Spin-transfer torque switching probability of CoFeB/MgO/CoFeB magnetic tunnel junctions beyond macrospin

J. Z. Sun
IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA

(Received 12 July 2021; revised 9 September 2021; accepted 17 September 2021; published 30 September 2021)

We present an empirical description of experimental spin-torque switching probability for the CoFeB/MgO/CoFeB type of magnetic tunnel junctions beyond macrospin limit, parametrizing measurement data for direct comparison with the corresponding macrospin asymptote expression. We show that, near 35 nm in diameter, spin-torque switching speed in these tunnel devices is faster than macrospin-limit predictions. These devices have a faster reduction of switching error rate versus spin-torque drive amplitude than macrospin. While the functional form similar to macrospin can still describe experimental data satisfactorily, the parameters no longer correspond to materials values. Instead they reflect the nonuniform nature of the switching process. Further, the parameters depend on the resistance-area product r_A of the junction, with higher r_A causing a steeper slope of switching error versus switching current. This r_A dependence could not originate from low-bias spin-dependent tunneling. These observations suggest that, in addition to nonuniform nonlinear dynamics during switching, it is also important to consider higher-order dynamic processes, including a high-bias tunnel electron's spin-flip scattering, voltage-induced change to interface magnetism, and possibly Joule heating.

DOI: 10.1103/PhysRevB.104.104428

I. INTRODUCTION

A spin-transfer torque (STT) switched magnetic tunnel junction (MTJ) with perpendicular magnetic anisotropy (PMA) is being pursued for magnetic random-access memory (MRAM) technology [1–12]. For these applications, the switching statistics relating to the "write-error" needs optimization to reduce the required switching current, and to improve switching speed. A physics-based quantitative description of the switching statistics at finite temperature, however, is currently limited to the so-called macrospin model, where the exchange-energy of the ferromagnet is assumed much larger than relevant energy scales involved in the dynamics. This assumption is overly simplified for most realistic devices in use today, as the corresponding magnetic exchange-length, defined as $\lambda_{\mathrm{ex}} \propto \sqrt{A_{\mathrm{ex}}/K_{\mathrm{u}}}$ [13,14], with A_{ex} the exchange-energy and K_{u} the net perpendicular anisotropy energy density, is of the same order as typical device sizes around 20 nm or larger [15,16]. Consequently, nonuniform, nonlinear dynamics play a central role, which in a certain parameter range yields switching characteristics *more desirable* than macrospin models would expect, with a lower requirement for switching spin-current for a given speed and error allowance.

Since many plausible mechanisms are at play in real-world devices, and all could potentially contribute to switching characteristics by a similar amount, a model-assumption based quantitative analysis is difficult due to the complexity of the problem, as models are severely underconstrained by measurable behavior, and causes of available observation are difficult to isolate. Instead, in this paper we introduce an empirical functional form to parametrize the experimental switching statistics of such "beyond macrospin" MTJs using a functional form convenient for direct comparison with the known macrospin asymptotic expression. The difference between the observation and the macrospin is then investigated against controllable materials and device parameters, such as the MTJ's resistance-area product r_A. The analysis provides a quantitative description of the general behavior of such beyond-macrospin devices, and it points to the roles of other mechanisms that are likely involved, including the effects from hot-electron spin-flip scattering and Joule heating.

In what follows, we review briefly the basic physics understanding of macrospin-based switching statistics. Then we describe a method for parametrizing experimental observations in a form convenient for direct comparison with macrospin. Using this quantitative method, we also examine the effect of MTJ's r_A on switching, demonstrating an r_A dependence that does not originate from simple spin-polarized tunneling. With these observations quantified, we examine the roles that a few mechanisms such as fast Joule heating and hot-electron spin-flip scattering can play that contribute to the unexpected behavior of an r_A dependence on switching statistics.

A. Macrospin model basics: A review

For spin-torque driven finite-temperature nanomagnet dynamics in switching MTJs, the constitutive relationships are by now well known [7,17–22]. The nanomagnet's dynamics is described based on the classical Landau-Lifshitz-Gilbert (LLG) equation. Analytical solutions can be obtained in a few limiting cases in a macrospin approximation. For spin-dependent tunneling, the simplest conduction models assume elastic tunneling in the low-bias voltage limit, where bias-dependent subband tunnel conductance and inelastic processes, magnetic or otherwise, are neglected for simplicity [23,24].

For an STT-switchable MTJ, the switching error probability ϵ_r (equivalently referred to, in memory technology terms, as "write error rate," or WER for short) is defined as the ratio between the number of failed switching events and the total number of switching trials for a given switching condition. The switching condition is defined by providing the MTJ with a spin-current bias I_s for a time duration of τ. Obviously, $\epsilon_r = 1 - P_{\rm sw}$, with $P_{\rm sw}$ the switching probability. To leading order, the spin-current $I_s \propto V_b$, the voltage bias across the MTJ at low voltage below ~0.2–0.4 V [19,23–26]. The spin- and charge-current in an MTJ is often written in a relationship of $I = \eta I_s$, with η the charge-to-spin current ratio of an MTJ, I the charge current, and I_s the spin-current represented in charge-current units (i.e., replacing $\hbar/2$ by electron charge e). At finite temperature T in thermal equilibrium for time $t \leqslant 0$, and with stepwise applied STT drive starting at time $t = 0$, the switching error ϵ_r of an MTJ free-layer (FL) in the macrospin-limit and with simple uniaxial anisotropy (usually perpendicular to the layer) collinear with the spin-polarization direction assumes asymptotic forms of [7,17,18,21,22,27–32]

$$\epsilon_r(\tau) \approx \begin{cases} \left(\frac{\pi^2\xi_b}{4}\right)\exp\left(-\frac{2\tau}{\tau_I}\right) + O\left[\exp\left(-\frac{\pi^2\xi_b}{4}\right)\right] & \text{(superthreshold: } I_s \gg I_{\rm sc0},\ \epsilon_r \ll 1 \text{ and } \xi_b \gg 1\text{)},\\ \exp\left\{-\left(\frac{\tau}{\tau_0}\right)\exp\left[-\xi_b\left(1-\frac{I_s}{I_{\rm sc0}}\right)^{\nu}\right]\right\} & \text{(subthreshold: } I_s \ll I_{\rm sc0},\ \epsilon_r \sim 1,\ \text{and } \xi_b \gg 1\text{)},\end{cases} \tag{1.1}$$

where $\xi_b = E_b/k_BT = mH_k/2k_BT$ is the thermal activation barrier height normalized by temperature, m is the total magnetic moment of the macrospin, $\tau_I = \tau_0/(I_s/I_{\rm sc0} - 1)$ is the characteristic timescale for STT switching above threshold, $\tau_0 \approx \hbar/2\alpha\mu_B H_k$ is related to the inverse attempt frequency, and α is the LLG-damping.

The STT instability threshold spin-current in charge-current units is $I_{\rm sc0} \triangleq \eta V_{\rm c0}/R_{\rm P} = (2e/\hbar)\alpha(mH_k)$, where $R_{\rm P}$ is the parallel (P) state MTJ low-bias resistance, and $V_{\rm c0}$ is the corresponding threshold in voltage across the MTJ. The corresponding charge-current threshold in P state is of course $I_{\rm c0} = I_{\rm sc0}/\eta$, with η the charge-to-spin current conversion ratio *in parallel state*. For MTJs with (i) symmetric tunnel interfaces, (ii) large tunnel magnetoresistance ($m_r \gg 1$), and (iii) an undisturbed reference layer tunnel electrode during switching, $\eta = \sqrt{m_r(m_r+2)}/2(m_r+1)$, with $m_r = (R_{\rm AP} - R_{\rm P})/R_{\rm P}$ the magnetoresistance ratio of the MTJ [19,22–24].

Throughout this writing, a "practical" unit set is being used, where the CGS-magnetic quantities are lumped together to yield a net energy in units of erg that is canceled by CGS-unit $\hbar$, whereas charge-related quantities are in SI units (Coulomb, Ω, Amp, or V). Length is in cm unless otherwise specified. Where possible, terms in expressions are grouped with same-unit ratios so the resulting unit would be obvious.

Strictly speaking, Eq. (1.1) only describes the probability of not switching at time τ *while* $I_{\rm s}$ *is still being applied*. The error probability of concern for memory technology is that of the end-state *after* the withdrawal of the write-pulse $I_{\rm s}$. We will neglect for now the difference between these two probabilities, but under special dynamic conditions these two probabilities can generally be different, albeit usually only by a small amount.

For subthreshold switching statistics, the exponent $\nu = 2$ for uniaxial anisotropy in collinear alignment with spin-current polarization based on an analytical solution to the corresponding Fokker-Planck equation [33–38]. It otherwise falls within a range of approximately 1–2, depending on the details of the potential shape of the anisotropy, and relative vector alignment with the spin-polarization direction, and for practical experimentally accessible parameter regions, as discussed in Refs. [7,39]. Experimentally, the value of ν is often entangled with the extracted value of $I_{\rm sc0}$, as the expression becomes linearized to a local slope due to a limited experimental time-span. The ν value in realistic MTJs depends further on details of the system's micromagnetics behavior when the junction is larger than a macrospin [40–45], and it can generally vary approximately around the range described above with some uncertainty from device to device, for different materials combinations, and for different measurement timescales involved. Since the main topic of the present work is on superthreshold high-speed switching probabilities, we will not discuss more details related to the exponent ν.

For superthreshold (fast) switching, the top line of Eq. (1.1) can be derived also from a Fokker-Planck equation related to macrospin dynamics [18,20,21,31]. Within our typical materials and device parameter space, the Fokker-Planck solution confirms the simpler picture that the switching probability distribution is primarily controlled by the distribution of the macrospin's initial angle at the moment the spin-torque I_s is turned on [7,17,18,21,22,27–32,46–48].

For a pulsed STT drive with bias voltage pulse-height $V_{\rm w}$, pulse width $\tau_{\rm w}$ in the super-threshold, and a fast-switching limit that results in a switching error of ϵ_r, Eq. (1.1)'s first line gives a characteristic charge unit of Q_0 as

$$Q_0 \triangleq \left(\frac{V_{\rm w} - V_{\rm c0}}{R_{\rm P}}\right)\tau_{\rm w} = \left(\frac{e}{\eta}\right)\left(\frac{m}{2\mu_B}\right)\ln\left(\frac{\pi^2\xi_b}{4\epsilon_r}\right), \tag{1.2}$$

with e the magnitude of an electron charge, and μ_B the Bohr magneton in the same units as m. Equation (1.2) describes a superthreshold linear relationship between the switching speed $1/\tau_{\rm w}$ and the drive amplitude $I_{\rm w}$, with a slope proportional to $1/Q_0$ determined by the error criteria ϵ_r and thermal activation energy barrier ξ_b, together with the total magnetic moment as measured by the number of Bohr magnetons in reflection of angular momentum conservation.

For MTJs with large magnetoresistance, the actual charge current passing through the MTJ would depend strongly on the relative magnetic orientation of the tunnel electrodes. However, in representing the corresponding STT spin-current, $I_{\rm s}$ and $I_{\rm sc0}$ used in Eq. (1.1) are independent of the MTJ's magnetic alignment, and they are directly associated with voltage across the MTJ, in the form of $(I_\mathrm{s}, I_\mathrm{sc0}) = \eta(V_\mathrm{w}, V_\mathrm{c0})/R_\mathrm{P}$. Values of I_s and I_sc0 thus defined represent the respective spin-current in charge-current units (i.e., replacing $\hbar/2$ by e/η), and they can differ from the actual instantaneous charge current flowing through the junction, reflecting a current spin-polarization that is generally dependent on the relative angle of the MTJ electrodes [19,23–25,49].

Writing Eq. (1.1) explicitly in the form of voltage across an MTJ, with $I_\mathrm{w} = V_\mathrm{w}/R_\mathrm{P}$,[1] the V_w versus τ_w (switching time) relationship becomes

$$V_\mathrm{w} \approx \begin{cases} V_\mathrm{c0}\left[1 + \frac{\tau_0}{2\tau_\mathrm{w}}\left(\ln \frac{\pi^2 \xi_b}{4} - \ln \epsilon_r\right)\right], & \text{superthreshold } (\tau_\mathrm{w} \ll \tau_0, V_\mathrm{w} \gg V_\mathrm{c0}, \epsilon_r \to 0^+), \\ V_\mathrm{c0}\left[1 + \frac{1}{\xi_b} \ln\left(\frac{\tau_0(1-\epsilon_r)}{\tau_\mathrm{w}}\right)\right], & \text{subthreshold } (\tau_\mathrm{w} \gg \tau_0, V_\mathrm{w} \ll V_\mathrm{c0}, \epsilon_r \to 1^-), \end{cases} \tag{1.3}$$

where V_c0 is the "antidamping" instability threshold voltage across an MTJ, defined from $I_\mathrm{c0} R_\mathrm{P}$ as

$$V_\mathrm{c0} = \left(\frac{2e}{\hbar}\right)\left(\frac{\alpha r_\mathrm{A}}{\eta}\right)(M_s t) H_k, \tag{1.4}$$

where $\xi_b = E_b/(k_B T)$ is the reduced uniaxial anisotropy energy, and r_A is the MTJ's resistance-area product, as defined earlier. A magnetic field bias, if applied along the anisotropy easy-axis, can be represented by $H_k \to H_k \pm H_\mathrm{applied}$ for its effect on instability threshold V_c0, with ± representing the two possible collinear directions.

For macrospin, $E_b = \xi_b k_B T = \frac{1}{2}(M_s t)(\frac{\pi}{4}a^2)H_k$ is the volume total anisotropy energy, assuming an MTJ with the switching nanomagnet free-layer (FL) of thickness t and diameter a, a saturation magnetization M_s, and a net uniaxial anisotropy field H_k in the film-normal direction.

B. Experimental observations beyond macrospin

For actual thin-film-based MTJs (typically of diameter $a >$ 15–20 nm) in the superthreshold, fast-switching limit, one finds that while the general form of Eq. (1.3) remains true in respective asymptotic forms, the various M_s terms in it do not correspond to simple materials parameters, but reflect the size, STT, and temperature-dependent nonuniform subvolume dynamics [40,42,44,50–52], which can become quite complex. A limited magnetic exchange-energy and exchange-length means that internal degrees of freedom of the magnetic electrodes cannot be safely ignored [42,44]. It also means that inelastic processes during tunneling may play a role in determining the charge-to-spin current conversion, such as spin-current from spin-flip scattering of tunnel electrons [49,53,54], leading to effects not captured by the macrospin model. In some cases, these can aid faster STT switching with better error statistics, which is important for applications [5,7,10,12,49,55].

Below, we show that the nonmacrospin effects of experimental MTJs in the fast switching superthreshold limit $V_\mathrm{w} \gg V_\mathrm{c0}$ can be empirically represented with the relation

$$V_\mathrm{w} \approx \left(\frac{4e}{\hbar}\right)\left(\frac{r_A}{\eta}\right)\left\{\frac{1}{2}(M_{\mathrm{s}1} t) H_k \alpha + \left(\frac{1}{8\mu_B}\right)\left(\frac{\hbar}{\tau_\mathrm{w}}\right) \times \left[(M_{\mathrm{s}2} t) \ln \frac{\pi^2 \xi_b}{4} - (M_{\mathrm{s}3} t) \ln \epsilon_r\right]\right\}, \tag{1.5}$$

which is a rewrite of Eq. (1.3)'s first line except with different moment coefficients in the three corresponding terms.

Equation (1.5) is not "derived" from theories for now, but is simply an attempt to parametrize observations we had on real devices. It is written in a way similar to the macrospin expression to easily observe the consequences of nonmacrospin behavior. The form is intentionally expanded in form for easy comparison with experimental data.

In the macrospin limit, $M_{\mathrm{s}1,2,3} = M_s$ simply reflects the switching nanomagnet's saturation magnetization, reverting Eq. (1.5) to Eq. (1.3)'s first line. In a nonmacrospin situation observed experimentally, the values of $M_{\mathrm{s}1,2,3}$ no longer simply correspond to M_s, but rather reflect the various excitations of internal dynamics.

$M_{\mathrm{s}1}$ in Eq. (1.5) relates to the V_c0 expression Eq. (1.4), reflecting the scaling of a spin-torque induced negative-damping instability threshold with lateral area of the MTJ. While the area-scaling of the STT spin-current instability threshold appears robust in larger-than-macrospin junctions [42,43], the value of M_s does not always assume the value of associated FL material, nor a simple dependence of the threshold on H_k, especially when measured by examining the applied field dependence (through $H_k \to H_k \pm H_\mathrm{applied}$) of V_c0. In such V_c0 versus H slope and intercept analysis [56], the M_s value often departs from the materials value, together with a changed apparent α (the damping parameter).[2] The incoherent nature of the regions initiating STT-switching in beyond-macrospin processes means that $M_{\mathrm{s}1}$ and its related α and H_k values are rescaled by the nucleating region (subvolume)'s dynamic exchange interaction with neighbors. As a result, they can assume values significantly different from their corresponding homogeneous materials.

$M_{\mathrm{s}2,3}$ describe the actual timescale involved in switching in the short τ_w and small ϵ_r asymptotic limit for a given spin-current density at V_w. Combined, they describe the $1/\tau_\mathrm{w}$ behavior of switching voltage increase, reflecting the thermal fluctuation-related cone-angle departure from easy-axis, and the coherent magnetic volume associated with such fluctuation [42].

$M_{\mathrm{s}2}$ describes the thermal fluctuation amplitude of the local cone-angle associated with a given thermal activation energy ξ_b measurement. Its departure from materials

[1] Here $|V_\mathrm{c0}|$ is, for simplicity, assumed symmetric for MTJ in parallel or antiparallel state [19,23–25,49].

[2] Reference [56] results were from in-plane magnetized spin-valve devices. Situations of perpendicularly magnetized MTJs in size ranges above 20 nm or so share this type of complexity in our unpublished data.

value reflects the dissociation of the local moment's thermal fluctuation amplitude and volume-integrated total anisotropy energy. The dissociation occurs when the finite wavelength spin-wave's thermal fluctuation amplitude can no longer be ignored [42,44].

M_{s3} describes the high STT-drive amplitude limit's residual error slope of $\delta \log_{10} \epsilon_r/\delta V_w$, associated mostly with the slowest transition paths occupying the switching state-space [7,17,21,31,46,47]. For thin-film MTJ devices we examined, M_{s3} is small compared to materials values [7], i.e., the actual write-error-rate (WER) slope of $\delta \log_{10} \epsilon_r/\delta V_w$ at the high V_w end is steeper (usually better for the sake of applications) than macrospin model prediction.

Note that Eq. (1.5) is an empirical attempt to "expand" the $V_w(\tau_w, \epsilon_r)$ relationship from the corresponding macrospin expression in Eq. (1.3), and that the macrospin expression Eq. (1.3) itself is only an asymptote for the limiting case of $\tau_w \ll \tau_0, V_w \gg V_{c0}, \epsilon_r \to 0^+$. For experimental comparison, such limiting conditions are at best only asymptotically approached, and they are not strictly met in most cases. This would also add some uncertainties to the apparent departure of $M_{s2,3}$ from materials parameters. The amount of uncertainties arising from the nonasymptotic nature of measurement parameters can be estimated by running the same data extraction procedure on finite-temperature macrospin model-generated WER curves. For our materials set at the measurement speeds, such an estimate points to an M_s uncertainty of typically no more than ∼15% due to departure from the asymptotic limit. It could not account for the observed departure of $M_{s2,3}$ from materials values input into the macrospin model used for generation of WER curves.

We also experimentally checked the accuracy of this asymptote analysis by examining extracted $M_{s2,3}$ values as a function of experimental write voltage pulse width τ_w. In our τ_w dependence shown for this set of MTJ layer materials, the WER slopes' extracted $M_{s2,3}$ values for $\tau_w \leqslant 10$ ns do not depend strongly on τ_w beyond data scatter, which is of the order 15–30 %. In what follows, we use $\tau_w = 10$ ns extracted $M_{s2,3}$ values for our attempt at quantitative analysis.

By the empirical nature of Eq. (1.5), the values of $M_{s2,3}$ thus obtained should only be viewed as a means to parametrize experimental observations, and to describe the next-order nonlinear dynamics related processes occurring in these MTJ FLs. They should not be overinterpreted as physical quantities without much more direct evidence. Some related considerations are discussed below.

II. EXPERIMENT

A. Brief description

Experiments are done on MTJs of the CoFeB-MgO-CoFeB type, with perpendicular magnetization [1,2], and measured at ambient temperature environment only.[3] The MTJs are patterned from thin films with sufficient perpendicular magnetic anisotropy (PMA) similar to those in Refs. [49,55]. Films used for this study have an ∼1.8-nm-thick free layer (FL) of $Co_{16}Fe_{57}B_{27}$ and a synthetic antiferromagnetic (SAF) reference layer (RL) for reduction of dipolar coupling between FL and RL. Films are sputter-deposited at ambient temperature, followed by a vacuum anneal around 400 °C for 1 h prior to optical photolithography. A reactive ion etch is used for the main junction etching step, followed by a low-energy (< 200 eV) grazing incidence Ar ion-beam etch for trimming the junction sides to the desired dimensions. The measured device diameter ranges from 15 to 150 nm as estimated from actual junction resistance as well as with scanning electron microscopy (SEM) images and occasional cross-section transmission electron microscopy (TEM) calibrations. Tunnel magnetoresistance in these devices is of the order ∼100%. For junction resistance-area product (r_A) dependence studies, multiple wafers with different MgO thickness were used to produce devices with a range of intentionally varied r_A values.

Below we present a set of analyses of the room-temperature MTJ switching probabilities as a function of pulse width τ_w and pulse height V_w. The intent is to quantify the behaviors of common MTJ switching behaviors beyond macrospin with a robust method of parametrization, and to understand the basic cause-and-effect dependences of the behavior on known and controllable device parameters. Temperature dependence studies are for now beyond the scope of our discussion, as that would involve another level of sophistication and complexity, both for experiments and for understanding. This would be better dealt with once we have the basic descriptions of room-temperature behavior at hand.

B. Switching probability, write-voltage, and pulse-width

The WER value ϵ_r is operationally defined as the number of failed write events divided by the number of write attempts at a given pulse height V_w and pulse width τ_w. Figure 1(a) shows $\log_{10} \epsilon_r$ as a function of write pulse voltage V_w for several different pulse-widths τ_w. Here, the positive V_w value corresponds to spin-torque polarity that drives a parallel to antiparallel (P-AP) transition of the MTJ, or in memory technology language, a "write-1" direction, or W1 for short; whereas a negative V_w drives an AP-P transition, or a W0. For our materials set, the W0 polarity corresponds to electrons that tunnel from RL into FL. Thus, for W0, FL is in the "downstream" direction of the electron particle current.

Typical parameter ranges are for $0 \leqslant |V_w| \leqslant 0.8$ V and for $2 \leqslant \tau_w \leqslant 500$ ns. The number of repeats was determined by the need to establish a statistical readout of ϵ_r below 10^{-6} [3,5,57], as shown in Fig. 1. Measurements at each V_w and τ_w were all done with the same constant number of repeats, around 10^6. Devices with visible anomalies are excluded from this analysis—anomalies such as reference-layer related low-$\epsilon_r - V_w$ "back-hopping" (such as shown in Ref. [9]), and defect-induced free-layer standing-wave related WER anomalies (often called "ballooning"), as shown in Refs. [5,58,59].

[3] While temperature dependence studies would open up a much wider window for new knowledge, the added complexity of temperature dependence is better addressed *after* one has a basic description of the behavior of ambient switching characteristics beyond macrospin. Ambient temperature studies also give us access to a large body of data based on technology thrusts. Therefore, here we concentrate on results obtained from ambient temperature studies.

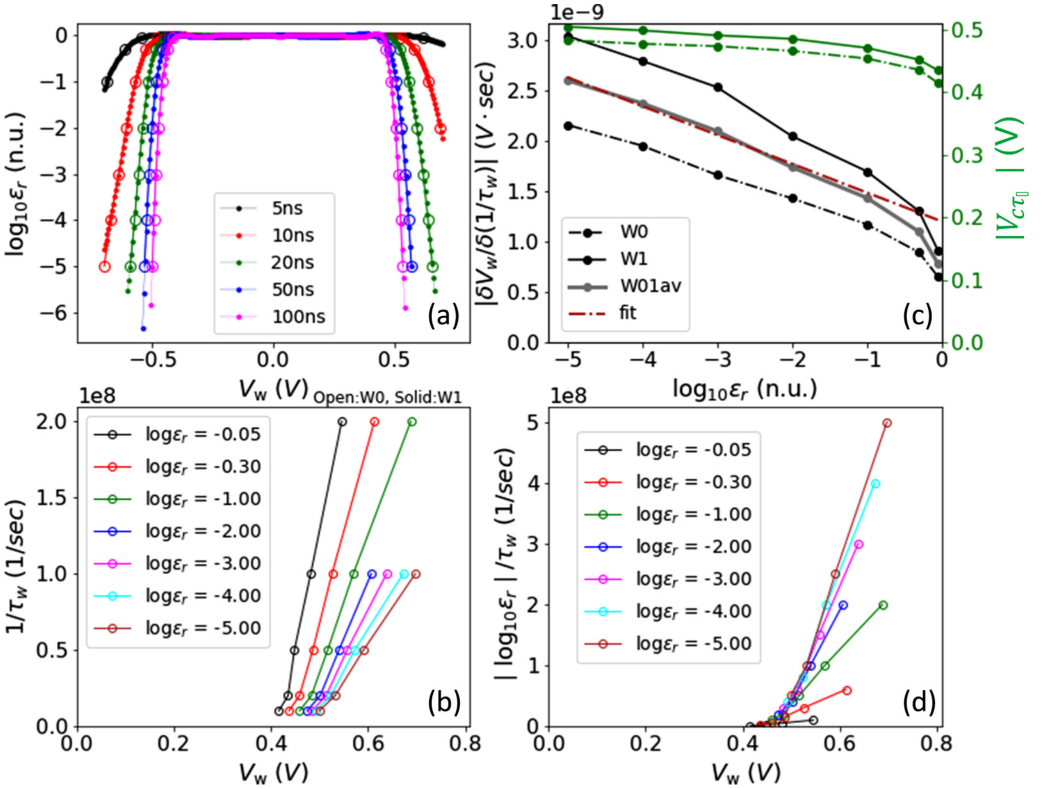

FIG. 1. A representative WER data set for a well-behaved MTJ. (a) The measured WER probability vs pulse height, at various pulse widths from 5 to 100 ns. (b) The relationship between switching speed ($1/\tau_w$, where τ_w is the pulse width) and write voltage V_w, at various error-floor criteria of $\log_{10}\epsilon_r$ of $-0.05, -0.30\ldots$ to -5.0. For clarity, only W0 direction data were shown. (c) the linear-fit slope and intercept (left and right y-axes) of data in (b) as a function of $\log_{10}\epsilon_r$. (d) A replot of data in (b) with $\log_{10}\epsilon_r/\tau_w$ as the y-axis variable. If the $\log_{10}\epsilon_r$ term dominates in Eq. (1.5), the data should collapse onto a single line. This device has an $E_b = 69\ k_BT$, $R_{\min} = 11.8\ \text{k}\Omega$, corresponding to a size of approximately 35 nm.

Below, we present *two separate analyses* for parametrizing experimental observations such as Fig. 1(a)'s based on Eq. (1.5). The first is the $1/\tau_w$ dependence of V_w for a given ϵ_r; the second is the $\log_{10}\epsilon_r$ asymptotic behavior for short pulse widths and at a high-V_w, strong STT-drive limit.

C. $1/\tau_w$ dependence of V_w at a given ϵ_r

The large open circles in Fig. 1(a) represent measured error-curve voltages of V_w corresponding to a fixed set of error rates. These map out the switching time τ_w versus write voltage V_w at different error levels in Fig. 1(b), with corresponding ϵ_r values as labels. For visual clarity, only switching data in the W0 direction are shown, and the V_w axis is plotted with magnitude only.

The slope and intercept of Fig. 1(b)'s data for different ϵ_r are extracted through a linear fit, and results are plotted in Fig. 1(c), where the left-y axis gives $|\delta V_w/\delta(1/\tau_w)|$ values, and the right-y axis gives the intercept values on the V_w axis (i.e., V_w for $\tau_w \to +\infty$, or τ_∞), both as a function of the log-error-rate $\log_{10}\epsilon_r$ as the labels show in Fig. 1(b). In Fig. 1(c), data points connected by solid lines indicate the $+V_w$ (W1) direction, and by dashed lines, the W0 direction.

For a visual check of the $|\log_{10}\epsilon_r|/\tau_w$ dependence of the functional form Eq. (1.5), Fig. 1(d) plots $|\log_{10}\epsilon_r|/\tau_w$ versus V_w. If the $\log_{10}\epsilon_r$ term in Eq. (1.5) dominates in the prefactor to $1/\tau_w$, all data for different ϵ_r would collapse onto one curve. Data in Fig. 1(d) do demonstrate such a tendency in the large $|\log_{10}\epsilon_r|$ limit. They also show that the collapse is incomplete for lower speed and smaller $|\log_{10}\epsilon_r|$ regions where thermal fluctuation related rounding becomes obvious, in the lower part of Fig. 1(d). Visually the collapse of curves in Fig. 1(d) does not occur until $\log_{10}\epsilon_r < -2$. It is consistent with the value of $\log_{10}(\pi^2\xi_b/4) \sim 2$ in Eq. (1.5) being non-negligible until at least when $\epsilon_r < 10^{-2}$. The $\delta V_w/\delta(1/\tau_w)$ versus $\log_{10}\epsilon_r$ data shown in Fig. 1(c) can give $M_{s2,3}t$ through Eq. (1.5). For simplicity, we use the bidirectional averaged amplitude of V_w in Fig. 1(c) (the gray data), and we draw a linear-fit through to obtain its slope and intercept, which relates to $(M_{s2}t)$ and $(M_{s3}t)$. The linear fit to Fig. 1(c) is taken from $\epsilon_r < 10^{-2}$, so data are close to the asymptotic limit[4] of Eq. (1.5). All other material-dependent parameters in Eq. (1.5) are derived from measured values: junction r_A is estimated from bin-average SEM-confirmed junction size [42,55] and the corresponding bin-averaged junction parallel state resistance R_P; $\eta = \sqrt{m_r(m_r+2)}/2(m_r+1)$ is obtained from junction tunnel magnetoresistance defined as $m_r = (R_{AP} - R_P)/R_P$, assuming a symmetric electrode limit at the

[4] This choice is still nonideal, limited by data availability in the deep ϵ_r region.

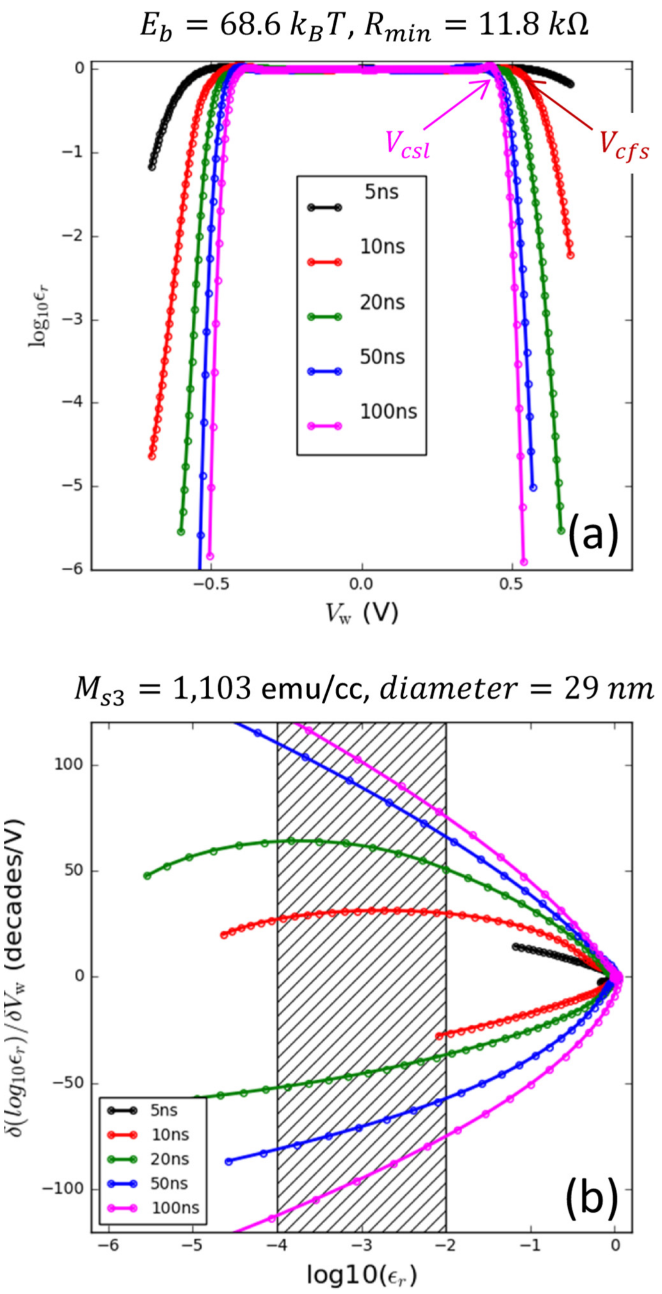

FIG. 2. (a) Raw WER traces, same as in Fig. 1(a); and (b) their derivatives giving the slope as a function of $\log_{10}\epsilon_r$. The shaded band is where the slope values are taken and averaged for a representative value of $\delta\log_{10}\epsilon_r/\delta V_w$, and the resulting M_{s3} calculated.

MgO tunnel barrier interfaces [19,22–24]. $\xi_b = E_b/k_BT$ is the reduced thermal activation barrier height derived from subthreshold pulse-width to switch voltage log-linear fit [42].[5] The resulting data across a few wafers of devices with different sizes and r_A values are shown in Figs. 3(a) and 3(b) and will be discussed later.

D. $\log_{10}\epsilon_r$ dependence of V_w at a given pulse width τ_w

In this approach, we focus on the low-error tail slope of V_w versus $\log_{10}\epsilon_r$ by taking a numerical derivative of data at the shortest available pulse-width τ_w. For consistent availability across different junctions and wafers, we focus on the $\tau_w = 10$ ns branch. This is not ideal, as $\tau_w = 10$ ns is not safely within Eq. (1.5)'s asymptotic limit of $V \gg V_{c0}$ and $\tau_w \ll \tau_0$. Experiments, however, are limited by junction breakdowns and sometimes by reference-layer instabilities at high V_w values [9,57]. We limit ourselves here to $\tau_w = 10$ ns for the maximum number of devices measurable across available size and r_A range.

This data evaluation procedure is illustrated in Fig. 2(b). The slope data are evaluated for $10^{-4} \leqslant \epsilon_r \leqslant 10^{-2}$, as shown by the shaded region, whose average value within the shaded region is then used to deduce M_{s3} using the last term of Eq. (1.5).

Threshold values V_{csl} and V_{cfs} near $\epsilon_r \sim 0.5$ of the 10 and 100 ns traces, as marked in Fig. 2(a), are compared with the first two terms of Eq. (1.5) by using

$$V_{cfs} - V_{csl} \approx \left(\frac{4e}{\hbar}\right)\left(\frac{r_A}{\eta}\right)\left(\frac{1}{8\mu_B}\right) \times \left(\frac{\hbar}{\tau_w}\right)(M_{s2}t)\ln\frac{\pi^2\xi_b}{4}. \qquad (2.1)$$

While these estimates of V_{cfs} and V_{csl} are not ideal in terms of asymptotic limit considerations, they do give an estimate to the values, and especially its systematic variations against r_A, of M_{s2} through Eq. (2.1).

E. $M_{s2,3}$ deduced from the two different methods

Figure 3 summarizes the parameters $M_{s2,3}$ deduced using the two methods described in Secs. II C and II D. Each data point is the average from the same size-bin per mask-design, with median size confirmed by scanning electron microscopy. Each size-bin contains the test results of about 10 devices. The small number of junctions each bin-size contains, and the large variation of device behavior, are responsible for data scatter. The scatter is made worse also in part by occasional junction reference-layer instability in the high-V_w, small-ϵ_r region [9,57]. Devices with gross defects of such "WER-rises" have been removed from the data pool. However, those with only a slight hint of such instabilities at the bottom of the ϵ_r are difficult to automatically catch and remove, causing some scatter in the estimate of ϵ_r versus V_w relationship. Despite such variations, some general trend can be gleaned.

Figures 3(a)–3(c) and 3(b)–3(d) give side-by-side comparisons of the $M_{s2,3}$ derived from these two different methods. Both methods yield similar values, albeit with uncertainties related to data scatter. To have a systematic comparison for same-wafer, same-device-size results, a linear interpolation is made through the data-points of various size-bins, and an interpolation value is derived for 35-nm-diam devices, shown in Figs. 3(a)–3(d) as large open circles.

[5] Joule-heating of the junction could add some error to the E_b measured this way. Our estimate of this measurement error is that it is nearly linearly increasing with r_A, and at $r_A \sim 20\ \Omega\,\mu\text{m}^2$ it gives an overestimation of E_b by about 0.05 eV for a junction of $E_b \sim 1.3$ eV, or $\sim 50\ k_BT_{\text{ambient}}$, which is not a dominant effect on r_A dependencies for the discussion below.

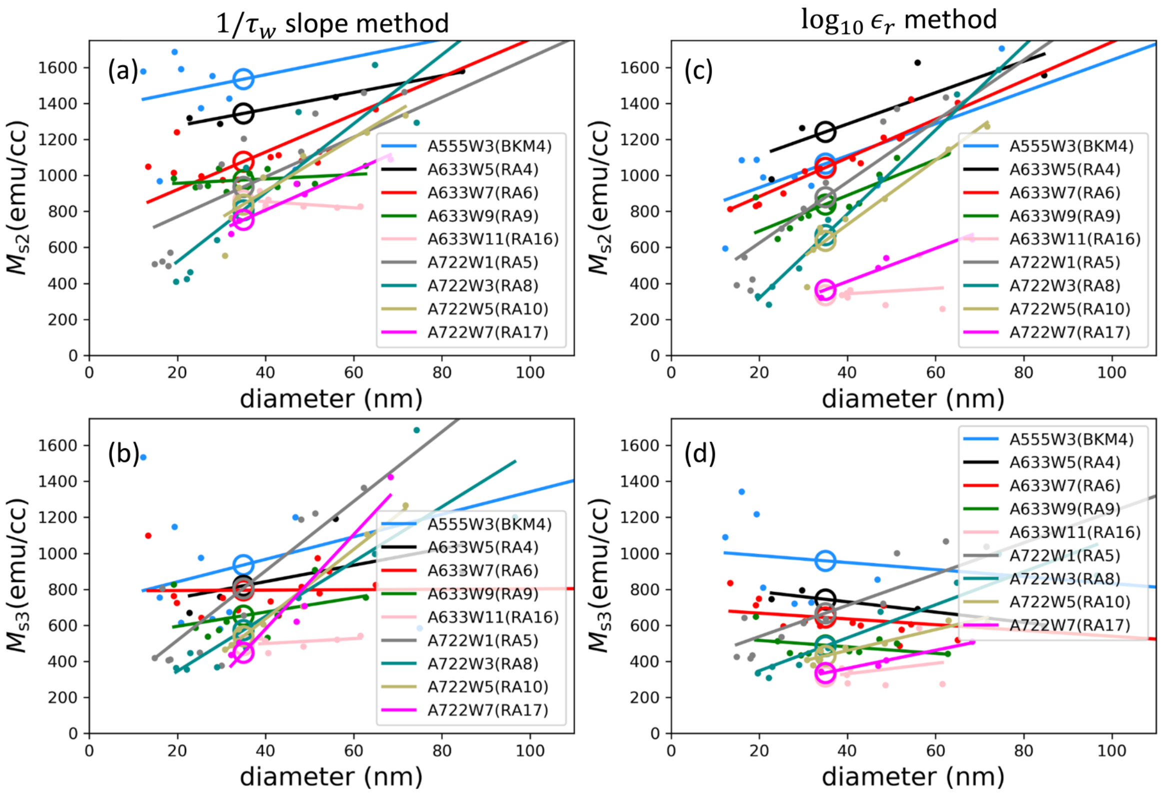

FIG. 3. (a) M_{s2} and (b) M_{s3} deduced from linear fits of $|\delta V_w/\delta(1/\tau_w)|$ vs $\log_{10}\epsilon_r$ as shown in Fig. 1(c). Each data point is the average of devices of the same size-group. Each color represents one particular sample wafer. Open circles are values at 35 nm by linear interpolation of data from various sizes as shown. The interpolation along linear regression lines shown here is done to control the leading-order experimental size-dependence, so that the average $M_{s2,3}$ values are compared at the same sample size for subsequent discussions. (c),(d) The same quantities deduced using the 10 ns $\log_{10}\epsilon_r$ slope values and thresholds, as described in Sec. II D. The inset sample labels contain the wafer-lot names. These wafers have different values of RA by design, as further discussed in Fig. 4.

The dominant dependence of $M_{s2,3}$ at 35 nm appears to be wafer-to-wafer differences, which corresponds to a junction r_A change.

F. Effect of MTJ r_A on the WER-deduced $M_{s2,3}$

The junction r_A dependence of $M_{s2,3}$ from data described in Secs. II C–II E is shown in Figs. 4(a) and 4(b), with r_A values for each corresponding wafer obtained by the measured junction resistance R_P and SEM-measured mean size-bin junction area. The $1/\tau_w$-slope based analysis is shown as open-circles, whereas the $\log_{10}\epsilon_r$ slope-based analysis is shown as solid circles. These two methods give consistent numbers at 35 nm for r_A dependence. The size 35 nm is chosen because we have the largest amount of data near this region from routine device-screening measurements. Of the data in Figs. 4(a) and 4(b), M_{s3} especially shows clearly a systematic decrease upon the increase of r_A. Data have more scatter for M_{s2}, but they remain consistent in trend with M_{s3}'s r_A dependence.

In Fig. 4(c), solid data points show the size-bin-averaged values of E_b at 35 nm junction size. They are all within a range of $55k_BT < E_b < 75k_BT$ using the same bin-size interpolation method from individual wafer data. Junction TMR-deduced spin-polarization factor η as defined by the discussion below Eq. (1.1) is shown as open circles in Fig. 4(c). These indicate that, other than a variation of r_A, these devices do not have significantly different magnetic nor tunneling characteristics.

Figure 4(d) gives the corresponding intercept-defined V_{c0av} during the linear fitting of Fig. 1(b), as shown in Fig. 1(c) by green data-points to the right-y axis. These data were taken for their W0-W1 average first, and then averaged for their values in the region of $\epsilon_r \leqslant 10^{-2}$ and over all devices in the size-bin, followed by a size-interpolation to arrive at their 35 nm values. V_{c0av} in Fig. 4(d) scales linearly with junction r_A, with a slight positive intercept at zero r_A. The intercept is small and is not much beyond the data noise limit. This linear dependence of V_{c0av} with r_A is reassuring, supporting the validity of the basic scaling relation of V_w versus r_A assumed by Eq. (1.5). Also, the finite intercept of V_{c0av} at $r_A \to 0$ in Fig. 4 is consistent with another observation of the I_{c0}-defined "switching efficiency" increasing slightly upon an increase of device r_A [55].

G. Asymmetry of the W0 and W1 direction's RA dependence

The W0,W1 asymmetry of WER slopes such as M_{s3} can reveal the upstream versus downstream tunneling related hot-electron effects. Using the $\log_{10}\epsilon_r$ versus V_w slope from the W0 and W1 branches separately gives results shown in Fig. 5 for wafer-averaged 35-nm-diam devices. Data were processed using the same methodology as in Fig. 4. The only difference

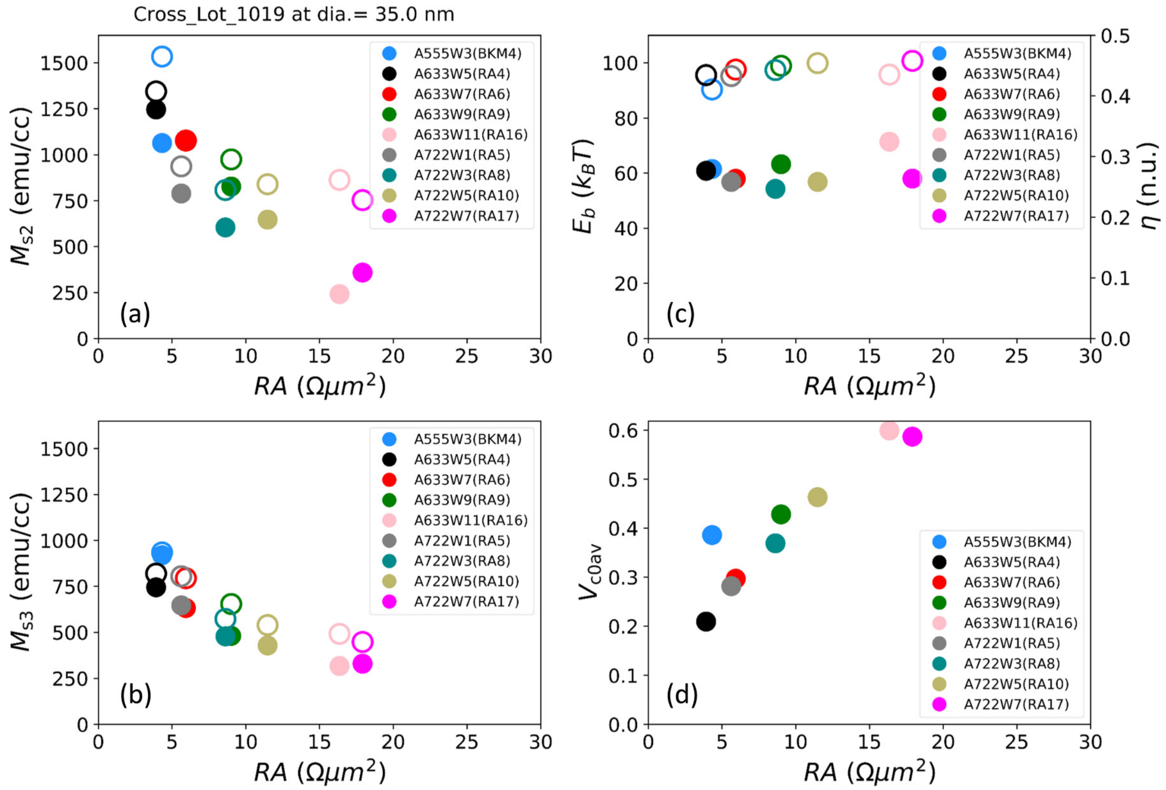

FIG. 4. (a),(b) $M_{s2,3}$ as a function of junction r_A, derived from time-domain $1/\tau_w$ slope analysis of Fig. 3 in open circles, and from $\log_{10}\epsilon_r$ slope below $\epsilon < 10^{-2}$ of 10 ns pulse width curves, in solid circles. The two methods generally yield consistent values. M_{s3} shows a better defined dependence on junction r_A, decreasing in value as r_A of the device increases. (c) The mean value of E_b for all wafers (solid circles, on the left-y scale), and the mean value of η as defined in Eq. (1.1) from junction MR (open circles, on right-y), both interpolated at 35 nm diam. (d) The average V_c values, from Fig. 1(c)'s W0-W1 averaged data with $\epsilon_r \leqslant 10^{-2}$, as a function of junction r_A, size-interpolated at 35 nm.

is that for Fig. 4 the $\log_{10}\epsilon_r$ slopes are averaged where data from both sides are available, while in Fig. 5 they are separated into W0 and W1 directions individually.

In Figs. 5(a) and 5(c), the M_{s3} values for the W0 and W1 branches were plotted separately. In Figs. 5(b) and 5(d), the ratio of M_{s3W0}/M_{s3W1} was plotted against wafer r_A.

From these results, two points can be made. (i) The leading-order process reduces M_{s3} associated with WER slope for both W0 and W1 with increasing r_A. (ii) Beyond leading order, there is a weak trend of M_{s3W0}/M_{s3W1} decreasing for higher r_A, i.e., higher r_A makes W0 easier more than for W1. This suggests that the switching efficiency gain in high-bias (high-r_A) switching is preferentially more for W0 than for W1.

This asymmetry effect of W0 versus W1 thus appears consistent with high-bias tunnel electron spin-flip related magnon spin-current enhancing STT in W0 relative to the W1 direction. This expected asymmetry has recently been experimentally observed [49], and is consistent with theories by, for example, Levy and Fert [53].

The observation of high-bias W0-W1 asymmetry, although observed, is in agreement with other asymmetry effects. First, the low-bias asymmetry of the W0 versus W1 process may include that of Joule-heating. This could result in a lower W1 (starting from R_P) state M_{s3} due to additional Joule dissipation power input. Secondly, and especially for junction sizes that are large and becoming comparable to the relevant exchange-length of the layers, edge- vs center-nucleation initiated switching dynamics could significantly affect differently the WER versus V_w slopes in the W0 and W1 directions. A signature junction diameter dependence is expected of this effect, which is indeed observed experimentally as well.

III. POSSIBLE ORIGINS OF AN r_A-DEPENDENT WER SLOPE

Within the context of Eq. (1.5) as discussed in Sec. I B, an r_A increase simply rescales V_w, and should not result in systematic reduction of $M_{s2,3}$. There might be higher-order effects of MgO tunnel barrier thickness affecting spin-polarization, but judging from the relatively constant MR of around 80–100%, and the resulting constant η of these wafers in this r_A region [Fig. 4(c)], that is unlikely a main contributor to the reduction of effective $M_{s2,3}$ at higher r_A.

Such analysis, however, has relied so far on the linear dependence between V_w and the $1/\tau_w$ term as stated by Eq. (1.5). This linear dependence may not be robust.

One way for this linear V_w versus $1/\tau_w$ dependence to break down is, for example, there is a reduction of the STT threshold V_{c0} for high $V_w \gg V_{c0}$. This can happen due to a reduction of anisotropy field H_k via Joule heating, or from a FL's magnetic moment reduction at the onset of V_w due either to Joule heating or to spin-flip scattering from tunnel electrons. Since $V_{c0} \propto M_s H_k$ in Eq. (1.4), both could lead to a reduction of V_{c0} upon application of V_w, and introducing

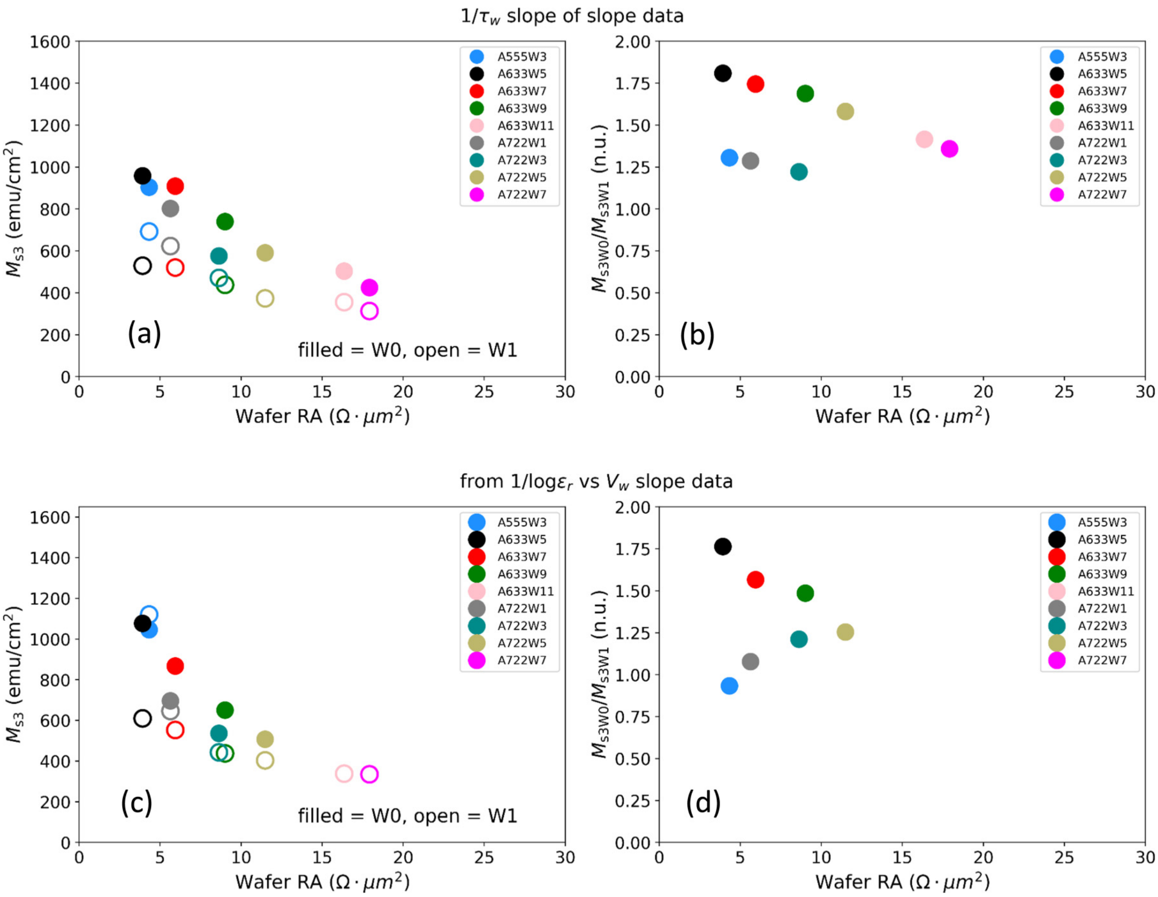

FIG. 5. W0-W1 separated values of M_{s3} as a function of junction r_A. (a) From $1/\tau_w$ slope extraction. (b) The ratio of M_{s3W0}/M_{s3W1} from (a). (c) From $\log_{10}\epsilon_r$ slope extraction. (d) The ratio from (c). Note that for sample A722W7, there is not enough voltage range in the W1 direction for deduction of M_{s3} from the $\delta\log_{10}\epsilon_r/\delta V_w$ slope as shown in (c), and thus no data point in (d).

nonlinearity into Eq. (1.5).[6] Below we examine a few of these hypotheses quantitatively.

A. Joule-heating as a possible source for V_{c0} reduction and WER improvement

The power-dissipation induced Joule-heating in our MTJs has been examined quantitatively, using the SAF reference layer's characteristic exchange field behavior in R(H) loops as a function of bias voltage,[7] and compared to the same obtained at different temperatures. A typical estimate is to have a temperature-power coefficient of the form

$$T_w = T_{amb}\left[1 + s_T\left(\frac{V_w^2}{r_A}\right)\right]. \tag{3.1}$$

For our devices near 35 nm in diameter, $s_T \approx 5.5\times10^{-8}$ (cm^2/W) [61]. This gives a junction local temperature rise of about 70 K from ambient at 0.5 V bias with an $r_A \sim 5.9\ \Omega\,\mu$m^2. Here T_{amb} is our ambient temperature around 300 K, T_w is the device-temperature at V_w, assumed to have reached thermal steady-state within the first nanosecond or so upon pulse application [62], and treated as instantaneous in discussions below.

One can also define a temperature-dependent uniaxial anisotropy field $H_k(T)$. That is,

$$H_k(T_w) = H_k(T_{amb})\left[1 - s_k\left(\frac{V_w^2}{r_A}\right)\right], \tag{3.2}$$

where T_w is the device temperature during write-current bias, and T_{amb} the ambient temperature prior to the application of write-current. The same analysis as done above gives $s_k \approx 3.9\times10^{-8}$ cm^2/W, which is further confirmed by direct ferromagnetic resonance measurements at elevated bias voltages [61]. This s_k value gives a linearly extrapolated $H_k \rightarrow 0$ temperature around 450 °C for our sample set.

B. V_w versus $1/\tau_w$ nonlinearity from Joule-heating induced V_{c0} reduction due to H_k reduction

The leading contribution of Joule heating is a V_{c0} reduction due to H_k reduction as described above. This gives a modification of the first term in Eq. (1.5), turning it into a second-order

[6] Such a description is also very approximate in nature, as Eq. (1.5) describes only a switching process that has a prethermalized initial condition. For fast Joule heating induced by bias, this initial condition is no longer precisely valid. The subsequent analysis here assumes that such fast Joule heating's modification to the process is minor compared to its effect on the material parameters such as H_k.

[7] The SAF exchange-field is one easily measurable quantity of an MTJ that is the *least* sensitive to the tunnel current's spin-current influence, thus making it a good proxy for temperature change [60].

equation for V_w, with

$$\frac{1}{2}M_sH_k \rightarrow \frac{1}{2}M_sH_k\left[1 - s_e\left(\frac{V_\text{w}^2}{r_A}\right)\right], \tag{3.3}$$

where $s_e = s_k$ if H_k is the only temperature-dependent quantity under consideration. We would also consider the effect of $M_s(T)$ on s_e later in Secs. III C and III D, in which case $s_e = s_k + s_{M_s}$.

With Eq. (3.3), one rewrites Eq. (1.5) as

$$V_\text{w} = V_{\text{c}0}\big(1 - X_{\text{s}1}V_\text{w}^2\big) + C_1Y_s - C_1(X_{\text{s}1} + X_{\text{s}2})V_\text{w}^2. \tag{3.4}$$

The various terms are defined as $C_1 = (r_A/2\eta)(e/\mu_B)(M_st/\tau_\text{w})$, $X_{\text{s}1} = s_e/r_A$, $X_{\text{s}2} = s_T/r_A$, and $Y_s = \ln(\pi^2\xi_b/4\epsilon_r)$, and $V_{\text{c}0}$ by Eq. (1.3). Here we have used the macrospin definition for $\xi_b = (1/2k_BT)M_sH_k$ and did a leading-order expansion for temperature rise at V_w as $\xi_b \rightarrow \xi_b(1 - \frac{s_e+s_T}{r_A}V_\text{w}^2)$. In doing so, we assumed a thermal and magnonic equilibrating time short compared to our pulse width, which seems reasonable for $\tau_\text{w} \geqslant 10$ ns.

With these definitions, Eq. (3.4) is solved to give

$$V_\text{w} = \frac{\sqrt{1 + 4[V_{\text{c}0}X_{\text{s}1} + C_1(X_{\text{s}1} + X_{\text{s}2})](V_{\text{c}0} + C_1Y_s)} - 1}{2V_{\text{c}0}X_{\text{s}1} + 2C_1(X_{\text{s}1} + X_{\text{s}2})} \tag{3.5}$$

The consequence of this added nonlinearity is illustrated in Fig. 6. The range of Joule-heating coefficient s_k (and corresponding s_T) encompasses the experimentally estimated value of around 4×10^{-8} W/(cm^2 K).

For a given parameter region of τ_w and ϵ_r, one takes the various slope and intercept measures of the $V_\text{w}(\tau_\text{w}, \epsilon_r)$ data and using the linear dependence of Eq. (1.5) to deduce from data the values of $M_{\text{s}2,3}$. As described around Fig. 1, for example, the "$1/\tau_\text{w}$ slope method" proceeds with $\text{SL1} = \frac{dV_\text{w}}{d(1/\tau_\text{w})}$, $\text{INT1} = V_{\text{c}0}$, $\text{SL2} = \frac{d\text{SL1}}{d\ln(1/\epsilon_r)}$, $\text{INT2} = \text{SL1} - \text{SL2}\ln(1/\epsilon_r)$.

These give

$$M_{\text{s}2} = \frac{\text{INT2}}{\left(\frac{r_A}{2\eta}\right)\left(\frac{e}{\mu_B}\right)t\ln\left(\frac{\pi^2\xi_b}{4}\right)},$$
$$M_{\text{s}3} = \frac{\text{SL2}}{\left(\frac{r_A}{2\eta}\right)\left(\frac{e}{\mu_B}\right)t}. \tag{3.6}$$

The resulting $M_{\text{s}2,3}$ versus r_A behavior is shown in Fig. 7. Also shown by dashed lines is the experimentally observed $M_{\text{s}2,3}$ reduction upon an increase in r_A. If this comparison is taken at face value, one concludes that even if one increases the Joule-heating coefficient s_k (corresponding to H_k's sensitivity to physical Joule-heating) by three times from the experimentally deduced value, one could only account for about 1/3 of the experimentally observed $M_{\text{s}2,3}$ reductions.

Such an observation remains qualitative and suggestive at present, as opposed to quantitative or fully conclusive. The quantitative detail can be affected by other practical considerations that increase complexity. A short list of such factors includes the following: First, Eq. (1.5) and its derivative Eq. (3.4) are both taken from macrospin's "short time" asymptote limits. This is generally not true for $\tau_\text{w} \sim 10$ ns. The approximation becomes more stable and less dependent on τ_w for small ϵ_r, but the residual error is uncontrolled, and its consequence on Joule-heating related modification is not quantitatively known. Secondly, Joule-heating's real-life behavior is different from simple V_w^2, as the actual power dissipation at elevated voltage is a function of the nonlinear *I*-*V* characteristics of the MTJ, which (especially in its AP state) is with a sizable voltage dependence, resulting in further modification to V_w that is not captured by the approximation leading to Eq. (3.4). Third, the amount of H_k reduction due to Joule-heating as estimated above is of a similar order of magnitude compared to voltage-induced interface magnetic anisotropy (VCMA) change at the MgO-CoFeB tunnel interface [63]. For an MgO-CoFeB FL, the two processes partially cancel each other (due to the corresponding asymmetry from Joule heating of the P and AP junction resistance difference, and the linear voltage dependence in VCMA), making the residual $H_k(V_\text{w})$ more difficult to establish. Lastly, the macrospin assumption of $\xi_b \propto H_k$ is *not true* experimentally within our experimental samples of MTJ at 35 nm. Instead, the device at 35 nm is already in (or at least near) the so-called subvolume thermal activation regime [42–44,55], where the observed ξ_b contains as much of a contribution from exchange energy A_ex as H_k. This makes the assumption $\xi_b \rightarrow \xi_b(1 - \frac{s_e+s_T}{r_A}V_\text{w}^2)$ questionable at best.

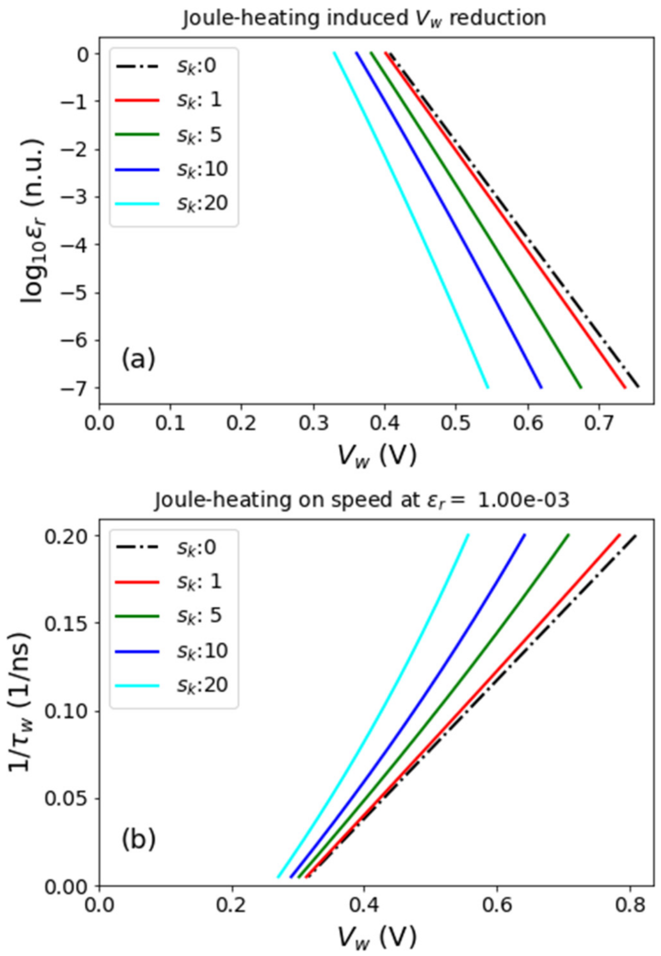

FIG. 6. A numerical illustration of Joule heating induced nonlinearity V_w from Eq. (3.4) on WER- and write-speed vs V_w behaviors. (a) WER vs V_w at different Joule heating coefficient s_k, assuming $r_A = 10\ \Omega\,\mu\text{m}^2$. (b) Switching speed $1/\tau_\text{w}$ vs V_w at $\epsilon_r = 10^{-3}$. All in the macrospin limit with $H_k = 8$ kOe.

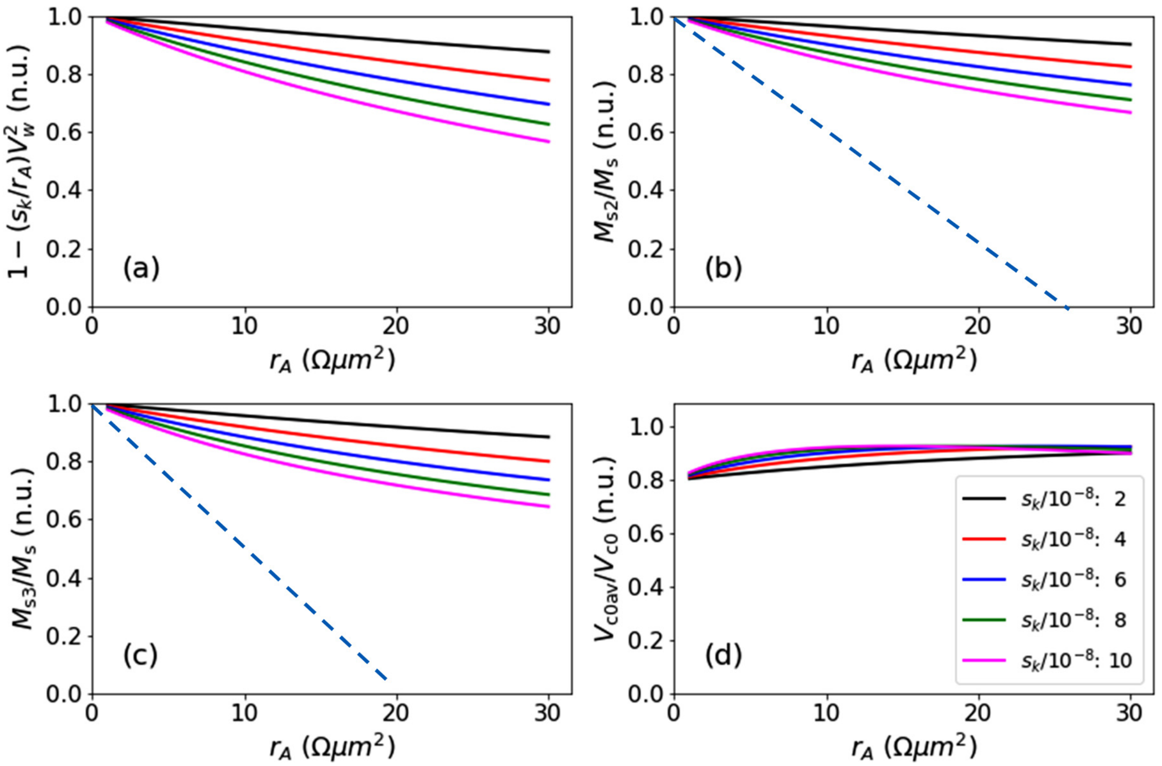

FIG. 7. Numerically emulating the analysis using the V_w expression Eq. (3.4) with Joule-heating. (a) The value of Joule-heating related M_sH_k reduction per Eq. (3.3). (b),(c) $M_{s2,3}/M_s$ as from Eq. (3.6). The dashed lines correspond to the deduced $M_{s2,3}$ trend from experimental data in Fig. 4. (d) $1/\tau_w$ vs V_w fit deduced V_{c0} intercept along V_w axis normalized to V_{c0} defined by Eq. (1.4), using the same analysis method as in Fig. 4(d).

Even with these considerations, however, it seems likely that the very simple estimate based on Eqs. (3.4) and (3.5) captures the general trend of Joule heating, and its rough order-of-magnitude effect on WER. As it stands now, since much of the cancellation effects (such as VCMA and bias-dependent r_A) is not included in the estimate, it is probable that the Joule-heating effect does not dominate the r_A dependence observed. There would be other mechanisms worth considering. Below we consider two leading terms from magnon density-related considerations.

C. The role of thermal magnons on M_s reduction at high bias

The same temperature increase described above gives an increase in the thermal magnon population, reducing M_s of the FL. In this context, a temperature rise of ΔT from ambient would result in an M_s reduction of the order

$$\Delta M_s \approx C_m \Delta T. \tag{3.7}$$

In the low-temperature limit $T \ll T_c$, with T_c being the FL's Curie temperature, the Bloch $T^{3/2}$ law [64] of $M = M_{T=0} - [\zeta(3/2)\mu_B/8\pi^{3/2}](k_BT/D_a)^{3/2}$ gives

$$C_m = \frac{dM_s(T)}{dT} = -0.176\mu_B\left(\frac{k_B\sqrt{k_BT}}{D_a^{3/2}}\right) \tag{3.8}$$

with D_a in units of erg cm^2 as the spin-wave stiffness. Since the Bloch $T^{3/2}$-law is a low-temperature limiting behavior, Eq. (3.7) at ambient temperature only gives an estimate whose accuracy would depend on whether the FL's Curie temperature satisfies $T_c \gg T$. For our typical material such as CoFeB, $D_a \sim 350$ meV Å^2 [16], giving a $C_m \sim -0.11$ emu/cm^3 K. Combining these with the estimated temperature rise from Eq. (3.1) gives

$$M_s(T_w) \approx M_{s,\mathrm{amb}}\left[1 - s_{\mathrm{Ms}}\left(\frac{V_w^2}{r_A}\right)\right], \tag{3.9}$$

with $s_{\mathrm{Ms}} \approx (C_m/M_{\mathrm{samb}})T_{\mathrm{amb}}s_T \sim 1.8\times10^{-9}$ cm^2/W. This value is an order of magnitude smaller than s_k. Thus thermal magnon-related direct M_s reduction could be safely excluded from being a contributing source of WER improvement.

In addition to Joule-heating related thermal processes, the tunnel electrons are of sufficiently high energy (~ 0.5–1 eV) to induce a significant inelastic scatter event. Chief among them is high-bias spin-flip scattering related magnon generation. This is a significant process at least at the MgO-CoFeB interface, as it severely reduces an MTJ's high-bias TMR [53,54,65].

D. Hot-electron spin-flip scattering-related magnon and related M_s reduction

We covered the spin-flip scattering process recently in a separate paper [49]. Here we only state that it is possible for spin-flip scattering of hot tunnel electrons to induce an M_s reduction of a form similar to Eq. (3.9). The same spin-flip scatter process also brings a bias-dependent tunnel magnetoresistance, which was well understood [53,54,65]. The bias-dependent TMR can be related to that of spin-flip induced magnon generation, although the relationship is complex, and involves many materials and physics related assumptions with parameters that are indirect at best.

We take an empirical approach to relate the bias-dependent TMR to spin-flip scatter induced magnon current [49,66]. We assume spin-flip scattering is the main source responsible for TMR's bias dependence. This is acceptable for a well-formed MgO-type of tunnel barrier with high TMR and in the low-r_A region of around 10 $\Omega\,\mu\text{m}^2$. The combination of high TMR and low r_A makes charge-defect related shunt conductance small when compared to the main tunnel conductance. With this assumption, one may experimentally extract a voltage-dependent P and AP state tunnel conductance of $g_{\mathrm{P,AP}}(V) \approx g_{(\mathrm{P,AP})0} + \Gamma_{\mathrm{P,AP}}(V) g_{(\mathrm{P,AP})0}$, where $g_{(\mathrm{P,AP})0}$ are the low-bias tunnel conductance, and $\Gamma_{\mathrm{P,AP}}(V)$ is the spin-flip scattering rate determined additional conduction path, which controls the leading term of voltage dependence in the tunnel conductance.

We have experimentally examined the values of $\Gamma_{\mathrm{P,AP}}(V)$ in our devices. To the leading order of V and at ambient temperature around 300 K, $\Gamma_{\mathrm{P,AP}}(V) \sim \Gamma(V) \sim \eta_m |V|$, with $\eta_m \sim 0.5$–1 (1/V) [49,66]. With it, one can relate such spin-flip scattering current to magnon-generations in the tunnel electrodes. Assuming the magnons reach steady state over a timescale fast compared to our τ_{w}, then a steady state M_s reduction can result, giving an estimate of the moment reduction in the form of

$$s_{\mathrm{Ms}} \approx \left(\frac{2\mu_B}{M_s t}\right)\left(\frac{\eta_m \tau_{\mathrm{sm}}}{e}\right), \tag{3.10}$$

where τ_{sm} is a magnon-lattice relaxation-related timescale.

To account for our observation in terms of parameters deduced in Sec. III B, one needs to have $s_{\mathrm{Ms}} \to s_k \approx (2\text{–}10)\times 10^{-8}$ cm^2/W as per descriptions surrounding Eq. (3.3). That is, one needs a $\tau_{\mathrm{sm}} \sim 0.2$–$0.5$ ns. This is a similar timescale to the long-wavelength magnon relaxation time of the order $\hbar/\alpha(2\mu_B H_k + D_a \pi^2/a^2)$, where α is the damping parameter, D_a is the exchange-stiffness parameter (around 0.2–0.3 eV Å^2, for example), and a is the junction diameter. Therefore, the spin-flip scatter induced spin-current could in principle sustain a long-wavelength magnon density sufficient to account for the demagnetization necessary to explain our r_A-dependent WER data in Sec. II F.

IV. SUMMARY

High-speed (10 ns or faster) and low error-rate ($\epsilon_r < 10^{-3}$) STT switching in an experimental CoFeB-MgO-CoFeB type of MTJs can be parametrized using an asymptotic expression, Eq. (1.5). This equation shares a form with the corresponding macrospin solution, but with different relationships of parameters. The nonmacrospin nature of devices around 35 nm in size is captured by assigning each term in Eq. (1.5) its own effective M_s values different from the material's.

From WER dependence on $V_{\mathrm{w}}/R_{\mathrm{P}}$, we deduced values of $M_{s2,3}$. $M_{s2,3}$ deduced this way show a dependence on junction r_A, reflecting a better WER performance at a given $I_{\mathrm{w}} = V_{\mathrm{w}}/R_{\mathrm{P}}$ for higher r_A MTJs.[8]

One possible cause for such a junction r_A dependence of parameters $M_{s2,3}$ is an STT threshold V_{c0} reduction upon the application of the write pulse V_{w}. Such threshold reduction leads to a nonlinear relationship between V_{w} and the $1/\tau_{\mathrm{w}}$-dependent terms in Eq. (1.5) in the form of Eqs. (3.4) and (3.5). This nonlinear relationship can cause the deduced $M_{s2,3}$ values to depend on r_A.

Such STT threshold V_{c0} reduction could be caused either by junction heating-related H_k reduction, or by spin-flip scattering-related M_s reduction due to a large amount of magnon generation by hot tunnel electrons, or more likely by both. Spin-flip induced magnons can have about the right density to account for the observed bias dependence, and give a W0/W1 asymmetry that is consistent with observations.

Beside effects from spin-flip scattering and Joule heating, there are other contributing factors worthy of consideration that have not been convincingly ruled out. For example, another factor involved is voltage-controlled magnetic anisotropy (VCMA) [63,67,68]. This usually has a linear V-dependence across $V_{\mathrm{w}} = 0$ to the leading order, and at the Fe-MgO interface it is reported [63] to cause an increase in interface perpendicular anisotropy in the "W1" bias voltage direction in our convention for the FL. Thus for W1 bias direction there is a V_{w}-dependent increase in H_k. The same sign for VCMA is observed in stacks more similar to our current geometries [68]. This sign of VCMA-related H_k change is in the opposite direction to Joule-heating, since the W1 direction starts with a lower junction resistance (in P state), and it causes relatively more Joule heating, resulting in more H_k reduction. Just how these two effects balance out is not quantitatively known, as are considerations of the added nonequilibrium initial condition's effect on STT-driven switching.

In addition to VCMA, the FL's exchange energy A_{ex} has recently been postulated to carry a linear V_{w} dependence across $V_{\mathrm{w}} = 0$ as well [69]. This in combination with VCMA will likely give rise to a more complex dependence of the exchange length λ_{ex} on V_{w}, affecting our WER discussion, too.

A clear resolution of these factors will likely require better refined sample control as well as a more definitive experimental design, as VCMA is highly sensitive to interface conditions [70], and its competition with A_{ex}'s V_{w} dependence can lead to different behaviors at different junction diameters. A quantitative understanding of WER at elevated V_{w} much above 0.5 V is still to be developed for real-life devices. The methodology of parametrizing experimental switching characteristics discussed in this paper takes a step in that direction.

ACKNOWLEDGMENTS

Work done with the MRAM group at IBM T. J. Watson Research Center was supported in part by a partnership with Samsung Electronics. I especially thank Janusz Nowak and Eric Edwards for valuable discussions, and for making experimental data available for this work.

[8] We are not advocating the use of high r_A as a technology solution for better MTJ performances. Rather, we are emphasizing the importance of understanding this r_A dependence as it is unexpected from macrospin and simple elastic spin-dependent tunneling, and it may reveal other physical mechanisms present that could help with switching current reduction.

[1] D. C. Worledge, G. Hu, D. W. Abraham, J. Z. Sun, P. L. Trouilloud, J. Nowak, S. Brown, M. C. Gaidis, E. J. O'Sullivan, and R. P. Robertazzi, Appl. Phys. Lett. **98**, 022501 (2011).

[2] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, Nat. Mater. **9**, 721 (2010).

[3] J. J. Nowak, R. P. Robertazzi, J. Z. Sun, G. Hu, D. W. Abraham, P. L. Trouilloud, S. Brown, M. C. Gaidis, E. J. O'Sullivan, W. J. Gallagher, and D. C. Worledge, IEEE Magn. Lett. **2**, 3000204 (2011).

[4] M. Krounbi, V. Nikitin, D. Apalkov, J. Lee, X. Tang, R. Beach, D. Erickson, and E. Chen, ECS Trans. **69**, 119 (2015).

[5] J. J. Nowak, R. P. Robertazzi, J. Z. Sun, G. Hu, J. H. Park, J. H. Lee, A. J. Annunziata, G. P. Lauer, C. Kothandaraman, E. J. O'Sullivan, P. L. Trouilloud, Y. Kim, and D. C. Worledge, IEEE Magn. Lett. **7**, 3102604 (2016).

[6] A. D. Kent and D. C. Worledge, Nat. Nanotechnol. **10**, 187 (2015).

[7] J. Z. Sun, Proc. SPIE **9931**, 993113 (2016).

[8] D. Apalkov, B. Dieny, and J. M. Slaughter, Proc. IEEE **104**, 1796 (2016).

[9] G. Jan, L. Thomas, S. Le, Y.-J. Lee, H. Liu, J. Zhu, J. Iwata-Harms, S. Patel, R.-Y. Tong, V. Sundar, S. Serrano-Guisan, D. Shen, R. He, J. Haq, Z. J. Teng, V. Lam, Y. Yang, Y.-J. Wang, T. Zhong, H. Fukuzawa *et al.*, in *Proceedings of the 2018 IEEE Symposium on VLSI Technology* (IEEE, Piscataway, NJ, 2018), pp. 65–66.

[10] G. Hu, J. J. Nowak, M. G. Gottwald, J. Z. Sun, D. Houssameddine, J. Bak, S. L. Brown, P. Hashemi, Q. He, J. Kim, C. Kothandaraman, G. Lauer, H. K. Lee, T. Suwannasiri, P. L. Trouilloud, and D. C. Worledge, IEEE Magn. Lett. **10**, 4504304 (2019).

[11] L. Thomas, G. Jan, S. Serrano-Guisan, H. Liu, J. Zhu, Y.-J. Lee, S. Le, J. Iwata-Harms, R.-Y. Tong, S. Patel, V. Sundar, D. Shen, Y. Yang, R. He, J. Haq, Z. Teng, V. Lam, P. Liu, Y.-J. Wang, T. Zhong *et al.*, in *Proceedings of the 2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Piscataway, NJ, 2018), pp. 27.3.1–27.3.4.

[12] E. R. J. Edwards, G. Hu, S. L. Brown, C. P. D'Emic, M. G. Gottwald, P. Hashemi, H. Jung, J. Kim, G. Lauer, J. J. Nowak, J. Z. Sun, T. Suwannasiri, P. L. Trouilloud, S. Woo, and D. C. Worledge, in *Proceedings of the 2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Piscataway, NJ, 2020), pp. 24.4.1–24.4.4.

[13] C. A. F. Vaz, J. A. C. Bland, and G. Lauhoff, Rep. Prog. Phys. **71**, 056501 (2008).

[14] C. L. Dennis, R. P. Borges, L. D. Buda, U. Ebels, J. F. Gregg, M. Hehn, E. Jouguelet, K. Ounadjela, I. Petej, I. L. Prejbeanu, and M. J. Thornton, J. Phys.: Condens. Matter **14**, R1175 (2002).

[15] M. Yamanouchi, A. Jander, P. Dhagat, S. Ikeda, F. Matsukura, and H. Ohno, IEEE Magn. Lett. **2**, 3000304 (2011).

[16] C. J. Safranski, Y.-J. Chen, I. N. Krivorotov, and J. Z. Sun, Appl. Phys. Lett. **109**, 132408 (2016).

[17] J. Z. Sun, IBM Internal Memo (2006).

[18] J. He, J. Z. Sun, and S. Zhang, J. Appl. Phys. **101**, 09A501 (2007).

[19] J. C. Slonczewski, Phys. Rev. B **71**, 024411 (2005).

[20] W. H. Butler, T. Mewes, C. K. A. Mewes, P. B. Visscher, W. H. Rippard, S. E. Russek, and R. Heindl, IEEE Trans. Magn. **48**, 4684 (2012).

[21] H. Liu, D. Bedau, J. Z. Sun, S. Mangin, E. E. Fullerton, J. A. Katine, and A. D. Kent, J. Magn. Magn. Mater. **358-359**, 233 (2014).

[22] J. Z. Sun, in *Handbook of Spintronics*, edited by Y. Xu, D. D. Awschalom, and J. Nitta (Springer, Berlin, Heidelberg, 2016), Vol. XII.

[23] J. Z. Sun and D. C. Ralph, J. Magn. Magn. Mater. **320**, 1227 (2008).

[24] J. C. Slonczewski and J. Z. Sun, J. Magn. Magn. Mater. **310**, 169 (2007).

[25] J. C. Sankey, Y.-T. Cui, J. Z. Sun, J. C. Slonczewski, R. A. Buhrman, and D. C. Ralph, Nat. Phys. **4**, 67 (2008).

[26] C. Wang, Y.-T. Cui, J. A. Katine, R. A. Buhrman, and D. C. Ralph, Nat. Phys. **7**, 496 (2011).

[27] J. Z. Sun, T. S. Kuan, J. A. Katine, and R. H. Koch, Proc. SPIE **5359**, 445 (2004).

[28] J. Z. Sun, IBM J. Res. Dev. **50**, 81 (2006).

[29] J. Z. Sun, in *Spin Angular Momentum Transfer in Magnetoresistive Nanojunctions*, edited by H. Kronmüller and S. Parkin, Spintronics and Magnetoelectronics Vol. 5 (John Wiley & Sons, Chichester, 2007).

[30] H. Tomita, T. Nozaki, T. Seki, T. Nagase, K. Nishiyama, E. Kitagawa, M. Yoshikawa, T. Daibou, M. Nagamine, T. Kishi, S. Ikegawa, N. Shimomura, H. Yoda, and Y. Suzuki, IEEE Trans. Magn. **47**, 1599 (2011).

[31] H. Liu, D. Bedau, J. Z. Sun, S. Mangin, E. E. Fullerton, J. A. Katine, and A. D. Kent, Phys. Rev. B **85**, 220405(R) (2012).

[32] A. D. Kent, H. Ohldag, H. A. Dürr, and J. Z. Sun, in *Handbook of Magnetism and Magnetic Materials*, edited by M. Coey and S. Parkin (Springer, Cham, 2021).

[33] D. M. Apalkov and P. B. Visscher, Phys. Rev. B **72**, 180405(R) (2005).

[34] D. M. Apalkov and P. B. Visscher, J. Magn. Magn. Mater. **286**, 370 (2005).

[35] P. B. Visscher and D. M. Apalkov, J. Appl. Phys. **99**, 08G513 (2006).

[36] T. Taniguchi and H. Imamura, Phys. Rev. B **83**, 054432 (2011).

[37] T. Taniguchi and H. Imamura, Phys. Rev. B **85**, 184403 (2012).

[38] K. A. Newhall and E. Vanden-Eijnden, J. Appl. Phys. **113**, 184105 (2013).

[39] D. Pinna, A. D. Kent, and D. L. Stein, Phys. Rev. B **88**, 104405 (2013).

[40] L. Thomas, G. Jan, S. Le, and P.-K. Wang, Appl. Phys. Lett. **106**, 162402 (2015).

[41] E. Hirayama, H. Sato, S. Kanai, F. Matsukura, and H. Ohno, IEEE Magn. Lett. **7**, 3104004 (2016).

[42] J. Z. Sun, R. P. Robertazzi, J. Nowak, P. L. Trouilloud, G. Hu, D. W. Abraham, M. C. Gaidis, S. L. Brown, E. J. O'Sullivan, W. J. Gallagher, and D. C. Worledge, Phys. Rev. B **84**, 064413 (2011).

[43] J. Z. Sun, P. L. Trouilloud, M. J. Gajek, J. Nowak, R. P. Robertazzi, G. Hu, D. W. Abraham, M. C. Gaidis, S. L. Brown, E. J. O'Sullivan, W. J. Gallagher, and D. C. Worledge, J. Appl. Phys. **111**, 07C711 (2012).

[44] J. Z. Sun, S. L. Brown, W. Chen, E. A. Delenia, M. C. Gaidis, J. Harms, G. Hu, X. Jiang, R. Kilaru, W. Kula, G. Lauer, L. Q. Liu, S. Murthy, J. Nowak, E. J. O'Sullivan, S. S. P. Parkin, R. P. Robertazzi, P. M. Rice, G. Sandhu, T. Topuria *et al.*, Phys. Rev. B **88**, 104426 (2013).

[45] L. Desplat and J.-V. Kim, Phys. Rev. Lett. **125**, 107201 (2020).

[46] D. Bedau, H. Liu, J.-J. Bouzaglou, A. D. Kent, J. Z. Sun, J. Katine, E. E. Fullerton, and S. Mangin, Appl. Phys. Lett. **96**, 022514 (2010).

[47] D. Bedau, H. Liu, J. Z. Sun, J. A. Katine, E. E. Fullerton, S. Mangin, and A. D. Kent, Appl. Phys. Lett. **97**, 262502 (2010).

[48] H. Liu, D. Bedau, D. Backes, J. A. Katine, and A. D. Kent, Appl. Phys. Lett. **101**, 032403 (2012).

[49] J. Z. Sun, Phys. Rev. B **103**, 094439 (2021).

[50] H. Sato, M. Yamanouchi, K. Miura, S. Ikeda, R. Koizumi, F. Matsukura, and H. Ohno, IEEE Magn. Lett. **3**, 3000204 (2012).

[51] G. D. Chaves-O'Flynn, E. Vanden-Eijnden, D. L. Stein, and A. D. Kent, J. Appl. Phys. **113**, 023912 (2013).

[52] G. D. Chaves-O'Flynn, G. Wolf, J. Z. Sun, and A. D. Kent, Phys. Rev. Appl. **4**, 024010 (2015).

[53] P. M. Levy and A. Fert, Phys. Rev. B **74**, 224446 (2006).

[54] T. Balashov, A. F. Takacs, M. Dane, A. Ernst, P. Bruno, and W. Wulfhekel, Phys. Rev. B **78**, 174404 (2008).

[55] J. Z. Sun, Phys. Rev. B **96**, 064437 (2017).

[56] J. Z. Sun, D. J. Monsma, T. S. Kuan, M. J. Rooks, D. W. Abraham, B. Oezyilmaz, A. D. Kent, and R. H. Koch, J. Appl. Phys. **93**, 6859 (2003).

[57] J. J. Nowak, G. Hu, M. G. Gottwald, R. Robertazzi, P. L. Trouilloud, Y. Kim, E. O'Sullivan, R. Kothandaraman, B. Doris, and J. Sun (unpublished).

[58] T. Min, Q. Chen, R. Beach, G. Jan, C. Horng, W. Kula, T. Torng, R. Tong, T. Zhong, D. Tang, P. Wang, M. Min Chen, J. Z. Sun, J. K. Debrosse, D. C. Worledge, T. M. Maffitt, and W. J. Gallagher, IEEE Trans. Magn. **46**, 2322 (2010).

[59] E. R. Evarts, R. Heindl, W. H. Rippard, and M. R. Pufall, Appl. Phys. Lett. **104**, 212402 (2014).

[60] A. Chavent, C. Ducruet, C. Portemont, L. Vila, J. Alvarez-Hérault, R. Sousa, I. L. Prejbeanu, and B. Dieny, Phys. Rev. Appl. **6**, 034003 (2016).

[61] P. L. Trouilloud, G. P. Lauer, C. Safranski, J. Z. Sun, G. Hu, and D. C. Worledge (private communication).

[62] M. Kerekes, R. C. Sousa, I. L. Prejbeanu, O. Redon, U. Ebels, C. Baraduc, B. Dieny, J.-P. Noziéres, P. P. Freitas, and P. Xavier, J. Appl. Phys. **97**, 10P501 (2005).

[63] A. Rajanikanth, T. Hauet, F. Montaigne, S. Mangin, and S. Andrieu, Appl. Phys. Lett. **103**, 062402 (2013).

[64] C. Kittel, in *Introduction to Solid State Physics*, 6th ed. (John Wiley & Sons, Chichester, 1986), p. 435.

[65] S. Zhang, P. M. Levy, A. C. Marley, and S. S. P. Parkin, Phys. Rev. Lett. **79**, 3744 (1997).

[66] J. Z. Sun, P. L. Trouilloud, G. P. Lauer, and P. Hashemi, AIP Adv. **9**, 015002 (2019).

[67] T. Maruyama, Y. Shiota, T. Nozaki, K. Ohta, N. Toda, M. Mizuguchi, A. Tulapurkar, T. Shinjo, M. Shiraishi, S. Mizukami, Y. Ando, and Y. Suzuki, Nat. Nanotechnol. **4**, 158 (2009).

[68] S. Kanai, M. Gajek, D. C. Worledge, F. Matsukura, and H. Ohno, Appl. Phys. Lett. **105**, 242409 (2014).

[69] T. Dohi, S. Kanai, F. Matsukura, and H. Ohno, Appl. Phys. Lett. **111**, 072403 (2017).

[70] M. K. Niranjan, C.-G. Duan, S. S. Jaswal, and E. Y. Tsymbal, Appl. Phys. Lett. **96**, 222504 (2010).