Conquering Process Variability: A Key Enabler for Profitable Manufacturing in Advanced Technology Nodes (Keynote Speech)

Andrzej J. Strojwas*
PDF Solutions, Inc.
San Jose, California, USA
andrzej.strojwas@pdf.com
* also with Carnegie Mellon University, Pittsburgh, PA, USA

Abstract – Achieving the required time to market with economically acceptable yield levels and maintaining them in volume production has become a very challenging task in the most advanced technology nodes. One of the primary reasons is the relative increase in process variability in each generation. This paper will describe a comprehensive study of the main sources of variability and their effects on active devices, interconnect and ultimately product performance and yield. We will present benchmarking of yield loss components for different product classes. We will then propose several approaches for variability reduction in the design, yield ramp and volume manufacturing phases.

EVOLUTION OF YIELD LOSS MECHANISMS

In the older technology generations, manufacturing yield loss was dominated by random defects. By the time volume manufacturing started, systematic yield loss was typically insignificant. This situation started to change rapidly at the 130nm technology node in which the product layout systematic effects became more critical. More recently, due to challenging product performance requirements and increased process variability, parametric yield losses have become significant as well. This evolution in yield loss mechanisms is shown in Figure 1 for the most recent technology nodes in production, namely 130, 90 and 65nm.

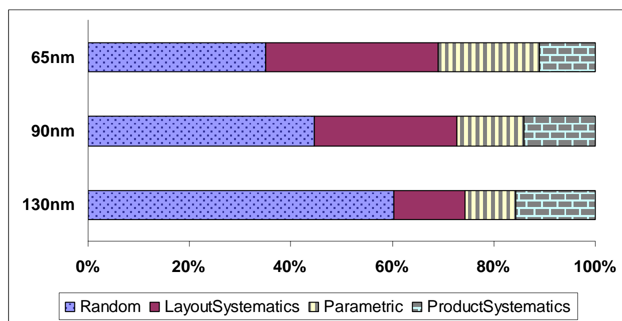


Figure 1 Evolution of yield loss mechanism breakdown.

As can be seen, for 90 and 65nm nodes, random defect limited yield losses contribute less than 50% to the overall yield loss and the layout systematic and parametric yield losses continue to increase. There is also a fourth category of manufacturing yield losses: product systematics that are only observable for specific products. These effects may be caused by the product reticle effects or specific product sensitivities to within-wafer non-uniformities.

This breakdown of yield loss mechanisms is, however, very much dependent on the class of products. Clearly, memory products with abundant redundancy exhibit different Paretos than SOC products. Moreover, even within SOC products, this breakdown varies among different product classes. Figure 2 demonstrates this finding for four different SOC product types: wireless, microprocessor, game chip and DSP. The data represents a snapshot of current 90nm fabrication processes with at least 3 different products in each category.

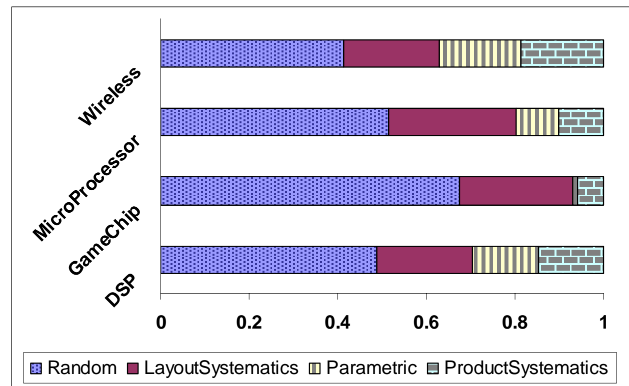


Figure 2. Yield loss mechanism breakdown per product type.

In the next section, we will examine the sources of process variability for the most recent technology nodes.

PROCESS VARIABILITY SOURCES

Process variability sources can be categorized based on the spatial hierarchy: lot-to-lot, wafer-to-wafer, within-wafer or within-die, or root causes: random or systematic. These sources create a complicated distribution of parameters that must be addressed by circuit designers. One of the key parameters is poly linewidth since it has the dominant effect on MOS transistor electrical performance. For 90nm technologies, more than 50% of the variance in poly linewidth comes from within-die (within field) variations; the next component is die-to-die. The key question is: what is the percentage of variance due to random variations (line edge roughness, optical aberrations) versus systematic effects due to neighborhood pattern dependent printability? Figure 3 demonstrates that the percentage of systematic variations increases with device scaling; for 90nm NMOS transistors, it reaches 40% of the overall Across Chip Variance (ACV). Random line edge roughness (LER) is actually an important contributor to the overall variations and can be on the order of

15% of Ldrawn (3σ) for 90nm technology. Variations in critical dimensions are, however, not the only contributors to transistor performance variability. New materials (e.g., NiSi) and stress/strain engineering contribute to the increased variability in scaled-down technology nodes. This is illustrated in Figure 4 which shows that the normalized standard variation in narrow 65nm NMOS transistor drive current can be greater than 10%. The variation is even greater in the SRAM transistors which use the push-rules.

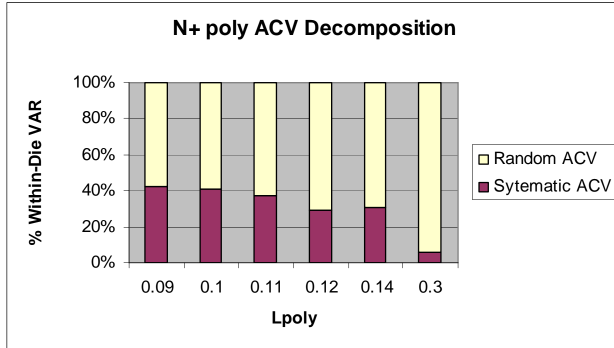


Figure 3. Random vs. systematic variations for Across Chip Variance (ACV) of Lpoly for NMOS transistor.

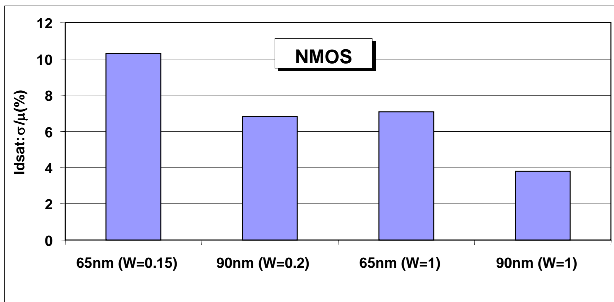


Figure 4. Normalized standard variation in drive current for wide and narrow NMOS transistors in 90 and 65nm technologies.

As we stated before, transistors behave differently based upon the neighborhood layout pattern due to printability and stress/strain effects. Figure 5 shows the Idrive vs. Ioff currents for identical transistors in three different layout environments in 65nm technology. As depicted in the figure, there is a 40% difference in the mean values of Idrive and two orders of magnitude difference in the mean values of Ioff between these transistors placed in environments I and III, respectively.

Moreover, printability and Chemical Mechanical Polishing (CMP) cause significant variations in interconnect parameters such as resistance and capacitance. Figure 6 depicts metal resistance variations as a function of local neighborhood width and spaces, and intra-layer density. This variability is even more significant if the pattern density in the underlying layers is considered.

To deal with this level of variation, circuit designers should employ very accurate statistical process characterization, performance verification such as statistical timing closure via Statistical Static Timing Analysis (SSTA) or statistical optimization on analog components (PLLs, memory periphery circuitry). Unfortunately, most current methodologies and design flows are not capable of handling this level of complexity, making the designs vulnerable to process variations and thus contributing to yield loss.

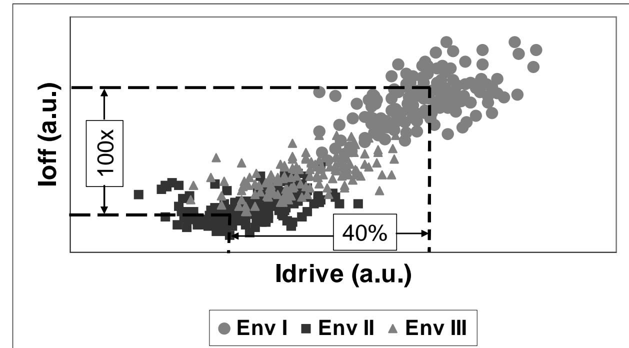


Figure 5. Idrive vs. Ioff plots for identical transistors in three different layout environments.

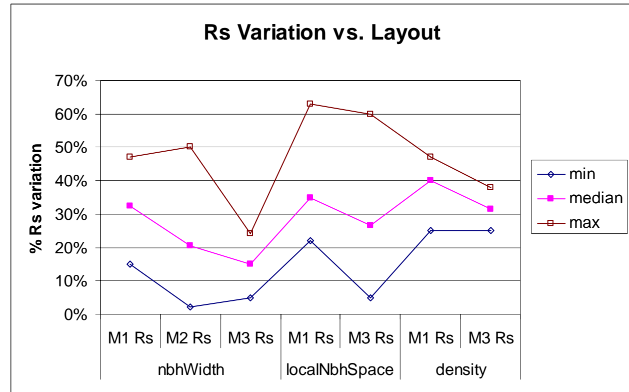


Figure 6. Metal resistance variations as a function of local neighborhood width and spaces, and intra-layer density.

YIELD VARIABILITY IN PROCESS RAMP

In the previous section, we focused on the parametric variability of transistor and interconnects. In this section, we will describe the main reasons for functional yield losses. In the process of ramping yield, the initial focus is on the intrinsic module variability due to basic process integration issues and process windows. It is extremely important not only to lower the layer defectivity and fail rates for contact/via holes but also to reduce their variability by properly centering the process. Adequate module characterization is an absolute must in order to observe defectivity levels below single parts per billion from very few wafers. To achieve these observability levels, full reticle test structures are required which have a rich set of structures that characterize all random defectivity and which have comprehensive layout patterns that represent the particular product type in terms of density and density gradient ranges, and difficult-to-print patterns. This requires a sophisticated Design Of Experiments (DOE) to optimally utilize the reticle area. To increase the learning rate, short flow test structures are required for both FEOL and

BEOL parts of the process flow. These short flows are sufficiently efficient to fully characterize the distributions of key module parameters across the entire hierarchy of variations, as shown in Figure 7 for two examples of key module failure types, namely Poly shorts and Via 1 opens.

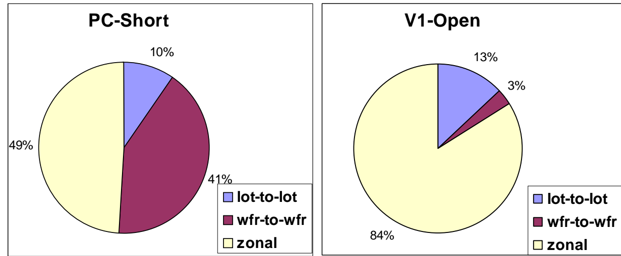


Figure 7. Spatial distribution of Poly short and Via 1 open module failures.

As can be seen from this figure, the within-wafer spatial variations (zonal variations) and wafer-wafer variations dominate, which can be explained by the challenges of maintaining process uniformity for 300 mm wafers and the large percentage of single wafer process steps. The same conclusion is true for product yield. The ability to accurately extract module defectivity/failure rates and their spatial distributions is crucial for ramping yield because the key factors driving module parameter variability must be identified and prioritized.

All the failure rates (e.g., for contact/via opens and also random defect characteristics) necessary to predict yield of the actual product can be determined since the product design features (number of non-redundant contacts, number of redundant contacts, and critical areas per layer) can be extracted. The detailed yield loss breakdown per layer and even the root cause mechanisms can now be presented in a Yield Impact (YIMP) table**.**

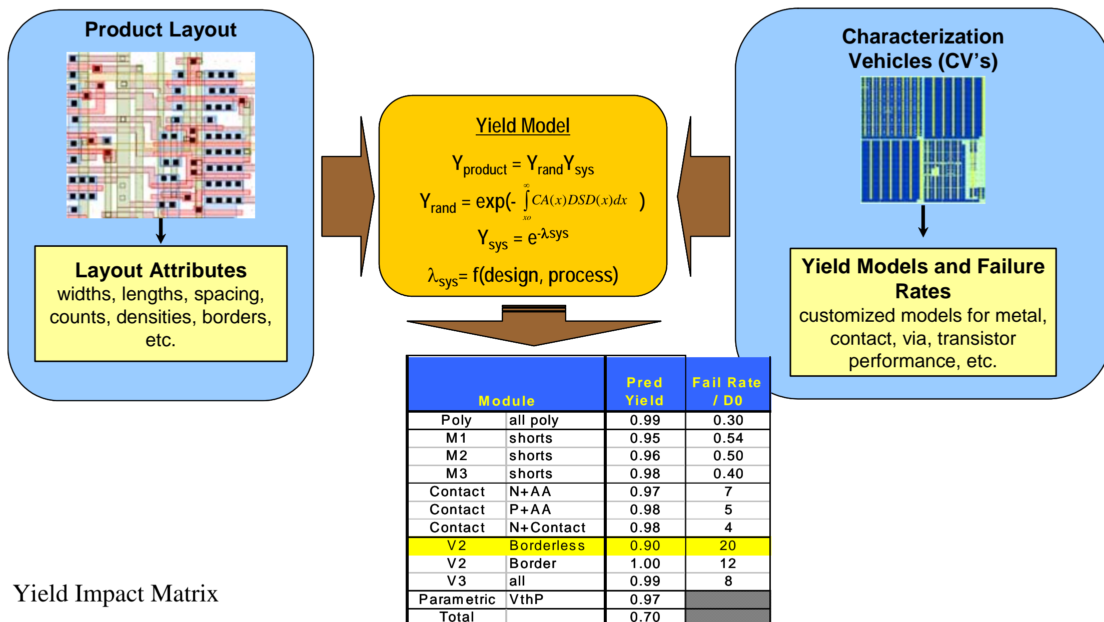
The details of the limited-yield modeling approach for the individual yield loss mechanisms are beyond the scope of this

paper but can be found in [1]. The main benefit of the YIMP methodology is that it *quantifies* the yield issues for a particular product, or set of products. This allows for better prioritization of resources, more accurate production planning and scheduling, and finally, faster time-to-volume.

The key difference between the YIMP methodology and traditional methods is that YIMP accounts for the product design [2]. Typically, fabs use the yield of test vehicles to determine which process modules are causing problems. For example, if the yield of stacked via structures is lower than the yield of contact structures, the conclusion is that there is a problem with the stacked vias; subsequently, resources are put on this issue. The YIMP methodology, however, may draw a different conclusion because it accounts for the impact on overall product yield rather than simply the yield of an individual process step. Information from the YIMP analysis can be used to create a Pareto chart of yield loss mechanisms for a particular product manufactured with a given fabrication process. Efforts can then be prioritized to focus on improving yield by improving the process module (e.g., NiSi), modifying layout design rules, or even modifying the product design.

YIELD VARIABILITY IN VOLUME PRODUCTION

While the main objective in a yield ramp is to reduce the intrinsic module variability by making sure that the lead product can be robustly manufactured within the available process windows, the ramp is typically performed using a limited set of tools. Hence, issues like tool/chamber matching, optimization of preventative maintenance (PM) or consumables (e. g., slurry and polishing pads in CMP) cannot



Module		Pred Yield	Fail Rate / D0
Poly	all poly	0.99	0.30
M1	shorts	0.95	0.54
M2	shorts	0.96	0.50
M3	shorts	0.98	0.40
Contact	N+AA	0.97	7
Contact	P+AA	0.98	5
Contact	N+Contact	0.98	4
V2	Borderless	0.90	20
V2	Border	1.00	12
V3	all	0.99	8
Parametric	VthP	0.97	
Total		0.70	

Figure 8. Yield Impact Matrix

be fully addressed for the entire equipment set over an extended period of time. Moreover, there will be excursions caused typically by equipment malfunction. The popular myth is that most of the production yield variability is caused by the excursions and this is the focus of metrology, defect inspection and SPC.

Let us now examine the validity of this myth. First, we will define the concepts of baseline wafers and excursion wafers. We will define baseline as the largest population of wafers classified by bin and spatial signatures. Note that it can include low and high yielding wafers if they have the same wafer map and same ratio of bin fallouts. Figure 9 illustrates spatial distributions in baseline wafers.

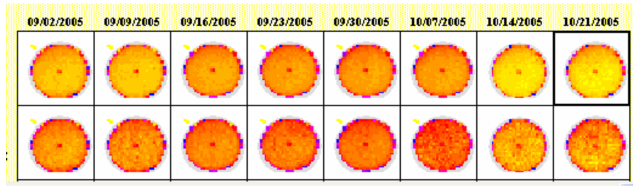


Figure 9. Baseline wafers with different wafer yield values.

Excursion wafers, on the other hand, are the ones with significantly different spatial and bin signatures (as shown in Figure 10).

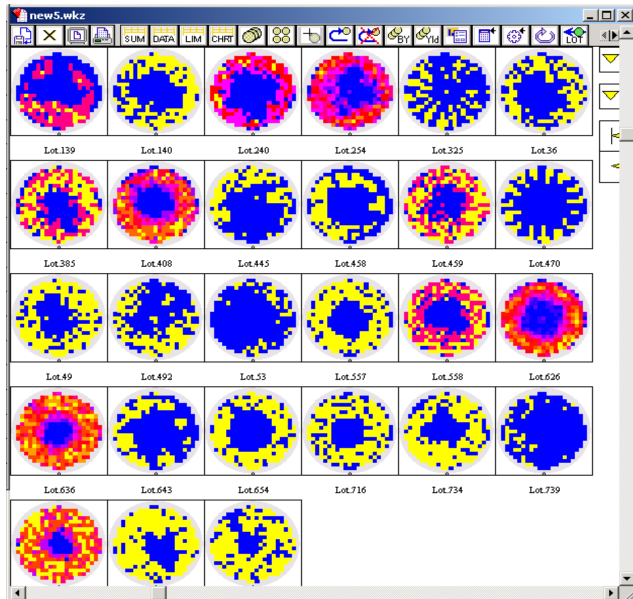


Figure 10. Spatial distributions in excursion wafers.

Based upon this classification, we can now verify the percentage of wafers in both populations. Figure 11 shows this for 6200 wafers manufactured in a state-of-the-art 300mm fabline.

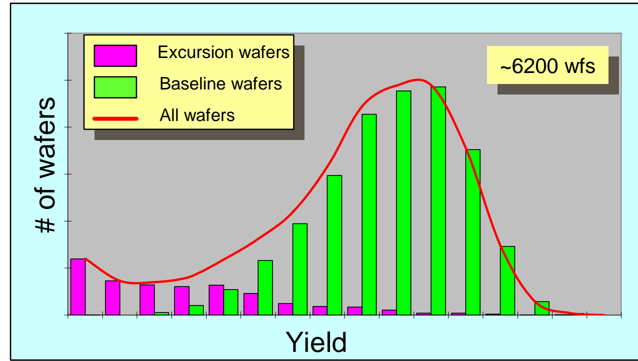


Figure 11. Baseline wafers with different wafer yield values.

In this benchmarking, only 15% of all wafers were classified as excursions, mostly due to equipment malfunctions. The remaining 85% were the baseline wafers with most of the variability distributed between within-wafer and wafer-wafer.

Finally, if we decompose the total product yield by the limited yield for baseline and excursion so Y_total =LY_baseline * LY_excursions, then, in this study, the excursion limited yield is 90% while the best fabs achieve the value of 95%. Hence, even for volume production there is a big incentive to focus on baseline variability reduction.

Again, however, if we cannot characterize this variability accurately and identify the root causes, there is no clear path to reducing the tails of the distributions. It is then interesting to examine how many wafers of short flow vehicles, called Characterization Vehicles (CVs), are required to fully characterize the baseline variability. Figure 12 demonstrates that it takes only 80-120 short flow CV wafers (which corresponds to 25-40 full flow equivalent wafers) to predict 85% of the baseline yield variability for thousands of product wafers in the 130nm process.

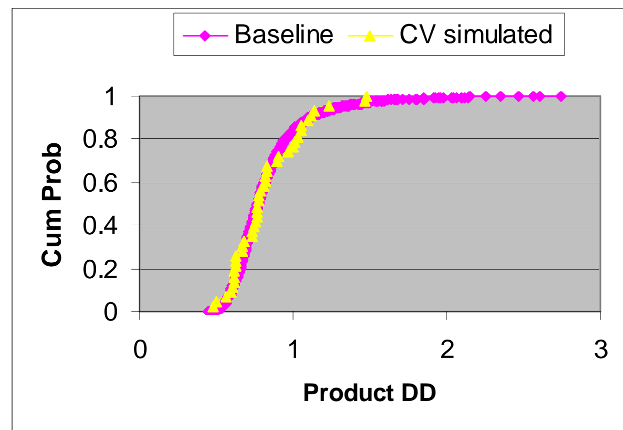


Figure 12. Prediction of baseline variability from CV wafers

Figure 13 demonstrates that it takes only 80-120 short flow CV wafers to predict 95% baseline yield variability for thousands of product wafers in the 90nm process.

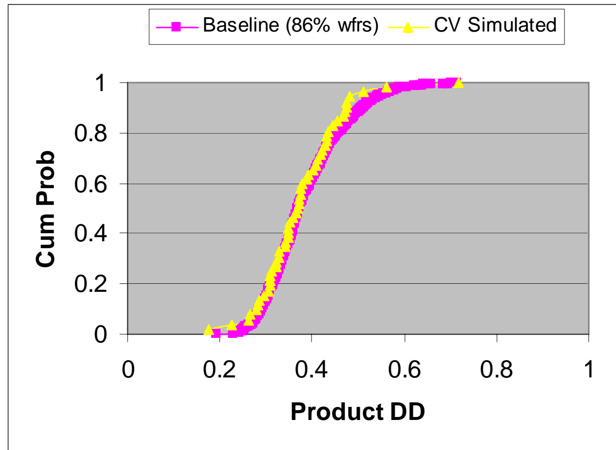


Figure 13. Extraction of baseline variability

DESIGN BASED PRODUCT YIELD VARIABILITY REDUCTION

The yield impact modeling-based approach can be integrated in the IC design flow to create manufacturable-by-construction designs. A DFM methodology has been developed [3], [4] that selects the most manufacturable design block for a target SOC based on accurate and continuously updated yield models. This methodology was applied to optimize the standard cell logic of complex SOCs, but it can be generalized to any type of structured design.

This DFM methodology has three main components. The first component is a modification of the IP library in order to include variants of the basic standard cell implementations to address different yield loss mechanisms for the same logic functionality and driving strength. These alternative implementations are called *Yield Strength Variants* because they provide different strengths of tolerance to the various yield loss mechanisms.

The second component is a data file containing the accurate characterization of the yield attributes of the library that is obtained by evaluating the yield models integrated into a suitable characterization engine. These models are calibrated to silicon by using dedicated CV test chip data. The third component is an extension of the physical synthesis tools to incorporate yield in the logic optimization cost function. Because logic block size is often dominated by routing, the synthesis tools can exploit placement and timing slacks to optimize for yield without increasing block size or delay.

The key idea in this methodology is the incorporation of the DFM guidelines in the design of the IP library ahead of knowing which mechanisms dominate yield loss in a particular process. The system relies on dynamically evolving yield characterization data to optimize logic block mapping on physical libraries and thus maximize the manufacturing yield.

It is important to note that a certain fraction of the yield strength variants can be designed to relax printability issues due to inherent variations such as defocus, exposure variations, misalignment and mask error enhancement factor (MEEF). Figure 14 shows printability the hot spots (necking and poor contact coverage) when litho process variations are considered. By characterizing and modeling the relative yield loss contribution of systematic yield loss due to poor printability and by applying the described DFM methodology, it is possible to significantly improve IC yields.

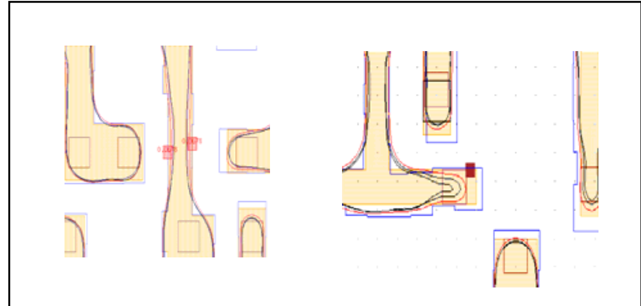


Figure 14. Potential systematic chip failures due to cell printability issues

The practical feasibility of this DFM methodology has been thoroughly verified in silicon on many complex SOC designs [3]. Figure 15 shows the actual percent of good die per wafer improvement for 5 different ICs implemented using this DFM methodology. The improvement is measured with respect to the previous revision of the same chip designed using traditional methodologies.

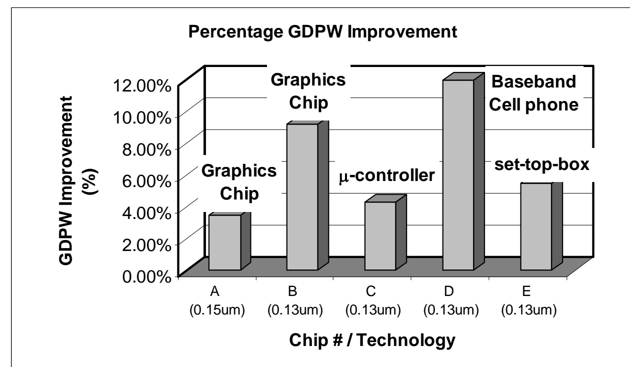


Figure 15. Measured good die per wafer improvement

Interconnect yield losses play a very significant role in the overall yield Pareto. While in older Al interconnect technologies, yield loss was dominated by inter-layer random defects that caused shorts, the state-of-the-art Cu BEOL technologies have a more complex set of yield loss mechanisms. Contact and via hole failures are very significant contributors due to printability problems related to non-robust OPC (layout pattern sensitivity, minimum metal island size), delamination or underlying topography effects caused by non-uniform layout density which results in dishing or erosion effects in CMP. These yield losses may be reduced by applying more robust OPC taking into account all significant variations in the printability process.

Another approach to minimize these yield losses is based on contact and via doubling. It must be noted, however, that failure rates for the doubled vias cannot be assumed to be always lower than for single vias (especially for small pitches) and must be carefully characterized. Figure 16 shows the comparison between the single and double via failure rates. It is possible to observe an increase in failure rates of double vias, especially in the lowest k porous ILD processes.

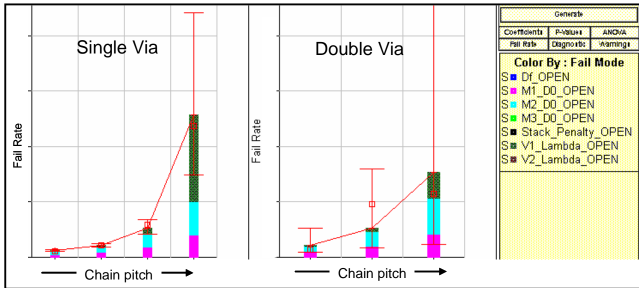


Figure 16. Single vs. double via fail rates as a function of a via pitch.

Along with logic, memories represent another significant source of yield loss for a large class of ICs, and SOCs in particular. Yield loss in memories can be due to failures occurring in the core or the periphery, with the former being usually most dominant unless for very small arrays.

Memory bit-cell still plays a critical role during technology development because it is often used to tune and optimize the process flow and OPC. The need for larger embedded memories drives the implementation of very high-density bit-cell layouts and of aggressive ad-hoc design rules along with custom, hand-optimized OPC. In fact, these dense SRAMs are the first victims of increased process variability; achieving acceptable noise margins and stability for the required values of Vmin has been extremely challenging for 65 and 45nm technologies. To preserve the 6-transistor architecture, layout of SRAM bit-cells had to be restricted to uni-directional poly and some patterning requirements had to be relaxed. Figure 17 shows Intel's bitcell in 65nm technology [5].

Moreover, shorter bitline and full metal wordline with wider spacing had to be employed to improve timing performance. Furthermore, many circuit techniques, ranging from multiple, Vcc, adaptive array biasing and better error detection/correction schemes, are implemented to provide robust SRAMs.

Finally, we should also address the issue of parametric yield losses in the analog parts of SOCs. It is not uncommon to observe failures in PLLs or even memory periphery circuitry. Again, the key is to provide a thorough statistical characterization of transistor performance variability, including both random and systematic effects including device mismatch characteristics. Then it becomes possible to perform statistical optimization of these analog functional blocks to center the designs for the actual process windows.

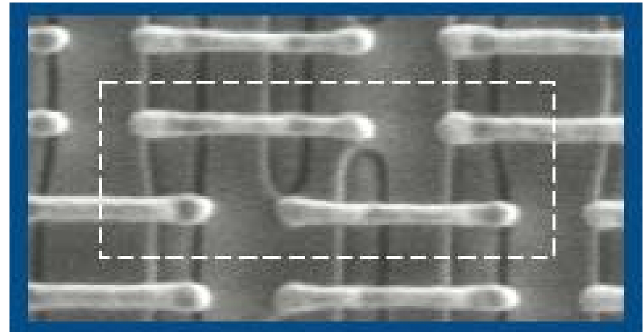


Figure 17. Intel's DFM SRAM

MANUFACTURING YIELD VARIABILITY REDUCTION

The short flow vehicles (CVs) described above provide excellent infrastructure to reduce process variability. Due to the small number of short flow wafers required for accurate characterization of process modules, efficient experiments can be run to center the process for the type of leading products being ramped up in the process. The same short flow infrastructure can also be used for equipment/chamber matching and even for PM/consumable optimization. This will allow for a significant reduction in baseline variability as illustrated in Figure 18.

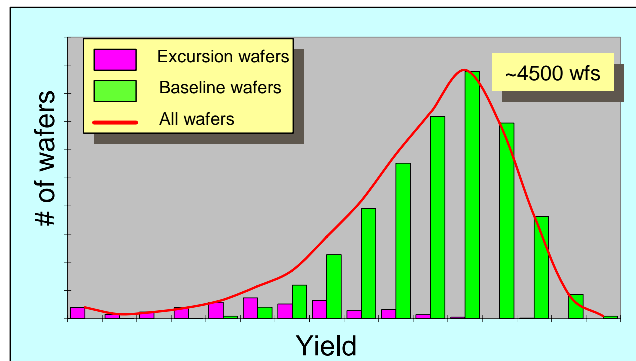


Figure 18. Baseline variability reduction after a focused 3-month program

This 3-month program focused on baseline variability reduction resulted in significant improvement in both the mean value of wafer yield and also in the reduction of normalized standard deviation from 22 to 18%; this baseline improvement accounted for 75% of the overall yield improvement. In this focused effort, the process has become more stable and even the excursion limited yield increased to 95%. After 3 quarters of this program, the normalized standard deviation was reduced by a factor of 2.

It must be clearly stated, however, that the short flows are not sufficient to improve yield stability in volume production since there are many tools of the same type (e.g., scanners, CVD, CMP) used for manufacturing product wafers; equipment malfunctions or drifts, and all of these tools must be monitored continuously for product wafers. This is accomplished by collecting data from in-situ sensor

measurements, parametric and defect in-line metrology, e-tests from PCMs placed on scribe lines, and finally, product tests. Due to the time-consuming nature of the in-line metrology and limited capacity of the capital intensive metrology equipment, only sampling of a few wafers per lot can be afforded. As a result, it is extremely difficult to catch excursion wafers this way. PCM measurements are also relatively slow and only a few structures are measured per wafer and only a few wafers per lot. This is a problem since, as we stated before, most of the variability comes from within-wafer or wafer-to-wafer sources. Although product test provides very thorough coverage, diagnosis of yield loss root causes from product test is extremely difficult.

Fortunately, it is possible to utilize the scribe area on production wafers in a much more efficient fashion by designing test structures that can be stacked on top of each other underneath the pads. Hence, the entire scribe area can be utilized for each layer (active, poly, metal, via,…) providing very good coverage for random defectivity/fail rates (large critical area) and also key systematic defects. An efficient switching matrix can be implemented using active devices so all of these structures can be tested. Moreover, this Scribe CV design allows for identification of defective layer and defect type (short, open metal or via, resistive via, etc.). These scribe structures can be tested on a massively parallel tester in such a way that 120 65x4000 micron locations on the wafer can be finished in 10 minutes, which is equivalent to testing a small number of PCM structures in 9 locations per wafer. Such a coverage allows for quick detection of excursions, spatial distribution and wafer-to-wafer variations, thus accelerating root cause identification. Figure 19 shows an example of Scribe CV results for poly stack opens and their correlation to the product yield.

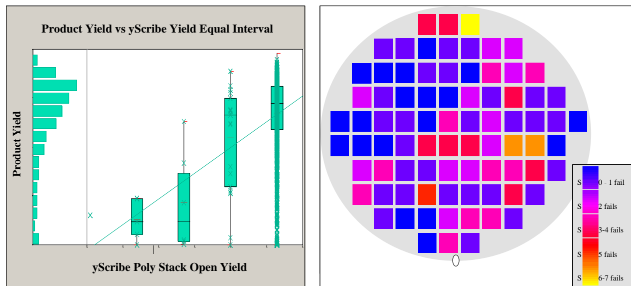


Fig 19. Example of Scribe CV results for poly stack opens and their correlation to the product yield.

This concept of fast, testable scribe structures has also been extended to the parametric characterization of variability. Fast, parallel-testable device arrays are placed in the scribe area (again, in 120 locations per wafer, for example) in the Device Scribe CV and can fully characterize the spatial distribution of device electrical parameters. In the current implementation, 320 device I-V characteristic sweeps in an array can be performed in 11 seconds.

The above mentioned techniques provide very valuable characterization of variability components and can lead to root cause determination. Although scribe test structures (like PCMs) can be tested in-line (after low level metallization steps), this is, for all practical purposes, post-mortem analysis (product test is for sure) and equipment excursions (or drifts), if undetected earlier, can result in a large number of wafers/lots with very low yield thus contributing to the overall yield variability.

Yield Relevant Statistical Process Control

In recent years, there has been tremendous progress in providing equipment with an array of in-situ sensors which perform thousands of measurements in real-time. Most of the modern fabs invested in Fault Detection and Classification (FDC) systems which analyze the multi-variate distributions of the in-situ sensor data to determine if the equipment set performs within the prescribed specifications. Until now, however, these FDC systems have been focused mostly on equipment health monitoring independent of the product yield impact requirements. If we utilize the concepts described above (i.e., short flow and scribe CV test structures, and the ability to map their results into a product yield impact matrix), it becomes possible to develop a new yield-relevant approach to SPC, and even APC. This is of crucial importance for an overall fab operation efficiency. If the process is fully tuned to the product requirements and statistical process control assures that the equipment shut downs occur only in the case of yield relevant events, the economic impact can be quite significant.

Although mapping of FDC parameters to product yield has been attempted, it is virtually mission impossible because of the mapping of many thousands of in-situ parameters into a single response. With the ability to extract module level characteristics using short flow and scribe CV test structures, it is now possible to build two-level hierarchical models with the first layer providing the mapping of the most relevant FDC variables into the module characteristics (defectivity, fail rates, layer resistivities, electrical measurements of CDs, layer thicknesses, etc.). This mapping can even include spatial within-wafer distribution parameter and, of course, wafer-to-wafer level variability. The second layer is the mapping of the module characteristics to product yield via the Yield Impact Matrix. Actually, this mapping can be performed for all limited yields for all key modules, as well as product functional blocks.

This allows for construction of very robust models which can capture the main dependencies and can be frequently updated based on new scribe test structure data from the production wafers spanning the entire set of equipment. Such models can now be inverted to determine the acceptability region in the FDC parameter space and thus yield-relevant SPC Out of Control (OOC) Limits can be derived [6]. Figure 20 depicts the predictive capability of such models for the contact to N+ active opens vs. the key FDC parameters of the RIE module. These models can be also applied to derive yield-relevant spec limits for virtual metrology, feedback (run-to-run) and feed-forward APC.

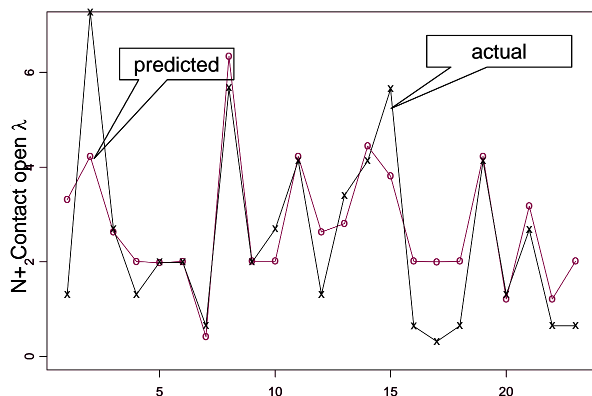


Figure 20. Model prediction vs. actual wafer data for contact to N+ active fail rate as a function of RIE FDC parameters

LOOKING INTO THE FUTURE

Conquering process variability will continue to be the most challenging task in future technology generations. Chip makers have gotten away with pushing traditional methodologies to their limits thus far. We are now seeing these limits become binding constraints and there is a need for a new paradigm in DFM and process control. The key enabler in this paradigm shift is the accurate and complete silicon characterization. This paper showed how the challenges at 65 and 45nm nodes can be overcome in an economically efficient manner by employing these comprehensive characterization-based approaches.

Looking further down the road, printability will become the key challenge. Minimizing printability variations is especially daunting for at least the two upcoming technology generations employing immersion lithography until/if EUV lithography becomes a reality. There two ways in which this challenge can be tackled. One can opt for either expensive manufacturing solutions such as polarized illumination, double exposure techniques for contact printing (pack-and-cover or pitch splitting techniques [7]) or separate exposures for SRAM array and periphery. The alternative approach is to impose extreme layout regularity [8] where the critical layers (poly, M1, M2) are laid out as unidirectional gratings and contacts are on grid. Such an approach allows for exploration of push rules and therefore minimizes or may eliminate the area/performance penalty which is the problem in the current Restricted Design Rule approaches. It also simplifies tremendously the OPC/RET procedures and even allows for optimization of illumination conditions for specific gratings.

Finally, with the advances in process observability via in-situ equipment data collection, there is also an opportunity to advance process control methodology. It should be possible to detect excursions on the spot and quickly identify and fix the root causes. It should also be feasible to act on the tails of parameter distributions by effective feed-forward and possibly even real-time control. If the control limits are derived in a yield-relevant fashion, it would then be feasible to control yield and performance variability for each high volume product which would lead to a significant increase in fab profitability.

ACKNOWLEDGEMENTS

The author wants to acknowledge numerous colleagues at PDF Solutions, Inc. who provided the leading edge results for this paper. In particular, extremely helpful discussions and assistance in assembling the data of Sherry Lee, Rakesh Vallishayee, Sharad Saxena and Yun Lin are really appreciated. The guidance of Dave Joseph and feedback from Marcin Strojwas were extremely valuable to the author. The author is also grateful to Dana Russell for assistance in paper preparation.

REFERENCES

[1] D. Ciplickas, S. Lee and A.J. Strojwas, "A New Paradigm for Evaluating IC Yield Loss", Solid State Technology, October 2001

[2] D. Ciplickas, A. Joshi, S. Lee and A. J. Strojwas, "Designing for High Product Yield", Semiconductor International, October 2002

[3] J. Kibarian, C. Guardiani and A. J. Strojwas, "Proactive DFM for Nanometer SOC Designs", Proc. 2005 ISSCC

[4] C. Guardiani, N. Dragone and A. J. Strojwas "Design For Manufacturability", Chapter 19, Electronic Design Automation for Integrated Circuits Handbook, CRC Press, 2006.

[5] Y. Borodovsky, "Marching to the beat of Moore's Law", SPIE-Microlithography, 2006

[6] A. J. Strojwas, "Yield Relevant Statistical Process Control of ULSI Manufacturing", European AEC/APC Conference, 2006

[7] B. J. Lin, "Lithography for manufacturing sub-65nm nodes and below", IEDM, 2005

[8] V. Kheterpal, V. Rovner, T.G. Hersan, D. Motiani, Y Takegawa, L. Pileggi and A. J. Strojwas, "Design Methodology for IC Manufacturability Based on Regular Logic-Bricks", Design Automation Conference, 2005

AUTHOR BIOGRAPHY

ANDRZEJ J. STROJWAS is Joseph F. and Nancy Keithley Professor of Electrical and Computer Engineering at Carnegie Mellon University. Since 1997 he has served as Chief Technologist at PDF Solutions, Inc. He has held positions at Harris Semiconductor Co., AT&T Bell Laboratories, Texas Instruments, NEC, HITACHI, SEMATECH, KLA-Tencor and PDF Solutions, Inc. His research interests include Design For Manufacturability, modeling of IC equipment, fabrication processes and devices, and control and diagnosis of the IC fabrication processes. In 1990 he was elected IEEE Fellow.