Architecting ASIC Libraries and Flows in Nanometer Era

Clive Bittlestone, Anthony Hill, Vipul Singhal, Arvind N.V.

Texas Instruments Inc. Dallas Texas 75243

ABSTRACT

This paper is in response to the question 'ASIC Design the nm era – dead or alive' from an ASIC library architecture and library flow point of view. The authors believe it is certainly significantly harder to design in the nm era but ASIC design is not dead. ASIC Design is much more challenging in the nanometer era. This paper will present some of the main effects that have become significant in terms of library architecture and library creation flow. Some full chip level effects will be discussed. Example solutions to some of these dramatic trends will also be presented. This is presented in a 'stories from the trenches' format – from the team that architects and delivers TI ASIC libraries. The majority of the data presented comes from development of TI ASIC 130, 90 and 65nm libraries.

Categories and Subject Descriptors

B.7.1 [**Types and Design Styles**] advanced technology, Standard Cells.

General Terms: Design, Performance.

Keywords: Standard Cell, nanometer design, libraries.

1. INTRODUCTION

As the industry moves down the ITRS roadmap, an increasing number of effects that were previously insignificant are becoming first order and can no longer be ignored in standard cell library design. Therefore it is becoming much more difficult to design in the nm Era whilst maintaining expected performance, power (static and active), area, design cycle time, manufacturability and cost scaling. These effects are attacked at all levels: Process, modeling, extraction, library architecture, circuit design, synthesis, place and route, clock distribution architecture, verification, probe and final test. Architecting a cell library with these effects in mind can help reduce the impact of these effects for chip designers. Many of these effects have been known for some time and have increasingly impacted chip design flows since 180nm. Serving a wide market place with only one or two libraries is becoming difficult due to increasingly divergent multiple market needs. The compromises lead to possible solution points that are suboptimal compromises. Different library architectures or some form of scalable/unified architecture are needed to serve the needs of these markets.

2. PROCESS EFFECTS

2.1 Non Linear Resistance (NLR)

NLR as a function of width first appeared in the form of a local density based thickness adjustment. In earlier processes, e.g., 250nm, the impact was primarily wide wires. Wire slotting and guard banding was the generic solution used. With the introduction of copper in 130nm, nonlinear resistance due to liners, dishing and sidewall scattering became significant for all interconnects.[1] This has required parasitic extraction tool upgrades.

2.2 Selective Process Bias (SPB)

Simple modeling of width variation with global width reduction became inaccurate in 130nm with local density effects. Adjacent metal space dependant width, thickness and conductor edge angle. 130 nm used a simple 1-D per-edge nearest-neighbor model for this effect. [1] For sub-90nm processes a single adjacency adjustment is insufficient in some cases.

2.5 Narrow Width Effects (NWE)

Transistors with near minimum widths need to be used with care at 90nm technology node. Poly and diffusion turn patterning issues later are worse for narrow-width transistors in terms of relative impact. Random dopant fluctuations and L control are more severe in small transistors. The min-width 1-sigma Vt variation in 90nm is almost 40% higher than it was in 130nm. Narrow width transistors have poor drive strength and are more susceptible to antenna charging and SER events (in sequential-logic elements, particularly in reduced voltage scenarios). Vt roll off at near minimum-width transistors can cause mismatches between simulation and silicon, even if the effect is modeled

2.3 Length Control

Transistor length variation on die can cause significant leakage and I-drive variation. The impact of on-die variation in timing must be minimized and remaining uncertainty handled with margin tools. The cumulative path impact of these effects on margin is becoming an intolerable percentage of performance entitlement. However within die variation of leakage is more difficult to design with and manage across the full range of process seen in high volume devices.

3. EXTRACTION

For many years, standard-cell intrinsic performance has been dominated by intra-cell capacitance from various sources. With continued scaling other effects have had to be introduced to extraction to maintain cell level timing accuracy.

3.1 Resistance

Poly, contact, active and metal resistances have been gradually introduced starting with 250nm. In 90 nm, min R via and max R

via corners are added. NLR and SPB aware extraction is now used for all 90nm cell extraction.

3.2 Capacitance

Gate overlap capacitance, contact-to-gate capacitance, conformal dielectrics, multi-layer dielectrics, poly topology and shallow trench isolation (STI) have all driven the need for increasingly more complex cell extraction. Some effects that may have been included the spice model in the past now need to be dealt with by extended spice models (e.g. BSIM4 and new STI-stress models), or are handled within the extractor itself.

3.3 Field Solver

Tiled cells such as RAM bit cells or data path elements require field-solver quality extraction. The cumulative tiling error in C and R in fast extraction modes is unacceptable because the geometry being extracted is often so small that it is in the error bound of these extraction modes. In 65nm, all standard cells may be extracted using full field solvers.

3.4 Context emulation

Library cells are often extracted and characterized in an abstract context. Statistical context emulation is important to avoid adding large uncertainty to chip level margins. The table below illustrates some of the emulation modes available to designers. Designers may use different context emulation for different corners. For example fast corner may use min density metal, no through cell routes, a slow corner may use full dummy poly/active insertion, high density metal over the cell route.

Table 1. Context emulation roadmap

	250nm	180nm	130nm	90nm
Metal over cell	Y	Y	Y	Y
Metal thru cell	N	Simple	Y	Y
Lateral active/poly	Y	Y	Stat	Stat
Corner splits	N	Y	Y	Y
Dummy insertion	Simple	Simple	Exact	Exact
Port routing	N	Y	Y	Y
Misalignment	N	N	Manual	Y
Context coupling	N	N	Simple	Y

However this has limitations such as within-die placement variation, route and fill density variation. True in-context extraction is still a requirement for extreme performance.

3.5 Fill geometry

In the good old days, fill geometries were automatically inserted at tape out. The impact on cell or design timing was handled via statistical dummy models used in extraction or handled in timing margins. In some extreme performance examples, the tape out dummy was inserted into the database prior to extraction. Tight density ranges now needed to reduce in-die thickness variation have forced the use of finer grained fill. This fill may also be closer to actual metal and has driven a need for better dummy metal modeling and extraction.

3.6 OPC and Process model

Completed cells are run through OPC, pseudo tape out and full process modeling including depth of field (DOF), exposure

variance and misalignment exploration. The results are auto analyzed to identify sensitive devices [4]. Critical active, poly and gate areas undergo further analysis. The data also goes through a manual process/designer group review. Corrective action involves re-layout, circuit changes and additional design rules.

4. SIMULATION

4.1 Statistical simulation

Except in very simple cases exact correlation between transistor strong (weak) corner and circuit best (worst) case may not occur . Of particular concern is the case where circuit performance corners are beyond the predicted transistor corners. Some reasons for this miscorrelation include: stronger transistors may present larger input loads, strong transmission gates may present more capacitance than they provide in reduced channel resistance, and internal mismatch may cause setup/hold changes.

It is no longer sufficient to characterize the library at the strong and weak transistor corners alone. The results of characterization must be checked against statistical simulations to ensure robust functionality across the entire process spread. The results are also used to guide top level design margins. Some circuits may need redesign or usage constraints to avoid problems. This analysis is of particular interest across extreme ranges of voltage operation.
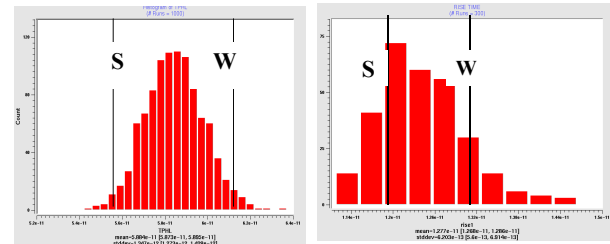

Figure 1: Statistical Simulation, Circuit with (a) acceptable (b) unacceptable excursions outside characterization corners.

4.2 Corners

Traditional Slow/Fast corners are no longer sufficient for cell characterization. Cell designers may extract 26 corners or more for characterization. Transistor, Metal, Vias, Temp are variables for fixed corner extraction. Statistical interconnect extraction and statistical simulation are becoming critical.
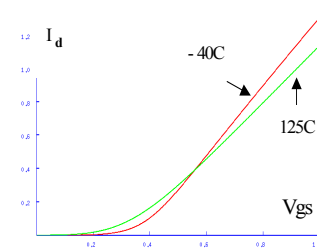

Figure 2: Drain current crossover

One new corner example is illustrated with figure 2. This shows the variation of the MOS drain current with gate-to-source voltage (Vgs) at two different temperatures (assuming constant Vds). At low values of Vgs the transistor has higher drive current at 125C than at –40C. This can cause the timing at lower temperature to be worse than higher temperature. This means that the setup

violations need to be checked at two corners: strong, low-voltage, high-temp and strong, low-voltage, low-temp. Similarly, hold violations need to be checked at two corners: strong, low-voltage, high-temp and strong, low-voltage, low-temp.

4.3 NBTI and EOL models

Transistor performance degrades over time from a variety of effects. The designer has to ensure that the degradation does not cause functional or performance failures during of after the degradation. This includes differential degradation with is particularly difficult to deal with. Commercial tools are becoming available to allow simulation of these effects. At 90nm technology Negative Bias Temperature Instability (NBTI) [9] is expected to cause significant P-transistor performance degradation over device-lifetime. It degrades the threshold voltage (Vt) as well as the drive current (Idrive) of the P transistors. This causes larger delays and a higher minimum voltage for circuit operation (Vmin). The degradation increases with the value and the duration of negative bias across the PMOS gate, and the temperature. In circuits like PLL the degradation over lifetime can be calculated for each transistor and fed-back into circuit-simulations to ensure the lifetime. In digital-logic circuits, since the exact degradation cannot be estimated for each transistor, cells are characterized with "end-of-life" PMOS models, and timing analysis is done at this additional corner.

4.4 Miller effect

Traditionally the capacitance seen at the input of a CMOS standard cell has been considered independent of all factors external to the cell. The increasing drain-to-gate capacitance has changed this. It acts like a Miller capacitance making the effective input capacitance a function of the load capacitance and (to some extent) the transition time.
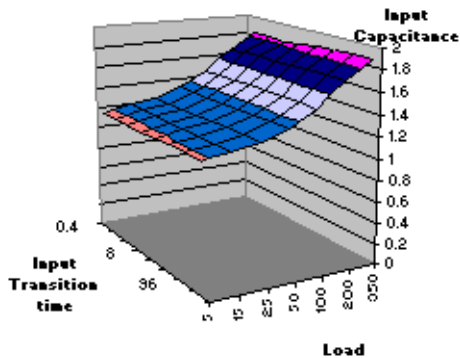

Figure 3. Miller capacitance

5. PHYSICAL ARCHITECTURE

5.1 New effects in Architecture Selection

Physical architecture selection has also been impacted by scaling trends in deep-submicron design. Selection of track height and routing pitch are now more complicated and heavily influenced by several effects. Nominal routing pitch is impacted by higher clock rates and small wire cross-sections. Noise effects in deep-submicron impact route width and space selection. Intra-cell parasitics have an increasing effect on total-path performance.

5.2 Cell height selection

The importance of intra-cell parasitics and topology are becoming increasingly significant to system-level performance targets in scaled libraries. Gate cross-sections, high-density lowest-level metal and contact cross sections all continue to decrease. The end result is that intra-cell resistances have become much more significant. One consequence of increased intra-cell R effects is that cell height starts to impact optimal repeater delay. In figure 3 we consider various track heights, transistor sizes, wire width and space, etc. and have plotted the optimal delay as a function of total wire length for various track height cells. Increasing track heights reduces total resistance and improves performance. The decision to use double or single row cells and the importance of repeater delay versus cell delay are more decision points for analysis.
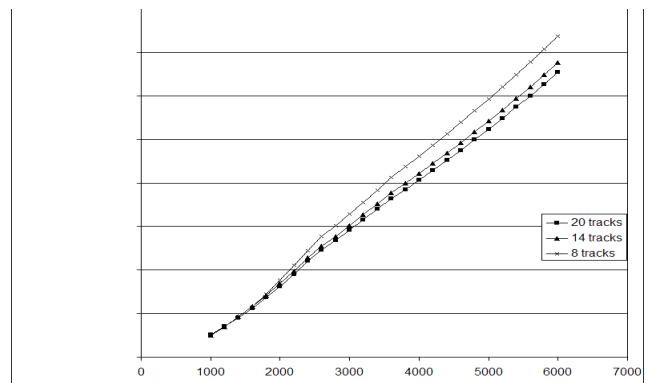

Figure 3. Optimal repeater distance

5.3 Power rail selection

Continual scaling of metal cross sections has also impacted design of power supplies in standard cells. Traditional standard cells distribute power on large metal-1 straps running at top and bottom of cells. Some recent libraries still use this architecture. However, there are now many libraries which distribute power on higher-level lower-resistance metals.

Often there are now two or more power rails in libraries, e.g., one 'always on' power supply for state retention and another 'active on' supply for the remainder of logic which can be powered down. Some styles may have header and or footer switches in the cells. Some cells may have additional power supplies for body bias control. Body bias control may allow increased performance and/or reduced leakage consumption depending on process corner.

5.4 Universal architecture

In order to allow designers to mix different cell types easily in the same design, a universal architecture may be used that allows cells to be mixed at a row level or block level. This allows header cells to be used next to footer cells, back bias, retention, high-Vt, low-Vt etc. This flexibility needs specific flow support.

6. CELL SELECTION AND DESIGN

Cell selection is heavily influenced by the intended tool stream. Different tools have 'interesting' behavior in function selection,

design optimization and logic mapping. An automated tool exercise flow (Crank) is an integral part of library architecture. Crank runs various benchmark circuits through the selected tools with different constraints, tool switches and strategies. Genetic Algorithm techniques may be used to optimize performance, area, power strategy, tool settings, versions and library content.

6.1 Power management

The need for power management is strongly reflected in flow and libraries at 90nm. As channel lengths shrink, leakage is increasing sharply. Leakage current is no longer predominantly the traditional sub-threshold current. Major components of leakage include Gate Induced Drain Leakage (GIDL), Gate enhanced drain leakage (GEDL), and the Gate leakage i.e. the leakage across the gate oxide.[10] In 90nm technology the Gate leakage is of the same order as the sub-threshold drain current, and occurs in ON transistors as well as OFF transistors. At the same time dynamic power is increasing the effective power density and driving power management for traditional 'power no issue' high performance designers.
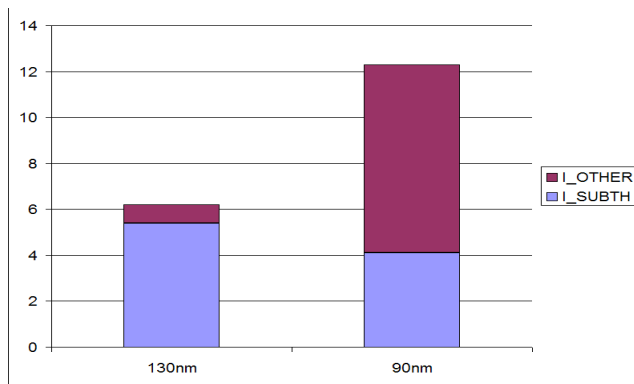

Figure 4. Leakage breakdown Strong, 25C

In 90nm technologies and below the transistor Vt is not scaling as it did in previous technologies. Coupled with demands for high performance, supply-voltage scaling is screeching to a near halt at about 1V. There are no longer power-saving benefits from static voltage scaling.

6.2 Leakage and Multi-Vt

Increasing leakage currents have also forced low power library designers to augment traditional libraries with new design elements. One very common feature found in recent libraries is high-threshold (HVT) transistors.

High-Vt transistors reduce leakage currents at the sacrifice of performance. Depending on process design rules, cells can be designed which intermix HVT and low-Vt (LVT) transistors or cells may be either HVT or LVT. As a consequence of dual and multi-Vt processes, library sizes have also increased typically doubling or more in size for a 3-Vt library.

Most timing paths in a design (say 75%) have small path delays and can benefit from HVT cells. Of the remainder, most can still benefit from some HVT cells, and then there is a small section of paths (say 5%) which exclusively require LVT cells.

In technologies where multiple-Vt is not feasible one possibility is to use different gate lengths. This is conceptually similar to multiple Vt, though it can be complicated and the results may be

less optimal. Non-monotonic Vt roll-off may lead to greater leakage than expected unless care is taken in design.

6.3 Clock gating

Since voltage scaling has slowed, other mechanisms are used to bring down the active switching power. Clock gating is almost mandatory in all power-conscious designs in 90nm. This requires several clock gating cells. Typically these cells need to have equal rise and fall delays to maintain clock duty cycle.

6.4 Cells for supply voltage variation

Multiple voltages can be used in different ways to reduce leakage while maximizing performance. One way is to create different design blocks working at different voltages based on the performance requirements of each block. Another is to have storage elements that retain data using an auxiliary supply while the main supply is turned off whenever the circuit is in standby mode. Yet another approach is to lower the power supply when the circuit is not active. In general the special cells created in 90nm libraries for such schemes are – 1. Level shifters: For interfacing logic at two different voltage levels. 2. Data Retention flops: Flops that hold the state while the main power supply is switched off. 3. Special scan flops for efficient scan dumping of state or setting special leakage reduction logic states. 4. SER hardened flops: To ensure operation at low operation/retention voltages

6.5 Cells for high performance

To stay on the expected performance curve, ASIC libraries will have to include faster but higher risk cells that have not normally been used in ASIC designs. Examples are unbuffered transmission gates, 1-hot muxes, alternate flop styles like pulsed flop, and dynamic flop. Special characterization and enhancement to design flows to control and check use of these cells is critical to avoid problems.

6.6 Cells for Density

At 90nm designers are forced to think of new layout architectures to regain density. This includes different cell-heights, power bus architectures, use of higher-level metal in cells and power rails, positioning of the substrate connections etc. Library composition has also changed. Cells with more complex functionality, multi-bit logic cells, and functions integrated into flops are 'must have' for density recovery.

6.7 Cells for low cost

Costs in terms of masks, cycle time have driven the use of Gate Array style cells for some time. These may be in blocks for later configuration or scattered as filler for ECO purposes. Some market segments may be able to use silicon platform style design to further reduce costs. Platforms with mixed GA/ SOC were quite prevalent in the early 90's. Configurable logic and interconnect techniques offer a myriad of possibilities across the density, performance, mask count and cycle time spectrum.

7. LAYOUT RULES

The complexity of layout rules from 250 to 130 to 90nm has been increasing. This is reflected in increasing runtime and run deck complexity.

Table 2. Layout rule complexity increase

Process(nm)	250	180	130	90	65
Code lines	9.5k	10k	13K	18k+	24k growing
Run time	1x	1.1	1.6	2.2x	2.3x+

7.1 Metal bin rules

Metal spacing bins (width dependant spacing and via overlap) have been in use for several years, primarily wide wire driven. However recent nodes drove a significant increase in the number of different bins. These bins are photo driven and affect almost all cell layouts and auto routing. Complex context and router emulation design checks along with chip router enhancements.

7.2 Min area / hole

Minimum area rules have been an increasing challenge for scaling, leading to new tool requirements for via stacks and a requirement to oversize some cell ports if via route connections are used.

7.3 Phase generation

130nm used simple phase rule emulation. In 90nm and below, all cells are run through full phase generation and assignment to ensure clean layout.[5] This provides flexibility in phase strategy for library reuse with different fabrication locations.

7.4 Density

There are increasing restrictions on density ranges for metal, active, poly and vias. Cell designers use tools to insert intentional and/or auto dummy and then run cell density checks. Only then will the cell be placed in the library. Density checks are repeated through the design flow to tape out.

7.5 Via rules

There are several different sets of rules that have to be aligned with each other. Most of these rules are a challenge for chip creation with limited impact in small cell architectures. Asymmetric metal overlap of vias and in-direction of overlap, are comprehended by, most tools. Via stress migration rules, Via cluster rules, Via array rules. Max via spacing. Redundant vias , and dummy via insertion have 'evolving' support.

7.6 Gate shape control rules

The most aggressive OPC, SRAF techniques are limited in the amount of correction they can apply. These OPC limitations are captured in extra drawn layout rules to force creation of OPC friendly layout. One example is in control of poly and active turns near critical gates. There can be significant L variation close to the gate ends. Additional design rule checks are needed to ensure OPC friendly layouts are created. Inspection friendly layout is also a new requirement.

7.7 Different

Many rules are simply different and represent a challenge for tools, flows and learning curve for personnel. Examples include exact gate overlap, granular gate width, Max via space, non contiguous pitch rules, gate orientation, decreasing metal density ranges, noise control rules and high voltage spacing.

# 8. CHIP LEVEL ISSUES

8.1 Integration

Sub-100 nm CMOS technologies offer the capabilities to create true systems-on-chip (SoCs). However, major challenges remain in developing design tools, flows and methodologies that can be used to create differentiated SoCs in a short design cycle-time. Key among these challenges are IP reuse, import, analog integration, RF and power management subsystems.

8.2 Data

The volume of data required is straining everything from OS, disk, memory, to network capacity. In particular post-OPC database size if it is fractured flat and transmitted to Mask shops.

8.3 Crosstalk

With technology scaling, the spacing between wires has shrunk faster than the thickness and has lead to increased dominance of coupling capacitance. The distribution of percentage coupling capacitance across nets on the same design in 90nm and 130nm technology is shown in figure 5, which indicates the continuing dominance of coupling capacitance in 90nm.
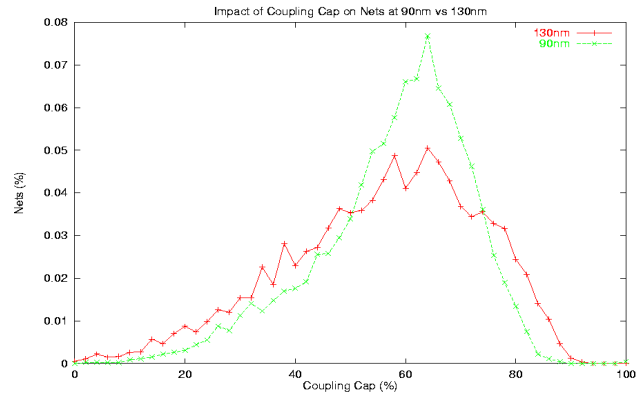

Figure 5: Distribution of coupling capacitance in 130nm and 90nm

This coupling capacitance can induce noise on a silent victim or cause speedup/slowdown of signals when the aggressor and victim switch together. Noise threshold have been decreasing with technology/power supply scaling, making designs even more crosstalk noise prone. Crosstalk also causes signals to rise above power supply (overshoot noise) leading to more gate oxide integrity issues.

Figure 6 shows the percentage stage delay change due to coupling capacitance in 90nm. A large percentage of nets have significant delay change. It is essential to capture this delay change during STA. However current flows are more focused on addressing setup/hold violations with all coupled aggressors silent. Crosstalk based timing analysis represents silicon timing closer than the conventional lumped timing analysis and needs to become *the* timing closure approach for 90nm designs.

A common method used today for obtaining switching window information for crosstalk analysis, is STA arrival windows. With increased on-chip variation in 90nm, computation of accurate switching windows and bounded crosstalk analysis is even more difficult.
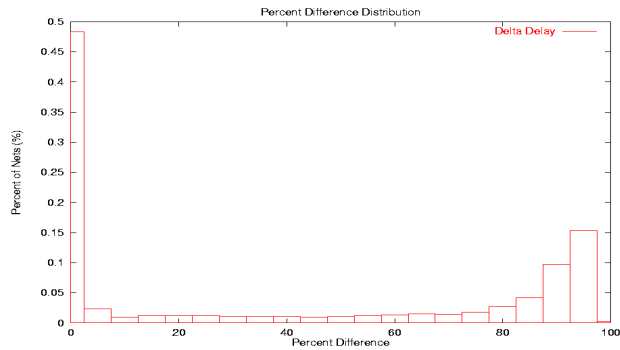
Figure 6: Percentage delay change due to crosstalk in 90nm design

Crosstalk needs to be considered as early as placement to ensure tight slew control and during global route phase to ensure adjacent nets do not switch together.

8.4 Electromigration(EM)

Conventional approaches to ensure EM safety are based on keeping Average/RMS current density within limits on every metal lead. As shown in figure 7, current density in min width leads has been rapidly increasing with technologies. The shift to copper did increase the EM current density threshold. However in 90nm a min-width wire carrying the same current is closer to violating current density thresholds than at 130nm.

The large number of violations in 90nm designs indicates that EM issues can no longer be handled by checking leads in isolation. Newer approaches are needed for EM analysis on 90nm designs which takes a system reliability perspective.
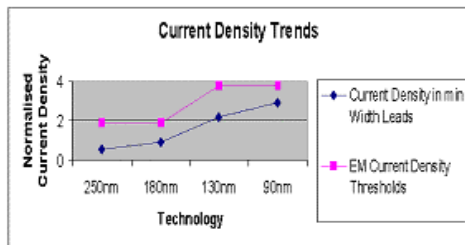
Figure 7: Current density trends across technologies

8.5 On die variation (OCV), CTS

On-chip variation effects are increasing as processes continue to scale. Handling these as uncertainties in margins is no longer acceptable. True statistical extraction, simulation and STA[7,8] are needed to set up static corners or for actual chip design. On-chip variation impacts the entire design methodology. One area of particular importance is clock-tree synthesis (CTS). Modern clock-tree design practices include; Equalizing gate delay and interconnect delay to all CTS endpoints to minimize sensitivity to non-symmetric variation; CTS buffer design to reduce variation, e.g., dummy poly for line width variation reduction and uniformity in layout; Minimization of CTS insertion delay to reduce the impact of on-chip variation.

Timing signoff flows have also evolved over time to improve handling of on-chip variation. One simple approach is use of a fixed percent variation for gate or wire delay. Another approach

commonly used is to characterize at some given process corner and at a corner 'close to' the target corner.

9. CONCLUSIONS

In reference to the original question, the authors believe that 90 nm ASIC is not dead. This statement would not be possible without extensive changes in almost every aspect of process design, layout, extraction, characterization, modeling, margins and of course chip design. Creating a 90nm library was more difficult than 130nm by an order of magnitude. There are new and exciting challenges for 65nm that will also be overcome by creative engineers and new tools. The gaps between OPC, lithography, etch performance and geometric requirements are becoming so large that more restricted or structured design may be inevitable although there are promising alternatives. We expect the continuing struggle of uncertainty management - constraints to reduce uncertainty and tools to model the effects. Increasing mask costs is also driving towards platform solutions, structured fabrics and heavily constrained layout for low volume markets. However Standard Cell/SOC still provides clear performance and density benefits that will still drive high end markets through 65nm.

10. ACKNOWLEDGMENTS

11. REFERENCES

[1] Nagaraj N.S. et al. Benchmarks for Interconnect Parasitic Resistance and Capacitance ISQED 2003

[2] Asenov et al. Increase in the Random Dopant Induced Threshold Fluctuations and Lowering in sub-100nm mosfets…. IEEE Trans on Electron Devices Apr 2001

[3] Diaz C.H. Process and Circuit Design Interlock for Application-Dependant Scaling Tradeoffs and Optimization in the SOC Era. IEEE JSSC (March 2003 vol. 38), 444-449.

[4] O Brien S., Aton T. Mason M. Vickery C. Randall J. OPC on Real World Circuitry Proc. Proc. SPIE, MM02 vol. 5042 2003

[5] Sanie M. et al. Practical Application of Full-Feature Alternating Phase shifting Technology for a Phase aware Std Cell design flow. Proc. DAC 2001.

[6] Kocher M and Rappitsch G. Statistical Methods for the Determination of process Corners Proc. ISQED 2002

[7] Tsukiyama et al. A statistical STA considering correlations between delays. Proc DAC 2001

[8] H.F. Jyu, Malik S et al. Statistical timing analysis of combinatorial logic circuits IEEE trans. VLSI Systems vol. 1, no 2 1993

[9] N. Kimizuka et. al. "The Impact of Bias temperature instability …" Symposium on VLSI technology 1999, 73-74

[10] Keshavarzi, A. et. al., Intrinsic leakage in low power deep submicron CMOS ICs Test Conference, 1997. Proceedings., International , 1-6 Nov 1997, 146 -155