University of Pittsburgh

Loadsa[1] : A Yield-Driven Top-Down Design Method for STT-RAM Array

Wujie Wen, Yaojun Zhang, Lu Zhang and Yiran Chen

University of Pittsburgh

Loadsa: a slang language means "lots of "

Outline

- **Introduction**
  - **STT-RAM basics**
  - **Statistical design challenges**
- **Overview of top-down statistical method-*Loadsa***
- **Hierarchical semi-analytical model of *Loadsa***
  - **Generic yield mapping model for ECC/Red.**
  - **Statistical failure-probability model for STT-RAM cell**
- **Case study of *Loadsa*: Yield-Driven Array Opt.**
- **Conclusions**

Memory Technologies

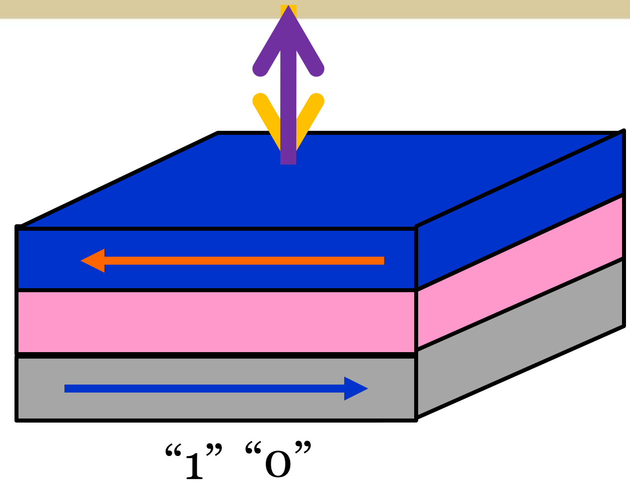
	SRAM	DRAM	NAND Flash	STT-RAM	PCRAM	R-RAM	MRAM
Data Retention	N	N	Y	Y	Y	Y	Y
Memory Cell Factor (F^2)	50-120	6-10	2-5	4-20	6-12	<1	16-40
Read Time (ns)	1	30	50	2-20	20-50	<50	3-20
Write /Erase Time (ns)	1	50	10^6-10^5	20	50-120	<100	3-20
Number of Rewrites	10^{16}	10^{16}	10^5	10^{15}	10^{10}	10^{15}	10^{15}
Power Consumption – Read/Write	Low	Low	High	Low	Low	Low	Med/ High
Power Consumption – Other than R/W	Leakage Current	Refresh Power	None	None	None	None	None

- **Spin-Transfer Torque RAM(STT-RAM), a promising candidate for future universal memory technologies.**
- **Combing the speed of SRAM, the density of DRAM, and the non-volatility of Flash.**

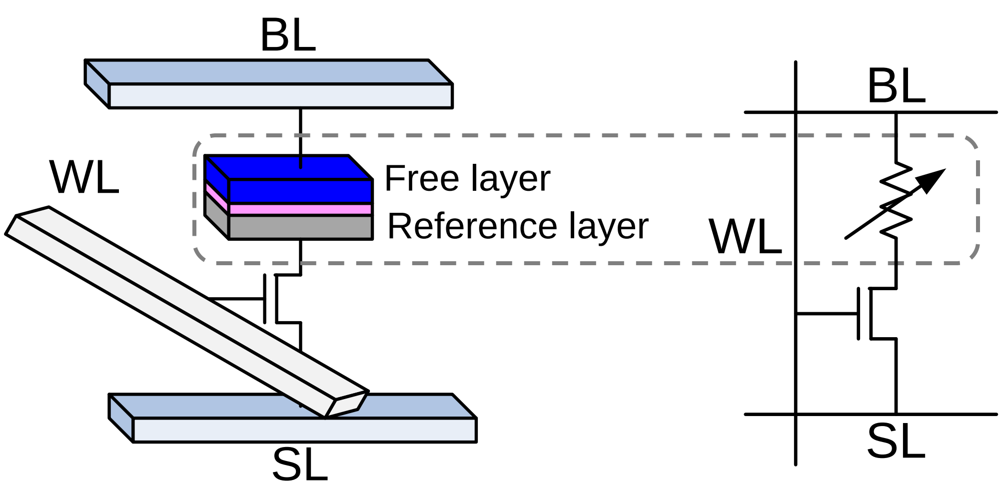
Reference: ITRS 2009

STT-RAM basics

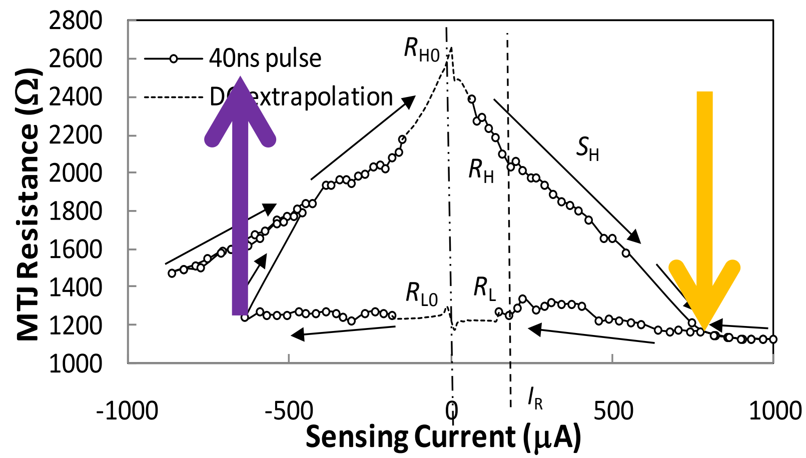
The nonvolatile data storage device in an STT-RAM cell is MTJ.



MTJ –Magnetic Tunneling Junction



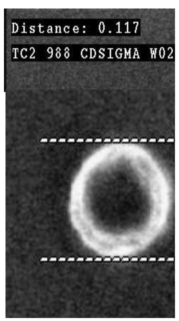
1T-1MTJ Schematic



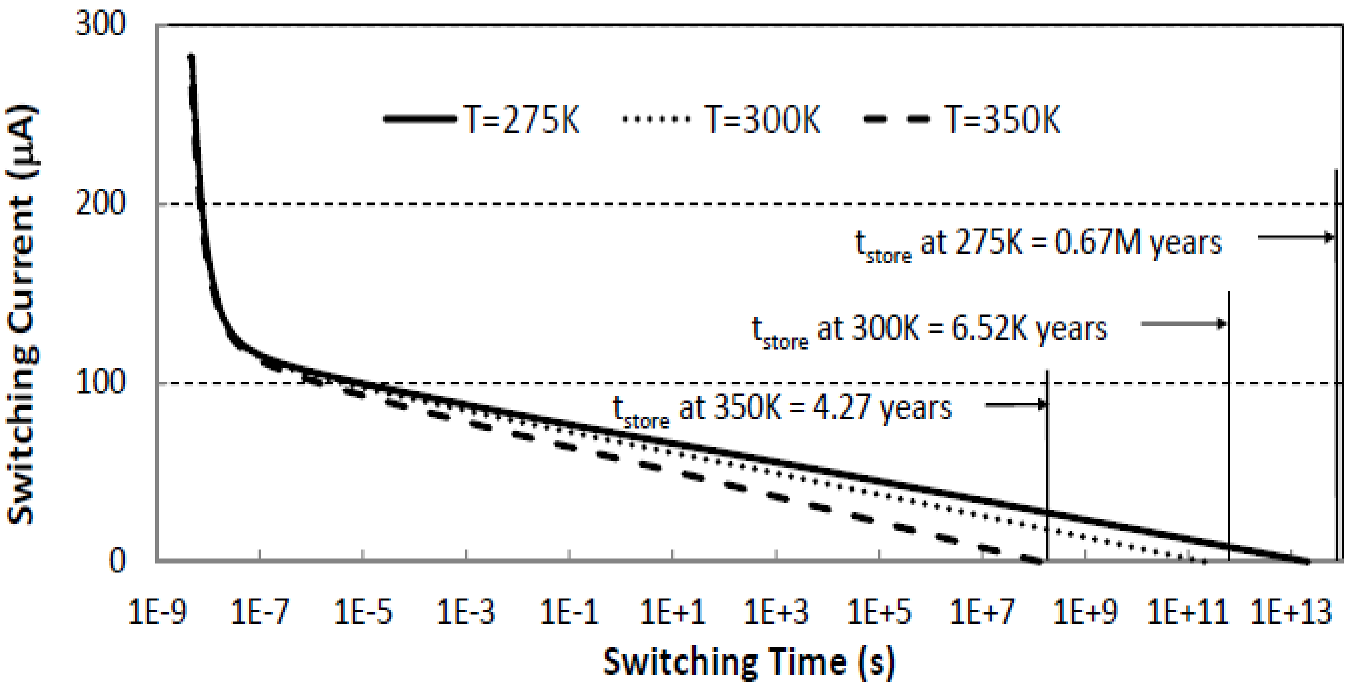
The measured static R-I curve of an MgO-based MTJ

Statistical design challenges (1)

- **More prominent statistical factors under scaled technology**
  1 CMOS+Device process variations → Persistent errors
  2 Probabilistic MTJ devices → Non-persistent errors
- **Expanded design space: read/write reliability/retention time/endurance.**



Geor



ɪabilistic
ɪvior

Statistical design challenges (2)

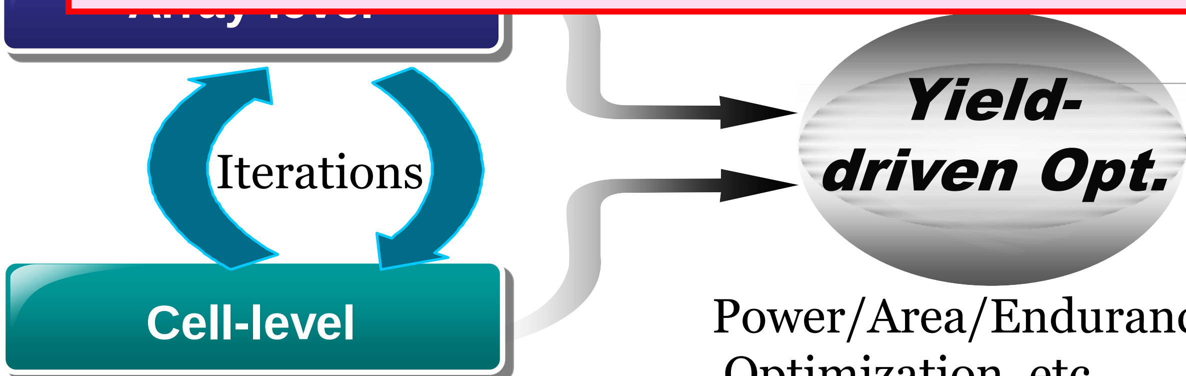
- **For system architects,**

Array-level reliability enhancement techniques, Error Correction Code (ECC)/Red. to relax the robustness requirement of single cell, like transistor size/cell failure rate (Huge exponential computation)

- **For device/circuit designers,**

Cell-level repair techniques, like size up the transistor size to tolerate the process variations/thermal fluctuations, to lower the cost of ECC/Red (Expensive Monte-Carlo simulations +magnetic-CMOS models)

Bottom-up design method is hardly integrated into system design.



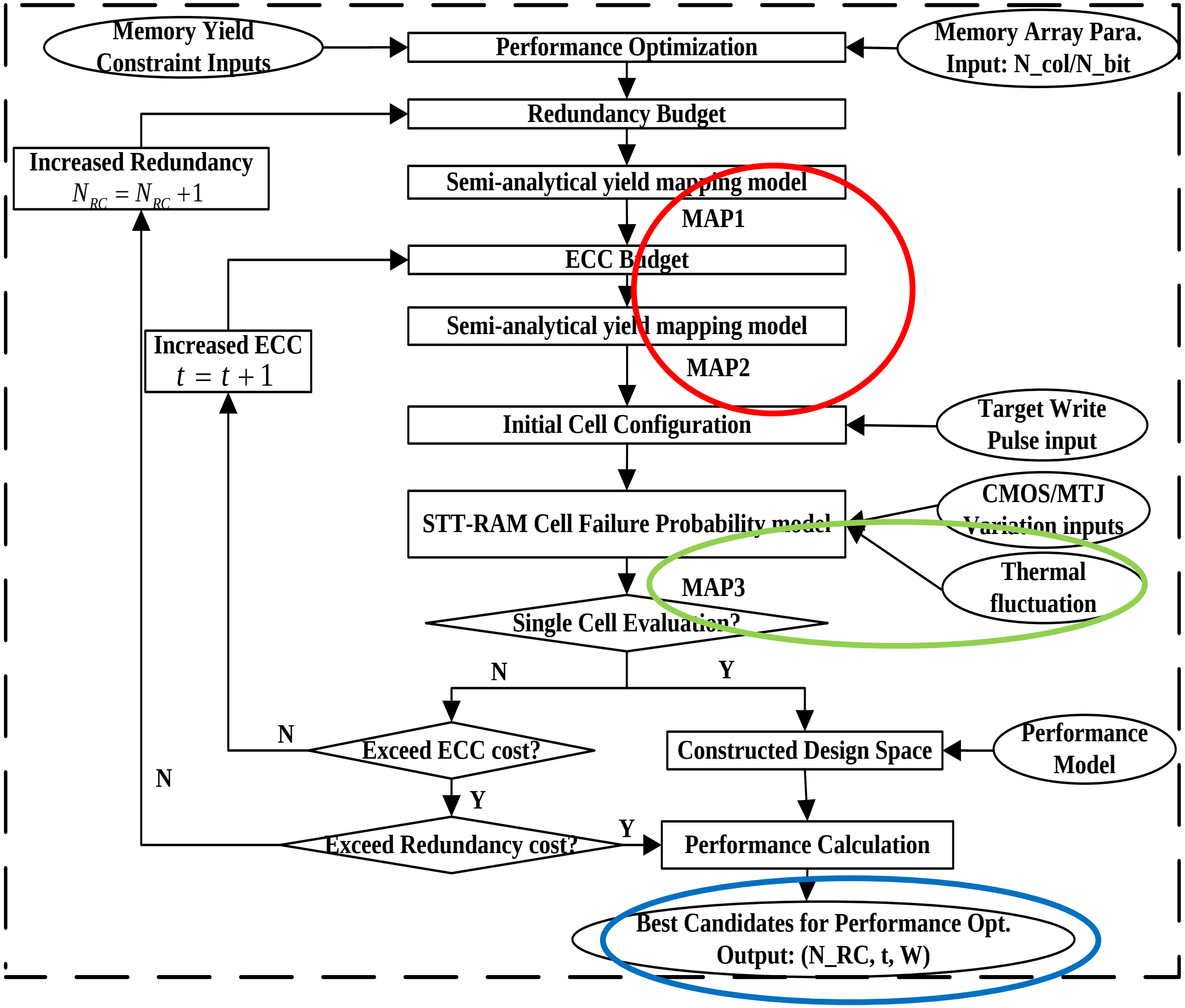
Power/Area/Endurance, Optimization, etc

Traditional bottom-up design method incurs costly iterations, even the cell-level reliability estimation is too costly

Outline Revisit

- Introduction
  - STT-RAM basics
  - Statistical design challenges
- **Overview of top-down statistical method-*Loadsa***
- **Hierarchical semi-analytical model of *Loadsa***
  - **Generic yield mapping model for ECC/Red.**
  - **Statistical failure-probability model for STT-RAM cell**
- **Case study of *Loadsa*: Yield-Driven Array Opt.**
- **Conclusions**

Overview of *Loadsa*



Top-down flow:

MAP1/MAP2: Generic mapping for ECC/Red. from array yield to Col/Row yield, then to cell failure rate.

MAP3: Variation-aware cell failure model mapping from cell failure rate to cell design para.

Best combination of array-cell design space for yield driven Opt.

Outline Revisit

- Introduction
  - STT-RAM basics
  - Statistical design challenges
- Overview of top-down statistical method-Loadsa
- Hierarchical semi-analytical model of *Loadsa*
  - Generic yield mapping model for ECC/Red.
  - Statistical failure-probability model for STT-RAM cell
- Case study of *Loadsa*: Yield-Driven Array Opt.
- Conclusions

Generic yield mapping model for ECC/Red.

- **Unaffordable computation cost of MAP1/MAP2 , especially the exponential computation**

1 Array yield Y_{mem} to column/row failure rate P_C under given Red.

2 Translation from P_C to cell failure probability P_F under selected ECC Schemes.

3 Map1/Map2 are switchable, generic expression (n_t,k,t), n_t=f(k,t),

Take ECC as example, then extend to a special case of ECC

Redundancy $n_t = f(k,t) = k + t.$

$$t = 1 : Y = (1 - P_1)^{n_1} + n_1 P_1 (1 - P_1)^{n_1 - 1}$$

$$t = 2 : Y = (1 - P_2)^{n_2} + n_2 P_2 (1 - P_2)^{n_2 - 1} + \frac{n_2(n_2-1)}{2} P_2^2 (1 - P_2)^{n_2 - 2}$$

$$t = 3 : Y = \sum_{i=0}^{t} C_{n_t}^{i} P_t^{i} (1 - P_t)^{n_t - i},$$

$$C_{n_t}^{t} = \frac{n_t(n_t-1)\cdots(n_t-t+1)}{1\cdot 2 \cdots t}$$

$$\ldots$$

Generic yield mapping model for ECC/Red.

- **Low cost Heuristic direct Model deduction (ECC example)**

1 Mathematatic deduction based on the P_0 without ECC

$$\begin{aligned} Y &= (1-P_0)^k \\ &= \sum_{i=0}^{t} C_{n_t}^i P_t^i (1-P_t)^{n_t-i} \quad \forall t \in [1, t_{\max}] \end{aligned}$$

2 Approximated Heuristic expression deduction (t=1,2, ECC)

$$\begin{aligned} (1-P_0)^k &= 1 - kP_0 + \frac{k(k-1)}{2!}P_0^2 \\ &- \frac{k(k-1)(k-2)}{3!}P_0^3 + O\left(P_0^4\right) \end{aligned}$$

$$\sum_{i=0}^{1} C_{n_1}^i P_1^i (1-P_1)^{n_1-i} = 1 - \frac{n_1(n_1-1)}{2}P_1^2$$

$$+\left(\frac{1}{2!}-\frac{1}{3!}\right)n_1(n_1-1)(n_1-2)P_1^3 + O\left(P_1^4\right)$$

t=1 $\quad P_1 = a_{1,1}P_0^{1/2} + a_{1,2}P_0.$

$$a_{1,1} = \left(\frac{2k}{n_1(n_1-1)}\right)^{1/2}, \quad a_{1,2} = \frac{1}{3}a_{1,1}^2(n_1-2)$$

t=2 $\quad P_2 \approx a_{2,1}P_0^{1/3} + a_{2,2}P_0$

$$a_{2,1} = \left(\frac{6k}{n_2(n_2-1)(n_2-2)}\right)^{1/3}, \quad a_{2,2} = \frac{n_2-3}{4}a_{2,1}^2.$$

$$P_t = a_{t,1}P_0^{1/(t+1)} + a_{t,2}P_0^{2/(t+1)} + \cdots + a_{t,t+1}P_0$$

Generic yield mapping model for ECC/Red.

- **High accurate Heuristic logarithm Model deduction**

Proposed for the reduced accuracy of direct mapping model if P_t is high (i.e.>1e-2), because of the inaccuracy of Taylor expansion

1 Approximated Heuristic expression deduction (t=1, ECC)
"ln" denotes natural logarithm function

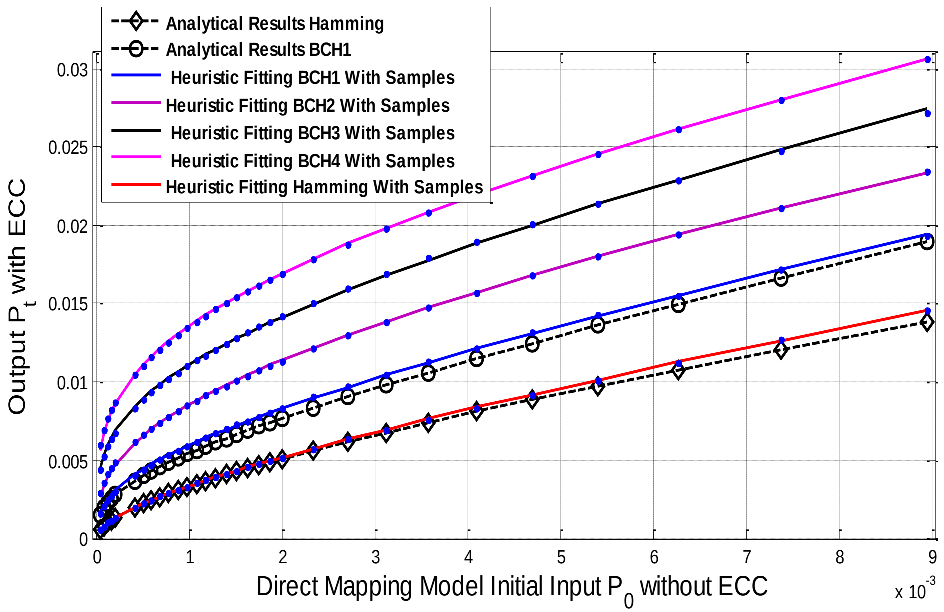
t=1 $$k\ln(1-P_0) = (n_1-1)\ln(1-P_1)+\ln(1+(n_1-1)P_1)$$ $$x_1 = b_{1,1}x_0 + b_{1,2}$$

$$-k\left(e^{x_0}+\frac{e^{2x_0}}{2}\right) \approx \frac{n_1(n_1-1)}{2}e^{2x_1} + \frac{n_1(n_1-1)(n_1-2)}{3}e^{3x_1}$$ $$b_{1,1}\approx 1/2, b_{1,2}\approx \frac{1}{2}\ln\left(\frac{2k}{(n_1-1)n_1}\right)$$

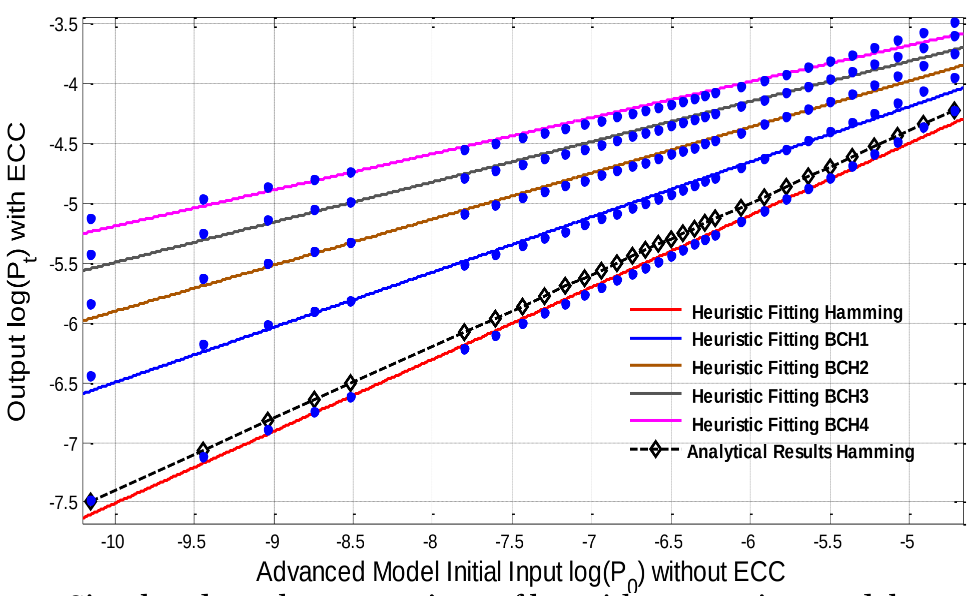
Heuristic Linear relationship $$x_t = b_{t,1}x_0 + b_{t,2}.$$

Validation-Generic yield mapping model for ECC

- Heuristic fitting/analytical results agree well with the golden direct computed samples in both Direct model and logarithm model.
- Logarithm model is more accurate in high error rate zone.



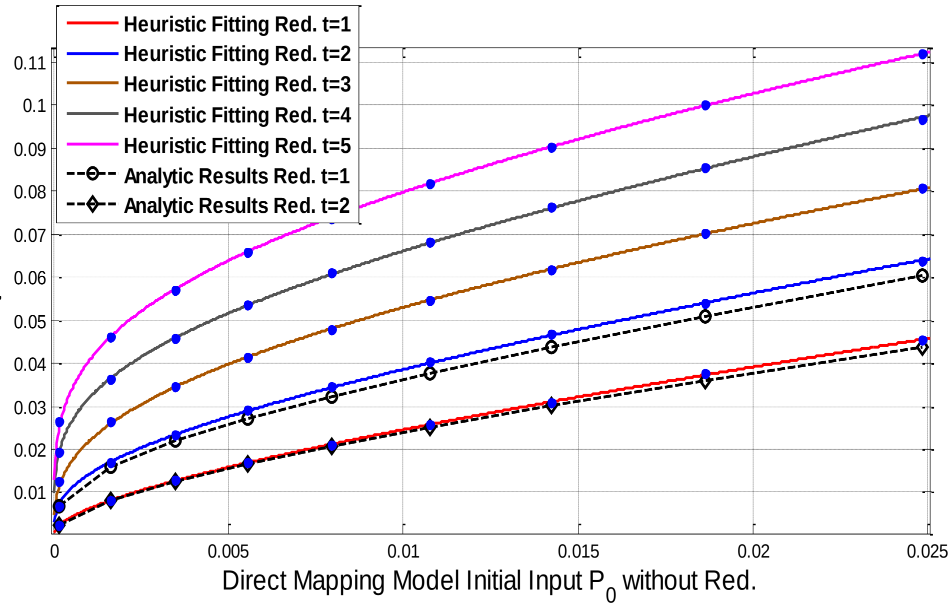
Simulated results comparison of direct mapping Model under different ECCs (Hamming, BCH1, BCH2, BCH3, BCH4).



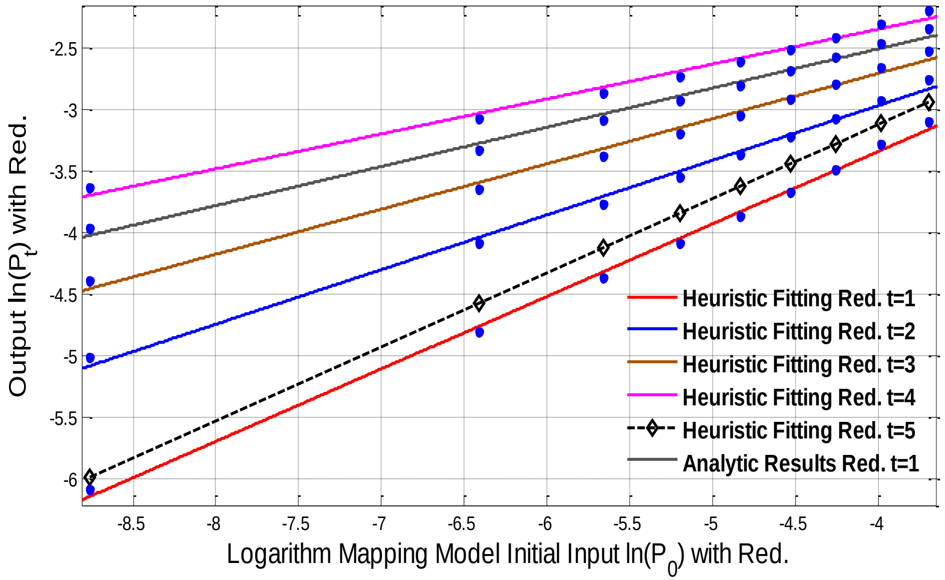
Simulated results comparison of logarithm mapping model under different ECCs (Hamming, BCH1, BCH2, BCH3, BCH4).

Validation-Generic yield mapping model for Red.

- Redundancy is a special case from ECC, can be seamlessly integrated in previous ECC yield mapping model $n_t = f(k,t) = k + t.$
- Results of Generic model for Redundancy have similar accuracy as ECC's.



Simulated results comparison of direct mapping model under different redundancy configuration (k = 64, t = 1, 2, 3, 4, 5).



Simulated results comparison of logarithm mapping model under different redundancy configuration (k = 64, t = 1, 2, 3, 4, 5).

Failure-probability model for STT-RAM Cell

- Translation from cell failure rate P_F to cell design parameters
Require an analytical model to characterize both process variations and probabilistic behavior of MTJ device for statistical design.

- Fast (significantly reduce the traditional expensive hybrid spice & macro-magnetic simulation)
- Scalable (independent of technology)
- Variation-Aware (statistical analysis for expanded design space exploration)
- Expendable (more design parameters and variability inputs)
- Smart enough for integration and multi-level optimization

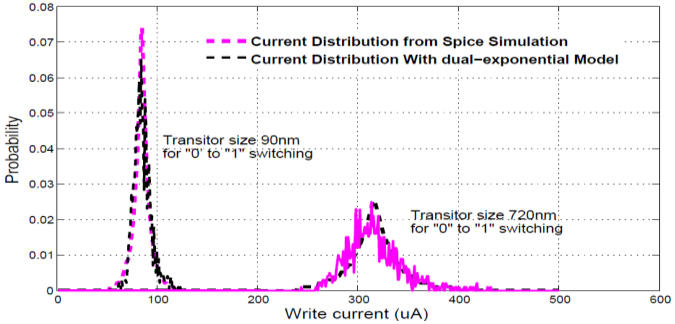
Failure-probability model for STT-RAM Cell

❑ Semi-analytical model deduction

A. Statistical Characterization of MTJ Switching Current (sensitivity analysis+ dual exponential current model for process variations)

$$p(I_{sw}) = \begin{cases} a_1 e^{b_1(I_{sw}-u)} & I_{sw} \le u \\ a_2 e^{b_2(u-I_{sw})} & I_{sw} > u. \end{cases}$$

$$\int p(I_{sw})dI_{sw} = 1$$
$$\int p(I_{sw})I_{sw}dI_{sw} = \mu_{I_{sw}}$$
$$\int {I_{sw}}^2 p(I_{sw})dI_{sw} = \mu_{I_{sw}}^2 + \sigma_{I_{sw}}^2.$$



B. STT-RAM Cell Failure under Thermal Fluctuation

Switching Current-time mapping $$I_{sw}(\tau_{th}) = I_{c0}\left(1 - \frac{\ln(\tau_{th}/\tau_0)}{\Delta}\right), \quad \tau_{th} > 10\text{ns}$$

Switching time distribution $$f(\mu_{\tau_{th}}) = \begin{cases} a_1 k_1 k_2 e^{b_1(k_1-\mu)} \mu_{\tau_{th}}^{-b_1 k_1 k_2 - 1}, & \mu_{\tau_{th}} \ge e^{\frac{1}{k_1} - \frac{\mu}{k_1 k_2}} \\ a_2 k_1 k_2 e^{b_2(\mu-k_1)} \mu_{\tau_{th}}^{b_1 k_1 k_2 - 1}, & \mu_{\tau_{th}} < e^{\frac{1}{k_1} - \frac{\mu}{k_1 k_2}} \end{cases}$$

STT-RAM Cell failure-rate $$P_F = \int_{e^{\frac{1}{k_1} - \frac{\mu}{k_1 k_2}}}^{+\infty} f(\mu_{\tau_{th}}) d\mu_{\tau_{th}} \int_{T_w}^{+\infty} f_g(\tau_{th}, \mu_{\tau_{th}}, \sigma_{\tau_{th}}) d\tau_{th} + \int_0^{e^{\frac{1}{k_1} - \frac{\mu}{k_1 k_2}}} f(\mu_{\tau_{th}}) d\mu_{\tau_{th}} \int_{T_w}^{+\infty} f_g(\tau_{th}, \mu_{\tau_{th}}, \sigma_{\tau_{th}}) d\tau_{th}$$

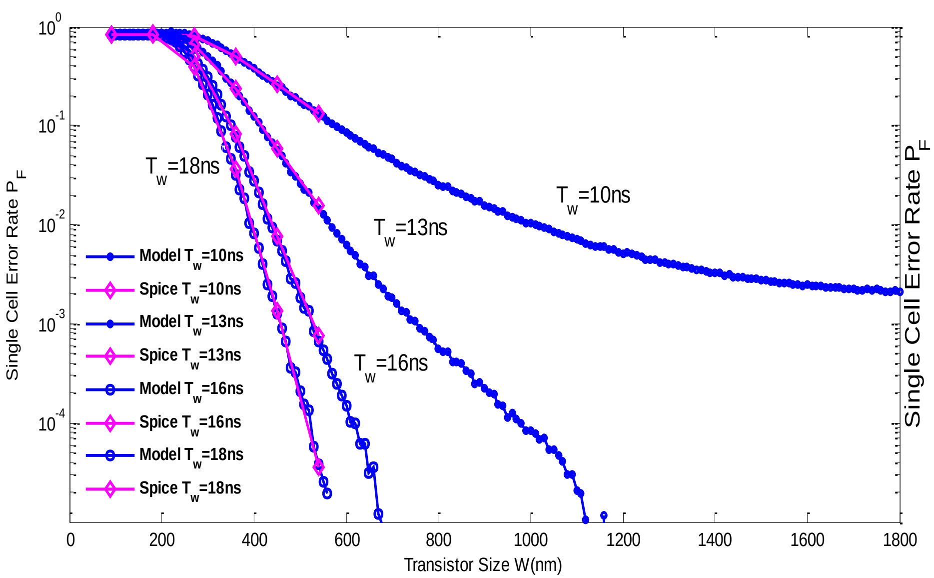
Validation-Failure-probability model for STT-RAM

- **Simulation settings at T=300K**

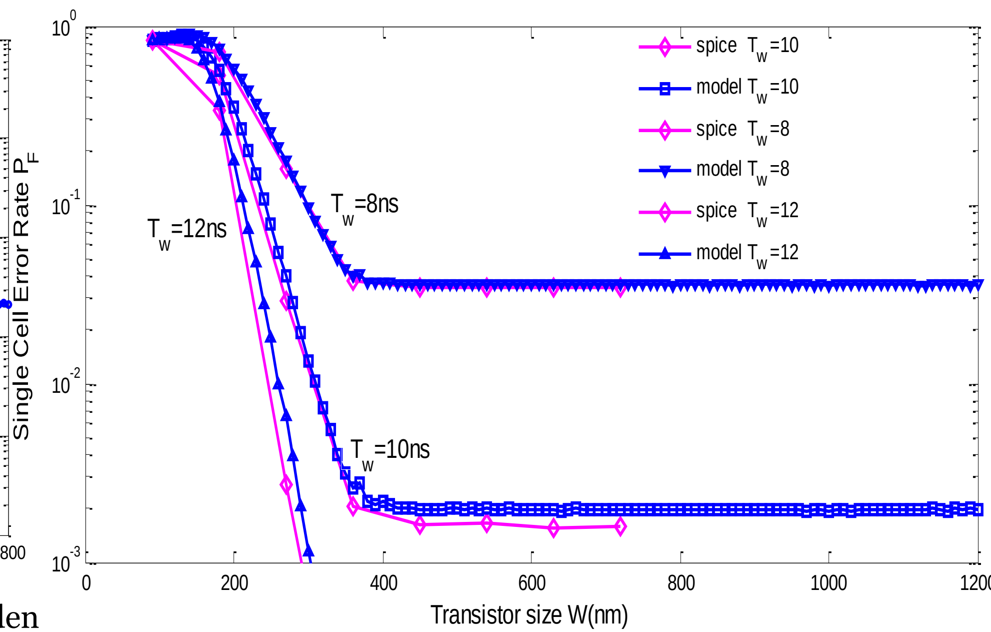
Parameters	Mean	Std.
Channel width	$\overline{W} = 90 \sim 1800\text{nm}$	$\sigma_W = 5\%\overline{L}$
Channel length	$\overline{L} = 45\text{nm}$	$\sigma_L = 5\%\overline{L}$
Threshold voltage	$\overline{V}_{th} = 0.466\text{V}$	Calucaltion
Mgo thickness	$\overline{\tau} = 2.2\text{nm}$	$\sigma_\tau = 2\%\overline{\tau}$
MTJ surface area	$\overline{A} = 45 \times 90\text{nm}^2$	Calculation
Resistance high	$R_H = 2000\Omega$	Calculation
Resistance low	$R_L = 1000\Omega$	Calculation

Validation-Failure-probability model for STT-RAM

- Accurate translation from P_F to cell design parameters at both directions under both process variations and thermal fluctuations.



The comparison of PF STT-RAM cell failure model v.s. golden spice for different Tw under T=300k at '1' to '0' switching.



The comparison of PF STT-RAM cell failure model v.s. golden spice for different Tw under T=300k for '0' to '1' switching

Outline Revisit

- Introduction
  - STT-RAM basics
  - Statistical design challenges
- Overview of top-down statistical method-Loadsa
- Hierarchical semi-analytical model of Loadsa
  - Generic yield mapping model for ECC/Red.
  - Statistical failure-probability model for STT-RAM cell
- **Case study of *Loadsa*: Yield-Driven Array Opt.**
- **Conclusions**

Case study-Loadsa

- **Mathematical Model formulation for performance opt.**

A. *F(X) is the target performance need to be optimized, such as power/area etc, we need to obtain the best combination of transistor size, redundancy/ECC configurations under yield/write pulse/variations(both process + thermal), the optimized value X.*

$$U_{opt} = \min(F(\mathbf{X}))$$

Where $\mathbf{X} = [W \quad N_{RC} \quad t]$,

Subject to:

Yield Constraint: $Y_{mem} \leq Y_{con}$ for $T_w \leq T_{w\_con}$,

Redundancy budget: $N_{RC} \in [1, N_{RC\_con}]$,

ECC budget: $t \in [1, t_{con}]$

Variations: $\sigma = [\sigma_{W\_con}, \sigma_{L\_con}, \sigma_{V\_th}, \sigma_{A\_con}, \sigma_{T\_con}]$

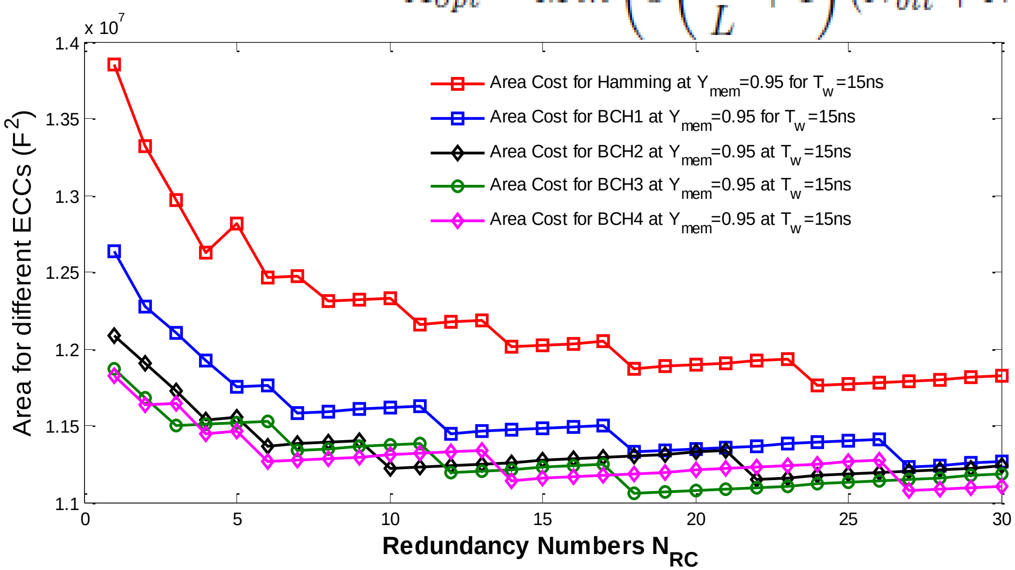
For all X, $X \in [X_{\min}, X_{\max}]$.

Case study-Loadsa

- **Case study: Yield-driven area optimization.**

Nbit=256bit, Ncol=1024, N_{RC_con}=30. Hamming code (265, 256, 1) and four BCH codes -BCH1 (274, 256, 2), BCH2 (283, 256, 3), BCH3 (292, 256, 4) and BCH4 (301, 256, 5), with the error correction capability t from 1 to 5.

$$A_{opt} = Min\left(3\left(\frac{W}{L}+1\right)(N_{bit}+N_{ECC})(N_{col}+N_{RC})\right)$$



Simulated results of area optimization for the budget ECCs, Redundant numbers NRC under Ymem = 95% for Tw = 15ns.

1 Benefit of increasing the strength of ECC for area optimization monotonically decreases when the ECC scheme changes from Hamming code to BCH1 – BCH4 with any simulated redundancy configurations.

2 Among all the configurations, the minimum area is acheived at BCH3 with 18 redundant columns.

Outline Revisit

- Introduction
  - STT-RAM basics
  - Statistical design challenges
- Overview of top-down statistical method-Loadsa
- Hierarchical semi-analytical model of Loadsa
  - Generic yield mapping model for ECC/Red.
  - Statistical failure-probability model for STT-RAM cell
- Case study of Loadsa: Yield-Driven Array Opt.
- **Conclusions**

Conclusion

- We developed a fast and accuracy generic semi-analytical yield mapping algorithm to hierarchically map the required memory array yield to the cell-level failure probability under certain ECC and redundancy configurations.
- We proposed using the sensitivity analysis technique and the dual-exponential model of MTJ switching to simplify the derivation of PF from the cell designs by considering both process variations and thermal fluctuations. The accuracy and cost of semi-analytical STT-RAM cell model are demonstrated.
- We demonstrated the possibility of developing a top-down statistical design method for STT-RAM and the efficiency of our proposed *Loadsa* technique in our experiment results and case studies.

Thank you!