Realizing a 45nm System on Chip in the Age of Variability

Laurent Le Cam[1], Andy Appleby[2],
Philippe Hurat[3], Benoit Carpentier[4], Kuang-Han Chen[3], Nishath Verghese[3],
[1]NXP Semiconductors, Gerstweg 2, 6534 AE Nijmegen, The Netherlands
[2]NXP Semiconductors, Millbrook Technology Campus, Southampton SO15 0DJ, UK
[3]Cadence Design Systems, Inc., 2655 Seely Ave. San Jose, CA 95134, USA
[4]Cadence Design Systems, 1080 Route des Dolines, 06560 Valbonne, France

ABSTRACT

In this paper, we present the challenges of the realization of a large 45nm modern Media Processing SoC with multiple design teams distributed across many countries and time zones. We also describe the complex design methodology deployed to ensure the design is "closable" in the timing and manufacturability domain.

Silicon variability impacts both the physical integrity and the parametric performance of the design. Lithography and CMP can cause enough context-dependent systematic variations, requiring exhaustive lithography and CMP physical verification and optimization of the layout.

We present the physical and electrical DFM methodology at NXP. We will show how NXP has developed a manufacturing-aware design flow based on early prevention, detection and fixing using a hierarchical approach for model-based lithography checks and model-based CMP checks, from IP level to full-chip. We also present results of variability-aware timing sign-off.

1. INTRODUCTION

NXPs latest 45nm design contains >150 individual soft and hard IPs many of which were being developed concurrently with the host SoC. All these design activities must converge predictably at tape-out time. With Time to Market such a crucial factor in today's fast moving semiconductor business there is often no time for a re-spin. First Time Right is the name of the game.

During the definition phase, the SoC Physical Design Team must underpin commitments on cost, performance, area, package choice, power and the tape-out schedule to the customer. These commitments must be made long before the IP content has matured into stable design components. How do we gain confidence that our design will meet timing within such a "framework of uncertainty"?

The "traditional" approach of using complex STA sign-off-tools that analyse mature design timing data to ensure the performance targets are met, pre-supposes that with a low level of optimisation in the Physical Design space the performance requirements will be met during the "Closure Phase". Real problems occur if they are not! Sign-off tools are run so late in the activity that if any "out of band" problems are found, fixing them will result in schedule slip and potentially, missed market opportunity!

In 45nm the variability is so large that it must be taken into account at all stages of the design.

Silicon variability can impact both the physical integrity and the parametric performance of the design. On top of random variations, lithography and Chemical Mechanical Polishing (CMP) can cause enough context-dependent systematic variations, requiring exhaustive lithography and CMP physical verification and optimization of the layout.

Variability due to random sources and systematic effects such as Lithography, CMP and stress create a real challenge at 45nm. We can no longer presume that timing will be met across all modes and for all process, voltage and temperature combinations when we emerge at the back end of our flow. We need to manage the convergence process very carefully to ensure the design is "closable" in the timing ECO phase. We also need to talk into account the context-dependent systematic proximity effects of lithography and stress.

Design for Manufacturability through Design-Process Integration IV, edited by Michael L. Rieger, Joerg Thiele,
Proc. of SPIE Vol. 7641, 764106 · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.848023

In this paper, we present the physical and electrical DFM methodology at NXP. We will show how NXP has developed a manufacturing-aware design flow based on early prevention, detection and fixing. A hierarchical approach, from IP level to full-chip, was used to reduce the runtime impact on design schedule of the model-based lithography checks without compromising accuracy. We will also describe how CMP checks are applied at full-chip to take into account the long-range effects of CMP, and how model-based fixing can make design CMP-clean. We also present results of variability-aware timing sign-off.

Combining the complexity of large scale SoC, the uncertainty of concurrent design and the variability of 45nm is a recipe for sleepless nights for the Physical Designer. We try to show how this multi-dimensional problem can be managed in the real world of product creation.

2. BACKGROUND AND MOTIVATION

To speed-up Time-To-Market and optimize the die price for the sub-65nm products, NXP is addressing both systematic and random defects. Focusing on systematic defects will help for Time to Market and addressing random defect will help to get the most dies from a wafer. Systematic defects are commonly due to lithography process window, Chemical Mechanical Polishing (CMP) and stress issues and can be captured by dedicated lithography and CMP analysis tools. A local optimization of the layout for lithography variations and a smart tiling will correct both issues. Random defects generated during the Back-End processing in vias and metal lines can be detected by Critical Area Analysis tools. Via doubling, wire spreading and wire widening can limit the sensibility of the product to these random defects. The products can be analyzed for these possible process issues using the DFM kits provided by the foundries. These kits contain the raw information from the process and need to be converted with the corresponding EDA analysis tool. This is a key element for a good integration into the design flow so that the design teams have access to these DFM analysis and optimizations. These DFM tools must be qualified by the foundries, they must run fast and as they are meant for design teams spread all over the world, they must be fully integrated in the design flow and very robust to limit any delay in the design process.

3. VIA DOUBLING

3.1 Via doubling motivation

Via processing is known to be a major challenge in advanced node processes. It is sensitive to lithography variations and to random defects present during the processing. One issue during the via process step is commonly named “open via” and corresponds to a broken connection between the two metal layers (no current can flow). Another type of failing via is commonly called “resistive via” and in this case some current can still flow between the two metal layers, but the connection might deteriorate with time and end up in an “open via”. These “resistive via” are very difficult to detect. A via has a certain failure rate, that can be measured by the foundry using dedicated test structures. With modern SoC chips containing more that 80 Million metal transitions, using via doubling methods is essential to get a functional chip (yield). Good test coverage of the product will help in finding dies with open or resistive vias, but test coverage is usually not 100%, which means that chips with open or resistive vias will be sent to the customer. This is very critical for products such as automotive applications, where zero defects are required. Customer satisfaction getting reliable products is therefore another reason to focus on via doubling during the design phase.

3.2 Via doubling flow

The flow described below is focusing on digital blocks. All the layout blocks are assembled together via the Place&Route (P&R) tool Encounter. The P&R tools is already concurrently implementing during routing double via as much as possible but is limited by its intrinsic “on grid” placement. Depending on the layout congestion, a typical via doubling percentage achieved by Encounter is between 75% and 90%. In our flow, we are using Cadence Chip Optimizer (CCO) to increase further the double via percentage. CCO has the advantage of running “off grid”, so has much more flexibility for inserting doubled vias.

The flow is shown in Figure 1. Once the layout is placed and routed in Encounter, the database is imported into CCO, to perform the grid-less via doubling optimization. The next paragraph will detail the different options of this step. When the via doubling optimization is done, the database is converted back to Encounter to run the timing and DRC/LVS sign off.

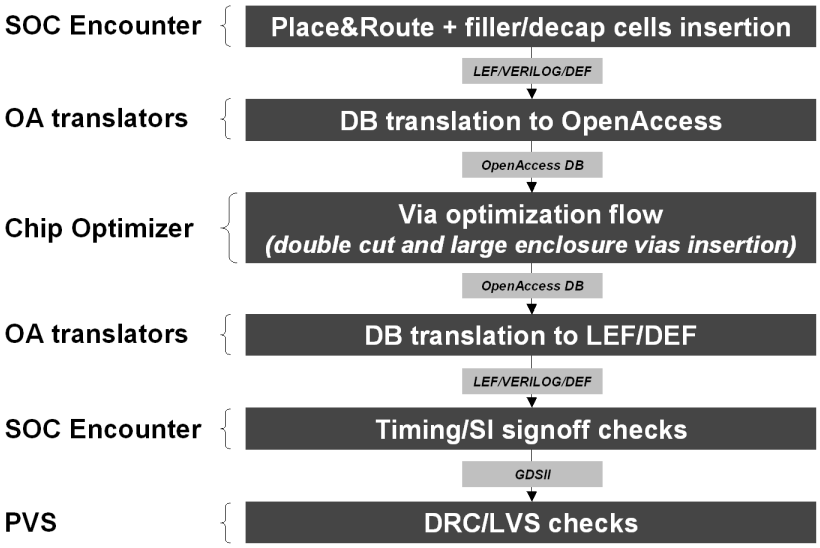

Figure 1: The via optimization flow using Cadence Chip Optimizer

Different options can be used to insert the doubled via, as illustrated in Figure 2. “on wire” and “off wire” have the advantage of having the less impact on the neighborhood lines, while “on wire push” and “off wire push” give more options if the previously mentioned techniques are not applicable. CCO has a grid-less and space-based architecture for optimal optimization of the results and is at the same time timing aware for minimal impact on electrical characteristics. Two modes are available: a “timing preserved” mode where CCO will check that inserting an extra via will keep the timing change inside the tolerance and a “non timing preserved” mode where CCO will double has much as possible. The best approach for via doubling is to use first the “non timing preserved” mode as in this mode one can achieve the most efficient via doubling and check for any timing impact. If the impact on timing is too large, CCO is then used in “timing preserved” mode where the number of via doubling is a bit less but the timing is guaranteed. In both cases, the clock tree nets are excluded from the via doubling flow to limit any possible large variation on the final timing.

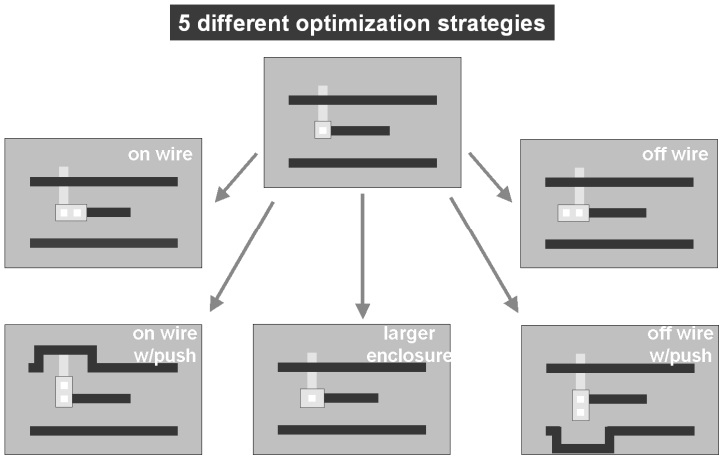

Figure 2: The via optimization strategies using Cadence Chip Optimizer

3.3 Via doubling results

The Table 1 illustrates the efficiency of via doubling, with an initial via doubling ranging from 75.5% to 89.1%. CCO can increase via doubling above 90% in almost any cases with low impact on timing.

Name	Initial double cut %	Final double cut %	Initial Nb single cut Vias	Final Nb single cut Vias	Timing Impact	New DRC
U5	75.5	87.9	5.039.670	2.495.545	-10ps	0
U3	81.6	90.9	3.609.820	1.790.046	-20ps	0
U9	83.9	90.1	154.159	95.002	-17ps	0
U6	89.1	93.1	2.234.371	1.463.639	-18ps	0

Table 1: illustration of the improvement in via doubling with low impact on timing

4. LITHOGRAPHY CHECKS

4.1 Lithography check motivation

As technology migrates from 90-nm to 45-nm node, fast yield ramp-up is increasingly difficult to achieve due to the sub-wavelength effects of lithography. While minimum feature size in IC design has decreased with each process node, the wavelength of steppers used in lithography has remained constant at 193 nm. This so-called "lithography gap" makes it difficult to produce an accurate image of the reticle on the wafer, resulting in qualitative defects in the layout that lead to catastrophic or parametric yield loss.

Litho Physical Analyzer (LPA) is a full-chip model-based design manufacturability checker that designers can use to identify hotspots and predict contours across process windows. LPA accurately predicts manufacturing variations associated with lithography and etch and still meets the runtime requirements of sub-90nm IC design implementation.

Lithography analysis is part of mandatory foundry requirements. Level-1 lithography hotspots like SPACING_LEVEL_1, WIDTH_LEVEL_1, and LINEEND_LEVEL_1 are fatal hotspots; they do not have a large enough process window in production and must be fixed prior to tape-out. This section covers the learning of DFM signoff criteria for Lithography analysis and fixing flow implemented for 45nm projects.

4.2 Lithography-aware flow

As for any DFM implementation in the NXP product creation, the goal of the lithography analysis is to correct any hotspot as early as possible in the design flow. Following this concept, NXP has developed a hierarchical design flow where lithography analysis is done at multiple stages (Figure 3). In the first place, foundation IP's like Standard Cells, Memories, Analog and IO blocks are analyzed for possible lithography issue, from the Active area up to the upper metal layers (Mx) if applicable. Then these foundation IP's are used to generate IP blocks, more complex IP's with more functionality. At that stage, design teams check their IP blocks for possible lithography hotspot (Active up to Mx), before delivery to the SoC team for final assembly. Once the lithography checks have been performed and that the IP is "litho clean", the IP block is tagged with a specific cover layer that is used by LPA to skip already "litho clean" areas.

The SoC team is responsible of generating the final product, collecting and assembling all the IP blocks together. As the IP blocks have been checked for lithography issue and marked as "litho clean", the SOC team only checks the inter block Mx routing (Figure 4). This hierarchical approach is extremely useful to reduce the top level runtime where it has the most impact on product schedule. Leveraging our hierarchical approach and the scalability of LPA, the top level lithography checks ran in about 2 hours for our latest large 45nm design. It is of course still possible to run a full SoC lithography analysis, which is usually done by the tape-out office.

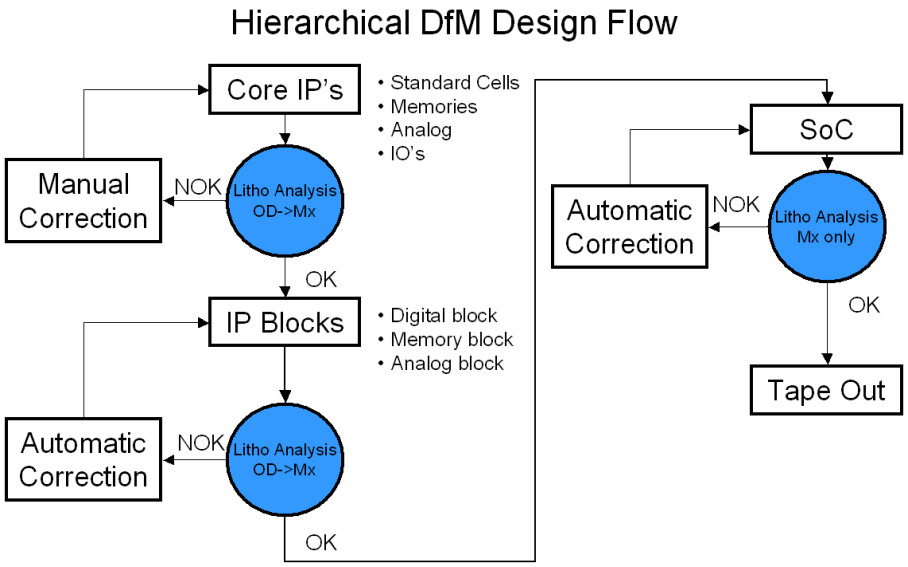

Figure 3: The hierarchical DFM design flow

Applying a systematic check at tape-out is the minimum requirement to this prevents taping out a product with potential yield loss and/or variability issues. This hierarchical DFM flow prevents a lithography hotspot to propagate till the tape-out of the product. There are several advantages to use this hierarchical design flow. First of all, when a hotspot is discovered, the design team that ran the lithography analysis is also the owner of the layout. So they can correct it before it is used in any other block. Also, it speeds up the product creation cycle time. If this hierarchical flow is not used and a lithography hotspot located deep in the hierarchy is found during the tape-out of the product, it might be too difficult to correct the IP block and re-inject it in the SoC or it simply will cost too much delay and make the product loose its market window.

Figure 4: Hierarchical approach allows IP blocks already analyzed by design teams to be skipped at SoC level using a "NO LPC" marker

To facilitate the lithography analysis inside this hierarchical design flow, Cadence lithography prevention, fast screening and correction flows are available. For the prevention action, the lithography prevention switch is activated in the P&R tool. It prevents the P&R tool to generate layout configurations known to be difficult to print in advanced nodes. To be more efficient and convergent, the latest version of the lithography tool is directly integrated in the P&R tool. This allows a very fast incremental check after fixing. It also enable designers to interactively run the lithography analysis

directly from Encounter and CCO and see back the results (hotspots and guidelines) in their design tool. For lithography screening, an accelerated version of the lithography sign-off has been integrated into the P&R tool. The P&R user has direct access to the accelerated lithography check for fast screening, as shown in Figure 5. The fast lithography screening does not replace the sign-off lithography checks to be performed on all shapes of every layer, but it allows early detection and fixing of lithography hotspots on interconnect layers. On our design, the accelerated lithography screening was 110 times faster than the full lithography signoff. For the corrective aspect, auto correction loops are possible in the P&R tool or in the CCO tool. The coordinates of the hotspots and guidelines on how to correct the hotspots are outputted with the lithography run. The P&R tool can use the coordinates to re-route around the hotspot and generate a new routing that is lithography clean. The CCO tool can use the coordinates and the guidelines to solve the hotspots with a local approach. Two examples of auto-correction on Metal2 hotspots are illustrated in Figure 6.

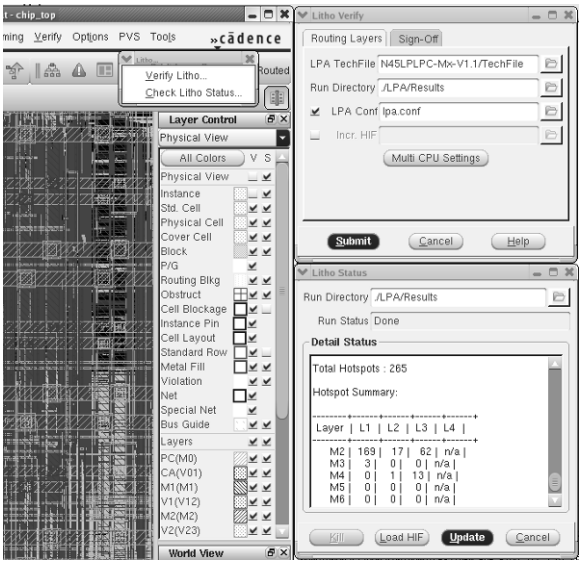

Figure 5: Integrated accelerated lithography check in P&R

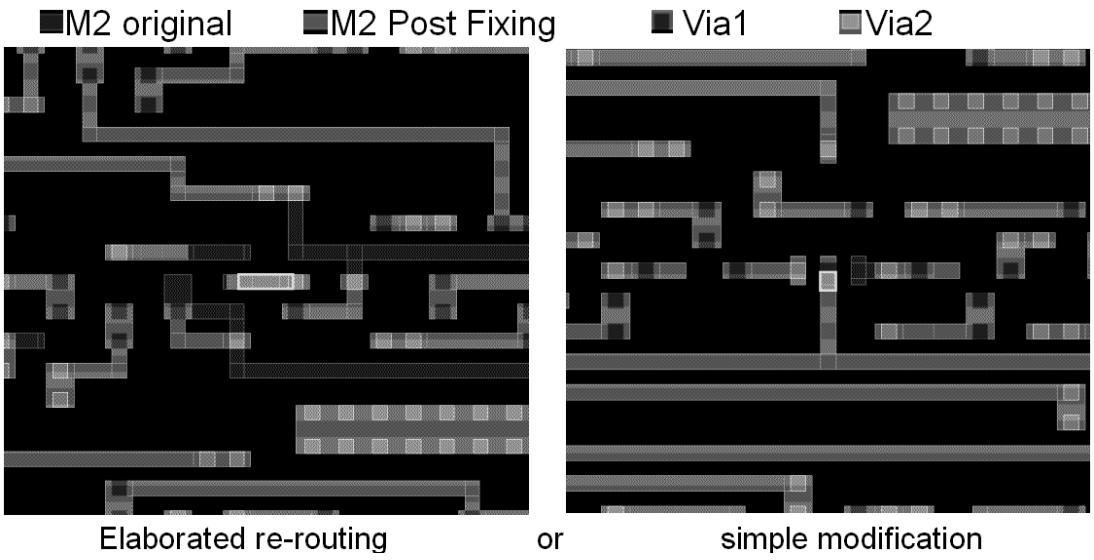

Figure 6: Examples of lithography hotspot auto-correction with a P&R tool

5. CMP CHECKS

5.1 CMP Check Motivation

CMP is known to suffer from pattern dependencies such as dishing, and erosion, which can lead to potential yield and performance problems, including copper pooling and excessive copper loss. To minimize the pattern-dependency of the CMP process, some rule-based dummy-fill insertion, also called tiling, must be done on the design database prior to tape-out. However, in most advanced CMP processes this is not sufficient anymore. In most of CMP processes, wide wire tends to experience more dishing than narrow wire, and with certain layout configuration, like several narrow wires overlay on top of two large plates; the excessive recess from the two large plates below can cause the narrow wires to be shorted; this is called cooper poolig. CMP can also introduce additional challenges to subsequent manufacturing steps, like lithography. As an example, areas with high or low surface height can cause lithography defocusing issue in the above layers; hence, severe topography variations can cause printing problems. CMP induced thickness and topography variations become an increasing concern for today's designs running through advanced process nodes, 45nm and below; leading foundries have spent a great deal of effort to develop CMP physics-based models to complement the insufficient rule-based methods. Some foundries are mandated their customers to run their designs through CMP model based check, and have qualified the Cadence CMP Predictor (CCP) to enable such model-based checks during designs.

5.2 CMP-Aware Design Flow

Once the rule-based dummy fill insertion has been performed, designers use CCP in the design flow at sign-off for a model-based screening of the designs for potential CMP hotspots. Once hotspots are detected, designers can deploy a host of tools to fix these potential problem areas, before the mask for the design is made. NXP get access to DFM kits directly from foundries. DFM kits contain several high accuracy physics-based models, such as lithography and CMP models. The design database must be streamed out to a gds file format in order to be used by CCP, and a technology map file is then created by user to map the gds layer number and data type info to physical manufacturing process steps. CMP models together with design database and technology map file are used for CMP simulation in CCP. CMP simulation flow consists of two steps, geometry extraction and prediction. Geometry extraction step extracts geometry info, such as density, from a layout using a default window size; and only after the geometry extraction is completed for the entire layout, user can proceed with the prediction step. In the CMP simulation flow, CCP enables users to run CMP simulation on a full chip level and it takes into account multi-level and long-range effects.

5.3 CMP Fixing Flow

Any CMP hotspot predicted by CCP can be seen on a heatmap. With this approach design teams can quickly pinpoint the location of the hotspots and investigate the possible root cause. The Figure 7 illustrates physical level 1 hotspots highlighting a possible Depth of Focus issue in the process. X and Y axis are the chip coordinates while the vertical axis is the surface height. Too much surface height differences between the higher (dark areas at the periphery) and lower (dark areas in the center) surface height within a given band width of current metal level can cause undesirable depth of focus issues for the next metal level.

With the help of the density heatmaps, one can understand the root cause of the CMP issue. These are mainly due to wrong tiling implementation or too much difference between high density areas like memory blocks and lower density areas like logic blocks. The goal is to re-balance the densities to reach homogenous values all over the chip. This can be done by modifying the tiling in the needed regions, either manually with the tiling scripts or with CMP aware DFM tools like CCO. In the later case, the layout is imported into CCO, which calls the CMP model-based simulation tool. The CMP tool extracts the density of the layout, runs a prediction analysis, and generates hints that help CCO to adjust the tiling to correct the CMP hotspots.

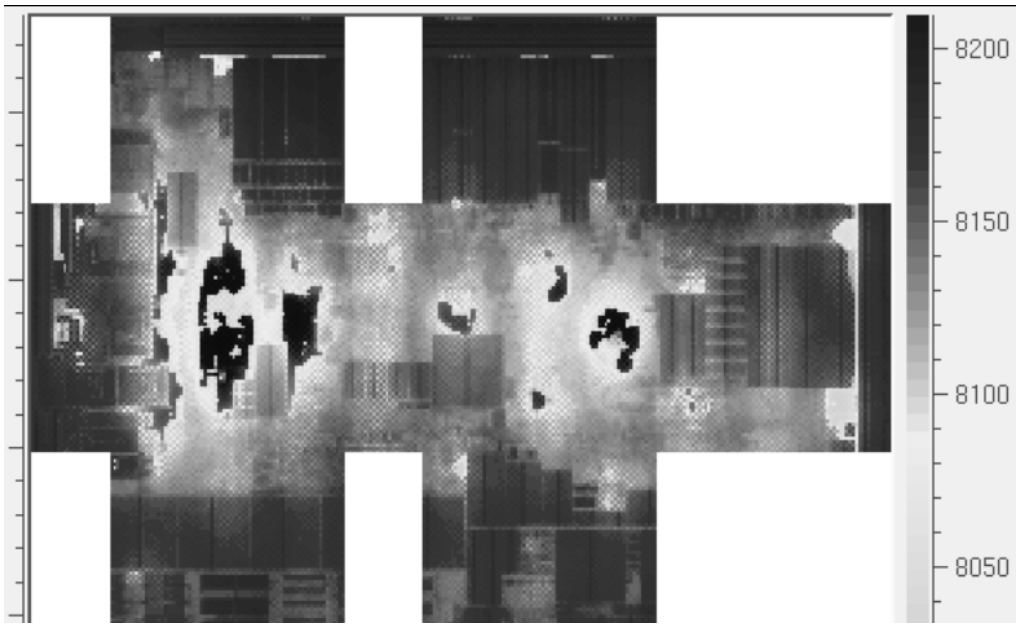

Figure 7: Snapshot of CPM Depth of Focus hotspots (x,y: chip coordinates; heat map: surface height)

6. VARIABILITY ANALYSIS

6.1 Variability analysis motivation

In advanced nodes, many layout dependent effects (LDE) such as well-proximity effects (WPE), lithography (flaring, rounding, defocus, etc) and stress effects have a significant impact on the transistor performance because these effects change the transistor parameters such as length, width, mobility of the carriers, etc. To take in consideration the LDE in the timing views of the standard cells, the cell delay characterization is not done using an isolated cell; the cell under characterization is surrounded with other cells, called characterization context. So the traditional timing analysis is done with the timing views extracted from the characterization context.

However the cells in the real design have different context that the context used during characterization, and therefore the traditional timing analysis and sign-off is done without taking into account the context dependent timing variations.

In order to quantify the context dependent timing variations, we performed two different variability analyses using the Cadence Litho Electrical Analyzer (LEA). LEA can extract the timing variations due to context and report the delay variations on a cell instances or full path.

6.2 Variability Analysis Flows and Results

We used a two fold approach. First we quantified the variability at the instance level. This step can be done after placement and enables to quantify the variability due to context. It's a light-weight approach since only the library data and DEF are needed. If the context-induced variation is large enough, then a path-level approach is required. The path-level looks at the top critical paths and check if the timing variations due to context change enough the critical path delays and might cause a set-up or hold violation. This path-level variability analysis is typically used after routing when the block or design is close to sign-off. It requires the following additional data: interconnect parasitics (SPEF), and critical path information from the static timing analysis tool (STA).

6.2.1 Instance level variability analysis

Unlike approaches described [1], [2] and [3], where random contexts where created, we used LEA to read the P&R database (DEF) and to extract the real placement of some specific cells. We picked some of the most frequently-used cells to estimate the potential variability impact of context on these cells. Also we focused on stress induced variations [6] since this is a dominant factor in 45nm.

The results shown in Figure 8, represent the context variations due to stress of an AOI cell instantiated 403 times in the analyzed block. The similar analysis was done across multiple cells and we analyze both delay and slew variations. The timing report indicates that up to 11 ps delay variations and 32ps slew variations were induced across the 403 contexts. In the histogram, the darker box in the middle of the histogram represents the characterized delay of 255ps, and characterized slew of 468ps using the ideal context, while the light gray bars show the delays and slews for the real contexts extracted from the design. We can observe that some contexts make the cell faster with a minimum delay of 248ps and a minimum slew of 449ps, while some contexts make it slower with a maximum delay of 259ps and a maximum slew of 481ps.

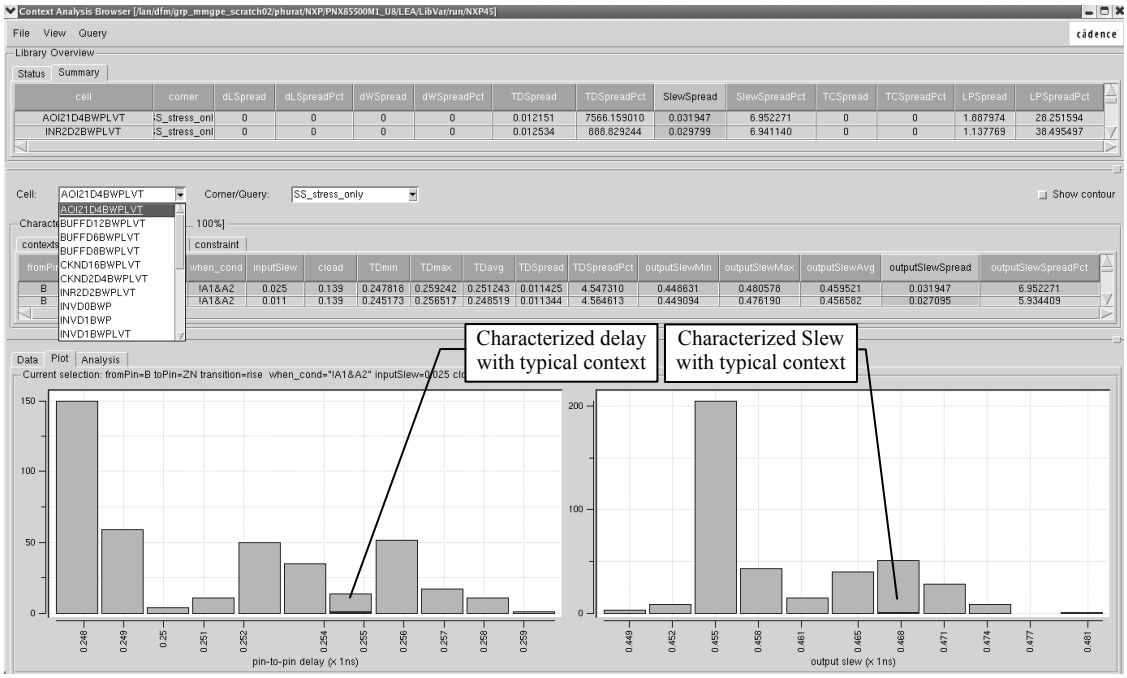

Figure 8: Cell Variability Analysis Results of LEA

6.2.2 Path variability analysis

Based on this amount of variations reported by the cell-level analysis, we decided to check how context-induced LDE impact the delay of the most critical paths. We used an approach similar to the one described in [4] where path timing analysis was done using lithography-based context variations, but instead of extracting the lithography-induced variations we looked at the stress-induced variations, as described in [5]. We extracted the top critical paths for set-up and hold and then used LEA to carve out the cells of interest with their contexts, as shown in Figure 9. Then stress effects due to context are extracted and the delay variations are calculated by LEA. LEA reports the delay variations and then creates an incremental SDF for back-annotation in the timing flow. We analyzed the top 1000 critical paths of our design and the context-induced LDE impacted the path delay by up to 29ps but no timing violation was reported.

In case of timing violations found during in context-induced stress variability analysis, then LEA creates a incremental SDF that can be loaded in the implementation tool for a timing fix.

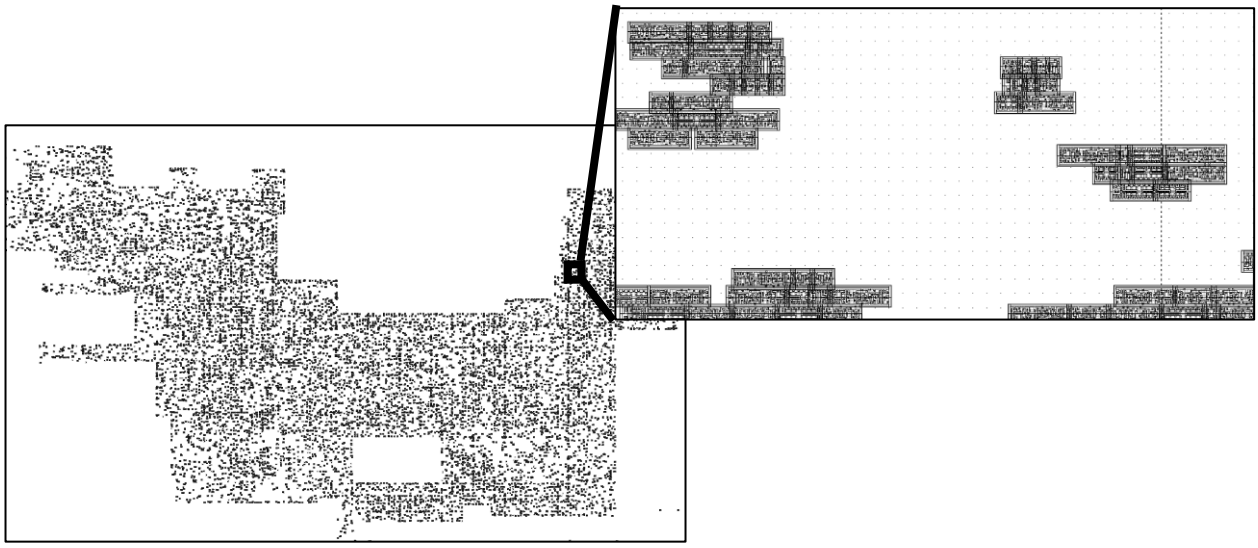

Figure 9: Extracted contexts for variability analysis of top 1000 critical paths

7. SUMMARY

In this paper we have described the DFM methodology and infra-structure developed and deployed to serve our sub-65nm products designed world-wide. We addressed both random and systematic defects. For random defects we introduced layout optimization such as via doubling. We addressed lithography and CMP issues with model-based analysis and automatic fixing integrated into a hierarchical design flow. We also quantified the impact of systematic variability due to stress proximity effects at the cell and path levels. This comprehensive deployment of DFM techniques has been successfully integrated in the implementation and verification flow of our advanced products, which are now in full production without yield issue.

REFERENCES

[1] Rajagopal A., Rajaram A., Damodaran R., Cano F., Swaminathan S., Bittlestone C., Terry M., Mason M., Ran Y., Chen H., Ritchie R., Kasthuri B., Condella J., Hurat P., Verghese N., "Context Analysis and Validation of Lithography Induced Systematic Variations in 65nm Designs," Proc. SPIE vol. 6925, (2008).

[2] Sadra K., Terry M., Rajagopal A., Soper R., Kolarika D., Aton T., Hornung B., Khamankar R., Hurat P., Kasthuri B., Ran Y., Verghese N., "Variations in Timing and Leakage Power of 45nm Library Cells due to Lithography and Stress Effects," Proc. SPIE vol. 7275, (2009).

[3] Bingert R., Aurand A., Marin J.-C., Balossier E., Devoivre T., Trouiller Y., Vautrin F., Verghese N., Rouse R., Cote M., Hurat P., "Implementation of Silicon-Validated Variability Analysis and Optimization for Standard Cell Libraries," Proc. SPIE vol. 6925, (2008).

[4] Yanagihara T., Hamamoto T., Sato K., Okamura A., Matsunaga T., Kobayashi N., Maekawa T., Verghese N., Condella J., Hurat P., "Microprocessor Chip Timing Analysis Using Extraction of Simulated Silicon-Calibrated Contours", Proc. SPIE vol. 6925, (2008)

[5] TSMC, [TSMC Reference Flow 10.0], (2009)

[6] Faricelli J., "Layout-Dependent Proximity Effects in Deep Nanoscale CMOS," IEEE Solid-State Circuit Society, Seminar in Denver, (2009)