Internship Report: Analysis of Wafer Test Data

Karthi Srinivasan, EE17B070

July 25, 2020

Contents

1 Background

Testing is an integral part of the semiconductor manufacturing workflow. Any part that is designed needs to verified and tested before being shipped out to a customer, to ensure that all specifications are being met and that there are no defective parts being shipped. The testing process of a wafer, which typically consists of a few thousand dice, happens in three steps: Parametric Test (or) E-test, Multiprobe Test, and Final Test. Each testing process is progressively more expensive and time-consuming per test than the previous one.

1.1 Parametric Test

This is a process to ensure integrity of the wafer and process itself. The tests are not dependent on which part is being manufactured on that particular wafer, but only on the process node that is used for that particular wafer.
During fabrication, certain devices of pre-determined specifications (scribe devices) are etched onto the wafer, as determined by a Process Control Document(PCD). These devices are typically one each of the different kinds of resistors, capacitors and FETs that can be manufactured in the given process. The PCD also lists out the different measurements that need to be made on these devices and what are limits within which they should fall.
Parametric Tests (also known as PCM, short for Process Control Monitor) are typically done only at around a dozen locations per wafer.

1.2 Multiprobe Test

This is a part-specific test program that is defined by the designers and test engineers that are responsible for the particular part that is being manufactured. The multiprobe, as the name suggests, contains multiple probes and hence a few dice are tested at the same time. This test typically tests most of the items that are listed in the part's datasheet, among others. Some specifications which are guaranteed by design or characterization are not tested. Every die on the wafer is tested completely, in this program. For example, in the case of the LIN transceiver, various currents in different modes of operation, transition times from one mode to another and duty cycle values are measured, among other common measurements like contact resistances and leakage currents through various pins.

1.3 Final Test

This is the last stage of testing that is done before the part is shipped out. This test happens after assembly and packaging of the die. For parts that have an extensive multiprobe test program, the number of tests in this program is usually very low as it is expensive to perform. However, there are some parts which do not undergo multiprobe test and hence have extensive final tests. One of the intents of this test is to detect defects that may have been caused during the packaging process itself.

2 Objective

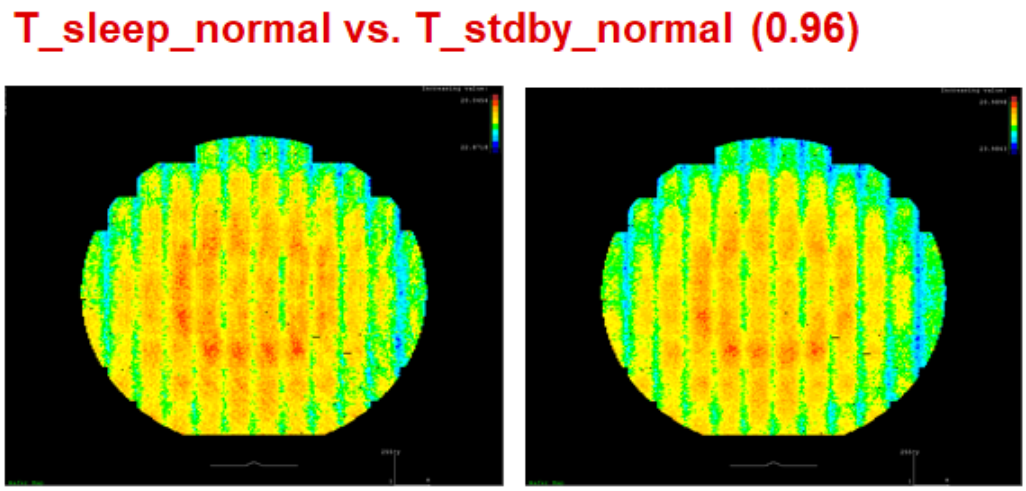
Given that there is a large amount of data available for each of these types of tests and a large number of tests themselves, it is inevitable that there will be some degree of interdependence/redundancy in the tests that are being performed, and any such relations that can be discerned will be true. The aim of this project is to find these correlations and see if any causal relationships can be found, to use these relationships to understand the variations in more detail, and to make the test program itself more efficient.

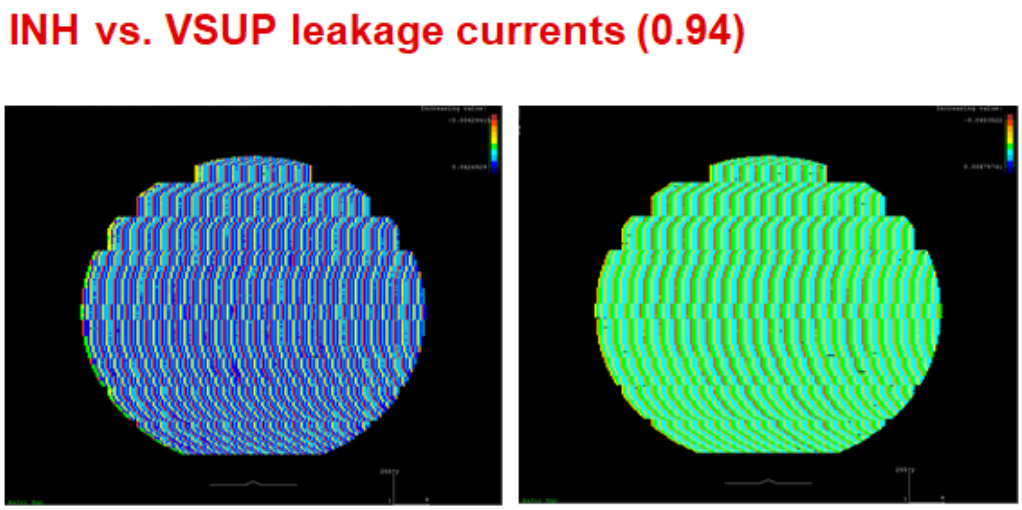
3 Initial Results

This section details some of the initial findings on spatial correlations seen on Multiprobe-to-Multiprobe and Multiprobe-to-PCM tests.

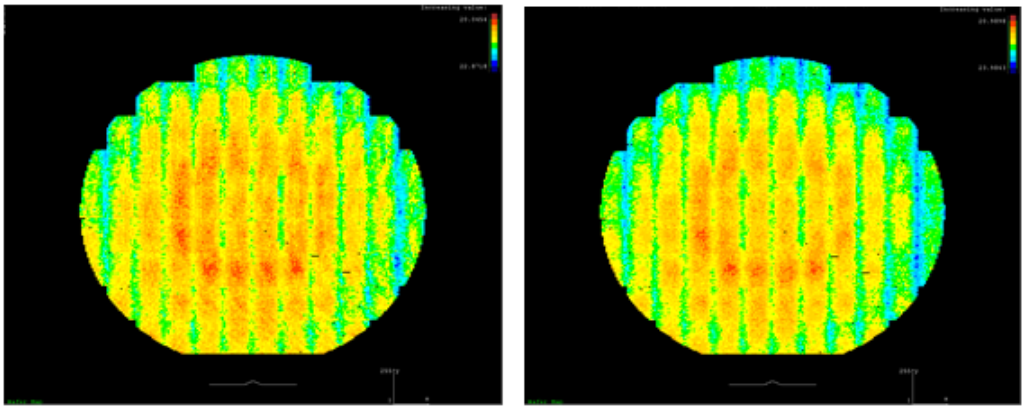
3.1 Spatial Correlations: Multiprobe-Multiprobe

The wafermaps shown below are from one fablot of the part. Each datapoint represents one die. The spatial correlation coefficient is also shown, which is a number that ranges from -1 to 1. It is the familiar Pearson's correlation coefficient over this spatial data.

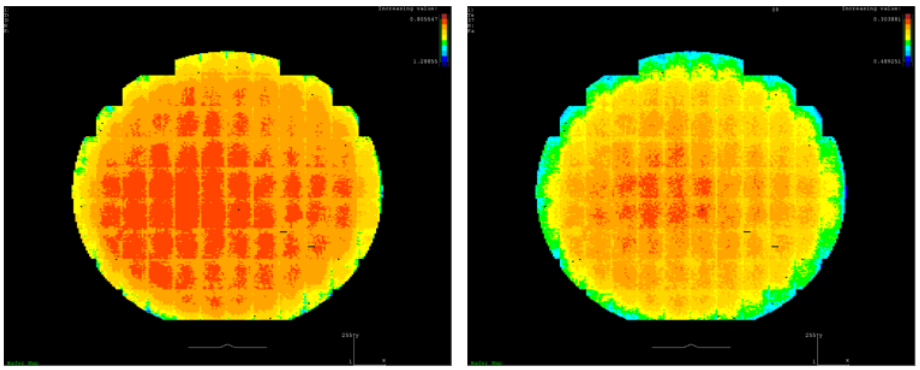




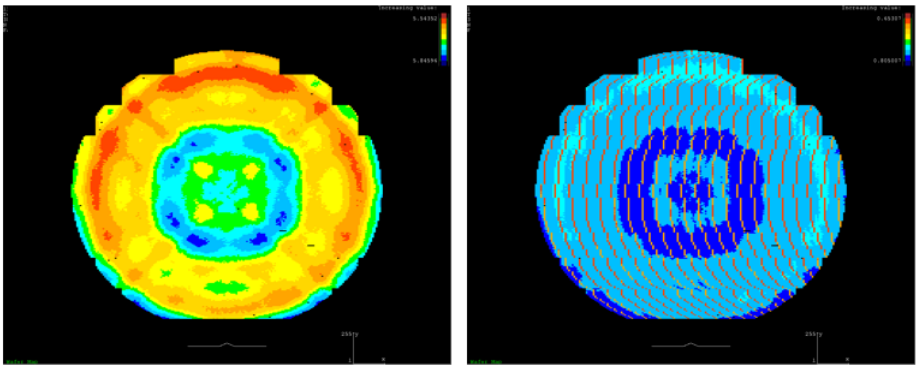
T_sleep_normal vs. T_stdby_normal (0.96)



LIN current vs. LIN Defect Type 1 (0.97)



Type 2 Defect vs. Wake current (0.94)



A few inferences can be made from the above wafermaps, and a few patterns can be explained.

1. Parameters like leakage currents and contact resistances depend on the probe movement pattern, which is shown by the fine striations in the second wafermap. The quality and extent of contact that the probe makes with the wafer itself influences the measured value of these parameters.

2. Type 1 defect tests and their associated measurements are dependent on the reticle/photomask movement pattern across the wafer itself. During fabrication, the reticle is moved in discrete steps across the wafer. The discontinuities that happen due to the edges of the reticle influence the values of these parameters, as can be seen from the rectangular grid pattern in the fourth wafermap.

3. Stress tests in general, are indicative of currents drawn in different modes of operation of the device itself. Type 1 defect tests for VSUP quite strongly indicate the normal and sleep mode currents. Type 2 defect tests correlate well with standby mode currents. This is expected, as stress tests essentially test the strength of the transistors, by applying high voltage stress to gates of the transistors and measuring the drawn current through the pin. Hence, they would also predict the currents drawn through the pins in different modes of operation.

4. There are, however, some complicated and slightly chaotic patterns such as the one seen in the last wafermap that does not yield itself to a straightforward explanation. It it likely a combination of many of the patterns seen earlier, indicating that all of the factors described above play a role in determining the pattern itself.

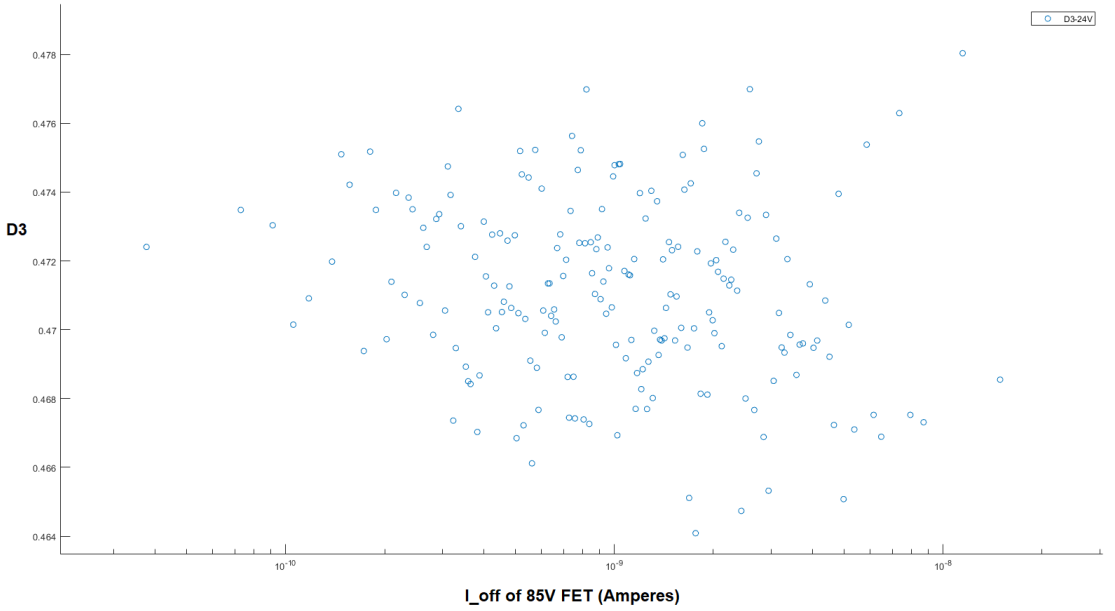
3.2 Multiprobe - Parametric Correlations

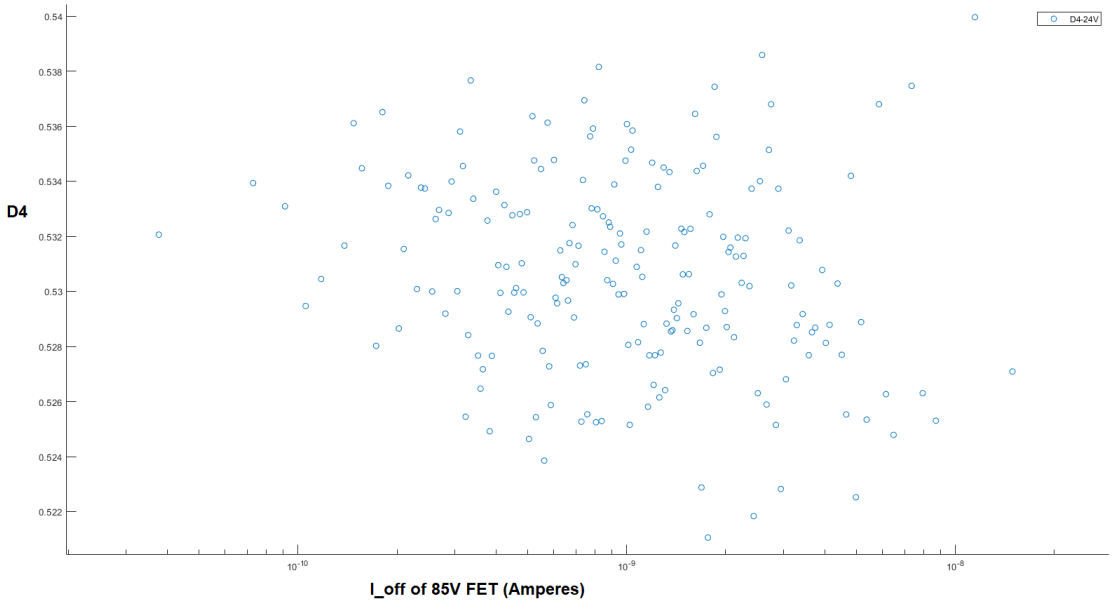
The correlation between multiprobe tests and parametric tests cannot be done on a simple one-to-one basis as the number of scribe devices, i.e. number of parametric tests done per wafer is orders of magnitude smaller than the number of dice per wafer. Hence, the multiprobe tests that are done need to be spatially averaged around the locations of the scribe devices.
This process results in a few hundred multiprobe test locations being mapped to one PCM test location, although the exact number depends on the reticle and die size. There are around a dozen PCM test locations per wafer, typically. There are less than 10 wafers of this part in this fablot and hence, this results in less than 100 data-points in total.
As is evident, this is not a large enough number to draw conclusions from. However, for the sake of completeness, correlation analysis for this data was also done.
Most correlations that were found were between leakage currents and contact resistances in multiprobe and off-state currents and sheet resistances in parametric. These were deemed to be insignificant as contact resistance and leakage current variations don't affect the functioning of the device in any way and are just characterization parameters.
One parameter, however showed significant correlation with duty cycle tests (Multiprobe), the off-state current of the 85 volt DE-NMOS transistor (PCM). The duty cycle tests are treated in detail in the following section. However, the number of datapoints for this was extremely scarce, less than 20 over the entire fablot. This is due to this PCM test parameter being tested only at much fewer locations per wafer, than the typical amount. To add to this, all these locations are spatially concentrated towards the north-east corner of the wafer, rather than being spread apart uniformly.

3.3 Monte Carlo Simulations

To combat the sparsity of data described in the previous section, and to verify if the correlations that are being seen are actually true, Monte Carlo simulations were used to gather more data from simulation itself, as more wafer data was not available for the original part.
To measure the off-state current of the 85V NMOSFET, an extra device was added to the schematic with gate and source grounded, and drain connected to VSUP, which is the supply voltage. Next, the process parameters of all 85V FETs in the designed were varied through a set of Monte Carlo simulations (300 runs).

The plots of two duty cycle tests that initially seemed to correlate with the off-state current vs. the off-state current itself, of the 85V NMOSFET is shown below.





As can be plainly seen, there is no actual correlation between the two parameters that were originally thought to be correlated. The perceived correlation has been confirmed to be the consequence of the sparsity of data.

4 The Duty Cycle Tests

The duty cycle tests are large set (64 for the original part, 70 for for another similar part, described later) of tests that measure the duty cycle values of the PWM wave at a certain crucial pin, under various combinations of supply and load conditions. The exact tests conform to ISO standards.
These tests are quite time-consuming to perform. They take about 1 second out of the 2 seconds that the entire multiprobe test program takes, despite being only 64 in number, out of 250+ multiprobe tests.
Upon closer examination of the correlation matrix of the data, it is evident that many of these duty cycle tests are closely correlated with each other. But also, more importantly, there are no significant correlations between duty cycle tests and any of the other multiprobe tests, that are not duty cycle tests. This yields a useful fact: The problem of analysis of duty cycle tests can be viewed as a self-contained, isolated problem. The goal of this analysis is to, if possible, find a method to perform a reduced number of duty cycle tests, thereby making savings on test-time but still capture all the variation in the data itself.

5 Methodology

This section outlines the mathematical techniques that have been used for the analysis of the duty cycle test data.

5.1 Principal Component Analysis

PCA is a statistical method for finding a low-dimensional approximation, described by an orthogonal set of vectors, for a set of data-points that are inherently higher dimensional. Given a collection of points in two, three, or higher dimensional space, a "best fitting" line can be defined as one that minimizes the average squared distance from a point to the line. The next best-fitting line can be similarly chosen from directions perpendicular to the first. Repeating this process yields an orthogonal basis in which different individual dimensions of the data are uncorrelated. These basis vectors are called principal components, which is where the method gets its name from.
PCA is done by performing the following two steps:

- Calculating the data covariance (or correlation) matrix of the original data.
- Performing eigenvalue decomposition on the covariance matrix.

5.2 Finding Minimum Distance from a Vector to a Linear Vector Space

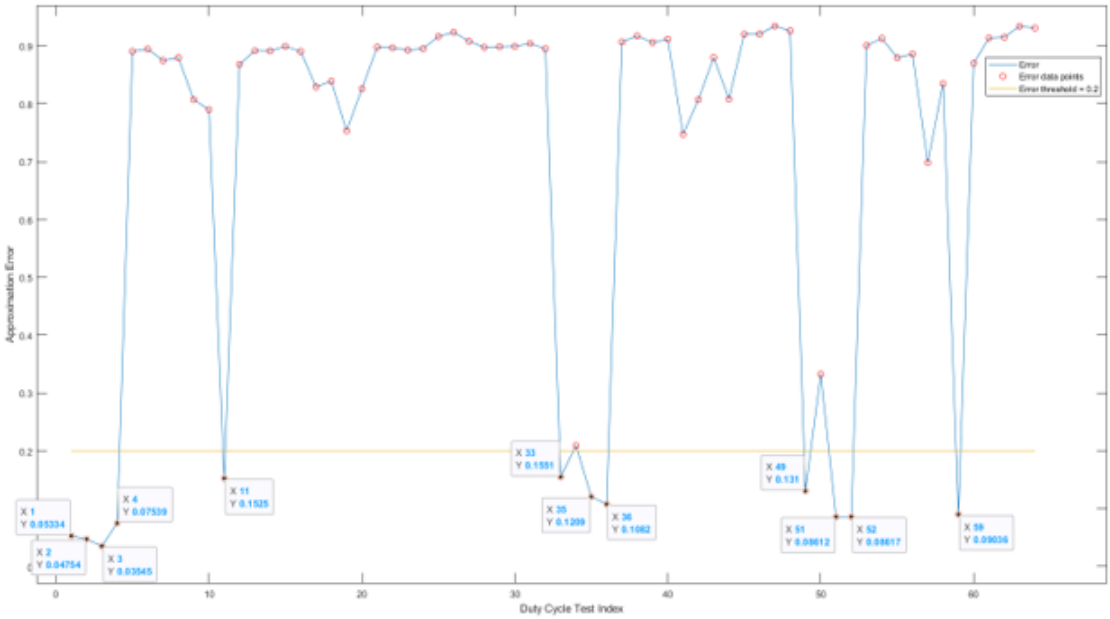
Given a vector and linear vector space defined in terms of its basis set, the problem of finding the minimum distance from the given vector to any point in the LVS can be formulated as a convex optimization problem as follows:

$$\text{minimize} \quad ||Bw - v||_2 \tag{1}$$

where B is matrix consisting of the basis vectors as its columns. w is the variable vector being minimized, which provides the weights that need to be assigned to each basis vector to obtain the point that corresponds to that of minimum distance from the given vector. v is the given vector for which we want to find the minimum distance.

6 Data Analysis and Results

As mentioned earlier, PCA has been used to analyze the duty cycle test data. PCA, however only gives 'directions' along which the variation is maximum, via the eigenvectors, and the extent of this variation, via the eigenvalues. This is not sufficient since these directions will not necessarily correspond to a 'physical direction', i.e. it will not be along the directions that represent any of the single duty cycle tests that can be performed. It will be a composite of many. To resolve this problem, we compute which 'physical directions' lie close to the space spanned by the eigenvectors.
Note that, if all eigenvectors are considered, then all duty cycle vectors will lie within the space, by virtue of completeness, i.e. the 64 eigenvectors span all of $\mathbb{R}^{64}$. Hence, we consider only those eigenvectors that have an eigenvalue above a certain threshold, to find which duty cycle vectors actually lie even within the approximate space.
The graph of trying to fit all 64 duty cycle test vectors, which are essentially represented as the 64 elementary basis vectors, is shown below. The error on the y-axis ranges from 0 to 1, where 0 indicates the duty cycle test vector lies exactly within the space spanned by the eigenvectors, and 1 indicates that it is orthogonal to the space.

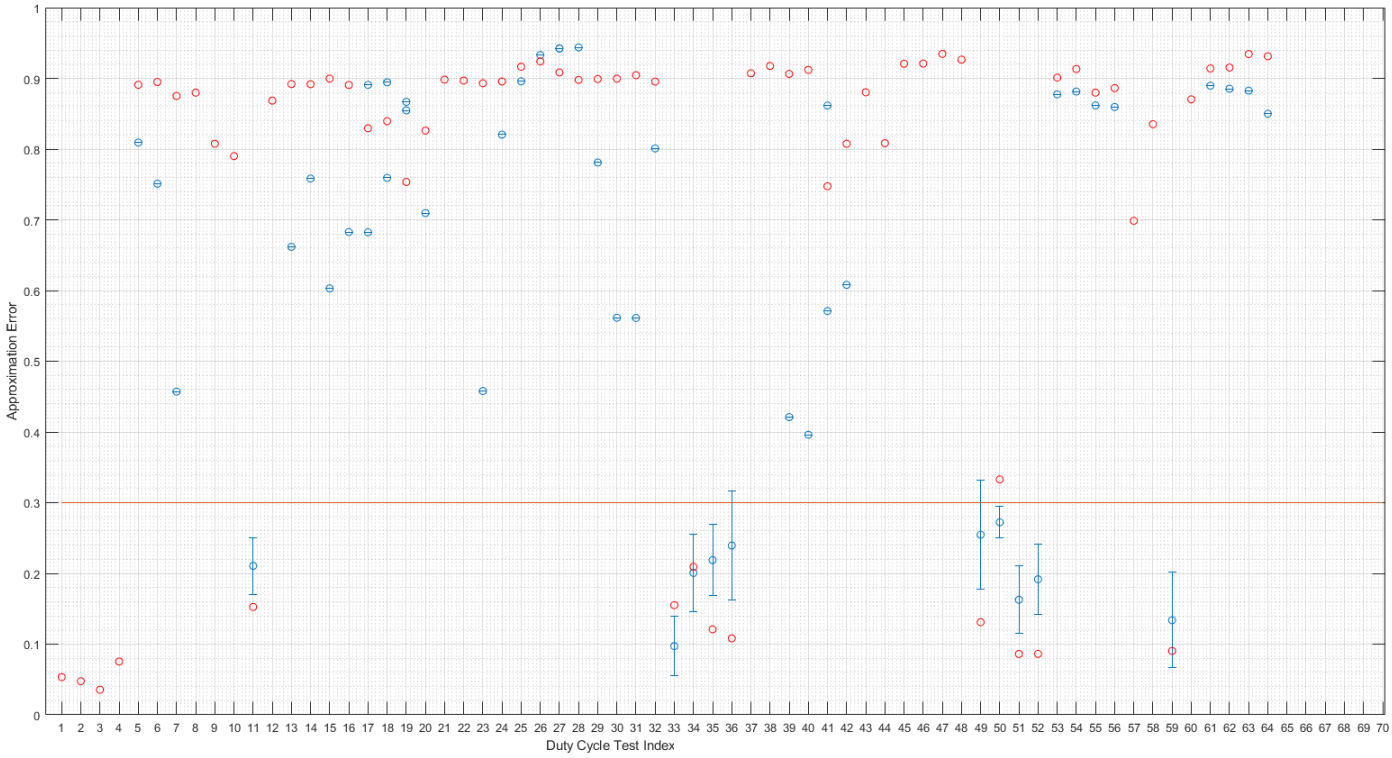


An error threshold of 0.2 was used to choose which tests to consider in the optimal test set.

An optimal test set of 12 tests, marked in the graph above was found.

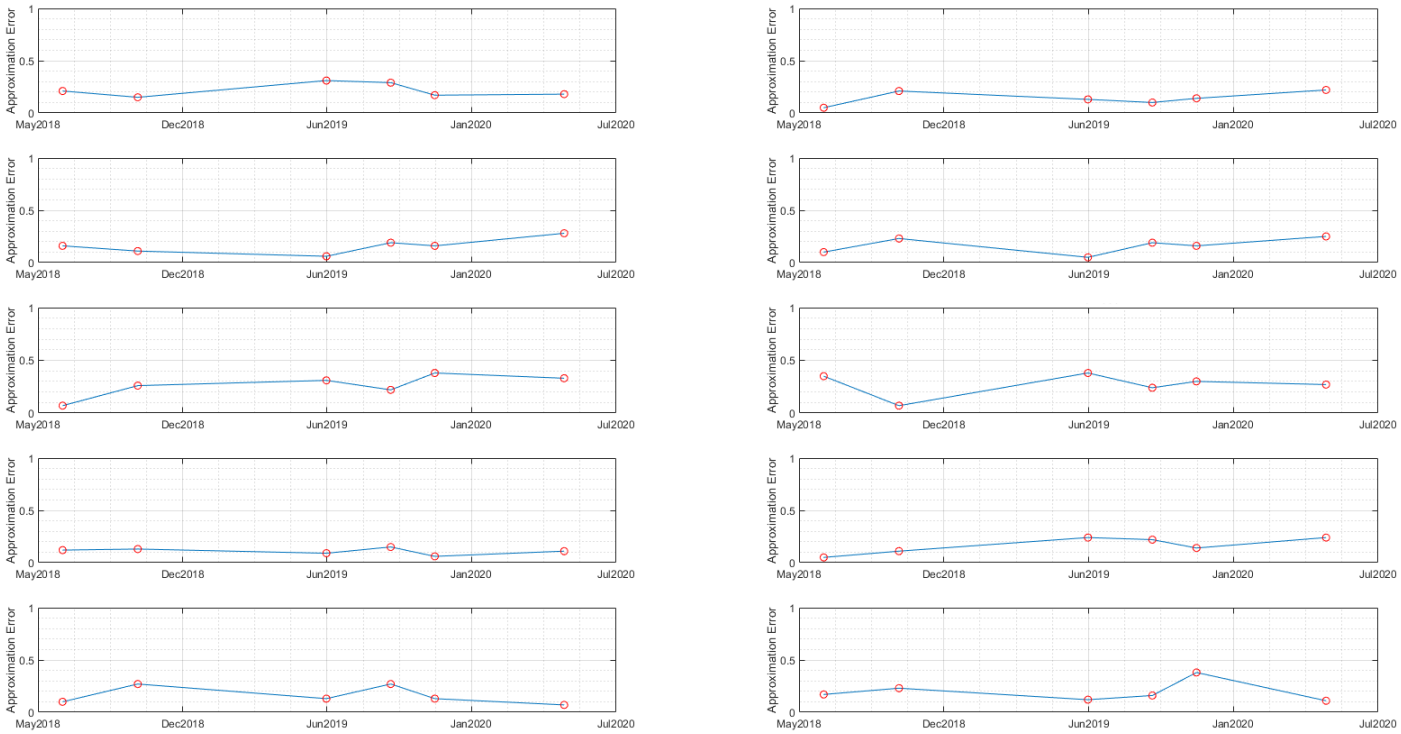
7 Verification with Data from a Similar Part

As mentioned earlier, the data from which the model was derived only consists of one fablot. Further, all tests are done at only one temperature. Hence, there is a need to test the robustness of the model as process parameters vary from fablot to fablot. Since, there was no additional data available for the original part, data from a similar part, was used to verify whether the model still worked.
This data is from 6 different fablots, spanning roughly two years, 2018-2020. The model was first refitted to 1 fablot from the similar part, and then was tested on the remaining 5 fablots.

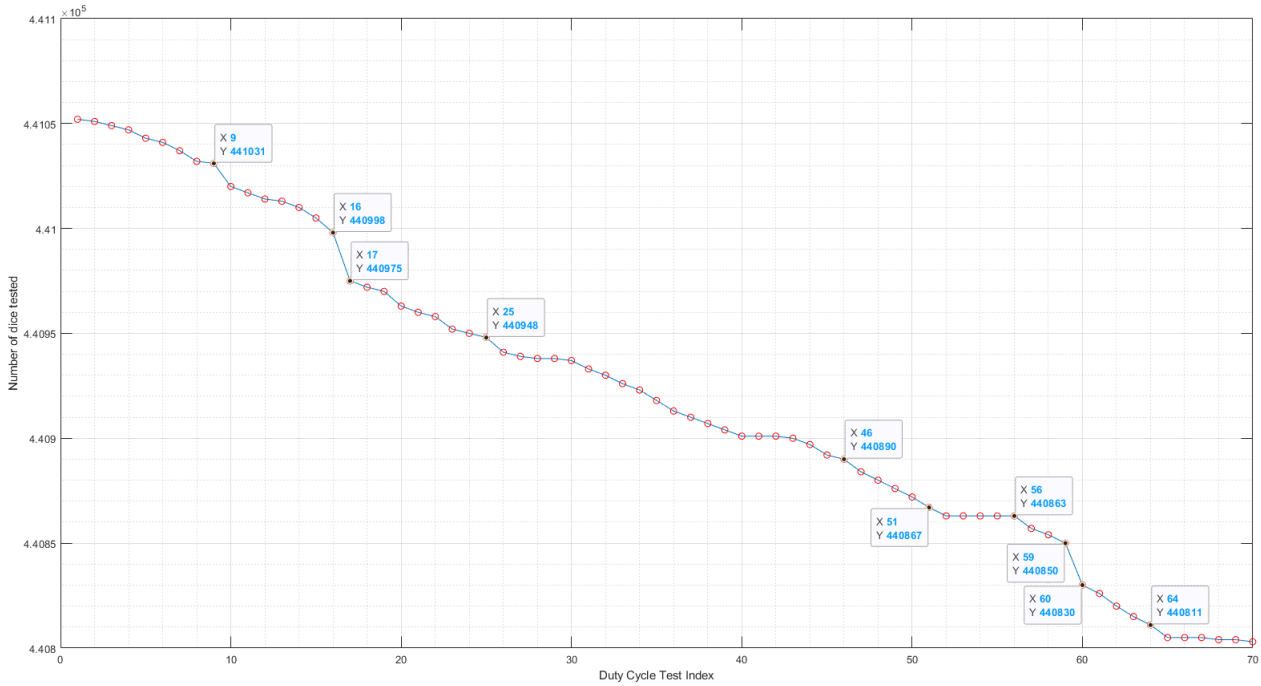


The above graph is the analogue to the earlier graph showing errors in approximation. Note that the tests for the two parts are not in the same order, and hence re-indexing needs to be done to get an accurate picture. Data for teh new part is shown in blue, and the original is in red. As can be seen, many of the same duty cycle tests that form the minimal test set in the original case also form the minimal set in the new part. However, there are a few exceptions. The cluster of 4 tests on the extreme left have no matches with the new part's tests. This is due to the fact that these exact testes are not being performed in the new part's test program. Hence, the new optimal test set needs to be changed slightly to account for this. All tests that lie below the 0.3 error threshold for the new part have been considered to form the optimal test set.
The error bars in the graph represent the $max - min$ of the error over the 6 different fablots.
A few other points to note from the graph:

- Not all tests in the first part's program have a corresponding one in the new one, and vice versa.
- Some original part's tests have 2 corresponding tests in the new part's test program. This is due to the fact that the same test is sometimes repeated in the new part's test program, for hitherto unknown reasons.



The above graph shows the error over time of the 10 tests in the optimal test set. The error does have some variation over fablots but largely stays within reasonable bounds. The above graph was generated by first refitting/retraining the model on one fablot of the new part's data, and applying it to the remaining 5 fablots for verification. The important inference from this is that the optimal test set does not change over time.



The above graph shows the number of dice tested versus the test number, in chronological order, i.e. the order in which they are performed.
In the test program, a part that fails a certain test is immediately deemed defective and is not passed through the subsequent tests. Hence, the difference in number of dice between a certain test and its immediate successor is the number of dice that failed that test.
If the optimal test set does indeed capture most of the variation in the test data, then it is logical to expect

it to also cumulatively catch the most failures also. This translates to the chosen tests occurring just before steep dips in the graph, i.e. a lot of failures.
As can be seen, this is largely true, although not perfect. The chosen tests do precede the steepest dips in the graph, but there aren't sufficient tests to cover the entire range. Note that the test indices here will be different from those in the previous graphs as these aren't re-indexed to the old part's test sequence, to preserve the chronological order.

8 Conclusions and Suggestions

The parametric and multiprobe test programs of two parts were analyzed in this project.

8.1 Duty Cycle Test Program

The analysis performed shows that there is a significant level of redundancy in the duty cycle tests of the parts' multiprobe test program. Given that these tests take a significant chunk of time, around 50% of the entire multiprobe test time, it would be important to consider whether some tests can be eliminated. Initially, the test program can be restructured to have the chosen tests occur at the beginning. Once this is done, the tests that occur later in the test program would hypothetically have a zero/negligible yield loss. This would be the final confirmation required for actually eliminating these extraneous tests themselves.

8.2 Generalizing the Method

The method (PCA, specifically) used to analyze the duty cycle test data is a general statistical method, with a wide range of applications. The analysis does not hinge upon the fact that the tests were 'duty cycle tests' in any way. Any similar set of tests can be analyzed using the same method, with largely similar results.
Further, the method is quite easy to implement, and is computationally efficient. Not all steps in the workflow are available in TI's toolkit as of now. A large part of the analysis was done through custom scripts, written primarily in Python and MATLAB. These can be quite easily integrated into the existing set of tools, so that it may have a larger reach in terms of usage and usefulness.

8.3 Final Thoughts

Although the methods mentioned above may sound very promising, it is important to know what its possible shortcomings are, and whether these can be dealt with, effectively.

- PCA requires a significant amount of data to find a good model. This is typically not a problem as there are hundreds of thousands of dice even in one fablot, and hence, there is an ample amount of data. However, care must be taken to ensure that this data isn't biased in any particular way. For example, the original part's data that was initially used was only from one fablot, and one temperature, so any variations across these parameters would be effectively invisible. Further verification of this model lead to the model being changed slightly. Thus, unbiased data of significant amount is necessary.

- The underlying data could be quite uncorrelated. Although this is an extremely rare case, it is possible to construe a scenario where the test set is already optimal, and every characteristic of the part is captured by this test program. Although not explicitly a disadvantage, this type of scenario is one where PCA cannot add any additional value.

- There may not be a large set of similar tests over which to run the analysis. In this case, teh part had 64-70 duty cycle tests, which are essentially measuring the same underlying quantity under different conditions. If such a set does not exist, then it may be hard to justify the use of PCA. However, the definition of 'similar tests' itself is quite vague here, and tests that superficially do not look to be related could be revealed by data to be so.

9 Impacts and Future Steps

The statistical analysis of the part's data revealed some key points about the nature of the design and the testing program itself. In particular, the duty cycle tests analysis could potentially lead to reordering or removal of tests, thereby making savings on cost and time during the multiprobe testing. More importantly, the methods discussed and implemented would be valuable additions to the existing toolkit for analyzing wafer data and deriving useful results from them.

The analysis also led to understanding of existing redundancies, intentional or otherwise, in the test program, and spatial distributions of various test parameters, some of them exhibiting complex patterns. Some of these could be explained trivially, while others required deeper study. Further, the relative independence of multiprobe tests on parametric tests was also uncovered.

There are a few important future steps to be taken. The foremost among them is reordering the test program for the this part and checking whether the yield loss for the trailing tests goes to zero sufficiently quickly. This would be a certificate for the validity of the reduced test set being adequate. Upon this verification, the tests could be removed entirely from the program.

Secondly, the method could be extended to other parts that have a large number of similar tests, and to sets of seemingly dissimilar tests also. The fact that the analysis was to be run only on the duty cycle test set was deduced somewhat heuristically. There is nothing that prevents the analysis from being run on the entire test set, although, the results might be harder to make sense of.

Finally, the scripts can be integrated into existing software to streamline the whole process of running the analysis, without having to deal with implementation details each time.

10 IEC-to-HBM Waveform Conversion Network

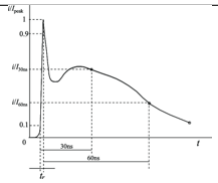
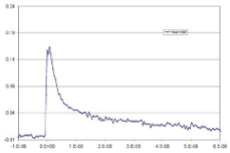
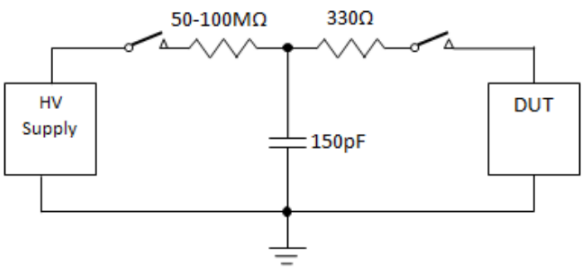
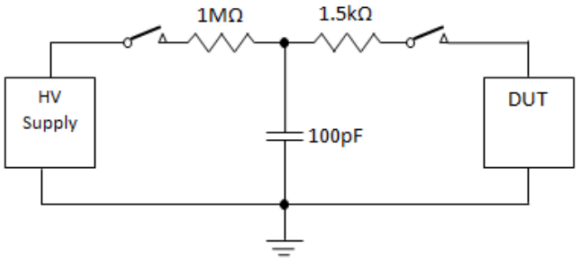
This section outlines a side project whose purpose was to design a network that can produce an approximate HBM ESD waveform (for debugging purposes only) when excited with an IEC standard ESD waveform.

10.1 Background

Electrostatic discharge is the sudden flow of electricity between two electrically charged objects caused by contact, an electrical short, or dielectric breakdown. Integrated circuit components are particularly sensitive to electrostatic discharges, and can suffer irreparable damage if not protected from these events properly. There are many standards for testing whether an IC can withstand electrostatic discharges. An internationally used standard is the IEC 61000-4-2, for system level testing. Apart from, there are also pin-to-pin testing standards for the Human Body Model (HBM). One such standard is the JEDEC JS-001. As of the moment, the lab has an IEC ESD simulator gun, but no similar provision for HBM. Hence, the goal is to design a system that sits between the IEC ESD gun and the device under test (DUT) such that the DUT effectively sees an HBM-like discharge.

10.2 IEC vs. HBM

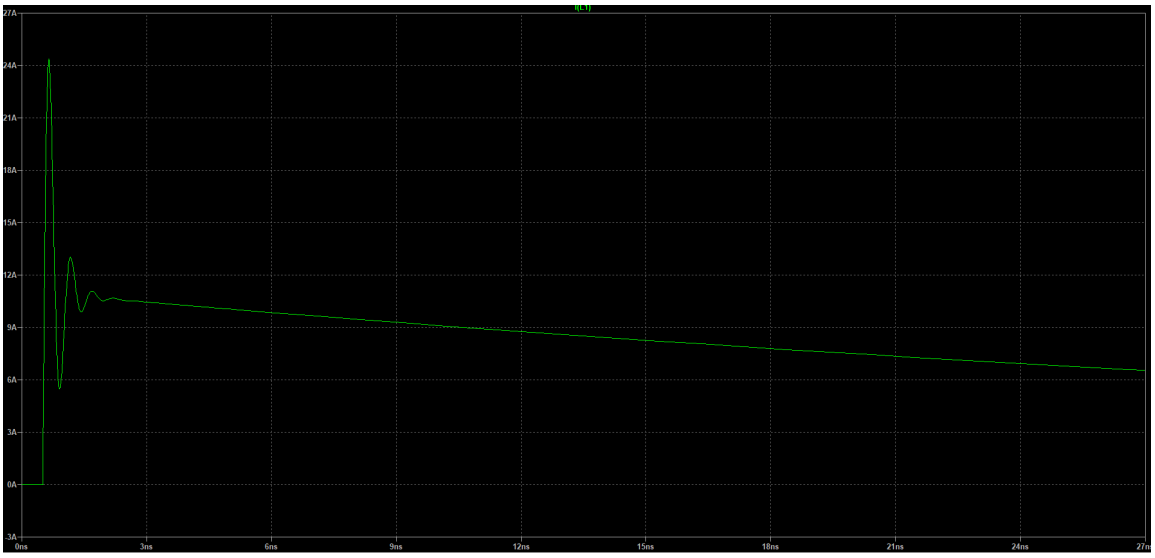
The following table outlines the key differences between the two standards.

IEC-61000-4-2	HBM JEDEC JS-001
System Level Test	Pin-to-pin test
Fast rise time (0.7-1ns)	Slower rise time (5 - 10ns)
Faster decay time: 50ns	Slower decay time: 130-170ns
Wave shape determined by parasitics	Relatively insensitive to parasitics
Looks like higher order response	First order response
	
 Simulation circuit of IEC6100-4-2	 Simulation circuit of HBM, MIL-STD-883

Note that the impedance that the DUT sees when looking into the discharge network is not the same for both cases. This needs to be taken into account when designing the network. Also, since the IEC waveform itself is very sensitive to parasitics, the new network should not disturb this waveform.

10.3 IEC Parasitic Simulation

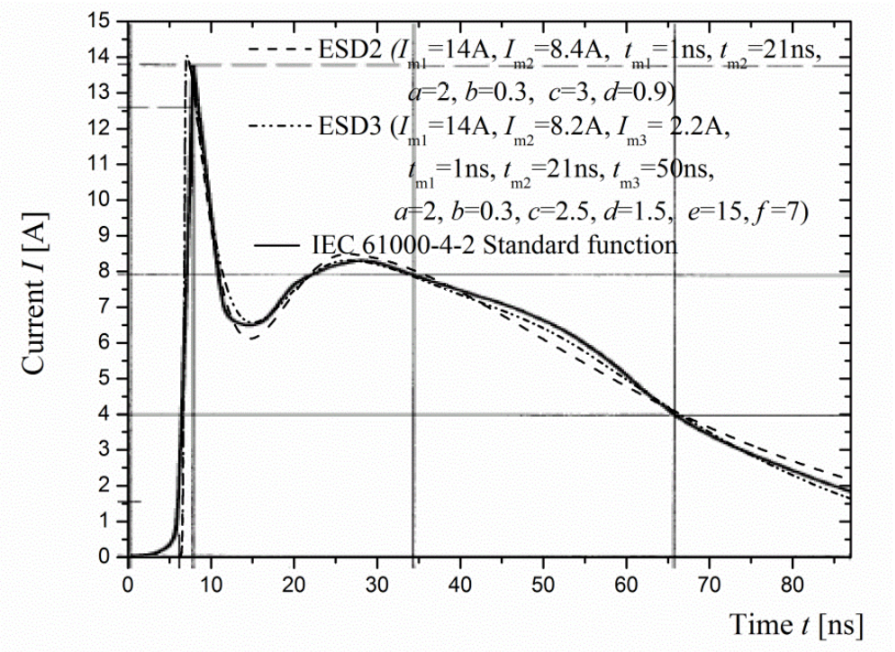
This section describes the simulation of the IEC discharge circuit with parasitics modeled, approximately. The current waveform through the 330-Ohm resistor is shown below:



The specifications of the wave are not exact, but the general shape is visible. This verifies the fact that the parasitics are the principal determiner in the discharge waveform of the IEC network.

10.4 Analysis and Design

Since we are dealing with extremely high voltages and currents, it is better to have a completely passive network, as these usually have high rated voltages and currents. The first step in this process was to find a functional approximation of the IEC ESD waveform itself. This work has already been done in an academic research paper. The functional approximation is shown below. ESD3 is the model that has been considered for this analysis.



Differentiable Functional Approximation:

$$i(t)=\begin{cases} I_{m1}(t/t_{m1})^{a}\exp[a(1-t/t_{m1})]+I_{m2}(t/t_{m2})^{c}\exp[c(1-t/t_{m2})] \\ \qquad +I_{m3}(t/t_{m3})^{e}\exp[e(1-t/t_{m3})], & 0<t<t_{m1} \\ I_{m1}(t/t_{m1})^{b}\exp[b(1-t/t_{m1})]+I_{m2}(t/t_{m2})^{c}\exp[c(1-t/t_{m2})] \\ \qquad +I_{m3}(t/t_{m3})^{e}\exp[e(1-t/t_{m3})], & t_{m1}<t<t_{m2} \\ I_{m1}(t/t_{m1})^{b}\exp[b(1-t/t_{m1})]+I_{m2}(t/t_{m2})^{d}\exp[d(1-t/t_{m2})] \\ \qquad +I_{m3}(t/t_{m3})^{e}\exp[e(1-t/t_{m3})], & t_{m2}<t<t_{m3} \\ I_{m1}(t/t_{m1})^{b}\exp[b(1-t/t_{m1})]+I_{m2}(t/t_{m2})^{d}\exp[d(1-t/t_{m2})] \\ \qquad +I_{m3}(t/t_{m3})^{f}\exp[f(1-t/t_{m3})], & t_{m3}<t<\infty \end{cases}$$

Once this form has been found, the further analysis is a simple Laplace domain analysis. The Laplace domain equivalent of this function and that of the HBM (which is straightforward, since it is just a first order exponential decay) was found. The ratio of the these two functions is the required transfer function. This results in a 4th order transfer function. However, the transfer function need not hold for all inputs, only for a pre-defined input which we already know, the IEC wave. This enables us to make an approximation to reduce it to a second order transfer function. This transfer function was then implemented through an RC network. The input and output current waveforms are shown below. IEC input current through source B4 is in green, while output current through the 2-ohm resistor is in purple.