Predictive method of charge storage memory degradation on a dedicated 4kb test vehicle

S. Perrin, Khaled Alkema, V. Della Marca, Thibault Kempf, O. Paulet, Martin Arteaga Castillo, M. Akbal, B. Arrazat, J. Metz, J. Moragues, et al.

Predictive method of charge storage memory degradation on a dedicated 4kb test vehicle

S. Perrin[1,2*], K. Alkema[1,2], V. Della Marca[2], T. Kempf[1], O. Paulet[1], M. Arteaga[2], M. Akbal[1], B. Arrazat[1], J. Metz[1], J. M. Moragues[1], A. Regnier[1], M. Bocquet[2], J. M. Portal[2]
[1]STMicroelectronics, 190 avenue Celestin Coq, 13106 Rousset, France
[2]Aix-Marseille University, IM2NP, CNRS, UMR 7334, 5 rue Enrico Fermi, 13453 Marseille, France
*sebastien.perrin01@st.com

***Abstract*— This study investigates the variability in HCI (Hot Carrier Injection) degradation within a 4 kb charge storage memory array, manufactured on a dedicated wafer split process. A standard set of experiments is conducted to extract electrical features of the cells before and after stress. A statistical analysis method, based on the Principal Component Analysis (PCA) approach, is introduced to enhance the comprehension of the cell degradation prior to reliability testing. Additionally, a graphical model is developed to identify extrinsic cells in the memory array prior to stress, as well as to assess the overall technology yield. Finally, others classifier models are explored aiming to improve extrinsic cells detection before running the reliability test.**

Keywords— charge-storage memory, reliability, variability, principal component analysis.

I. Introduction

In the context of advanced semiconductor technology development, it is imperative for foundries to validate their manufacturing processes to ensure compliance with reliability specifications. To achieve this, a series of electrical tests are performed at the wafer level across various test structures, with the primary goal of obtaining statistically significant data. The Cell Array Stress Test (CAST) [1] is commonly employed to detect process-related failures by assessing a large amount of memory bit cells in parallel. In complement, complex test chip, close to product functionality, offers also statistical measurements and spatial analysis of electrical performance [2], [3], but at the cost of an intense design and test development. So, during the R&D phase, a specific 4-kilobit addressable test vehicle, named SuperCAST [4], has been developed for EEPROM technology, keeping the design complexity at the level of a CAST, whereas proposing single cell addressing capability.

In this study, the SuperCAST is utilized to investigate the statistical impact of intentional process variations on the reliability of EEPROM technology, specifically focusing on the modification of channel width (W). To do so, this work examines first the electrical degradation observed during endurance testing, with the aim of elucidating the underlying physical mechanisms associated with this process variation. Then, the development of a statistical model is proposed to predict, at the initial test (t_0), which memory bit cells will likely belong to the extrinsic population following the reliability tests. The extrinsic population is defined as the lower threshold voltage (V_T) memories and the intrinsic population corresponds to the higher V_T memories. Principal Component Analysis (PCA) is employed to accomplish this objective. Multivariate analysis methods, particularly PCA, have been demonstrated in numerous studies over the past two decades to be effective in detecting performance outliers in CMOS circuits [6] and in identifying potential failures during final testing [7], [8], [9]. This paper is an extended version of an initial study presented at the International Integrated Reliability Workshop (IIRW) 2024 [10]. The previous work aimed to demonstrate that it is possible to detect the extrinsic population at t_0 using PCA, and the first result is discussed. This paper reviews the experiment and PCA method in Section II, explains the physical mechanism involved in the reliability test in Section III, defines the PCA implementation in Section IV.A, and assesses the performance of a linear classifier model designed to detect the entire extrinsic population at t_0 in Section IV.B. The additional value of this present work is first depicted in Section IV.C, where three different models are automatically calculated to improve the precision/recall trade-off compared to the one obtained in Section IV.B. Then, the impact of dataset features on these models' performance is studied. Finally, in Section IV.D, we show that the trained models in Section IV.C demonstrate good performance predictions in terms of the number of weak bit memory cells, regardless of whether the die is highly impacted by hot carrier degradation or not.

II. Experiment and Method

The process modification consists in a width decrease of the p-doped active well implants, of the floating gate memory device. This is realized all over the wafer to obtain memory cells with a narrow channel effect. The modification implies several activations of the cell and its degradation during reliability studies due to the decrease of injection surface into the floating gate and the narrower channel. To correlate the experimental results, we integrated on each wafer die different EEPROM test structures: single cells, CAST (100k cells all connected in parallel), CAST of TREQ (CAST of memory transistors where the floating gate and the control gate are shorted) and SuperCAST (4k bit fully addressable array). Firstly, to detect the process modifications realized on the whole wafer, a set of drain current (I_D) versus control gate voltage (V_{CG}) characteristics has been carried out on the CAST of TREQ to measure the threshold voltage (V_T) variability. The V_T wafermap is presented in Fig. 1. As we are interested in the extrinsic population in the cells array, the threshold voltage has been extracted at a drain current of 100 nA. The spatial variability of the narrow cells shows coherent value compared to the process modification that we implemented. This enables the dies identification most impacted by the process. Therefore, three dies have been selected according to their V_T and W values, to perform reliability tests and to investigate the intra-array variability, firstly on a CAST of 100k cells and then on a fully addressable memory array (SuperCAST). The dies 1 and 3, that present a CAST of TREQ with low V_T, are cycled with different electrical conditions to verify

the presence of a possible extrinsic population when the cells are not stressed enough. The results of these characterizations will be used in the next section to generate

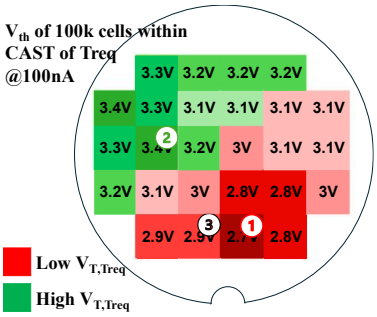


Fig 1: Threshold voltage spatial distribution of 100k CAST of TREQ. Three dies are identified for the following reliability test.

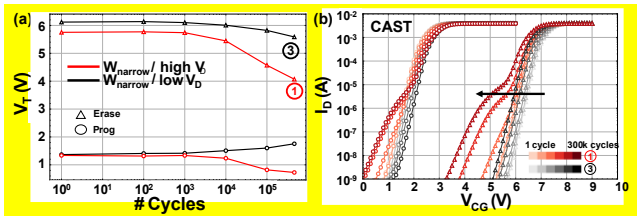


Fig 2: Endurance of 100k cells CAST up to 300K cycles at low V_D (die 3) and high V_D voltage (die 1). (a) Programming window and (b) I_D-V_{CG} evolution during stress.

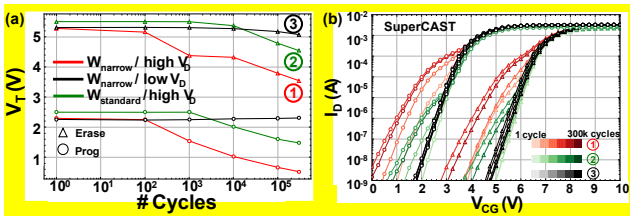


Fig 3: Endurance of 4k cells SuperCAST up to 300K cycles at low V_D (die 3) and high V_D (die 1) with narrow channel and standard width using a high drain voltage (die 2). (a) PW and (b) $I_D(V_{CG})$ evolution during stress.

the dataset for the PCA. The aim will be to show the graphical model predictability for the die 2 that is considered as not impacted by the process variation from the classical analysis (Fig. 1). In Fig. 2 endurance tests are thus performed with program/erase Fowler-Nordheim operations, on CAST of two dies with a narrow channel width (W_{narrow}). The evolution of extracted V_T curves show that programming window shift is more important by cycling die with higher drain bias, as shown in Fig. 2.a. This performance can be attributed to the channel interface degradation due to the hot hole generation during programming operation. Moreover, the drain current measurements by sweeping the control gate voltage during stress illustrate the presence of extrinsic population for die 1 while die 3 seems to have a normal behavior (Fig. 2.b).

Following these, in Fig. 3, we reported the results of endurance tests, adding the die 2 with standard active width. In this case, the electrical characterizations have been performed on the SuperCAST of 4kb memory array. This structure enables us to study the intra-array phenomena previously detected on the CAST array. In Fig. 3.a, the V_T evolution confirms the behavior observed on CAST for the dies 1 and 3. Moreover the die 2 presents a strong V_T degradation despite the standard active shape. In Fig. 3.b extrinsic signature on the I_D-V_{CG} characteristics has been identified for both standard and narrow channel meaning that there is an intra-array phenomenon that can be dissociated using SuperCAST structure. In particular, the SuperCAST enables us to measure any relevant electrical parameter for each single memory cell in the array. From these measurements, a statistical analysis method named Principal Component Analysis (PCA) [11] has been implemented to identify clusters of failed and non-failed cells in the memory array at each step of the reliability test. Therefore, a statistical model could be created, to predict the number of extrinsic devices before an endurance test.

III. FAILURE MECHANISMS

To establish the relevant parameters for PCA, the physical mechanism underlying this kind of degradation has been investigated at the single memory cell level. During the ejection of electron from the floating gate, the n+ implant between the select and memory transistors is biased to a high positive potential. The junction with the substrate is thus in strong inversion while the channel is pinched-off, generating hot electron-hole pairs. Some of these holes can be directed toward the tunnel oxide increasing the interface defects generation (Fig. 4). Quasi-static characterizations are performed to elucidate the role of V_D potential in the cell degradation following several thousand cycles on single narrowed cell compared to a standard. In Fig. 5 and 6, bulk and drain current (I_B and I_D respectively) as a function of V_{CG} are plotted. The measurements are carried out at various V_D voltages from 0.7V to 4V maintaining the memory state unchanged, while the other terminals are grounded. At the beginning of the experiments the floating gates were emptied to avoid the memory state changing during the quasi-static tests at high V_D voltages. It is important to notice that normally the EEPROM memory device is used with higher drain biases, but the amounts used in this paper allow to highlight the failure mechanism without perturbations. The Fig. 5.a presents the I_D-V_{CG} for different V_D voltage, showing on a standard device no leakage is generated at low V_{CG}. The $|I_B|$ current is measured at the same time and Fig. 5.b shows, for a $W_{standard}$ cell, a first peak indicating an electron-hole pair generation [10], around $V_{CG}=V_T$, the current is still high when the control gate voltage increases. For a narrow cell, the drain current at low V_{CG} increases with the drain bias (Fig. 6.a), as well as a higher bulk current peak is measured, as shown in Fig. 6.b. This demonstrates that the process variation improves the impact-ionization increasing the oxide interface degradation during the electron ejection operation when V_D reaches high voltages. This explains the threshold voltages drop during the endurance test and the differences between die 1 and the die 2. However, for the narrowed channel, degradation will occur at lower VD voltage. In conclusion, these quasi-static measurements have been used to know if single memory devices suffer from hot carrier injection induced by narrow

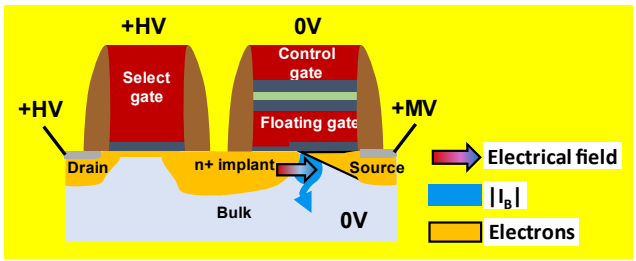


Fig 4: Schematic of the studied memory technology during the ejection of electron from the floating gate.

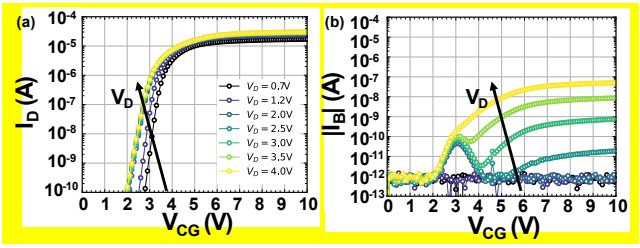


Fig 5: Quasi-static characteristics of a $W_{standard}$ cell: (a) I_D-V_{CG} evolution and (b) I_B-V_{CG} evolution for various V_D.

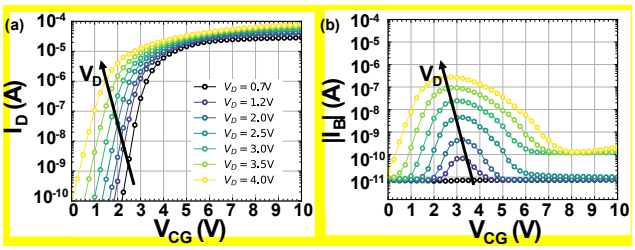


Fig 6: Quasi-static characteristics of a W_{narrow} cell: (a) I_D-V_{CG} evolution and (b) I_B-V_{CG} evolution for various V_D.

channel effect even if Fowler-Nordheim program/erase operations are used. Indeed, hot hole injection is possible due to the electron-hole pairs generation at the tunnel oxide interface in the n-implant junction region. So, to build a predictive model using the PCA, the electrical parameters that discriminate narrow channel characteristics have been extracted for each memory cell of the 4kb SuperCAST.

IV. PCA IMPLEMENTATION FOR RELIABILITY PREDICTION

In this section, the PCA method is first presented (A) to observe the correlation of electrical parameters. Next, a graphical linear model is evaluated to predict the extrinsic population at t_0 (B). After this, the impact of an additional parameter in the dataset from die 1 is examined by comparing the training results obtained from several models (C). Finally, these models are tested on datasets with varying numbers of extrinsic samples (D).

A. PCA methodology

PCA is a statistical analysis methodology used in Machine Learning algorithm to simplify the data analysis. Indeed, this method is known to reduce the dataset dimensionality while minimizing the information loss for a best data representation to analyze [5]. To implement PCA, it is important to create a dataset storing the electrical properties of each addressable memory bitcell in the SuperCAST. The relevant parameters considered for the dataset are: the transconductance (G_M), the peak drain current (I_{ON}), V_T, the subthreshold slope (Ss) and drain current measured for higher V_D at five control gate voltage ($I_{D,Vcg=0V}$, $I_{D,Vcg=0.5V}$, $I_{D,Vcg=1V}$, $I_{D,Vcg=1.5V}$, $I_{D,Vcg=2V}$). The measurements are performed for dies 1, 2 and 3 (see Fig. 1) and a dataset is built at each step of the endurance test (t_0, 10kcycles, 100kcycles and 300kcycles). Then, the PCA can be applied to the entire dataset. This mathematical technique linearly combines the features of the dataset to create a new set of uncorrelated variables called principal components (PC). These components are calculated such that they fit the data while capturing the maximum amount of variability. Thus, N features (electrical parameters) can be analyzed and

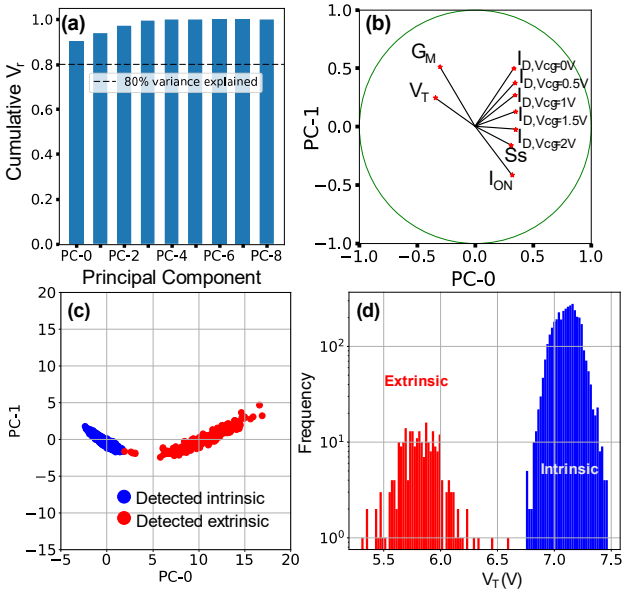


Fig. 7: (a) The cumulative variance explained ratio (Vr) as a function of the principal components that were extracted from the dataset of die 1 after 300kcycles. PC-0 and PC-1 are selected for a two-dimensional analysis while limiting information loss (b) Correlated parameters after stress represented as vectors in a normalized circle in the principal components space (c) After stress dataset from die 1 transformed and projected in the principal component space where intrinsic and extrinsic populations are highlighted (blue and red respectively) (d) V_T distribution to identify extrinsic (low V_T) and intrinsic (high V_T) memory cells after 300kcycles.

plotted thanks to only 2 or 3 PC. In this study, PCA was applied at first to the dataset obtained after 300kcycles on die 1. The PCA method involves four preliminary steps: i) a dataset is created; ii) the electrical parameters are normalized in the preprocessing phase to make variables comparable [5]; iii) during the PCA process, the covariance matrix is calculated, iv) which is used to obtain the eigenvalues and eigenvectors. The latter are defined as principal components [5],[11]. In addition, the variance explained ratio (Vr) of a principal component is calculated as the ratio of its corresponding eigenvalue to the sum of all eigenvalues. Hence, Vr indicates the proportion of the total variance in the dataset explained by that principal component. As a result, principal components are chosen so that the cumulated Vr is bigger than 80 %. This empirical threshold assists in determining the minimum number of principal components needed to account for at least 80% of the original data variance. As illustrated in Fig. 7.a, the cumulative variance ratio reaches 90% with just PC-0. Nonetheless, a two or three-dimensional representation facilitates data analysis more effectively than a one-dimensional representation and maximizes the variance captured from the original dataset. This approach enhances the optimization of a predictive model within that space. Therefore, the first two principal components, PC-0 and PC-1, are selected to capture 94% of the variability from the original data. They constitute a two-dimensional space where the parameters are correlated in a normalized circle (Fig. 7.b). Indeed, each parameter is represented as a vector, and their collinearity directly describes their correlation with each other. To achieve the data representation shown in Fig. 7.c, the electrical parameters are transformed to be displayed in that 2D-space. First, the data are standardized to remove the scale differences

between the electrical parameters. The standardized data are subsequently projected into the principal component space by multiplying them with a projection matrix derived from the selected principal components. This transformation is analogous to a change of reference frame, where each point represents all electrical parameters of a memory cell in the array. The intrinsic and extrinsic population are identified in that space thanks to the V_T distribution (Fig 7.d).

B. Graphical linear model

In Fig. 7.c, a majority group is illustrated by the intrinsic population (3843 points in blue) and separated from a minority group illustrated by the extrinsic population (253 points in red). According to the circle of correlation (Fig. 7.b), the high V_T population is concerned by a good correlation of parameters V_T and G_M with Ss and I_{ON} respectively while the low V_T population is concerned by the drain current measured at high V_D for multiple gate voltage. It is important to note that the parameter $I_{D,Vcg=2V}$ is along the PC-0 direction, meaning that it participates to the horizontal separation of both population.

The dataset obtained on the same die before stress was transformed and projected in the same principal component space to conserve the similar relationship between electrical parameters. The extrinsic and intrinsic populations are initially identified from post-stress data (see Fig. 8.a). A linear classifier model is then manually constructed to separate these populations at t_0. The slope and intercept of this model are graphically adjusted so that the linear boundary positions all extrinsic samples on the right side of the principal component space in both pre-stress and post-stress cases, as illustrated in Fig. 8.a and Fig. 8.b. As the dataset obtained from die 1 is used for training the model, this procedure was repeated with the measured data from die 2 before and after stress to validate this classifier model in the case where less extrinsic population was observed (Fig. 9.a,b). This model allows extrinsic and intrinsic population classification without the need of a reliability test. Fig. 8.b and Fig. 9.b show a large population of memory cells on the left side of the linear model (black line) corresponding to the predicted intrinsic population. Moreover, a smaller population is on the right side of the model corresponding to the predicted extrinsic population. It is important to remark that some memory cells identified as intrinsic are on the right side of the model. That population correspond to false extrinsic predicted at t_0. Therefore, to assess the model's prediction robustness, the accuracy, precision, and recall metrics were determined.

Accuracy metric is the fraction of the detected extrinsic and intrinsic memory cells that have been correctly predicted over the total population. Precision metric represents the fraction of the detected extrinsic cells on the right side of the model over the total population on the right side of the black line. Finally, recall corresponds to the fraction of the detected extrinsic population on the right side of the model over the whole detected extrinsic population (all the red points).

The metrics of this model are compared in Table I for both dies before stress.

Higher accuracy is observed for die 1 because more samples are correctly predicted in both categories, while in the case of die 2, the fraction of the correctly predicted samples is smaller. However, we can consider that the accuracy is similar for both cases because they are affected by imbalanced dataset. An imbalanced dataset is characterized by one class being largely represented by many examples, while the other class is represented by only a few examples. It has been published that an imbalanced dataset can be an obstacle (though not the only one) for machine learning algorithms when inducing a classifier model [12]. So, it is possible to achieve high accuracy without correctly predicting the extrinsic population. Thus, precision and recall are additionally calculated to validate the model aiming to predict the extrinsic memory cells before stress.

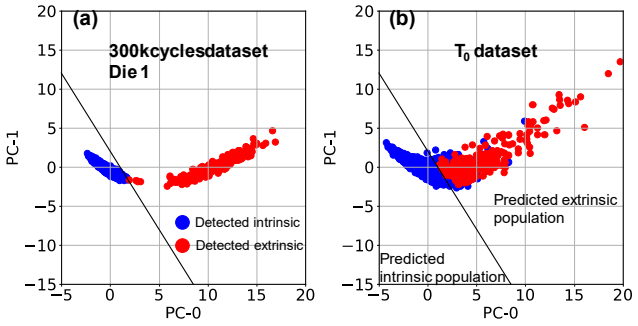


Fig. 8: Dataset from die 1 transformed and projected in the principal component space (a) after 300k cycles and (b) at t_0. Extrinsic population (red) is detected after the reliability test and is identified in a minority

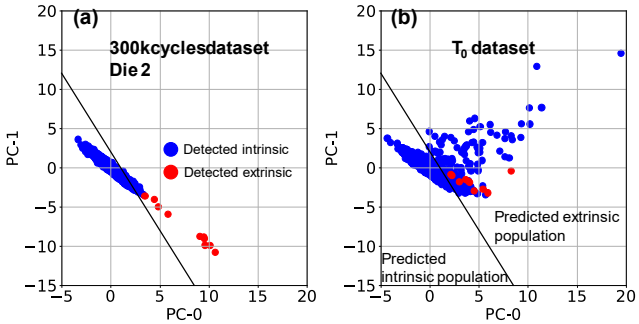


Fig. 9: Dataset from die 2 transformed and projected in the principal component space calculated with dataset from die 1 after 300kcycles (a) after 300k cycles and (b) at t_0. Extrinsic population (red) is detected after the reliability test and is identified in a minority population on the right side of the linear classifier model (black line). At t_0, the model can predict extrinsic devices in the minority population.

TABLE I. METRICS OF THE CLASSIFIER LINEAR MODEL

Dataset @t_0	Accuracy	Precision	Recall	Predicted extrinsic population size
Die 1	78%	22%	100%	1166
Die 2	73%	1%	100%	1095

For both dies, the measured precision of the model is low especially in the case of die 2, because its corresponding detected extrinsic population is very small in respect to the predicted extrinsic population size. Moreover, the model is more precise for die 1 because the detected extrinsic population is larger. Therefore, this metric is directly

related to the number of degraded devices during the endurance test. Fig. 10. displays the memory bit cells at t_0 in the principal component space, indicating those that will join the extrinsic population during the endurance test. It highlights that the earliest degrading memories (red) are statistically associated with higher drain current at t_0, whereas the latest degrading memories (orange and yellow) tend to appear in the overlap zone between intrinsic and extrinsic populations. So, we could imagine that increasing the number of cycles would have created more extrinsic samples in the overlay zone and improve the precision metric.

Finally, the measured recall is 100% for both datasets, which means that all detected extrinsic devices are in the predicted extrinsic population. These results are consistent with the manual tuning of the model. They represent how many samples belong to the extrinsic population at t_0 according to the model for both dies.

Thus, this model can indicate whether a die is affected by process variation, but it cannot predict the extent to which one die is impacted compared to another, as a classical endurance test would (as shown in Fig. 3.b). Consequently, to improve the model, the precision must be strongly increased to enhance the failure prediction. In this case, precautions must be taken to ensure that the recall metric is not adversely affected. Thus, in the following section, a fourth metric called the F1-score is introduced to quantify the trade-off between precision and recall.

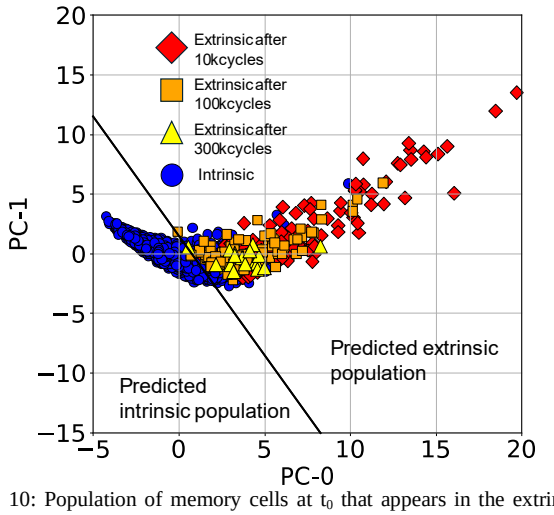


Fig. 10: Population of memory cells at t_0 that appears in the extrinsic population during cycling depicted in the principal component space calculated in Section A.

Additionally, we aim to develop a model capable of predicting the population that will fail and distinguishing between more and less impacted dies at t0.

C. Dataset impact on trained models

This section aims to improve the model described in Section IV.B. Initially, three standard Support Vector Classifier (SVC) models are trained on the dataset obtained from die 1. Each model is automatically fitted to the data as accurately as possible, ensuring maximal optimization of the metrics for classifying both populations at t_0. Subsequently, the process is repeated with a derived dataset that includes an additional feature. This experiment aims to enhance the precision/recall trade-off and to study the impact of the additional parameter on the model's performance.

The additional studied parameter corresponds to the slope of the bulk current ($|I_B|$), which is calculated for each memory bit cell and added to the dataset described in Section IV.B. As illustrated in Fig. 11, the slope is extracted for $1.5V \leq Vcg \leq 2V$ from the $|I_B|=f(Vcg)$ characteristics of both extrinsic and intrinsic classes. The slope is extracted at this control gate conditions because the bulk current variation is more sensitive to the impact ionization peak.

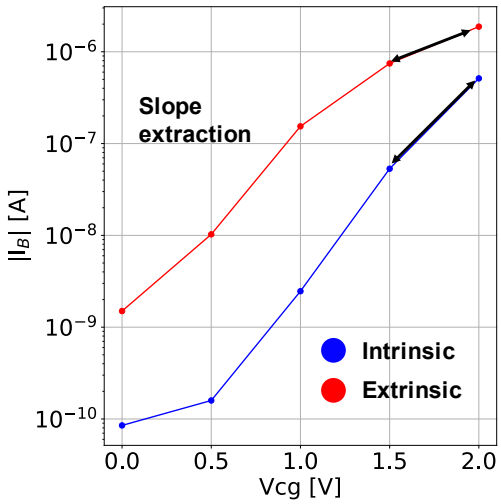


Fig. 11: Absolute bulk current ($|I_B|$) vs. control gate voltage (Vcg) characteristics of extrinsic (red) and intrinsic (blue) single memory bit cell at t_0. The slope extraction corresponds to the slope of $|I_B|$ between V_{cg} = 1.5V and V_{cg} = 2V.

Indeed, a lower calculated slope suggests that the maximum of $|I_B|$ is nearly reached, indicating that the memory is more susceptible to impact ionization compared to a higher slope. Thus, the PCA procedure was reproduced adding this new feature to the dataset from die 1, before and after degradation. As described in Section IV.B, the principal component space was calculated using the dataset obtained after stress. The dataset corresponding to the t_0 data was then transformed and projected into this same space. The PCA results obtained with the new dataset are depicted in Fig. 12. In Fig. 12.a, 93% of the data variance is captured by the first two principal components, suggesting that a two-dimensional space is sufficient for analyzing the data and assessing the classifier model. The slope relationship with the other parameter is illustrated by circle of correlation (Fig. 12.b). This parameter makes a significant contribution along the PC-1 axis and has a similar impact along the PC-0 axis as V_T. Consequently, the intrinsic population becomes aligned with the slope direction (Fig. 12.c) and appears more concentrated than it would be without this parameter in the dataset (Section IV.B). Similar slope impact is observed for the t_0 dataset (Fig. 12.d). Thus, to evaluate the relevance of this parameter for failure prediction quality at t_0, the model metrics are compared across three classifier types. The first implemented model is linear, the second is polynomial (degree 3) and the third one is Radial Basis Function (RBF) model. The latter performs classification based on an exponential model.

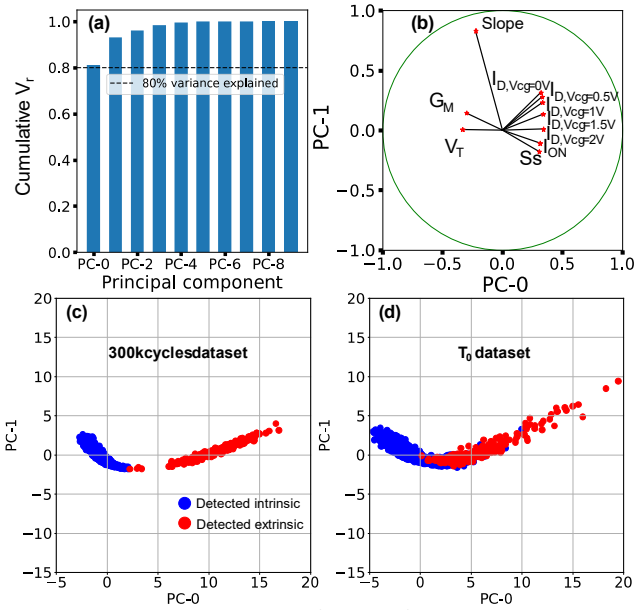


Fig. 12: (a) The cumulative Vr as a function of the principal components calculated from the die 1 dataset after 300kcycles including the slope parameter after 300kcycles. PC-0 and PC-1 are selected to capture 93% of the original variance (b) The circle of correlation representing the relationship between measured parameters after stress (c) After stress dataset from die 1 transformed and projected in the principal component space where intrinsic and extrinsic populations are highlighted (blue and red respectively) (d) The t0 dataset projected in the same space obtained in (c).

TABLE II. CLASSIFIER MODEL METRICS PERFORMANCE FOR THE DATASET NOT INCLUDING THE SLOPE PARAMETER

Dataset (without Slope) @t_0	Accuracy	Precision	Recall	F1-score
Linear	96%	82%	53%	64%
Polynomial	96%	89%	39%	54%
RBF	96%	80%	58%	68%

TABLE III. CLASSIFIER MODEL METRICS PERFORMANCE FOR THE DATASET INCLUDING THE SLOPE PARAMETER

Dataset (with Slope) @t_0	Accuracy	Precision	Recall	F1-score
Linear	96%	83%	51%	63%
Polynomial	96%	84%	45%	59%
RBF	96%	80%	58%	68%

For both datasets, with and without the slope parameter, the metrics are calculated for all model types and summarized in Table III and Table II, respectively.

The influence of the slope parameter, although minor, is particularly noticeable in the polynomial model with respect to the F1-score. Specifically, when the slope parameter is excluded (Table II), the F1-score is 54%, whereas it increases to 59% when the slope parameter is included (Table III). This result indicates that the polynomial model is weakly improved by incorporating the slope parameter in the dataset for calculating principal components. On the contrary, the slope parameter does not impact the linear and RBF models. Additionally, the RBF model appears to be slightly more optimized than the linear model with respect to the F1-score.

The higher recall observed in Table I, as compared to Tables II and III, regardless of the inclusion of the slope parameter, can be attributed to the manual tuning of the graphical model. This manual adjustment was specifically performed to ensure that all extrinsic population is positioned on the right side of the principal component space. However, this approach, while enhancing recall, adversely affects precision, rendering the process variation impact from one die to another indistinguishable.

In contrast, the automatic fitting methods, as demonstrated in Tables II and III, yield higher precision. This is because the algorithm employed in automatic fitting is designed to optimize the precision/recall trade-off, thereby enhancing the prediction performance and enabling the discernment of impacted dies.

The small impact of the slope parameter is related to the bulk current measurement. This electrical parameter is measured at the programmed state of the memory bit cell and exhibits variability across the entire memory array. Additionally, the applied control gate voltage (V_{cg}) does not control directly the channel due to the floating gate. Consequently, these conditions are not optimal for highlighting the impact ionization mechanism for each individual memory cell. To achieve accurate measurement conditions, the programmed threshold voltage ($V_{th,prog}$) should be considered, and V_{cg} should be adjusted to maintain an identical vertical electric field across the tunnel oxide for each memory bit cell.

This experiment has demonstrated that, despite the significant impact of the slope parameter after stress, it is not a decisive factor in distinguishing between extrinsic and intrinsic populations at t_0. In Section IV.D, the training model results summarized in Table III are tested on two datasets with varying numbers of extrinsic samples. For the remainder of this paper, the slope parameter is retained, as it slightly improves the training metrics.

D. Test models

The first dataset used for testing the models is significantly impacted by hot carrier degradation and contains 341 extrinsic samples, whereas the second dataset, corresponding to die 2, contains only 10 extrinsic samples. As the role of the trained model is to discern the more impacted dies over the wafer at t_0, the number of predicted failed bits (P_{extr}), in the predicted extrinsic zone (see Fig. 13.), are compared for each model and both datasets. From this population, it is possible to forecast statistically the number of samples that should be in the extrinsic population after stress (N_{pred}).

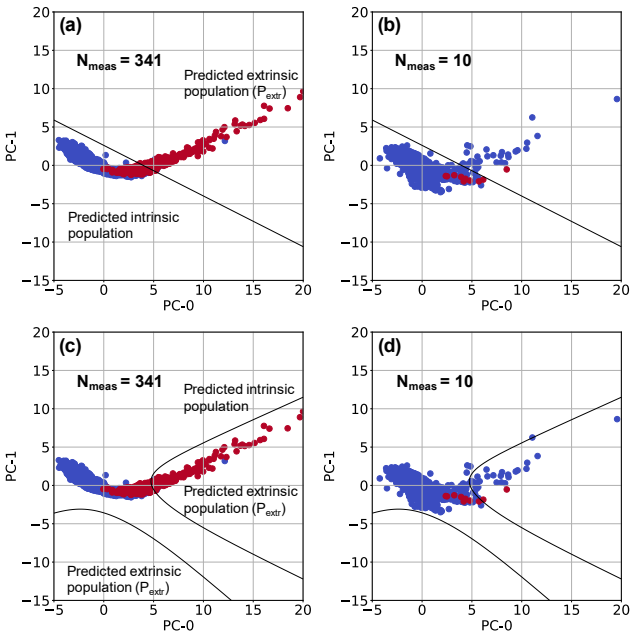
N_{pred} is calculated from the entire population in the predicted extrinsic zone (P_{extr}), weighted by the metrics of the trained model. For example, in the case of the linear model trained by die 3 that contains N_{meas}=253 failed bit cells, P_{extr} is 155 samples. However, according to its corresponding metrics in Table II, the precision indicates that 83% will statistically fail (128 samples) and the recall informs to us that P_{extr} represents 51% of the total degraded bit cells population (253 samples). Thus, for a given t_0

dataset, the extrinsic population size is statistically calculated thanks to the formula (1).

$$N_{pred} = P_{extr} \times Precision / Recall \quad (1)$$

The accuracy is not considered in (1) because it is primarily influenced by the intrinsic population due to the imbalanced dataset configuration.

Both t_0 datasets are transformed and projected into the principal component space determined in Section IV.C and the calculated models are traced. The tested linear model is illustrated in Fig. 13.a,b, the tested polynomial (degree 3) model is depicted in Fig. 13.c,d and the RBF model test result is shown in Fig. 13.e,f. Therefore, for each model and dataset, P_{extr} is measured and N_{pred} is calculated by using (1) and the metrics in Table II. The results summarized in Table IV show that the number of failing memory cells is similarly predicted by all three models for both datasets at t_0. It is evident that no models predicted exactly the real number of extrinsic memories, but they are consistent. Indeed, in the case of large extrinsic population dataset, the models have roughly predicted 300 failed memories instead of 341, and in the other case, 60 failed memories were roughly predicted instead of 10. Consequently, the models appear to provide more accurate predictions when P_{extr} is high, while greater errors are observed when P_{extr} is low. This uncertainty can be reduced by minimizing the population overlap in the principal component space at t_0. One approach would be to measure the drain current at a higher drain voltage than that used in this experiment. Nonetheless, the models are sufficiently effective in distinguishing the more impacted dies from the less impacted ones. Thus, these results demonstrate that the three models trained with the dataset from die 1 enable the prediction of the extrinsic population and provide reliable estimates for anticipating the number of memory bit cells that will degrade before stress.



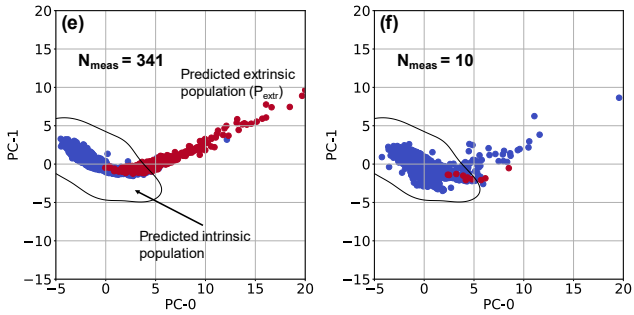


Fig. 13: Comparison of the predicted extrinsic population (P_{extr}) at t_0 for two different datasets: one with more extrinsic memories (a, c, e) and one with fewer extrinsic memories (b, d, f). The three classifier models trained in Section IV.C. are tested: Linear (a, b), Polynomial (c, d), and RBF (e, f).

TABLE I. MULTIPLE MODEL COMPARISON FOR PREDICTING EXTRINSIC POPULATION SIZE FOR TWO T_0 DATASETS

N_{meas}	Linear		Polynomial (degree 3)		RBF	
	P_{extr}	N_{pred}	P_{extr}	N_{pred}	P_{extr}	N_{pred}
341	189	307	161	300	222	306
10	36	58	29	54	47	65

V. CONCLUSION

In this paper, we have demonstrated the versatility of the SuperCAST test vehicle as an effective tool for extensive data collection. Its ability to measure multiple memory cells in parallel yielded reliability results that are comparable to those obtained using the CAST for equivalent memory transistors. Furthermore, the addressable nature of the SuperCAST enabled the measurement of various electrical parameters for each single cell, facilitating a comprehensive analysis of hot carrier degradation variability. Principal Component Analysis (PCA) was employed to establish correlations between electrical properties measured before and after stress, revealing that the extrinsic and intrinsic populations could be effectively separated using a classifier model at the initial test (t_0). Several prediction models were explored to assess the impact of hot carrier degradation on the memory bit cells. While a graphical linear model was found to have limitations in accurately identifying weak memory cells and distinguishing affected dies, a set of three models—automatically determined by the algorithm—were evaluated. The comparison of these models demonstrated similar prediction accuracy, with the quality of the electrical measurement conditions being the primary factor influencing performance. Overall, the results show that the SuperCAST test vehicle, coupled with PCA, is a powerful methodology for statistical analysis and predicting reliability failures based on initial memory bit cell variability. This approach is highly valuable in the context of semiconductor technology development, as it aids in the early identification of process failures, thereby accelerating the process flow qualification.

REFERENCES

[1] P. Capeletti R. Bez, D. Cantarelli, D. Nahmad, L. Ravazzi. "Cast: An electrical stress test to monitor single bit failure", *Micro. Rel.* vol 37, No 3, pp.473-481, Mar, 1997, doi: 10.1016.0026-2714(95)00214-6

[2] J. Plantier, H. Aziza, J. M. Portal, C. Reliaud, A. Regnier, J.L. Ogier. "Retention test and electrical stress correlation to anticipate EEPROM tunnel reliability issues", *Int. Semi. Dev. Res. Symp.*, College Park, MD, USA, 2009.

[3] T. Kempf, V. Della Marca, L. Baron, F. Maugain, F. La Rosa, S. Niel, A. Regnier, J.-M. Portal, P. Masson, "Threshold voltage bitmap analysis methodology: Application to a 512kB 40nm Flash memory test chip", *Int. Rel. Phys. Symp.*, Burlingame, CA, USA, 2018, pp. 6E.41– 6E.48, 2018.

[4] V. Della Marca, J. Guillau-Tavernier, P. Laine, F. Melul, M. Bocquet, T. Kempf, L. Welter, J. M. Moragues, A. Regnier, J. M. Portal. "SuperCAST: a full free addressable memory array," *Int. Conf. on Micro. Test Struc.*, Cleveland, OH, USA, 2022, pp. 1-4.

[5] R. Bro R, A. K. Smilde. Principal component analysis. Analytical Methods, vol 6, 2014, pp 2812-2831.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] P.M. O'Neil, "Production Multivariate Outlier Detection Using Principal Components," *Int. Test Conf.*, Santa Clara, CA, USA, 2008.

[7] D. Drmanac; N. Sumikawa, L. Winemberg, L. Wang, M.S. Abadir, "Multidimensional parametric test set optimization of wafer probe data for predicting in field failures and setting tighter test limits," *Design, Auto. & Test in Europe*, Grenoble, France, 2011.

[8] A. Haggag; N. Sumikawa, A. Shaukat; J.K. Jerry Lee, N. Aghel, C. Slayman, "Mitigating "No trouble found" component returns," *Int. Rel. Phys. Symp.*, Monterey, CA, USA, 2015.

[9] N. Sumikawa, D. G. Drmanac; L. Wang, L Winemberg, M. S. Abadir, "Forward prediction based on wafer sort data — A case study," *Int. Test Conf.*, Anaheim, CA, USA, 2011.

[10] S. Perrin, K. Alkema, V. Della Marca, T. Kempf, O. Paulet, M. Arteaga, M. Akbal, B. Arrazat, J. Metz, J. M. Moragues, A. Regnier, M. Bocquet, J. M. Portal, "Predictive method of charge storage memory degradation on a dedicated 4kb test vehicle," in *Proc. IEEE Int. Integr. Rel. Work.*, Fallen Leaf Lake, CA, USA, 2024.

[11] B. Szelag, M. Dutoit, F. Balestra, "New Findings on Hot Carrier Effects in Bulk Silicon MOSFETs," ESSDERC '96: Proceedings of the 26th *Euro. Solid State Dev. Res. Conf.*, Bologna, Italy, 1996, pp. 859-862.

[12] C. M. Bishop. "Continuous Latent Variables" in *Pattern Recognition and Machine Learning*. Berlin, Germany, Springer, 2006, ch.12, sec.1, pp. 561-570.

[13] G. Batista, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explorations Newsletter*, Vol. 6, no. 1, pp. 20-29, June, 2004, doi: 10.1145.1007730.1007735.