steps ΔV. For small steps the sum approximates an integral, from which Eq. 4.26 is derived. For ramp-rates of more than 300 points, the difference between the integral approximation and the sum is less than 5 %.

An example of an RVS analysis is shown in Fig. 4.8. In Fig. 4.8(a) the breakdown voltage distributions are plotted. Since the voltage ramp starts from 0 V, even the weakest devices, which may result in an early failure in a

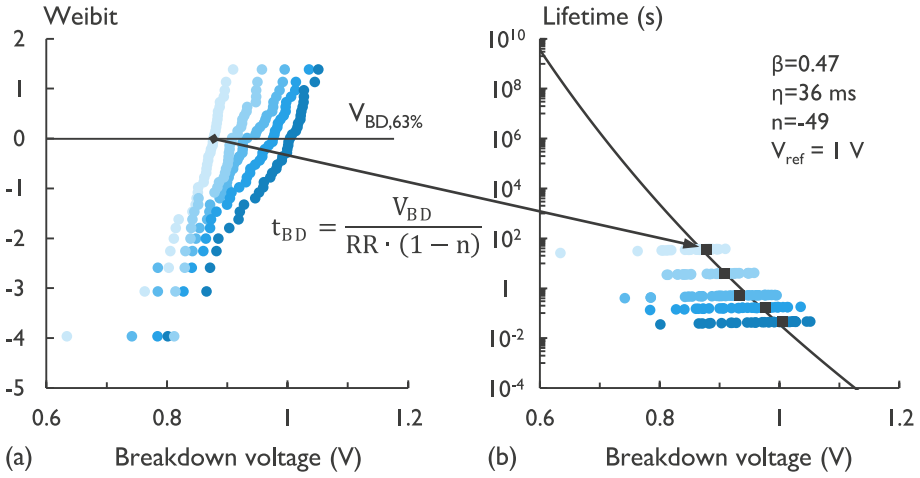

Figure 4.8: *(a) breakdown voltage distributions for 185 ⌀ 150 nm devices with 1.0 nm MgO for 5 different ramp-rates (500, 100, 40, 4, 0.4 mV/s). The breakdown voltages V_{BD} are converted to an equivalent breakdown time for constant stress bias $V_{CVS} = V_{BD}$, using Eq. 4.26. (b) Lifetime extrapolation derived from the all-in-one fit (see Sec. 4.3.3). The squares represent the equivalent 63 % breakdown time at the 63 % breakdown voltage.*

CVS measurement, can be characterized. Also the strongest devices can be measured, that otherwise in a CVS would not break down before time out. For this reason, RVS is capable of characterizing devices that exhibit large variability in breakdown time.

One of the disadvantages of the RVS measurements is that, as for the CVS, the voltage range over which breakdown occurs is small. To improve the accuracy on determining the voltage acceleration exponent n, multiple ramp-rates are mandatory. Slower ramp-rates result again in a prolonged measurement time, thus moderating the time advantage over CVS.

4.5.3 Pulsed breakdown stress

In order to increase the total measurable breakdown time range we make use of a pulsed breakdown stress. In addition, short voltage pulses mimic the real STT-MRAM operation. The device will break down after a certain number of pulses. The breakdown time is then calculate as the cumulative pulse time to

breakdown $t_{BD,pulsed} = N_{BD} \cdot \tau_{pw}$.

In this section, we will introduce the experimental setup, which is capable of detecting breakdown within a single voltage pulse, resulting in an increased measurable breakdown time range, with high time resolution in the nanosecond range. In addition, the setup is capable of measuring the actual stress voltage over the MTJ. Furthermore, we study the influence of the pulse width and duty cycle. We find no significant influence of pulse width and duty cycle, in contrast to some literature, and conclude therefore that $t_{BD,pulsed} = t_{BD,CVS}$.

Experimental setup

To establish pulsed BD measurements and exploit them to the fullest, we have built our own setup (see Fig. 4.9(a)). The setup consist of one pulse generator and one 10 GSamples/s oscilloscope, all 50 Ω terminated. The pulse generator applies voltage pulses with a pulse width τ_{pw}. The applied voltage pulse sees one 50 Ω oscilloscope channel (CH1) that monitors the applied stress voltage V_{in} over the in-parallel DUT. The other 50 Ω oscilloscope channel (CH2) is in series with the DUT. Therefore, part of V_{in} drops over the DUT and the remainder over CH2, i.e. V_{out}. As a result, the stress voltage over the DUT is $V_{DUT} = V_{in} - V_{out}$.

To correctly make use of this setup, three important notes are discussed. (i) we introduce a shorting loop as close as possible to our devices to mimic a ground-signal transmission line. (ii) we make use of oscilloscope CH2 to carefully monitor our voltage pulses and allow within-pulse breakdown detection. (iii) prior to these 2-point measurements, the series resistance of the devices is measured with a 4-point probe sense resistance check, in order to correct V_{DUT} for the series resistance.

(i) Shorting loop
For the first note we clamp a short wire to the grounds of the probes. In total the ground loop around the device is now estimated to be 8 cm. The smaller the ground loop the better the setup can approach a lumped element circuit, where the physical dimensions are small compared to the wavelength λ. In this case λ is 60 cm, derived by the rise time of the pulse generator, which is 2 ns. In addition we make use of parametric DC probe needles that have a frequency response (3 dB point) of only 150 MHz.

In Fig. 4.9(b) the effect of rise time through our setup is shown. We compare

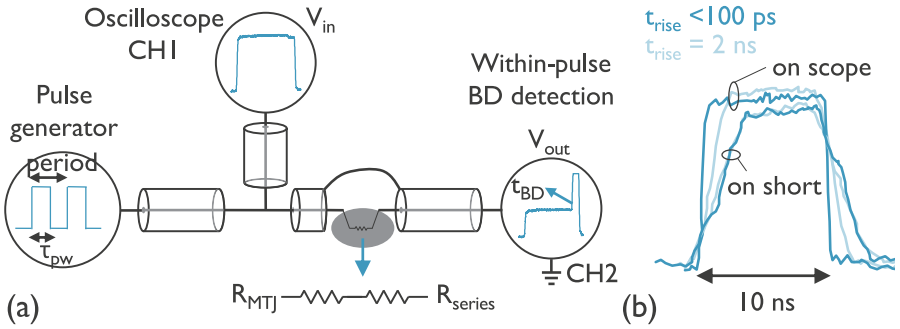

Figure 4.9: *(a) Experimental setup for pulsed BD. An oscilloscope monitors the incoming and outgoing pulse, to determine the stress voltage and enabling within pulse breakdown detection. (b) Characterization of the limited bandwidth of our setup. The rise time is reduced to 3-4 ns when pulses pass through the setup and a short.*

10 ns pulses that go directly from the pulse generator to the oscilloscope, with 10 ns pulses that first pass a shorted device on the wafer. Both the voltage and rise time are reduced by our setup. We define the rise time as the transition time from 10 % to 90 %.

We tested two different pulse generators, an Agilent 8133A and 81110A, with a rise time of <100 ps and 2 ns, respectively. When the pulse generator is directly connected to the oscilloscope, the difference in rise time is obvious. However, when the signal first passes through the device, the rise time is reduced to ≈ 4 ns in both cases, i.e. the frequency limit of the setup.

(ii) Within-pulse breakdown detection
The incoming and outgoing pulses are evaluated for stress voltage and breakdown time, respectively. During the first 10 pulses, meausring with 100 ps steps, the oscilloscope detects breakdown within the pulse (Fig. 4.9(a)). As a result the measurable range of t_{BD} is increased with high time resolution down to the nanosecond range.

(iii) V_{DUT} correction for series resistance
To correctly determine the stress voltage over the MTJ, the voltage has to be corrected for the series resistance R_{ser}, that needs to be measured prior to the pulsed BD. The used 4-point structures are measured in DC to calculate R_{ser}. The MTJ voltage V_{MTJ} is determined by a voltage division between the MTJ resistance R_{MTJ} and R_{ser}.

Influence of pulse width and duty cycle

Large impact of the pulse width and duty cycle have been reported in literature [49, 2]. In [49] a change of more than 3 orders of magnitude in breakdown time is reported for pulse widths between 10 ns and 1 μs. Moreover, in [2] the duty cycle is reported to have an even more significant impact. To evaluate the influence of pulse width and duty cycle, we performed breakdown measurements for 4 different pulse widths (in range 10 ns-1μs) and 3 different duty cycles (from 1 % to 77 %).

In Fig. 4.10(c), the cumulative breakdown distributions are shown. All devices are stressed with an applied voltage of 1.5 V and a duty cycle of 33 %. From the raw breakdown distributions one might conclude that for short pulse widths, a different breakdown behavior is observed. However, if we study the applied voltage pulses in more detail (see Fig. 4.10(a)-(b)), we observe an increase in stress voltage with pulse width. The input voltage, V_{IN}, which is monitored at the oscilloscope (Fig. 4.9(a)), does not reach a steady state level after 10 ns. Therefore, the average stress voltage is not the same for the different pulse widths.

We can correct for the voltage differences in two steps. First, $V_{DUT} = V_{IN} - V_{OUT}$ has to be corrected with the series resistance to obtain V_{MTJ} (see Fig. 4.10(b)). Second, we rescale the breakdown times for each R_{series}-corrected stress V_{MTJ}, to a stress V_{ref}, by applying a power-law (Eq. 4.27) with a typical exponent n of -50, see Fig. 4.8(a). This way all distributions are rescaled to an effective voltage stress of $V_{ref} = 1\,V$ (Fig. 4.10(d)). All rescaled distributions except the 10 ns one fall within the 95 % confidence bounds. However, as stated in subsection 4.5.3, for 10 ns we reach the bandwidth limit of the setup. Hence, the cumulative time can be significantly impacted by the relative longer rise/fall times.

Furthermore, we observe a drop-off at small cumulative breakdown times (Fig. 4.10(c)). However, this drop-off is reduced after correcting each breakdown time for both the applied voltage and the R_{series}-correction. The drop-off occurs because of the lower average stress voltages at short breakdown times. To emphasize this effect, in Fig. 4.10(b,c,d) the devices with a $t_{BD} < 10$ ns are plotted as open symbols. These devices broke down seeing lower average voltage and correcting with voltage acceleration to $V_{ref} = 1$ V, reduces the effective

t_{BD}, resulting in a more straight cumulative breakdown distribution. Hence, the drop-off is a measurement artifact we can correct for.

We performed similar measurements to study the influence of the duty cycle. The pulse width remains fixed at 100 ns, and the duty cycle is set to 1 %, 33 % or 77 %. In contrast to different pulse widths, the voltage is similar for different duty cycles (see Fig. 4.11(a,b)). In Fig. 4.11(c) the breakdown distributions rescaled to $V_{ref} = 1$ V are plotted. We find no significant influence of duty cycle on cumulative breakdown time. The distributions are tested for equivalence by a Kolmogorov-Smirnov test and all have p-values > 0.68.

This result is quite different compared to other reports in literature [2]. In Amara et al., up to 6 orders of magnitude difference in breakdown time is found when changing the delay between the pulses. The authors attribute this effect to charge trapping and detrapping in the MgO barrier. Our study of the influence of duty cycle and pulse width has been repeated at Qualcomm on high quality devices [58]. The authors make use of a Gbit array and the devices show good uniformity, having a Weibull slope even larger than 1. Each breakdown

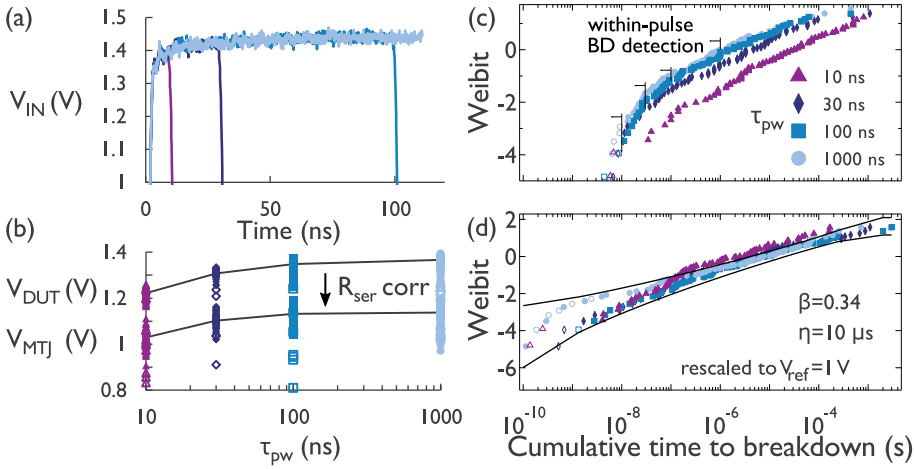

Figure 4.10: *Study of influence of pulse width on breakdown time of ⌀100 nm MTJs with 1 nm MgO. (a) Example curves of input voltage as a function of time for 4 different pulse widths (10 ns-1µs). In a 10 ns pulse the voltage is not yet saturated and hence the average voltage on the DUT is lower (b). (c) The cumulative breakdown distributions. (d) Rescaled distribution to $V_{ref} = 1$ V, taking into account different stress voltages.*

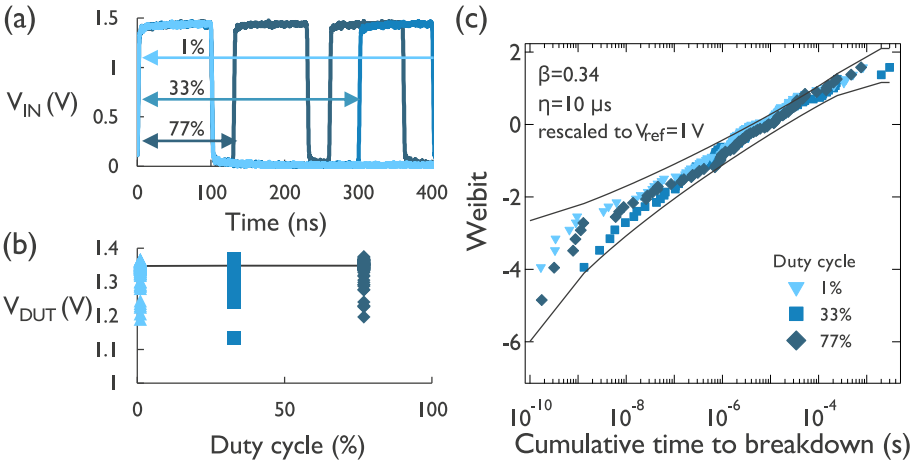

Figure 4.11: *Study of influence of duty cycle on breakdown time of ⌀100 nm MTJs with 1 nm MgO. (a) Example curves of input voltage as a function of time for 3 different pulse widths (1 %-77 %). The stress voltage does not depend on the duty cycle (b). (c) Breakdown distributions as a function of the cumulative breakdown time rescaled to 1 V. No significant difference is observed, the data falls between the 95 % confidence bounds.*

condition contains at least 75-480 samples randomly selected. The authors find no significant influence of duty cycle in the range 1 %-90 %, in agreement with our own study. In addition, also no impact of pulse width is found down to 200 ns. The discrepancy at smaller pulse widths (< 200 ns) might be related to a reduction of stress voltage on the device for smaller pulse widths as in Fig. 4.10(b).

Therefore we hypothesize that the charging effects seen in the samples from [2], could be related to the use of a different stack and process. For example the interface between CoFeB and MgO is based on Ru 0.8 nm/CoFeB 2 nm/CoFe 0.5 MgO. The CoFe cannot be deposited amorphous and could inflict an increased lattice mismatch at the MgO interface. In addition, the CoFeB capping layer is Ru, which could ineffectively scavenge the boron out of the CoFeB during a high temperature anneal. Therefore, an increased boron diffusion towards the MgO takes place and introduces defects at the MgO interface.

In summary, we conclude that the breakdown time can be expressed as the cumulative pulse time, independent of duty cycle or pulse width down to 30 ns. This demonstrates that the oxide degradation process is cumulative in nature

and has no measurable relaxation mechanisms.

4.5.4 Comparison between breakdown measurements

In this section, we compare the techniques of CVS, RVS and pulsed BD. In previous section, it has been shown that oxide degradation is cumulative in nature and the breakdown time can be expressed as the cumulative pulse time. As such, the CVS and pulsed BD are shown to be equivalent. We will demonstrate that RVS also results in the same breakdown parameters within the error bar of the measurements.

In a first test, we compare CVS with RVS, and in a second test, RVS with pulsed BD. We make sure the devices are equally distributed across the wafer for the different techniques, such that the measured devices have the same wafer location variability for the different techniques.

In Fig. 4.12(a), the breakdown distribution of a CVS measurement is compared to the equivalent breakdown time for an RVS performed at 4 ramp-rates (1 - 1000 mV/s). In this case, there was not enough CVS data to fit accurately the voltage acceleration. Therefore, we use the power-law exponent extracted from the RVS to plot the rescaled distribution at median voltage of the CVS. Both distributions are equivalent, proving that also for MgO both techniques are equivalent in this measured voltage range.

In a pulsed BD measurement, the voltage stress range is extended to higher voltages [4]. However, the extracted breakdown parameters are still comparable with those from RVS, see Fig. 4.12(b).

To summarize, a 1 nm MgO exhibits a large intrinsic percolation statistic, as well as process variability. As a result, breakdown can occur over a time range of more than 5 orders of magnitude for 1 CVS stress condition. Therefore, a DC CVS measurement is very time consuming. CVS is thus not easy to automate and it is more difficult to accurately fit the voltage acceleration. On the contrary, RVS is easy to automate, since it guarantees breakdown within the measurement time. By introducing multiple ramp-rates, it is also possible to extract a lifetime. The larger the ramp-rate range, the more accurate the voltage acceleration can be fitted, however, the more time consuming the measurement becomes. Pulsed BD can substantially extend the measurable time range up to 10 orders of magnitude. Although this measurement technique is not easy

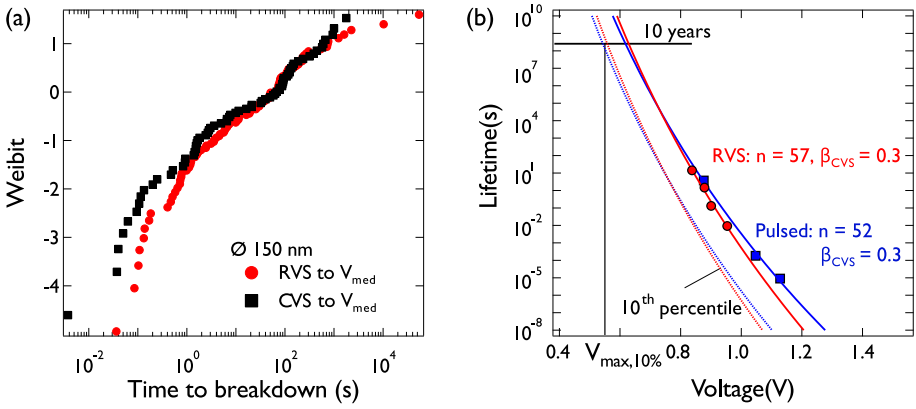

Figure 4.12: *Comparison between 3 breakdown measurements, CVS, RVS and pulsed BD. (a) The rescaled breakdown distribution at the median CVS stress The RVS (red) data is based on 4 ramp-rates (1 - 1000 mV/s). (b) Lifetime extrapolation of RVS (red) and pulsed BD (blue) data. The same wafer variability is in each full measured technique.*

to automate, it is suited for large variability in breakdown time and suited for accurately studying the voltage acceleration, see Sec. 4.6.

4.5.5 Breakdown measurements in a Mbit array

In previous measurements, we have made use of single device structures (0T1MTJ), where an MTJ is measured in a 4-point cross structure. Unfortunately this structure consumes a lot of die area and is only repeated 18 times on each die for every MTJ dimension. In order to fit the Weibull breakdown time or voltage distribution, sufficient statistics are required. We therefore need to measure across different dies, introducing process-related variability across the wafer, like oxide thickness variation, roughness variation, CD variation, etc.

These process-related variations can be reduced using the Mbit array. One Mbit array has 1024 x 1024 cells available to measure with a 4-point measurement. All cells are contained in an area of only $6 \times 7\,\mathrm{mm}^2$. Each cell is connected to 2 transmission gate transistors, which are the drive transistors. One transmission gate (drive transmission gate A) has large transistors capable of delivering currents of more than 1 mA when open. The other transmission gate (drive transmission gate B) is smaller and can be used for a four point measurement,

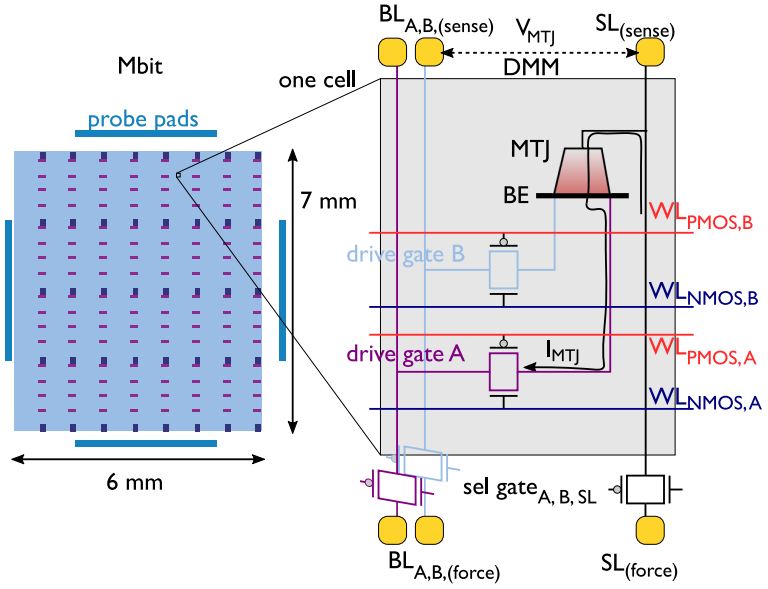

Figure 4.13: *Schematic of a MTJ cell in the Mbit array, with 2 transmission gate transistors and selector transistors. A 4 point measurement is possible via digital multimeter (DMM).*

where the voltage is measured by a digital multimeter (DMM), see Fig. 4.13. This is done as follows: the stressing current flows from the force of the source line (SL), to the MTJ and through the drive transmission gate A, to the bit line A (BL_A). The voltage is measured by the DMM between the sense SL, i.e. the TE of the MTJ, and the BE of the MTJ is sensed via drive transmission gate B at the bit line B (BL_B).

Furthermore, each cell has 3 selector transmission gate transistors, 2 for the bit lines (BL_A and BL_B) and 1 for the source line (SL). These transmission gates are opened, depending on which cell of the 1024 x 1024 array is connected.

In the Mbit array, stressing the device happens in two cycles, a write and a read cycle. Both cycles are controlled by the clock frequency, which is chosen between 1 kHz and 20 MHz. The equivalent CVS and RVS are shown in Fig. 4.14(a,b). During a cycle we keep the driving and selector gate transistors fully open, to ensure the maximum current can be delivered, this is around 2.5 mA.

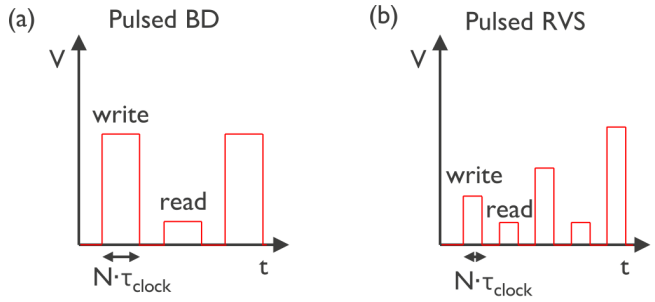

Figure 4.14: *(a) and (b) are equivalent pulsed BD and RVS, respectively, in the Mbit array. Each cycle contains a stress (write) and read cycle, the pulse width is controlled by the clock frequency and a chosen amount of repeats N ($N \cdot \tau_{clock}$).*

All the transistors contribute to the series resistance of the measurement and are considered as the periphery. Besides the resistance of the periphery, which can be measured and adjusted for, there also is a parasitic series resistance. This parasitic resistance originates from the process flow and is situated most probably between the M3 and the BE layer, see appendix A. Because of the unknown parasitic resistance, we will only report on Mbit array measurements for thick MgO. In this case the series resistance is negligible compared to the MTJ resistance. For breakdown results using the Mbit array, see chapter 5, section 5.5.2.

4.6 Breakdown acceleration and lifetime extrapolation

The desired knowledge to extract from a breakdown analysis is the lifetime of MTJs at operating voltage. Will the devices pass the breakdown criterion? The estimated lifetime at operating voltage ought to be $> 10^{15}$ pulses, for 100 ns pulses this translates into approximately 10 years of constant stress time. It is impossible to measure the breakdown distribution at this stress condition. Therefore, the tests occur at accelerated conditions and are extrapolated to the operation conditions.

In previous sections it was shown how the thin MgO tunnel barrier can fail by the formation of a percolation path of defects. We know that these defects are generated due to the applied stress, however, how exactly this generation

mechanism can be accelerated by an electric or thermal stress is not known.

Since there is no extended knowledge acquired on breakdown acceleration for the MgO oxide, we will discuss acceleration models established for gate oxides and test whether they suffice for the thin MgO in MTJs. The acceleration models are divided in field-driven models and voltage-driven models. Among the field-driven models we will consider the bond breaking E-model and hole-induced 1/E-model for thick oxides. Among the voltage models we will consider hole-induced and hydrogen-induced models for thin oxides.

For thick gate oxides, acceleration is described well by the E-model, which has found some theoretical support in the framework of the so-called thermo-chemical model, see sub section 4.6.1. Although there is strong evidence disproving this model, it is still being used for MgO lifetime extrapolations [134, 2, 123, 89].
In the 1/E-model, also called Anode hole injection (AHI), not the electric field, but the hole fluence (see Fig. 4.4(b)) causes trap generation, see section 4.6.2. In this case, the electron energy at the anode determines the oxide degradation. Consequently, the acceleration factor changes from thick to thin oxides, since the electron transport changes from non-ballistic to ballistic tunneling. Therefore, for thin oxides the voltage determines the electron energy and thus the trap generation.
In the hydrogen-induced model, hydrogen is released due to the electrons. The released hydrogen diffuses through the oxide and can generate electron traps, resulting in a voltage power-law dependence (sec. 4.6.3).

In MgO, other mechanisms might play. Because of the self-heating, diffusion mechanisms could be activated, either towards the MgO or oxygen diffusion from the MgO. Currently there is no consensus on what acceleration model to use of MgO breakdown, herefore we use our pulsed BD measurement, with a measurable range of breakdown times of >10 orders of magnitude. This extended breakdown time range is used to discriminate between the different acceleration models, to see which one most accurately describes the full dataset (Sec. 4.6.4).

4.6.1 E-model or thermo-chemical model

The E-model has been historically used for gate oxide reliability extrapolation as the worst case approach, before any theoretical support existed. It is the most pessimistic acceleration model. It is still used today in MgO-based MTJ [134, 2, 123, 89], although there is compelling experimental evidence against

this model.
In the thermochemical theory, the generation of defects is attributed to the interaction between the electric field and the inter-atomic bonds [74, 96, 75]. The bond breakage rate depends on the interaction frequency and on the probability that an interaction will break the bond. The external field E_{ox} lowers the necessary energy required to break the bonds and thus the breakdown time follows:

$$t_{BD} = A_0 \exp\left[\frac{\Delta H_0}{k_b T} - \gamma E_{ox}\right], \tag{4.29}$$

where γ is referred to as the field acceleration parameter. ΔH_0 is the zero-field activation energy and although this is directly linked to the nature of the chemical bonding in an oxide, this parameter is mostly fitted by absorption into the pre-exponential factor A_0 or the η_r in the Weibull distribution.

There are several strong arguments against the E-model: it does not explain any polarity dependence [37, 127, 33], it predicts a limited lifetime at 0 stress and it does not support the dependence on the electron fluence and energy [29, 57, 129]. In addition, measurements with an extended dataset down to low stress show a field dependent acceleration parameter, seriously questioning the validity using an exponential E-law. In [129], a decreasing voltage acceleration is found with increasing thickness. On extrapolation to low voltages, the lifetime of thin oxides could exceed that of thick oxides, which is very counterintuitive and seems to be unphysical. Furthermore, the field acceleration shows non-parallel slopes for different areas, hereby the E-model violates Poisson area scaling [129]. To summarize, large datasets over a large voltage and time range show that the exponential E-law is not suited for oxide breakdown: this is the case for both gate oxides [129] and BEOL oxides [26]. In Sec. 4.6.4, we show that this is also the case for thin MgO. As a result, the E-model should not be used for accurately predicting the lifetime of MgO-based MTJ.

4.6.2 1/E or anode hole injection model

The Anode hole injection model (AHI) is a current-driven model. It is based on the idea that electrons at the anode can generate holes that travel back through the oxide and cause damage by creating neutral electron traps [19, 97]. According to [97], the field dependence of α, which is the probability to create a hole that can tunnel back into the oxide (see Eq. 4.24), can be described as:

$$\alpha = \alpha_0 \cdot \exp\left(\frac{-H}{E_{ox}}\right) \quad and \quad Q_{BD} = Q_0 \cdot \exp\left(\frac{H}{E_{ox}}\right), \tag{4.30}$$

with α_0 and H constants (for a fixed oxide thickness). Considering conduction in thick oxides to be FN-tunneling, with J_{FN} the current density, the time to breakdown becomes:

$$t_{BD} = \frac{Q_{BD}}{J_{FN}} = A_0 \cdot \exp\left(\frac{B+H}{E_{ox}}\right). \tag{4.31}$$

The main criticisms on the 1/E model are threefold: (1) since FN tunneling is very weakly temperature dependent, the AHI model in its original form, fails to explain the strong temperature dependence of dielectric breakdown [57, 129]. (2) The defects generated by hot holes have little or no influence on breakdown during CVS [38, 122, 48]. In addition, (3) the defect generation is assumed to be that of the anode hole injection probability, following the FN 1/E dependence (α in Eq. 4.30). In thick oxides, in the measured breakdown range, α is not strongly voltage dependent and therefore the breakdown time follows the 1/E dependence from FN tunneling. However, for thin oxides this defect generation seems to follow a strong power-law $\alpha \sim V^{38}$ and $J_{DT} \sim \exp(aV) \approx V^7$ [131]. Therefore the 1/E dependence for t_{BD} is no longer valid. The AHI model was mainly used for thick oxides, but appears unsuited for thin oxides. Therefore, it is not further considered to be valid for breakdown of 1 nm MgO.

4.6.3 Power-law or Anode hydrogen release model

As for the AHI model, the power-law model is fluence- and energy-driven. On decreasing oxide thicknesses the conduction mechanism changes from FN-tunneling to direct tunneling (DT). For oxide thicknesses below 4 nm, the energy of the electron at the anode is determined by the voltage instead of the oxide field. The electron energy can release holes or hydrogen [107, 110]. The released species travels back through the oxide and reacts with the oxide to generate defects that will finally form a percolation path and trigger breakdown. A power-law model describes the failure time as [128]:

$$t_{BD} = \eta_r \left(\frac{V}{V_r}\right)^n, \quad with\ n < 0\,. \tag{4.32}$$

Like the AHI model, the power-law or anode hydrogen release (AHR) model, is also a two step model. In a first step some positively charged species (holes or protons) are released at the anode. In a second step this species has a probability to react with the oxide and create defects. A schematic picture of the defect generation process is shown in Fig. 4.15(a). Each step has its own probability of occurrence, represented by ζ_1, ζ_2 and ζ_3. The defect generation efficiency can then be written as the product of the efficiency of the injected

electrons to release species from the anode (ζ_1) and the probability (k) that these released species react with the precursors to generate a defect [130]

$$\zeta(V, T) = \zeta_1 \cdot k(\zeta_2, \zeta_3) \tag{4.33}$$

In the generalized hydrogen release-reaction model of Wu and Suñé, temperature effects are taken into account. At elevated temperature the Si-H bond is no longer in the ground state, see Fig. 4.15(b). Release of hydrogen from these excited vibrational states, requires less energy and results in a lower power-law exponent [94]. Therefore, ζ_1, needs to include a Boltzmann probability of occupation of the Si-H bond vibrational eigenstates n according to their energy level E_n

$$\zeta_1 = \sum_{n=0}^{n=N} (N-n)! \left(\frac{I_1}{I_0}\right)^{(N-n)} \exp\left(-\frac{E_n}{k_b T}\right), \tag{4.34}$$

where N is the number of energy levels, $N \sim 10-12$ for Si-H [130]. $\frac{I_1}{I_0}$ is the inelastic tunneling fraction. I_0 is the total elastic current flowing through the Si-H bond and I_1 is the inelastic current associated to the excitation of the bond to the first excited state. It is empirically shown that the inelastic tunneling

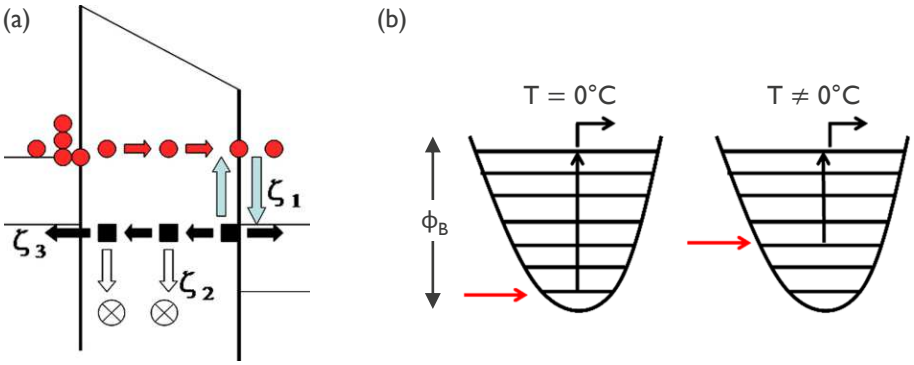

Figure 4.15: *(a) Schematic picture of the defect generation process driven by electron fluence and energy. Electrons travel through the oxide and partially loose their energy at the anode interface to release some positively charged species (holes or hydrogen) with an efficiency ζ_1. In addition, ζ_2 represents the probability of the released species to create defects and ζ_3 the probability of loosing the generated species towards the electrodes. (b) Schematics of the thermally assisted hydrogen release model. Si-H (or in our case Mg-H) breakage occurs through multiple excitations by electrons. For higher temperatures, the number of energy levels to overcome is reduced, leading to a smaller power-law exponent. Replotted from [130].*

fraction depends on the applied voltage as ([94]):

$$\frac{I_1}{I_0} = A\left(\frac{V}{V_0}\right)^4, \tag{4.35}$$

where A is the inelastic tunneling fraction prefactor. As such, the power-law exponent is equal to $4(N-n)$ and as the probability to occupy excited states with larger n, increases with temperature, at the same time the power-law exponent will decrease.
The reaction probability is assumed to be a simple Arrhenius expression

$$k = k_0 \exp\left(-\frac{\Delta H}{k_b T}\right) \tag{4.36}$$

where ΔH is the activation energy responsible for this reaction process and k_0 a material dependent parameter. The generalized hydrogen release model describes the experimental data well for a large range of voltages and temperatures in the case of 2.15 - 2.67 nm gate oxides [130]. Also in the case of MgO, the hydrogen could be introduced by the high-temperature anneals. However, it is not straightforward to apply this model directly to MgO. For example, what is the energy barrier for hydrogen release of Mg-H? How many energy levels should be taken? In addition, the authors assume a linear evolution of the density of defects as a function of time [130]. From experimental data, however, the defect generation is observed to follow a power-law in time (Eq. 4.24) [29]. Furthermore, instead of hydrogen, other atoms or analogous mechanisms can cause breakdown. In Sec. 5.4, a study of spacer layer material shows that diffusion of oxygen out of the MgO is a major contributor to the susceptibility to breakdown.

We will use the voltage power-law model, since the breakdown results of MgO correlate well with a power-law model (see Sec. 4.6.4). However, the described generalized model includes the temperature dependence based on the knowledge of the Si-H bond vibrational energy levels E_n. Since, this knowledge is not yet established for Mg-H or an analogous mechanism could be the major contributor, we will include the temperature dependence in an empirical manner in Section 5.5.4. Here it is shown that breakdown in MgO behaves similar as the breakdown in SiO_2.

4.6.4 Acceleration for MgO

In MTJ breakdown studies, the power-law and E-model are the most frequently used models [134, 2, 123, 89]. There is no further investigation of the voltage

acceleration in MgO and a model seems arbitrarily chosen. Therefore, in this section we use a large dataset to be able to discriminate between the different acceleration models. Pulsed breakdown with extended breakdown time range, discussed in section 4.5, is suited for this analysis.

We have performed breakdown measurements on 855 devices in total at 4 different stress voltages [4]. The devices are chosen equally spread over the wafer for all stress voltages, such that the die-to-die variability can be assumed identical for each stress voltage. Pulses with a width of 100 ns are applied until breakdown. At logarithmically spread detection times, we evaluate for breakdown. The breakdown time distributions as a function of the nominal applied stress voltage are plotted in Fig. 4.16(a). Overall, we measure a cumulative breakdown time range of more than 11 orders of magnitude (2 ns - 1000 s). Firstly, the wider time range allows to study the large variability that comes with a 1 nm thin MgO dielectric. Secondly, it allows us to compare the different acceleration models with high discriminative power.

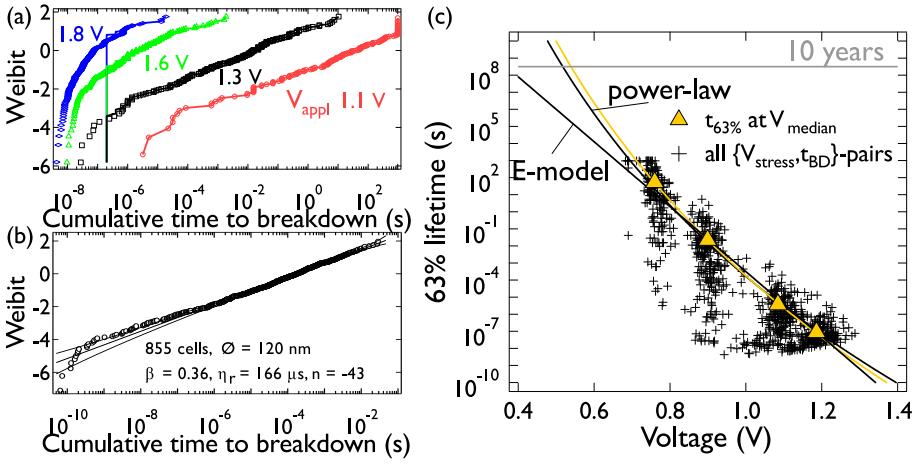

Figure 4.16: *(a) t_{BD} distributions from pulsed BD: BD evaluation at logarithmically spread inspection times (solid lines) and within pulse BD detection (open symbols). (b) residual t_{BD} distribution: the data are rescaled to 1 V using the power-law. (c) The black lines are the fits of the E-model and the power-law. The triangles are the $t_{63\,\%}$ percentiles at the median stress voltage of the 4 stress conditions. The yellow line represents the conventional lifetime extrapolation through only the $t_{63\,\%}$ values, not taking into account each $\{V_{stress}, t_{BD}\}$-pair. The power-law fits best the data.*

We simultaneously fit with a maximum likelihood method, all $\{V_{stress}, t_{BD}\}$-pairs for a voltage acceleration model and a Weibull distribution. We obtain the actual V_{stress} for each device in 2 steps. First we determine the $V_{DUT} = V_{in} - V_{out}$. Then, we correct for R_{series} and losses in the cables. Each device has a different V_{stress}. When a device breaks within the first pulse, the V_{stress} is reduced (see Sec. 4.5.3), which explains the apparent deviation from a Weibull distribution in Fig. 4.16(a). Longer breakdown times occur than expected at the fixed applied V_{stress}, due to the reduced actual V_{stress}.

As discussed in section 4.3.1 and Eq. 4.16, each V_{stress}-value is related to an η_i-value. The fitted Weibull distribution is then:

$$\ln\left[-\ln\left(1-F(t_{BD,i})\right)\right] = \beta \cdot \ln t_{BD,i} - \beta \cdot \ln \eta_i \tag{4.37}$$

the η_i in this case is different for each acceleration model:

$$Power-law: \quad \eta_i = \eta_r \left(\frac{V_i}{V_r}\right)^n \tag{4.38}$$

$$E-model: \quad \eta_i = \eta_r \left(\frac{exp(V_i)}{\exp(V_r)}\right)^{\frac{\gamma}{t_{ox}}} = \eta_r \exp\left(\frac{\gamma}{t_{ox}}(V_i - V_r)\right). \tag{4.39}$$

For each acceleration model β, η_r and n are fitted with a maximum likelihood method for non-continuous monitoring data at fixed inspection times (see Section 4.3.1 and Eq. 4.18).

Unlike in [134, 2, 123, 89], we investigate the goodness-of-fit for each acceleration model. In order to find which acceleration model best describes our extended dataset, we perform a likelihood ratio test [26]. We compare the likelihood of the best scoring model (power-law) with another model. The ratio can be written as:

$$Likelihood\,ratio = \frac{\Lambda_j}{\Lambda_{power-law}} = \exp\left(\ln \Lambda_j - \ln \Lambda_{power-law}\right), \tag{4.40}$$

where Λ_j is the likelihood function of the other acceleration model.

In order to attribute a significance to the likelihood ratio, we approximate the likelihood of the power-law by a multivariate normal distribution in terms of the 3 fitting parameters (β, η_r, n). Next, we determine the significance by comparing with the quantile of the χ^2 distribution. In case the likelihood ratio is large, we reject the j^{th}-model. This concept is illustrated for 2 parameters in

Fig. 4.17(a). Currently, there are 3 model parameters, consequently the critical significance becomes:

$$Critical\ significance = exp(-\frac{\chi^2(95\%, 3)}{2}) = 0.02\,. \tag{4.41}$$

The maximum likelihood values for E-model are far below the one of the power-law, which results in a significant likelihood ratio $< 10^{-20}$ (Fig. 4.17(b)). Thus the likelihood ratio test shows that the power-law best describes voltage acceleration.

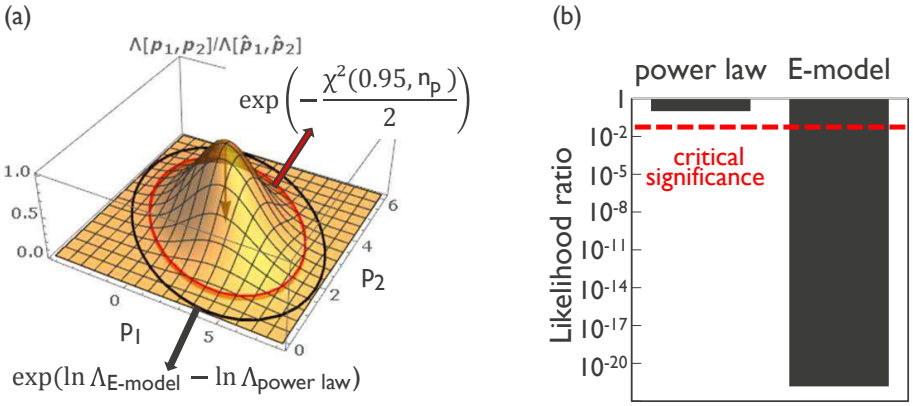

Figure 4.17: *(a) Approximating the likelihood function with a multivariate normal distribution in terms of the fit parameters defines a confidence contour (red line),* $\exp(-\chi^2/2)$ *times below the maximum likelihood level. In case the likelihood ratio test (black line) lies below the red line, we reject this model. (b) Likelihood ratio of the power-law and E-model for the data of Fig. 4.16.*

4.7 Conclusions

In this chapter, we have introduced the Weibull statistic, which is present in our experimental breakdown data. Next, we have discussed the breakdown mechanism. By the formation of a percolation path between anode and cathode, the generated defects create a conductive filament across the MgO barrier and hence irreversible breakdown occurs. For the studied ultra-thin MgO dielectric, a single or two defect path can be sufficient to break the dielectric, resulting in low Weibull slopes (< 1).

We have developed an all-in-one maximum likelihood fitting method that simultaneously fits all measured breakdown data at different stress conditions.

Fitting the Weibull parameters and the voltage acceleration parameter in one go, is more robust than first fitting one common Weibull slope and for each stress condition a 63 % breakdown time, continued by a second fit of the voltage acceleration using only these 63 % breakdown times, instead of the full dataset of breakdown times.

We have demonstrated the equivalence between the CVS, RVS and pulsed breakdown measurements. In addition, for pulsed breakdown the average stress voltage is carefully measured on an oscilloscope and each device is corrected for series resistance. Applying these corrections lets us explain the apparent deviation from a Weibull distribution. We conclude that the breakdown time can be expressed as the cumulative pulse time, independent of duty cycle or pulse width down to 30 ns. This demonstrates that the oxide degradation process is cumulative in nature and has no measurable relaxation mechanisms.

In the last section, a review of the most important acceleration models for thin MgO is given. For thin MgO the power-law acceleration, based on the anode hydrogen release model, seems the most plausible to explain the voltage and temperature acceleration of breakdown. Instead of hydrogen, other atoms or analogous mechanisms could cause breakdown, again resulting in a power-law model. A maximum likelihood ratio test is used to show this power-law model best describes the data. By using our pulsed BD measurement technique, we can measure a cumulative breakdown time range of more than 11 orders of magnitude, allowing to compare the different acceleration models with high discriminative power. We find that the power-law best describes the voltage acceleration with a likelihood ratio $< 10^{-20}$.

In the next chapter, we will use the studied measurement techniques, the developed all-in-one maximum likelihood fit and the power-law voltage acceleration, to study the effect of variability and self-heating in the breakdown analysis of MTJs.

Chapter 5

Variability & self-heating analysis of MgO-breakdown

Apart from process-related extrinsic defects, the processing also influences diffusion mechanisms in the device, which in turn impacts the breakdown mechanisms. Furthermore, it is demonstrated that taking self-heating into account is imperative to make accurate lifetime predictions.

5.1 Introduction

In this chapter the techniques elaborated in chapter 4 are applied to investigate the impact of processing and different stack configurations on the reliability of STT-MRAM devices. We find that ion beam etching induces less variability than reactive ion etching in Section 5.2.1. However, the variability of the reactive ion etch can significantly be reduced by a post-etch treatment (Sec. 5.2.2). In addition, the MgO deposition technique, by natural oxidation or sputtering, does not significantly impact breakdown, as will be shown in section 5.3. In contrast, in section 5.4, we find that the spacer layers in the stack greatly impact breakdown in the MTJ. We propose an oxygen scavenging model to explain the increased susceptibility to breakdown for standard Ta spacers. In the final section, 5.5, the temperature acceleration is discussed. Self-heating has a strong impact on the breakdown in thin MgO barriers and results in so-called "beyond area scaling". By comparing 1.7 nm thick MgO with low self-heating to thinner MgO with significant self-heating, we estimate a temperature acceleration coefficient and use this coefficient to calculate the thermal resistance

of our devices. We find that in conventional breakdown measurements, the MTJs break down at temperatures between 200 and 300 °C (Sec. 5.5.4).

5.2 Impact of the MTJ patterning on reliability

One of the most crucial processing techniques is the patterning of the MTJ pillar via an ion beam etch (IBE) or a reactive ion etch (RIE). In Fig. 5.1, the simplified etch process is shown, consisting of three steps. Firstly, the hard mask is opened [Fig. 5.1(a)], secondly the stack is etched [Fig. 5.1(b)] and, finally the stack receives a post-etch treatment and in-situ encapsulation of the cell [Fig. 5.1(c)]. The post-etch treatment removes the damage inflicted on the sidewall and prevents the creation of a short across the MgO barrier. In addition, the encapsulation provides protection against moisture.

For the etching step, two frequently used techniques are compared, namely RIE and IBE [Fig. 5.1(b)]. IBE physically sputters the deposited stack under a pre-determined angle. Therefore, obtaining high areal density remains a difficult task, due to the shadowing effect, i.e. when the MTJ pillars are positioned at narrow pitch, the beam may sputter neighboring pillars. Currently, pitches below 100 nm can be obtained with optimized IBE conditions. RIE, however, does not use the directivity of IBE. RIE makes use of a high density plasma for stimulated physical and/or chemical interactions. With RIE small pitches can be obtained, but controlling the redeposition of etched material on the side walls is challenging. Optimizing the post-etch treatment improves the RIE process, see subsection 5.2.2. The impact of both techniques on breakdown is evaluated on the same single MgO stack in subsection 5.2.1.

5.2.1 Reactive versus ion beam etch

Etching of a pillar is a complex process. In order to simplify and investigate the impact of the etch process, we have measured breakdown after both RIE and IBE on a single MgO stack, with an MgO thickness of 1 nm [see Fig. 5.1(a)]. We have studied 3 RIE and 3 IBE etching gases and will note them hereafter by RIE (IBE) A, B and C (1, 2 and 3). All samples received a post-etch oxidation for 300 s at 500 W. The IBE results are depicted as different shades of blue triangles, whereas the RIE results as different shades of red circles. For all process flows, 60 devices have been measured across 6 adjacent dies. The devices have 3 nominal sizes (⌀ 60, ⌀ 100 and ⌀ 150 nm), and we have used the RVS technique at 3 ramp-rates (10, 100, 1000 mV/s), see section 4.5.2. In Fig. 5.2

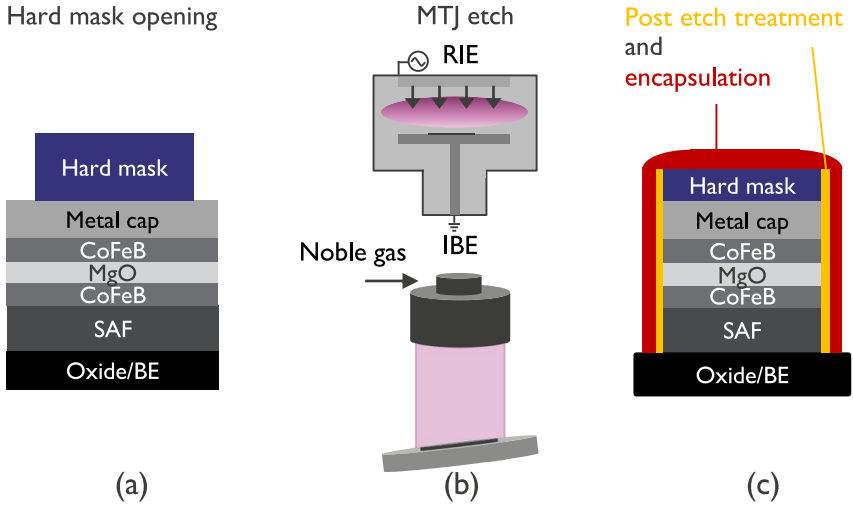

Figure 5.1: *Schematic of the MTJ etch consists of (a) the hard mask (HM) opening, (b) the MTJ pillar etch and (c) a post-etch treatment and encapsulation.*

we show the main findings. The Weibull slope (β) is larger for the IBE process than the RIE process [Fig. 5.2(a)]. For these typical process conditions the IBE β are significantly higher. The error bars represent the 95 % confidence error as derived from the maximum likelihood fit (Sec. 4.3.2). A higher β means less process variability and less influence from extrinsic effects or damage on the perimeter. Due to a higher β, IBE devices will outperform the RIE ones in case we tighten the requirements. We schematically illustrate this in Fig. 5.2(b), where a breakdown distribution is shown for a high β and a low β (upper panel) and the resulting lifetime extrapolation. Even though both breakdown distributions have the same $V_{BD,63\%}$, when scaling to lower percentiles the lifetime of the high β breakdown distribution will be significantly higher.

Note that there are no results for IBE gas 3 for ⌀ 60 and 100 nm, since this process has low yield, because of shorts.

Furthermore, the $V_{max,10\,\%}$ for 10 years lifetime and 10 % failed devices (i.e. 1000 FIT) are consistently higher for the IBE case (Fig. 5.2(c)). The $V_{max,10\,\%}$ of all sizes is plotted as a function of the median of the measured resistance before breakdown. For each size and etch process the median resistance of all 60 devices is used. Small devices have a higher resistance, and a higher $V_{max,10\,\%}$, as expected from the area scaling rule (Eq. 4.9). The $V_{max,10\,\%}$ of the IBE and

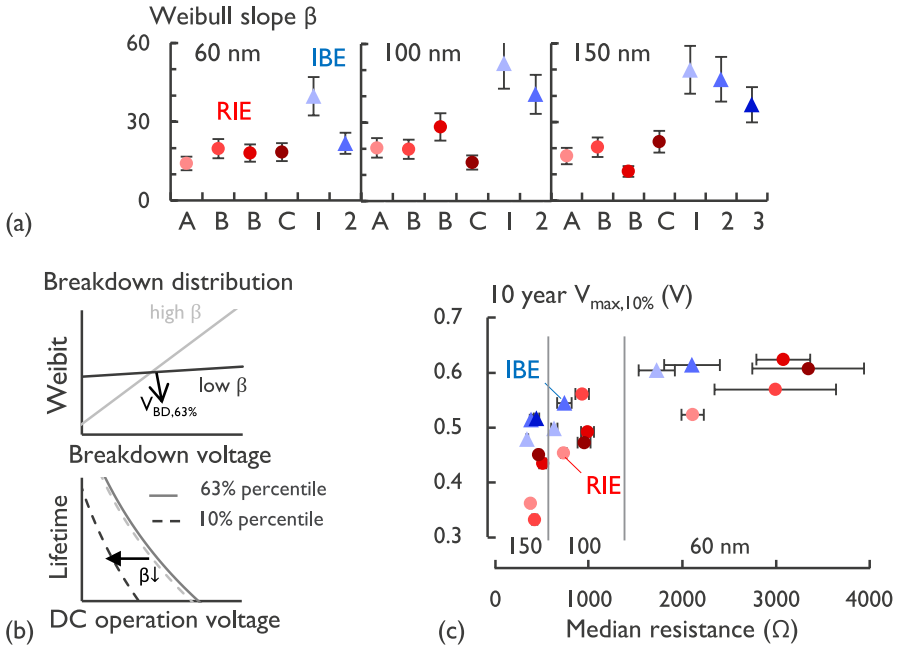

Figure 5.2: *(a) Extracted Weibull slopes (β) for various etch processes and 3 nominal sizes on a single MgO stack (see Fig. 5.1). IBE shows increased Weibull slopes (i.e. reduced variability). The error bars correspond with the error on β. (b) Illustration of the influence of β on the lifetime extrapolation for various lifetime criteria. (c) The maximum tolerable operating voltage for 1000 FIT as a function of median resistance. The error bars indicate the standard deviation in the measured resistance distribution.*

RIE are depicted as blue triangles and red circle, respectively. The different etch gases follow the same color code as in Fig. 5.2(a).

The error bars on this plot indicate the standard deviation in the measured resistance before breakdown. The resistance distribution of the RIE samples is not significantly different from that of IBE samples. For the larger sizes, i.e. small resistance, the error bars are within the symbol. For smaller sizes there is a larger spread in the resistance distribution. A small variation in diameter, has a larger impact on the smaller devices. As such, a variation ± 3 nm could explain the variation in resistance. This can also explain why a smaller β is observed for the smaller devices, because the increased variability in measured areas for the same nominal size will reduce the β. As a result, the β for the $\varnothing$ 60 nm devices is lower than the β of the $\varnothing$ 100 and $\varnothing$ 150 nm devices.

In summary, we find that the etching process causes more damage to the MTJ cell in the case of an RIE process than an IBE process.

5.2.2 Influence of post-etch treatments on RIE

The RIE damages the side walls of the MTJ pillars, leading to worse breakdown characteristics compared to IBE (section 5.2.1). However, by optimizing the post-etch treatment, a reduction of side wall damage and thus an improvement of the RIE process can be obtained. In this section, we consider the influence of sidewall oxidation method and the influence of oxidation energy for three different device sizes. All post-etch treatments are performed in-situ, i.e. within the controlled environment of the etch reactor.

Firstly, the effect of the side wall oxidation method is studied. This method can be either direct or remote. For direct oxidations a high density plasma and radicals are used. The difference between direct and remote is that in the remote case, radicals are filtered out of the plasma used for the oxidation. In addition, for the remote method, the temperature can be controlled. We have studied two temperatures, namely 65 and 125 °C. Samples without treatment only have an encapsulation, in this case the encapsulation was ex-situ and thus the samples were exposed for a long time (> hours) to air and uncontrolled environments before encapsulation.

For the following tests we have used 24 devices out of 4 adjacent dies for each size (⌀ 60, 100, 150 nm). Breakdown is measured with RVS at a ramp-rate of 100 mV/s.

The Weibull slope and the voltage where 63 % of the devices broke down (63 % V_{BD}) of two direct methods (1) 500 W, 300 s, (2) 0 W, 100 s, and two remote methods (1) 65 °C, (2) 125 °C (both 2000 W, 30 s), and no treatment, are shown in Fig. 5.3(a). The samples from a remote treatment and no treatment have higher 63 % V_{BD} and β values than the samples from a direct method. At these high oxidation energies, direct oxidation likely causes more damage to the MgO. Oxidation is required, because the wafer oxidized at 0 W is seriously influenced by shorts and shows low breakdown voltages. The reason why no treatment gives good results is not known. Possibly due to the ex-situ encapsulation, the side wall still oxidizes at a low oxidation energy, i.e. an "uncontrolled" post-etch treatment.

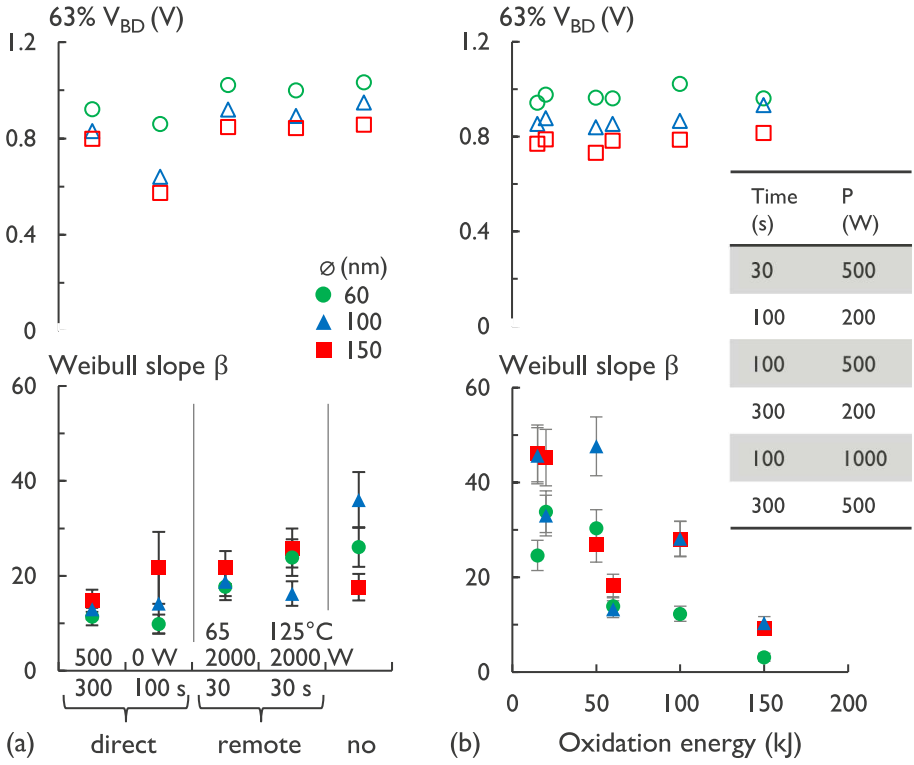

Figure 5.3: *Extracted Weibull slopes (closed symbols) and 63 % breakdown voltage (open symbols) for various post-etch treatment methods (a) and processing conditions (b) for RIE and 3 nominal sizes. (b) The Weibull slope increases as a function of a the oxidation energy. Inset: the different oxidation times and powers used.*

Secondly, the effect of the oxidation energy for direct oxidation is studied. Different post-etch oxidation powers and oxidation times are evaluated for a direct method. For these measurements, we have used 36 devices out of 2 adjacent dies for each size (⌀ 60, 100, 150 nm). Breakdown is measured with RVS at a ramp-rate of 100 mV/s. The β systematically increases for lower oxidation energies, regardless of the MTJ diameter, where oxidation energy is the product of oxidation power and oxidation time ($E = P \cdot t$) [Fig. 5.3(b)]. We attribute this improvement to reduced damage at the perimeter, causing reduced variability.

Furthermore, there is no clear trend in the 63 % V_{BD} as a function of the oxidation energy.

5.3 MgO barrier

For interest of breakdown reliability, a first assumption would be that the MgO layer and CoFeB/MgO interface are the most important. In this section, we compare two different MgO deposition techniques (Sec. 5.3.1). Also the impact of an MgO treatment during the deposition is studied in section 5.3.2. Finally, in section 5.3.3, we elaborate on the reliability trends when changing the MgO thickness. We find that by reducing the MgO thickness from 1 nm to 0.8-0.9 nm the reliability margin, i.e. between breakdown and write operations, is increased. For these studies a standard double MgO stack has been used (see Sec. 2.3.3).

5.3.1 Deposition technique

With respect to breakdown, only the MgO layer is important. To investigate the quality of the MgO layer, we compare two MgO deposition techniques: (1) RF-MgO sputtered from an MgO target and (2) naturally oxidized Mg in O_2 (i.e. in-situ DC-MgO) [42]. In DC-MgO the cycle of Mg deposition and oxidation is repeated until the desired thickness is obtained [see Fig. 5.4(a)]. Benefits for DC-MgO are a more uniform deposition over the wafer and multiple tuning parameters, with the most important parameters: Mg deposition thickness, oxidation time at each cycle and oxygen flow. For RF-MgO, the only control knobs are the sputtering time and deposition power. The RA product across the wafer shows less variability for the DC-MgO process, see Fig. 5.4(b).

For both techniques we compare the breakdown distributions of devices with the same area and comparable parallel resistance. In Fig. 5.4(c), the $V_{BD,63\%}$ is plotted as a function of the median resistance before breakdown. We have studied 50 devices across 25 dies close to the center, within a range of 70 mm from the center (the location of this die-range is indicated by the grey area in Fig. 5.4(b)). 3 nominal sizes ($\varnothing$ 60, 100 and 150 nm) have been measured with RVS at ramp-rate of 100 mV/s. The $V_{BD,63\%}$ increases with median resistance, as is expected from the area scaling rule, higher resistance is a smaller device. The error bars indicate the spread of the resistance and represent the standard deviation of the resistance distribution of the measured devices. The breakdown characteristics are quite similar for both deposition techniques and fall within each others confidence bounds [Fig.5.4(d)]. This suggest that another stack property dominates breakdown.

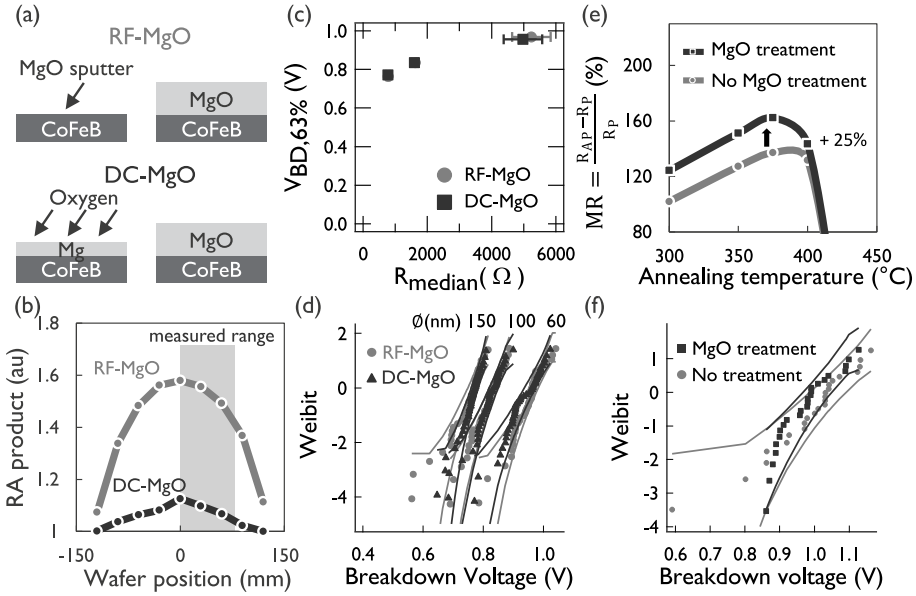

Figure 5.4: *(a) schematic of RF- and DC-MgO deposition. (b) DC-MgO has a better thickness uniformity across the wafer compared to DC-MgO. (c) $VBD_{63\%}$ as a function of median resistance for RF- and DC-MgO, the values are comparable. (d) The breakdown voltage distribution for 3 different sizes are identical within the confidence bounds for RF- and DC-MgO. (e) MR improves as a function of annealing temperature when applying an MgO treatment. (f) This MgO treatment has no significant impact on the breakdown voltage distribution.*

5.3.2 MgO treatment

In order to increase the TMR, an additional MgO treatment after deposition is added to the process. After MgO-deposition the wafer is cooled down to -100 °C, before CoFeB is deposited. The TMR is impacted by the interface quality between MgO and CoFeB. The introduction of the MgO treatment, results in a 25 % increase in TMR [Fig. 5.4(e)]. The cooling will minimize the atomic mobility during film growth and improves the interface quality. Besides the improvement in TMR, no significant impact on the breakdown characteristics is observed, since the distributions are identical within the confidence bounds [Fig. 5.4(f)]. For this study 24 devices of ⌀ 60 nm are used, picked from 4 adjacent dies.

5.3.3 MgO thickness

In this section, we discuss the influence of the MgO thickness on the reliability of breakdown. 1 nm MgO is optimal to easily integrate the memory element in the CMOS technology, because of an ideal resistance range. Further increase in MgO thickness will result in resistance values which are excessively high for the surrounding circuitry. 1 nm MgO corresponds to an RA product around 10 $\Omega\mu m^2$.

The MTJs are constantly scaled down to smaller dimensions, and smaller areas correspond with higher resistances. For these sizes, a lower RA will result in a better integration with the CMOS technology, lower resistances and lower switching voltages. In this section, we show that going to lower RA also increases the reliability margin, where we compare breakdown values with switching values.

In Fig. 5.5(a), the RA and 63 % V_{BD} trends are shown on a logarithmic scale as a function of MgO thickness. The MgO is sputtered by RF-MgO deposition. By tuning the sputtering time, we can alter the MgO thickness. The breakdown voltages are obtained by an RVS measurement with a ramp-rate of $100\,mV/s$ on 36 devices ($\varnothing$ 150 nm).

The RA depends exponentially on the MgO thickness, whereas the V_{BD} only linearly. From Fig. 5.5(a), one might wrongly conclude that the breakdown is field-driven, since $E_{BD} = \frac{V_{BD}}{t_{ox}} = cst$. This is not the case, however, because self-heating is not correctly taken into account and significantly affects the breakdown. We will show, in section 5.5.1, that there is more self-heating in the thinner MgO samples before breakdown than in the thicker MgO, making it unfair to make conclusions based on the observed linear trend of V_{BD} with MgO thickness. The main reason for the difference in self-heating is the linear trend of V_{BD} with MgO thickness, compared to the exponential trend of RA with MgO thickness, which is equivalent with the MgO resistance. Because of these trends, less power is dissipated before breakdown in thicker MgO, which means less self-heating, since an the exponential trend of RA increases faster than the quadratical trend of V_{BD}^2 and $P = \frac{V_{BD}^2}{RA} A$.

In the more relevant RA range (3 - 15 $\Omega\mu m^2$), we compare the switching to the breakdown characteristics, see Fig. 5.5(b). For 3 nominal areas, we perform an RVS with the devices initialized in the the most stable state, i.e. the AP-state. The bias direction favors switching to the P-state. As such, both V_{SW} and

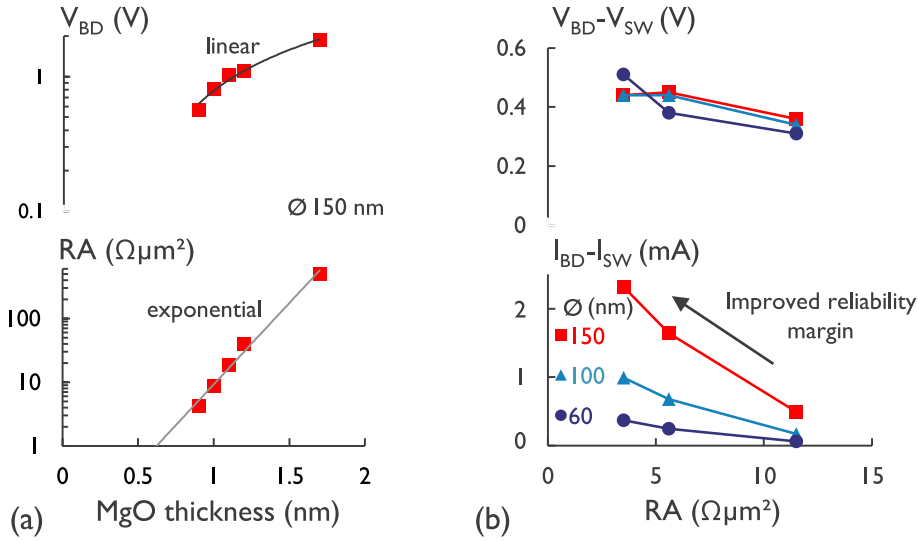

Figure 5.5: *MgO thickness study. (a) 63 % V_{BD} values (top) and the RA product (bottom) depend linearly and exponentially on the MgO thickness, respectively. (⌀ 150 nm) (b) Reducing the RA, increases the reliability margin between breakdown and switching. For 3 nominal sizes, the difference of the breakdown value with the switching value is increasing. (top) 63 % V_{BD} - median V_{SW} and (bottom) 63 % I_{BD} - median I_{SW}.*

V_{BD} are measured in the same RVS trace. For automatic detection of both parameters, we adjusted our triggering algorithm, as presented in [88].

Decreasing the RA, increases the reliability margin. The difference between the breakdown and switching voltage is plotted in the top plot in Fig. 5.5(b). In the bottom plot, the breakdown current is subtracted by the switching current. In both cases, the margin between breakdown and switching increases at lower RA, for all areas.

In conclusion, the reliability margin can be increased going to lower RA, by thinning down the MgO to 0.8-0.9 nm, because the switching voltage decreases faster than the breakdown voltage. In addition, these low RA values 3-5 $\Omega\mu m^2$ are necessary to achieve resistance values acceptable for technology, using ultra-scaled devices with sub-20 nm diameter.

5.4 Influence of stack configuration

As previously discussed in section 2.4, thermal robustness of the stack is imperative to enable the STT-MRAM technology. There has been an evolution in increasing the thermal budget from 300 °C to 375 °C and finally to 400 °C for more than 180 min, which resembles the potential temperature in BEOL process. To ensure the STT-MRAM stack can withstand this thermal budget, it is applied during an annealing step. The STT-MRAM-stack is continuously engineered to achieve higher thermal budget, higher TMR, lower switching voltages, while maintaining sufficient reliability. Changing the seed layers, spacer layers, thickness and atomic concentration, influences the MgO crystallization, TMR, PMA, reliability etc.

In this section, we will study the impact of the stack configuration on the reliability. For this we will use a thermal robust stack annealed at 375 °C, which is different from the stacks used in previous sections. A higher boron concentration is used in the CoFeB layers, to achieve a higher thermal budget [see Fig. 5.6(a)], where in the conventional CoFeB stack, the TMR quickly degrades around 400 °C, and in the B-rich remains stable at 400 °C.

The main cause of the TMR degradation is the activation of diffusion mechanisms at these high temperatures [77]. Nevertheless, high temperatures are necessary to enable high TMR. At high temperatures the amorphous CoFeB crystallizes to a bcc structure with (001) orientation, thus forming a good lattice match with the MgO(001) [137]. The MR depends on the interface quality between MgO and CoFeB. In order to achieve high MR, an optimal annealing temperature is needed to form a good interface without activating diffusion mechanisms. For this reason, we make use of a B-rich FL and RL. B-rich CoFeB provides a higher thermal robustness and maintains high MR up to temperatures above 400° C, see Fig. 5.6(a). It is hypothesized that the higher boron concentration can delay the crystallization of CoFeB as well as the activation of diffusion mechanisms. No breakdown penalty is observed going to a high boron FL (CoFeB+) [Fig. 5.6(a)].

Changing the RL and FL spacer significantly improves the reliability. In the top panel of Fig. 5.6(b) we compare the breakdown results for ⌀ 60 nm nominal sized MTJs with a Ta-based FL spacer and 3 different RL spacers, not disclosed. There is a clear improvement using the RL spacers compared to the Ta-based reference, resulting in a higher 63 % breakdown voltage and in a higher Weibull slope. In addition, if the FL spacer is changed, the breakdown voltage increases even further [bottom panel in Fig. 5.6(b)].

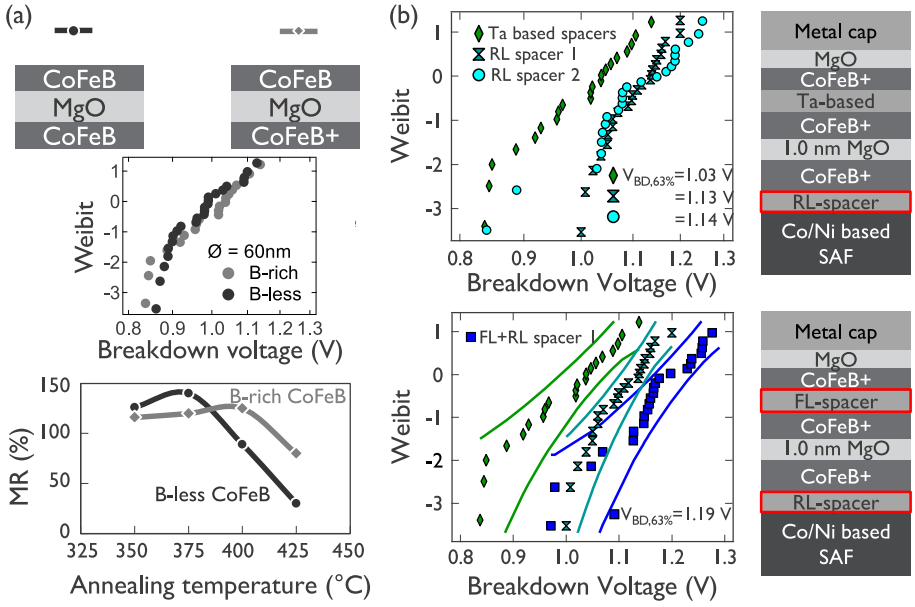

Figure 5.6: *(a) MR as a function of the annealing temperature. The B-rich stack shows higher thermal robustness against annealing and therefore maintains the MR up to higher temperatures. (b) Breakdown voltage distribution for different reference layer and free layer spacers. RL/FL spacer 1 has significantly better breakdown properties (lines are the 95 % confidence bounds).*

In order to explain these results, we explore the role of oxygen outdiffusion from the MgO layer. We propose a diffusion based model [Fig. 5.7(a)]. During processing, oxygen diffuses out of the MgO, leaving behind oxygen vacancies. Simultaneously, spacer layer material diffuses towards the MgO. Both diffusion mechanisms reduce the MgO resilience to breakdown when a high thermal budget is applied (BEOL processing). Diffusion of Ta in CoFeB is already reported in [77]. Moreover, when Ta diffuses towards MgO, it will more easily scavenge O from MgO, again reducing the MgO resilience to breakdown.

In order to find experimental proof of these mechanisms, we have performed element selective x-ray absorption measurements at different annealing temperatures. These absorption measurements provide information on the amount of oxygen present in CoFeB. The oxygen in this case could only originate from the MgO layer.

We compare the impact of two spacers on O-scavenging. In Fig. 5.7(b) the oxygen concentration in CoFeB is shown as a function of annealing temperature. For Ta-based spacers, more oxygen diffuses out of the MgO. The reduced amount of oxygen in the MgO leads to a lower RA product, as is experimentally measured and shown in Fig. 5.7(c). Finally, from Oxide RAM technology, Ta is known to be a strong oxygen scavenger [20]. As such, an oxygen diffusion mechanism can explain why higher breakdown voltages and Weibull slopes are found for non-Ta-based spacers.

In conclusion, spacer design impacts the formation of oxygen vacancies in the MgO. Proper selection of the spacer material in addition to the MgO etch optimization is as such crucial to enable good breakdown properties.

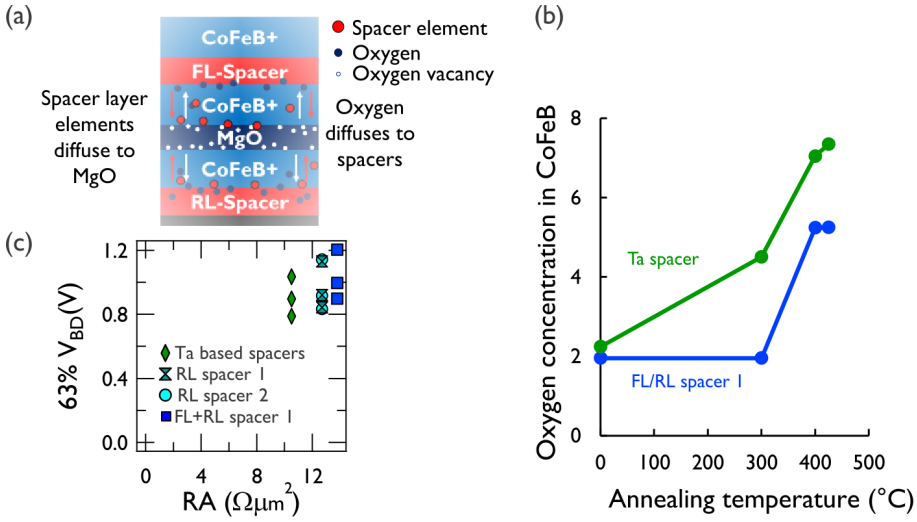

Figure 5.7: *(a) Schematic picture representing the diffusion of oxygen out of the MgO to the spacer layers or spacer layer material into the MgO, causing defects in the MgO. (b) Results of X-ray absorption spectroscopy measurements at the Fe edge of simplified stacks Co/X/CoFeB(1 nm)/MgO(nm)/X/Co, where X is either 4 Å Ta or 5 Å RL or FL spacer 1. The fraction of oxygen as a function of annealing temperature (i.e. increasing diffusion rate), is shown. (c) 63 % breakdown voltage as a function of device resistance area product (RA). The metallic spacer layers influence the RA via oxygen scavenging from the MgO. In particular, reducing RA and breakdown resilience, for Ta-based spacers.*

5.5 Derivation of self-heating via temperature acceleration

Self-heating plays an important role in the reliability of STT-MRAM. As such, the self-heating causes an apparent deviation from the Poisson area scaling rule. Smaller devices are more reliable than one would expect solely from area scaling. For a 1 nm MgO barrier, this results in breakdown improvement "beyond area scaling", i.e. small areas have higher V_{BD} than expected from area scaling of large devices. In Fig. 5.8, we illustrate this discrepancy for a 375 °C annealed stack with RA = $13\,\Omega\mu m^2$ (1 nm MgO). Scaling down the MTJ area is thus beneficial for device reliability.

This effect can be explained by self heating as is suggested in [51, 3] and in more detail later in Kan et al. [58]. At the same voltage, small areas will heat up less than large devices, causing an additional temperature-induced V_{BD} increase. Based on simulations, see section 3.2, the estimated self-heating at the same voltage level before breakdown depends on area (see Fig. 3.6(a)). However, no complete model has been derived to correctly include the self-heating effects on breakdown.

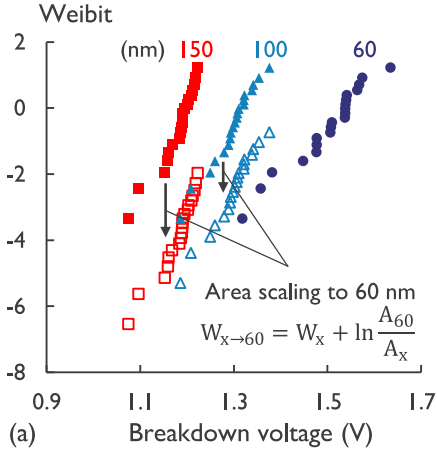

Figure 5.8: *Illustration of "beyond area scaling": Breakdown distributions are plotted for 3 nominal sizes (⌀ 60, 100, 150 nm) of a 375 °C annealed stack with RA = $13\,\Omega\mu m^2$ (1 nm MgO). The open symbols are the corresponding distributions rescaled to the ⌀ 60 nm size, following Eq. 4.9, and do not fall on top of each other. Moreover, the smaller devices are more reliable than one would expect solely from area scaling, i.e. "beyond area scaling".*

In this section, the concept of self-heating and its effect on breakdown is more thoroughly discussed. Firstly, we will extend the present breakdown model to include temperature acceleration, in section 5.5.1. Secondly, in section 5.5.2, 1.7 nm thick MgO is used to minimize the effect of self-heating. For thick MgO it is possible to observe Poisson perimeter and area scaling. Now, we experimentally find that RIE samples show perimeter scaling, whereas IBE samples show area scaling, proving that RIE inflicts damage on the sidewalls, weakening the MTJ. Thirdly, we discuss the effect of oxide thickness on temperature acceleration and introduce a phenomenological model in section 5.5.3. Last in section 5.5.4, the self-heating is calculated for a 1 nm MgO.

5.5.1 Temperature acceleration in the breakdown model

As described in chapter 4, breakdown can be characterized by the formation of a percolation path. This formation leads to a Weibull distribution for time-to-breakdown and voltage-to-breakdown, and follows the Poisson area scaling (Eq. 4.9). In section 4.6, the acceleration of defect generation was discussed. For thin oxides we use the power-law to describe the voltage acceleration. Combining Weibull distribution, area scaling and voltage acceleration results in following breakdown distribution:

$$W = \beta\Big[\ln(t_i) - \ln(\eta_{ref})\Big] - \beta \cdot n \ln\left(\frac{V_i}{V_{ref}}\right) + \ln\left(\frac{A_i}{A_{ref}}\right) \tag{5.1}$$

Adding the temperature acceleration, the Weibull distribution of breakdown becomes:

$$W = \beta\Big[\ln(t_i) - \ln(\eta_{ref})\Big] - \beta \cdot n \ln\left(\frac{V_i}{V_{ref}}\right) + \ln\left(\frac{A_i}{A_{ref}}\right) + \beta \cdot \alpha_T \Big[\frac{1}{k_b T_{BD,i}} - \frac{1}{k_b T_{ref}}\Big] \tag{5.2}$$

with α_T the temperature acceleration coefficient, $T_{BD,i}$ the temperature at breakdown for sample i and T_{ref} the reference temperature.

In literature, temperature acceleration has been extensively studied for SiO_2 [57, 129, 130] and MgO [51, 58, 21]. A linear dependence of $\ln(t_{BD})$ with T is found in [57], in contrast to the often reported Arrhenius-like T-acceleration [129, 58], Eq. 5.2. In the measured temperature range, both models are indistinguishable.

In section 5.5.2 we will report on our findings for MgO, that support an Arrhenius-like temperature acceleration. Furthermore, for thicker MgO, self-heating effects will become less important than for 1 nm thin MgO. Simplified, the breakdown temperature depends on the external temperature and the self-heating temperature:

$$\begin{aligned} T_{BD} &= T_{ext} + T_{self-heating} \\ &= T_{ext} + R_{th} \cdot P_{BD} \\ &= T_{ext} + R_{th} \frac{V_{BD}^2}{RA} A \end{aligned} \tag{5.3}$$

As seen in section, 5.3.3 the V_{BD} increases linearly with MgO thickness, but RA increases exponentially with MgO thickness. As a result, the impact of self-heating temperature will reduce for thick MgO ($t_{MgO} \uparrow \rightarrow \frac{V_{BD}^2}{RA} \downarrow$). Therefore, measuring thick MgO, will enable the determination of the breakdown temperature acceleration coefficient α_T, since self-heating can be minimized and thus T_{BD} mostly depends on the applied T_{ext}.

5.5.2 Reduced self-heating effect in thick MgO

To perform this temperature acceleration study, we make use of the Mbit array vehicle and a standard Co/Pt bottom pinned MTJ with Ta spacers, discussed in section 5.4 without an optimized post-etch treatment. The estimated MgO thickness is 1.7 nm. Within the Mbit array, we assume negligible oxide thickness variation. We perform the test on 500 randomly chosen MTJs for each condition. We measure 4 different ramp-rates at 4 different temperatures (25, 50, 75 and 100 °C). These measurements are performed on both RIE and IBE-etched arrays.

Note that for the thick oxide, the parasitic effect of an R_{series} will not significantly impact our analysis. One of the most important parasitic effects is an unknown R_{series}. We hypothesize, in the appendix A.3, that it originates from oxide present between M3 and the BE. Since for thin MgO samples, the resistance of the MTJs is low with respect to the parasitic R_{series}, the stress voltage cannot be correctly extracted. For thicker MgO, the MTJ resistance is much larger than this unknown R_{series}, hence V_{stress} is known with higher accuracy.

The results of the extracted breakdown reliability parameters as a function of temperature are plotted in Fig. 5.9. For the RIE, we find low Weibull slopes,

not expected for such thick MgO barriers [Fig. 5.9(a)]. The Weibull slopes have a limited temperature dependence. On the other hand, Ln η shows an Arrhenius behavior as a function of temperature, with an activation energy around $\alpha_T \approx 0.4\,eV$ [Fig. 5.9(b)]. The power-law exponent n also shows an Arrhenius behavior as a function of temperature with an activation energy $\phi_T \approx -1.4\,eV$ [Fig. 5.9(c)]. This Arrhenius dependency will prove to be important to include in the breakdown model (Sec. 5.5.3).

We test the thick MgO for perimeter and area scaling in Fig. 5.10. The breakdown voltage distribution of $\varnothing$ 150 nm is rescaled using the Poisson scaling rule to $\varnothing$ 75 nm. For thick MgO, we do not observe "beyond area scaling", as seen in 1 nm thick MgO (Fig. 5.8). The fact that we observe area scaling supports the hypothesis that the difference in self-heating for different areas have a negligible effect on breakdown in the 1.7 nm thick MgO. We will further elaborate on this below. We can also conclude that the defects responsible for breakdown are uniformly distributed, i.e. not edge-damage related, for IBE.

The thermal resistance does not depend on the MgO thickness, but more on the MgO surrounding layers and the thermal boundaries (Sec. 3.3.2). In addition, in section 5.3.3, it was shown that V_{BD} increases linearly with oxide thickness, whereas RA increases exponentially with oxide thickness. As a result, the power generated before breakdown decreases, as well as the self-heating (Eq. 5.3),

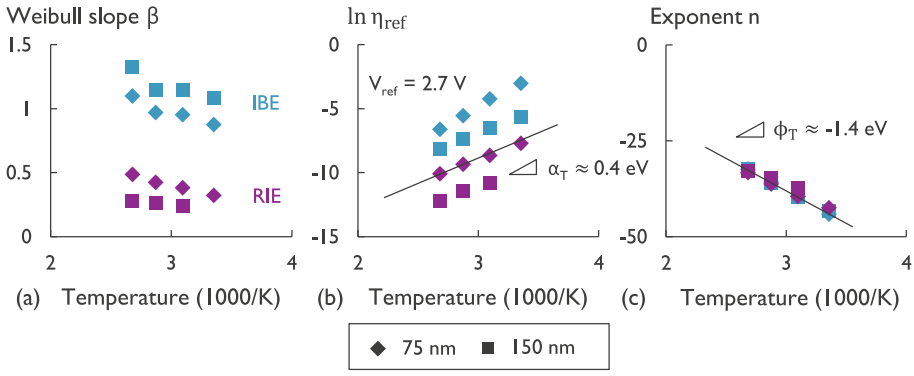

Figure 5.9: *Temperature acceleration in 1.7 nm thick MgO for 2 nominal sizes 75 nm (diamond) and 150 nm (square) and for a RIE (purple) and IBE (blue) etch (measured at 4 external temperatures 25, 50, 75 and 100 °C). (a) Weibull slope β, (b) $\ln \eta_{ref}$ and (c) exponent n. $\ln \eta_{ref}$ and n show Arrhenius behavior with activation energy $\alpha_T \approx 0.4\,eV$ and $\phi_T \approx -1.4\,eV$, respectively.*

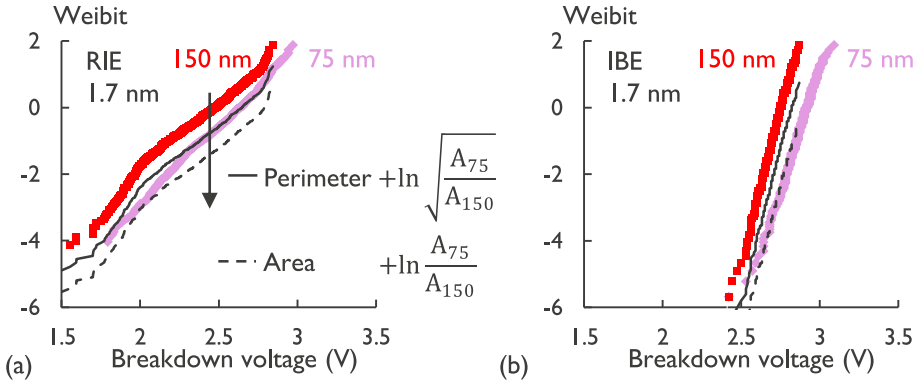

Figure 5.10: *Test for perimeter or area scaling for 1.7 nm thick MgO with ⌀ 150 (red, squares) and 75 nm (pink, diamonds). The distribution with ⌀ 150 nm has been replotted using the Poisson scaling rule. This results in the black line in case of perimeter scaling is assumed, or the dashed line in case area scaling is assumed. (a) For RIE, the bulk of the distribution fits better with perimeter scaling than area scaling, indicating preferential breakdown around the perimeter, (b) in the case of IBE area scaling is observed.*

when increasing the MgO thickness. Indeed, for the 1.7 nm thick MgO, the power before breakdown is between 10 μW - 100 μW, whereas for the 1.0 nm MgO this is between 100 μW - 1 mW (Fig. 5.11(a)). These powers result in self-heating for both the 1.0 nm as the 1.7 nm MgO, however, where the self-heating is negligible compared to room temperature for the 1.7 nm MgO, this is certainly not the case for the devices with 1.0 nm MgO (Fig. 5.11(b)). This self-heating for the different areas is estimated using simulated thermal resistances R_{th} and the power at breakdown condition. The large absolute differences in self-heating for different areas in 1.0 nm MgO lie at the base of the so-called "beyond area scaling".

In summary, for the 1.7 nm MgO we observe perimeter scaling and area scaling for RIE and IBE samples, respectively. The low Weibull slope β was already an indication of edge damage caused by the RIE and now the observed perimeter scaling confirms this edge damage. No "beyond area scaling" is observed, because the powers before breakdown are too low to let self-heating affect the area scaling.

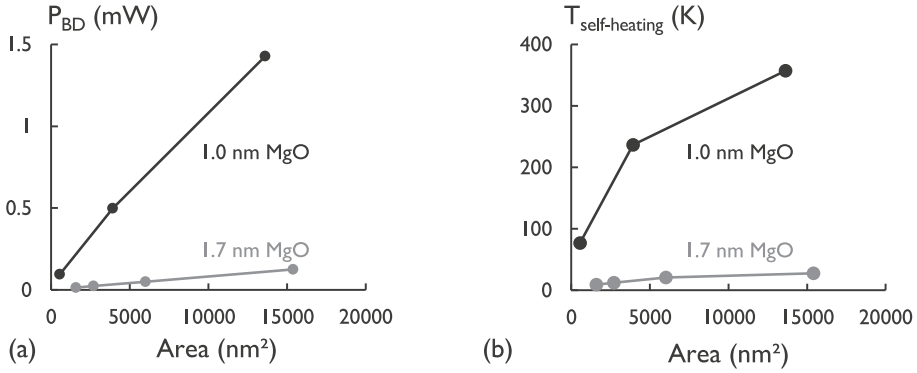

Figure 5.11: *(a) Power and (b) self-heating temperature before breakdown in 1.0 nm and 1.7 nm MgO. These results indicate the reduced effect of self-heating in thicker MgO. The 1 nm data comes from the same datasets used in Fig. 5.8, and the 1.7 nm dataset from the IBE samples (Fig. 5.10(b).)*

5.5.3 Effect of oxide thickness on temperature acceleration

To estimate how the coefficient α_T will change when reducing MgO thickness, we take a look at the vast dataset of SiO_2. In most studies, the temperature dependence of breakdown is evaluated for oxide thickness in the range of 2 nm up to more than 10 nm. In these studies it was found that the temperature acceleration α_T depends on the SiO_2 layer thickness. In [129], however, Wu et al. point out that the stress voltage is the actual controlling factor for the temperature dependence. Specifically, the temperature dependence of the power-law exponent, see Fig. 5.12(a), explains the overall temperature dependence of various oxide thicknesses. A similar dependence is also found for our thick MgO. We empirically fit the temperature dependence of the voltage power-law exponent by an Arrhenius model:

$$n(T) = \frac{\phi_T}{k_b T} + n_0 \,. \tag{5.4}$$

Here ϕ_T characterizes the Arrhenius activation constant, and n_0 is the power-law exponent at infinite temperature. Using the power-law voltage acceleration (Eq. 4.27), the temperature dependence of the exponent n (Eq. 5.4) will result in a

voltage dependent temperature acceleration coefficient $\alpha'_T(V)$:

$$\begin{aligned}\ln t_{BD} &\sim n(T) \ln \frac{V}{V_{ref}} + \alpha_T \cdot \left(\frac{1}{k_b T} - \frac{1}{k_b T_{ref}} \right) \\ &\sim n_0 \ln \frac{V}{V_{ref}} - \alpha_T \frac{1}{k_b T_{ref}} + \frac{\phi_T \cdot \ln \frac{V}{V_{ref}} + \alpha_T}{k_b T} \\ &\sim C + \frac{\phi_T \cdot \ln V - \phi_T \cdot \ln V_{ref} + \alpha_T}{k_b T} \\ &\sim C + \frac{\alpha'_T(V)}{k_b T} \,.\end{aligned} \tag{5.5}$$

Here the terms independent of temperature are combined in the constant C. The Arrhenius coefficient $\alpha'_T(V)$ scales linearly with $\ln V$. In Fig. 5.12(b) this coefficient is plotted for a large dataset collected from IBM and imec samples for SiO_2 [127, 129, 130, 57]. The coefficient α'_T is the slope of a linear fit of $\ln t_{BD}$ as a function of $\frac{1}{T}$. Each point is thus the Arrhenius slope α'_T fitted from a collection of $t_{BD,63\%}$ extracted from a CVS at different temperatures. These measurements are performed for oxide thicknesses ranging from 2 nm to 8 nm. For thicker oxides and higher voltages >4 V, FN-tunneling and AHI are the major contributors to the oxide breakdown (see 4.6.2). Below 4 V, direct tunneling and a power-law-based defect generation are better suited to explain the breakdown results. In this case, α'_T depends logarithmically on voltage (see the blue line in Fig. 5.12(b) representing Eq. 5.5).

In conclusion, with an empirical Arrhenius dependence we fit the temperature dependence of the power-law exponent n, which in its turn causes the temperature acceleration of breakdown to be voltage dependent, i.e. $\alpha'_T \sim \ln V$. We will use this voltage dependence of α'_T to extrapolate α'_T to stress conditions used in 1 nm MgO.

5.5.4 Derivation of self-heating in 1 nm thick MgO

In previous section, we empirically derived the temperature dependence of the power-law exponent (see Eq. 5.4). For oxides, where DT dominates, this results in an increasing temperature dependence of the time to breakdown with $\ln V$. We also concluded, in Sec. 5.5.2, that self-heating significantly impacts the temperature at stress conditions in 1 nm MgO (Fig. 5.11(b)). However, no direct measurements of this Joule-heating can readily be performed. In the following,

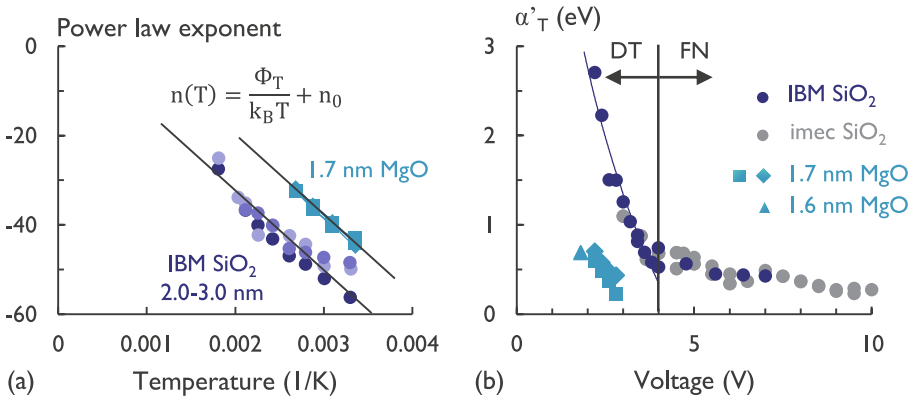

Figure 5.12: *(a) The power-law exponent as a function of temperature can be approximated with an Arrhenius law. (b) Each point is the extracted slope of an Arrhenius fit of* $\ln t_{BD}$ *data from* SiO_2 *and MgO data for different oxide thickness and voltage. At low voltage the direct tunneling takes over from FN-tunneling. With this a power-law model results in a* $\ln V$ *dependence of* α'_T*.* SiO_2 *data from [127, 129, 130, 57]*

we propose an indirect method to statistically estimate the temperature at stress conditions:

(Step 1) We estimate $\alpha'_T(V)$ for the 1 nm MgO.

(Step 2) We fit the thermal resistance, such that the temperature dependence matches with the estimated coefficient $\alpha'_T(V)$.

$$\begin{aligned} \ln t_{BD} &\sim C + \frac{\alpha'_T(V)}{k_b T} \\ &\sim \frac{\alpha'_T(V)}{k_b(T_{ext} + T_{SH})} \\ &\sim \frac{\alpha'_T(V)}{k_b(T_{ext} + R_{th}\frac{V_{BD}^2}{RA}A)} \end{aligned} \tag{5.6}$$

(Step 3) As a result, using the thermal resistance, we calculate the temperature at stress condition.

The general understanding of self-heating will open doors to better include the temperature acceleration in breakdown models and as such, perform more accurate lifetime predictions.

In Fig. 5.13.(a) the same thick MgO data is plotted as in Fig. 5.12(b), together with the extracted coefficient from the temperature analysis of 1.0 nm MgO stressed at 1.05 V, presented in [58]. The coefficients α'_T (purple circles) are determined from the data in the inset, where we add a self-heating term T_{SH} to represent the unknown self-heating at the constant stress voltage of 1.05 V.

The MTJ temperature becomes $T_{MTJ} = T_{ext} + T_{SH}$. For higher T_{SH}, the slope increases. We change T_{SH} from 0 K to 250 K. At 250 K, $\alpha'_T \approx 1.7$ eV, this is in line with the expected $\ln V$ dependence (Eq. 5.5). The power during CVS is estimated to be around 420 μW, since the V_{stress} is 1.05 V and the resistance around 2.6 kΩ ($\varnothing$ 70 nm device). Using Eq. 5.3 we then find an R_{th} between 0.5 and $0.6{\cdot}10^6$ K/W, which is in reasonable agreement with the simulations on our stack, that result in an R_{th} of $0.7{\cdot}10^6$ K/W. Also, in [58], the authors simulate a similar value of $R_{th} \approx 0.4{\cdot}10^6$ K/W.

In a final step, we determine the breakdown temperature of our devices. Therefore, we take the slope α'_T to be 1.7 eV. We fit R_{th}, such that the added self-heating $T_{SH} = P_{BD} \cdot R_{th}$ results in the slope α'_T =1.7 eV, as depicted in Fig. 5.13(b). We performed RVS measurements on single devices across the wafer for 4 different temperatures (-10, 25, 75, 125 °C). Applying this methodology

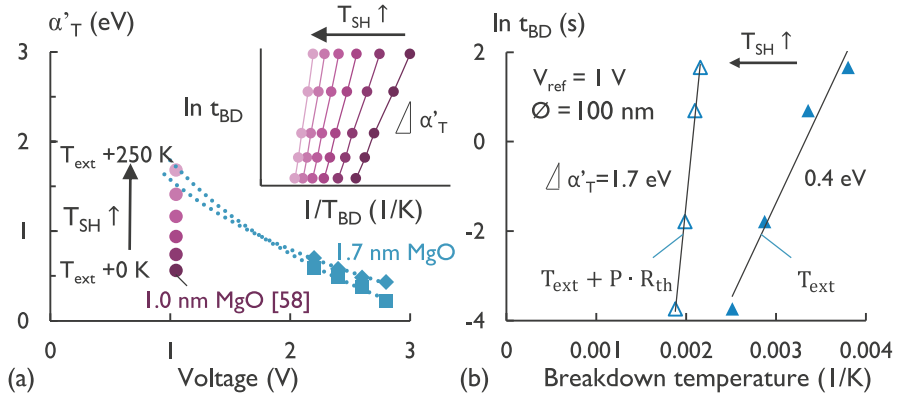

Figure 5.13: *(a) Breakdown temperature acceleration slope α'_T for MgO data. The dashed line is a fit assuming $\alpha'_T(V) \sim \ln V$ from the breakdown voltage range of the 1.7 nm MgO to the breakdown voltage range of the 1.0 nm MgO (data from [58]). Inset: adding self-heating temperature T_{SH} to the data, increases the Arrhenius slope α'_T. $T_{SH} \approx 250$ K results in an activation energy of ≈ 1.7 eV. (b) Methodology where R_{th} is fitted for a 1 nm MgO dataset (closed symbols) to match an $\alpha'_T \approx 1.7$ eV at temperature $T_{ext} + P \cdot R_{th}$ (open symbols).*

results in stress temperatures between 200 and 300 °C. The stress temperature slightly increases with the external temperature (see Fig. 5.14).

With the extracted R_{th}, the temperature at operation condition can be calculated, resulting in more accurate predictions. The extracted R_{th} are slightly lower than our simulations, see Fig. 5.14(b). The absolute value of the power-law exponent will thus increase when stressed at lower voltages, increasing the reliability. Thus when stressing at low voltage the lifetime is better than expected from the breakdown data at high voltage. This last result also explains why there are reports on unlimited endurance tests at operation voltage [59].

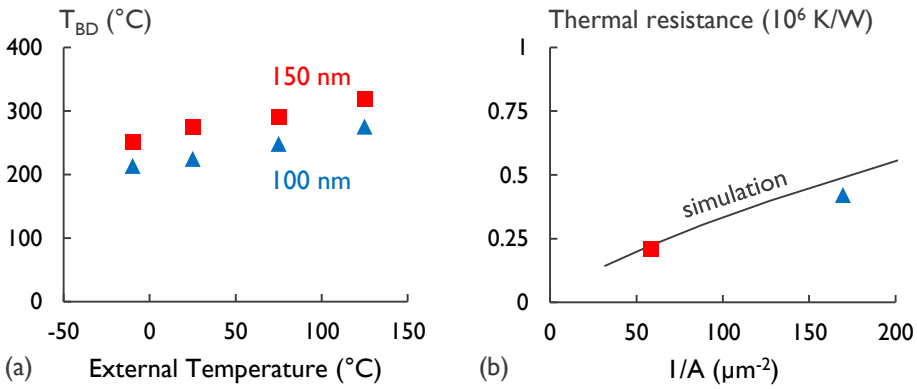

Figure 5.14: *(a) Calculated temperature before breakdown as a function of external temperature for RVS measurements on 2 different sizes. (b) Corresponding fitted thermal resistance as a function of* $\frac{1}{Area}$*, correlates well with our simulation.*

Ultimately, incorporating the self-heating into the breakdown model, results in an elevated temperature at breakdown conditions and a reduced temperature at operating conditions. The temperature differences will affect the power law exponent and the temperature acceleration. In Fig. 5.15, both effects are taken into account (red curves) and result in up to a factor of 1000 difference in 10 year lifetime (Fig. 5.15(b)). When extrapolating from the 63-percentile to 1 ppm, this difference increases up to 10^7 (Fig. 5.15(c)).

We conclude that breakdown measurements at stress conditions underestimate the lifetime, because increased self-heating reduces the lifetime compared to operation conditions. As such, incorporating self-heating and its effect on the lifetime extrapolation becomes imperative.

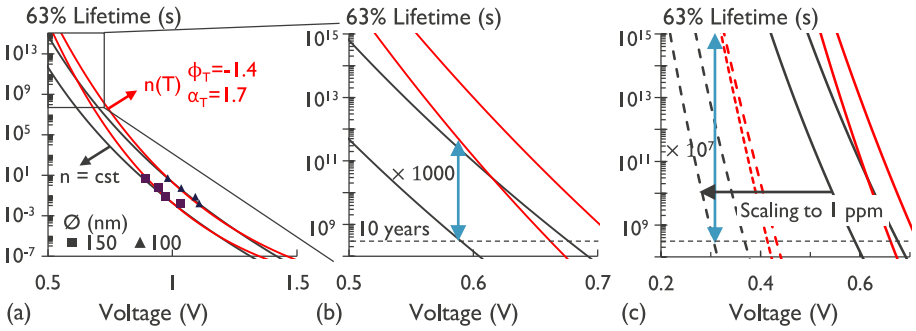

Figure 5.15: *(a) 63% lifetime extrapolation without self-heating and constant n (black lines), and with self-heating and (i) a temperature dependent n(T) determined by ϕ_T and (ii) temperature acceleration determined by α_T (red lines). (b) Zoom in to the 10 year lifetime region, for the 140 nm data the lifetime is underestimated by a factor of 1000. (c) The underestimated factor increases to 10^7 when performing percentile scaling to 1 ppm.*

5.6 Conclusions

In this chapter, we have discussed the impact of processing and stack configuration on breakdown. We observe that RIE induces more t_{BD} and V_{BD} variability than IBE, indicating more significant edge damage during a RIE etch. However, the variability of RIE can still be improved by an optimized post-etch treatment. In addition, we find only a minor impact of the MgO deposition technique on V_{BD}. A more significant impact is found for changing the thin spacer layers in the STT-MRAM-stack. For these last results, we propose an oxygen scavenging model to explain the increased susceptibility to breakdown for the standard Ta spacers.

The reliability margin can be increased going to lower RA, by thinning down the MgO to 0.8-0.9 nm, because the switching voltage decreases faster than the breakdown voltage. In addition, these low RA values of 3-5 $\Omega\mu m^2$ are necessary to achieve reasonable resistance values in ultra-scaled devices with sub-20 nm diameter.

In the final section we have discussed how self-heating has a large impact on breakdown. At typical RVS conditions, the degradation in thin MgO occurs at temperatures around 200 to 300 °C. Reducing self-heating will improve the breakdown characteristics. To achieve this, a reduction of the thermal resistance is required, such that for the same power there is less self-heating. Another

way is to decrease the area. Small areas have reduced self-heating at the same breakdown voltage. This temperature difference between different MTJ sizes explains the failure of the area scaling rule in 1 nm MgO, and the observation of so-called "beyond area scaling".

Furthermore, in thick MgO samples, self-heating is small and therefore it can be safely assumed that samples with different areas degrade and breakdown at the same temperature. For 1.7 nm thick MgO samples we experimentally observed Poisson perimeter scaling and Poisson area scaling for RIE and IBE-etched samples, respectively. The observed perimeter scaling confirms the hypothesis of edge damage inflicted for the RIE.

The temperature at operating conditions will be lower than at breakdown measurements conditions. As a result, the lifetime extrapolation to operation conditions is underestimated. It is therefore imperative to include self-heating effects in the breakdown model in order to perform accurate lifetime extrapolations from breakdown stress conditions to operating conditions.

Chapter 6

Measurement and modeling of retention

Accurately extracting the thermal stability, and thus characterizing the data retention in STT-MRAM, requires an accurate switching model and large switching statistics obtained by accelerating switching with magnetic field, current or temperature.

6.1 Introduction

The thermal stability factor ($\Delta = E_b/kT$) is a measure for the information retention time at operating temperatures, and to correctly characterize the data retention of STT-MRAM, its accurate determination is crucial. To achieve this, switching models are used to describe switching at accelerated conditions. Next, based on the switching models, Δ is then extracted by extrapolation to off-state conditions, i.e. zero magnetic field, zero current and ambient temperature. The correctness of Δ relies on how accurate the switching model can be extrapolated. There are two main switching models in literature, one is based on uniform switching of the magnetization, the other is based on domain wall nucleation and propagation. In addition, a recent study shows that a significant error in Δ is made using the conventional methods for Δ extraction [118]. In this chapter, we investigate the origin of this low accuracy. In addition, we compare the extracted Δ, using different acceleration techniques to obtain a better understanding in to which simplified switching model is best suited to describe

the energy barrier in STT-MRAM.

In this chapter, we introduce the magnetization dynamics that cause the switching statistics (Sec. 6.2). In Sec. 6.3, we discuss the concept of thermal stability and the necessity of accelerated testing. In Sec. 6.4, the methodology of accelerating switching with temperature, magnetic field and current is discussed, and accelerated switching measurements are performed and analyzed. Next, in Sec. 6.5, conventional switching models, namely the Macrospin and the domain wall switching model, are described in detail. We use a maximum likelihood fit to estimate the switching model parameters to accurately describe the accelerated switching data. The fitted parameters are found to be highly correlated and the accuracy of the fitted parameters is evaluated in Sec. 6.6. The Macrospin and domain wall switching model are experimentally compared in Sec. 6.7. In Sec. 6.8, current acceleration is compared with magnetic field acceleration at different external temperatures..

6.2 Magnetization dynamics in micromagnetics

In micromagnetics, the ferromagnetic body is treated as a continuous medium with slowly varying magnetic properties. In this continuum approximation, elementary volumes are used, small enough with respect to the scale over which magnetization varies significantly, but large enough with respect to the atomic lattice parameters [8].

Each elementary volume has a local magnetization vector M(r). The magnetic state of the body is then described by the magnetization of all elementary volumes. The orientation of the magnetization of each elementary volume can vary from point to point, but always tries to minimize the total energy of the body [23]:

$$G = \int_V \big(\epsilon_{ex} + \epsilon_{an} + \epsilon_d + \epsilon_{zeeman} + ...\big) dV\,, \tag{6.1}$$

with G the total free energy, defined by a volume integration of all energy contributions: the exchange energy ϵ_{ex} that tends to align neighboring magnetic moments, the anisotropy energy ϵ_{an} that tends to align the magnetization with preferential orientations, like the easy-axis, the energy contribution ϵ_d caused by the demagnetization field H_d that tends to form magnetic domains, and the Zeeman energy ϵ_{zeeman} that tends to align the magnetization with the applied external field. There are other energy contributions like stress/strain

and magnetostrictive self-energy, however we do not consider them, because the associated energies are small in the case of interest here.

The energy minimization explains the tendency for a specific magnetic orientation, corresponding with a local energy minimum. The energy minimization, however, does not explain how the system will evolve to this local minimum, i.e. magnetization dynamics. The Landau-Lifshitz-Gilbert (LLG) equation provides this extension to describe the magnetization dynamics.

To achieve switching in the MTJ, the magnetization of the FL has to reverse. Depending on the magnetization configuration of the FL and RL, the resulting state will have a high resistance or low resistance corresponding with an anti-parallel and parallel configuration, respectively.

In Sec. 6.2.1 we introduce the LLG equation, which can model the magnetization dynamics. Furthermore, we discuss the effect of the spin-transfer torque on the magnetization dynamics in Sec. 6.2.2. Finally, we discuss how to numerically simulate the magnetization dynamics using micromagnetic simulations and, the simplified approach, using a Macrospin model.

6.2.1 Introduction of LLG equation

An isolated magnetic moment μ precesses around an external applied magnetic field H_{appl} according to the equation (vectors are in bold):

$$\frac{d\boldsymbol{\mu}}{dt} = -\gamma_0 \cdot \boldsymbol{\mu} \times \boldsymbol{H_{appl}}\,, \tag{6.2}$$

where γ_0 is the product of the gyromagnetic ratio γ and the permeability of vacuum μ_0. In micromagnetics, the isolated magnetic moment is replaced by the magnetization M(r) for each elementary volume. In addition, the interactions inside the body are taken into account by the micromagnetic effective field H_{eff}. This precession contribution, together with the relaxation toward equilibrium, which is described by a phenomenological damping term, results in the so-called Landau-Lifshitz-Gilbert (LLG) equation [99]:

$$\frac{\partial \boldsymbol{M(r,t)}}{\partial t} = -\gamma_0\big(\boldsymbol{M} \times \boldsymbol{H_{eff}}\big) + \frac{\alpha}{M_s}\big(\boldsymbol{M(r,t)} \times \frac{\partial \boldsymbol{M(r,t)}}{\partial t}\big)\,, \tag{6.3}$$

where $\boldsymbol{M}$ is the magnetization of the free layer, $\boldsymbol{H_{eff}}$ is the effective magnetic field and α is the Gilbert damping term. Equation 6.3 consists of a precession and

a damping term. The magnetization will precess around $\boldsymbol{H_{eff}}$ at a frequency $\omega_0 = \gamma_0 H_{eff}$, like it was the case for the precession of an isolated magnetic moment around an external applied field (Eq. 6.2). The precession, which is energy-conservative, cannot go on forever and eventually $\boldsymbol{M}$ must align with $\boldsymbol{H_{eff}}$ (taken into account by the dissipative damping term α).

It is important to understand all the contributing fields that can change the magnetization dynamics in order to correctly model the thermal stability. The most important contributions to the effective magnetic field are an externally applied field H_{appl}, an anisotropy field H_{an}, an exchange field H_{ex} and a demagnetizing field H_d. In what follows we discuss in more detail these fields.

The externally applied field H_{appl} can be used to switch the magnetization state of the free layer, as will be further clarified in 6.4.2.

The anisotropy field H_{an} arises from the preference of the magnetization to align with certain magnetization orientations. It includes magnetocrystalline anisotropy, shape anisotropy, surface anisotropy and strain anisotropy.

The exchange field H_{ex} and demagnetizing field H_d are always present. The exchange field tends to align the magnetization, whereas the demagnetizing field tends to oppose the magnetization. This counteraction causes the formation of magnetic domains, where the energy is always minimized. The relationship between H_d and M is given by [23]:

$$H_{di} = -N_{ij} M_j \;\; i,\, j = x, y, z\,, \tag{6.4}$$

where N_{ij} is the demagnetizing tensor, which is generally represented by a symmetric 3 x 3 matrix. N_{ij} depends on the geometry of the cell.

6.2.2 Spin-transfer torque

Later on, the LLG-equation (Eq. 6.3) has been updated to take into account the spin-transfer torque (STT) effect [138, 98]:

$$\frac{\partial \boldsymbol{M}}{\partial t} = -\gamma_0 \big(\boldsymbol{M} \times \boldsymbol{H_{eff}}\big) + \frac{\alpha}{M_s}\big(\boldsymbol{M} \times \frac{\partial \boldsymbol{M}}{\partial t}\big) + \gamma_0 a_j \boldsymbol{M} \times \big(\boldsymbol{M} \times \boldsymbol{M_{RL}}\big) + \gamma_0 b_j \big(\boldsymbol{M} \times \boldsymbol{M_{RL}}\big)\,, \tag{6.5}$$

where $\boldsymbol{M_{RL}}$ is the magnetization of the RL, and a_j and b_j are proportional to the current density. The two additional terms arise from the torque that spin-polarized electrons can exert on the free layer. This torque has two main contributions, an in-plane and out-of plane torque. The in-plane component goes by names like Slonczewski torque, anti-damping torque and spin-transfer

torque. We will use spin-transfer torque or STT. The out-of-plane torque we will name field-like torque. The magnitude and effect of the field-like torque is not yet fully understood, it has a similar effect as if there would have been an extra magnetic field, hence the name [86, 10]. The spin-transfer torque can add or subtract to the damping torque, according to the direction of the current. In the nanoworld, it is more effective to exert torque by spin-transfer than by the magnetic fields created by currents in nearby bit lines, i.e. Oersted fields [23]. A simplified picture of the LLG-equation explaining the magnetization dynamics, is depicted in Fig. 6.1.

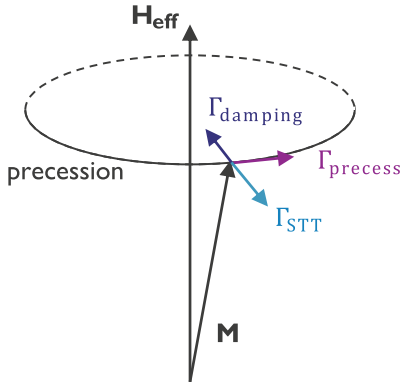

Figure 6.1: *Magnetization dynamics described by the LLG equation and Eq. 6.5. The magnetization* $\boldsymbol{M}$ *precesses around the effective field* $\boldsymbol{H_{eff}}$ *influenced by several torques.* $\Gamma_{damping}$ *is the Gilbert damping torque,* Γ_{STT} *is the torque caused by STT and* $\Gamma_{precess}$ *the torque causing the precession from* H_{eff}*. The field-like torque normal to the STT is not shown.*

In Fig. 6.2 we schematically explain how the effect of STT can lead to free layer switching. In the top schematic, a current flows from the free layer to the reference layer. In that case, STT favors parallel state. Electrons get polarized flowing through the thicker RL. In a reaction to the polarization, the electrons exert a torque on the RL. However, the RL, in theory, is stable enough not to be influenced by this torque. Next, the polarized spin-current tunnels through the barrier and will now exert a torque stabilizing parallel alignment.
Reversing the current, the STT will favor anti-parallel alignment. Now, electrons acquire spin polarization from the FL and most of them tunnel through the barrier to ineffectually exert torque on the RL. However, some, predominantly those with spins opposite to the RL, are reflected, and flow back to the FL, exerting a torque trying to stabilize the AP-state. Thus, depending on the direction of the current, the MTJ can be switched from AP-to-P or P-to-AP.

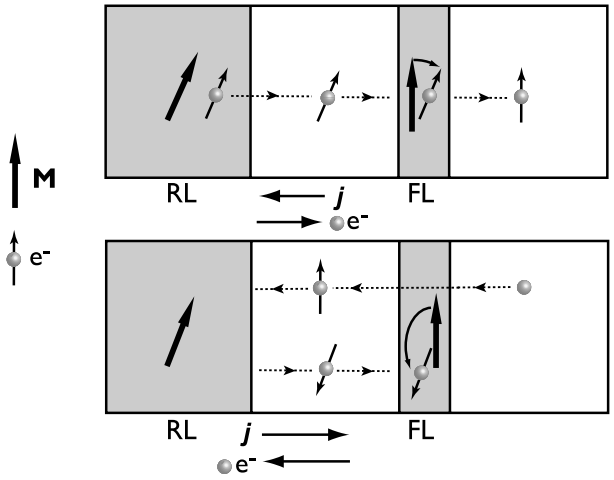

Figure 6.2: *Spin-transfer torque associated with flow of angular momentum in an MTJ. The sense of the torque, exerted by the electron spin and acting upon the magnetization of the free layer (FL), is indicated by the rotating arrows and depends on the current direction. The unshaded layers are nonmagnetic. Modified from [23].*

This writing mechanism is slightly asymmetric. In principal, it is easier to switch from AP-to-P, since this is a majority pass-through spin effect. Switching from P-to-AP, however, is more difficult, since this is a minority accumulating spin effect. This asymmetry can be tuned by engineering the stray field, which the pinned layer exerts on the free layer, resulting in an offset field. Therefore, in reality the most favorable state and the most difficult state to switch from, depends on the subtle engineering of this offset field.

The spin-transfer torque $\boldsymbol{M} \times \boldsymbol{M_{RL}}$ is zero when both magnetizations are perfectly collinear. Therefore, thermal fluctuations and non-ideal fabrication are important for efficient switching, since they will cause an initial angle between the magnetization and the polarized electrons.

6.2.3 Introduction of micromagnetic simulations and Macrospin approach

There are two main methods to learn more about the magnetization dynamics in STT-MRAM devices, via micromagnetic simulations or the Macrospin approach.

In micromagnetic simulations, the LLG equation is numerically solved for elementary volumes in the MTJ. Material parameters like exchange stiffness A_{ex}, anisotropy energy K_{eff}, magnetization saturation M_s and damping constant α are used as input, together with the device geometry and an initial configuration of the magnetizations.

Based on these inputs, the simulation computes all contributing fields and magnetic states necessary to solve the LLG-equation (Eq. 6.5). This can be very time consuming and relies on knowledge of the material parameters. The full micromagnetic simulation approach gives the highest accuracy and is therefore suited for physical understanding of the magnetization dynamics. It is, however, not straightforward to extract knowledge about the energy barrier as a function of applied acceleration conditions. Reducing the switching voltages results in longer switching times, significantly impacting the simulation time, which uses time steps based on the magnetization precession frequency ω_0, which is in the order of tens of gigahertz [23, 68].

The Macrospin approximation reduces the complexity and requires much less computation time. The magnetization of the cell is simplified to a single magnetic moment, i.e. a single domain with infinitely high exchange energy. This way, the exchange field drops out of the LLG-equation, and the demagnetizing field is simplified. The demagnetizing field makes micromagnetic simulations very time consuming, since it needs to be computed everywhere in the structure and at every time-step The single domain assumption is reasonable for very small dimension, i.e. MTJs with a diameter $< 20\,\text{nm}$, depending on the exchange constant. A higher exchange constant will favor the formation of Macrospin. For the Macrospin approximation it is possible to derive the energy barrier as a function of applied acceleration conditions. As a result, a lot of the measurement techniques used in literature are based on this Macrospin assumption, which we will discuss in Sec. 6.4.

Unfortunately, for larger dimensions a single domain approximation is not valid. The demagnetization energy increases with the volume of the magnet. At a certain dimension the system minimizes its energy, by forming domains and domain walls. This way, the demagnetization energy is reduced, but these non-uniform magnetization configurations cost a significant amount of exchange energy. The exchange energy for a Bloch wall per unit domain wall area is $\sigma = 4\sqrt{A_{ex}K_{eff}}$ (J/m^2) [23]. This energy we will call the domain wall energy. Since the demagnetization energy increases with volume, and the domain wall energy with area, at a certain dimension it will be more favorable to form a domain wall to minimize the system energy. A schematic illustration of this

effect is shown in Fig. 6.3, where the dimensions are too large for a uniform magnetization. For these dimensions, reversal of the magnetization occurs preferably by the nucleation and propagation of domain walls, i.e. the domain wall model (see Fig. 6.4). To summarize, for larger dimensions the domain wall switching mechanism is expected to be energetically more favorable than uniform switching of the full system magnetization. In section 6.5, we compare the Macrospin model with a domain wall model.

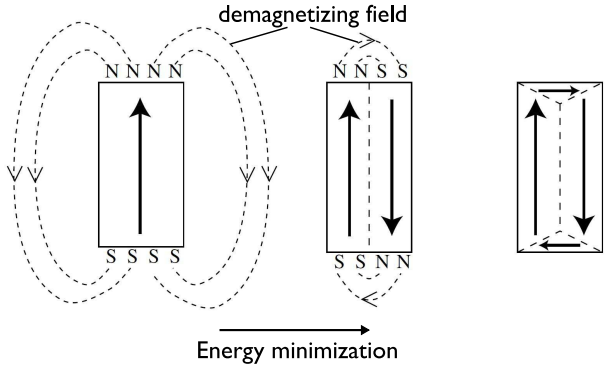

Figure 6.3: *Reduction of the total energy by domain formation in a large ferromagnet. As such, the demagnetization energy is minimized. Replotted from [102].*

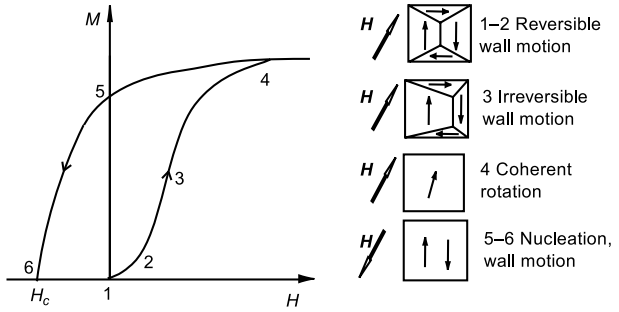

Figure 6.4: *Initial magnetization curve and demagnetization curve of a ferromagnet. (1) In equilibrium, energy is minimized by forming a multi-domain state. (2)→(3) By applying a magnetic field the domain walls start to move, eventually eliminating all but the one most favorably oriented domain. (4) For even higher fields, the domain will coherently rotate to align with the magnetic field. (4)→(5) At some point on the reverse segment, domains nucleate and begin to propagate, resulting in a multi-domain state (6). Replotted from [23].*

6.3 Thermal stability and the need for accelerated testing

The influence of thermal fluctuations on the magnetization was introduced by Néel [82], and further developed by Brown [13]. In Néel and Brown's model of thermally activated magnetization reversal, an energy barrier separates two equivalent ground states of opposite magnetization in a single domain magnetic particle. At high temperature, the system can transition from one state to the other by thermal activation over the barrier, i.e. the "thermally activated regime". At very low temperatures, close to absolute zero, the system can switch by quantum tunneling. In this thesis, we will only concentrate on the thermally activated regime. The time necessary to switch can then be explained by a model of thermally assisted crossing of the energy barrier (Arrhenius-Néel formalism) [82]. In the thermally activated regime, the switching rate λ is well described by an Arrhenius law[13], similar as the escape of particles over barriers:

$$\lambda(t) = f_0 exp\Big(- \frac{E_b(t)}{k_b T}\Big)\,, \tag{6.6}$$

with f_0 the attempt frequency and E_b the energy barrier. For the derivation of the switching probability it is assumed that the probability $P(t)$ that a particle did not switch after a time t follows [41]:

$$\frac{dP}{dt} = -\lambda(t) P\,. \tag{6.7}$$

The cumulative switching distribution F is then $1 - P$ and given by:

$$F_i = 1 - \exp\left(\int_0^{t_i} -\lambda\, dt \right). \tag{6.8}$$

Thermal stability Δ is defined as the energy barrier E_b divided by $k_b T$

$$\Delta = \frac{E_b}{k_b T}\,. \tag{6.9}$$

Using Eqs. 6.9 and 6.8, we find the switching probability in case of thermal reversal with a constant energy barrier in time:

$$F(t) = 1 - exp\Big[- f_0 t\, exp\Big(- \Delta\Big)\Big]\,, \tag{6.10}$$

This switching distribution is an exponential distribution and if plotted on Weibit scale results in a Weibull slope $\beta = 1$:

$$\ln\big(- \ln[1 - F(t)]\big) = 1 \cdot \ln t + \ln f_0 - \Delta\,. \tag{6.11}$$

The switching probability depends on a double exponential of the thermal stability factor Δ. Therefore, small changes in Δ, have a large impact on the switching time. We elaborate on this impact in the following example. Cells in an array need to retain their state for typically 10 years. In order to only have 1 failure in a full Mbit array operating at 80°C, following Eq. 6.10, a $\Delta > 54\,k_bT$ is required or $64\,k_bT_{ref}$ when referenced at room temperature. Depending on the application the required data retention will be more or less strict, e.g. storage memories require high thermal stability, whereas for working memories Δ can be lower. If the Mbit array fulfills the requirements, testing at 80°C, would result in only 1 failure after 10 years. That is why to extract Δ, accelerated testing to higher temperatures is required. As such, the observed failure distribution can be fitted with Eq. 6.10, resulting in an estimation of Δ.

Applying a magnetic field or current can lower the energy barrier and therefore accelerate switching. In addition, thermal reversal can be accelerated by increasing the temperature. Several methods have been developed to accelerate switching and extrapolate Δ: based on current pulses [119], magnetic field sweeps [41] or temperature [50]. In these methods, E_b is expressed by a simplified Macrospin model, see Sec. 6.5. The thermal stability is then extrapolated, based on this Macrospin model, from the accelerated measurement range, to the operation range, i.e. zero field, zero current and at reference temperature. The accuracy of the estimated thermal stability depends on how valid this Macrospin model is to extrapolate from the accelerated conditions (temperature, field and current) to the operation range. Later in Sec. 6.5, next to the Macrospin model, an additional model, called domain wall model, is introduced. In the next section (Sec. 6.4), we will elaborate on the different acceleration measurement techniques.

6.4 Acceleration measurement techniques used to extract Δ

In this section, we explore the different measurement techniques for magnetic field, current and temperature acceleration, for a single device and a Mbit array. Each measurement technique obtains switching data at accelerated conditions. The obtained switching distributions can then be fitted with switching models. In temperature acceleration, a high temperature is applied to faster induce reversal by thermal activation. Typically this reversal mechanism, is modeled by a simple Néel-Brown model (Sec. 6.4.1) [117, 121, 50]. At high temperatures, however, the material parameters change [46]. This temperature dependence of the parameters is not taken into account in the simple Néel-Brown model resulting in incorrect extrapolation. In contrast, there is no contribution from

temperature effects with magnetic field acceleration (Sec. 6.4.2). Finally in section 6.4.3, we discuss current acceleration, where the extrapolation to zero current also depends on the switching mechanisms, and in addition is impacted by the current-induced self-heating. The correctness of the extrapolation thus relies on the used switching model. Later in Sec. 6.5, we compare the models to find which model best describes the extrapolation.

6.4.1 Temperature acceleration

Due to the thermal energy, the magnetization can rotate around the local magnetic state. The higher the temperature, the larger the magnetization fluctuations, resulting in an increased probability to jump over the energy barrier. This form of switching is well described by an Arrhenius law and is modeled by Néel [82] and Brown [13] for single-domain magnetization.

The switching distribution on the same device is obtained by repeating switching cycles at high temperature. The switching distribution follows Eq. 6.8 and depends on the switching rate λ, which in turn depends on the energy barrier. Each switching cycle is composed of three steps: (1) Set and verify whether the device is in a known state, (2) bring the device at high temperature, and (3) monitor the device state and determine when the state reverses. Fitting the distribution allows determination of the energy barrier and hence Δ. The more switching cycles, the better the accuracy of the fit and the estimation of Δ.

For a single device, it can be very time consuming to obtain a high number of switching cycles, because of a low, unity Weibull slope and the large impact Δ has on the switching time (Eq. 6.11). Therefore, it is more convenient to measure a large number of devices at the same time. Each device, however, has its own Δ, not necessarily the same. In [50], it is assumed that the distribution of Δ is Gaussian. In that case Eq. 6.10 is transformed as follows:

$$F_{array}(t) = 1 - \int_{-\infty}^{+\infty} \exp\left(\int_0^t -\lambda\, dt\right) \cdot g(\Delta, \mu_\Delta, \sigma_\Delta)\, d\Delta\,, \qquad (6.12)$$

$$with\, \lambda = f_0 e^{-\Delta} = f_0 e^{-\frac{E_b(\Delta,T)}{k_b T}}$$

$g(\Delta, \mu_\Delta, \sigma_\Delta)$ is the probability density function of a Gaussian distribution with mean μ_Δ and standard deviation σ_Δ. In the short time regime this equation

simplifies to [117]:

$$\begin{aligned} F(t \to 0) &= 1 - \int_{-\infty}^{+\infty} \exp(-f_0 t \cdot e^{-\Delta}) \cdot g(\Delta, \mu_\Delta, \sigma_\Delta)\, d\Delta \\ &\approx 1 - \int_{-\infty}^{+\infty} (1 - f_0 t \cdot e^{-\Delta}) \cdot g(\Delta, \mu_\Delta, \sigma_\Delta)\, d\Delta \\ &\approx \int_{-\infty}^{+\infty} \frac{f_0 t}{2\sqrt{2\pi\sigma^2}} \cdot \exp\Big(-\frac{\Delta^2}{2\sigma_\Delta^2} + \big(\frac{\mu_\Delta}{\sigma_\Delta^2} - 1\big)\Delta - \frac{\mu_\Delta^2}{2\sigma_\Delta^2} \Big)\, d\Delta \\ &\approx f_0 t \cdot \exp\big(-\mu_\Delta + \frac{\sigma_\Delta^2}{2} \big) \end{aligned} \tag{6.13}$$

For low switching probability or short times, Eq. 6.12 is well fitted by Eq. 6.13 (Fig. 6.5(a,b)). In that regime, a full array can be characterized by a single Δ_{eff}-value [117]. On a log-log plot, the last approximation in Eq. 6.13 depends linearly on $\ln t$, with a slope 1:

$$\ln F = 1 \cdot \ln t - \mu_\Delta + \frac{\sigma_\Delta^2}{2}\,. \tag{6.14}$$

However, this approximation is only valid in a limited range of the parameter space. The validity of the parameter space is estimated using the slope $ln(F)/ln(t)$, which is close to 1 in the linear approximation and decreases rapidly when non-linearity becomes more pronounced. In Fig. 6.5(c), the probability below which the linear slope is higher than 0.95, is shown. For narrow distributions, $\sigma_\Delta/\Delta < 5\%$, probabilities smaller than 10^{-2} are already sufficient to be in the linear regime. Whereas, for wider distributions, $\sigma_\Delta/\Delta > 10\%$, the linear approximation only applies for probabilities below 10^{-6}, making the method less practical. Generally, σ-values closer to 10 % are found for STT-MRAM [116, 117]. In summary, deviations from the linear approximation start occurring already at lower probabilities for higher Δ and wider distributions, i.e. large σ_Δ (Fig. 6.5(a,b)).

Note that the authors in [117, 121] also include bi-directional switching, meaning that besides switching from "1" to "0", bits can also switch back from "0" to "1". In that case the switching probability without Δ-distributions becomes [121]:

$$F(t) = 1 - \frac{e^{-\Delta_0} + e^{-\Delta_1} \cdot \exp\big(-t \cdot f_0 (e^{-\Delta_0} + e^{-\Delta_1}) \big)}{e^{-\Delta_0} + e^{-\Delta_1}}\,, \tag{6.15}$$

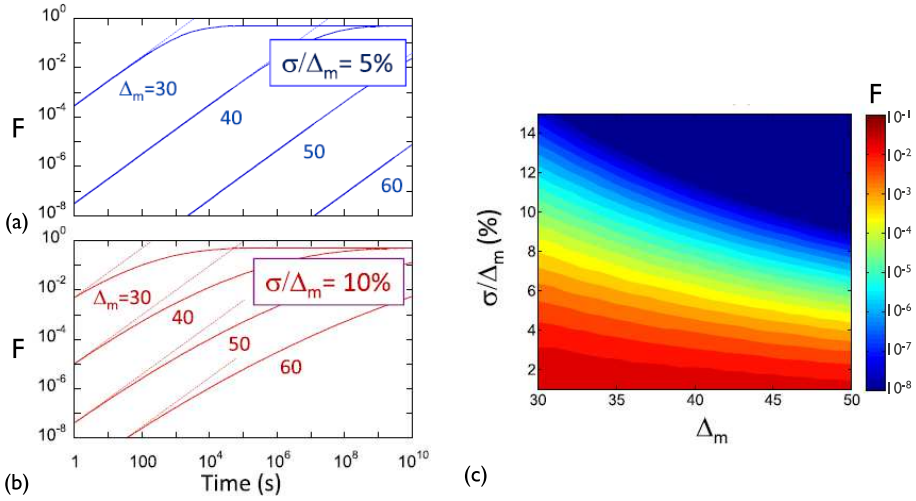

Figure 6.5: *Failure probability calculated for normal distributions of* Δ *having mean* Δ_m*, i.e.* μ_Δ*, and fixed relative standard deviation of 5 % (a) and 10 % (b). The dotted lines show the regime for short time, given by Eq. 6.13. note that the authors take into account bi-directional switching (*$\Delta_0 = \Delta_1 = \mu_\Delta$ *and* $F(t \to \infty) = 0.5$*)(c) Maximum failure probability below which the linear approximation applies (*$ln(F)/ln(t) > 0.95$*), as a function of the mean and relative standard deviation of* Δ*. Modified from [117].*

with Δ_0 and Δ_1 the thermal stability of bit state "0" and "1", respectively. That is also the reason why at $t \to \infty$, F approaches 0.5 for $\Delta_0 = \Delta_1 = \mu_\Delta$ (Fig. 6.5(a,b)).

We have extracted the thermal stability using temperature acceleration in the Mbit array. 60 nm devices are studied in a single MgO stack with a Co/Ni SAF, instead of the conventional used Co/Pt SAF. Characterization parameters of this array are TMR 70 %, RA 7.9 $\Omega\mu m^2$, coercive field $H_c = 58$ mT and a large offset field H_{off} =32 mT, favoring P-state. The baking experiments are performed as follows: (1) 500 000 bits are put in AP-state, (2) the wafer is baked in the oven, (3) the bits are checked for switching and (4) repeat of step (2,3). The results are shown in Fig. 6.6. Due to the high offset field, there is a large difference in thermal stability of the two switching directions ($\Delta_{P-to-AP} >> \Delta_{AP-to-P}$). As such, it is only possible to characterize the unstable switching direction by thermal reversal, i.e. $\Delta_{AP-to-P}$. Since switching from the most stable state (P-state) requires high temperatures, that once the device switches to the unstable state (AP-state), would be immediately cause switching back to the more stable P-state, before we check switching at a specific detection time in

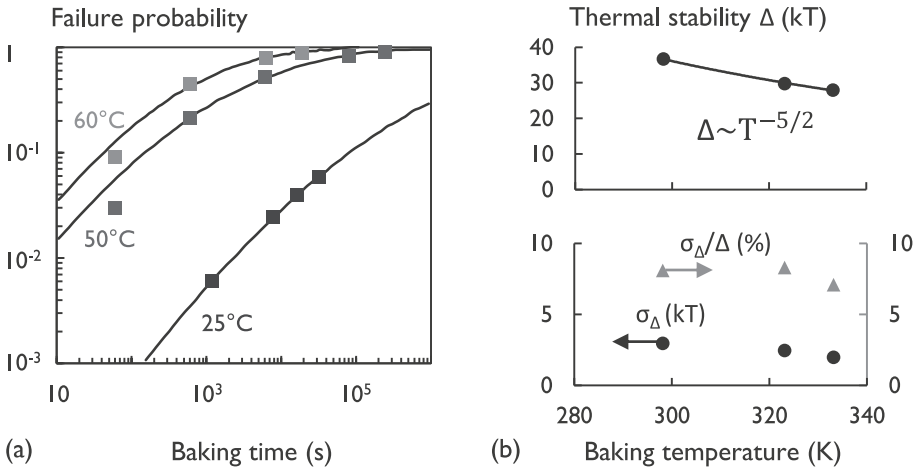

Figure 6.6: *Thermal stability extraction via temperature acceleration in ⌀ 60 nm devices in a single MgO Mbit array with low coercivity $H_c = 58\,mT$ and high offset field ($H_{off} = 30\,mT$). (a) Failure probability in a baking test for the weaker state of 500 000 devices as a function of baking time with baking temperatures of 25, 50, 60° C (solid lines are fits of Eq. 6.12). (b) Fitted thermal stability and standard deviation as function of baking temperature.*

step (3).

Due to the low coercivity and high offset field, the devices could already switch significantly even at room temperature, resulting in a low $\Delta \approx 30kT$. Equation 6.12 fits the data nicely (the linear approximation could not be used), and the resulting $\frac{\sigma_\Delta}{\Delta} \approx 8\,\%$ (Fig. 6.6(b) bottom plot). The deviation of the 2 datapoints at 60 s (for 50 and 60° C) is most probably because the actual baking time/temperature cannot be accurately controlled at these "short" times. Note that in this measurement all bits will switch to the most stable state ($F(t \to \infty) = 1$).

Following Eq. 6.9 and 6.13, a sole $1/T$ dependence of Δ is expected, however, a faster variation is observed (Fig. 6.6(b)), following a power-law with exponent -5/2 ($\Delta \sim T^{-\frac{5}{2}}$), this trend corresponds with the trends observed in [117]. These results indicate that the energy barrier itself depends on temperature. For example, the temperature dependence of the effective anisotropy energy [1]. In addition, we find that the stray field also depends on the temperature [132]. As a result, using high temperature to accelerate the switching cannot be modeled by a simple Néel-Brown model. The temperature dependence of the

energy barrier needs to be correctly modeled to achieve accurate extrapolations back to the reference temperature. Thermal reversal can only be experimentally observed in a limited range of temperatures. Therefore in [46], the authors study the temperature dependence of the anisotropy field and energy barrier using *magnetic field* acceleration in a temperature range from 20 to 400 K. A similar study should be performed on STT-MRAM devices in order to incorporate the temperature dependencies into the thermal reversal switching model.

6.4.2 Magnetic field acceleration

The main advantage of magnetic field acceleration is that no additional temperature effects take place, like self-heating or the high external temperature necessary to cause switching when using temperature acceleration (Sec. 6.4.1).

There are three different experimental methods for studying the stochastic switching process, accelerated by a magnetic field, namely, waiting time, ramped switching field and telegraph noise measurements (Fig. 6.7).

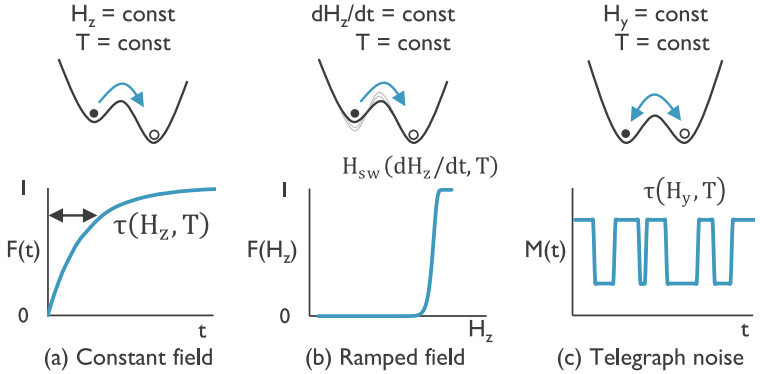

Figure 6.7: *Schematic of three methods for studying the jump across an energy barrier: constant field, ramped field and telegraph noise measurements. (a,c) give direct access to the switching probability in time $F(t) = 1 - P(t)$, whereas ramped switching field measurement (b) to the switching probability as a function of switching field $F(H_z)$, but can be correlated with time via the applied ramp-rate. Modified from [125].*

Constant field measurements

In constant field measurements, a constant magnetic field is applied and the time until switching is monitored. In case of single devices, this methods is repeated multiple times to obtain the cumulative distribution function (CDF) $F(t)$. From this CDF the probability that the magnetization has switched after a time t is fitted (Fig. 6.7(a)). Performing this method at different fields and temperatures, enables the derivation of the barrier height and the thermal activation energy. However, the time range in which the measurements can be performed is limited. In the short-time (< milliseconds) range, experiments are limited by the inductance of the field coils, and for long-time (minutes, hours) experiments are limited by the stability of the experimental setup and available time. Thus for single device measurements, the waiting time measurements are not feasible. Although, when measuring a large number of devices, it is possible to use this method, following a similar approach as in section 6.4.1 with the temperature acceleration.

We performed large statistical measurements on the Mbit array test vehicle. In a similar way as Eq. 6.12, we define the failure probability as:

$$F_{array}(t, H_z) = 1 - \iint_{-\infty}^{+\infty} e^{\left(\int_0^t -\lambda\, dt\right)} g(\Delta, \mu_\Delta, \sigma_\Delta, H_{sw}, \mu_{H_{sw}}, \sigma_{H_{sw}})\, d\Delta dH_{sw}\,,$$

$$with\, \lambda = f_0 e^{-\frac{E_b(\Delta, H_z, H_{sw})}{k_b T}} \tag{6.16}$$

with g the Gaussian distribution in Δ and H_{sw}. H_{sw} is a device parameter and is not necessary equal for the different devices. Therefore, also for H_{sw}, a Gaussian distribution is assumed. In Fig. 6.8(a), the results are shown for 500 000 devices. The test is composed of accelerating switching with 6 different constant, applied magnetic fields (110, 140, 150, 160, 180 and 200 mT), measured at 8 logarithmically spread inspection times. During inspection, the magnetic field is turned off. At each inspection time the number of switched devices is counted. The number of switched, i.e. failed, devices are plotted on a Weibit scale in Fig. 6.8(a). The solid lines are the fit of the switching distribution (Eq 6.16), considering a Macrospin model (see Sec. 6.5). The fit matches well with the bulk of the array, but not at the tails. At low applied fields (110 mT) more devices have switched than expected from the Gaussian fit, and at high fields (200 mT), less devices have switched than expected from the Gaussian fit. Gaussian sigma-values are found of approximately 10 % and 5 % for σ_Δ and $\sigma_{H_{sw}}$, respectively.

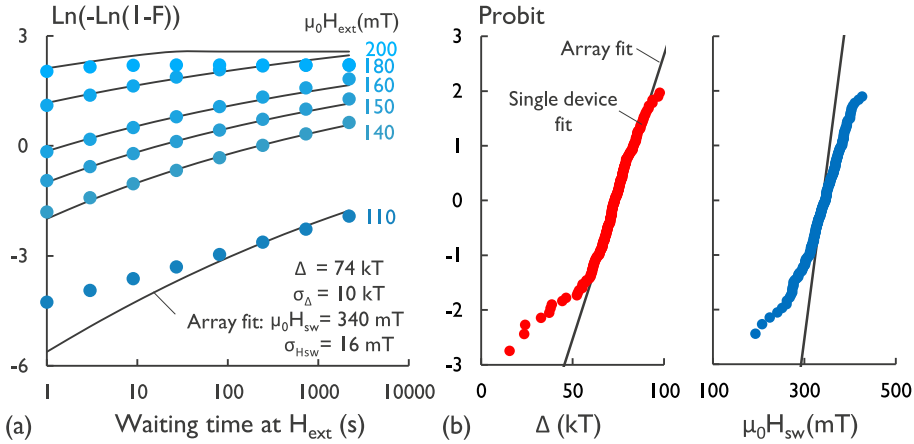

Figure 6.8: *(a) Switching measurements for constant field acceleration in the Mbit array. 500 000 ⌀ 150 nm devices are tested at 6 constant reversal fields ($\mu_0 H_{ext}$ =110, 140, 150, 160, 180 and 200 mT). The fraction of switched devices are plotted on a Weibit scale as a function of the waiting time at H_{ext}. Solid lines are the fit of Eq. 6.16. (b) Δ and H_{sw} distribution from a repeated ramped field test (300 repeats), on 150 randomly chosen devices in the same Mbit array. The median values match well with the array fit, but the tails are deviating.*

In order to explain the deviation from the fit, we will test the validity of using a Gaussian distribution on Δ and H_{sw}. We have tested 150 cells, randomly chosen in the same Mbit array, and performed repeated ramped switching field measurements, see next subsection. Each ramped field cycle is repeated 300 times, for every device. The distributions of the fitted Δ and H_{sw} are shown in Fig. 6.8(b) and compared with the array fit. The median values are well matched, but the tails deviate. We conclude that using a Gaussian distribution to fit the bulk of the array is reasonable, however, to correctly study the effect of the tail bits, single device measurements of these tail bits are necessary.

Ramped magnetic field measurements

For single device studies, it is more convenient to study switching by ramping a magnetic field at a given ramp-rate (RR). The field is increased until the device switches. The certainty of switching, makes this method more suited for automation. Repeating the ramp cycle several hundred times, results in a switching field distribution [Fig. 6.7(b)]. For our setup the ramp-rate can vary between 0.01 and 10 Hz or 6 mT/s and 6 T/s. These ramp-rates are still

slow compared to the switching dynamics of the MTJ, which is in the order of nanoseconds. Therefore, switching still occurs in the thermally activated regime, as in the case of the waiting time measurements.

We developed an experimental setup to measure switching at these different ramp-rates. In our conventional magnetic field ramp setup, the field is increased stepwise, and at each field step, the resistance is measured. In this setup however, we use a waveform to generate a linear magnetic field ramp and detect switching with an oscilloscope.

We will further explain and use this linear field ramp setup. The setup makes use of a 20 GSamples oscilloscope and is depicted in Fig. 6.9(a). In this oscilloscope the switching is detected in channel 1 (C1) and the magnetic field ramp is monitored in channel 2 (C2).
Channel 1: A 10 mV DC bias is applied over the MTJ and a variable resistance R_{div} [Fig. 6.9(a)]. $R_{div} \approx R_{MTJ}$ to optimize the transition signal (P-to-AP or AP-to-P) dumped in the oscilloscope.
Channel 2: The magnetic field is measured via a calibrated Hall sensor, placed on top of the probe card, and dumped in the oscilloscope.
Both the Hall sensor and the transition signal are stored on the oscilloscope to have a very accurate determination of the switching time. The switching time is determined by correlating V_{osc} with a stepped function (V_{osc} is depicted in the bottom plot of Fig. 6.9(b)). The switching field corresponding to this switching time, is then extracted from the magnetic field data sensed by the hall sensor (channel 2). With this setup we can measure very accurately, fast and at low bias, i.e. with negligible self-heating and STT.

Note that (i) the Hall sensor is calibrated at wafer level, such that the measured field, which is above the probecard can be corrected to the field present at wafer level. (ii) A delay of two seconds is introduced, in order to safely use automated triggering and saving of the data on the oscilloscope. (iii) The DUT is biased only with 10 mV, in order to eliminate any self-heating effect.

We have tested the effect of the bias voltage and impact of the magnetic cycle (results are shown in Fig. 6.10). Each cycle we repeated 2000 times. In Fig. 6.10(a) a device of diameter 200 nm and switching direction AP-to-P, is measured for 4 different biases (-200, -25, 25, 200 mV). The applied bias influences the switching distribution for larger bias, but no significant effect below 25 mV is observed (Fig. 6.10(a)). With a negative voltage, the STT tends to stabilize the AP-state and results in higher switching fields, whereas for a positive voltage, the STT aids reversal to P-state, resulting in lower switching fields.

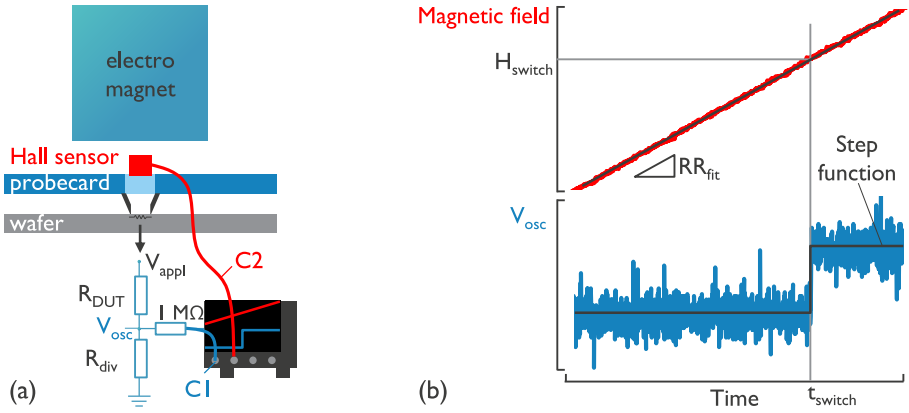

Figure 6.9: *(a) Schematic of the ramped magnetic field switching measurement. 1 Oscilloscope channel is used to trigger the switching time, using a voltage divider between R_{DUT} and R_{div}, and the other channel to monitor the magnetic field, using a calibrated Hall sensor. (b) Extraction of magnetic switching field H_{switch} and switching time t_{switch}. (top graph) magnetic field measured by the Hall sensor, which is fitted to extract the RR_{fit} around H_{switch}. (bottom graph) The oscilloscope voltage V_{osc} is fitted by a step function, in order to determine t_{switch}.*

In Fig. 6.10(b), a ⌀ 150 nm device and switching direction P-to-AP, is measured at 10 mV for 2 ramp-rates and 2 times the 2000 cycles/ramp-rate. We see no effect of degradation of the switching distribution after 2000 cycles.

Telegraph noise measurement

The telegraph noise measurement is based on stochastic fluctuations between the two stable states[24, 55]. This time a magnetic field is applied in the direction of the hard axis, i.e. in plane. This will lower the energy barrier and make stochastic switching between two states possible at a reasonable temperature. The time spent in each state follows the same exponential switching probability law given by Eq. 6.10. Due to the unavailability of an in-plane field, this method is not further studied or used.

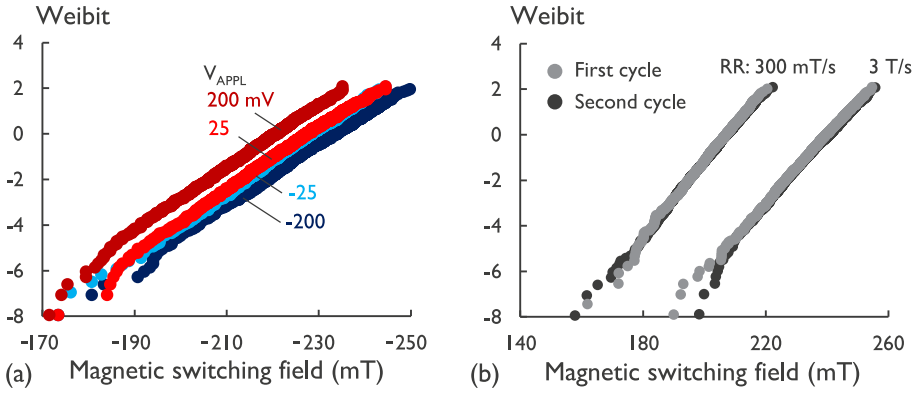

Figure 6.10: *(a) AP-to-P Switching field distributions from 2000 cycles at 4 applied bias voltages (-200, -25, 25, 200 mV) for a ⌀ 200 nm device. STT influences the distribution at higher bias, but no significant effect below 25 mV is observed. Note: the minus sign in front of the magnetic field on the x-axis is to indicate the direction of the field. (b) P-to-AP switching distributions from 2000 cycles at 10 mV and 2 ramp-rates (300 mT/s and 3 T/s). The measurement is repeatable.*

6.4.3 Current acceleration

A last method of acceleration is by current, i.e. the actual operation mechanism of the STT-MRAM. In contrast to the other acceleration methods, current acceleration can be studied on a much shorter time scale. Namely, switching can be induced sub-ns, but at high current densities. In this short time range, however, a different switching regime is observed, a precessional regime compared to the thermally activated regime. As such, three different physical regimes are distinguished for spin-torque-induced switching [71] (Fig. 6.11(a)):
(i) Short-time ballistic limit below a few nanoseconds, i.e. precessional regime
(ii) Long-time limit, i.e. thermal regime
(iii) Cross-over regime
These regimes have been studied for perpendicular spin-valves [119, 71]. In contrast to an MTJ, a spin-valve has no tunnel barrier, and is fully metallic. The switching mechanisms are similar, but a spin-valve has low MR $\approx 10\,\%$. In general, for faster switching, i.e. decreasing pulse widths, an increasing amount of current is necessary to induce switching.

In STT-MRAM, however, the sub-ns switching regime or precessional switching, is difficult to measure, due to the very high currents required. These high currents can result in breakdown of the MgO tunnel barrier, which is not the case for an all metal spin-valve. Most of the reports of sub-ns switching in MTJs

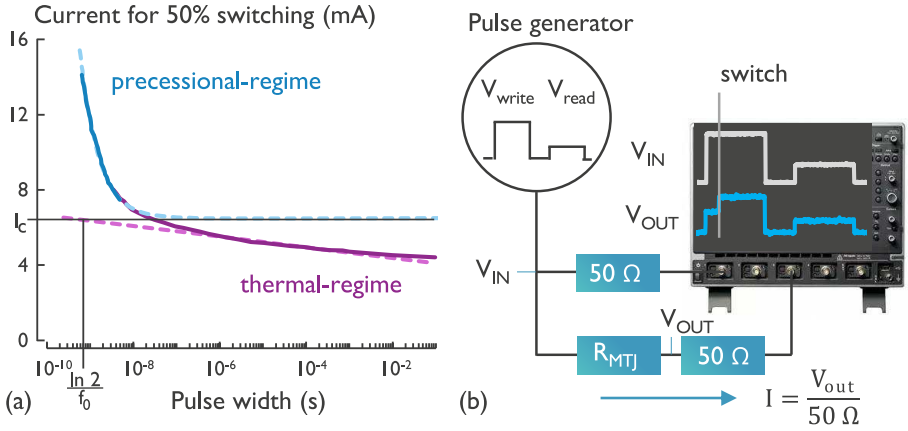

Figure 6.11: *(a) Pulse amplitude for 50 % switching as a function of pulse width from 300 ps to 1 s in a 100 x 100 nm^2 spin-valve structure. Dotted line are fits of the precessional and thermally activated switching regime. The critical current I_c intersects with the thermal-regime at a pulse width of $\frac{\ln 2}{f_0}$ (Eq. 6.18). Replotted from [71]. (b) Current accelerated experimental setup.*

are with a polarizer magnet, which destabilizes the FL, such that switching occurs already at low currents [27, 67]. Other results are on a different type of device, which switches with spin-orbit torque [47], we will not further elaborate on this type of devices. The few recent reports on ultra fast switching in MTJs are from a state-of-the-art technology [53] or for very small cells (<20 nm) [93], which are more resilient to breakdown, because of the small area. Data retention is defined at zero current, however, therefore the regime which is used for current acceleration is of secondary importance. In this thesis, we only investigate the long-time limit switching with pulse widths $> 100\,ns$.

The current switching distribution is then well described using a critical switching current I_c, in the thermal-regime:

$$F_{array}(t,I) = 1 - \iint_{-\infty}^{+\infty} e^{\left(\int_0^t -\lambda\, dt\right)} g(\Delta, \mu_\Delta, \sigma_\Delta, I_c, \mu_{I_c}, \sigma_{I_c})\, d\Delta dI_c\,, \tag{6.17}$$

$$with\, \lambda = f_0 e^{-\frac{E_b(\Delta,I,I_c)}{k_b T}}$$

with g the Gaussian distribution in Δ and I_c. In the thermal-regime the required current to switch, can be smaller than I_c, or $I_{sw} < I_c$, because of the statistics, whereas in the precessional regime, $I_{sw} > I_c$, because of the physics. At the

critical current, the energy barrier is reduced to zero, hence the switching distribution of a single device in the thermal regime simplifies to ($E_b = 0$):

$$\ln - \ln\big(1 - F(I = I_c)\big) = \ln f_0 + \ln t - 0 \tag{6.18}$$

In the thermal-regime I_c is the switching current at a pulse width of $t = \frac{\ln 2}{f_0}$, with 50 % switching $F = 0.5$ (see Fig. 6.11(a)), where f_0 is around 1 GHz. $E_b(\Delta, I, I_c)$ depends on the used switching model (see Sec. 6.5) and is in acceleration conditions, i.e. $I \approx I_{sw}$, around zero. The extracted δ, however, which is defined as the energy barrier at $I = 0$, largely depends on the used switching model.

There are multiple ways one can measure the current accelerated distribution (Eq. 6.17), similar to magnetic field acceleration, namely, by monitoring the waiting time at a fixed current amplitude, by changing the current amplitude for various pulse widths or by ramping a current. In the following we will develop the method where the current pulse amplitude is incremented until switching occurs and this measurement is performed using multiple pulse widths.

We use the experimental setup, shown in Fig. 6.11(b), to measure current-induced switching in MTJs. The setup makes use of a pulse generator and an oscilloscope. The 50 Ω-terminated pulse generator applies a write and a read pulse. The read pulse has a fixed bias of 100 mV, whereas the write pulse can change voltage amplitude. One oscilloscope channel measures the input voltage V_{IN}, in parallel with the MTJ and channel 2. Channel 2 is then in series with the MTJ and measures the voltage at the output V_{OUT} of the MTJ. The MTJ stress voltage and current passing through the MTJ are then given by:

$$V_{MTJ} = V_{IN} - V_{OUT} \tag{6.19}$$

$$I_{MTJ} = \frac{V_{OUT}}{50\,\Omega} \tag{6.20}$$

When the device switches, it is visible in the write pulse, and will be verified in the successive read pulse.

In the thermal regime, the shortest time we can obtain for current switching (t > 100 ns) is much smaller than the shortest time for magnetic field switching, which is in the order of milliseconds in a typical setup. However, the current can significantly heat the cells during switching, see Sec. 6.8. A high STT, together with the temperature increase due to self-heating, can destabilize the RL, causing back-hopping [65]. This back-hopping is an important reliability

concern when deep error-rates are necessary, i.e. high currents need to be applied to ensure writing [76, 54]. We do not further discuss the effect of back-hopping in this thesis.

6.4.4 Summary

Different acceleration methods can be used to study switching in STT-MRAM. Temperature, magnetic field and current are used to reduce the energy barrier and accelerate the switching.

Temperature acceleration is best performed using large sample sizes, the model is assumed to be the simple Néel-Brown model, but the energy barrier and other material parameters, like the offset field, are also temperature dependent and are not included in the models used throughout literature.

Magnetic field acceleration does not introduce any temperature effects and can be performed on single devices as well as Mbit arrays. The Δ extraction, however, largely depends on the switching model, see Sec. 6.5.

Current acceleration is well suited to study single devices. The time range in which switching can be studied is the largest of the acceleration methods considering our experimental setups. The Δ extraction, however, is influenced by self-heating, since large currents are required to switch the MTJs.

Considering the experimentally accessible time range of the measurements. The upper limit is similar for the three acceleration methods and mostly determined by the stability of the experimental setup and available time (hours). The lower limit, however, is different for each acceleration method:

Temperature In the case the thermal bake and resistance verification are in separated setups (oven + measurement setup), an accurate temperature control will only be obtained starting from several minutes.

Magnetic field The short-time range is limited to milliseconds, because of the inductance of the magnet.

Current The largest time range can be measured with current acceleration. In case we only consider the thermally activated regime, this starts from times $> 100\,\text{ns}$.

6.5 Macrospin and domain wall switching models

As discussed in the previous sections, the energy barrier depends on different acceleration parameters. The dependencies on these parameters are modeled into a function $E_b(\Delta, I, H, T)$, and this function is used to extract the thermal stability from the switching distribution (Eqs. 6.12, 6.16, 6.17).

In this section, we will derive these switching distribution functions for the two most common models of switching: uniform switching, i.e. Macrospin (MS), and domain wall (DW) mitigated switching.

6.5.1 Macrospin model

In the Macrospin approximation the magnetization of the free layer can be approximated to one magnetic spin. This spin will uniformly rotate and the switching dynamics is simplified to the dynamics of this one spin. In general, the Macrospin model is only valid for sizes below a critical diameter, which depends on material parameters like exchange stiffness and effective anisotropy. Although around reversal conditions, this model perfectly describes the switching distribution, and because of its simplicity it is frequently used. Later, in Sec. 6.7, we will show that although the reversal data is well described, the extracted Δ is not comparable when using a Macrospin or a domain wall model. First, we derive the Macrospin model for magnetic field and current switching.

Magnetic field switching

For a uniform rotation of the magnetization, Stoner and Wohlfarth have developed a model in 1947 to describe the reversal mechanism at 0 K [104]. It is based on a Stoner-Wohlfarth particle, i.e. a uniformly magnetized ellipsoid with uniaxial anisotropy. In this particle, the exchange energy forces all spins parallel to each other. Therefore, the magnetization can be modeled by the switching of only one all-encompassing spin. In that case, the exchange energy has no role in the energy minimization (Eq. 6.1). In addition, the now simplified demagnetization contribution is included into the effective anisotropy energy term. Consequently, there is only competition from the effective anisotropy energy and the Zeeman energy between the external field and the magnetization. In the simplest case of uniaxial anisotropy, the energy is given by [125]:

$$E_{MS} = K_{eff} V sin^2\phi - \mu_0 M_s V H cos(\phi - \theta)\,, \tag{6.21}$$

where $K_{eff}V$ is the effective uniaxial anisotropy energy, V is the volume of the particle, M_s is the spontaneous magnetization, H is the magnitude of the applied field, and ϕ and θ are the angles of the magnetization and the applied field, wrt the easy axis of magnetization (Fig. 6.12(a)). Eq. 6.21 leads to two minima separated by an energy barrier. At zero field, these two minima occur at an angle $\phi = 0$ or π, which correspond to up or down magnetization. Applying a reversal field, changes the energy landscape locally around the minimum, until magnetization reversal. This reversal is defined by the minimum magnetic field at which the energy barrier vanishes, i.e. at $\partial E/\partial\phi = \partial^2 E/\partial\phi^2 = 0$.

The energy minimization can as such explain the origin of the hysteresis curve. For a uniform magnetization and an external field applied in parallel with the easy-axis, i.e. $\theta = 0$, the resulting energy diagrams are plotted in Fig. 6.12(b), with corresponding hysteresis curve on Fig. 6.12(c). When moving across the hysteresis curve (point 1 to 5), the energy landscape changes drastically:
(1) At zero applied field, 2 local minima are observed separated by an energy barrier.
(2) The reversal field, destabilizes the present magnetization. The energy barrier is reduced, but the present state remains a local minimum.
(3) Magnetic reversal occurs. The energy barrier is reduced to zero and a new local minimum is found in the reversal state.
(4) The reversed state is further stabilized.
(5) At zero applied field, a new magnetic state is obtained and stable.

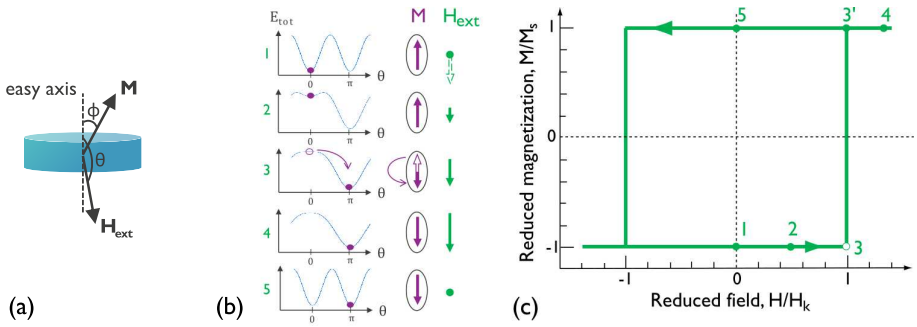

Figure 6.12: *(a) Schematic of Macrospin switching, where the magnetization is represented with one vector with magnitude M and angle ϕ with the easy-axis. The external field is applied with an angle θ. Schematics of the energy landscape (b) and magnetization curves (c) for the Stoner-Wohlfarth model for a field applied parallel to the easy axis. The energy landscape, and orientation and magnitude of the uniform magnetization M and external field H_{ext} (b), are illustrated for 5 different points on the hysteresis curve (c). Replotted from [23].*

When the external field is applied under a non-zero angle with the easy-axis, the hysteresis curve changes. The coercivity decreases down to zero for the case the external field is applied along the hard-axis, in-plane axis with $\theta = 90°$(Fig. 6.13(a)).

In later studies from Thiaville et al., the Stoner-Wohlfarth model is generalized, and the anisotropy energy is extended with shape, magnetocrystalline, surface and magneto-elastic anisotropy energy [114, 115]. Thiaville's calculation also predicts the field dependence of the energy barrier height $E_{b,MS}$ close to switching to be:

$$E_{b,MS} = \Delta_{MS}\left(1 - \frac{H_{ext}}{H_{sw}}\right)^{\eta}, \tag{6.22}$$

with $\Delta_{MS} = \frac{K_{eff}V}{2k_bT}$ and $\eta = 2$ for fields applied along the easy axis, i.e. $\theta = 0$. Only for this case H_{sw} matches the anisotropy field H_k:

$$H_{sw}(\theta = 0) = H_k = \frac{2K}{\mu_0 M_s} \tag{6.23}$$

Note that in reality, the state will already reverse before reaching H_{sw}, due to thermal effects, not taken into account by the Stoner-Wohlfarth model, which is modeled at 0 K.

In Fig. 6.13(b) the resulting exponent η from a numerical calculation of the energy barrier approximated by Eq. 6.22 is shown. η quickly decreases from 2 to 1.5 when the angle of the external field is not perfectly parallel with the

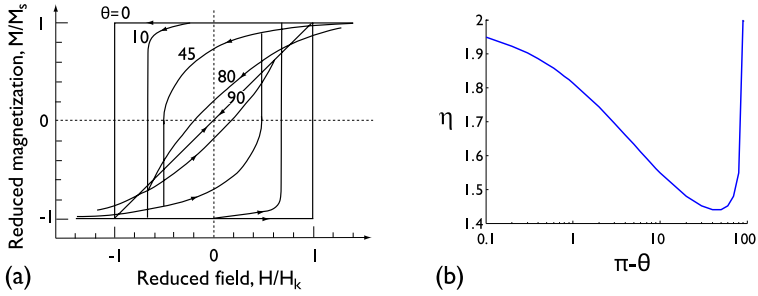

Figure 6.13: *(a) Magnetization curves for the Stoner-Wohlfarth model for various angles θ between the external field and the easy-axis. Replotted from [23]. (b) Numerical calculation of the exponent η in the field dependence of the energy barrier in Eq. 6.22. η changes rapidly as a function of the angle of the applied field H_{ext} with the easy-axis.*

easy-axis (1.5 already at $\theta = 10\,°$). That is why there are also reports on the Macrospin model using an exponent $\frac{3}{2}$ [46].

Current switching

Including current in the Macrospin model, results in an energy barrier derived from the LLG-equation with STT:

$$E_{b,MS} \propto \left(1 - \frac{I}{I_c}\right)^b, \tag{6.24}$$

where the exponent b varies between 1 and 2. The difference in b results from how the spin-torque is treated in the LLG-equation (Eq. 6.5). In [66, 109], the authors included the spin-torque in the damping term as an effective temperature argument. Whereas in [113, 119, 90], the spin-torque is an extra term in the effective magnetic field. It is difficult, however, to experimentally prove which approach is correct (b = 1 or b = 2). Since in the limited range of switching current amplitudes, plotting the median switching current $I(F = 0.5)$, will appear linear as a function of the logarithm of time, i.e. a Taylor expansion of $(1 - I/I_c)^b$ is linear for small variations in I relative to Ic for both b = 1 and b = 2. Extending the current range to $I \to 0$, is extremely time consuming. In general, the fits of a linear dependence will underestimate the thermal stability. This once again illustrates the impact the switching model has on the Δ-extraction.

The Macrospin model used to extract the thermal stability data in this thesis is then defined as:

$$E_{b,MS} = \Delta\left(1 - \frac{H}{H_{sw}}\right)^2\left(1 - \frac{I}{I_c}\right)^2. \tag{6.25}$$

The advantage of this model is that it is simple and analytic. The disadvantage is that it oversimplifies certain matters:

Magnetic field acceleration exponent $\eta = 2$ is defined for an applied field collinear with the easy axis. Deviations from this collinearity reduces η to approximately 1.5.

Current acceleration exponent $b = 2$ assimilate the spin-torque as an extra term in the effective magnetic field. On the other hand including the spin-torque in the damping term as an effective temperature argument would result in $b = 1$.

A Macrospin behavior is assumed, i.e. uniform magnetization configuration with all elementary volumes aligned, and predicts coherent rotation of all the spins during magnetization reversal. This approximation is only valid in the case of strong exchange energy, that forces the uniform magnetization and does not allow the formation of domain walls and multi-domain configurations.

6.5.2 Domain wall model

For larger MTJ dimensions, the assumption that all magnetizations are aligned by the exchange energy, is no longer valid. It will require less energy to reverse the state by nucleating a domain wall and propagating the domain wall through the MTJ. The transition point for a conventional MTJ is found at a critical diameter d_c around 20-30 nm [17]. As can be seen in Fig. 6.14(a), the energy for Macrospin, increases quadratically with MTJ diameter, whereas the energy for domain wall switching only linearly.

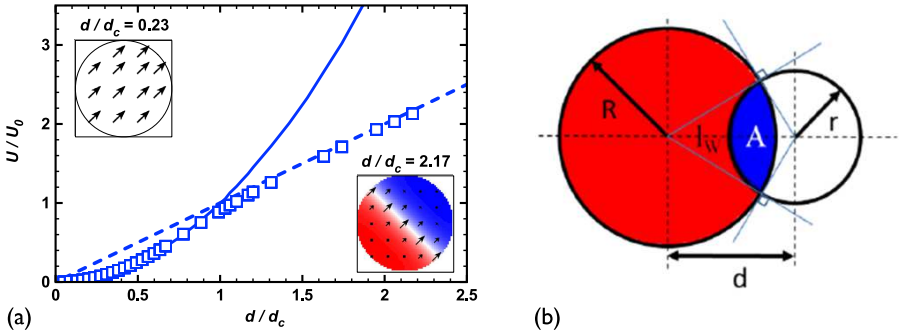

Figure 6.14: *(a) Rescaled energy barrier U/U_0 for Macrospin reversal (solid line) and for a domain state (dashed line) as a function of the normalized diameter d/d_c. The blue open squares represent energy barriers obtained from micromagnetic simulations using the string method. The insets show the micromagnetic configuration of the transition states for the coherent reversal $d/d_c = 0.23$ and domain wall $d/d_c = 2.17$ cases. Replotted from [17]. (b) Schematic of the domain wall reversal model, represented by two orthogonal intersecting circles [116].*

Magnetic field switching

In the case of a domain wall-induced switching, the system becomes more complex. Now, there is a reduction to the energy barrier due to the Zeeman

energy from the already reversed volume of the cell, but also an energy increase by the domain wall energy [17]. The energy for a collinear applied field in a circular device is then given by:

$$E_{DW} = \left(\pi R^2 - A\right) \cdot \mu_0 M_s H t - A \cdot \mu_0 M_s H t + \sigma_w l_w \cdot t\,. \tag{6.26}$$

Here R is the device radius, t the FL thickness, A the area of the already reversed volume, σ_w the domain wall energy and l_w the domain wall length. Micromagnetic simulations show that the domain wall is defined by the intersection of the circular MTJ with a fictive circle of radius r at distance d from the center of the device, see Fig. 6.14(b) [17, 116]. Furthermore, considering the energy landscape, the energy critical point to overcome corresponds to the case of orthogonal intersecting circles for which $d^2 = R^2 + r^2$. The authors use this condition, such that A and l_w can be simplified [116]:

$$A = r\left(\frac{\pi}{2} r - R\right) + \left(R^2 - r^2\right) \cdot \tan^{-1}\left(\frac{r}{R}\right) \tag{6.27}$$

$$l_w = \pi \cdot r - 2 \cdot r \cdot t \cdot \tan^{-1}\left(\frac{r}{R}\right) \tag{6.28}$$

Calculating the energy barrier follows a similar approach as for the Stoner-Wohlfarth model. The energy barrier is determined by the difference of a fully non-switched device ($r \to 0$) and the saddle point determined for $r = r_0$, such that $\partial E_{DW}/\partial r = \partial^2 E_{DW}/\partial r^2 = 0$. It follows that $r_0 = \frac{\sigma_w}{2\mu_0 M_s H}$ and the energy barrier:

$$\begin{aligned} E_{b,DW} &= E_{DW}(r_0) - E_{DW}(0) \\ &= -2 \cdot A(r_0) \cdot \mu_0 M_s H t + \sigma_w l_w(r_0) t \\ &= \sigma_w R t + \frac{\sigma_w^2 t}{2\mu_0 M_s H}\left(\frac{\pi}{2} - \tan^{-1}\left(\frac{\sigma_w}{2\mu_0 M_s H R}\right)\right) \\ &\quad - 2\mu_0 M_s H t R^2 \tan^{-1}\left(\frac{\sigma_w}{2\mu_0 M_s H R}\right) \end{aligned} \tag{6.29}$$

The energy barrier only approaches zero at infinite magnetic field:

$$E_{b,DW}(H \to \infty) = \frac{\pi \sigma_w^2 t}{4\mu_0 M_s H} \tag{6.30}$$

That the barrier only approaches zero asymptotically is in contrast with Macrospin model, where $E_{b,MS} = 0$ for a magnetic field equal to H_{sw} (Eq. 6.25). This is because the domain wall in this model (Eq. 6.29) has no width. In reality, the domain wall has a finite width and thus within the domain wall

the magnetization is partially in the plane and therefore has a smaller Zeeman contribution than from the fully reversed and non-reversed domains, since they are parallel with the external field. To include a finite domain wall width, the authors define a length 2δ within which the Zeeman contribution is assumed zero. With this approximation the magnetic energy becomes [116]:

$$E_{DW} = \left(\pi R^2 - A(r+\delta)\right) \cdot \mu_0 M_s H t - A(r-\delta) \cdot \mu_0 M_s H t + \sigma_w l_w(r) t\,, \quad (6.31)$$

Unfortunately, due to the introduction of a domain wall width, there is no analytical solution for the energy barrier. It can be expressed as a function of the exchange constant A_{ex}, the effective anisotropy K_{eff} and the applied field H as follows:

$$\begin{aligned} \delta &= 4\sqrt{\frac{A_{ex}}{K_{eff}}} \\ \sigma_w &= 4\sqrt{A_{ex} K_{eff}} \\ E_{b,DW}(A_{ex},\, K_{eff},\, H) &= E_{DW}(r_0) - E_{DW}(0) \end{aligned} \quad (6.32)$$

where δ is determined by a Bloch wall [23]. To calculate the energy barrier when A_{ex}, K_{eff} and H are given, r_0 is determined by numerically solving the roots of Eq. 6.31 ($\frac{\partial E_{DW}}{\partial r} = 0$). $E_{DW}(0)$ is again the energy calculated for a fully non-switched device, i.e. no domain wall present.

In Fig. 6.15(a), we compare the energy barrier for a simulated $K_{eff} = 270\,\mathrm{kJ/m^3}$, $M_s = 1100\,\mathrm{kA/m}$, $A_{ex} = 20\,\mathrm{pJ/m}$, $t = 2.6\,\mathrm{nm}$ and $R = 60\,\mathrm{nm}$ (based on thin film data and literature[36]) for the Macrospin and domain wall model (Eqs. 6.25,6.29 and 6.32). For the domain wall model, the energy barrier drops to zero at finite magnetic fields when the domain wall width is included, and drops to zero only at infinity without a domain wall width (w δ and w/o δ). Furthermore, the domain wall model has a lower energy barrier at zero field than the Macrospin model, note that both models are simulated with the same effective anisotropy. In addition, the energy barrier of the domain wall model (w δ) drops to zero at lower magnetic fields as in the case of a Macrospin model.

Of course, in reality the device parameters are not known, H_{sw} and Δ for the Macrospin model, and K_{eff} and A_{ex} for the domain wall model. The parameters are extracted by fitting the measured switching distributions with a maximum likelihood fit (see next section). As discussed in the section on acceleration measurements (Sec. 6.4), the switching is only obtained in a limited measurement range. For a standard 2 ramp-rate field sweep on 300 switching

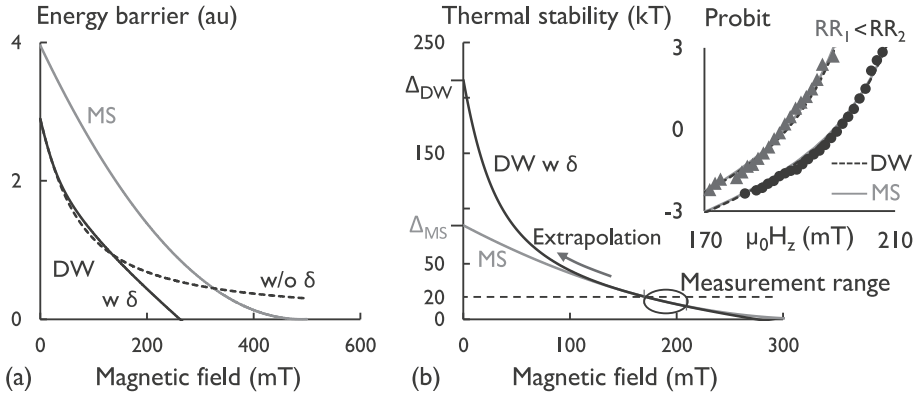

Figure 6.15: *The energy barrier calculated for Macrospin (MS) and domain wall (DW) model. (a) Calculated energy barrier for MS and DW model, with (w) and without (w/o) domain wall width δ included. used parameters:* $K_{eff} = 270\,kJ/m^3$, $M_s = 1100\,kA/m$, $H_{sw} = H_k = \frac{2K_{eff}}{\mu_0 M_s}$, $\Delta = \frac{K_{eff}V}{2}$, $A_{ex} = 20\,pJ/m$, $t = 2.6\,nm$ *and* $R = 60\,nm$. *(b) Fitted thermal stability for DW and MS model. Within the given measurement range, no difference between the models is observed, however in the extrapolation to zero-field condition,* $E_{b,DW} > E_{b,MS}$. *(inset) switching distributions obtained for 2 ramp-rates (1 Hz, 10 Hz) with 300 cycles, DW and MS-fit cannot be distinguished in the measurement range (solid lines).*

cycles per ramp-rate, this limited measurement range is depicted in the inset of Fig. 6.15(b). In this measurement range the energy barrier is estimated to be between 10 and $20\,k_bT$, such that thermal activation over the lowered energy barrier is possible. It is from this measurement range that the thermal stability is extrapolated to zero field, following a Macrospin or domain wall model. Both models fit the switching distribution perfectly, but there is a large difference in extracted thermal stability (see Fig. 6.15(b)).

Current switching

There is no equivalent simple domain wall model for current-induced switching. A lot of literature exists on how the domain wall dynamics will proceed at switching conditions [35]. However, there are limited studies, up to our knowledge, on how the energy barrier evolves going to zero current conditions, and preferably implemented in a simplified model. As a starting point, the conceptual analysis of Munira and Visscher, can be used, where the Macrospin model is followed up to a critical instability point (extra fitting parameter) from

this point on the energy barrier depends linearly on the current [79].

In summary, the simplified models, Macrospin model and domain wall model, elaborated in this section are indistinguishable given a conventional measurement range. However, the extrapolated thermal stabilities, i.e. extrapolated to zero-field, do not match. Extending this measurement range to obtain switching values at low fields, i.e. $\Delta > 20\,\mathrm{kT}$, is very time-consuming, because of the double exponential dependence of F on Δ (Eq. 6.10). In section 6.7, we will further compare the Macrospin and domain wall model, but first we discuss how to fit the switching distribution using a maximum likelihood method.

6.6 Correlation between parameters, error analysis

Since the retention time is predicted by a double exponential of Δ (Eq. 6.10), it is very important to make the error on the fitted Δ_{fit} as small as possible. Therefore, in this section the fitting of the distribution function is discussed for the case of magnetic field acceleration. We find that the model parameters are highly correlated. As a result, large statistics are necessary to decrease the error on the extracted thermal stability. Following equation 6.8, the cumulative switching distribution F for a Macrospin model is given by

$$\begin{aligned} F_{MS,i} &= 1 - \exp\left(\int_0^{t_i} -f_0 exp\Big(-\Delta\Big(1 - \frac{Rt}{H_k}\Big)^2\Big)\, dt \right), \\ &= 1 - exp\Big(\frac{f_0 H_k}{2R}\sqrt{\frac{\pi}{\Delta}}\Big[erf\Big(\sqrt{\Delta}\Big[1 - \frac{Rt_i}{H_k}\Big]\Big) - erf\Big(\sqrt{\Delta}\Big)\Big]\Big), \end{aligned} \tag{6.33}$$

with R the magnetic field ramp-rate ($H_i = Rt_i$). For the domain wall model this switching distribution function has to be derived numerically using Eq. 6.32 and Eq. 6.8 to get F_{DW}:

$$F_{DW,i} = 1 - \exp\left(\int_0^{t_i} -f_0 exp\Big(-\frac{E_{b,DW}\left(A_{ex},\, K_{eff},\, Rt\right)}{kT}\Big)\, dt \right). \tag{6.34}$$

The probability density function is then the derivative of F. The switching distribution is measured by repeating switching events. The MTJs are switched with a magnetic field ramp between -300 mT and 300 mT. The switching data

are fitted with a maximum likelihood method (ML) [83] to extract the model parameters H_k and Δ for MS and A_{ex} and K_{eff} for DW. The likelihood function is the product of the probability density functions f_i for all measured switching times t_i:

$$\Lambda = \prod_i^N f(t_i) \tag{6.35}$$

The maximum of the ML, results in a best fit for the full measured dataset, i.e. measured switching times, for the given model. In Fig. 6.16(a),(b), a contour plot of the logarithm of the likelihood function is shown for MS and DW, respectively. The fitted parameters, Δ and H_k for MS and A_{ex} and K_{eff} for DW, are highly correlated [5]. This correlation expresses itself by the elliptic banana-shape contours. On these contours, large ranges in the parameter space describe the fitted data with an equal likelihood. If there was no correlation, the contours would be circles.

To estimate the error of the fit, we use the variance of an ML estimator ($\hat{\theta}_{ML}$). In the MS case, $\hat{\theta}_{ML}$ can be $\hat{\Delta}$ or $\hat{H}_k$. The variance is calculated by the inverse of the Fisher information matrix $I(\theta)$, where this matrix is the negative of the expected value of the Hessian matrix $H(\theta)$ of the maximum likelihood function

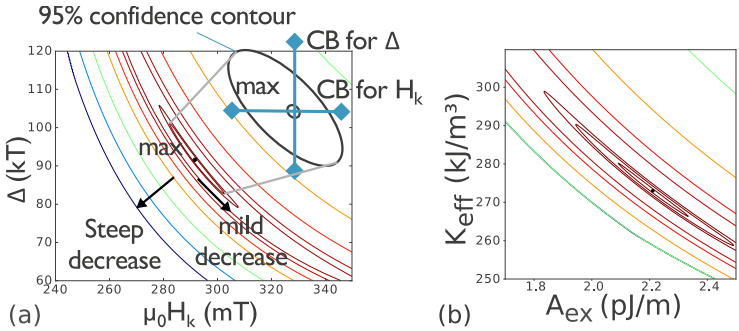

Figure 6.16: *The maximum likelihood function shows highly correlated parameters, which results in large errors. Inset: 95 % confidence contour and extracted confidence bounds in Δ and H_k. (a) Macrospin model fits Δ and H_k, whereas (b) the domain wall model fits K_{eff} and A_{ex}.*

around the maximum (in $\hat{\Delta}$, $\hat{H_k}$)[44]:

$$\begin{aligned} var(\theta) &= \big[I(\theta)\big]^{-1} \\ &= \big(-E\big[H(\theta)\big]\big)^{-1} \\ &= \left(-E\left[\frac{\partial^2 \ln \Lambda(\theta)}{\partial\theta\partial\theta'}\right]\right)^{-1}. \end{aligned} \tag{6.36}$$

In short, the likelihood function around the maximum is approximated by a multivariate normal distribution. The steeper the likelihood function around the maximum, the higher the accuracy of the fit. This way a 95 % elliptical confidence contour can be determined [see inset Fig. 6.16(a)], which correlates well with the actual contour of the likelihood function. The 95 % confidence bound (CB) on the model parameters is then determined by the square root of the diagonal elements of the variance and a factor taking into account the 95 % confidence and the number of fitted parameters:

$$CB_{95\%,ML} = \hat{\theta} \pm \sqrt{\chi^2(95\,\%, 2) \times var(\hat{\theta})}\,, \tag{6.37}$$

with χ^2 the chi-square distribution.

To further study the error of the extracted thermal stability, we have performed Monte-Carlo (MC) simulations. The MS-model is used, but a similar approach is possible for a DW-model. We study the effect of the dataset size on the error of the extracted parameters as follows:
(1) Generate switching dataset consisting of (1 to 4) ramp-rates and (10 to 10 000) generated datapoints, i.e. switching cycles.
(2) Fit each dataset for Δ_{fit} and $H_{k,fit}$.
(3) Repeat step (1) and (2) 500 times, to obtain a distribution of Δ_{fit} and $H_{k,fit}$.
(4) Determine the 95 % confidence bound $CB_{95\,\%}$ based on the obtained Δ_{fit}- and $H_{k,fit}$-distribution, i.e. MC-based error.

For (1), the simulated datasets are based on the inverse of Eq. (6.33) and have a $\Delta_{sim} = 70\,kT$ and $\mu_0 H_{k,sim} = 300\,mT$. With a random number generator we choose a value for F between 0 and 1. Each value represents a switching cycle and results in a switching field $H_i = Rt_i$. Each switching cycle (AP-to-P/P-to-AP) occurs at a fixed ramp-rate. Multiple ramp-rates are used and combined into one full dataset. (2) Each full dataset is then fitted to give a (Δ_{fit}, $H_{k,fit}$). (3) This process is repeated 500 times. As input parameters we thus have $\Delta_{sim} = 70\,kT$, $\mu_0 H_{k,sim} = 300\,mT$, the number of generated

datapoints for each ramp-rate and the amount of ramp-rates (between 0.01 Hz and 10 Hz). Example distributions are shown for one ramp-rate in Fig. 6.17(a).

(4) We find that the (Δ_{fit}, $H_{k,fit}$) are Gaussian distributed around the simulated (Δ_{sim}, $H_{k,sim}$), see Fig. 6.17(b). The 95 % CB is determined by the Δ_{fit}-distribution, approximated by a normal distribution. This 95 % CB is what we will call the MC-based error. A high number of cycles and using multiple ramp-rates, reduce the error, see Fig. 6.18. With 95 % confidence the error can only be reduced below 5 % by having more than 1000 cycles if you perform switching cycles with only one ramp-rate. In another case, we use two ramp-rates, with speeds spaced one decade, i.e. 1 and 10 Hz. The 5 % error line is now crossed for 600 cycles, i.e. 300 cycles per ramp-rate.

In addition, the fastest two ramp-rates, namely 10 Hz and 1 Hz, have similar total measurement times, due to external delays. Therefore, the increase in accuracy can go hand in hand with a reduction in measurement time. Furthermore, the error based on the Monte-Carlo simulation is compared with the one derived out of the maximum likelihood fit, $CB_{95\,\%,ML}$ in Eq. (6.37). The ML error is derived for each of the 500 fits, and the median of the error is depicted for 1 ramp-rate by the black open circles and matches with the Monte-Carlo-based error [Fig. 6.18]. Eq. (6.37) is thus an accurate way to estimate the error of the extracted parameters, considering the used model is correct.

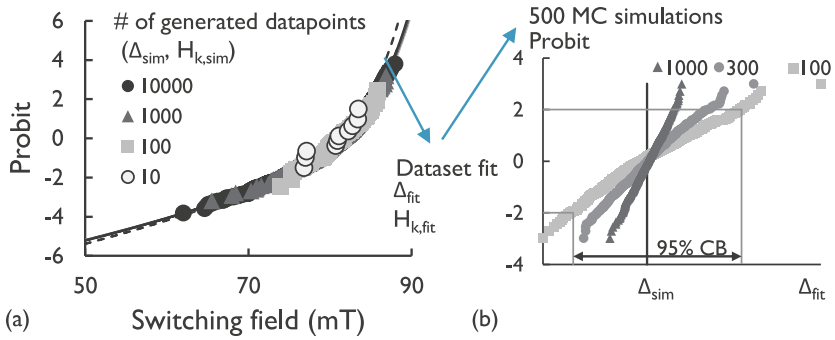

Figure 6.17: *Monte-Carlo (MC) simulations. (a) Example of fitting procedure for one ramp-rate and different number of generated datapoints. These datasets are then fitted for a Δ_{fit} and $H_{k,fit}$. (b) For 500 different MC simulations, we find that the (Δ_{fit}, $H_{k,fit}$) are Gaussian distributed close to the simulated (Δ_{sim}, $H_{k,sim}$)(# generated datapoints 100 (square), 300 (circle) and 1000 (triangle). The 95 % CB is depicted for the Δ_{fit}-distribution using 100 datapoints.*

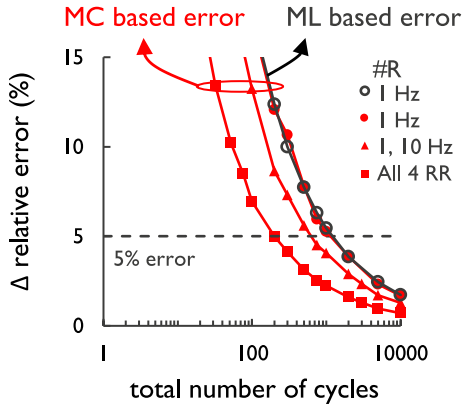

Figure 6.18: *Monte-Carlo simulations predict the relative error in Δ (red curves). Applying multiple ramp-rates and increasing the number of switching events reduce the error. The error derived by the maximum likelihood method (ML) (black open circles, Eq. 6.37) is similar compared to Monte-Carlo (MC) simulations (red filled , Fig. 6.17(b)).*

6.7 Experimental comparison of domain wall and Macrospin model

In this section, we investigate more in depth the differences between the domain wall and Macrospin model introduced in section 6.5. In recent literature, fits of both models cannot be distinguished in the measured magnetic field range [116]. However, only small datasets are used by Thomas et al. Therefore in this section we try to distinguish the fitting with large datasets, having multiple ramp-rates.

To compare the DW and MS model, we use them to fit a large switching sample size, i.e. > 2500 switching events and multiple ramp-rates. A large device size of $\varnothing\, 500\,\mathrm{nm}$ is used, where domain wall nucleation and propagation is highly expected. We find a significant difference, in favor of DW, using a likelihood ratio test, see below, with a log-likelihood ratio of 20. Within the measured field range, both models differ only slightly [see Fig. 6.19(a)]. However, there is a significant difference for the extracted thermal stability (at zero-field).

We compare the Macrospin and domain wall model for different sizes in a range $60\,\mathrm{nm}$ to $500\,\mathrm{nm}$ (Fig. 6.19(b)). Δ is expected to increase quadratically or linearly with the FL diameter for the Macrospin or domain wall model,

respectively (as seen in Fig. 6.14(a)). For the domain wall model, the extracted Δ follows this linear trend, however, for the Macrospin model, the extracted Δ remains constant. The apparent lack of size dependence in extracted Δ with the MS model has been reported and attributed to sub-volume excitation [108], where only a sub-volume is thermally excited. Another explanation could be the nucleation and propagation of a domain wall. Note that the very high Δ extracted for very large devices like $\varnothing$ 500 nm might be an overestimation, because in the simplified domain wall model, only 1 domain wall is considered in a 2 domain configuration, and at these large dimensions reversal could be by multiple domain walls in a multi-domain configuration. The extrapolation to zero-field can be overestimated.

Furthermore, we investigate with Monte-Carlo simulations (like in Sec. 6.6) how large the dataset should be to start seeing statistically significant differences between both models, in the given measurement range. To achieve this, we make use of the maximum likelihood ratio test used already in chapter 4. The ratio can also be written as a difference of the log-likelihood function, which, for numerical reasons, is preferred:

$$\frac{\Lambda_{MS}}{\Lambda_{DW}} = e^{\left(\ln \Lambda_{MS} - \ln \Lambda_{DW}\right)}, \tag{6.38}$$

where, Λ_{MS} and Λ_{DW} are the maximum likelihood value of the Macrospin and domain wall model, respectively. In order to attribute a significance to the likelihood ratio, we approximate the likelihood of the MS model by a multivariate normal distribution around the maximum. Next, we determine the

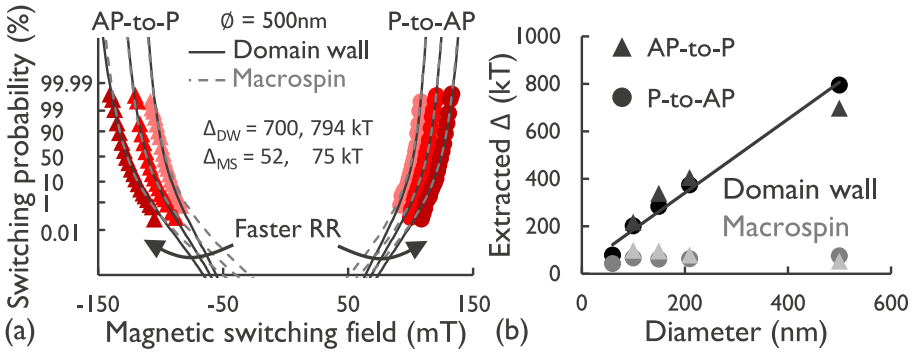

Figure 6.19: *(a) Comparison of domain wall and Macrospin model fitted on a large experimental dataset of* $>$ 2500 *switching events for 3 ramp-rates (0.1 Hz, 1 Hz and 10 Hz) and both AP-to-P as P-to-AP. (b) Comparison of the extracted* Δ *for different sizes.*

significance by comparing with the quantile of the χ^2 distribution:

$$\textit{Significance level} = e^{\left(-\frac{\chi^2(CL,n_p)}{2}\right)}\,, \tag{6.39}$$

where CL and n_p are the confidence level and number of parameters, respectively. For likelihood ratio′s larger than the Eq. 6.39 we will reject the DW model in this case.

$$\ln \Lambda_{MS} - \ln \Lambda_{DW} > -\frac{\chi^2(CL, n_p)}{2}\,. \tag{6.40}$$

For 2 fitting parameters, the 95 % confidence level is at $\frac{\chi^2(95\%,2)}{2} \approx 3$.

Our approach is summarized as follows:
(1) Simulate a dataset composed of number of switching cycles x at 1 RR, using a Macrospin model ($\Delta = 70\,\mathrm{kT}$ and $H_{sw} = 250\,\mathrm{mT}$).
(2) Fit the simulated dataset with the Macrospin and domain wall model with a maximum likelihood method.
(3) Calculate the maximum likelihood ratio (Eq. 6.38).
(4) Repeat (1)-(3) for the same number of switching cycles x.

The results are shown in Fig. 6.20. In Fig. 6.20(a), we observe an increased likelihood ratio for larger number of cycles within the dataset, meaning the simulated Macrospin model fits statistically better than the domain wall model. The likelihood ratio distributions can be estimated by a Gaussian distribution. The horizontal solid line in Fig. 6.20(a) depicts the log-likelihood ratio above which 95 % of the Monte-Carlo simulations are found. The 95 % likelihood ratio, or the ratio above which 95 % of the simulations lie, is then plotted as a function of the number of cycles in a dataset in Fig. 6.20(b). We find that statistically differentiating between domain wall and Macrospin model is only possible for datasets with more than 30 000 switching cycles, given the simulated parameters $\Delta = 70\,\mathrm{kT}$ and $H_{sw} = 250\,\mathrm{mT}$.

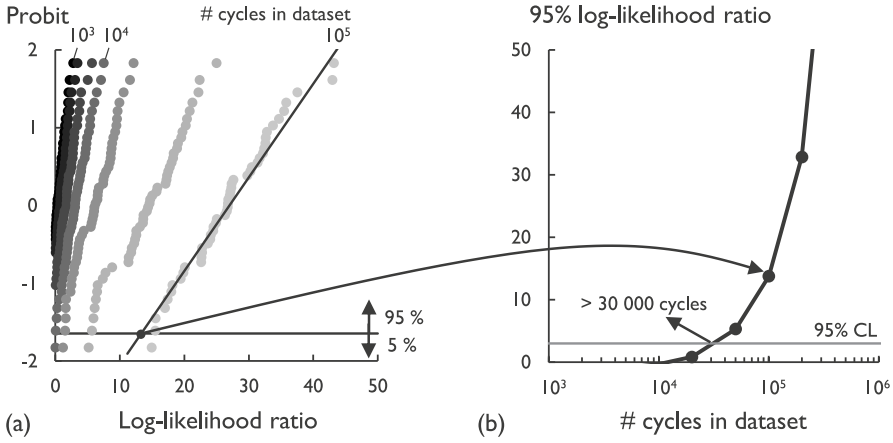

Figure 6.20: *Likelihood ratio tests to study to the required dataset size to statistically differentiate between the fitting of a simulated Macrospin and a fitted domain wall model. The likelihood ratio is determined by Eq. 6.38, the higher the log-likelihood ratio, the better the Macrospin model fits the data compared to a domain wall model, the critical significance for 95 % confidence is* ≈ 3. *(a) The log-likelihood ratio distributions composed of 60 Monte-Carlo simulations with a number of switching cycles in the dataset in the range 1 000, 2 000, 5 000, 10 000, 20 000, 50 000, 100 000. Increasing the #cycles in the dataset increases the likelihood ratio. For the 95 % confidence bound, we take corresponding likelihood-ratio-value out of the distribution, see intersection of the two solid lines. (b) This 95 % likelihood-ratio-value is compared with the 95 % critical significance level as a function of the #cycles in the dataset. Statistically differentiating between domain wall and Macrospin model is only possible for datasets with more than 30 000 switching cycles.*

6.8 Combining acceleration methods to validate the switching model and include self-heating

We have found that due to a large correlation of the fitting parameters, large statistics are required to reduce the error on the extracted thermal stability. In addition, we have found that the extracted thermal stability depends on the used switching model. We have compared a Macrospin and a domain wall model. Both models fit the data perfectly within the accelerated measurement range. However, the extracted thermal stability strongly differ. Besides the required extension of the measurement range, i.e. measuring more switching cycles, we discuss here another way to validate using a combination of acceleration methods.

In order to validate the switching models and find which extracted thermal stability is more correct, in this section, we combine the different acceleration methods and based on their comparison we conclude that the Macrospin model for the studied $\varnothing$ 120 nm size is not justified to use to extrapolate the thermal stability. To achieve this we have proceeded as follows:
(1) Measure current- and magnetic field acceleration at different temperatures.
(2) Extract the temperature dependence of thermal stability from magnetic field acceleration.
(3) Include self-heating to explain differences between current- and magnetic field acceleration.

(1) Current- and field acceleration at different temperatures

We do an in-depth analysis on several $\varnothing$ 120 nm single devices. The studied stack was a conventional 375 °C annealed wafer, with around 160 % TMR and an RA of 10 $\Omega\mu m^2$. There is however a large offset field of about 30 mT, favoring the AP-state, as can be seen from the TMR loops in Fig. 6.21.

Current- and magnetic field-accelerated measurements are performed at 3 different temperatures (300, 350 and 400 K). For Magnetic field-induced switching, we make use of two ramp-rates (1 and 10 Hz) with each 300 repeats. The thermal stability is extracted as discussed in Sec. 6.6 for the domain wall and Macrospin model. For current-induced switching, the thermal stability is extracted using the switching current value of 20 repeats per pulse width (τ_{pw}) for 4 different pulse widths (between 1 μs and 30 ms), i.e. in the thermal regime. The median switching current with correction of the common model by Koch et

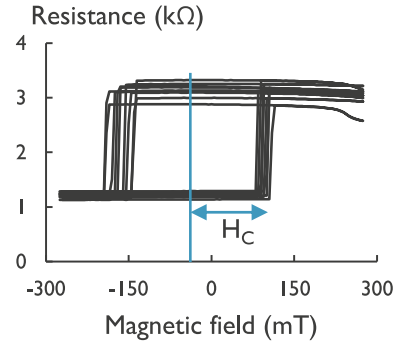

Figure 6.21: *TMR loops of the ⌀ 120 nm devices studied in Sec. 6.8. TMR is close to 160 % and the offset field is approximately 30 mT favoring the AP-state.*

al.[66] is then given by Taniguchi et al. [113]:

$$I_{sw,med}(\tau_{pw}) = I_c\left[1 - \sqrt{\frac{1}{\Delta}\Big(ln[f_0\tau_{pw}] - ln\Big[ln2\Big]\Big)}\right]. \tag{6.41}$$

The ln[ln 2] term results from the fact that the median $I_{sw,med}$ is fitted. As such, this equation can be derived by substituting $F = 0.5$ in the switching distribution (Eq. 6.17), with the energy barrier given by the Macrospin model $E_b = \Delta\big(1 - \frac{I_{sw}}{I_c}\big)^2$ (Eq. 6.25).

An example of the effect of temperature on the switching current and the magnetic field switching distribution are shown in Fig. 6.22 for a typical device of the measured ⌀ 120 nm devices. The switching current and magnetic switching field decrease when increasing the temperature for both directions. However, the effect of temperature is more pronounced in the magnetic field accelerated case. We emphasize this difference by depicting the normalized median switching current and median switching field as a function of temperature for both switching direction (Fig. 6.23). In the AP-to-P case (black), the median switching current (open circle) only drops to 80 % at 400 K, whereas the median magnetic switching field (solid triangle) to 50 %. For P-to-AP, the drop with temperature is even more higher. The large difference between AP-to-P and P-to-AP can be explained by the influence of a large offset field (see Fig. 6.21). Note that no current switching data is extracted at 400 K for P-to-AP, because the impact of the V_{read}-pulse at 100 mV would not be negligible anymore. Next, we will fit the thermal stability at different temperatures.

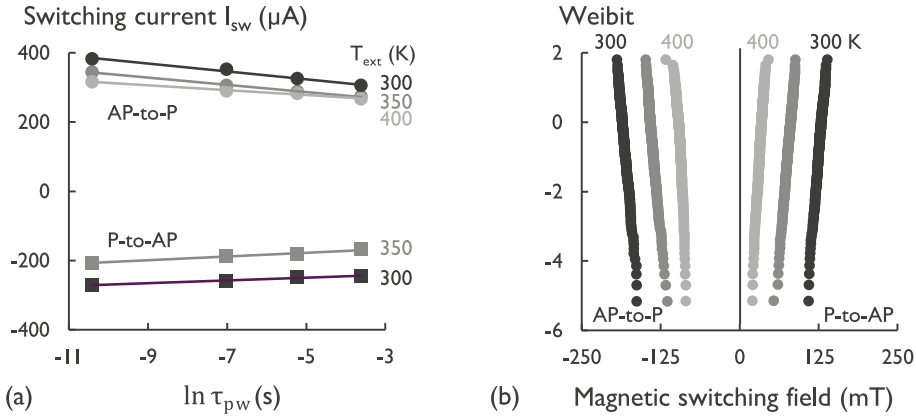

Figure 6.22: *Current and magnetic field-accelerated switching measurements for a typical ⌀ 120 nm device at 3 temperatures (300, 350, 400 K). (a) Median switching current, derived from 20 repeats, as a function of the pulse width. (b) Magnetic field switching distribution for 300 repeats at a ramp-rate of 10 Hz.*

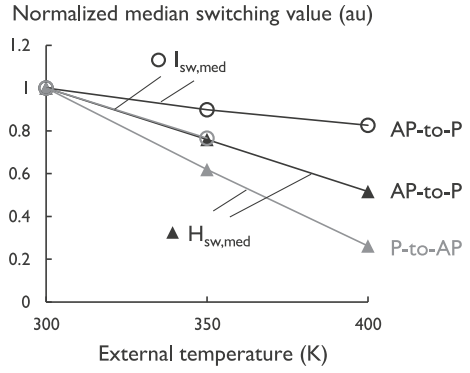

Figure 6.23: *Normalized median values for current and magnetic field-accelerated switching measurements for a typical ⌀ 120 nm device as a function of external temperature (300, 350, 400 K). The median values are normalized with respect to 300 K. The color refers to the switching direction, AP-to-P (black) and P-to-AP (gray). The symbol refers to the acceleration method, current (open circles) and magnetic field (solid triangles). The switching current has a lower dependence on external temperature than the magnetic field. The large difference between AP-to-P and P-to-AP can be explained by the influence of a large offset field (see Fig. 6.21).*

(2) Extract the temperature dependence of thermal stability from magnetic field acceleration

The thermal stability is extracted for both magnetic field-accelerated measurements as current-accelerated measurements. The procedure for the magnetic

field-accelerated switching is based on a maximum likelihood fitting method elaborated in Sec. 6.6. Since we make use of 2 ramp-rates and 600 cycles in total, we reduced the relative error on Δ below 5 %, considering the used switching model is correct.

The fitting for current switching is based on the median switching current dependence on the pulse width. A similar method cannot be applied for magnetic fields, since it is very difficult to apply a magnetic pulse in a large time range with our setup (μs to seconds, see Sec. 6.4.2). This is a common method to fit Δ in literature. 95 % confidence bound errors, considering a 2 parameter fit (Δ, I_c) results in relative errors on Δ between 10-20 %. As from Eq. 6.41, the thermal stability Δ depends on the slope of the switching current as a function of pulse width. This slope decreases for AP-to-P for higher temperature (Fig. 6.22(a)), and hence Δ increases with temperature for AP-to-P switching, which is counter-intuitive.

The results of the thermal stability fitting are summarized for the Macrospin model in Fig. 6.24. A first observation is that the extracted Δ via magnetic field (Δ_{field}) correlates well with the coercive field, however, Δ extracted from current ($\Delta_{current}$) for AP-to-P does not [see solid lines, i.e. linear fits, in Fig. 6.24(a,b)]. The coercive field describes how resilient the magnetic state in the FL is against magnetic field-induced switching. It is therefore hypothesized to be correlated with the thermal stability of the free layer. The reason why $\Delta_{current}$ does not correlate with H_c is unknown. One possibility is because the model does not incorporate self-heating, another because the switching mechanism is not well described by a Macrospin model.

A second observation is that for a Macrospin model the extracted $\Delta_{current}$ are higher than Δ_{field}. There are two major differences between the acceleration methods. One is a difference in temperature due to self-heating of the MTJ when applying the write pulses. The other is the switching mechanism, i.e. by STT instead of a direct external magnetic field. Assuming the STT mechanism has a similar effect on the FL, then an elevated temperature should cause the difference in Δ.

In Fig. 6.25, the temperature dependence on extracted Δ_{field} for AP-to-P is shown. A linear fit results in a slope φ_T of around -0.3 kT/K for the Macrospin model (see Fig. 6.25). We use the slope of each device to trace back the temperature difference that results in the difference in Δ_{field} and $\Delta_{current}$.

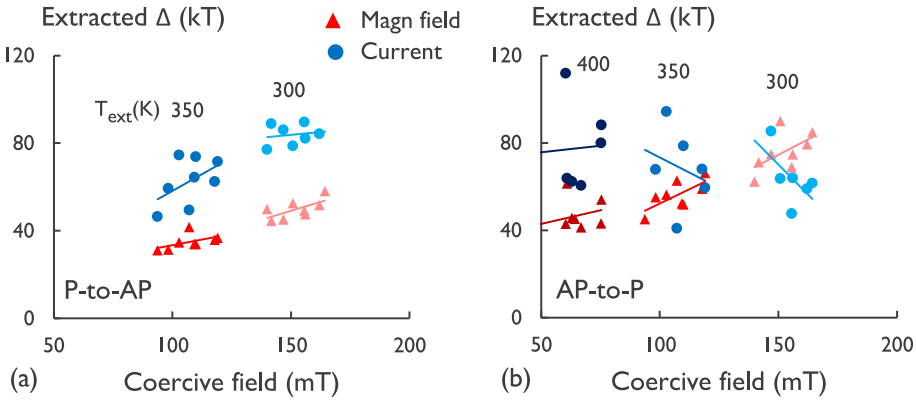

Figure 6.24: *Temperature analysis of the Δ-extraction with magnetic field and current for MTJs with nominal diameter of 120 nm. (a, b) Extracted Δ (AP-to-P, P-to-AP) for magnetic field and current switching as a function of coercive field, which correlates for field, but not for current in the AP-to-P case.*

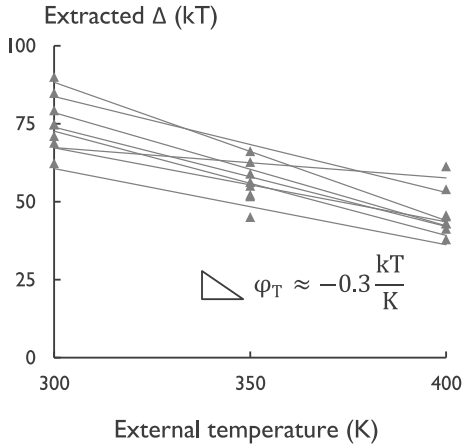

Figure 6.25: *Determination of the slope φ_T of the extracted Δ at external temperature, for Macrospin model, as a function of external temperature.*

(3) Include self-heating to explain differences between current and magnetic field acceleration

During the magnetic field acceleration only a 10 mV bias is applied to the devices, resulting in negligible self-heating. As such, the Δ_{field} is characterized for the given external temperature. The temperature dependence of Δ_{field} is measured

at 3 temperatures (300, 350, 400 K) and shown previously in Fig. 6.25. The temperature dependence fits well with a linear slope φ_T. If we hypothesize that the Macrospin model is correct for the derivation of Δ, the differences could be explained by a difference in temperature. We therefore use φ_T, measured for each device, to trace back the self-heating temperature that explains the difference between Δ_{field} and $\Delta_{current}$ by:

$$T_{self-heating} = \frac{\Delta_{current} - \Delta_{field}}{\varphi_T}. \tag{6.42}$$

The extracted $T_{self-heating}$ for a Macrospin model is depicted in Fig. 6.26. The self-heating temperature becomes negative, i.e. cooling, which is not possible. Our simulations suggest that for a similar MTJ stack a self-heating temperature between 25 K to 100 K is found for the observed switching conditions, i.e. powers of 50 and 200 μW for P-to-AP and AP-to-P, respectively. This is a clear indication that for our $\varnothing$ 120 nm devices the Macrospin approximation does not hold and will result in a wrong estimation of the thermal stability at zero-current and zero-field.

We also extracted the Δ_{field} for a domain wall model. As seen in section 6.7, Δ_{field} is larger for a domain wall model as for the Macrospin model. Similar results are seen in this case, and shown in Fig. 6.27. We could also extract a self-heating temperature from the domain wall fitted Δ, however, no equivalent domain wall model currently for current-induced switching was

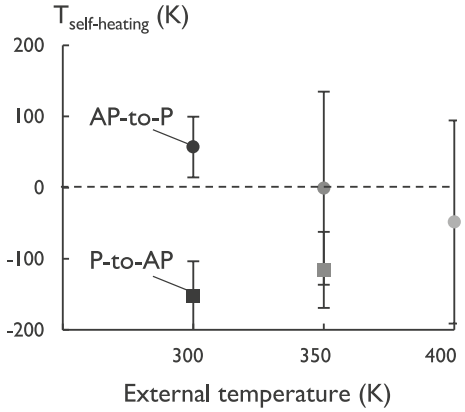

Figure 6.26: *Self-heating temperature derived using Eq.(6.42) as function of external temperature. Negative self-heating temperatures, i.e. cooling, are predicted, which is not possible and an indication that the Macrospin model is not valid. The error bars represent the standard deviation between the 8 measured $\varnothing$ 120 nm devices.*

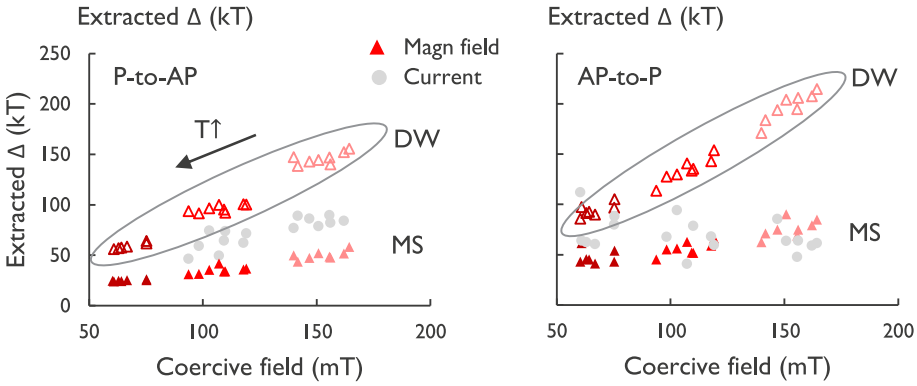

Figure 6.27: *Extracted Δ as a function of coercive field. In addition to Fig. 6.24, the magnetic field extracted Δ for a domain wall (DW, open symbols) model are included. The domain wall model predicts higher Δ and a steeper temperature dependence, i.e. slope φ_T.*

implemented. The conceptual approach of Munira and Visscher could be used for this, although it also adds an additional fitting parameter, which will affect the accuracy significantly [79].

In summary, we find by comparing magnetic field and current acceleration for the extraction of thermal stability using a Macrospin model, that (i) the Macrospin model is not valid. (ii) Significant differences are found in the extracted Δ from magnetic field and current acceleration. (iii) These differences in Δ cannot be explained by the influence of self-heating in the current acceleration, because this results in unphysical negative self-heating temperatures.

Therefore, the Macrospin model should not be used for these ⌀ 120 nm devices. Instead, a domain wall model should be used. In the relevant sizes, the simplified domain wall model discussed in section 6.5, correctly estimates the energy landscape when switching is induced by a magnetic field. However, no equivalent domain wall model exists for current-induced switching and therefore the self-heating could not be extracted by comparing current and field acceleration.

6.9 Conclusions

The thermal stability parameter defines the data retention at off-state conditions, i.e. zero field, zero current and operation temperature. The data retention cannot be characterized, however, at these off-state conditions. Therefore acceleration is necessary. Relying on switching models, the accelerated switching distributions are fitted and extrapolated to off-state conditions. Three acceleration methods are considered, temperature, magnetic field and current acceleration.

Temperature acceleration is very time consuming and is generally modeled with a simple Néel-Brown model. Unfortunately, this simple model does not take into account a temperature dependence of different magnetic properties, like offset field, magnetization saturation and anisotropy field. In order to use this type of acceleration, the temperature dependence of these magnetic properties needs to be characterized and included into the model. Furthermore, considering available time, temperature acceleration is only reasonable using arrays and extracting bulk parameters. A single device study and thus a study of the tail bits is not possible.

Magnetic field acceleration is suited to study both single devices as well as arrays. The Δ extraction is based on at least 2 fitting parameters, which are highly correlated. Because of the high correlation, large statistics, i.e. number of switching cycles, are necessary to accurately fit the switching data. Based on a Monte-Carlo approach and a maximum likelihood fitting method, we find that more than 1000 switching cycles are necessary to reduce the relative error down to 5 %. Furthermore, the Δ depends on the used switching mechanism, by Macrospin or by domain wall.

Current acceleration is well suited to study single devices. In the thermal regime, experimental time ranges starting from 100 ns can be used. As for the magnetic field acceleration, the Δ extraction, relies on simplified models, that depend on the switching mechanism. Furthermore, the large currents will cause self-heating of the MTJ. This self-heating has to be included in the models in order to correctly derive Δ.

In the accelerated measurement range the switching mechanisms are difficult to differentiate. For example, based on Monte-Carlo simulations and a maximum likelihood ratio test, we find that for ramped magnetic field switching more than 30 000 cycles are required to see statistical differences.

Furthermore, the Macrospin and domain wall models have been validated based on the Δ extracted by current and magnetic field acceleration. Using the Macrospin model, these Δ do not match. These differences in Δ cannot be explained by the influence of self-heating in the current acceleration, because this results in unphysical negative self-heating temperatures. Therefore we conclude that the Macrospin model is not suited to extract the Δ for the studied device sizes. In addition, the Δ has been extracted using a domain wall model. However, no equivalent domain wall model to extract Δ for current switching exists. As such, the domain wall model could not be validated for magnetic field and current acceleration.

Chapter 7

Conclusions and outlook

Conclusions

We have started this thesis explaining the high potential STT-MRAM has to be introduced in the memory market and beyond. STT-MRAM is targeted for various applications like embedded Flash replacement, or SRAM/DRAM replacement, or in emerging markets like the Internet-of-Things. For this, STT-MRAM needs to operate reliably. We have considered two important reliability concerns:

1. Breakdown of the tunnel barrier causes limited endurance specifications. STT-MRAM operates at high current densities, causing significant self-heating and breakdown of the tunnel barrier. By performing an in-depth analysis corroborated with large measured breakdown statistics using (1) a breakdown time range of more than 11 orders of magnitude and (2) different MgO thicknesses, we extended the breakdown model by incorporating the self-heating effect. Moreover, we find that scaling down the MTJ dimension and MgO thickness will further increase the reliability margin between breakdown and switching.

2. Insufficient data retention results in loss of the non-volatility property. Data retention is determined by the energy barrier needed to overcome at off-stress conditions. There are many publications on switching mechanisms, switching models and acceleration techniques. There is, however, no consensus on which model is correct and more importantly none of the existing models take into account the effect self-heating will have on the extraction of the energy

barrier. We perform an in-depth statistical evaluation of the measurement techniques and switching models, and provide a baseline for validating what switching mechanism describes best the extrapolation to off-stress conditions. For this, a combination of temperature, magnetic field and current acceleration, including self-heating, is implemented.

In the following we report on the main findings categorized per chapter:

Chapter 2: A reliable STT-MRAM technology is composed of many nanometer thick layers. All layers contribute to a 400 °C compatible process and contribute to achieve good performance for all basic properties like PMA, reference layer stability, high TMR, low switching power and reliability.

Chapter 3: The many nanometer thick layers in STT-MRAM affect not only the electric and magnetic, but also the thermal properties. Since excessive temperature deteriorates all the important MTJ properties, a theoretical model on MTJ cannot be complete without accurate self-heating characterization.

We find that the poorly thermal conducting layers, close to the MgO, cause an ultra-fast heating in sub-ns. In order to reduce the thermal resistance, the TaN bottom electrode, the TiN hard mask and top electrode are crucial parts. Thermal simulations indicate that self-heating can result in temperatures of 200° C at breakdown stress conditions. In these simulations the thermal boundaries affect mostly the time constants needed to reach a steady-state temperature. The thermal resistance derived from these simulations depends, however, on poorly known thermal conductivity values for the multiple thin film layers of which the STT-MRAM is composed.

In order to assess the thermal resistance, an indirect conceptual method has been developed and demonstrated. This method relies on the breakdown statistic, and the results match well with our simulations.

Chapter 4: The breakdown distributions in STT-MRAM are well described by a Weibull distribution and a power-law voltage acceleration. By the formation of a percolation path between anode and cathode, the generated defects create a conductive filament across the MgO barrier and hence irreversible breakdown occurs. For the studied ultra-thin MgO dielectric, a single or two defect path can be sufficient to break the dielectric, resulting in low Weibull slopes (< 1).

We have developed a more robust all-in-one maximum likelihood fitting method that simultaneously fits all the measured breakdown data, collected at different stress conditions, for the Weibull parameters and the voltage acceleration parameter.

We have demonstrated the equivalence between a constant voltage stress, a ramped voltage stress and a pulsed breakdown measurement of the MgO in STT-MRAM. We conclude that the breakdown time can be expressed as the cumulative pulse time, independent of duty cycle or pulse width down to 30 ns. This demonstrates that the oxide degradation process is cumulative in nature and has no measurable relaxation mechanisms.

For thin MgO the power-law acceleration, based on the anode hydrogen release model, describes best the voltage and temperature acceleration of breakdown. This is supported by a maximum likelihood ratio test with a likelihood ratio $< 10^{-20}$. Instead of hydrogen, other atoms or analogous mechanisms could cause breakdown, again resulting in a power-law model. For example the effects of oxygen diffusion discussed in Chapter 5.

Chapter 5: Apart from process-related extrinsic defects, the processing also influences diffusion mechanisms in the device, which in turn impacts the breakdown mechanisms. Furthermore, it is demonstrated that taking self-heating into account is imperative to make accurate lifetime predictions.

We observe that RIE induces more breakdown time and breakdown voltage variability than IBE, indicating more significant edge damage during a RIE etch. However, the variability of RIE can still be improved by an optimized post-etch treatment. In addition, we find only a minor impact of the MgO deposition technique on breakdown voltage. A more significant impact is found for changing the thin spacer layers in the STT-MRAM-stack. For these last results, we propose an oxygen scavenging model to explain the increased susceptibility to breakdown for the standard Ta spacers.

The reliability margin can be increased going to lower RA, by thinning down the MgO to 0.8-0.9 nm, because the switching voltage decreases faster than the breakdown voltage. In addition, these low RA values of 3-5 $\Omega\mu m^2$ are necessary to achieve reasonable resistance values in ultra-scaled devices with sub-20 nm diameter.

Self-heating has a large impact on breakdown. At typical RVS conditions,

the degradation in thin MgO occurs at temperatures around 200 to 300 °C. Reducing self-heating will improve the breakdown characteristics. To achieve this, a reduction of the thermal resistance is required, such that for the same power there is less self-heating. Another way is to decrease the area. Small areas have reduced self-heating at the same breakdown voltage. This temperature difference between different MTJ sizes explains the failure of the area scaling rule in 1 nm MgO, and the observation of so-called "beyond area scaling".

An important consequence of the self-heating is that the temperature at operating conditions will be lower than at breakdown measurements conditions. As a result, the lifetime extrapolation to operation conditions is underestimated. It is therefore imperative to include self-heating effects in the breakdown model in order to perform accurate lifetime extrapolations from breakdown stress conditions to operating conditions. As such, the breakdown model consist of the breakdown distribution parameters β and η, the temperature dependent power-law voltage acceleration and the thermal resistance of the MTJ.

Chapter 6: Accurately extracting the thermal stability, and thus characterizing the data retention in STT-MRAM, requires large switching statistics, obtained by accelerating switching with magnetic field, current or temperature.

Temperature acceleration is very time consuming and the benefits from a simple Néel-Brown description are partly lost, since the material parameters determining the thermal stability show temperature dependence. Furthermore, temperature acceleration is only reasonable using arrays and extracting bulk parameters. A thorough study of the tail bits is not possible.

Magnetic field acceleration is suited to study both single devices as arrays. Thorough analysis of single devices allows to measure the exact Δ-distribution in large statistics and as such take into account and study the tail bits. The Δ extraction is based on at least 2 fitting parameters, which are highly correlated. Because of the high correlation, large statistics, i.e. number of switching cycles, are necessary to accurately fit the switching data. Based on a Monte-Carlo approach and a maximum likelihood fitting method, we find that more than 1000 switching cycles are necessary to reduce the relative error down to 5 %. Furthermore, the Δ depends on the used switching mechanism, by Macrospin or by domain wall.

Current acceleration is well suited to study single devices. In the thermal regime, experimental time ranges starting from 100 ns can be used. As

for the magnetic field acceleration, the Δ extraction, relies on simplified models, that depend on the switching mechanism. Furthermore, the large currents will cause self-heating of the MTJ. This self-heating has to be included in the models in order to correctly derive Δ.

In the accelerated measurement range the switching mechanisms (Macrospin or domain wall) are difficult to differentiate. For example, based on Monte-Carlo simulations and a maximum likelihood ratio test, we find that for ramped magnetic field switching more than 30 000 cycles are required to see any statistical difference. Due to the large inductance of the magnetic setup, generating this number of cycles takes a large amount of time.

Furthermore, the Macrospin and domain wall models have been validated based on the Δ extracted by current and magnetic field acceleration. Using the Macrospin model, these Δ do not match. Moreover, the difference in Δ cannot be explained by the influence of self-heating in the current acceleration, because this results in unphysical negative self-heating temperatures. In addition, Δ extracted for different MTJ areas is inconsistent with a Macrospin model. Therefore we conclude that the Macrospin model is not suited to extract the Δ for the studied device sizes ($\varnothing$ 120 nm).

Outlook

There are still a lot of opportunities for further research. Related to the thesis we consider the following:

Breakdown of the MgO tunnel barrier: Self-heating significantly contributes in the acceleration of breakdown. Reducing the self-heating by reducing the thermal resistance is crucial. Several approaches are possible, for example incorporate good heat conducting encapsulation materials, reduce the thickness of the TaN bottom electrode and TiN hard mask and top electrode, e.g. in a top electrode less approach, where the TiN is replaced by Cu. Making use of the extended breakdown model, the effect of these adjustments can be accurately studied and further corroborated with thermal simulations.

The empirical generalized power-law model can be improved by more accurately determining the Mg-H bonds vibrational energy levels, and or investigate potential effects of oxygen diffusion at elevated on-state device temperatures.

In this thesis unipolar pulses are used for breakdown measurements to make sure the device remains in parallel state. By preventing switching form AP-to-P and P-to-AP, we can have a more stable device stressing, without potential perturbation due to switching. However, several studies report a reduced breakdown time for bipolar pulse stress [2, 59]. From [2], a charge trapping and detrapping mechanism explains the difference, where for bipolar pulses all charges are trapped/detrapped at each alternating pulse, and hence this strong modulation of trapped charges deteriorates the breakdown. However, the authors used this charge model also to explain differences seen in unipolar stress with different pulse widths and duty cycles. For unipolar stress our results are not consistent with a charge trapping/detrapping model. Therefore, the physical mechanism in bipolar pulses is still under debate.

How will breakdown be influenced by sub-nanosecond pulses? Characterizing sub-ns breakdown will be interesting to link with the sub-nanosecond switching regimes.

Data retention: The simple Néel-Brown model for temperature acceleration should be extended and tested with the temperature dependence of several material parameters like offset field, coercive field and magnetization saturation. Furthermore, it will be interesting to study the Macrospin boundary in devices with MTJ diameters smaller than 20 nm, i.e. the expected dimension for Macrospin to be valid. With our indirect method the Macrospin model can be validated combining all acceleration methods.

There is a need to build a current-induced domain wall switching model for the extraction of the thermal stability via current acceleration. As a starting point, the model from Munira and Visscher could be used, where via an extra instability parameter the energy landscape is empirically adjusted [79]. This model can then be used to extract self-heating using the combination of the different acceleration techniques as presented in this thesis.

Study new type of devices: In **top-pinned** devices the MgO barrier is at the bottom of the pillar and therefore is less influenced by the patterning. In addition, it is possible to deposit the MgO layer close to the ultra-smooth bottom electrode. As a result, the TMR of top-pinned structures is higher than that of bottom-pinned structures, but there is no BEOL compatibility in these top-pinned structures. However, recent results enable a 400° C top-pinned approach [111]. These achievements are also an important step in enabling emerging devices like the spin-orbit torque MRAM.

Spin-orbit torque MRAM is a three-terminal, top-pinned MTJ-based device, where the switching is induced by means of spin-orbit torque. The read and write path are decoupled, and as such, the high current pulses required to switch the free layer of the MTJ do not need to pass through the thin tunnel barrier, because current flows in the metal path under the MTJ. This way, the risk of breaking down the tunnel barrier is significantly reduced. Moreover, the switching mechanism is faster and sub-ns at potentially lower power.

Variability: Variability analysis was still lacking in this thesis, mainly because the in-house STT-MRAM technology cannot compete with the process of first-class companies, resulting in tail bits. Nevertheless, designing new type of structures like mini-array structures with neighboring MTJs connected at very dense pitch ($\leq 100\,\text{nm}$), can allow characterization of advanced nodes for MTJs with diameters smaller than 20 nm. Effects of variability should also be incorporated in the thermal simulations. Ultimately to make accurate predictions on the STT-MRAM performance.

Appendix A

Parasitic series resistance in the Mbit array test vehicle

In this thesis we make use of a Mbit array vehicle, where 1024x1024 cells are connected by transistors in a 4T1MTJ configuration. This way we can stress and read out each MTJ individually, as benefit we find all 1024 x 1024 MTJs within a 6x7 mm^2 area. As such, the device to device variability is reduced, e.g. variability coming from different oxide thickness or differences in hard mask remaining. In addition, large statistics can be obtained using the Mbit array, making it easier to test various stress conditions and test the conventional fitting models.

The 4T1MTJ layout makes a 4-point measurement possible, such that the series resistance of the periphery and the transistors can be measured. However, we find that within the 4-point measurement path, there is parasitic resistance coming from an oxidized layer between the BE and the metal 3 (M3) layers. With this unknown parasitic series resistance it is not possible to calculate the stress voltage over the MTJ in case the MTJ resistance is low, like in the 1 nm MgO devices.

In section A.1, we discuss the layout of the Mbit array. Next, in section A.2, we compare the design of the 4T1MTJ with the single device structure (0T1MTJ) in a non-CMOS wafer. Furthermore, we discuss the origin of the parasitic resistance and show the effect on the MTJ resistance.

A.1 layout of the Mbit array

One Mbit array has 1024 x 1024 cells available to measure with a 4-point measurement. All cells are contained in an area of only 6 x 7 mm^2. Each cell is connected to 5 transmission gate transistors. 2 transmission gate transistors serve as driving transistors, the other 3 as selector transistors. A transmission gate transistor is made of an NMOS in parallel with a PMOS transistor. 2 transmission gate transistors are the drive transistors (Fig. A.1). One transmission gate (drive transmission gate A) has large transistors, capable of delivering currents of more than 1 mA when fully open. The other transmission gate (drive transmission gate B) is smaller and can be used for a 4-point measurement, where the voltage is measured by a digital multimeter (DMM). This is done as follows: the stressing current flows from the force of the source line (SL), to the MTJ and through the drive transmission gate A, to the bit line A (BL_A). The DMM measures the voltage over the TE and BE of the MTJ. This is done via the sense SL, i.e. the TE, and the BE of the MTJ is sensed via drive transmission gate B at the bit line B (BL_B).

Furthermore, each cell has 3 selector transmission gate transistors, 2 for the bit lines (BL_A and BL_B) and 1 for the source line (SL). These transmission gates are opened, depending on which cell of the 1024 x 1024 is connected.

A.2 MTJ cell (4T1MTJ)

In the Mbit array the MTJ is connected to the transistor level via the M3 layer. In Fig. A.2, the main difference between the 4T1MTJ and 0T1MTJ structure is the absence of the BE-M3 connection, which in the cross structure of the 0T1MTJ is replaced with a connection upwards to the contact pad. The BE-M3 is necessary to connect to the lower transistor level.

A.3 Parasitic resistance between BE and M3 layers

We find that the 4-point resistance, measured in the Mbit array, is significantly larger than in the 0T1MTJ single device cross structures. We have measured for the same wafer devices from the 0T1MTJ single device structure and devices from the 4T1MTJ structures in the Mbit array. The Mbit array resistance is

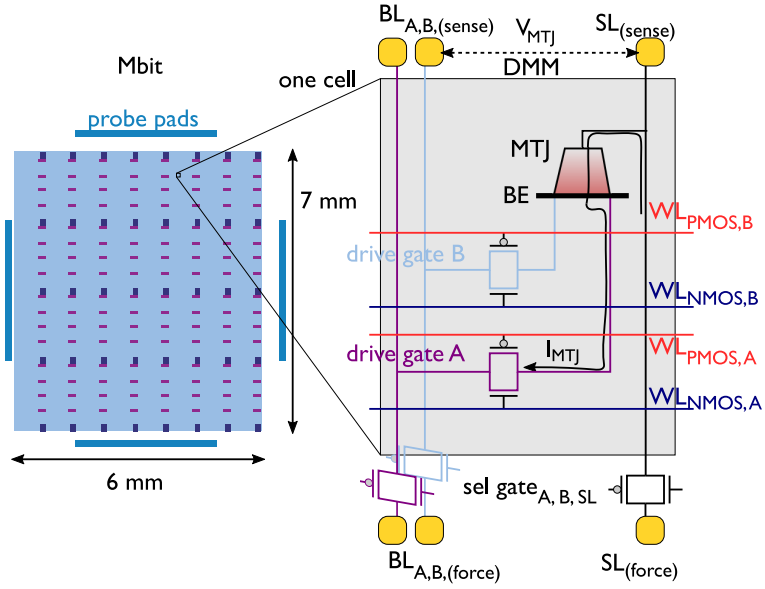

Figure A.1: *Schematic of a MTJ cell in the Mbit array, with 2 drive transmission gate transistors and 3 selector transmission gate transistors. Current flows through the source line (SL_{force}), the MTJ, drive transmission gate A and to bit line A (BL_A), see black arrow. A 4-point measurement is possible via the digital multimeter (DMM) over the sense source line (SL_{sense}) and bit line B (BL_B).*

approximately 650 Ω higher, see Fig. A.3(a). If we compare the TMR as a function of resistance, we can explain the drop in TMR by a series resistance effect (Fig. A.3(b)):

$$TMR(R_{series}) = \frac{R_{p,median}}{R_{p,median} + R_{series}} \cdot TMR_0 \,, \tag{A.1}$$

here $R_{p,median}$ and TMR_0 are the median parallel resistance and the TMR of the single device dataset, respectively. When adding a series resistance term (R_{series}), which does not contribute to the TMR, the TMR will drop, see solid line in Fig. A.3(b), according to Eq. A.1.

Since these measurement make use of a 4-point measurement, the parasitic series resistance has to be within the shared current path (Fig. A.4). Only the MTJ, BE and M3 share the measurement current path. Elsewhere, the voltage and current path are separated, such that in the voltage path no additional resistance drop is caused by the resistance of the periphery. In this way the voltage over the top of the MTJ and the bottom of M3 is correctly measured.

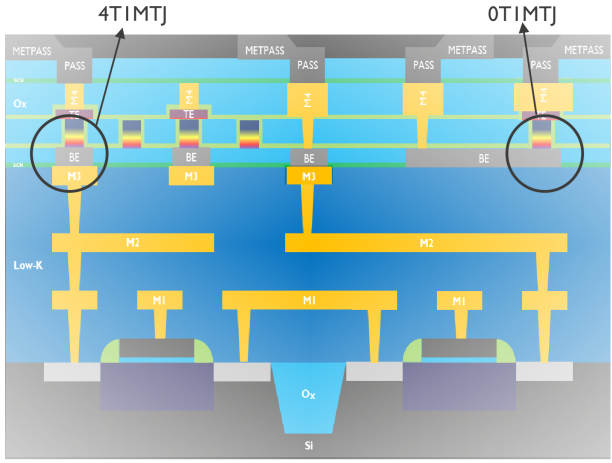

Figure A.2: *Comparison of device design between 4T1MTJ (left) and 0T1MTJ (right). The 4T1MTJ structure is connected to the transistor level via M3, whereas for the 0T1MTJ the BE is directly connected upwards to the contact pad.*

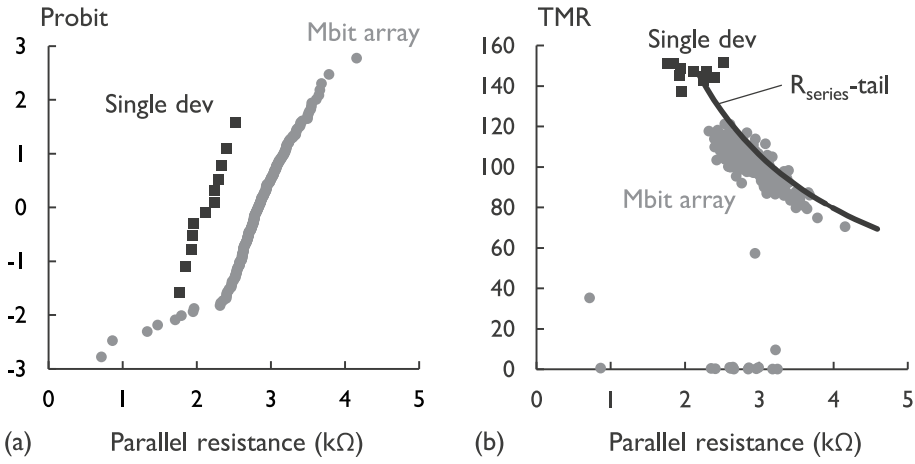

Figure A.3: *Measurement of the parasitic series resistance of ⌀ 60 nm devices in the Mbit array (grey circles) compared to single devices (black squares). (a) Parallel resistance distribution. (b) TMR as a function of the parallel resistance. The solid line represents the TMR-drop caused by series resistance effect based on the median resistance and TMR of the single devices using Eq. A.1.*

We performed Energy-dispersive X-ray spectroscopy (EDS) on a TEM sample. The results and location of the linescan located at the BE-M3 interface are

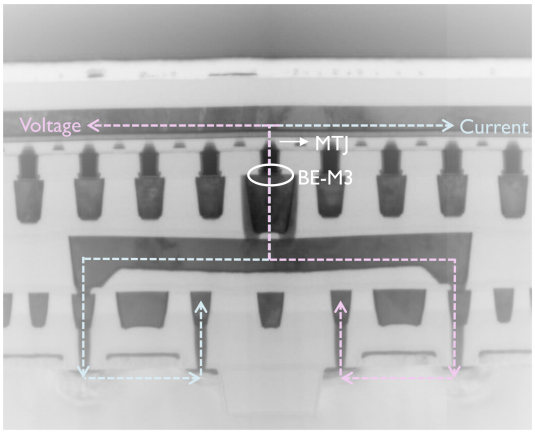

Figure A.4: *TEM micrograph of the 4-point probe path in the Mbit array. In the common shared current path, the BE-M3-interface is oxidized, causing the parasitic series resistance.*

depicted in Fig. A.5. The Ta interface is oxidized, causing a parasitic series resistance. During the process of the BEOL, currently after a reactive pre-clean of the Cu, Ta is deposited by physical vapor deposition. Next, TiN is deposited in a different tool with atomic layer deposition. In between there is a moment the Ta is exposed to air and can oxidize. Preventing any oxidation is not straightforward. The approaches to resolve the oxidation issues are not further discussed.

The parasitic series resistance can also not be extracted post-breakdown, since the instantaneous current increase when the MgO breaks down, also causes the parasitic oxide layer to break. In addition, the current compliance is limited to $\approx 2.5\,\mathrm{mA}$, which is the maximum current the NMOS transistor can deliver. As seen in section 3.3.1, this is not large enough to fully short the MTJ within the 4-point measurement path.

At the time of writing, this parasitic series resistance issue has not been resolved. For this reason, we do not make use of the Mbit array for electrical measurements of the 1 nm thick MgO. However, for thick MgO (e.g. 1.7 nm), the MTJ resistance is much higher than the parasitic resistance and will not affect the analysis. Furthermore, magnetic field-based measurements do not get influenced by the parasitic series resistance and can be performed on arrays with 1 nm MgO.

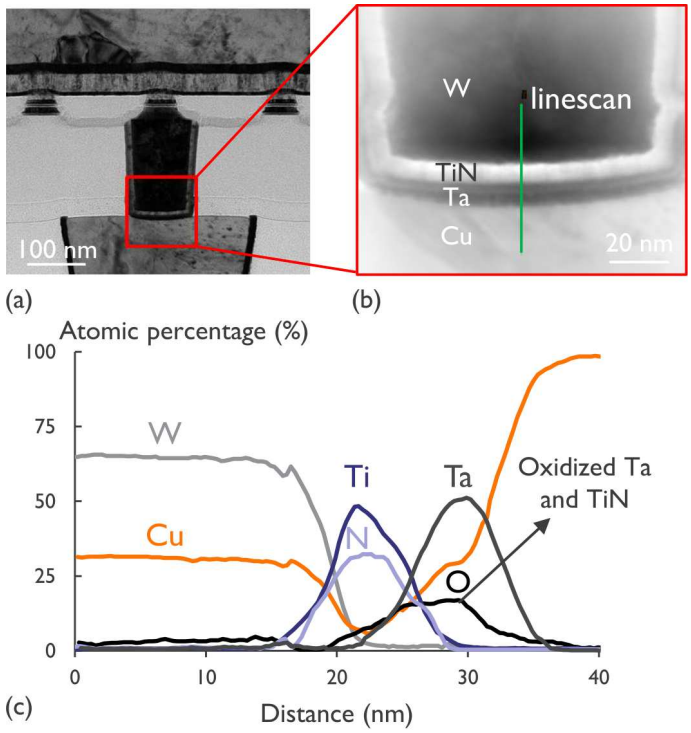

Figure A.5: *TEM micrograph (a-b) and EDS linescan of the M3 and BE interface (c). The TiN and Ta layers between the W-BE and Cu M3 layers are oxidized, resulting in an unknown parasitic series resistance.*

Bibliography

[1] Alzate, J. G., Khalili Amiri, P., Yu, G., Upadhyaya, P., Katine, J. A., Langer, J., Ocker, B., Krivorotov, I. N., and Wang, K. L. Temperature dependence of the voltage-controlled perpendicular anisotropy in nanoscale mgo| cofeb| ta magnetic tunnel junctions. *Applied Physics Letters 104*, 11 (2014), 112410.

[2] Amara-Dababi, S., Bea, H., Sousa, R., Mackay, K., and Dieny, B. Modelling of time-dependent dielectric barrier breakdown mechanisms in mgo-based magnetic tunnel junctions. *Journal of Physics D: Applied Physics 45*, 29 (2012), 295002.

[3] Beek, S. V., Martens, K., Roussel, P., Donadio, G., Swerts, J., Mertens, S., Kar, G., Min, T., and Groeseneken, G. Four point probe ramped voltage stress as an efficient method to understand breakdown of stt-mram mgo tunnel junctions. In *2015 IEEE International Reliability Physics Symposium* (April 2015), pp. MY.4.1–MY.4.6.

[4] Beek, S. V., Martens, K., Roussel, P., Donadio, G., Swerts, J., Mertens, S., Thean, A., Kar, G., Furnemont, A., and Groeseneken, G. Voltage acceleration and pulse dependence of barrier breakdown in mgo based magnetic tunnel junctions. In *2016 IEEE International Reliability Physics Symposium (IRPS)* (April 2016), pp. MY–4–1–MY–4–4.

[5] Beek, S. V., Martens, K., Roussel, P., Wu, Y. C., Kim, W., Rao, S., Swerts, J., Crotti, D., Linten, D., Kar, G. S., and Groeseneken, G. Thermal stability analysis and modelling of advanced perpendicular magnetic tunnel junctions. *AIP Advances 8*, 5 (2018), 055909.

[6] Benard, A., and Bos-Levenbach, E. C. Het uitzetten van waarnemingen op waarschijnlijkheids-papier1. *Statistica Neerlandica 7*, 3 (1953), 163–173.

[7] Berger, L. Emission of spin waves by a magnetic multilayer traversed by a current. *Phys. Rev. B 54* (Oct 1996), 9353–9358.

[8] Bertotti, G. Magnetization dynamics. `http://ieeemagnetics.org/images/stories/SummerSchool/Assisi/Bertotti1.pdf`, 2013. Online; accessed 17 April 2018.

[9] Bloemen, P. J. H., van Kesteren, H. W., Swagten, H. J. M., and de Jonge, W. J. M. Oscillatory interlayer exchange coupling in co/ru multilayers and bilayers. *Phys. Rev. B 50* (Nov 1994), 13505–13514.

[10] Bose, A., Shukla, A. K., Konishi, K., Jain, S., Asam, N., Bhuktare, S., Singh, H., Lam, D. D., Fujii, Y., Miwa, S., et al. Observation of thermally driven field-like spin torque in magnetic tunnel junctions. *Applied Physics Letters 109*, 3 (2016), 032406.

[11] Bozorg-Grayeli, E., Li, Z., Asheghi, M., Delgado, G., Pokrovsky, A., Panzer, M., Wack, D., and Goodson, K. E. High temperature thermal properties of thin tantalum nitride films. *Applied Physics Letters 99*, 26 (2011), 261906.

[12] Brinkman, W., Dynes, R., and Rowell, J. Tunneling conductance of asymmetrical barriers. *Journal of applied physics 41*, 5 (1970), 1915–1921.

[13] Brown, W. F. Thermal fluctuations of a single-domain particle. *Phys. Rev. 130* (Jun 1963), 1677–1686.

[14] Bury, E. *Assessing bias-temperature instabilities and self-heating effects in advanced semiconductor nodes.* PhD thesis, KU Leuven, May 2016.

[15] Chang, J. Embedded meomries for energy efficient computing. Short course presentation IEDM, December 2017.

[16] Chappert, C., Fert, A., and Nguyen Van Dau, F. The emergence of spin electronics in data storage. *Nature materials 6*, 11 (2007), 813.

[17] Chaves-O'Flynn, G. D., Wolf, G., Sun, J. Z., and Kent, A. D. Thermal stability of magnetic states in circular thin-film nanomagnets with large perpendicular magnetic anisotropy. *Physical Review Applied 4*, 2 (2015), 024010.

[18] Chen, C.-Y. *Understanding and improving reliability of oxide-based resistive RAM for embedded application and storage class memory.* PhD thesis, KU Leuven, Sep 2017. Groeseneken, Guido (supervisor).

[19] Chen, I. C., Holland, S., Young, K. K., Chang, C., and Hu, C. Substrate hole current and oxide breakdown. *Applied Physics Letters 49*, 11 (1986), 669–671.

[20] Chen, Y. Y., Goux, L., Clima, S., Govoreanu, B., Degraeve, R., Kar, G. S., Fantini, A., Groeseneken, G., Wouters, D. J., and Jurczak, M. Endurance/retention trade-off on $hfo_2/metal$ cap 1t1r bipolar rram. *IEEE Transactions on Electron Devices 60*, 3 (March 2013), 1114–1121.

[21] Choi, C.-M., Sukegawa, H., Mitani, S., and Song, Y.-H. Tddb modeling depending on interfacial conditions in magnetic tunnel junctions. *Semiconductor Science and Technology 32*, 10 (2017), 105007.

[22] Chung, S. W., Kishi, T., Park, J. W., Yoshikawa, M., Park, K. S., Nagase, T., Sunouchi, K., Kanaya, H., Kim, G. C., Noma, K., Lee, M. S., Yamamoto, A., Rho, K. M., Tsuchida, K., Chung, S. J., Yi, J. Y., Kim, H. S., Chun, Y. S., Oyamatsu, H., and Hong, S. J. 4gbit density stt-mram using perpendicular mtj realized with compact cell structure. In *2016 IEEE International Electron Devices Meeting (IEDM)* (Dec 2016), pp. 27.1.1–27.1.4.

[23] Coey, J. M. *Magnetism and magnetic materials*. Cambridge University Press, 2010.

[24] Coppinger, F., Genoe, J., Maude, D. K., Gennser, U., Portal, J. C., Singer, K. E., Rutter, P., Taskin, T., Peaker, A. R., and Wright, A. C. Single domain switching investigated using telegraph noise spectroscopy: Possible evidence for macroscopic quantum tunneling. *Phys. Rev. Lett. 75* (Nov 1995), 3513–3516.

[25] Couet, S., Swerts, J., Mertens, S., Lin, T., Tomczak, Y., Liu, E., Douhard, B., Elshocht, S. V., Furnemont, A., and Kar, G. S. Oxygen scavenging by ta spacers in double-mgo free layers for perpendicular spin-transfer torque magnetic random-access memory. *IEEE Magnetics Letters 7* (2016), 1–4.

[26] Croes, K., Roussel, P., Barbarin, Y., Wu, C., Li, Y., Bömmels, J., and Tőkei, Z. Low field tddb of beol interconnects using 40 months of data. In *2013 IEEE International Reliability Physics Symposium (IRPS)* (April 2013), pp. 2F.4.1–2F.4.8.

[27] de Castro, M. M., Sousa, R. C., Bandiera, S., Ducruet, C., Chavent, A., Auffret, S., Papusoi, C., Prejbeanu, I. L., Portemont, C., Vila, L., Ebels, U., Rodmacq, B., and Dieny, B. Precessional spin-transfer switching in a magnetic tunnel junction with a synthetic antiferromagnetic perpendicular polarizer. *Journal of Applied Physics 111*, 7 (2012), 07C912.

[28] Degraeve, R. *Time dependent dielectric breakdown in thin oxides: mechanisms, statistics and oxide reliability prediction.* PhD thesis, KU Leuven, May 1998.

[29] Degraeve, R., Groeseneken, G., Bellens, R., Depas, M., and Maes, H. E. A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides. In *Proceedings of International Electron Devices Meeting* (Dec 1995), pp. 863–866.

[30] Degraeve, R., Groeseneken, G., Bellens, R., Ogier, J. L., Depas, M., Roussel, P. J., and Maes, H. E. New insights in the relation between electron trap generation and the statistical properties of oxide breakdown. *IEEE Transactions on Electron Devices 45*, 4 (Apr 1998), 904–911.

[31] Degraeve, R., Kauerauf, T., Cho, M., Zahid, M., Ragnarsson, L. A., Brunco, D. P., Kaczer, B., Roussel, P., Gendt, S. D., and Groeseneken, G. Degradation and breakdown of 0.9 nm eot sio/sub 2/ ald hfo/sub 2/metal gate stacks under positive constant voltage stress. In *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.* (Dec 2005), pp. 408–411.

[32] Degraeve, R., Roussel, P., Groeseneken, G., and Maes, H. A new analytic model for the description of the intrinsic oxide breakdown statistics of ultra-thin oxides. *Microelectronics Reliability 36*, 11 (1996), 1639 – 1642. Reliability of Electron Devices, Failure Physics and Analysis.

[33] Degrave, R., Kauerauf, T., Kerber, A., Cartier, E., Govoreanu, B., Roussel, P., Pantisano, L., Blomme, P., Kaczer, B., and Groeseneken, G. Stress polarity dependence of degradation and breakdown of sio2/high-k stacks. In *2003 IEEE International Reliability Physics Symposium Proceedings, 2003. 41st Annual.* (March 2003), pp. 23–28.

[34] Devolder, T. Scalability of magnetic random access memories based on an in-plane magnetized free layer. *Applied Physics Express 4*, 9 (2011), 093001.

[35] Devolder, T., Kim, J.-V., Garcia-Sanchez, F., Swerts, J., Kim, W., Couet, S., Kar, G., and Furnemont, A. Time-resolved spin-torque switching in mgo-based perpendicularly magnetized tunnel junctions. *Phys. Rev. B 93* (Jan 2016), 024420.

[36] Devolder, T., Kim, J.-V., Nistor, L., Sousa, R., Rodmacq, B., and Diény, B. Exchange stiffness in ultrathin perpendicularly magnetized

cofeb layers determined using the spectroscopy of electrically excited spin waves. *Journal of Applied Physics 120*, 18 (2016), 183902.

[37] DiMaria, D. J. Explanation for the polarity dependence of breakdown in ultrathin silicon dioxide films. *Applied Physics Letters 68*, 21 (1996), 3004–3006.

[38] DiMaria, D. J., and Stathis, J. H. Anode hole injection, defect generation, and breakdown in ultrathin silicon dioxide films. *Journal of Applied Physics 89*, 9 (2001), 5015–5024.

[39] Everspin Technologies. Mram applications & case studies in aerospace. `https://www.everspin.com/aerospace`, 2018. Online; accessed 7 May 2018.

[40] Everspin Technologies. Mram applications & case studies in automotive. `https://www.everspin.com/automotive`, 2018. Online; accessed 7 May 2018.

[41] Feng, X., and Visscher, P. Sweep-rate-dependent coercivity simulation of fept particle arrays. *Journal of applied physics 95*, 11 (2004), 7043–7045.

[42] Ferreira, R., Wisniowski, P., Freitas, P., Langer, J., Ocker, B., and Maass, W. Tuning of mgo barrier magnetic tunnel junction bias current for picotesla magnetic field detection. *Journal of applied physics 99*, 8 (2006), 08K706.

[43] Fourier, J. *Theorie analytique de la chaleur, par M. Fourier.* Chez Firmin Didot, père et fils, 1822. p. 54 (Chapter 1).

[44] Frieden, B. R. *Science from Fisher information: a unification.* Cambridge University Press, 2004.

[45] GlobalFoundries. Globalfoundries strengthens 22fdx emram platform with evaderis' ultra-low power mcu reference design. `https://www.globalfoundries.com/news-events/press-releases/globalfoundries-strengthens-22fdxr-emram-platform-evaderis-ultra-low`, 2018. Online; accessed 7 May 2018.

[46] Gopman, D. B., Bedau, D., Wolf, G., Mangin, S., Fullerton, E. E., Katine, J. A., and Kent, A. D. Temperature dependence of the switching field in all-perpendicular spin-valve nanopillars. *Phys. Rev. B 88* (Sep 2013), 100401.

[47] Grezes, C., Ebrahimi, F., Alzate, J., Cai, X., Katine, J., Langer, J., Ocker, B., Khalili Amiri, P., and Wang, K. Ultra-low switching

energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product. *Applied Physics Letters 108*, 1 (2016), 012403.

[48] Heh, D., Vogel, E. M., and Bernstein, J. B. Impact of substrate hot hole injection on ultrathin silicon dioxide breakdown. *Applied Physics Letters 82*, 19 (2003), 3242–3244.

[49] Herault, J., Sousa, R. C., Papusoi, C., Conraux, Y., Maunoury, C., Prejbeanu, I. L., Mackay, K., Delaet, B., Nozieres, J. P., and Dieny, B. Pulsewidth dependence of barrier breakdown in mgo magnetic tunnel junctions. *IEEE Transactions on Magnetics 44*, 11 (Nov 2008), 2581–2584.

[50] Hofmann, K., Knobloch, K., Peters, C., and Allinger, R. Comprehensive statistical investigation of stt-mram thermal stability. In *VLSI Technology (VLSI-Technology): Digest of Technical Papers, 2014 Symposium on* (2014), IEEE, pp. 1–2.

[51] Hosotani, K., Nagamine, M., Ueda, T., Aikawa, H., Ikegawa, S., Asao, Y., Yoda, H., and Nitayama, A. Effect of self-heating on time-dependent dielectric breakdown in ultrathin mgo magnetic tunnel junctions for spin torque transfer switching magnetic random access memory. *Japanese Journal of Applied Physics 49*, 4S (2010), 04DD15.

[52] Ikeda, S., Miura, K., Yamamoto, H., Mizunuma, K., Gan, H., Endo, M., Kanai, S., Hayakawa, J., Matsukura, F., and Ohno, H. A perpendicular-anisotropy cofeb–mgo magnetic tunnel junction. *Nature materials 9*, 9 (2010), 721–724.

[53] Jan, G., Thomas, L., Le, S., Lee, Y.-J., Liu, H., Zhu, J., Iwata-Harms, J., Patel, S., Tong, R.-Y., Serrano-Guisan, S., Shen, D., He, R., Haq, J., Teng, J., Lam, V., Annapragada, R., Wang, Y.-J., Zhong, T., Torng, T., and Wang, P.-K. Achieving sub-ns switching of stt-mram for future embedded llc applications through improvement of nucleation and propagation switching mechanisms. In *2016 IEEE Symposium on VLSI Technology* (June 2016), pp. 1–2.

[54] Jan, G., Thomas, L., Le, S., Lee, Y.-J., Liu, H., Zhu, J., Iwata-Harms, J., Patel, S., Tong, R.-Y., Sundar, V., Serrano-Guisan, S., Shen, D., He, R., Haq, J., Teng, J., Lam, V., Yang, Y., Wang, Y.-J., Zhong, T., Fukuzawa, H., and Wang, P.-K. Demonstration of ultra-low voltage and ultra low power stt-mram designed for compatibility with 0x node embedded llc applicaitons. In *2018 IEEE Symposium on VLSI Technology, in press* (June 2018), pp. 1–2.

[55] Jiang, X., Dubson, M. A., and Garland, J. C. Giant discrete resistance fluctuations observed in normal-metal tunnel junctions. *Phys. Rev. B 42* (Sep 1990), 5427–5432.

[56] Julliere, M. Tunneling between ferromagnetic films. *Physics Letters A 54*, 3 (1975), 225 – 226.

[57] Kaczer, B., Degraeve, R., Pangon, N., and Groeseneken, G. The influence of elevated temperature on degradation and lifetime prediction of thin silicon-dioxide films. *IEEE Transactions on Electron Devices 47*, 7 (Jul 2000), 1514–1521.

[58] Kan, J., Park, C., Ching, C., Ahn, J., Xue, L., Wang, R., Kontos, A., Liang, S., Bangar, M., Chen, H., et al. Systematic validation of 2x nm diameter perpendicular mtj arrays and mgo barrier for sub-10 nm embedded stt-mram with practically unlimited endurance. In *Electron Devices Meeting (IEDM), 2016 IEEE International* (2016), IEEE, pp. 27–4.

[59] Kan, J. J., Park, C., Ching, C., Ahn, J., Xie, Y., Pakala, M., and Kang, S. H. A study on practically unlimited endurance of stt-mram. *IEEE Transactions on Electron Devices 64*, 9 (Sept 2017), 3639–3646.

[60] Kapur, K., and Lamberson, L. *Reliability in Engineering Design John Wiley & Sons*. 1977.

[61] Kar, G. S., Kim, W., Tahmasebi, T., Swerts, J., Mertens, S., Heylen, N., and Min, T. Co/ni based p-mtj stack for sub-20nm high density stand alone and high performance embedded memory application. In *2014 IEEE International Electron Devices Meeting* (Dec 2014), pp. 19.1.1–19.1.4.

[62] Kasuya, T. A theory of metallic ferro-and antiferromagnetism on zener's model. *Progress of theoretical physics 16*, 1 (1956), 45–57.

[63] Kerber, A., Pompl, T., Rohner, M., Mosig, K., and Kerber, M. Impact of failure criteria on the reliability prediction of cmos devices with ultrathin gate oxides based on voltage ramp stress. *IEEE Electron Device Letters 27*, 7 (July 2006), 609–611.

[64] Kim, I. *Interface and Size Effects on TiN-based Nanostructured Thin Films*. PhD thesis, Texas A & M University, 2012.

[65] Kim, W., Couet, S., Swerts, J., Lin, T., Tomczak, Y., Souriau, L., Tsvetanova, D., Sankaran, K., Donadio, G. L., Crotti, D., Beek, S. V., Rao, S., Goux, L., Kar, G. S., and Furnemont, A.

Experimental observation of back-hopping with reference layer flipping by high-voltage pulse in perpendicular magnetic tunnel junctions. *IEEE Transactions on Magnetics 52*, 7 (July 2016), 1–4.

[66] Koch, R., Katine, J., and Sun, J. Time-resolved reversal of spin-transfer switching in a nanomagnet. *Physical review letters 92*, 8 (2004), 088302.

[67] Lacoste, B., de Castro, M. M., Sousa, R., Prejbeanu, I., Buda-Prejbeanu, L., Auffret, S., Ebels, U., Rodmacq, B., and Dieny, B. Control of sub-nanosecond precessional magnetic switching in stt-mram cells for sram applications. In *Memory Workshop (IMW), 2016 IEEE 8th International* (2016), IEEE, pp. 1–4.

[68] Le Goff, A., Nikitin, V., and Devolder, T. Spin-wave thermal population as temperature probe in magnetic tunnel junctions. *Journal of Applied Physics 120*, 2 (2016), 023902.

[69] Lee, K., Kan, J. J., and Kang, S. H. Unified embedded non-volatile memory for emerging mobile markets. In *2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)* (Aug 2014), pp. 131–136.

[70] Li, F.-f., Li, Z.-z., Xiao, M.-w., Du, J., Xu, W., Hu, A., and Xiao, J. Q. Bias dependent tunneling in ferromagnetic junctions and inversion of the tunneling magnetoresistance from a quantum mechanical point of view. *Journal of applied physics 95*, 11 (2004), 7243–7245.

[71] Liu, H., Bedau, D., Sun, J., Mangin, S., Fullerton, E., Katine, J., and Kent, A. Dynamics of spin torque switching in all-perpendicular spin valve nanopillars. *Journal of Magnetism and Magnetic Materials 358-359*, Supplement C (2014), 233 – 258.

[72] Lü, X. Thermal conductivity modeling of copper and tungsten damascene structures. *Journal of Applied Physics 105*, 9 (2009), 094301.

[73] Lu, Y., Zhong, T., Hsu, W., Kim, S., Lu, X., Kan, J. J., Park, C., Chen, W. C., Li, X., Zhu, X., Wang, P., Gottwald, M., Fatehi, J., Seward, L., Kim, J. P., Yu, N., Jan, G., Haq, J., Le, S., Wang, Y. J., Thomas, L., Zhu, J., Liu, H., Lee, Y. J., Tong, R. Y., Pi, K., Shen, D., He, R., Teng, Z., Lam, V., Annapragada, R., Torng, T., Wang, P. K., and Kang, S. H. Fully functional perpendicular stt-mram macro embedded in 40 nm logic for energy-efficient iot applications. In *2015 IEEE International Electron Devices Meeting (IEDM)* (Dec 2015), pp. 26.1.1–26.1.4.

[74] McPherson, J. W. Stress dependent activation energy. In *24th International Reliability Physics Symposium* (April 1986), pp. 12–18.

[75] McPherson, J. W., and Mogul, H. C. Underlying physics of the thermochemical e model in describing low-field time-dependent dielectric breakdown in sio2 thin films. *Journal of Applied Physics 84*, 3 (1998), 1513–1523.

[76] Min, T., Sun, J. Z., Beach, R., Tang, D., and Wang, P. Back-hopping after spin torque transfer induced magnetization switching in magnetic tunneling junction cells. *Journal of Applied Physics 105*, 7 (2009), 07D126.

[77] Miyakawa, N., Worledge, D. C., and Kita, K. Impact of ta diffusion on the perpendicular magnetic anisotropy of ta/cofeb/mgo. *IEEE Magnetics Letters 4* (2013), 1000104–1000104.

[78] Multiphysics, C. Comsol 5.

[79] Munira, K., and Visscher, P. B. Calculation of energy-barrier lowering by incoherent switching in spin-transfer torque magnetoresistive random-access memory. *Journal of Applied Physics 117*, 17 (2015), 17B710.

[80] Nahas, J., Andre, T., Subramanian, C., Garni, B., Lin, H., Omair, A., and Martino, W. A 4mb 0.18 mu;m 1t1mtj toggle mram memory. In *2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519)* (Feb 2004), pp. 44–512 Vol.1.

[81] NamKoong, J. H., and Lim, S. H. Temperature increase in nanostructured cells of a magnetic tunnel junction during current-induced magnetization switching. *Journal of Physics D: Applied Physics 42*, 22 (2009), 225003.

[82] Néel, L. Thermal fluctuations of a single-domain particle. *Ann. Geophys. 5* (1949), 664.

[83] Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. *Applied linear statistical models*, vol. 4. Irwin Chicago, 1996.

[84] Novotnỳ, R., Kadlec, J., and Kuchta, R. Nand flash memory organization and operations. *Journal of Information Technology & Software Engineering 5*, 1 (2015), 1.

[85] Nowak, J. J., Robertazzi, R. P., Sun, J. Z., Hu, G., Park, J. H., Lee, J., Annunziata, A. J., Lauer, G. P., Kothandaraman, R., O'Sullivan, E. J., Trouilloud, P. L., Kim, Y., and Worledge,

D. C. Dependence of voltage and size on write error rates in spin-transfer torque magnetic random-access memory. *IEEE Magnetics Letters 7* (2016), 1–4.

[86] Oh, S.-C., Park, S.-Y., Manchon, A., Chshiev, M., Han, J.-H., Lee, H.-W., Lee, J.-E., Nam, K.-T., Jo, Y., Kong, Y.-C., et al. Bias-voltage dependence of perpendicular spin-transfer torque in asymmetric mgo-based magnetic tunnel junctions. *Nature Physics 5*, 12 (2009), 898–902.

[87] Ono, K., Kawahara, T., Takemura, R., Miura, K., Yamamoto, H., Yamanouchi, M., Hayakawa, J., Ito, K., Takahashi, H., Ikeda, S., et al. A disturbance-free read scheme and a compact stochastic-spin-dynamics-based mtj circuit model for gb-scale spram. In *Electron Devices Meeting (IEDM), 2009 IEEE International* (2009), IEEE, pp. 1–4.

[88] O'Sullivan, B., Beek, S. V., Roussel, P., Rao, S., Kim, W., Couet, S., Swerts, J., Yasin, F., Crotti, D., Linten, D., and Kar, G. Extended rvs characterisation of stt-mram devices: Enabling detection of ap/p switching and breakdown. In *IEEE International Reliability Physics Symposium (IRPS)* (March 2018), pp. P–MY–5–1 – P–MY–5–6 (in press).

[89] Park, C., Kan, J. J., Ching, C., Ahn, J., Xue, L., Wang, R., Kontos, A., Liang, S., Bangar, M., Chen, H., Hassan, S., Gottwald, M., Zhu, X., Pakala, M., and Kang, S. H. Systematic optimization of 1 gbit perpendicular magnetic tunnel junction arrays for 28 nm embedded stt-mram and beyond. In *2015 IEEE International Electron Devices Meeting (IEDM)* (Dec 2015), pp. 26.2.1–26.2.4.

[90] Pinna, D., Mitra, A., Stein, D. L., and Kent, A. D. Thermally assisted spin-transfer torque magnetization reversal in uniaxial nanomagnets. *Applied Physics Letters 101*, 26 (2012), 262401.

[91] Roussel, P., Degraeve, R., Kerber, A., Pantisano, L., and Groeseneken, G. Accurate reliability evaluation of non-uniform ultrathin oxynitride and high-k layers. In *2003 IEEE International Reliability Physics Symposium Proceedings, 2003. 41st Annual.* (March 2003), pp. 29–33.

[92] Ruderman, M. A., and Kittel, C. Indirect exchange coupling of nuclear magnetic moments by conduction electrons. *Phys. Rev. 96* (Oct 1954), 99–102.

[93] Saida, D., Kashiwada, S., Yakabe, M., Daibou, T., Fukumoto, M., Miwa, S., Suzuki, Y., Abe, K., Noguchi, H., Ito, J., and

Fujita, S. 1× - to 2× -nm perpendicular mtj switching at sub-3-ns pulses below 100 μ a for high-performance embedded stt-mram for sub-20-nm cmos. *IEEE Transactions on Electron Devices 64*, 2 (Feb 2017), 427–431.

[94] Salam, G. P., Persson, M., and Palmer, R. E. Possibility of coherent multiple excitation in atom transfer with a scanning tunneling microscope. *Phys. Rev. B 49* (Apr 1994), 10655–10662.

[95] Sato, H., Yamamoto, T., Yamanouchi, M., Ikeda, S., Fukami, S., Kinoshita, K., Matsukura, F., Kasai, N., and Ohno, H. Comprehensive study of cofeb-mgo magnetic tunnel junction characteristics with single- and double-interface scaling down to 1x nm. In *2013 IEEE International Electron Devices Meeting* (Dec 2013), pp. 3.2.1–3.2.4.

[96] Schlund, B., Messick, C., Suehle, J., and Chaparala, P. A new physics-based model for time-dependent-dielectric-breakdown. In *IEEE 1995 International Integrated Reliability Workshop. Final Report* (Oct 1995), pp. 72–80.

[97] Schuegraf, K. F., and Hu, C. Hole injection sio2 breakdown model for very low voltage lifetime extrapolation. *IEEE Transactions on Electron Devices 41*, 5 (May 1994), 761–767.

[98] Shpiro, A., Levy, P. M., and Zhang, S. Self-consistent treatment of nonequilibrium spin torques in magnetic multilayers. *Phys. Rev. B 67* (Mar 2003), 104430.

[99] Slonczewski, J. Current-driven excitation of magnetic multilayers. *Journal of Magnetism and Magnetic Materials 159*, 1 (1996), L1 – L7.

[100] Song, Y. J., Lee, J. H., Shin, H. C., Lee, K. H., Suh, K., Kang, J. R., Pyo, S. S., Jung, H. T., Hwang, S. H., Koh, G. H., Oh, S. C., Park, S. O., Kim, J. K., Park, J. C., Kim, J., Hwang, K. H., Jeong, G. T., Lee, K. P., and Jung, E. S. Highly functional and reliable 8mb stt-mram embedded in 28nm logic. In *2016 IEEE International Electron Devices Meeting (IEDM)* (Dec 2016), pp. 27.2.1–27.2.4.

[101] Sousa, R. C., Kerekes, M., Prejbeanu, I. L., Redon, O., Dieny, B., Nozières, J. P., and Freitas, P. P. Crossover in heating regimes of thermally assisted magnetic memories. *Journal of Applied Physics 99*, 8 (2006), 08N904.

[102] Spaldin, N. A. *Magnetic materials: fundamentals and applications*. Cambridge University Press, 2010.

[103] Stathis, J. H. Percolation models for gate oxide breakdown. *Journal of Applied Physics 86*, 10 (1999), 5757–5766.

[104] Stoner, E. C., and Wohlfarth, E. P. A mechanism of magnetic hysteresis in heterogeneous alloys. *reprinted version of Philos. Trans. London Ser. A 240, 599 (1948), in IEEE Transactions on Magnetics 27*, 4 (July 1991), 3475–3518.

[105] Stratton, R. Volt-current characteristics for tunneling through insulating films. *Journal of Physics and Chemistry of Solids 23*, 9 (1962), 1177–1190.

[106] Suñé, J., Wu, E., and Tous, S. A physics-based deconstruction of the percolation model of oxide breakdown. *Microelectronic Engineering 84*, 9 (2007), 1917 – 1920. INFOS 2007.

[107] Suñé, J., and Wu, E. Y. Hydrogen-release mechanisms in the breakdown of thin sio_2 films. *Phys. Rev. Lett. 92* (Feb 2004), 087601.

[108] Sun, J., Robertazzi, R., Nowak, J., Trouilloud, P., Hu, G., Abraham, D., Gaidis, M., Brown, S., O'Sullivan, E., Gallagher, W., et al. Effect of subvolume excitation and spin-torque efficiency on magnetic switching. *Physical Review B 84*, 6 (2011), 064413.

[109] Sun, J. Z. Spin-current interaction with a monodomain magnetic body: A model study. *Phys. Rev. B 62* (Jul 2000), 570–578.

[110] Sune, J., and Wu, E. Y. Mechanisms of hydrogen release in the breakdown of sio/sub 2/-based gate oxides. In *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.* (Dec 2005), pp. 388–391.

[111] Swerts, J., Liu, E., Couet, S., Mertens, S., Rao, S., Kim, W., Garello, K., Souriau, L., Kundu, S., Crotti, D., Yasin, F., Jossart, N., Sakhare, S., Devolder, T., Beek, S. V., O'Sullivan, B., Elshocht, S. V., Furnemont, A., and Kar, G. S. Solving the beol compatibility challenge of top-pinned magnetic tunnel junction stacks. In *2017 IEEE International Electron Devices Meeting (IEDM)* (Dec 2017), pp. 38.6.1–38.6.4.

[112] Taiwan semiconductor manufacturing company (TSMC). Tsmc to start emram production in 2018. `https://www.mram-info.com/tsmc-start-emram-production-2018`, June 2017. Online; accessed 7 May 2018.

[113] Taniguchi, T., and Imamura, H. Thermally assisted spin transfer torque switching in synthetic free layers. *Physical Review B 83*, 5 (2011), 054432.

[114] Thiaville, A. Extensions of the geometric solution of the two dimensional coherent magnetization rotation model. *Journal of Magnetism and Magnetic Materials 182*, 1-2 (1998), 5–18.

[115] Thiaville, A. Coherent rotation of magnetization in three dimensions: A geometrical approach. *Physical Review B 61*, 18 (2000), 12221.

[116] Thomas, L., Jan, G., Le, S., Lee, Y.-J., Liu, H., Zhu, J., Serrano-Guisan, S., Tong, R.-Y., Pi, K., Shen, D., et al. Solving the paradox of the inconsistent size dependence of thermal stability at device and chip-level in perpendicular stt-mram. In *Electron Devices Meeting (IEDM), 2015 IEEE International* (2015), IEEE, pp. 26–4.

[117] Thomas, L., Jan, G., Le, S., and Wang, P.-K. Quantifying data retention of perpendicular spin-transfer-torque magnetic random access memory chips using an effective thermal stability factor method. *Applied Physics Letters 106*, 16 (2015), 162402.

[118] Tillie, L., Nowak, E., Sousa, R., Cyrille, M.-C., Delaet, B., Magis, T., Persico, A., Langer, J., Ocker, B., Prejbeanu, I., et al. Data retention extraction methodology for perpendicular stt-mram. In *Electron Devices Meeting (IEDM), 2016 IEEE International* (2016), IEEE, pp. 27–3.

[119] Tomita, H., Miwa, S., Nozaki, T., Yamashita, S., Nagase, T., Nishiyama, K., Kitagawa, E., Yoshikawa, M., Daibou, T., Nagamine, M., et al. Unified understanding of both thermally assisted and precessional spin-transfer switching in perpendicularly magnetized giant magnetoresistive nanopillars. *Applied Physics Letters 102*, 4 (2013), 042409.

[120] Tsai, M.-C., Cheng, C.-W., Tsai, C. C., and Chern, G. The intrinsic temperature dependence and the origin of the crossover of the coercivity in perpendicular mgo/cofeb/ta structures. *Journal of Applied Physics 113*, 17 (2013), 17C118.

[121] Tsunoda, K., Aoki, M., Noshiro, H., Iba, Y., Fukuda, S., Yoshida, C., Yamazaki, Y., Takahashi, A., Hatada, A., Nakabayashi, M., Tsuzaki, Y., and Sugii, T. Area dependence of thermal stability factor in perpendicular stt-mram analyzed by bi-directional data flipping model. In *2014 IEEE International Electron Devices Meeting* (Dec 2014), pp. 19.3.1–19.3.4.

[122] Vogel, E. M., Edelstein, M. D., and Suehle, J. S. Defect generation and breakdown of ultrathin silicon dioxide induced by substrate hot-hole injection. *Journal of Applied Physics 90*, 5 (2001), 2338–2346.

[123] Wang, X., Wang, Z., Hao, X., Zhou, Y., Zhang, J., Gan, H., Jung, D. H., Satoh, K., Yen, B., Malmhall, R., and Huai, Y. Different dielectric breakdown mechanisms for rf-mgo and naturally oxidized mgo. *Applied Physics Express 7*, 8 (2014), 083002.

[124] Weibull, W., et al. A statistical distribution function of wide applicability. *Journal of applied mechanics 18*, 3 (1951), 293–297.

[125] Wernsdorfer, W. Classical and quantum magnetization reversal studied in nanometer-sized particles and clusters. *Advances in Chemical Physics 118* (2001), 99–190.

[126] Worledge, D. C., and Trouilloud, P. L. Magnetoresistance measurement of unpatterned magnetic tunnel junction wafers by current-in-plane tunneling. *Applied Physics Letters 83*, 1 (2003), 84–86.

[127] Wu, E., and Sune, J. New insights in polarity-dependent oxide breakdown for ultrathin gate oxide. *IEEE Electron Device Letters 23*, 8 (Aug 2002), 494–496.

[128] Wu, E. Y., Aitken, J., Nowak, E., Vayshenker, A., Varekamp, P., Hueckel, G., McKenna, J., Harmon, D., Han, L. K., Montrose, C., and Dufresne, R. Voltage-dependent voltage-acceleration of oxide breakdown for ultra-thin oxides. In *International Electron Devices Meeting 2000. Technical Digest. IEDM (Cat. No.00CH37138)* (Dec 2000), pp. 541–544.

[129] Wu, E. Y., and Suñé, J. Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability. *Microelectronics Reliability 45*, 12 (2005), 1809 – 1834.

[130] Wu, E. Y., and Suñé, J. Generalized hydrogen release-reaction model for the breakdown of modern gate dielectrics. *Journal of Applied Physics 114*, 1 (2013), 014103.

[131] Wu, E. Y., Sune, J., and Vollertsen, R. P. Comprehensive physics-based breakdown model for reliability assessment of oxides with thickness ranging from 1 nm up to 12 nm. In *2009 IEEE International Reliability Physics Symposium* (April 2009), pp. 708–717.

[132] Wu, Y. C., Kim, W., Siddharth, R., Garello, K., Beek, S. V., Couet, S., Enlong, L., Swerts, J., Kundu, S., Souriau, L., Yasin, F., Crotti, D., Kar, G. S., Furnemont, A., Jochum, J. K., Bael, M. V., Houdt, J. V., and Groeseneken, G. Impact of temperature on the switching behavior of scaled perpendicular magnetic tunnel junction.

[133] Yakata, S., Kubota, H., Suzuki, Y., Yakushiji, K., Fukushima, A., Yuasa, S., and Ando, K. Influence of perpendicular magnetic anisotropy on spin-transfer switching current in co fe b/ mg o/ co fe b magnetic tunnel junctions. *Journal of Applied Physics 105*, 7 (2009), 07D131.

[134] Yoshida, C., and Sugii, T. Reliability study of magnetic tunnel junction with naturally oxidized mgo barrier. In *2012 IEEE International Reliability Physics Symposium (IRPS)* (April 2012), pp. 2A.3.1–2A.3.5.

[135] Yosida, K. Magnetic properties of cu-mn alloys. *Phys. Rev. 106* (Jun 1957), 893–898.

[136] Yuasa, S., Nagahama, T., Fukushima, A., Suzuki, Y., and Ando, K. Giant room-temperature magnetoresistance in single-crystal fe/mgo/fe magnetic tunnel junctions. *Nature materials 3*, 12 (2004), 868–871.

[137] Yuasa, S., Suzuki, Y., Katayama, T., and Ando, K. Characterization of growth and crystallization processes in cofeb/ mgo/ cofeb magnetic tunnel junction structure by reflective high-energy electron diffraction. *Applied Physics Letters 87*, 24 (2005), 242503.

[138] Zhang, S., Levy, P. M., and Fert, A. Mechanisms of spin-polarized current-driven magnetization switching. *Phys. Rev. Lett. 88* (May 2002), 236601.

Curriculum

Simon Van Beek
19/09/1990
Leuven, België

PhD Candidate in Engineering 10/2013 – Present
Topic: *Investigation of reliability aspects of STT-MRAM*
Supported by an IWT scholarship for strategic basic research
Device reliability and characterization (DRE), imec
Dpt. Of Electrical Engineering (ESAT), KU Leuven

Master of Science in Engineering 10/2011 – 06/2013
Nanoscience and Nanotechnology, *magna cum laude*
KU Leuven, Belgium

Bachelor of Science in Engineering 10/2008 – 06/2011
Electrical Engineering / Material Engineering, *cum laude*
KU Leuven, Belgium

Scientific contributions

Publications as first author

International journal publications

S. Van Beek, K. Martens, P. Roussel, Y. C. Wu, W. Kim, S. Rao, J. Swerts, D. Crotti, D. Linten, G. S. Kar and G. Groeseneken, "Thermal stability analysis and modelling of advanced perpendicular magnetic tunnel junctions", *AIP Advances*, vol 8, issue 5, pp. 055909-1 - 055909-6, 2018.

International conference proceedings

S. Van Beek, K. Martens, P. Roussel, G. Donadio, J. Swerts, S. Mertens, G. Kar, T. Min and G. Groeseneken, "Four point probe ramped voltage stress as an efficient method to understand breakdown of STT-MRAM MgO tunnel junctions", In Proc. IRPS, pp. MY 4.1-4.6, APR 2015.

S. Van Beek, K. Martens, P. Roussel, G. Donadio, J. Swerts, S. Mertens, A. Thean, G. Kar, A. Furnemont and G. Groeseneken, "Voltage acceleration and pulse dependence of barrier breakdown in MgO based magnetic tunnel junctions", In Proc. IRPS, pp. MY 4.1-4.4, APR 2016.

S. Van Beek, K. Martens, P. Roussel, S. Couet, L. Souriau, J. Swerts, W. Kim, S. Rao, S. Mertens, T. Lin, D. Crotti, R. Degraeve, E. Bury, D. Linten, G. Kar and G. Groeseneken, "Impact of processing and stack optimization on the reliability of perpendicular STT-MRAM", In Proc. IRPS, pp. 5A-1.1-5A-1.5, APR 2017.

S. Van Beek, P. Roussel, B. O'Sullivan, R. Degraeve, S. Cosemans, D. Linten and G. S. Kar, "Study of breakdown in STT-MRAM using ramped voltage stress and all-in-one maximum likelihood fit", In Proc. ESSDERC, SEP 2018 (accepted).

S. Van Beek, P. Roussel, B. O'Sullivan, R. Degraeve, E. Bury, J. Swerts, S. Couet, L. Souriau, S. Kundu, S. Rao, W. Kim, F. Yasin, D. Crotti, D. Linten and G. S. Kar, "Impact of self-heating on reliability predictions in STT-MRAM", In Proc. IEDM, DEC 2018 (submitted).

International conference abstracts

S. Van Beek, K. Martens, P. Roussel, W. Kim, J. Swerts, G. Donadio, G. Kar, A. Thean, A. Furnemont and G. Groeseneken, "Thermal stability extraction from single perpendicular MTJs and Mbit arrays: comparison and statistical analysis", MMM - intermag, JAN 2016.

S. Van Beek, K. Martens, P. Roussel, Y. C. Wu, W. Kim, S. Rao, J. Swerts, D. Crotti, D. Linten, G. S. Kar and G. Groeseneken, "Thermal stability analysis and modelling of advanced perpendicular magnetic tunnel junctions", MMM, NOV 2017.

Co-authored publications

International journal publications

W. Kim, S. Couet, J. Swerts, T. Lin, Y. Tomczak, L. Souria, D. Tsvetanova, K. Sankaran, G. Donadio, D. Crotti, **S. Van Beek**, L. Goux, G. Kar and A. Furnemont, "Experimental observation of back-hopping with reference layer flipping by high-voltage pulse in perpendicular magnetic tunnel junctions", *IEEE trans. on Magnetics*, vol 52, issue 7, pp. 1-4, JULY 2016.

E. Raymentants, A. Vaysset, D. Wan, J. Swerts, **S. Van Beek**, O. Zografos, D. E. Nikonov, S. Manipatruni, I. A. Young, D. Mocuta, I. P. Radu, M. Heyns and M. Manfrini, "Spin-torque-driven MTJs with extended free layer for logic applications", Journal of Applied Physics D, vol. 51, issue 27, pp. 275002, JUNE 2018.

International conference proceedings

J. Swerts, E. Liu, S. Couet, S. Mertens, S. Rao, K. Garello, L. Souriau, S. Kundu, D. Crotti, F. Yasin, N. Jossart, S. Sakhare, T. Devolder, **S. Van Beek**, B. O'Sullivan, S. Van Elshocht, A. Furnemont and G. S. Kar, "Solving the BEOL compatibility challenge of top-pinned magnetic tunnel junction stacks", In Proc. IEDM, DEC 2017.

K. Garello, F. Yasin, S. Couet, L. Souriau, J. Swerts, **S. Van Beek**, W. Kim, E. Liu, S. Kundu, D. Tsvetanova, N. Jossart, K. Croes, E. Grimaldi, M. Baumgartner, D. Crotti, A. Furnemont, P. Gambardella and G. S. Kar, "SOT-MRAM 300mm integration for low power and ultrafast embedded memories", In Proc. VLSI, JUNE 2018.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
ESAT-MICAS
Kasteelpark arenberg 10
B-3001 Leuven
simon.vanbeek@imec.be